Automatisierte bild- und videobasierte Mimikanalyse für die Messung von Schmerzen und Facial Action Units

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

von **Dipl.-Ing.-Inf. Philipp Werner**
geb. am 25. August 1986 in Dessau

genehmigt durch die Fakultät für Elektrotechnik und Informationstechnik
der Otto-von-Guericke-Universität Magdeburg

Gutachter:
Prof. Dr. Ayoub Al-Hamadi
Prof. Dr. Klaus-Dietz Tönnies
Prof. Dr. Harald C. Traue

Promotionskolloquium am 24. Mai 2022

Zusammenfassung

Schmerzen sind eine persönliche Erfahrung und lassen sich nur indirekt messen. Alle bisher bekannten Maße haben ihre Schwächen, was oft zur suboptimalen Behandlung von Patienten beiträgt. Die vorliegende Arbeit schlägt neue Methoden zur objektiven und automatisierten Messung von akuten Schmerzen vor. Diese könnten insbesondere Patienten zugutekommen, die sich nicht selbst zu ihren Schmerzen äußern können, wie Säuglingen, bewusstlosen Patienten oder Patienten mit schwerer Demenz. Die vorgeschlagenen Methoden erfassen und interpretieren Schmerzreaktionen im Gesicht mit Hilfe von digitalen Kameras und künstlicher Intelligenz. Primärer Forschungsgegenstand der Arbeit ist die bild- und videobasierte Erkennung und Messung von Mimik, zum einen von Schmerzmimik, zum anderen von Action Units (AUs) nach dem Facial Action Coding System (FACS), die zur Beschreibung von beliebigen Gesichtsausdrücken dienen. Dies geschieht im Kontext verschiedener Herausforderungen, von denen insbesondere nicht-frontale Kopfposen, die begrenzte Verfügbarkeit von Daten, die Charakteristik von Schmerzen, sowie das Ziel der kostengünstigen und einfachen Anwendbarkeit adressiert werden. Als maschinelle Lernverfahren werden Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Random Forests, Transferlernen und Multi-Task-Lernen angewendet und für die Problemstellungen adaptiert. Vorgeschlagen und evaluiert werden unter anderem auch verschiedene neue Verfahren zur Merkmalsextraktion aus Einzelbildern sowie zur zeitlichen Integration von Bildinformationen für die videobasierte Erkennung. Zusätzlich werden mit Methoden der Computergrafik Datensätze generiert, die beim maschinellen Lernen zum Einsatz kommen. Vergleiche mit Methoden anderer Autoren und mit der Leistungsfähigkeit von Menschen zeigen die Nützlichkeit der vorgeschlagenen Methoden.

Pain is a personal experience and can only be measured indirectly. All available measures have their weaknesses, which often contribute to suboptimal treatment of patients. This work propses new methods for measuring acute pain objectively and automatically. These may be beneficial especially for patients, who cannot utter on their pain experience, such as newborns, unconcious patients, or patients with severe dementia. The proposed methods capture and interpret facial pain reactions with digital cameras and artificial intelligence. The research addresses image and video based recognition of facial expression, namely of facial expression of pain and of Action Units (AUs) as defined in the Facial Action Coding System (FACS), which can be used to describe any facial expression. The recognition is associated with different challenges, of which the work especially addresses: non-frontal head poses, limited available data, charactertistics of pain, and the goal to develop an afforable and easy-to-use technology. Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Random Forests, transfer learing, and multitask learning are used and adapted for handling the problems. Among others, the work proposes and evaluates new methods for extracting features from single images and for temporal integration of image information for video-based recognition. Additionally, computer graphics is used for generating machine-learning datasets. The usefulness of the proposed methods is shown by comparing them with methods of other authors and with human performance.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungen und Notation

Abkürzungen

Bei Erwähnung von Personen, wie z. B. Probanden und Patienten, wird im Interesse der leichteren Lesbarkeit das generische Maskulinum verwendet, d. h. es sind immer Personen aller Geschlechter gemeint.

2D	2-dimensional, verwendet im Zusammenhang mit Bildern, die reale oder virtuelle 3D-Räume/Szenen abbilden
3D	3-dimensional, verwendet im Zusammenhang mit realen oder virtuellen Räumen/Szenen, die in 2D-Bildern abgebildet sind
Abb.	Abbildung
AU	Action Unit, „Einheit“ der Gesichtsbewegung nach FACS [EFH02]
bzw.	beziehungsweise
d. h.	das heißt
CNN	Convolutional Neural Network
dim.	dimensional
FACS	Facial Action Coding System [EFH02]
ICC	Intraclass Correlation [SF79]
MN-R / MN-K	MobileNetV3-large CNN [How+19] zur Regression / Klassifikation
MW, Mittel, Mittelwert	arithmetischer Mittelwert
LBP	Local Binary Pattern Merkmale [AHP06]
OPR	Observer Pain Intensity Rating [Luc+11b]
PSPI	Prkachin and Solomon Pain Intensity [Luc+11b]
RF / RF-R / RF-K	Random Forest / Random Forest Regression / Random Forest Klassifikation
SVM / SVM-E	Support Vector Machine / Support Vector Machine Ensemble (Klassifikation)
SVR / SVR-E	Support Vector Regression / Support Vector Regression Ensemble
VAS	Visuelle Analogskala
vgl.	vergleiche
vs.	versus (gegenüber, im Vergleich zu)
z. B.	zum Beispiel

Notation

$x, X, i, j, n, N, \alpha, \lambda, \ldots$	Skalar (Notation kursiv)
$\mathbf{x}, \mathbf{y}, \mathbf{p}, \mathbf{l}, \ldots$	Vektor (Notation fett, meist kleiner Buchstabe)
$\mathbf{W}, \mathbf{F}, \mathbf{K}, \ldots$	Matrix (Notation fett, großer Buchstabe)

$\mathbb{R}/\mathbb{Z}/\mathbb{N}$	Menge der reellen / ganzen / natürlichen Zahlen
$\mathbb{E}(X)$	Erwartungswert der Zufallsvariable X
$\lceil a \rceil$	Aufrundung von a auf die nächstgrößere Ganzzahl: $\lceil a \rceil = \min\{x \in \mathbb{Z}, x \geq a\}$
$\lfloor a \rfloor$	Abrundung von a auf die nächstkleinere Ganzzahl: $\lfloor a \rfloor = \max\{x \in \mathbb{Z}, x \leq a\}$
$\max\{\dots\}$	Größte Zahl der angegebenen Menge
$\min\{\dots\}$	Kleinste Zahl der angegebenen Menge

Veröffentlichungen

Folgende Publikationen des Autors bilden die Grundlage dieser Dissertation bzw. stehen in inhaltlich engem Zusammenhang. Die Liste ist unterteilt in Artikel, bei denen der Autor als Erstautor die zentralen Ideen und den Hauptteil der Arbeit beigetragen hat, und in weitere Artikel, in denen als Koautor mitgewirkt wurde.

Alle Artikel wurden vor der Veröffentlichung in einem Peer-Review-Prozess begutachtet. Der angegebene Impact Factor stammt aus den Journal Citation Reports des Jahres 2020 von Clarivate Analytics und basiert auf Zitationsdaten der Web of Science Core Collection. Die Acceptance Rate einer Konferenz gibt an, wie viel Prozent der zur Begutachtung eingereichten Artikel zur Präsentation auf der Konferenz und Veröffentlichung im Konferenzband akzeptiert wurden.

Die vollständige Publikationsliste des Autors sowie von Google Scholar ermittelte Zitationen können über `http://philipp-werner.info/` abgerufen werden.

Journalartikel mit Erstautorenschaft

1. Ph. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, R. W. Picard. **„Automatic Recognition Methods Supporting Pain Assessment: A Survey“**. In: *IEEE Transactions on Affective Computing*, 2019, DOI: 10.1109/TAFFC.2019.2946774. **`Impact Factor: 10.506`**

2. Ph. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, H. C. Traue. **„Head movements and postures as pain behavior“**. In: *PLOS ONE* 13(2), 2018. **`Impact Factor: 3.240`**

3. Ph. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, H. C. Traue. **„Automatic Pain Assessment with Facial Activity Descriptors“**. In: *IEEE Transactions on Affective Computing* 8(3), 2017. **`Impact Factor: 10.506`**

4. Ph. Werner, A. Al-Hamadi, R. Niese. **„Comparative learning applied to intensity rating of facial expressions of pain“**. In: *Int. J. Pattern Recognit. Artif. Intell.*, 2014. **`Impact Factor: 1.373`**

Konferenzartikel mit Erstautorenschaft

5. Ph. Werner, F. Saxen, A. Al-Hamadi. **„Facial Action Unit Recognition in the Wild with Multi-Task CNN Self-Training for the EmotioNet Challenge“**. In: *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. **`Zweiter Platz in Challenge.`**

6. Ph. Werner, A. Al-Hamadi, S. Gruss, S. Walter. **„Twofold-Multimodal Pain Recognition with the X-ITE Pain Database“**. In: *International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2019.

7. Ph. Werner, F. Saxen, A. Al-Hamadi, H. Yu. **„Generalizing to Unseen Head Poses in Facial Expression Recognition and Action Unit Intensity Estimation“**. In: *International Conference on Automatic Face and Gesture Recognition (FG)*, 2019. **Acceptance Rate: 40%**

8. Ph. Werner, S. Handrich, A. Al-Hamadi. **„Facial Action Unit Intensity Estimation and Feature Relevance Visualization with Random Regression Forests“**. In: *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017.

9. Ph. Werner, A. Al-Hamadi, S. Walter. **„Analysis of Facial Expressiveness During Experimentally Induced Heat Pain“**. In: *International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2017.

10. Ph. Werner, F. Saxen, A. Al-Hamadi. **„Landmark based Head Pose Estimation Benchmark and Method“**. In: *International Conference on Image Processing (ICIP)*, 2017. **Acceptance Rate: 45%**

11. Ph. Werner, F. Saxen, A. Al-Hamadi. **„Handling Data Imbalance in Automatic Facial Action Intensity Estimation“**. In: *British Machine Vision Conference (BMVC)*, 2015. **Acceptance Rate: 33%**

12. Ph. Werner, A. Al-Hamadi, S. Walter, S. Gruss, H. C. Traue. **„Automatic Heart Rate Estimation from Painful Faces“**. In: *International Conference on Image Processing (ICIP)*, 2014. **Acceptance Rate: 44%**

13. Ph. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, H. C. Traue. **„Automatic Pain Recognition from Video and Biomedical Signals“**. In: *International Conference on Pattern Recognition (ICPR)*, 2014.

14. Ph. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, H. C. Traue. **„Towards Pain Monitoring: Facial Expression, Head Pose, a new Database, an Automatic System and Remaining Challenges“**. In: *British Machine Vision Conference (BMVC)*, 2013. **Acceptance Rate: 30%**

15. Ph. Werner, A. Al-Hamadi, R. Niese. **„Pain Recognition and Intensity Rating Based on Comparative Learning“**. In: *International Conference on Image Processing (ICIP)*, 2012. **Acceptance Rate: 39%**

Weitere Journalartikel

16. E. Othman, Ph. Werner, F. Saxen, A. Al-Hamadi, S. Gruss, S. Walter. **„Automatic vs. Human Recognition of Pain Intensity from Facial Expression on the X-ITE Pain Database“**. In: *Sensors* 21, no. 9: 3273, 2021. **Impact Factor: 3.576**

17. S. Handrich, L. Dinges, A. Al-Hamadi, Ph. Werner, F. Saxen, Z. Al Aghbari. **„Simultaneous Prediction of Valence / Arousal and Emotion Categories and its Application in an HRC Scenario“**. In: *Journal of Ambient Intelligence and Humanized Computing*, 2020. **Impact Factor: 7.104**

18. S. Frisch, Ph. Werner, A. Al-Hamadi, H. C. Traue, S. Gruss, S. Walter. **„Von der Fremdbeurteilung des Schmerzes zur automatisierten multimodalen Messung der Schmerzintensität - Narrativer Review zum Stand der Forschung und zur klinischen Perspektive“**. In: *Der Schmerz*, 2020. **Impact Factor: 1.107**

19. S. Walter, A. Al-Hamadi, S. Gruss, S. Frisch, H. C. Traue, Ph. Werner. **„Multimodale Erkennung von Schmerzintensität und -modalität mit maschinellen Lernverfahren“**. In: *Der Schmerz*, 2020. **Impact Factor: 1.107**

20. E. Othman, F. Saxen, D. Bershadskyy, Ph. Werner, A. Al-Hamadi, J. Weimann. **„Predicting Group Contribution Behaviour in a Public Goods Game from Face-to-Face Communication“**. In: *Sensors*, Juni 2019. **Impact Factor: 3.576**

21. M. Rapczynski, Ph. Werner, A. Al-Hamadi. **„Effects of Video Encoding on Camera Based Heart Rate Estimation“**. In: *IEEE Transactions on Biomedical Engineering*, März 2019. **Impact Factor: 4.538**

22. S. Gruss, M. Geiger, Ph. Werner, O. Wilhelm, A. Al-Hamadi, S. Walter. **„Multimodal Signals for Analyzing Pain Responses to Thermal and Electrical Stimuli“**. In: *Journal of Visualized Experiments* (146), 2019. **Impact Factor: 1.355**

23. M. Kächele, M. Amirian, P. Thiam, Ph. Werner, S. Walter, G. Palm, F. Schwenker. **„Adaptive confidence learning for the personalization of pain intensity estimation systems“**. In: *Evolving Systems*, 2016. **Impact Factor: 1.908**

24. K. Limbrecht-Ecklundt, Ph. Werner, H. C. Traue, A. Al-Hamadi, S. Walter. **„Mimische Aktivität differenzierter Schmerzintensitäten“**. In: *Der Schmerz* 30(3), 2016. **Impact Factor: 1.107**

25. S. Gruss, R. Treister, Ph. Werner, H. C. Traue, S. Crawcour, A. O. Andrade, S. Walter. **„Pain Intensity Recognition Rates via Biopotential Feature Patterns with Support Vector Machines“**. In: *PLOS ONE* 10(10), 2015. **Impact Factor: 3.240**

26. S. Walter, S. Gruss, K. Limbrecht-Ecklundt, H. C. Traue, Ph. Werner, A. Al-Hamadi, A. O. Andrade, N. Diniz, G. M. da Silva. **„Automatic pain quantification using autonomic parameters“**. In: *Psychology and Neuroscience*, November 2014.

Weitere Konferenzartikel

27. E. Othman, Ph. Werner, F. Saxen, A. Al-Hamadi, S. Walter. **„Regression Networks for Automatic Pain Intensity Recognition in Video using Facial Expression on the X-ITE Pain Database“**. In: *International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, 2021.

28. N. Pandya, Ph. Werner, A. Al-Hamadi. **„Deep Facial Expression Recognition with Occlusion Regularization“**. In: *International Symposium on Visual Computing (ISVC)*, 2020.

29. S. Handrich, L. Dinges, A. Al-Hamadi, Ph. Werner, Z. Al Aghbari. **„Simultaneous Prediction of Valence/Arousal and Emotions on AffectNet, Aff-Wild and AFEW-VA“**. In: *International Conference on Emerging Data and Industry 4.0 (EDI40)*, 2020.

30. F. Saxen, S. Handrich, Ph. Werner, E. Othman, A. Al-Hamadi. **„Detecting Arbitrarily Rotated Faces for Face Analysis“**. In: *International Conference on Image Processing (ICIP)*, 2019.

31. E. Othman, F. Saxen, Ph. Werner, A. Al-Hamadi, S. Walter. **„Cross-Database Evaluation of Pain Recognition from Facial Video“**. In: *International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2019.

32. F. Saxen, Ph. Werner, S. Handrich, E. Othman, L. Dinges, A. Al-Hamadi. **„Face Attribute Detection with MobileNetV2 and NasNet-Mobile“**. In: *International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2019.

33. M. Rapczynski, Ph. Werner, F. Saxen, A. Al-Hamadi. **„How the Region of Interest Impacts Contact Free Heart Rate Estimation Algorithms“**. In: *International Conference on Image Processing (ICIP)*, 2018.

34. F. Saxen, Ph. Werner, and A. Al-Hamadi. **„Real vs. Fake Emotion Challenge: Learning to Rank Authenticity From Facial Activity Descriptors“**. In: *International Conference on Computer Vision Workshops (ICCVW)*, 2017. `Gewinner der Challenge.`

35. M. Rapczynski, Ph. Werner, A. Al-Hamadi. **„Continuous Low Latency Heart Rate Estimation from Painful Faces in Real Time“**. In: *International Conference on Pattern Recognition (ICPR)*, 2016.

36. L. Zhang, S. Walter, X. Ma, Ph. Werner, A. Al-Hamadi, H. C. Traue, S. Gruss. **„BioVid Emo DB: A multimodal database for emotion analyses validated by subjective ratings“**. In: *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2016.

37. M. Kächele, P. Thiam, M. Amirian, Ph. Werner, S. Walter, F. Schwenker, G. Palm. **„Multimodal data fusion for person-independent, continuous estimation of pain intensity“**. In: *Engineering Applications of Neural Networks (EANN)*, 2015. `Acceptance Rate: 43%`

38. M. Kächele, Ph. Werner, A. Al-Hamadi, G. Palm, S. Walter, F. Schwenker. **„Bio-Visual Fusion for Person-Independent Recognition of Pain Intensity“**. In: *Multiple Classifier Systems (MCS)*, 2015.

39. S. Walter, S. Gruss, H. C. Traue, Ph. Werner, A. Al-Hamadi, M. Kächele, F. Schwenker, A. Andrade, G. Moreira. **„Data fusion for automated pain recognition“**. In: *International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, 2015.

40. R. Niese, Ph. Werner, A. Al-Hamadi. **„Accurate, Fast and Robust Realtime Face Pose Estimation Using Kinect Camera“**. In: *International Conference on Systems, Man, and Cybernetics (SMC)*, 2013.

41. S. Walter, Ph. Werner, S. Gruss, H. C. Traue, A. Al-Hamadi, et al. **„The BioVid Heat Pain Database: Data for the Advancement and Systematic Validation of an Automated Pain Recognition System“**. In: *International Conference on Cybernetics (CYBCONF)*, 2013.

1. Einleitung

Schmerzen erfüllen eine wichtige Schutzfunktion für den Körper. Gleichzeitig führen sie bei Betroffenen oft zu gravierenden Einschränkungen der Lebensqualität, in nicht wenigen Fällen durch die Chronifizierung von Schmerzen bis hin zur dauerhaften Veränderung des gesamten Lebensgefüges [Wer+19b]. Für die Gesellschaft entstehen Kosten in Milliardenhöhe für die Behandlung, durch Arbeitsunfähigkeit und vorzeitige Berentung [NZ05].

Schmerzen sind der häufigste Grund, der Menschen veranlasst sich in ärztliche Behandlung zu begeben [TM11a; Män+01]. Bei einer Studie von Cordell et al. waren Schmerzen bei 52,2% der Patienten einer Notaufnahme die Hauptbeschwerde, nur 34,1% der Patienten hatten keine Schmerzen [Cor+02; Wer+19b]. Nach Gregory et al. berichten etwa 50% der Patienten im Krankenhaus von akuten Schmerzen [GM16]. Zoëga et al. fanden sogar eine Schmerzprävalenz von 83% [Zoë+15]. In beiden Studien litten 35% der Krankenhauspatienten unter *starken* Schmerzen [GM16; Zoë+15].

Die Messung von empfundenen Schmerzen ist nötig für die Diagnostik, um die richtige Behandlung zu wählen, den Behandlungsfortschritt zu überwachen und um zu entscheiden, ob eine Behandlung fortgesetzt oder geändert werden sollte [Wer+19b]. Die Schmerztherapie zielt nicht nur darauf ab, Leid zu mindern, sondern auch direkte und langfristige negative Konsequenzen zu für die Gesundheit des Patienten zu vermeiden, z. B. Schädigungen des Nervensystems, des Hormonsystems und des Immunsystems, die durch unbehandelte Schmerzen entstehen können [Wer+19b; Lyn11]. Eine adäquate Schmerztherapie trägt z. B. nach Operationen dazu bei, die Genesung zu beschleunigen, die Sterbefälle zu reduzieren und die Wahrscheinlichkeit für die Chronifizierung der postoperativen Schmerzen zu senken [JO05]. Bei der Behandlung kommt es jedoch auf das richtige Maß an, insbesondere bei Medikamenten. Der übermäßiger Gebrauch von Opioiden kann zu Atemdepression und Abhängigkeit führen, andere Medikamente können Nebenwirkungen wie Übelkeit, Erbrechen oder Verstopfung hervorrufen [Wer+19b; MMJ97].

Die Standardmethode zur Messung von Schmerzen ist die Selbsteinschätzung. Einige Patienten sind jedoch kognitiv oder sprachlich nicht in der Lage, sich über ihre Schmerzen zu äußern, z. B. Säuglinge, bewusstlose Patienten oder Menschen mit fortgeschrittener Demenz. Diese Patienten sind gefährdet keine angemessene Schmerzbehandlung zu erhalten. Zwar gibt es Skalen, mit denen ein Beobachter die Schmerzen eines Leidenden bewerten kann, jedoch ist diese Einschätzung subjektiv, d. h. abhängig vom Beobachter, und die Anwendung ist mit hohem Zeitaufwand verbunden, der im Klinikalltag mit Personalmangel und Zeitdruck schwer aufzubringen ist.

Hier setzt die vorliegende Arbeit an, die ein automatisiertes System zur Erkennung und Messung von akuten Schmerzen anhand der Mimik vorschlägt. Dieses filmt die Bewegungen im Gesicht der leidenden Person mit einer oder mehreren Kameras und wertet die Bilder automatisiert mit Hilfe von Computer-Vision-Methoden aus, um Schmerzreaktionen zu erkennen und von diesen auf die empfundenen Schmerzen zu schließen. Ein solches System könnte ein objektives und permanentes Monitoring der Schmerzen ermöglichen und zu Verbesserungen bei der Schmerzbehandlung beitragen, indem das medizinische Personal gleichzeitig unterstützt und entlastet wird, sowie zusätzliche und potentiell zuverlässigere Informationen über die Schmerzen bereitgestellt werden.

Bevor in Abschnitt 1.3 detaillierter auf die Zielstellung der Arbeit eingegangen wird, werden zunächst die nötigen Grundlagen bezüglich Schmerzen sowie der automatisierten Auswertung von Bildern dargelegt.

1.1. Schmerzen

Im Folgenden wird das Phänomen Schmerzen definiert und einige Eigenschaften beschrieben. Anschließend werden die Methoden der aktuellen klinischen Schmerzmessung thematisiert. In beiden Teilen wird auch auf die Herausforderungen für die Forschung und Praxis eingegangen. Die Ausführungen basieren in weiten Teilen auf Textabschnitten, die der Autor bereits in einem Überblicksartikel zur automatisierten Schmerzerkennung veröffentlicht hat [Wer+19b].

1.1.1. Definition und Charakteristik

Die weithin akzeptierte Definition von Schmerz stammt von der *International Association for the Study of Pain* (IASP):

> Pain is "an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage." [Raj+20]
>
> Übersetzung: Schmerz ist ein unangenehmes Sinnes- oder Gefühlserlebnis, das mit tatsächlicher oder potenzieller Gewebeschädigung einhergeht oder einen ähnlichen Eindruck vermittelt.

Schmerz ist eine persönliche Erfahrung und immer subjektiv. Die Bedeutung des Wortes wird von jedem Individuum anhand von eigenen Erfahrungen gelernt, beginnend mit dem Erleben von Verletzungen in frühen Lebensphasen [Mer79; Int21].

Schmerzen helfen uns Lebewesen, gefährliche Situationen zu erkennen, Gewebeschädigung zu vermeiden, und sie begünstigen die Heilung indem sie uns an Aktivitäten hindern, die weitere Gewebeschädigungen verursachen könnten [Wil02]. *Akute Schmerzen* erfüllen normalerweise diese Funktion und verschwinden mit der Heilung. Schmerzen nennt man *chronisch*, wenn sie länger andauern (meist wird von länger als 3 Monaten gesprochen). Sie haben sich dann zumeist von ihrer ursprünglichen, oben genannten Funktion losgelöst und haben gravierende negative Auswirkungen auf das Leben der Betroffenen.

Schmerz ist eine nicht öffentlich zugängliche persönliche Erfahrung, eine Empfindung, die im Gehirn entsteht. Sie hat sensorisch-diskriminative, affektiv-motivierende und kognitiv-bewertende Komponenten [MC68]: Schmerz ist charakterisiert durch Intensität, Ort, Dauer und Qualität; er ist unangenehm und motiviert zu Aktivitäten zur Schmerzlinderung; und er wird beeinflusst durch die Kognition, z. B. durch die Bewertung der Schwere einer Verletzung, durch Ablenkungen oder durch kulturelle Werte [TM11b; MC68]. Die Schmerzerfahrung muss gewissenhaft unterschieden werden von der Schmerzursache (z. B. eine Gewebeschädigung mit Nozizeption), der Schmerzreaktion (z. B. verbale und nicht-verbale Kommunikation) und der Schmerzbeurteilung (z. B. durch eine Pflegekraft). Die Schmerzursache kann oft diagnostiziert werden (z. B. ein Knochenbruch) und kann bei absichtlicher Schmerzstimulation gesteuert werden (z. B. für neurologische Untersuchungen). Sie kann jedoch auch unbekannt oder bereits behoben sein (insbesondere bei chronischen Schmerzen). Typischerweise haben Schmerzen ihrem Ursprung in schädlichen Reizen (wie z. B. Temperaturen die das Gewebe schädigen könnten). Die Reize führen zu

einer Reaktion des sensorischen Nervensystems, die Nozizeption genannt wird. Die Nozizeption entspricht jedoch nicht direkt der Schmerzerfahrung, denn letztere wird durch biologische, psychologische und soziale Faktoren moduliert. Infolge dessen kann der selbe Reiz zu sehr unterschiedlichen Schmerzerfahrungen führen. So kann z. B. eine Anästhesie (Betäubung) mit Medikamenten die sonst starken Schmerzen einer Operationswunde vollständig abstellen. Einige wenige Menschen können aufgrund eines Gendefektes gar keine Schmerzen empfinden [YB97], andere empfinden schon bei schwachen Reizen starke Schmerzen.

Schmerzen führen im Allgemeinen zu beobachtbaren Schmerzreaktion, jedoch moduliert durch persönliche und kontextabhängige Faktoren. Zu typischen Reaktionen gehören nozizeptische Reflexe (Versuche sich der Schmerzursache zu entziehen), schützende Bewegungen oder Posen, charakteristische Mimik (Gesichtsausdrücke), verbale Äußerungen zu den Schmerzen sowie paralinguistische Äußerungen, wie Weinen, Schreien, Jammern oder Stöhnen [Cra92]. Neben diesen Verhaltensreaktionen sind auch physiologische Reaktionen zu beobachten, wie z. B. Änderungen des Herzschlages oder der Schweißproduktion an den Poren der Haut [Wer+19b]. Schmerzreaktionen erlauben Beobachtern, z. B. Pflegekräften oder auch automatisierten Monitoring-Systemen, Rückschlüsse auf das Schmerzempfinden des Leidenden zu ziehen. Dabei sind jedoch weder eine mangelnde Fähigkeit über empfundene Schmerzen zu kommunizieren noch das Ausbleiben von nicht-verbalen Verhaltensreaktionen ein Beleg dafür, dass keine Schmerzen empfunden werden.

Es gibt charakteristische Mimik, die relativ konsistent bei klinischen Schmerzzuständen und verschiedenen experimentell angewendeten Schmerzreizarten zu beobachten ist [CPG11; PS08; Prk92; Kun+07]. Die Stärke der Mimik steigt dabei mit der Intensität des Schmerzreizes [Kun+07; CPG11; Had+02]. Forschung zu Gesichtsausdrücken nutzt zumeist das Facial Action Coding System (FACS) [EFH02], das die Mimik als eine Kombination von Mimikbausteinen, sogenannten Action Units (AUs), beschreibt, die basierend auf der Aktivität von Gesichtsmuskeln definiert wurden. Zu den AUs, die bei Schmerzen am häufigsten auftreten gehören: das Senken und/oder Zusammenziehen der Augenbrauen (AU 4), das Heben der Wangen (AU 6) und Anspannen der Augenlider (AU 7), das Rümpfen der Nase (AU 9), das Heben der Oberlippe (AU 10), sowie das Schließen der Augen (AU 43) [Prk92; PS08]. Es gibt jedoch verschiedene Varianten und die Wahrscheinlichkeit alle oben genannten AUs gleichzeitig zu beobachten ist gering [KL14]. Zu anderen AUs, die bei Schmerz beobachtet werden können, gehören: das Heben der Mundwinkel (AU 12), das Auseinanderziehen der Lippen (AU 20, engl. lip stretch), das Öffnen der Lippen (AU 25), das Öffnen des Mundes durch Senken des Unterkiefers (AU 26), und das weite Öffnen des Mundes (AU 27) [Wil02; CPG11], sowie das Heben der Augenbrauen (AU 1/2) [KL14]. Abb. 2.2 (a)-(c) auf Seite 23 zeigen Beispiele für Schmerzmimik. Einige Studien sprechen dafür, dass die Gesichtsausdrücke bei Schmerzen und Basisemotionen (Freude, Überraschung, Trauer, Wut, Angst, Ekel) unterscheidbar sind [Wil02; Sim+08; CPG11]. Jedoch kommen Schmerzen und Emotionen oft zusammen vor [Wil02], was zu veränderten oder vermischen Gesichtsausdrücken führen kann. Verschiedene Personen unterscheiden sich hinsichtlich ihrer mimischen Expressivität [WAHW17; LMRP17a] und ihres Reaktionsschwellwertes, d. h. der Schmerzintensität die überschritten werden muss um eine mimische Reaktion auszulösen [PC95; Kun+04].

Herausforderungen: Schmerzen sind ein sehr komplexes Phänomen mit vielen verschiedenen Arten, Ursachen und Folgen, das noch nicht vollständig verstanden wird und Gegenstand vieler Forschungsaktivitäten ist. Die empfunden Schmerzen sind (nach aktuellem Stand der Forschung) von außen nicht direkt messbar, so dass Aussagen über die Schmerzen einer anderen Person immer auf der Selbsteinschätzung des Betroffenen, der Beobachtung von Schmerzreaktionen

oder dem Wissen über potentielle Schmerzursachen basieren. Zwar gibt es eindeutig Zusammenhänge zwischen der Heftigkeit einer Verletzung, der Schmerzempfindung und der Schmerzreaktion, jedoch erscheinen diese auch oft inkonsistent [PC95]. Zumindest werden die komplexen Zusammenhänge noch nicht vollständig verstanden, da sie durch zahlreiche biologische, psychologische und soziale Faktoren beeinflusst werden [Int21].

Zu den Faktoren mit Einfluss auf das Schmerzempfinden zählen: die Genetik, die persönliche Lebensgeschichte, insbesondere frühere Erfahrungen mit Schmerzen und die Sozialisation in der Familie und der Kultur, das Alter und andere Faktoren mit Einfluss auf das Nervensystem, die Ernährung und der Kontakt mit toxischen Substanzen, schmerzlindernde Medikamente und die Kognition (z. B. Antizipation, Aufmerksamkeit, Aufregung, Angst oder die Interpretation der Situation) [TM11b; MC68; PC95; HC04; Cra09]. Einfluss auf das Schmerzverhalten haben neben den eigentlichen Schmerzen auch die Situation und der soziale Kontext, Persönlichkeitseigenschaften, Denkweisen und Bewältigungsstrategien, soziale und kulturspezifische Verhaltensregeln und kognitive Einschränkungen (z. B. aufgrund geistiger Behinderung oder einer Gehirnerkrankung) [PC95; HC04; Cra09; CPG11].

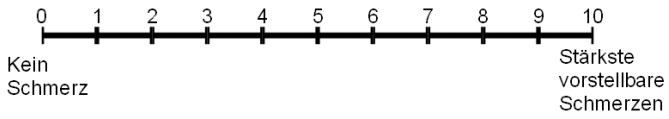
1.1.2. Klinische Schmerzmessung

In der aktuellen klinischen Praxis werden Schmerzen in der Regel anhand der Selbsteinschätzung (engl. self-report) des Patienten diagnostiziert. Dabei werden die Intensität, der Ort, die sensorische Schmerzqualität, zeitliche Eigenschaften sowie schmerzlindernde und -verstärkende Faktoren berücksichtigt. In dieser Arbeit liegt der Fokus auf der Intensität der Schmerzen. Auf Basis der Schmerzbeurteilung wird das Schmerzmanagement begonnen oder angepasst, welches Medikamente, Physiotherapie und Psychotherapie zur Schmerzlinderung einsetzt, wobei mögliche Nebenwirkungen mit abgewogen werden müssen.

Klinische Praxisleitlinien heben hervor, dass die Selbsteinschätzung die valideste Methode der Schmerzbeurteilung ist [Reg13; Her+11]. Validität ist ein zentrales Gütekriterium einer Messung und sagt aus, inwieweit mit dem Messinstrument genau das gemessen wird, was gemessen werde soll [Him07, S. 375]. Für Patienten, die nicht kommunizieren können, empfehlen Herr et al. [Her+11] (1) zunächst zu versuchen, eine Selbsteinschätzung einzuholen, (2) nach möglichen Schmerzursachen zu suchen, (3) das Verhalten des Patienten zu beobachten, (4) eine Einschätzung von einer Person einzuholen, die den Patienten gut kennt und (5) Schmerzmittel zu verabreichen bzw. neu einzustellen und den dadurch erzielten Effekt zu beurteilen. Im Folgenden werden klinisch genutzte Methoden zur Schmerzbeurteilung vorgestellt und diskutiert.

Selbsteinschätzung: Selbsteinschätzung meint die bewusste Kommunikation von schmerzbezogenen Informationen durch die leidende Person, typischerweise in gesprochener oder geschriebener Sprache oder mit Gesten, wie z. B. dem Nicken als Antwort auf eine Frage oder dem Zeigen auf ein Bild, das ihr Empfinden am besten beschreibt. Neben der Verwendung einfacher Fragen wie „Haben Sie Schmerzen?“ im persönlichen Gespräch wird auch auf standardisierte und wissenschaftlich evaluierte Werkzeuge wie Schmerzskalen, Fragebögen und Schmerztagebücher zurückgegriffen, um eine möglich hohe Validität und Reliabilität der Schmerzeinschätzung zu erreichen. Das Gütekriterium Reliabilität (Zuverlässigkeit) eines Messinstrumentes sagt aus, inwieweit die Messergebnisse bei wiederholter Messung unter gleichen Bedingungen reproduzierbar sind [Him07, S. 375]. Für die Messung der Intensität gibt es verschiedene Arten von Schmerzskalen [Nil19]: (1) *Verbale Skalen* nutzen leicht verständliche Kategorien wie „keine Schmerzen“,

Numerische Rating-Skala (NRS)



Quelle: https://commons.wikimedia.org/wiki/File:Numerische_Rating-Skala.png, abgerufen am 15.10.2020 um 16:40. Autor: Cvf-ps. Lizenz: CC BY 3.0.

Behavioral Pain Scale – nicht-intubiert (BPS-NI) [3, 4]

In der BPS-NI ist das Item „Adaption an das Beatmungsgerät" der BPS durch „Vokalisation" ersetzt. Mit der BPS-NI kann bspw. bei wachen, deliranten Patienten der Schmerzstatus bestimmt werden.

Merkmal	Beschreibung	Punkte	
Gesichtsausdruck	entspannt	1	
	teilweise angespannt	2	
	stark angespannt	3	
	Grimassieren	4	
obere Extremitäten	keine Bewegung	1	
	teilweise Bewegung	2	
	Anziehen mit Bewegung der Finger	3	
	ständiges Anziehen	4	
Vokalisation	keine Schmerz-Vokalisation	1	
	Stöhnen ≤3x/min *und* ≤3s	2	
	Stöhnen >3x/min *oder* >3s	3	
	Heulen *oder* verbale Äußerungen, inklusive "Au", "Autsch" *oder* Atemanhalten >3 s	4	
Der Punktwert/Score bzw. das Analgesie-Ziel sollte <6 sein.		**Total**	

Als Adaption der BPS ist für nicht-intubierte Intensivpatienten, die nicht in der Lage sind selbst Auskunft über ihren Schmerzstatus zu geben, die BPS nicht-intubiert (BPS-NI) validiert worden.

Quelle: http://www.pad-management.de/assets/dex_scoringblatt_bps_bps-ni_nrs_mai18_final.pdf, abgerufen am 15.10.2020 um 16:30.

Abbildung 1.1.: In der Klinik verwendete Schmerzskalen: Oft werden Patienten gebeten, ihre Schmerzen anhand einer Zahl der Numerischen Rating-Skala (oben) einzuschätzen. Bei Intensivpatienten, die sich nicht äußern können, beurteilt ein Beobachter die Schmerzen, z. B. anhand der Behavioral Pain Scale (unten).

„schwache Schmerzen", „starke Schmerzen" oder „unerträgliche Schmerzen". (2) Die *visuelle Analogskala* (VAS) besteht aus einer Linie mit den Endpunkten „kein Schmerz" und „stärkster vorstellbarer Schmerz", auf welcher der Patient die seiner Einschätzung der Schmerzintensität markiert (und die Pflegekraft die Position später exakt abmisst). (3) Bei der *numerischen Rating-Skala* (NRS, siehe Abb. 1.1, oben) gibt der Patient seine Schmerzempfindung als eine Zahl an (meist zwischen 0 und 10). Nilges [Nil19] empfiehlt die Nutzung der NRS, da diese eine ausreichende Differenzierung verschiedener Schmerzintensitäten erlaubt sowie einfach und schnell anzuwenden ist (auch ohne Papier, Stift und Lineal). Für Patienten, die mit der NRS nicht zurechtkommen, werden verbale Skalen eingesetzt, oder Skalen, die verschiedene Gesichtsausdrücke zeigen. Mit den Skalen können verschiedene Aspekte abgefragt werden, wie z. B. die aktuellen Schmerzen, die stärksten empfunden Schmerzen oder welche Schmerzintensität akzeptabel wäre.

Beobachtung des Verhaltens: Für Patienten, die ihre Schmerzen nicht selber einschätzen können, ist die Beobachtung des Verhaltens eine der wichtigsten Methoden zur Schmerzbeurteilung. Damit dies möglichst objektiv, zuverlässig und valide geschieht, werden Beobachterskalen eingesetzt, die zumeist für eine spezielle Patientengruppen entworfen und validiert wurden: Für Neugeborene und Kleinkinder z. B. NIPS, CRIES und FLACC [Zam+17a], für ältere Menschen mit schwerer Demenz z. B. PACSLAC, DOLOPLUS2 oder PAINAD [Zwa+06] oder für Intensivpatienten bzw. bewusstlose Patienten z. B. BPS, CPOT und NVPS [Her+11]. Abb. 1.1 zeigt unten als Beispiel die Behavioral Pain Scale (BPS), bei der der Gesichtsausdruck, die Bewegung der oberen Extremitäten sowie die Vokalisation beurteilt werden. Für jeden der drei Aspekte (Items der Skala) wird eine der vorgegebenen Klassen ausgewählt, die mit einer Punktzahl verknüpft ist. Die Summe der drei Punktzahlen ist die finale Einschätzung der Schmerzintensität.

Physiologische Informationen: Akute Schmerzen führen oft zu physiologischen Reaktionen, z. B. beschleunigtem Herzschlag, erhöhtem Blutdruck und verstärkter Schweißproduktion [Wer+19b]. Daher sind physiologische Items auch Teil einiger Beobachterskalen, z. B. von NVPS [KGN09] oder COMFORT [Amb+92]. Laut Herr et al. [Her+11] werden physiologische Indikatoren oft herangezogenen, um das Vorhandensein von Schmerzen zu belegen, was problematisch sei. Herr et al. betonen, dass die physiologischen Reaktionen nicht spezifisch für Schmerzen seien, d. h. unter anderem auch durch andere Formen von Leid, Krankheiten oder Medikamente verursacht werden könnten. Daher sollte die Physiologie nicht als primärer Indikator für Schmerzen herangezogen werden [Her+11].

Herausforderungen: Die Selbsteinschätzung wird aktuell als die valideste Methode der Schmerzbeurteilung angesehen, sie hat jedoch auch Schwächen. Sie spiegelt nicht nur die subjektive Schmerzempfindung wieder, sondern ist auch eine kontrollierte und zielorientierte Reaktion auf Schmerzen [CPG11]. Einfluss haben kann unter anderem die Stärke des Bedürfnisses nach Schmerzlinderung, das Erinnerungsvermögen und die Sprachfähigkeit [Cra92]. Einige Patienten sind zudem nicht in der Lage (oder bereit) eine Selbsteinschätzung abzugeben, so dass die Beobachtung des Verhaltens zur Schmerzeinschätzung nötig wird. Diese erfordert eine Ausbildung der Mitarbeiter, ist immer zu einem gewissen Grad subjektiv, ist zeitaufwändig und sollte regelmäßig sowie bei Interventionen erfolgen [Her+11; Deu09]. Zeit- und Kostendruck sowie Personalmangel im Krankenhausalltag erschweren jedoch die regelmäßige Beobachtung. Hier könnte eine automatisierte Schmerzmessung helfen, die zeitlich kontinuierlich und objektiv arbeitet und so im Idealfall gleichzeitig das medizinische Personal entlastet und die Schmerzmessung verbessert. Die Technologie sollte jedoch kostengünstig sein, um die Chancen der tatsächlichen Umsetzung in klinischen Anwendungen durch ein günstiges Kosten-Nutzen-Verhältnis zu verbessern.

Schmerzen führen nicht immer zu Schmerzreaktionen und die Reaktionen sind auch nicht immer spezifisch für Schmerzen. Ebenso gibt es Schmerzen ohne Schmerzursache (z. B. schädliche Reize) und schädliche Reize ohne Schmerzen. Diese schwachen Zusammenhänge zwischen Schmerzursache, Schmerzempfindung und Schmerzreaktion sind einer der wesentlichen Gründe, warum die Selbsteinschätzung als valideste Methode zur Schmerzmessung angesehen wird, obwohl sie als zielorientierte und kontrollierte Äußerung auch klare Schwächen bezüglich Validität und Zuverlässigkeit aufweist. Wie Schmerz gemessen werden sollte ist daher aus Sicht des Autors dieser Dissertation eine noch nicht zufriedenstellend beantwortete Forschungsfrage. Alle bekannten Maße sind mit gewissen Unsicherheiten behaftet, was für die Entwicklung und Validierung einer neuen, automatisierten Messmethode eine Herausforderung darstellt.

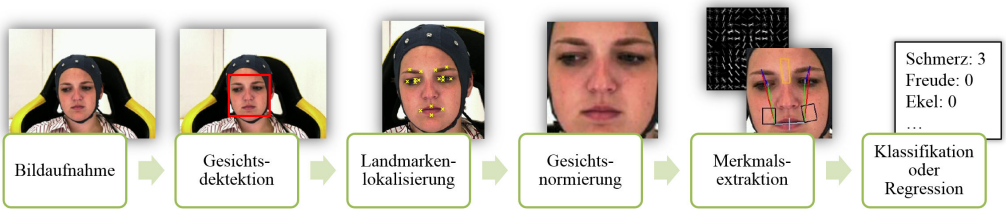


Abbildung 1.2.: Verarbeitungskette der automatisierten Gesichtsanalyse.

1.2. Automatisierte Auswertung von Bildern des Gesichtes

Gesichter vereinen verschiedene Aspekte, die auch Ziel einer automatisierten bildbasierten Erkennung sein können. Hierzu gehören neben der Mimik, d. h. sichtbaren Bewegungen im Gesicht durch Kontraktion der mimischen Muskulatur, auch die Identität, d. h. das permanente Aussehen einer Person, das sie von anderen Individuen unterscheidbar macht, sowie die Kopfpose, d. h. die Position und Rotation des Kopfes relativ zur Kamera. Im folgenden Abschnitt 1.2.1 wird zunächst die Verarbeitungskette vorgestellt, welche für die automatisierte Auswertung von Gesichtsbildern typischerweise angewendet wird. Die Verarbeitungskette nutzt mehrfach maschinelles Lernen, das in Abschnitt 1.2.2 eingeführt wird. In beiden Abschnitten wird auch auf die noch offenen Herausforderungen eingegangen.

1.2.1. Verarbeitungskette

Die automatisierte Analyse von Gesichtern zur Erkennung von Mimik, der Kopfpose, Attributen wie Geschlecht oder Alter, oder der Identität folgt meist der in Abb. 1.2 dargestellten typischen Verarbeitungskette oder einer leichten Variation dieser Kette. Im Folgenden wird genauer auf die Elemente der Kette eingegangen.

1. **Bildaufnahme:** Der erste Schritt ist die Bild- bzw. Videoaufnahme mit einer digitalen Kamera, d. h. die Abbildung der 3-dimensionalen Realität in digitale Pixelgrafiken. Wichtige Parameter hierbei sind unter anderem der Kameratyp, das Sichtfeld der Kamera, die Auflösung der Bilder und die Ausleuchtung der Szene. Der Autor hat für diese Dissertation existierende Farbkameradaten verwendet und auch neue Daten mit CCD-Farbkameras (mit Bayer-Sensoren) aufgezeichnet. Prinzipiell lassen sich die Konzepte jedoch auch auf Bilder anderer Kameratypen, z. B. Graustufenkameras, Nahinfrarotkameras oder Thermografiekameras übertragen. Die Wahl des Objektives entscheidet unter anderem über den Öffnungswinkel der Kamera und damit über die Größe des Sichtfeldes. Bei Vergrößerung des Öffnungswinkels bleibt der beobachteten Person einerseits mehr Bewegungsspielraum, andererseits reduziert sich (bei gleicher Auflösung) der Detailgrad der Gesichter, da sie einen kleineren Teil des Bildes einnehmen, weniger Bildpunkte (Pixel) des Sensors belichten und somit im Bild von weniger Pixeln beschrieben werden. Ein höher aufgelöster Sensor ist insofern wünschenswert, dieser erzeugt jedoch auch mehr Daten, die verarbeitet und meist gespeichert werden müssen, und ist daher in der Regel mit geringeren Bildwiederholraten und/oder höheren Kosten verbunden. Die Beleuchtung ist ein weiterer zentraler Aspekt der Datenaufnahme und steht im Zusammenhang mit weiteren Aspekten wie der Belichtungszeit, der Bewegungsunschärfe, dem Bildrauschen und der Schärfentiefe. In dieser Arbeit werden Bilder und Videos von gut und gleichmäßig ausgeleuchteter Szenen verwendet. Bei der Bildaufnahme zu bedenken ist auch die Wahl der Bild- bzw. Videokompression.

Mit verlustbehafteter Kompression lassen sich auch lange Videos in Dateien handhabbarer Größe speichern. Zu starke Kompression kann jedoch zu Bildartefakten führen, welche die Analyse erschweren.

2. **Gesichtsdetektion:** Der nächste Schritt besteht darin, in den aufgenommenen Bildern bzw. Videos das Gesicht (oder die Gesichter) zu finden. Die Gesichtsdetektion (engl. face detection) ist ein anspruchsvolles Computer-Vision-Problem, für dessen Lösung maschinellen Lernens angewendet wird. Das Ergebnis der Detektion ist eine Menge von Rechtecken, so genannte *Bounding Boxes* (engl. für begrenzende Rechtecke), wobei jedes Rechteck ein Gesicht markiert sowie den Ort und die Größe im Bild beschreibt. Die meisten Detektionsverfahren liefern mit ihren Rechtecken lediglich eine grobe Schätzung von Ort und Größe, die im nächsten Schritt verfeinert wird.

3. **Landmarkenlokalisierung:** Ausgehend von der Bounding Box der Gesichtsdetektion werden im Gesicht so genannte Landmarken lokalisiert. Dabei handelt es sich um charakteristische Punkte (auch Merkmalspunkte genannt) wie die Mundwinkel, Augenwinkel, Nasenspitze und weitere Punkte entlang der Konturen von Augen, Mund, Augenbrauen, Nase und Gesicht. Die Lokalisierung der Punkte ist ebenfalls ein anspruchsvolles Computer-Vision-Problem, für das maschinellen Lernens eingesetzt wird. Das Ergebnis ist ein Vektor von Punkten im 2-dimensionalen Bildkoordinatensystem, bei dem jedem Element eine Bedeutung zugewiesen ist, z. B. linker Augenwinkel, Mitte der Unterlippe oder äußerster Punkt der rechten Augenbraue. Einige Algorithmen zur Landmarkenlokalisierung bieten auch eine Funktion zur Verfolgung der Landmarken eines Gesichts in Videos, da sich die Punktpositionen in aufeinanderfolgenden Bildern meist nur wenig verändern und somit das Lokalisierungsproblem vereinfacht wird.

4. **Gesichtsnormierung:** Um später folgende Verarbeitungsschritte zu vereinfachen, werden das Bild des Gesichts und/oder die Landmarken registriert bzw. normiert, d. h. transformiert um sie mit einer Vorlage in gute Übereinstimmung zu bringen. Das Ziel ist es meist, spätere Schritte invariant gegenüber Einflüssen zu machen, die für die Erkennungsaufgabe keine Rolle spielen. Für die meisten Erkennungsaufgaben sind das die Translation, Skalierung und Rotation in der Bildebene, d. h. die Position, Größe und Drehung des Gesichts relativ zur Kamera. Eine Ausnahme bildet hier z. B. die Kopfposeschätzung, für die die Rotation des Kopfes eine zentrale Rolle spielt und nicht kompensiert werden sollte. Für andere Aufgaben wird oft durch eine Transformation mit Verschiebung, Skalierung, Rotation und Zuschneiden ein Bild vorgegebener Größe generiert, bei dem die Augen, Mund und andere Elemente des Gesichts so weit wie möglich immer an den gleichen Positionen liegen.

5. **Merkmalsextraktion:** Mit Hilfe der Bilder und/oder Landmarken werden Merkmale extrahiert, d. h. eine veränderte meist niedriger-dimensionale Repräsentation der Daten, die als Eingabe für die im nächsten Schritt angewendete Klassifikation oder Regression dient. Ziel der Merkmalsextraktion ist zumeist, die Erkennungsaufgabe zu vereinfachen. Hierfür wird Vorwissen darüber genutzt, welche Aspekte des Gesichts relevant oder irrelevant sind für die Erkennungsaufgabe. Um z. B. das Öffnen des Mundes zu erkennen, würde sich als ein Merkmal der Abstand der Landmarken von Ober- und Unterlippe eignen. Es codiert relevante Informationen und ignoriert irrelevante Informationen, wie die Farbe oder die genaue Position der Lippen im Bild. Gleichzeitig können Merkmale genutzt werden, um den Einfluss von Fehlern oder Ungenauigkeiten eines vorangegangenen Verarbeitungsschritt zu verringern. Außerdem können Merkmale extrahiert werden, die nicht nur Einzelbilder sondern ganze Videos oder Abschnitte eines Videos beschreiben und auf diese Weise zeitliche Informationen für die Lösung der Erkennungsaufgabe nutzbar machen.

6. **Klassifikation oder Regression:** Im letzten Element der Kette wird dem Gesicht (für ein Einzelbild oder eine Bildabfolge) eine Kategorie oder ein Wert zugeordnet, oder auch mehrere. Je nach Erkennungsproblem beschreiben diese Kategorien und/oder reellen Werte z. B. die Mimik, einzelne Teilaspekte der Mimik oder die Kopfpose. Aufgrund der Komplexität der meisten Erkennungsprobleme im Computer-Vision-Bereich wird auch hier maschinelles Lernen eingesetzt, d. h. ein Erkennungsmodell (eine spezielle mathematische Funktion) anhand von Beispieldaten parametrisiert. Im aktuellen Stand der Technik wird die Klassifikation bzw. Regression oft in Einheit mit der Merkmalsextraktion realisiert, insbesondere mit künstlichen neuronalen Netzen. Auf diese Weise kann die Merkmalsextraktion automatisch für die Erkennungsaufgabe optimiert werden, was meist zu besseren Erkennungsergebnissen führt, wenn genug repräsentative Daten vorliegen.

In dieser Verarbeitungskette sind viele Variationen möglich, insbesondere bei der genauen Umsetzung der verschiedenen Schritte. Diese Freiheitsgrade bilden einen sehr hoch-dimensionalen Lösungsraum für die Schmerz- und Mimikerkennung, von dem ein Unterraum in dieser Dissertation exploriert wird. Hierbei liegt ein klarer Fokus auf der zweiten Hälfte der Verarbeitungskette, für die diese Dissertation neue Verfahren vorschlägt. Für die Bildaufnahme, Gesichtsdetektion und Landmarkenlokalisierung werden Verfahren aus dem Stand der Technik genutzt und einige Vergleiche angestellt.

Herausforderungen: In Bildern stellen sich Gesichter als eine Komposition verschiedener Aspekte dar, insbesondere der Identität, der Mimik, der Kopfpose, der Beleuchtung und eventueller Teilverdeckungen. Die zentrale Herausforderung der automatisierten Auswertung von Gesichtsbildern ist es, diese Einflüsse zu trennen, denn ist man an einem interessiert, wirken die anderen als Störeinflüsse. Im Folgenden wird genauer auf Herausforderungen für die Mimikerkennung eingegangen.

Identität: Personen unterscheiden sich bezüglich ihres Geschlechts, ihres Alters, ihrer Ethnie, ihrer Gesichtsgeometrie und weiterer permanenter Erscheinungsbildmerkmale. Bilder von Gesichtern verschiedener Personen unterscheiden sich somit auch bei gleicher Mimik erheblich. Außerdem gibt es Individuen, die auch bei entspannter Muskulatur Anzeichen von Mimik zeigen, z. B. permanent hoch- oder heruntergezogene Mundwinkel. Auch können z. B. altersbedingte permanente Falten im Gesicht zum Teil schwer von vorübergehenden, durch Mimik erzeugten Falten unterschieden werden.

Kopfpose: Abhängig von Lage und Orientierung des Kopfes im 3-dimensionalen Raum relativ zur Kamera ergeben sich auch bei gleicher Mimik sehr verschiedene Bilder des Kopfes. Ist der Kopf nicht frontal zur Kamera gerichtet, können Teile des Gesichts wie Wangen, Augen oder Mundwinkel verdeckt sein. Durch die unebene 3-dimensionale Form des Kopfes und die perspektivische Verkürzung ändern sich außerdem die relativen Positionen und Abstände im Bild. So verkürzt sich z. B. der Bildabstand zwischen Augen und Mund wenn der Kopf wie beim Nicken nach oben oder unten geneigt wird. Gleichzeitig ändert sich die Position der Nasenspitze relativ zu Augen und Mund. Abb. 2.2e (S. 23) zeigt Beispielbilder der mit gleichen Mimik bei unterschiedlicher Kopfpose (oben und 3 Bilder unten links).

Beleuchtung: Zu schwache, zu starke, farbige oder sehr ungleichmäßige Beleuchtung können die automatisierte Mimikerkennung erschweren. In unterbelichteten Bildbereichen hat Bildrauschen einen großen Einfluss, in überbelichteten Bereichen können Details aufgrund des beschränkten Helligkeitswertebereichs „abgeschnitten“ (also unsichtbar) werden. Bei ungleichmäßiger Beleuchtung können in einem Bild beide Effekte auftreten. Zusätzlich können farbige oder ungleichmäßige Beleuchtung Bilder erzeugen, die sich stark

von den Trainingsdaten eines Erkennungssystems unterscheiden und darum falsch klassifiziert werden.

Verdeckungen: Durch Verdeckungen und Teilverdeckungen, z. B. durch Haare, Brillen oder eigene Hände, sind für die Mimikerkennung relevante Bereiche des Gesichts zum Teil nicht mehr sichtbar, so dass Informationen für die Klassifikation fehlen. Auch können Verdeckungen falsche Informationen in die Klassifikation einbringen, so z. B. Masken.

Große Anzahl möglicher Gesichtsausdrücke: Die mimische Muskulatur des Menschen umfasst mehr als 20 Muskeln [Sch+07], die jeweils verschieden stark kontrahiert werden können und dadurch vielfältige Mimik zahlreicher Intensitätsabstufungen erzeugen können. Auch wenn viele Muskeln typischerweise nicht unabhängig voneinander benutzt werden, ergibt sich trotzdem ein Raum mit zahlreichen Freiheitsgraden Es ist eine große Herausforderung Daten zu erheben, die diesen Raum hinreichend abdecken, um ein System zu trainieren und zu validieren, dass die Mimik von Interesse zuverlässig erkennen und von anderen Mimiken unterscheiden kann.

Interpretation der Mimik: Während sich die Mimikerkennung mit den rein visuell beschreibbaren, tatsächlich gezeigten Veränderungen im Gesicht beschäftigt, versucht die Erkennung von affektiven Zuständen, wie Schmerzen, Gefühlen, oder Aufmerksamkeit, von der gezeigten Mimik auf den inneren Zustand der Person zu schließen. Zahlreiche Studien haben Zusammenhänge zwischen der Mimik und dem inneren Zustand belegt [Wil02; KE00], jedoch sind die Zusammenhänge sehr komplex und in vielen Einzelfällen kann es zu falschen Schlussfolgerungen kommen. Zum einen kann Mimik unterdrückt oder gespielt werden [CPG11], zum anderen gibt es zum Teil Mehrdeutigkeiten, d. h. die selbe Mimik kann unterschiedlich gedeutet werden. So z. B., wenn nur Teile eines typischen Musters gezeigt werden, wie ausschließlich das Senken der Augenbrauen, was Teil der prototypischen Mimik von Schmerz, Wut, Angst und Trauer ist [Wil02; Sim+08]. Ebenfalls zum Teil nicht unterscheidbar ist die Mimik bei sehr intensiven positiven und negativen Gefühlen; Menschen greifen hier zur Einordnung vor allem auf Kontextinformationen zurück [ATT12].

1.2.2. Maschinelle Lernverfahren

In dieser Arbeit kommen vor allem überwachte Lernverfahren (engl. supervised learning) zum Einsatz. Ausgangspunkt ist hierbei, dass Objekten jeweils eine Klasse (Kategorie) oder ein Wert einer Skala zugeordnet werden. Die Objekte sind in dieser Arbeit Videoaufnahmen von Gesichtern, genauer gesagt Ausschnitte von Videos (Zeitfenster) oder Einzelbilder. Die zugeordneten Klassen bzw. Werte sind vor allem Schmerzeinschätzungen und Mimikintensitäten. Ein einzelnes solches Objekt i wird im Folgenden als *Sample* (engl. für Beispiel, Probe, Stichprobe) bezeichnet und durch einen meist hochdimensionalen Vektor $\mathbf{x}_i$ beschrieben (z. B. das Bild eines schmerzverzerrten Gesichtes oder ein davon abgeleiteter Vektor von Merkmalen). Die zugeordnete Klasse bzw. der Skalenwert wird als *Label* (engl. für Kennzeichnung, Etikett) y_i bezeichnet und meist als eine Zahl repräsentiert. Einem Objekt können jedoch auch gleichzeitig mehrere Label zugeordnet werden, z. B. die Intensitäten mehrerer Mimikbausteine. In diesem Fall ist y_i ein Vektor aus mehreren Labeln.

Beim überwachten Lernen sind die Samples mit Labeln als eine endliche Menge von Paaren $(\mathbf{x_i}, y_i)$ mit $i = 1, 2, ..., N$ gegeben. Das Ziel ist mithilfe einer solchen Menge (Trainingsdaten) ein Modell $f(\mathbf{x})$ so zu parametrisieren („lernen“ / „trainieren“), dass dieses bei einer zuvor ungesehenen Menge (Testdaten) für jedes $\mathbf{x_i}$ das korrekte Label y_i prädizieren kann. Das heißt im

Idealfall soll $f(\mathbf{x_i}) = y_i$ für alle i erreicht werden. In der Praxis ist dieses Ziel meist nicht im vollem Umfang erreichbar. Daher muss klar unterschieden werden zwischen der Menge der korrekten Label y_i, die auch Grundwahrheit genannt wird, und die der prädizierten, vom Modell ausgegebenen Label $f(\mathbf{x_i}) = \hat{y}_i$.

Ein Modell, das diskrete Kategorien ausgibt (z. B. $f(\mathbf{x}) \in \{1, 2, 3\}$), wird Klassifikator genannt. Eines, das reelle Zahlen ausgibt, nennt man Regressionsmodell bzw. Regressor. Es gibt zahlreiche Arten von Modellen und Lernverfahren, von denen für diese Arbeit drei vielversprechende Optionen gewählt wurden: Support Vector Machine (SVM), Random Forest (RF) und Convolutional Neural Network (CNN). Alle drei Verfahren wurden in zahlreichen Vorarbeiten sehr erfolgreich für die Lösung von Computer-Vision-Problemen eingesetzt [Lin+11; Kin15; Fan+13; Sho+13; KSH12; He+17]. Abschnitt 3.1.1 geht genauer auf die Lernverfahren ein.

Alle Lernverfahren optimieren eine sogenannte Loss-Funktion, die zumindest eine Fehlerfunktion und oft auch einen Regularisierungsterm umfasst. Die Fehlerfunktion misst die Abweichung der Prädiktionen $f(\mathbf{x_i})$ von den Zielwerten y_i. Der Regularisierungsterm hängt nicht von den Zielwerten ab, sondern beispielsweise von den gelernten Parametern des Modells [KGC17]. Meist dient er dazu, einer Überanpassung (engl. overfitting) auf den Trainingsdatensatz entgegenzuwirken, bei der auch Eigenschaften des Trainingsstichprobe mit gelernt werden, die in der zugrunde liegenden Verteilung so nicht vorkommen. Folge einer Überanpassung ist, dass das Modell auf den Trainingsdaten deutlich bessere Ergebnisse liefert als auf neuen Daten.

Entscheidend für die Nützlichkeit eines Erkennungsmodells ist jedoch, wie gut es generalisiert, d. h. wie gut es auf neuen Daten funktioniert. Um das zu bewerten, werden die verfügbaren annotierten Daten vor dem Lernen in Trainings- und Testdaten aufgeteilt. Anschließend wird für die Testdaten die Übereinstimmung der Prädiktionen $f(\mathbf{x_i})$ mit den Grundwahrheiten y_i mithilfe eines Performance-Maßes (engl. Leistungsmaß) beziffert, wodurch quantitative Vergleiche zwischen verschiedenen Modellen möglich werden. Mit statistischen Signifikanztests kann bewertet werden, ob die beobachteten Unterschiede oder Zusammenhänge „echt“ (signifikant) sind bzw. wie wahrscheinlich es ist, dass sie durch Zufall zustande gekommen sind. Details zur Evaluation von Erkennungssystemen und zu Signifikanztests sind in Abschnitt 3.1.2 zu finden, Details zu Performance-Maßen in Abschnitt 3.1.2 und Anhang A.

Herausforderungen: Das maschinelle Lernen versucht ein Modell f zu finden, das die Zusammenhänge von meist *unendlich vielen* verschiedenen Objekten i (z. B. alle möglichen Videos von Gesichtern, repräsentiert durch $\mathbf{x_i}$) und zugehören Labels y_i (z. B. Schmerzeinschätzungen) so genau wie möglich beschreibt. Es ergeben sich folgende Herausforderungen:

Begrenzte Verfügbarkeit von Daten: Für das Finden eines Modells steht nur eine endliche Stichprobe aus der zugrunde liegenden Verteilung zur Verfügung. Während einfache Zusammenhänge, wie ein linearer Zusammenhang in einem niedrigdimensionalen Raum, schon mit einer kleinen Stichprobe gefunden werden können, erfordern komplexere, nichtlineare und höherdimensionale Probleme deutlich größere Datenmengen. Für einige Computer-Vision-Aufgaben werden Bilder millionenfach aus dem Internet gecrawlt und anhand der Suchanfragen oder von Crowd-Sourcing mit Labeln versehen [Den+09; Cao+18]. Daten für die Erkennung von Schmerzmimik sind jedoch im Internet kaum zu finden, so dass man die Datensätze selbst aufzeichnen und annotieren muss, oder auf einen der wenigen öffentlich verfügbaren Datensätze anderer Forschungsgruppen zurückgreifen muss. Verglichen mit anderen Computer-Vision-Aufgaben ist die verfügbare Datenmenge insofern sehr beschränkt: Der größte dem Autor bekannte Datensatz mit Schmerzmimik ist die X-ITE Pain Database mit 134 Probanden und etwa 25.000 annotierten Schmerzreizen.

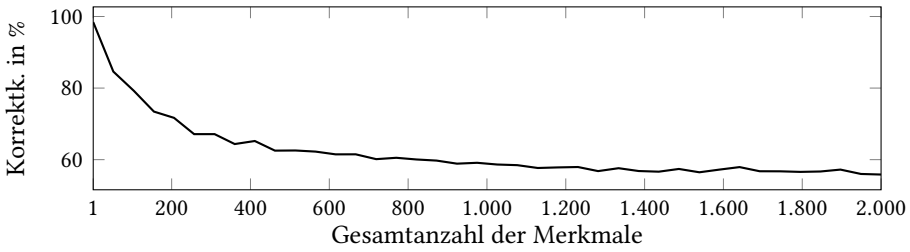


Abbildung 1.3.: Veranschaulichung des Fluchs der Dimensionalität beim Finden eines relevanten Merkmals. Simulationsbeispiel mit SVM, bei dem wiederholt mit $N = 100$ zufälligen Merkmalsvektoren verschiedener Dimensionalität trainiert und auf unabhängigen Daten getestet wurde. Als Label wurde eines der Merkmale mittels Schwellwert binarisiert. Gezeigt wird das Sinken der Korrektklassifikationsrate mit steigender Anzahl von irrelevanten Merkmalen.

Wahl der Repräsentation: Im Zusammenhang mit der begrenzten Verfügbarkeit von Daten steht die Herausforderung, wie die Daten repräsentiert werden sollen und wie viele Dimensionen diese Repräsentation $\mathbf{x_i}$ haben kann. Höherdimensionale Repräsentationen beinhalten potentiell mehr relevante Informationen, können zugleich auch mehr irrelevante Information einschließen. Nehmen wir zunächst an, wir kennen ein Merkmal mit dem ein Klassifikationsproblem fehlerfrei lösbar ist (Korrektklassifikationsrate 100%). Abb. 1.3 veranschaulicht anhand einer Simulation mit einem SVM-Klassifikator wie sehr sich die Korrektklassifikationsrate durch das Hinzunehmen von irrelevanten Merkmalen (Zufallswerten) verschlechtern kann. Bei gleichbleibender Anzahl von Trainingsbeispielen ist es für den Klassifikator mit zunehmender Merkmalsanzahl immer schwieriger das trennende Merkmal zu finden, da sich auch in den anderen Merkmalen durch Zufall scheinbare Zusammenhänge finden lassen (die sich jedoch in den Testdaten nicht bestätigen). Hier kommt der so genannte Fluch der Dimensionalität zum Tragen, denn mit jeder hinzugefügten Dimension vergrößert sich das Volumen des Merkmalsraums und er wird immer dünner besetzt.

Das eigentliche Problem in der Praxis ist, dass es das hier angenommene „perfekte“ Merkmal meist nicht gibt (oder es unbekannt ist) und der Klassifikator „schwächere“ Merkmale finden und kombinieren muss. Im Allgemeinen steigt daher die Erkennungsrate zunächst mit der Anzahl der in Betracht gezogenen Merkmale bis zu einem Optimum und fällt beim Hinzunehmen weiterer Merkmale wieder (bei gleichbleibender Sample-Anzahl). Dieser Sachverhalt wird auch *Hughes Phenomenon* bzw. *Peaking Phenomenon* bezeichnet [Tru79; Hug68]. Für einen konkreten Fall ist das Optimum zunächst unbekannt, hängt unter anderem vom Informationsgehalt der konkreten Merkmale ab und lässt sich nur experimentell finden. Häufig können durch die Wahl und Konstruktion von Merkmalen anhand von Vorwissen auch mit einer sehr geringer Anzahl von Merkmalen gute Erkennungsergebnisse erzielt werden.

Wahl des Modells: Im maschinellen Lernen finden momentan große Fortschritte bei der Entwicklung neuer Modelle und Lernmethoden statt. Bei vielen Problemen sind es aktuell Ansätze des Deep Learning, die am erfolgreichsten sind. Ein wesentlicher Grund hierfür ist, dass hier die Merkmalsextraktion vom Modell übernommen und für die konkrete Anwendung optimiert wird, so dass die Herausforderung der Wahl der Repräsentation im Wesentlichen entfällt. Typischerweise werden die Rohdaten (Bilder oder Videos) als Eingabe für das Modell genutzt und die Fragen bezüglich der Repräsentation beschränken sich

auf die Auflösung, eventuell anzuwendende Normalisierungen und ähnliches. Die Herausforderungen bei der Wahl des Modells hängen – auch bei Deep Learning – mit der begrenzten Datenmenge und dem Fluch der Dimensionalität zusammen. Wie bei der Anzahl der Merkmale gibt es auch bei der Anzahl der optimierbaren Parameter des Modells, bzw. der Kapazität des Modells, ein unbekanntes Optimum, das von vielen weiteren Faktoren abhängt und sich nur experimentell finden lässt. Bei zu hoher Kapazität kommt es leicht zur Überanpassung auf die Trainingsdaten und schlechter Generalisierung, bei zu niedriger kann die Erkennungsaufgabe selbst auf dem Trainingsdatensatz nicht hinreichend gut gelöst werden. Neben der Kapazität spielt auch die konkrete Art des Modells und des Trainings eine große Rolle. Z. B. kann ein CNN, das ausgehend von einer guten Initialisierung der Gewichte und mit Regularisierung trainiert wird, auch mit einem kleinen Lerndatensatz besser generalisieren als das gleiche CNN, wenn es ausgehend von einer zufälligen Initialisierung ohne Regularisierung trainiert wird.

Wahl der Grundwahrheit: Jedes bekannte Maß für Schmerzen ist mit gewissen Unsicherheiten behaftet (vgl. Herausforderungen in Abschnitt 1.1). Insofern ist die Wahl der richtigen Grundwahrheit für das Trainieren eines Schmerzerkennungssystems keine einfache Aufgabe. In vielen Fällen zeigt sich Schmerz nicht in der Mimik, so dass die Label der Schmerzreizung oder der Selbsteinschätzung für die Mimikerkennung in diesen Fällen wie fehlerhafte Annotationen (Label-Rauschen) wirken können. Mögliche Folgen sind unter anderem schlechtere Erkennungsergebnisse, Überanpassung der Modelle und dass mehr Trainingsbeispiele benötigt werden [FV14; AU20]. Aber auch, wenn man die Erkennungsaufgabe auf die visuell beobachtbaren Schmerzreaktionen beschränkt möchte, stellt sich die Frage, wie diese bemessen werden sollen. Schmerzmimik und andere Reaktionen sind vielfältig und es existieren zahlreiche Beobachterskalen zur Schmerzmessung, deren Validität nur schwer zu vergleichen ist [Wer+19b].

Ungleichverteilung: Die mimische Muskulatur ist die meiste Zeit entspannt. Daher ergibt sich in den meisten Mimikdatensätzen eine natürliche Ungleichverteilung, in der Mimik seltener vorkommt als das neutrale Gesicht. Für viele maschinelle Standardlernverfahren würde das Training mit einem solchen Datensatz bedeuten, dass die Erkennung der Mimik weniger wichtig ist als die Erkennung des neuralen Gesichts. Dies ist jedoch im Allgemeinen nicht der Fall, z. B. wenn es Schmerzereignisse unter allem Umständen erkannt werden sollen und dafür auch Fehlalarme in Kauf genommen werden können. Insofern ist die Handhabung der Ungleichverteilung der Klassenzugehörigkeiten eine Herausforderung, die besondere Beachtung verlangt [HG09; Lop+13].

1.3. Zielsetzung und Forschungsfragen

Diese Dissertation zielt auf die Verbesserung der Messung von Schmerzen ab und möchte damit langfristig zu einer besseren Schmerzbehandlung beitragen, insbesondere für Patienten, die sich nicht selbst zu ihren Schmerzen äußern können. Hierzu sollen bild- und videobasierte Methoden für eine objektive, zuverlässige und vollständig automatisierte Messung von akuten Schmerzen entwickelt und evaluiert werden. Forschungsgegenstand ist vor allem die Erkennung von Mimik, zum einen von Schmerzmimik, zum anderen von Action Units nach dem Facial Action Coding System (FACS) [EFH02], die zur Beschreibung von beliebigen Gesichtsausdrücken dienen.

Es gibt zahlreiche Vorarbeiten zur automatisierten Erkennung von Mimik und Schmerzen [Wer+19b; SGC15; Cor+16; ZLZ20]. Diese haben zwar bereits viel erreicht, sie genügen jedoch noch nicht den Anforderungen vieler realer Anwendungen, insbesondere im Zusammenhang mit

den oben erwähnten Herausforderungen. Ziel der vorliegenden Arbeit ist es, neue Erkenntnisse zu gewinnen und verbesserte Methoden zu entwickeln, insbesondere zu den Herausforderungen durch die Kopfpose, durch die begrenze Verfügbarkeit von Daten und durch die Charakteristik von Schmerzen.

Diese Arbeit zielt auf eine kostengünstige und einfach anwendbare Messtechnologie ab, um Hürden für einen späteren klinischen Einsatz zu vermeiden. Sie soll mit möglichst geringem Hardware- und Personalaufwand anwendbar sein. Dazu wird ein unimodaler Ansatz verfolgt, der in der Anwendung mit lediglich *einer* Standardfarbkamera auskommt. Diese könnte im Zuge der Produktentwicklung durch eine Kamera mit integrierter Verarbeitungseinheit ersetzt werden (z. B. OpenCV AI Kit), welche die gesamte Bildauswertung übernimmt und lediglich die Ergebnisse der Auswertung ausgibt, was Schwierigkeiten hinsichtlich des Datenschutzes reduzieren würde. Gegenüber einem multimodalen Ansatz mit biomedizinischer Sensorik würde der unimodale Ansatz Personal- und Zeitaufwand einsparen, der für das Anbringen und Überprüfen von Kontaktsensoren nötig ist. Insgesamt wird ein vollständig automatisiertes Monitoring angestrebt, das medizinisches Personal so gut es geht entlastet und unterstützt.

Folgende Forschungsfragen sollen beantwortet werden:

1. **Inwiefern sind die verfügbaren Schmerzerkennungsdatensätze von Unsicherheiten der Grundwahrheit betroffen? Was bewirken sie? Wie kann mir ihnen umgegangen werden?** Wie bereits angesprochen, ist die automatisierte Schmerzerkennung mit einigen nicht-technischen Herausforderungen konfrontiert. Diese zeigen sich insbesondere an den Grundwahrheiten, die untersucht und diskutiert werden sollen. Ziel ist es die verfügbaren Datensätze möglichst gut ausnutzen zu können, die Limitierungen der Datensätze zu verstehen und Erkenntnisse für die Erhebung neuer Datensätze zu gewinnen.
2. **Welche Computer-Vision- und Machine-Learning-Methoden erreichen die beste Performance?** Die Untersuchung der Generalisierung erfolgt dabei im Kontext vergleichsweise kleiner Datensätze, mehr oder weniger großer Unsicherheit in der Grundwahrheit und zum Teil starker Ungleichverteilungen der Klassenzugehörigkeit. Für jedes Element der Verarbeitungskette der Mimikerkennung (Abb. 1.2) gibt es zahlreiche Möglichkeiten, so dass sich durch deren Kombinationen ein sehr großer Raum von möglichen Erkennungssystemen aufspannt. Hier sollen explorativ verschiedene Möglichkeiten erforscht werden, insbesondere verschiedene Normierungsverfahren, Merkmale, sowie Lernansätze und Prädiktionsmodelle. Es stellt sich die Frage, ob Ende-zu-Ende-Lernen mit einem CNN oder die klassische Herangehensweise mit Merkmalsextraktion und anschließendem unabhängig gelernten Prädiktionsmodell besserer funktioniert. Im letzteren Ansatz werden drei Kategorien von Merkmalen betrachtet: (1) werden Merkmale mit allgemeinen oft verwendeten Verfahren extrahiert, (2) werden speziell für die Problemstellung eigene Merkmale entworfen und (3) es wird ein CNN zur Merkmalsextraktion genutzt. Zusätzlich werden einige Hypothesen untersucht: (1) Der Autor erwartet aufgrund der geringen Datenmenge eine bessere Generalisierung durch die Anwendung von Transferlernen und Multi-Task-Lernen (vgl. Abschnitt 3.1.1). (2) Es werden positive Effekte durch die Ausnutzung von zeitlichen Informationen erwartet, d. h. eine bessere Performance bei videobasierter Erkennung im Vergleich zu einzelbildbasierter Erkennung. (3) Es wird erwartet, dass durch Regression eine bessere Schätzung von Intensitäten erreicht werden kann als durch Klassifikation.
3. **Wie kann erreicht werden, dass das System bei möglichst vielen Kopfposen gut funktioniert?** Für ein permanentes Schmerzmonitoring kann nicht verlangt werden, dass der Patient dauerhaft in die Kamera schaut. Die meisten Vorarbeiten der Mimik- und Schmerzerkennung funktionieren jedoch nur zuverlässig, wenn die Person frontal oder

nahezu frontal in die Kamera blickt. Im Gegensatz dazu wird in dieser Arbeit Kopfposeinvarianz angestrebt, d. h. gute Performance unabhängig von der Kopfpose, um die Nützlichkeit und Akzeptanz der Systeme zu verbessern.

4. **Kann die Erkennung mit niedrigen Hardwarekosten realisiert werden, d. h. ohne Spezialkameras, mit geringer Anzahl von Kameras, mit niedriger Auflösung?** Mehrkamerasysteme und aktive Sensoren ermöglichen die Aufnahme von 3D-Daten, sind jedoch mit höheren Kosten und in der Auswertung mit größerem Rechenaufwand verbunden. Auch Thermografiesysteme, mit denen beleuchtungsunabhängige Bilder aufgenommen werden können, sind im Allgemeinen deutlich teurer als Standardfarbkameras, die z. B. in Smartphones oder als Webcams allgegenwärtig sind. Diese Arbeit verzichtet daher bewusst auf die Nutzung von Spezialkameras, um einfach und vielfältig einsetzbare, kostengünstige Erkennungssysteme zu entwickeln. Entsprechend wird auch das Ziel verfolgt, bei der Anwendung des entwickelten System mit möglichst wenigen Kameras auszukommen, im Idealfall mit nur einer Kamera. Um dem Patienten mit der gleichen Anzahl Kameras einen größeren Bewegungsspielraum zu ermöglichen oder alternativ preisgünstigere Kameras einsetzen zu können, wird untersucht, inwieweit die räumliche Auflösung des Bildes bzw. Gesichtes die Performance der Mimikerkennung beeinflusst.

5. **Kann das entwickelte Erkennungssystem eine ähnlich gute Beurteilung der Schmerzen erreichen wie ein Mensch?** Nach aktuellem Forschungsstand ist das Ideal einer fehlerfreien Messung empfundener Schmerzen nicht erreichbar, denn selbst für den Menschen ist die Beurteilung der Schmerzen einer beobachteten Person keine leichte Aufgabe. Für Erkennungssysteme kommen die technischen Herausforderungen aus Abschnitt 1.2 hinzu. Um die Nützlichkeit der entwickelten Systeme zu beurteilen, wird daher mit der Leistungsfähigkeit von Menschen verglichen.

Es wird angestrebt, datensatzübergreifende, möglichst allgemeingültige Aussagen treffen zu können. Daher werden die Untersuchungen größtenteils mit mehreren Datensätzen durchgeführt.

1.4. Beiträge und Gliederung der Arbeit

Ausgangspunkt der restlichen Arbeit ist die Beschreibung und Untersuchung der zur Verfügung stehenden Datensätze und Grundwahrheiten in Kapitel 2. Unter anderem werden dort Limitierungen der Datensätze und Grundwahrheiten aufgezeigt, Teildatensätze für spätere Untersuchungen erzeugt sowie die Performance ermittelt, die Menschen bei Erkennungsaufgaben dieser Arbeit erreichen.

Kapitel 3 beschäftigt sich mit der einzelbildbasierten Erkennung von Facial Action Units und Schmerzen. Hierfür werden zunächst Grundlagen und verwandte Arbeiten beschrieben, gefolgt von neuen Methoden zur Normierung des Gesichts, zur Merkmalsextraktion und zum Lernen von Erkennungsmodellen. Zur Merkmalsextraktion wird ein neues Verfahren zur Schätzung der Kopfpose im 3D-Raum aus 2D-Bildern (ohne 3D-Bilddaten) vorgestellt, sowie darauf aufbauende geometrische 3D-Merkmale für die Mimikerkennung entworfen. Als Lernmethoden werden insbesondere CNN mit Transferlernen und Multi-Task-Lernen vorgeschlagen, sowie SVM/SVR und RF in Kombination mit verschiedenen Merkmalen, auch CNN-Merkmalen, sowie die Fusion von Merkmalen. Für das Transferlernen und Multi-Task-Lernen wird der Datensatz Bosphorus3D erzeugt. Außerdem wird ein Verfahren zur Behandlung von Ungleichverteilung in Mehrklassenproblemen vorgeschlagen, das über einen Hyperparameter auf den Datensatz optimiert werden kann. Zahlreiche Untersuchungen mit mehreren Datensätzen zeigen, inwieweit die Performance von

der Reduzierung der Auflösung, verschiedenen Kopfposen und Gesichtsnormierungsverfahren, sowie verschiedenen Merkmalen und Lernverfahren beeinflusst wird. Vergleiche zu verwandten Arbeiten und der menschlichen Performance zeigen das sehr gute Abschneiden der vorgeschlagenen Methoden.

Kapitel 4 adressiert die videobasierte Schmerzerkennung auf Basis der Mimik, wobei es auf den Ergebnissen der einzelbildbasierten Erkennung aufbaut. Es werden verschiedene Methoden zur zeitlichen Integration von Einzelbildinformationen vorgeschlagen und evaluiert. Hierzu gehört zum einen ein Statistikdeskriptor, mit dem Zeitreihen von Einzelbildmerkmalen beschrieben werden. Dieser ist mit einem beliebigen Prädiktionsmodell kombinierbar. Zum anderen werden verschiedene Varianten eines Video-CNN vorgeschlagen, das mit Transferlernen auf dem Einzelbild-CNN des vorangegangenen Kapitels aufbaut. Zur Ausnutzung von Dynamikinformationen wie Geschwindigkeit, Beschleunigung oder Dauer einer Bewegung wird eine Variante entwickelt, die zeitliche Convolution ausnutzt. Für den Umgang mit Eigenheiten der Grundwahrheit bei der Messung der Schmerzintensität wird eine Gewichtung der Klassen vorgeschlagen. Die Performance der zeitlichen Integrationsmethoden wird ausführlich untersucht, mit mehreren Datensätzen, Einzelbildmerkmalen und Prädiktionsmodellen.

Kapitel 5 fasst diese Dissertation, ihre Ergebnisse und Schlussfolgerungen zusammen und schlägt vielversprechende Richtungen für weiterführende Arbeiten vor.

2. Datengrundlage

Für die Anwendung von maschinellem Lernen zur Schmerzerkennung und für die quantitative Evaluierung der gelernten Modelle sind geeignete Datensätze mit annotierter Grundwahrheit nötig. Im folgenden Abschnitt werden zunächst mögliche Grundwahrheiten diskutiert. Abschnitt 2.2 stellt die in dieser Arbeit verwendeten Datensätze vor. Diese werden in Abschnitt 2.3 mit dem Ziel untersucht, die Eigenschaften der Grundwahrheiten sowie das Auftreten und Fehlen von Schmerzverhaltensreaktionen besser zu verstehen und Erkenntnisse für eine bessere automatisierte Schmerzerkennung zu gewinnen.

2.1. Grundwahrheiten

Schmerz ist ein subjektives Empfinden und lässt sich nicht auf direktem Wege messen (vgl. Abschnitt 1.1.1). Meist gibt es jedoch Schmerzursachen und -reaktionen, die sich beobachten lassen, eine indirekte Messung möglich machen und sich somit als Grundwahrheit eignen. In den folgenden Abschnitten werden die in dieser Arbeit verwendeten Grundwahrheiten vorgestellt und diskutiert: Informationen über den angewendeten Schmerzreiz, die Selbsteinschätzung des vom Schmerz betroffenen Probanden bzw. Patienten, die Fremdeinschätzung des Schmerzes durch einen Beobachter und die Einschätzung der Schmerzmimik anhand des Facial Action Coding Systems (FACS). Zur Einordnung siehe auch Abb. 2.1. Teile der folgenden Abschnitte basieren auf zuvor veröffentlichten Artikeln des Autors dieser Dissertation [Wer+19b; Wer+17].

Neben diesen Optionen, die mit den genutzten Datensätzen zur Verfügung stehen, gibt es einige Weitere, die ein spezielles Studiendesign oder Vorwissen für die Gewinnung von Grundwahrheiten ausnutzen. Z. B. ist bekannt und gut belegt, dass sich die Schmerzintensität nach der Einnahme bewährter Schmerzmedikamente reduziert und dass postoperativer Schmerz über die Zeit aufgrund der Wundheilung abnimmt. Beide Sachverhalte ermöglichen die Gruppierung von Schmerzbeobachtungen in Klassen, die sich mittels maschineller Lernverfahren voneinander abgrenzen lassen. Eine derartige Idee wurde von Sikka et al. [Sik+15] bei einer Studie mit postoperativen Schmerzen umgesetzt.

2.1.1. Schmerzreiz

Einer der indirekten Zugänge zur nicht direkt messbaren Schmerzempfindung ist die Schmerzursache bzw. der Schmerzreiz. Insbesondere in experimentellen Schmerzstudien, in denen kontrollierte thermische, elektrische, mechanische oder chemische Schmerzreize angewendet werden [Ole+12], kann hierüber eine Aussage über empfundene Schmerzen getroffen werden. Intensität, Dauer, Häufigkeit und Ort der Schmerzreizung werden bei diesen Studien gezielt gesteuert. Eine höhere Reizintensität, z. B. durch extremere Temperaturen oder längere Anwendung bei einem thermischen Reiz, ist im Allgemeinen auch mit einer höheren Schmerzintensität verbunden. Verschiedene Personen unterscheiden sich jedoch hinsichtlich ihrer Schmerzsensitivität, die sich anhand ihrer Schmerzschwelle und ihrer Schmerztoleranz messen lässt. Die Schmerzschwelle ist die

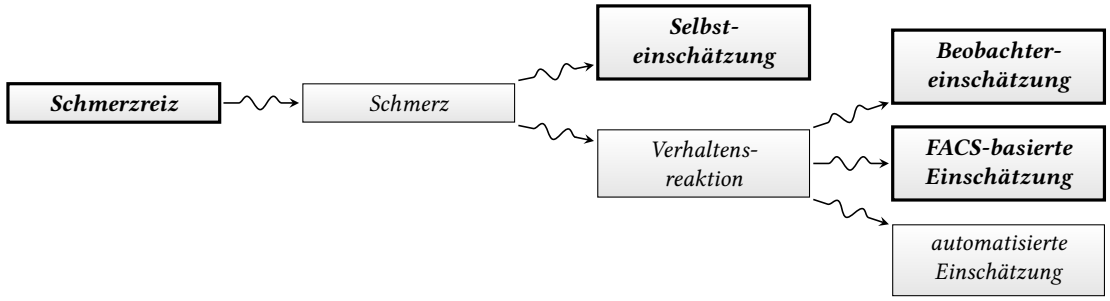


Abbildung 2.1.: Grundwahrheiten für die automatisierte Schmerzerkennung. Schmerz (das subjektive Schmerzempfinden) lässt sich nicht direkt messen, so dass indirekte Maße (□) als Grundwahrheiten genutzt werden müssen. Diese sind jedoch aufgrund komplexer und noch nicht vollständig verstandener kausaler Zusammenhänge mit zahlreichen weiteren Einflussfaktoren (⇝) mit Unsicherheit behaftet.

niedrigste Intensität eines Reizes, die als schmerzhaft wahrgenommen wird [Int21]. Die Schmerztoleranz ist die höchste Intensität eines schmerzhaften Reizes, die die Person zu tolerieren bereit ist [Int21]. Für experimentelle Studien ist es sinnvoll, die Reize im Bereich von Schmerzschwelle bis Schmerztoleranz anzuwenden, da weniger intensive Reize nicht als schmerzhaft empfunden werden und stärkere Reize die Versuchspersonen öfter zum Abbruch des Experimentes verleiten würden. Die Schmerzschwelle und Schmerztoleranz können mithilfe von Selbsteinschätzungen ermittelt werden, wobei jedoch auch die Schwächen der Selbsteinschätzung zum Tragen kommen können, z. B. bewusst oder auch unbewusst falsche Angaben gemacht werden können. Außerdem müssen bei der Reizung Verletzungen vermieden werden, wodurch beispielsweise bei Hitzereizen die maximal verwendbare Temperatur begrenzt ist.

Unter „Herausforderungen“ in Abschnitt 1.1.1 wurden persönliche und situationsabhängige Einflussfaktoren auf den Zusammenhang von Schmerzreiz und Schmerzempfindung aufgezählt. Da deren Auswirkungen noch nicht vollständig verstanden werden und somit mit den verfügbaren Daten nicht modelliert werden können, lassen sich Unsicherheiten im Zusammenhang von Schmerzreiz und Schmerzempfindung nicht vermeiden. Der Versuch, dieses Problem durch die personenspezifische Anpassung der Schmerzreize anhand von Schmerzschwelle und -toleranz zu umgehen, erbt jedoch die Unsicherheiten des Zusammenhangs zwischen Schmerz und Selbsteinschätzung.

2.1.2. Selbsteinschätzung

Aufgrund des persönlichen und subjektiven Charakters des Schmerzempfindens wird die Selbsteinschätzung des Schmerzes (engl. self-report) im Allgemeinen als die beste Möglichkeit zur Messung von Schmerzen angesehen (vgl. Abschnitt 1.1). Die Einschätzung ist jedoch eine kognitive Aufgabe, die nicht von allen Menschen und in allen Situationen bewältigt werden kann bzw. mit hoher Validität und Reliabilität bewältigt werden kann. Denn die meisten Messmethoden, die über die Frage „Haben Sie Schmerzen?“ hinausgehen, verlangen den Menschen viel ab, z. B. den Vergleich mit dem „stärksten vorstellbaren Schmerz“ oder früheren Schmerzerfahrungen, die Reduzierung einer komplexen Empfindung auf eine Zahl oder die Beschreibung der komplexen Empfindung mit Worten. Aber auch einfacher gestrickte Versuche, über Schmerzen zu kommunizieren können scheitern oder missverstanden werden, wenn die sprachlichen oder kognitiven

Fähigkeiten begrenzt sind, z. B. durch schwere Demenz, geistige Behinderung oder Bewusstlosigkeit. Neben den Fähigkeiten des Menschen spielen für die Validität und Reliabilität auch seine bewussten und unbewussten Ziele, der soziale Kontext, sowie Denkweisen und Bewältigungsstrategien eine zentrale Rolle (vgl. „Herausforderungen“ in Abschnitt 1.1.1). Diese können die Aussagen über die eigenen Schmerzen beeinflussen, z. B. um mehr Zuwendung oder Medikamente zu erhalten bzw. allgemein weniger leiden zu müssen, oder um nicht den Eindruck zu vermitteln wehleidig zu sein und „nichts aushalten“ zu können.

Es gibt verschiedene klinisch genutzte und vielfach validierte Skalen für die Schmerzintensität, wie z. B. die Visuelle Analogskala (VAS) oder die Numerische Ratingskala (NRS), die als Grundwahrheit für maschinelles Lernen verwendet werden können. Bei einer Studie kann die Aktivität, die seitens des Probanden zur Erhebung der Selbsteinschätzung nötig ist, jedoch als Belastung empfunden werden und das Experiment beeinflussen. Um dies zu vermeiden, kann bei experimentellen Schmerzstudien mit vielen Wiederholungen die Selbsteinschätzung genutzt werden, um die Schmerzreizung vor dem eigentlichen Versuch für den Probanden anzupassen [Wal+13; Gru+19].

2.1.3. Beobachtereinschätzung

Eine weitere Möglichkeit, Grundwahrheiten zur Messung der empfundenen Schmerzen zu erhalten, ist anhand der Beobachtung der Verhaltensreaktion. Die Schmerzeinschätzung kann mit einer Skala vom Likert-Typ erfolgen [Luc+11b; WAHW17], oder mit einer komplexeren validierten Beobachterskala (vgl. Abschnitt 1.1.2). Ein Problem der verhaltensbasierten Einschätzung ist, dass nicht alle Menschen im gleichen Maße Schmerzreaktionen zeigen. Z. B. wurde bei mehreren Schmerzstudien berichtet, dass 13-50% der Probanden gar keine mimische Reaktion auf Schmerzen gezeigt haben [PC95; Wil95; KL14; WAHW17; Wer+17]. Werden Schmerzen empfunden, aber nicht gezeigt, werden diese von der Beobachtereinschätzung nicht erfasst. Höhere Schmerzintensitäten, die in der Praxis wichtiger sind, führen jedoch häufiger zu Reaktionen und können somit auch zuverlässiger erfasst werden. Das Fehlen einer Reaktion impliziert nicht, dass der Beobachtete keine relevanten Schmerzen empfindet. Diese Unsicherheiten im Zusammenhang zwischen Schmerz und Verhaltensreaktion werden noch verstärkt durch Unsicherheiten, die auf der Seite des Beobachters entstehen. Dessen Einschätzung wird beeinflusst von persönlichen Erfahrungen und Einstellungen, seinem Wissen, der Beziehung zum Leidenden [Cra09] und auch von der Attraktivität der leidenden Person [Cra92].

2.1.4. FACS / PSPI

Mit dem Facial Action Coding System (FACS) [EFH02] codiert ein Beobachter im Nachhinein anhand einer Videoaufnahme die Muskelbewegungen im Gesicht als sogenannte Action Units (AUs). Ziel ist hierbei die objektive Beschreibung der Mimik, zunächst unabhängig von einer Deutung. Die FACS-Kodierung zu erlernen erfordert eine langwierige Ausbildung und die Anwendung durch einen Menschen ist sehr zeitaufwändig. Ein trainierter FACS-Coder benötigt etwa zwei Stunden um eine Minute Video zu annotieren [LBL09]. Daher eignet sich manuelle FACS-Kodierung nur für Forschungszwecke, jedoch nicht für die Anwendung in der klinischen Schmerzmessung. Durch automatisierte FACS-Kodierung [WSAH15; Wer+19c; WSAH20; Luc+12] könnte FACS jedoch auch für Echtzeitmonitoring anwendet werden. Wie bereits in Abschnitt 1.1.1 erwähnt, wurden mit FACS in zahlreichen Studien AUs und AU-Kombinationen identifiziert, die bei Schmerzen vermehrt zu beobachten sind. Diese sind nützlich zur objektiven Beschreibung von Schmerzmimik und der Abgrenzung von anderen mimischen Reaktionen.

Durch Lucey et al. [Luc+11b] und die weite Verbreitung ihres Schmerzdatensatzes UNBC wurde eine Grundwahrheit für Schmerzmimik etabliert, die für jedes Einzelbild anhand von AUs berechnet werden kann, die sogenannte *Prkachin and Solomon Pain Intensity (PSPI)*. Diese basiert auf Vorarbeiten von Prkachin and Solomon [PS08] mit einer Obermenge von UNBC. PSPI wird berechnet anhand der Intensitäten der AUs 4, 6, 7, 9, 10 (jeweils quantisiert und codiert als Ganzzahlen von 0 bis 5) sowie der binären Codierung des Auftretens der AU 43 (0 oder 1):

$$\text{PSPI} = \text{AU4} + \max\{\text{AU6}, \text{AU7}\} + \max\{\text{AU9}, \text{AU10}\} + \text{AU43}. \tag{2.1}$$

Es ergibt sich ein ganzzahliger Wertebereich von 0 bis 16, wobei 0 für kein Schmerz steht und größere Werte für Schmerzen größerer Intensität.

PSPI ist im Bereich der mimikbasierten Schmerzerkennung die am häufigsten verwendete Grundwahrheit. Werner et al. [Wer+17; Wer+19b] kritisieren den Fokus auf PSPI, da es verschiedene Schwächen hat und falsche Vorstellungen dazu existieren: Viele Autoren differenzieren beispielsweise nicht hinreichend zwischen der Schmerzmimik zu einem gewissen Zeitpunkt, die mit PSPI gemessen wird, und der Schmerzempfindung, dem eigentlichen Ziel der Schmerzmessung. Mit abwechselnder Anspannung und Entspannung der Gesichtsmuskeln steigt PSPI an und fällt ab, auch wenn der empfundene Schmerz stetig steigt [Wer+17], denn Schmerzmimik wird typischerweise nur kurz, 2 bis 3 Sekunden lang gezeigt, und selten länger als 5 Sekunden [PC95]. Insofern kann die zeitliche Auflösung in Bezug auf die Messung von *Schmerzen* (nicht Schmerzmimik) irreführend sein. PSPI kann trotz Schmerzen null sein, z. B. bei schwachen Schmerzen, geringer Expressivität [PC95; KL14; Kun+04] oder selteneren Mustern von Schmerzmimik [KL14]. PSPI kann auch größer null sein, obwohl keine Schmerzen empfunden werden. So z. B. wenn die Augen geschlossen sind (AU 43), was nicht spezifisch für Schmerzen ist, sondern auch bei Schlaf, Entspannung oder sehr hellem Licht vorkommt. Auch einige für Emotionen typische Gesichtsausdrücke haben AUs mit PSPI gemeinsam [Sim+08; Zha+14], z. B. Ekel (AU 9/10), Angst und Traurigkeit (AU 4) sowie Fröhlichkeit (AU 6). Insofern können, wenn PSPI in einem weiteren Kontext verwendet wird (z. B. Datensatz BP4D statt UNBC), viele Einzelbilder fälschlicherweise als Schmerzmimik gelabelt werden, was mit anderen Arten von Grundwahrheiten leicht vermieden werden kann. Eine weitergehende Diskussion inklusive Abbildungen ist im Supplemental Material von Werner et al. [Wer+17] zu finden. Auch wenn PSPI ein sehr objektives Maß ist, hat es dennoch Schwächen bezüglich der Validität bei der Messung von Schmerzen.

2.1.5. Diskussion

Es wurden verschiedenen Grundwahrheiten für Schmerzen vorgestellt, die mit Datensätzen zur mimikbasierten Schmerzerkennung zur Verfügung stehen. Jede Grundwahrheit hat Schwächen und ist mit Unsicherheit behaftet, insbesondere da die Zusammenhänge zwischen den messbaren Größen komplex sind, zahlreichen Einflussfaktoren unterliegen und noch nicht vollständig verstandenen werden. Zudem treten mit Schmerz assoziierte Reaktionen nicht bei jeder Schmerzempfindung auf und sie sind auch zumeist nicht spezifisch für Schmerzen. Hierdurch ist Validität der Grundwahrheiten für einen Teil der erhobenen Daten möglicherweise nicht gegebenen oder zumindest zu hinterfragen. Es stellt sich die Frage, inwieweit sich die Validität der Maße bei den zur Verfügung stehenden Datensätzen einschätzen lässt bzw. ob sich Teile der Daten identifizieren lassen, deren Grundwahrheiten wenig valide scheinen. Da ein perfekt valides Maß als Referenz fehlt, kann die Übereinstimmung verschiedener Maße im Sinne der Konvergenzvalidität [Him07, S. 383f] zur Bewertung genutzt werden.

Im folgenden Abschnitt werden die verwendeten Datensätze vorgestellt und anschließend weitere Voruntersuchungen zum Auftreten von Schmerzmimik durchgeführt, die Inkonsistenzen zwischen den beobachtbaren Schmerzreaktionen und den angewendeten Schmerzreizen bzw. der Selbsteinschätzung aufzeigen.

2.2. Datensätze

Im Folgenden werden die Datensätze BioVid, X-ITE, UNBC, BP4D und FERA 2017, sowie Bosphorus vorgestellt, die in dieser Arbeit verwendet werden. An der Erstellung von BioVid und X-ITE war der Autor dieser Dissertation maßgebend beteiligt, insbesondere bei der technischen Umsetzung der Datenaufnahme und der Planung und Einrichtung des Versuchslabors. Vorgestellt werden auch die Varianten der Datensätze, für diese Dissertation erzeugt wurden, um Teilaspekte der automatisierten Schmerzerkennung zu untersuchen. Ein weiterer Datensatz, der vom Autor komplett synthetisch erzeugt und SyLaFaN genannt wurde, wird in Abschnitt 3.2.2 vorgestellt.

Die folgenden Beschreibungen der Datensätze basieren in weiten Teilen auf den Publikationen, die zu Beginn des jeweiligen Abschnittes angegeben werden. Aus diesen können auch weitere Details zu den Datensätzen entnommen werden.

2.2.1. BioVid Heat Pain Database

Der Datensatz „BioVid Heat Pain Database“ [Wal+13; Wer+13] (kurz: BioVid) wurde in einer Studie mit 90 Teilnehmern erhoben, die in drei Altersgruppen rekrutiert wurden (18-35, 36-50 und 51-65 Jahre), wobei jede Gruppe aus 15 Männern und 15 Frauen zusammengesetzt war. Bei der Studie wurde der rechte Unterarm der Probanden mit einer Thermode (Medoc PATHWAY Model ATS) mit Hitze gereizt, um kontrolliert Schmerzen hervorzurufen. Die Temperaturen hierfür wurden individualspezifisch vor dem Hauptexperiment ermittelt, mit dem Ziel Unterschiede in der Schmerzsensitivität auszugleichen. Hierfür wurde der Proband aufgefordert bei ansteigender Temperatur in dem Moment eine Taste zu drücken, in dem er neben der Hitze auch Schmerz empfindet (Schmerzschwelle), sowie in dem Moment, in dem er die Schmerzen nicht mehr ertragen kann (Schmerztoleranz) [Wal+13]. Die persönlichen Temperaturen der Schmerzschwelle und -toleranz, sowie zwei weitere gleichmäßig verteilte Zwischenstufen wurden im Anschluss zur Stimulation der Schmerzen verwendet. Im ersten Teil des Hauptexperimentes wurde jede der vier Temperaturen, die im Folgenden auch Schmerzintensitäten genannt werden, 20 Mal in randomisierter Abfolge für jeweils vier Sekunden angewendet. Zwischen diesen Reizungen gab es randomisierte Pausen zwischen 8 und 12 Sekunden. Im zweiten Teil des Experimentes sollten die Probanden Schmerzen und Emotionen vortäuschen. Im Anschluss wurden den Studienteilnehmern Bilder und Videos gezeigt, mit dem Ziel authentische Emotionen hervorzurufen. Es folgte eine Wiederholung der Schmerzstimulation des ersten Teils, jedoch mit EMG-Elektroden an den Gesichtsmuskeln Zygomaticus Major und Corrugator Supercilii, um mimische Reaktionen kontaktbasiert zu messen. Im ersten Teil wurden diese weggelassen, um Teilverdeckungen des Gesichts zu vermeiden.

Das Hauptexperiment wurde mit Videokameras und biomedizinischen Sensoren aufgezeichnet. Hierfür wurden drei synchronisierte AVT Pike F145C Kameras (Auflösung 1388×1038 Farbpixel, 25 Hz) eingesetzt, eine direkt frontal vor dem sitzenden Probanden und zwei seitlich. Weitere Sensoren waren eine Microsoft Kinect zur Aufzeichnung von Tiefeninformationen (640×480 Pixel, ca. 30 Hz) sowie Kontaktsensoren zur Messung des Hautleitwertes (Schwitzen), des Elektrokardiogramms (Herzmuskelaktivität), des Elektromyogramms von drei Muskeln die mit Schmerz im

Zusammenhang stehen (Muskelaktivität), sowie des Elektroenzephalogramm (Aktivität des Gehirns). Aufgezeichnet wurde außerdem das Temperatursignal als Grundwahrheit der Schmerzreizung. Die Hard- und Software zur Videodatenaufnahme, zur Synchronisation der verschiedenen Sensoren sowie zur Aufbereitung der Daten wurde vom Autor dieser Dissertation konzipiert und implementiert.

Abb. 2.2a zeigt einige Beispielbilder der drei AVT-Kameras, die in dieser Dissertation genutzt werden. Die seitlichen Kameras waren so platziert, dass sie dann ein frontales Gesicht aufgezeichnet haben, wenn der Proband den Kopf um 45° nach links oder rechts gedreht hat. Ursprüngliche Idee bei der Datenaufnahme war es, die Kamerabilder gemeinsam zu betrachten und zumindest in einem der Bilder immer einen nahezu frontalen Blick auf das Gesicht zu haben, auch wenn der Proband den Kopf zur Seite dreht. Diese Dissertation strebt jedoch eine kostengünstige Lösung für das Schmerz-Monitoring an, die im Endeinsatz mit einer einzelnen Kamera auskommt. Um dies zu erreichen soll die Kopfposeunabhängigkeit der Erkennungsmodelle verbessert werden, so dass auch bei seitlicher Ansicht auf das Gesicht eine ähnliche gute Performance erreicht wird und eine frontale Ansicht nicht zwingend nötig ist. Hierfür werden die drei Ansichten einer Situation als unabhängige Samples genutzt, sowohl beim Training als auch bei der Bestimmung der Test-Performance.

Neben den vier stimulierten Schmerzintensitäten (auch PA1, PA2, PA3 und PA4 genannt) werden die Pausen als schmerzfreie Samples (Baseline, BLN) genutzt. Im Datensatz BioVid gibt es jedoch deutlich mehr schmerzfreie Abschnitte als Beispiele für die Schmerzintensitäten. Um diese Ungleichverteilung zu handhaben, wurden feste Zeitfenster der kontinuierlichen Aufnahme als Samples extrahiert. Werner et al. [Wer+14b] haben aus dem ersten Teil des Hauptexperiments passend zu den 20 Samples jeder Schmerzklasse auch 20 Samples aus den Pausen extrahiert, genauer gesagt aus den Pausen nach den Schmerzreizen der niedrigsten Intensität (PA1). Außerdem haben sich Werner et al. [Wer+14b] auf die 87 Probanden beschränkt, bei deren Versuchsdurchführung es keine technischen Probleme gab und die Daten *aller* Sensoren zur Verfügung stehen. Die Samples wurden als Zeitfenster mit 5,5 Sekunden Länge anhand der angewendeten Temperaturkurve extrahiert, bei Werner et al. [Wer+14b] jedoch nur für das Video der frontalen Ansicht (mittlere Kamera). Für diese Dissertation wurden nun zusätzlich die zugehörigen Videos der seitlichen Kameras extrahiert. Diese Teilmenge von BioVid bestehend aus 20 Videos je Proband für jede der fünf Klassen (vier Schmerzintensitäten + kein Schmerz) und jede betrachtete Ansicht. Sie wird im Folgenden *BioVid-A* genannt und in zwei Varianten betrachtet, mit nur der frontalen mittleren Kamera und mit allen drei Kameras.

BioVid-D ist wiederum eine Teilmenge von BioVid-A, bei der die Erkennungsaufgabe auf die Unterscheidung zweier Klassen reduziert wird, auf die höchste Schmerzintensität (PA4) und „kein Schmerz“ (BLN). *BioVid-S* betrachtet die gleichen Klassen wie BioVid-D, die Daten sind jedoch reduziert auf ein Einzelbild je Video. Der Endbuchstabe „S“ steht hier für *statisch* mit Einzelbildern, in Abgrenzung zu „D“ für *dynamisch* mit Videos bei BioVid-D. Für jede der Varianten von BioVid-A, BioVid-D und BioVid-S gibt es auch eine Variante, die lediglich die sieben expressivsten Probanden umfasst. Auf die Erstellung dieser Varianten wird in Abschnitt 2.3 genauer eingegangen. Tabelle 2.1 vergleicht die Datensätze, wobei jede Zeile einen Datensatz bzw. eine Datensatzvariante beschreibt.

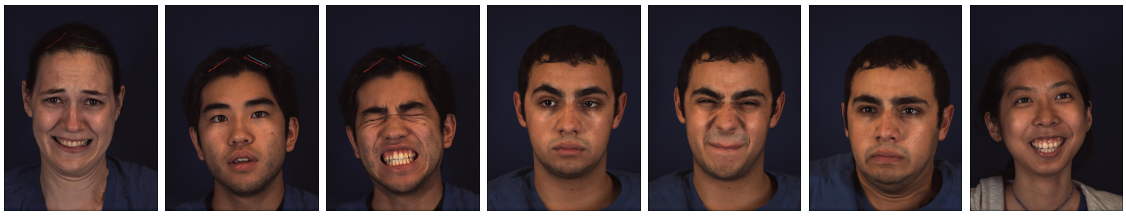
2.2.2. X-ITE Pain Database

Die „e**X**perimentally **I**nduced **T**hermal and **E**lectrical (**X-ITE**) Pain Database“ [Gru+19; Wer+19a] wurde in einer Studie mit 134 gesunden Probanden (67 Männer und 67 Frauen) im Alter zwischen

(a) BioVid-A

(b) X-ITE

(c) UNBC



(d) BP4D

(e) FERA 2017

Abbildung 2.2.: Beispielbilder der verwendeten Datensätze. Für Beispielbilder von Bosphorus3D siehe Abb. 3.5b.

Name	Inhalt	Annotationen (Anzahl Klassen)	Probanden	Schmerz-reize	Ansichten / Kameras	Annotierte ... Videos	Annotierte ... Einzelbilder	Experimentelle Verwendung
BioVid* [Wal+13; Wer+13]	Schmerz, Emot.	Schmerzreiz (5)	90	14.400	4	360	-	-
BioVid-A** [Wer+14b]	Schmerz	Schmerzreiz (5)	87	6.960	1	8.700	-	Abschn. 4.3.2
					3	26.100	-	„
BioVid-A7**	Schmerz	Schmerzreiz (5)	7	560	1	700	-	„
					3	2.100	-	„
BioVid-D**	Schmerz	Schmerzreiz (2)	87	1.740	1	3.480	-	Abschn. 4.3.1
					3	10.440	-	„
BioVid-D7**	Schmerz	Schmerzreiz (2)	7	140	1	280	-	„
					3	840	-	„
BioVid-S**	Schmerz	Schmerzreiz (2)	87	1.740	1	-	3.480	Abschn. 3.3
					3	-	10.440	„
BioVid-S7**	Schmerz	Schmerzreiz (2)	7	140	1	-	280	„
					3	-	840	„
X-ITE* (Hitze / elektrisch) [Gru+19]	Schmerz	Schmerzreiz (4 / 4)	131 / 132	12k / 12k	2	31k / 31k	-	Abschn. 4.3.2
X-ITE-E46** (elektrisch)	Schmerz	Schmerzreiz (4)	46	4.140	2	11k	-	„
UNBC [Luc+11b]	Schmerz	VAS (11), OPR (6)	25	200	1	200	-	Abschn. 4.3.2
		PSPI (17), FACS: 10 AUs	25	200	1	-	48k	Abschn. 3.3.3
BP4D [Zha+14]	Schmerz, Emot.	FACS: 7 AUs	41	41	1	-	147k	Abschn. 3.3.1
FERA 2017 (Training / Test) [Val+17]	Schmerz, Emot.	FACS: 7 AUs	41 / 20	41 / 20	9	-	1.322k / 680k	Abschn. 3.3.2-3
Bosphorus [Sav+08]	Emotionen	FACS: 26 AUs	105	-	1	-	3k	-
Bosphorus3D**	Emotionen	FACS: 26 AUs	105	-	49	-	142k	Abschn. 3.3.3

* Datensatz, bei dessen Erstellung der Autor dieser Dissertation maßgeblich beteiligt war.

** Datensatz, den der Autor auf Basis eines anderen Datensatzes erstellt hat.

Tabelle 2.1.: Vergleich der verwendeten Datensätze bezüglich Inhalt (Schmerz und/oder Emotionen), Art der Annotationen sowie Anzahl der Probanden, Schmerzreize (d. h. der unabhängigen Schmerzereignisse), Ansichten bzw. Kameras, annotierten Videos bzw. Einzelbilder (k steht für Kilo, d. h. den Faktor 1.000). Die letzte Spalte verweist auf die Abschnitte dieser Arbeit, in denen die Daten experimentell verwendet werden.

18 und 50 Jahren aufgezeichnet. Wie bei BioVid wurde mit der Thermode „Medoc PATHWAY Model ATS“ am Unterarm gezielt Hitzeschmerz verursacht. Zusätzlich wurden an Zeige- und Ringfinger Elektroden angebracht, um mit dem Digitimer DS7A gezielt elektrische Schmerzreize zu verabreichen. Sowohl Hitzereize als auch elektrische Reize wurden in zwei Varianten angewendet: phasische Reize mit 5 Sekunden Länge sowie tonische Reize mit 60 Sekunden Länge. Jede der vier Reizarten (Hitze phasisch, Hitze tonisch, elektrisch phasisch und elektrisch tonisch) wurde in drei Intensitäten stimuliert. Um Unterschiede in der Schmerzsensitivität zu handhaben, wurden die Temperaturen bzw. Stromstärken der vier Reizarten für jeden Studienteilnehmer vor dem Hauptexperiment individuell kalibriert. Hierfür wurde die Reizintensität schrittweise erhöht und der Proband jeweils nach seiner Selbstbeurteilung gefragt, um die persönliche Schmerzschwelle und -toleranz sowie die persönlichen Reizintensitäten zu ermitteln. Nach der Kalibrierung hat sich der Proband auf eine Untersuchungsliege gelegt und die etwa 90 Minuten dauernde Hauptphase des Experimentes über sich ergehen lassen. Dabei wurden die phasischen Reize jeder Modalität (Hitze und elektrisch) und Intensität je 30 Mal in randomisierter Reihenfolge wiederholt, jeweils unterbrochen von zufällig langen Pausen (8-12 Sekunden). Die tonischen Reize (je 1 Minute) wurden jeweils einmal angewendet, d. h. es gab sechs tonische Reize je Proband, jeweils gefolgt von einer Pause von fünf Minuten. Dabei folgte jeweils ein tonischer elektrischer Reiz auf einen tonischen Hitzereiz, die beiden mit der höchsten Intensität am Ende des Experiments, die mit den niedrigeren Intensitäten zufällig eingefügt in die Reihe der phasischen Reizungen. In dieser Arbeit werden nur die phasischen Reize betrachtet. Für Untersuchungen mit den tonischen Reizen siehe Werner et al. [Wer+19a].

In der Hauptphase des Experiments wurde das Gesicht des Probanden mit zwei Farbvideokameras aufgezeichnet, einer AVT Pike F145C Kamera (Auflösung 1384×1032 Pixel, 25 Hz) die direkt über der Liege befestigt war, sowie einer seitlich an der Decke des Raumes aufgehängten AVT Prosilica GT1600C (1620×840 Pixel, 25 Hz). Die Videos dieser beiden Kameras werden in dieser Arbeit verwendet. Abb. 2.2b zeigt einige Beispielbilder. In der seitlichen Ansicht ist das Gesicht des Probanden zweimal zu sehen, da hinter der Liege ein Spiegel platziert wurde. Durch den Einsatz eines Spiegels konnte der Autor dieser Dissertation die Anzahl der Kameras im Vergleich zu BioVid von drei auf zwei reduzieren, ohne den Bewegungsspielraum des Probanden einzuschränken bzw. ohne hinnehmen zu müssen, dass das Gesicht bei ungünstiger Kopfdrehung nicht mehr sichtbar ist. Wenn die Mimikerkennung auch bei Kopfdrehungen bis ±45° zuverlässig möglich wird, ist auf diese Weise eine seitliche Kamera mit Spiegel ausreichend um einen Bewegungsspielraum von ±90° abzudecken.

Neben den eben genannten Kameras wurden weitere Daten aufgezeichnet, die in dieser Dissertation nicht verwendet werden: (1) Videoaufnahmen des ganzen Körpers zur Analyse von Körperbewegungen, (2) Thermogrammvideos des Gesichts zur Analyse der Hauttemperatur, (3) Audiosignale zur Analyse paralinguistischer Äußerungen, sowie wie bei BioVid (4) Elektrokardiogramm, (5) Hautleitwert und (6) drei Oberfächen-Elektromyogramme. Die Hard- und Software zur Videodatenaufnahme, zur Synchronisation der verschiedenen Sensoren sowie zur Aufbereitung der Daten wurde vom Autor dieser Dissertation konzipiert und implementiert, ebenso wie der Einsatz des Spiegels.

Wie bei BioVid wurden die Baseline-Samples (`BLN`, kein Schmerz) von den Pausen nach der Hitzestimulation mit der niedrigsten Intensität extrahiert. Im Folgenden werden bei X-ITE die Teildatensätze der *Hitzereize* sowie der *elektrischen Reize* unabhängig voneinander betrachtet. In jedem gibt es jeweils drei Intensitäten von Schmerzstimuli und sowie „kein Schmerzstimulus“ (`BLN`). Beim Teildatensatz der elektrischen Reize ergeben sich die vier Klassen `BLN`, `PE1`, `PE2` und `PE3`,

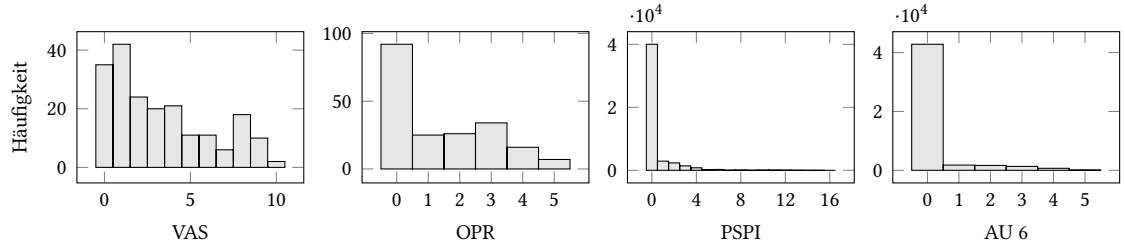


Abbildung 2.3.: Ungleichverteilung der Grundwahrheiten bei UNBC: Histogramme für die Label VAS, OPR, PSPI und die AU 6. Hohe Intensitäten sind selten. Bei OPR, PSPI, und AUs ist das Fehlen von Mimik (Klasse 0) dominant.

beim Hitze-Teildatensatz `BLN`, `PH1`, `PH2` und `PH3`. Von den 134 Probanden wird jeweils die Teilmenge genutzt, für welche die Videoaufnahme mit der frontalen AVT-Kamera sowie die Schmerzreizung fehlerfrei funktioniert hat. Für die elektrische Reizung sind das 132 Probanden und für die Hitzereizung 131 Probanden. Außerdem wird für die elektrische Reizung ein Teildatensatz der 46 expressivsten Probanden vorgeschlagen und untersucht, der *X-ITE-E46* genannt wird, vgl. die Abschnitte 2.3.3 und 4.3.2.

2.2.3. UNBC-McMaster Shoulder Pain Database

Der Datensatz „UNBC-McMaster Shoulder Pain Expression Archive Database“ [Luc+11b] (kurz: UNBC) umfasst 200 Videos von 25 Patienten mit Schulterschmerzen, die sich aktiven und passiven Beweglichkeitstests unterziehen (Beispielbilder siehe Abb. 2.2c). Nach jedem Test, d. h. für jedes Video haben die Patienten ihre empfundenen Schmerzen mit der Visuellen Analogskala (VAS) beurteilt. Die Ergebnisse der Selbstbeurteilung liegen im öffentlichen Datensatz jedoch nur diskretisiert, als ganzzahlige Werte im Wertebereich von 0 bis 10 vor. Zusätzlich wurde jedes Video im Nachhinein von unabhängigen, geschulten Beobachtern beurteilt, wofür sechs Auswahlmöglichkeiten vom Likert-Typ gegeben wurden, von 0 (kein Schmerz) bis 5 (starker Schmerz) [Luc+11b]. Die resultierenden Beobachterbeurteilungen (engl. oberver pain rating, OPR) stehen als weitere videobasierte Grundwahrheit für die Evaluierung von Erkennungssystemen zur Verfügung.

Der Datensatz wurde außerdem mit dem Facial Action Coding System (FACS) [EFH02] codiert [Luc+11b]. Hierbei wurden für jedes der 48.398 Einzelbilder zehn Action Units (AUs) annotiert, für die in vorangegangenen Studien ein potentieller Zusammenhang mit Schmerz gefunden worden war: AU 4, 6, 7, 9, 10, 12, 20, 25, 26 und 43. AU 43, das Schließen der Augen wurde lediglich binär codiert, alle anderen AUs mit Intensität (Ganzzahlige Werte von 0 bis 5). Außerdem wurde für jedes Einzelbild basierend auf den codierten AUs die *Prkachin-Solomon Pain Intensity (PSPI)* berechnet (Wertebereich 0 bis 16), die in vielen Arbeiten als Schmerzgrundwahrheit herangezogen wird (vgl. Abschnitt 2.1.4).

Die Label der Daten, insbesondere die mimikbasierten Label OPR und PSPI sind sehr ungleich verteilt, was für maschinelle Lernverfahren eine Herausforderung darstellen kann [HG09; Lop+13]. Abb. 2.3 zeigt die Verteilung der Label VAS, OPR und PSPI, sowie einer repräsentativen AU anhand von Histogrammen. Die Problematik der extremen Ungleichverteilung von PSPI und AU 6 gilt im Allgemeinen unabhängig vom Datensatz für beliebige AUs, die zumeist nicht präsent sind (Wert 0) und nur sehr selten in hohen Intensitäten auftreten.

2.2.4. BP4D und FERA 2017

Der Datensatz „BP4D-Spontaneous“ [Zha+14] (kurz: BP4D) wurde mit 41 Probanden erhoben, die acht verschiedenen Aufgaben (engl. Tasks) zu absolvieren hatten. Ziel der Aufgaben war es authentische Empfindungen und die zugehörige Mimik hervorzurufen. Zum Auslösen von Schmerzen war eine der Aufgaben der sogenannte *cold pressor test*, bei dem der Proband seinen Arm in Eiswasser taucht. Weitere Aufgaben sollten Fröhlichkeit, Traurigkeit, Überraschung, Verlegenheit, Angst bzw. Nervosität, Wut und Ekel auslösen. Hierfür wurde der Versuch von einem professionellem Schauspieler angeleitet. Für weite Teile des Datensatzes wurde die Intensität von sieben AUs mit FACS annotiert. Abb. 2.2d zeigt einige Beispielbilder des Videodatensatzes, der vorrangig frontale oder nahezu frontale Kopfposen beinhaltet. Neben dem Video stellt der Datensatz auch 3D-Dreiecksnetze für alle Einzelbilder zur Verfügung.

Valstar et al. [Val+17] haben diese 3D-Dreiecksnetze genutzt, um einen Datensatz für die Facial Expression Recognition and Analysis (FERA) Challenge 2017 zu generieren, mit dem Ziel die Entwicklung von Mimikerkennungsmethoden mit besserer Kopfposeunabhängigkeit zu fördern. Für diesen Datensatz, hier FERA 2017 genannt, wurden die texturierten 3D-Dreiecksnetze aus neun verschiedenen Ansichten gerendert, siehe Beispiele in Abb. 2.2e. Für die Generierung der Ansichten wurde das 3D-Modell bezüglich Gierwinkel (engl. yaw) um -40°, -20° und 0° gedreht, jeweils in Kombination mit Nickwinkeln (engl. pitch) von -40°, 0° 40°. Neben dem Trainingsteil des Datensatzes, der auf BP4D basiert, wurde auf Basis des ähnlichen Folgedatensatzes BP4D+ [Zha+16a] ein Testteil mit gleichen Posen jedoch 20 anderen Probanden gerendert.

2.2.5. Bosphorus Database

Der Datensatz Bosphorus [Sav+08] wurde mit 105 Probanden erhoben, die nach Anweisung bis zu 34 verschiedene Gesichtsausdrücke in frontaler Kopfpose gezeigt haben. Es wurden Einzelbilder mit Tiefeninformationen aufgenommen, und 2.902 der Bilder mit 26 FACS AUs codiert. In diesem Datensatz ist zwar keine Schmerzmimik enthalten, er umfasst jedoch eine große Anzahl von Personen und verschieden Gesichtsausdrücken und stellt umfangreiche Annotationen bereit. Daher wird er in Abschnitt 3.2.4 als Grundlage für das Rendering des Datensatzes *Bosphorus3D* genutzt, der in dieser Dissertation als Werkzeug zur Verbesserung der Kopfposeinvarianz der Mimikerkennung angewendet wird.

2.2.6. Vergleich der Datensätze

Tabelle 2.1 stellt verschiedene Eigenschaften der verwendeten Datensätze gegenüber. Die meisten sind Schmerzdatensätze. Bei BP4D und FERA 2017 zeigt etwa eines von acht Videos Schmerzen und die restlichen verschiedene Emotionen. Bei Bosphorus kommt gar keine Schmerzmimik vor, jedoch Emotionsmimik und viele andere instruierte Gesichtsausdrücke, die viele Mimikfreiheitsgrade des Gesichts abdecken. Bosphorus, BP4D, FERA 2017 und UNBC sind mit FACS annotiert, wobei bei Bosphorus deutlich mehr AUs annotiert sind als bei den anderen Datensätzen. BioVid und X-ITE sind mit Grundwahrheiten annotiert, die auf dem angewendeten Schmerzreiz beruhen. UNBC bietet auf Videoebene zwei Grundwahrheiten, sowohl eine Selbstbeurteilung als auch eine Beobachterbeurteilung. Für die Einzelbilder bietet es PSPI, eine AU-basierte Grundwahrheit für Schmerzmimik. Unter den Datensätzen mit videobasierter Grundwahrheit umfasst X-ITE die meisten Probanden, bei denen mit Einzelbildgrundwahrheit Bosphorus. Der Umfang der Datensätze bezüglich Schmerzreizen und Videos ist bei X-ITE am größten, gefolgt von BioVid. Die BioVid-Varianten mit den sieben expressivsten Probanden, sowie UNBC und insbesondere

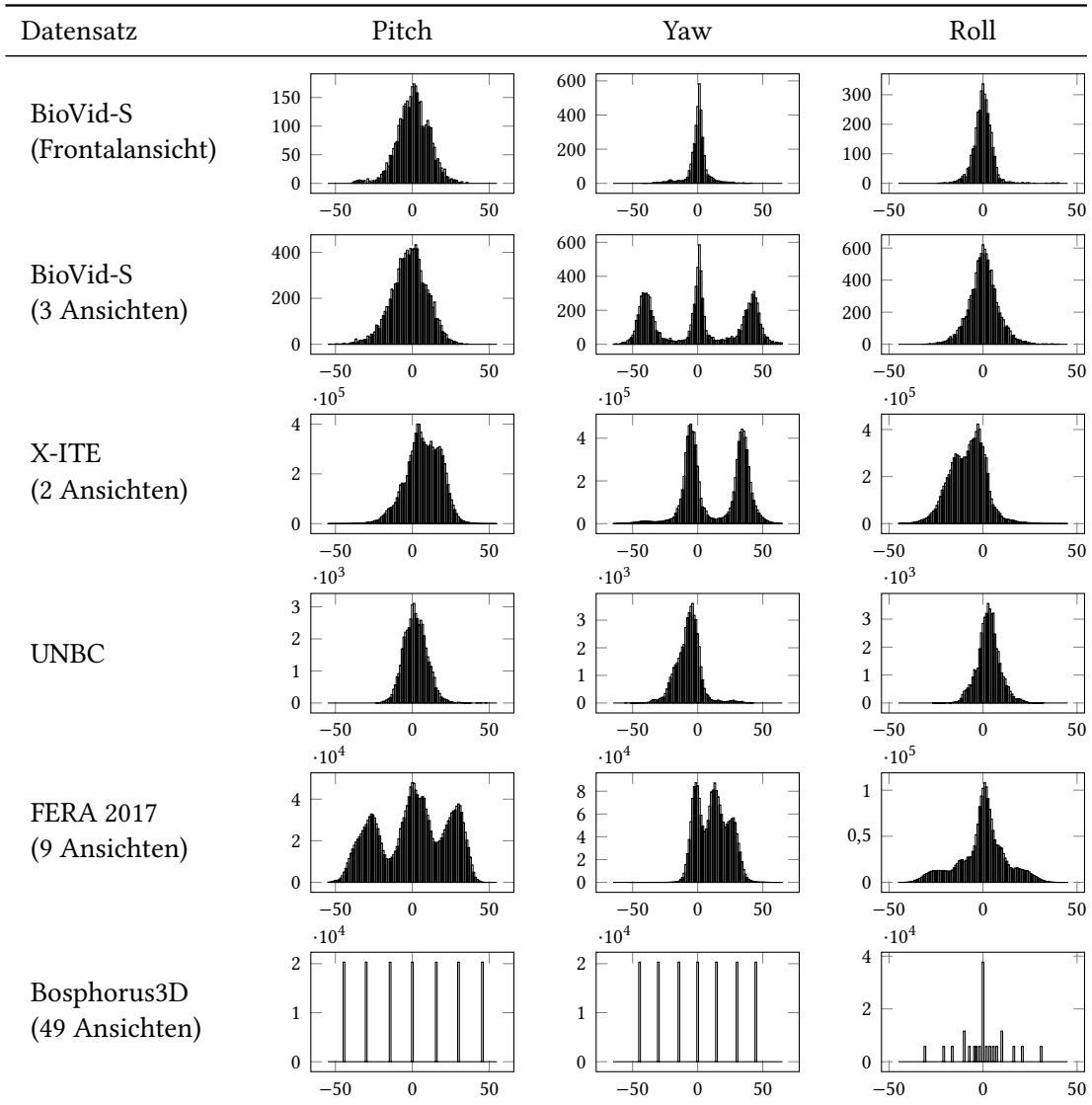


Abbildung 2.4.: Kopfposeverteilung der Datensätze: Histogramme für die Rotationswinkel Pitch (engl. für Nickwinkel), Yaw (engl. für Gierwinkel) und Roll (engl. für Rollwinkel) in Grad (Bedeutung siehe Abb. 3.2b). Für Bosphorus3D wurden die Winkel betrachtet, die für das Rendering genutzt wurden. Für die anderen Datensätze wurden die Winkel mit der in Abschnitt 3.2.3 beschriebenen Methode geschätzt.

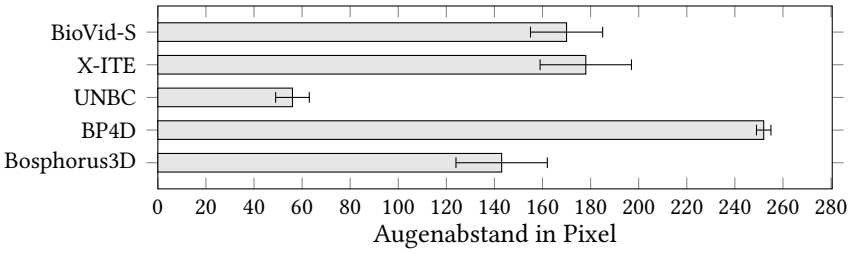


Abbildung 2.5.: Auflösung der Datensätze in frontaler Ansicht: Größe der Gesichter anhand von Mittelwert und Standardabweichung des Augenabstandes im Bild.

BP4D und FERA 2017 bieten für die videobasierte Schmerzerkennung deutlich weniger unabhängige Beispiele.

Im Vergleich von einer zu drei oder mehr Kameras bzw. Ansichten bei BioVid, BP4D / FERA 2017 und Bosphorus / Bosphorus3D vergrößert sich die Kopfposevarianz deutlich. Abb. 2.4 zeigt die Verteilungen der Kopfrotationswinkel für verschiedene Datenbanken. Die ersten zwei Zeilen der Tabelle zeigen insbesondere bei dem Yaw-Winkel die Erweiterung der Kopfposevarianz durch das Hinzunehmen der seitlichen Kameras. Bei X-ITE sind die zwei Ansichten im Yaw-Winkel gut zu erkennen, ebenso die jeweils drei Pitch-Winkel und drei Yaw-Winkel, die zum Rendern von FERA 2017 in 9 Ansichten genutzt wurden. UNBC hat nur eine Kameraansicht, so dass die Kopfposevarianz hier deutlich geringer ausfällt als bei X-ITE, BioVid und FERA 2017. Bosphorus3D hat eine sehr systematische Abdeckung des Kopfposeraumes mit 49 Ansichten.

Abb. 2.5 vergleicht die Auflösung einiger Datensätze bzw. der Gesichter in den Datensätzen, gemessen am Augenabstand in Pixeln (Mittelwert und Standardabweichung über alle Bilder). Hier zeigt sich insbesondere, dass die Gesichter bei UNBC mit nur 56 Pixeln mittlerem Augenabstand sehr gering aufgelöst sind. BP4D hat mit im Mittel 252 Pixeln Augenabstand deutlich detailliertere Aufnahmen, die sich gut zur Untersuchung des Einflusses der Auflösung eignen.

2.3. Voruntersuchungen zum Auftreten von Schmerzreaktionen

Bei der Betrachtung der Videos der Datenbanken BioVid, X-ITE, UNBC und BP4D fiel dem Autor dieser Dissertation auf, dass trotz Schmerzstimulation bzw. vom Patienten berichtetem Schmerz oft keine Schmerzreaktionen zu beobachten sind. Dies passt zu den Aussagen anderer Autoren, dass von einem erheblichen Anteil der Probanden in verschiedenen Schmerzstudien keine Schmerzmimik gezeigt wurde [PC95; Wil95; KL14]. Ohne Schmerzreaktion im Video ist über diese Sensormodalität auch keine automatisierte Erkennung eines stimulierten oder berichteten Schmerzes möglich. Im Folgenden werden zunächst Methoden zur Abschätzung der Gesichtsaktivität vorgeschlagen. Diese werden in den darauffolgenden Abschnitten 2.3.2 und 2.3.3 genutzt, um bei den Datensätzen BioVid und X-ITE die Aktivität bei verschiedenen Intensitäten und Probanden abzuschätzen und qualitative Beobachtungen auch quantitativ zu belegen. Abschnitt 2.3.2 zeigt außerdem die zeitliche Charakteristik der Schmerzreaktionen im Gesicht und beschreibt die Erstellung des Einzelbilddatensatzes BioVid-S. Abschnitt 2.3.3 beschreibt die Erstellung der Teildatensätze BioVid-A7, BioVid-D7, BioVid-S7 und X-ITE-E46, die jeweils nur die expressivsten Probanden beinhalten. Der Zusammenhang der Grundwahrheiten bei UNBC wird in Abschnitt 2.3.4 untersucht und diskutiert. In Abschnitt 2.3.5 wird die Leistungsfähigkeit von menschlichen Beobachtern thematisiert. Die Texte basieren in weiten Teilen auf Werner et al. [WAHW17].

2.3.1. Methoden zur Abschätzung der Gesichtsaktivität

Um abzuschätzen, inwieweit mimische Schmerzreaktionen auftreten, können verschiedene Maße herangezogen werden. Zum einen können diese auf automatisierter Bildanalyse aufbauen, wie dem optischen Fluss, zum anderen auf den Einschätzungen menschlicher Beobachter.

Optischer Fluss: Zur vollständig automatisierten Bild-für-Bild-Abschätzung der Gesichtsaktivität wird folgendes Vorgehen vorgeschlagen:

1. Das Gesicht wird in der frontalen Kameraansicht mit dem Viola-Jones Gesichtsdetektor von OpenCV [LKP03] detektiert. Um den Hintergrund auszuschließen wird nur der mittlere Teil der bounding box benutzt (60% der Breite und 80% der Höhe).
2. Diese Region wird in ein regelmäßiges 5×5 Raster unterteilt. In jeder Zelle wird der robuste Median Flow Tracker von Kalal et al. [KMM10] angewendet: Hierbei wird für 100 Punkte je Zelle der optische Fluss berechnet (mit dem Lucas-Kanade Algorithmus). Der Fluss wird vorwärts und rückwärts berechnet (jedes der beiden Bilder wird einmal als erstes Bild verwendet) und der Fehler zwischen beiden Ergebnissen berechnet, um Ausreißer zu entfernen, d. h. Vektoren mit wahrscheinlich fehlerhafter Bewegungsschätzung. Dazu werden die Vektoren, deren Fehler den Median der Gesamtmenge überschreiten, ignoriert. Weitere Ausreißer werden mithilfe eines ähnlichen Kriteriums basierend auf der normierten Kreuzkorrelation entfernt. Die übrigen Flussvektoren werden für jede Zelle in einen Bewegungsvektor kombiniert, indem der Median in x- und y-Dimension berechnet wird.
3. Es wird die Länge aller 25 Bewegungsvektoren berechnet und das ihr Maximum als Abschätzung der aktuellen Gesichtsaktivität als ein Punkt in der Zeitreihe des Videos gespeichert.

Abb. 2.6 veranschaulicht die beschriebene Schätzung des Flusses. Um die Berechnung zu beschleunigen, wurden die Bilder vor der Verarbeitung auf die Hälfte der Auflösung reduziert. Bei BioVid wurden neben allen frontalen Videos von BioVid-A auch zusätzliche `BLN`-Samples extrahiert, um den Einfluss gelegentlicher Bewegungen während der schmerzfreien Pausen zu reduzieren. Untersucht wurden somit 80 `BLN`-Samples und 80 Schmerz-Samples je Proband.

Das Verfahren hat den Vorteil, *vollständig automatisiert detaillierte Zeitreihen* der Bewegung für das gesamte Datenmaterial zu liefern. Eine Limitierung ist, dass keine Differenzierung zwischen rigiden Kopfbewegungen und Bewegungen durch Mimik stattfindet.

PSPI: Die *Prkachin and Solomon Pain Intensity (PSPI)* [Luc+11b] (vgl. Abschnitt 2.1.4) ist ein validiertes [PS08] und oft verwendetes [Wer+19b] Maß für Schmerzmimik. Für die Untersuchungen auf Videoebene wird hier der Maximalwert der PSPI aller Einzelbilder des Videos herangezogen, wie auch bei der Validierung von PSPI [PS08]. PSPI steht nur für UNBC zur Verfügung, sowie für einen Teil von BioVid, der mit FACS codiert wurde (1 Video je Intensität und Proband, d. h. 435 Videos bestehend aus 60.030 Einzelbildern aus der frontalen Ansicht von BioVid-A) [LE+16; Wer+17]. FACS und PSPI sind sehr objektiv, die Codierung ist jedoch sehr arbeitsaufwändig.

Subjektive Beurteilung jedes Samples: Als weitere Alternative wird vorgeschlagen, dass ein menschlicher Beurteiler eine subjektive Kategorisierung für jedes Sample vornimmt. Beim Datensatz UNBC fällt die Grundwahrheit OPR in diese Kategorie. Um ähnliche Informationen für den Datensatz BioVid zu sammeln, wurde wie folgt vorgegangen:

Zur Vorbereitung werden alle Videos bzw. Bilder in zufällige Reihenfolge gebracht und durchnummeriert in einem Dateisystemordner abgespeichert, so dass die Klassenzugehörigkeit nicht mehr ersichtlich ist. Die Zuordnung der Nummern zu den ursprünglichen Dateien wird für die spätere Auswertung in einer Textdatei gespeichert, die dem Beurteiler nicht zugänglich ist. Für jede Klasse wird ein Unterordner angelegt. Anschließend sieht der Beurteiler die Dateien in vorgegebener Reihenfolge an. Nach dem einmaligen Betrachten jeder Datei ordnet er diese unmittelbar einer Klasse zu, indem er sie in den entsprechenden Unterordner verschiebt. Nachdem alle Dateien zugeordnet worden sind, werden die vom Beurteiler zugeordneten Klassen mit den

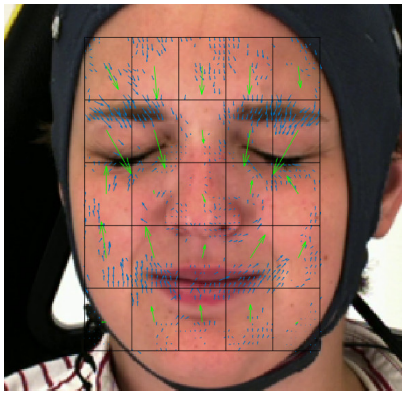
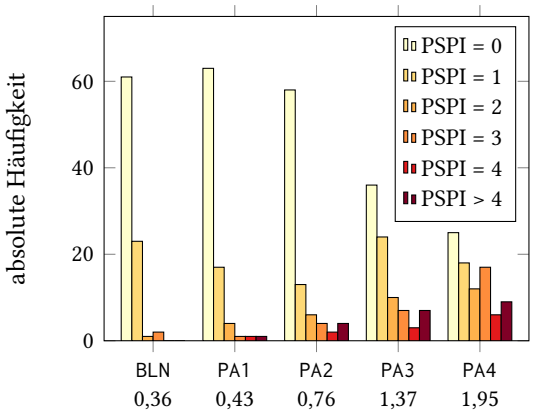
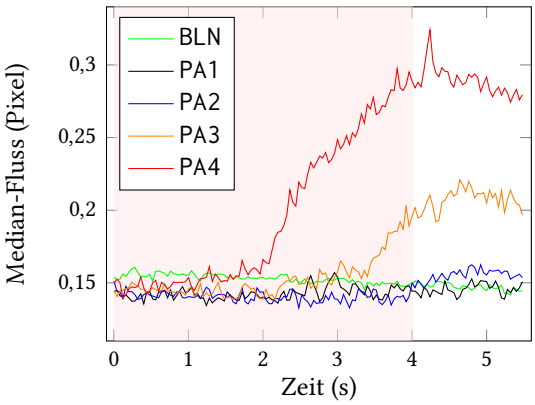


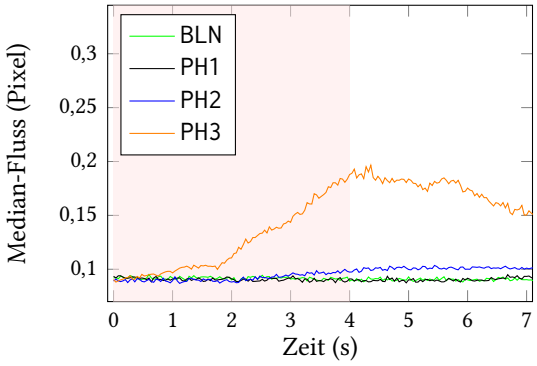
Abbildung 2.6.: Optischer Fluss zur Abschätzung der Gesichtsaktivität, geschätzt mit dem Median Flow Tracker [KMM10] in einem Raster von 5 × 5 Zellen (schwarz). Aus den gültige Flussvektoren (blau) wird für jede Zelle ein Bewegungsvektor berechnet (grün, skaliert für bessere Sichtbarkeit). Für jedes Videobild wird die maximale Vektorlänge der 25 Zellen als Teil einer Zeitreihe gespeichert, die die Gesichtsaktivität abschätzt. © 2017 IEEE [WAHW17].



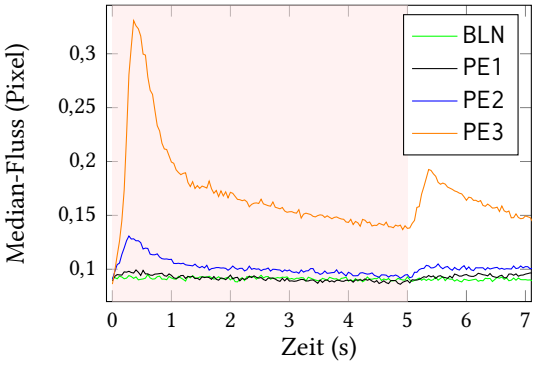
(a) BioVid: Verteilung der maximalen PSPI-Werte und deren Mittelwerte (unten) bei Hitzereizen (PA∗) und Pausen (BLN).



(b) BioVid: Median-Zeitreihen des optischen Flusses bei Hitzeschmerzreizen (PA∗) und Pausen (BLN).



(c) X-ITE: Median-Zeitreihen des optischen Flusses bei Hitzeschmerzreizen (PH∗) und Pausen (BLN).



(d) X-ITE: Median-Zeitreihen des optischen Flusses bei elektrischen Schmerzreizen (PE∗) und Pausen (BLN).

Abbildung 2.7.: Gesichtsaktivität in Abhängigkeit von Reizintensität und Zeit für den BioVid Datensatz (oben) und den X-ITE Datensatz (unten). Die Zeitreihen sind zeitlich an dem Plateau der Hitzestimulation bzw. der elektrischen Stimulation (beides roter Hintergrund) ausgerichtet. (a) und (b) © 2017 IEEE [WAHW17].

tatsächlichen Klassenzugehörigkeiten verglichen und die Übereinstimmung berechnet. Im Anschluss an die Beurteilung wird mittels Intra-Klassen-Korrelation (ICC), dem auch im Rest der Arbeit bevorzugten Maß (siehe Anhang A), die Übereinstimmung der Sample-Beurteilungen mit der durch die Schmerzstimulation gegebenen Klasse gemessen.

Die beschriebene Methode wurde für die Klassen BLN (Baseline, kein Schmerz) und PA4 (höchste Schmerzintensität) in Frontalansicht angewendet, sowohl für *BioVid-S* als auch für *BioVid-D*. Zunächst wurde aus jedem Video von BioVid-D ein Einzelbild extrahiert, um BioVid-S zu erzeugen (Details im folgenden Abschnitt). Der Beobachter hat anschließend jedes der 3.840 Einzelbilder einer der Klassen zugeordnet, analog zur in Kapitel 3 thematisierten automatisierten Erkennung anhand von Einzelbildern. Später wurden die 5,5 Sekunden langen Videos von BioVid-D (insgesamt 3.840) betrachtet und Klassen zugeordnet, analog zur automatisierten videobasierten Erkennung (Kapitel 4).

Die Bilder von BioVid-S wurden von zwei Beobachtern beurteilt: Beobachter 1 war der Autor dieser Dissertation. Er verfügt zwar nicht über klinische Erfahrung in der Schmerzbeurteilung, jedoch über umfangreiches theoretisches Wissen zur Schmerzmimik und Erfahrung aus der Betrachtung von Schmerzdatenbanken. Um ein eventuelles positives Bias in der Übereinstimmung der Beurteilung mit den tatsächlichen Labeln zu reduzieren, hat der Autor ab zwei Wochen vor Beginn der Annotation nicht mit den Daten gearbeitet und bis nach Abschluss der Annotation keine Label zu Bildern bzw. Videos gesehen. Beobachter 2 hatte keine besondere Ausbildung oder praktische Erfahrung in der Schmerzbeurteilung. BioVid-D wurde vom Autor beurteilt.

Ebenfalls vom Autor dieser Dissertation wurde eine Beurteilung von zufällig gewählten Videos des Datensatzes X-ITE vorgenommen. Hierbei wurden jeweils 1.000 Videos des Teildatensatzes mit phasischen Hitzereizen sowie 1.000 Videos des Datensatzes mit phasischen elektrischen Reizen ausgewählt und zugeordnet. Die Videos stammten dabei von allen Klassen sowie von der frontalen und seitlichen Kamera, wie bei der videobasierten Schmerzmessung in Abschnitt 4.3.2.

Die hier beschriebene Methode zur Beurteilung der Schmerzmimik jedes Samples ist weniger objektiv als FACS und PSPI, erfordert jedoch auch weniger Arbeitsaufwand und vorherige Ausbildung. Sie ermöglicht außerdem eine Einschätzung des menschlichen Leistungsvermögens bei der Erkennungsaufgabe und könnte auch als Beobachtergrundwahrheit verwendet werden.

Subjektive Kategorisierung der Probanden: Als schnellere Alternative zur Beurteilung jedes Samples wird die subjektive Kategorisierung der *Probanden* vorgeschlagen. Hierbei beurteilt ein Beobachter die mimische Expressivität eines jeden Probanden anhand einer zufälligen Stichprobe von etwa 50% der Videos mit der höchsten Schmerzintensität. Er ordnet den Probanden auf einer Skala vom Likert-Typ von 1 (keine Schmerzreaktion beobachtet) bis 5 (deutliche Schmerzreaktion in fast allen Videos) ein. Diese Annotation wurde wie die subjektive Beurteilung jedes Samples (siehe oben) vom Autor dieser Dissertation vorgenommen, für BioVid jedoch bereits ca. zwei Jahre zuvor abgeschlossen. Für den X-ITE-Datensatz mit elektrischer Reizung wurde ebenfalls eine solche Beurteilung vorgenommen, über 1 Jahr vor der subjektiven Beurteilung der einzelnen Samples.

Diese Methode zur Einschätzung der Expressivität ist weniger objektiv als FACS und weniger zuverlässig als die Beurteilung jedes einzelnen Samples. Sie ist jedoch deutlich weniger zeitaufwändig und sollte ausreichen, um eine grobe Kategorisierung der Probanden vorzunehmen.

2.3.2. Aktivität in Abhängigkeit von Reizintensität und Zeit

Im Folgenden wird für die Datensätze BioVid und X-ITE die Gesichtsaktivität bei verschiedenen Reizintensitäten (Klassen) verglichen. Abb. 2.7a veranschaulicht die Unterschiede bei BioVid-A anhand von Histogrammen der PSPI-Werte. Bei niedrigen Schmerzintensitäten sind mimische Schmerzreaktionen selten: Weit weniger als die Hälfte der Videos von PA1 und PA2 haben einen maximalen PSPI-Wert > 0. Zusätzlich steht PSPI = 1 in vielen Fällen für keine Gesichtsaktivität, da einige Probanden ihre Augen permanent geschlossen hatten [Wer+17], auch bei BLN (Pause ohne Schmerz). PSPI-Werte ≥ 2 treten bei höheren Reizintensitäten auch häufiger auf und bilden die Intensität der Schmerzmimik gut ab. Die Histogramme zeigen jedoch auch, dass selbst bei PA3 und PA4 sehr häufig PSPI = 0 vorkommt, d. h. keine typische Schmerzmimik gezeigt wird.

Für weitergehende Betrachtungen werden die Zeitreihen des optischen Flusses herangezogen (vgl. vorheriger Abschnitt). Die an der Stimulation zeitlich ausgerichteten Zeitreihen werden für jede Klasse separat betrachtet und jeweils die Median-Zeitreihe berechnet, d. h. die zeitliche Auflösung wird beibehalten und für jeden Zeitpunkt der Median aller Samples dieser Klasse berechnet. In der Median-Zeitreihe wird die gemeinsame Tendenz alles Samples sichtbar, auch wenn sich die Aktivität von Sample zu Sample stark unterscheidet. Abb. 2.7b vergleicht die zeitlich aufgelösten Gesichtsaktivitäten der Klassen bei BioVid. Jeder Punkt der Kurven repräsentiert den Betrag des Flusses, der von 50% der Samples dieser Klasse übertroffen wird. Die Beobachtungen passen zu denen der PSPI-Werte: Bei PA4 und PA3 sind die Gesichtsaktivitäten klar erkennbar. Die Aktivität bei PA2 übersteigt nur knapp BLN (am Ende des Zeitfensters) und PA1 is abgesehen von leichtem Rauschen konstant. Wie erwartet steigen die mimischen Reaktionen mit der Intensität der Schmerzreizung an. Am Anfang der Zeitfenster wird bei BLN jedoch sogar eine höhere Aktivität gemessen, als bei den Schmerzklassen (mit fallender Tendenz). Dies deutet auf ein Bias bei der Wahl der Baseline-Samples hin. Die Zeitfenster dieser Samples folgen Schmerzreizungen, die Gesichtsaktivität auslösen. Diese ist bei vielen BLN-Samples noch am ausklingen, was auch an der abfallenden Aktivität sichtbar wird. Wenn der nächste Schmerzreiz beginnt, hat die Aktivität oft einen niedrigeren Wert erreicht als während eines BLN-Samples. Ein solches Bias sollte in zukünftigen Schmerzstudien vermieden werden, indem zwischen den Reizen längere Pausen gelassen und die BLN-Samples vielfältiger gewählt werden.

Abb. 2.7b zeigt auch die Verzögerung zwischen Reiz und Reaktion, die mit steigender Intensität kürzer wird. Eine höhere Temperatur führt zu einer schnelleren Hitzeübertragung von der Thermode auf die Haut, so dass die Haut früher aufgeheizt ist. Auch führen höhere Intensitäten tendenziell zu schnelleren Reaktionen, um dem schädlichen Reiz zu entkommen.

Die Abb. 2.7 (c) und (d) zeigen die Median-Zeitreihen für X-ITE, mit Hitzereizen und mit elektrischen Reizen. Bei den Hitzereizen (Abb. 2.7c) ist die Aktivität ähnlich zu BioVid, insbesondere bei der höchsten Intensität PH3. Bei PH2 ist die Aktivität jedoch niedriger als bei der zweithöchsten Intensität von BioVid. Die niedrigste Intensität PH1 und BLN zeigen beide keine Aktivität und sind hier kaum unterscheidbar. Die elektrischen Reize (Abb. 2.7d) führen im Vergleich mit den Hitzereizen zu stärkeren Reaktionen, die unmittelbar nach dem Beginn des Reizes beginnen. Häufig werden dabei die Gesichtsmuskeln sehr schnell angespannt (erstes Maximum der Aktivität) und erst nach Ende des Reizes entspannt (zweites Maximum). Die qualitativen Unterschiede der Reize und Reaktionen lassen sich auch mit automatisierten Methoden erkennen, was Werner et al. [Wer+19a] gezeigt haben, in dieser Arbeit jedoch nicht weiter adressiert wird. Auch wenn die Reaktionen auf elektrische Reize insgesamt stärker ausfallen als bei den Hitzereizen von X-ITE, hebt sich auch hier die niedrigste Intensität PE1 kaum von BLN ab.

Die Ergebnisse deuten bei allen drei Datensätzen darauf hin, dass die Mehrheit der Samples der niedrigsten Reizintensität (PA1, PH1, PE1) keine mimische Schmerzreaktion beinhalten. Zudem

scheint bei Hitzereizen auch die nächstgrößere Intensität (PA2 und PH2) in den meisten Fällen mit keiner oder schwachen mimischen Schmerzreaktionen einherzugehen.

Einzelbildbasierter BioVid-Datensatz (BioVid-S): Zur späteren Gegenüberstellung von einzelbildbasierten und videobasierten Erkennungsmethoden wird zum Videodatensatz BioVid-D, der Teilmenge von BioVid-A mit den Klassen BLN und PA4, ein entsprechender Einzelbilddatensatz *BioVid-S* erstellt. Da für BioVid keine einzelbildbasierte Grundwahrheit vorliegt (FACS wurde nur für 5% der Videos codiert), wird von jedem Video *ein* Einzelbild gewählt. Die Wahl des Bildes orientiert sich an den Median-Zeitreihen des optischen Flusses der Klassen PA4 und BLN in Abb. 2.7b, welche die typische Gesichtsaktivität in den Videos abschätzen. Die Einzelbilder für BioVid-S wurden von Sekunde 4 im Abb. 2.7b extrahiert, was Sekunde 3 in den Videos von BioVid-D entspricht, da diese 1 Sekunde nach Erreichen des Hitzeplateaus beginnen. Zum gewählten Zeitpunkt endet das Plateau, d. h. die Hitze auf der Haut nimmt danach ab, und die Unterschiede zwischen der Gesichtsaktivität von PA4 und BLN erreichen ihr Maximum (abgesehen von zufälligen Schwankungen in der Kurve).

2.3.3. Aktivität der Probanden

Wie bereits mehrfach angesprochen gibt es viele personenspezifische Faktoren, die Schmerzreaktionen und Schmerzempfinden beeinflussen. Auch wurde von mehreren Autoren berichtet, dass ein Teil der Probanden ihrer Studien gar keine mimische Schmerzreaktionen gezeigt haben [PC95; Wil95; KL14]. Der Grad, inwieweit die Probanden Schmerzreaktionen zeigen (hier Expressivität genannt), wird im Folgenden mit dem Datensatz BioVid anhand der Maße der Gesichtsaktivität untersucht, die in Abschnitt 2.3.1 vorgestellt wurden. Um die Aussagekraft der Maße zu zeigen, wird die Konvergenzvalidität [Him07, S. 383f] untersucht, d. h. die Korrelation der unterschiedlichen Maße des selben Konstrukts Expressivität berechnet. Folge Maße werden betrachtet:

Fluss *p*: Zunächst wird von dem Betrag des optischen Flusses, der je Einzelbild zur Schätzung der Gesichtsaktivität vorliegt, der Mittelwert jedes Videos berechnet. Für jede Klasse ergibt sich je Proband eine Verteilung dieser Flussmittelwerte. Nun wird für jeden Probanden überprüft, ob während PA4 eine höhere Aktivität beobachtet wurde als während BLN. Hierfür wird ein Permutationstest [Moo+03] eingesetzt. Es wird von der Nullhypothese ausgegangen, dass der Proband bei BLN und PA4 eine gleichstarke Aktivität zeigt, und der p-Wert berechnet, der angibt, wie wahrscheinlich es ist, den beobachteten oder einen noch stärkeren Unterschied zwischen BLN und PA4 zu finden, wenn die Nullhypothese wahr ist. Je geringer der Wert von p, desto mehr widersprechen die Beobachtungen der Nullhypothese und desto stärker bestätigen sie die Hypothese, dass PA4 mit einer höheren Aktivität verbunden ist als BLN. Zu Permutationstest siehe auch Abschnitt 3.1.2.

PSPI PA4 und PSPI SD: Für den BioVid-Datensatz liegt ein PSPI-Wert für jede Klasse bei jedem Probanden vor (maximaler PSPI-Wert des Videos). Zur Abschätzung der Expressivität eines Probanden werden zwei Maße von PSPI betrachtet: PSPI im FACS-codierten Sample der Klasse PA4 des Probanden (PSPI PA4) sowie die Standardabweichung der PSPI-Werte aller fünf Klassen des Probanden (PSPI SD).

BR-ICC: Die Übereinstimmung von Beobachtereinschätzung (B) und angewendetem Reiz (R) wird für jeden Probanden als ICC-Wert berechnet und kann der Abschätzung seiner Expressivität dienen (siehe Anhang A zur Definition des ICC-Maßes). Die Beobachter-Reiz-Übereinstimmung für einen Probanden kann für jede der zwei Beobachtereinschätzungen

	Fluss p	PSPI PA4	PSPI SD	BR-ICC B. 1	BR-ICC B. 2	BR-ICC V.
PSPI PA4	-0,476					
PSPI SD	-0,465	0,876				
BR-ICC Bild 1	-0,507	0,629	0,605			
BR-ICC Bild 2	-0,509	0,561	0,577	0,849		
BR-ICC Video	-0,562	0,706	0,717	0,825	0,798	
Subj. Kat. d. Prob.	-0,607	0,683	0,634	0,777	0,709	0,797

PSPI PA4: Prkachin and Solomon Pain Intensity bei höchster Reizintensität PSPI SD: Standardabweichung der PSPI-Werte des Probanden BR-ICC: Beobachter-Reiz-Übereinstimmung (Intra-Klassen-Korrelation)

Tabelle 2.2.: Korrelation von Maßen der Gesichtsaktivität bei BioVid. Die aufgeführten Maße wurden für jeden Probanden der BioVid Datenbank berechnet. Anschließend wurde die Pearson-Korrelation der Maße bestimmt. Alle Korrelationen sind statistisch signifikant mit $p < 0{,}001$ (zu Signifikanztests siehe Abschnitt 3.1.2).

des Bilddatensatzes BioVid-S (BR-ICC Bild 1 und 2) sowie für die Beobachtereinschätzung des Videodatensatzes BioVid-D (BR-ICC Video) berechnet werden.

Subjektive Kategorisierung des Probanden: Hier liegt aus Abschnitt 2.3.1 für jeden Probanden bereits genau eine Zahl vor, welche seine Expressivität einschätzt.

Tabelle 2.2 listet die Korrelationen der Maße auf. Der p-Wert des Flusses ist als einziges Maß negativ mit den anderen korreliert, da ein niedrigerer Wert für einen größeren Unterschied der Gesichtsaktivität zwischen BLN und PA4 spricht. Bei allen anderen Maßen spricht ein größerer Wert auch für eine größere Expressivität des Probanden. Die Maße sind zum Großteil stark korreliert (Betrag des Korrelationskoeffizienten größer 0,5). Lediglich einige Korrelationen des Flusses p sind betragsmäßig leicht schwächer ausgeprägt. Alle Korrelationen sind statistisch hochgradig signifikant, was als Beleg für die Validität der Maße gesehen werden kann. Bemerkenswert sind insbesondere die hohen Korrelation der PSPI-Maße mit den BR-ICC-Maßen sowie mit der subjektiven Kategorisierung der Probanden, die mit unterschiedlichen Methoden und Beobachtern ermittelt wurden. Mit 0,876 besonders hoch ist die Korrelation von PSPI von PA4 und der Standardabweichung von PSPI (PSPI SD), die jedoch auf der gleichen FACS-Codierung basieren. Auf den Einschätzungen unterschiedlicher Beobachter beruhen die Expressivitätsbewertung „BR-ICC Bild 1“ und „BR-ICC Bild 2“, die mit 0,849 ebenfalls sehr stark korreliert sind. Auch sehr stark korreliert ist die Bewertung auf Basis der Videos aus BioVid-D (BR-ICC Video) mit den beiden bildbasierten Bewertungen aus BioVid-S (BR-ICC Video), was dafür spricht, dass zur Entnahme der Bilder für BioVid-S aus den Videos von BioVid-D ein geeigneter Zeitpunkt gewählt wurde.

In der FACS-codierten Teilmenge von BioVid zeigen 22% der Probanden gar keine Schmerzmimik nach PSPI, 8% zeigen Schmerzmimik lediglich für PA4, 16% ab der Intensität PA3, 15% ab der Intensität PA2 und lediglich 9% bei allen Reizintensitäten [Wer+17]. Diese Beobachtungen stimmen mit dem Modell der Schmerzkommunikation von Prkachin und Craig [PC95] überein, nach dem eine Schmerzempfinden einen personenabhängigen Intensitätsschwellwert überschreiten muss, um eine mimische Reaktion auszulösen. Bei BioVid und X-ITE könnte eine andere Ursache für das Fehlen von Schmerzmimik auch sein, dass die personenspezifische Kalibrierung der Reize insofern fehlgeschlagen ist, dass keine Schmerzen empfunden wurden bzw. weniger starke Schmerzen als nach dem Versuchsdesign vorgesehen. In diesem Zusammenhang spielt auch die personenunabhängige Maximaltemperatur eine Rolle, auf die die Hitzereizung beschränkt werden musste, um Verbrennungen der Haut auszuschließen.

Teildatensätze der expressivsten Probanden: Bei einige Probanden im Datensatz BioVid ist auch bei der stärksten Schmerzstimulation keine mimische Schmerzreaktion zu beobachten, was sich sowohl beim Betrachten der Videos als auch in den verschiedenen Maßen der Expressivität zeigt. Für einen solchen Probanden sind die Samples der Klasse PA4 nicht von denen der Klassen BLN zu unterscheiden und wirken für das maschinelle Lernen auf Basis der Mimik wie fehlerhafte Annotationen (Label-Rauschen). Um den Einfluss niedriger Expressivität und dem damit einhergehenden Label-Rauschen zu untersuchen, werden zusätzlich zu den vollständigen Datensätzen auch Teildatensätze betrachtet, die nur die expressivsten Probanden beinhalten. Zur Auswahl der expressivsten Probanden wird die subjektive Kategorisierung der Probanden herangezogen (vgl. Abschnitt 2.3.1) und die Probanden mit der höchsten Bewertung ausgewählt.

Für BioVid sind das sieben Probanden[1]. Für jede der Varianten BioVid-A, BioVid-D und BioVid-S wird in den Experimenten auch zusätzlich diese Teilmenge besonders expressiver Probanden betrachtet. Zur besseren Unterscheidbarkeit werden die Teilmengen BioVid-A7, BioVid-D7 und BioVid-S7 genannt.

Für X-ITE mit elektrischen Reizen hat sich aufgrund einer anderen Verteilung in der subjektiven Kategorisierung eine größere Teilmenge von 46 Probanden ergeben[2]. Der entsprechende Teildatensatz der 46 expressivsten Probanden wird als X-ITE-E46 bezeichnet.

2.3.4. Zusammenhang der Grundwahrheiten bei UNBC

Für den Datensatz UNBC liegen mehrere Grundwahrheiten vor und es kann die Korrelation von Selbsteinschätzung (VAS), Beobachtereinschätzung mit Likert-Typ Skala (OPR) und Beobachtereinschätzung mit PSPI berechnet werden. Bei PSPI wird jeweils der maximale PSPI-Wert je Video betrachtet, der auch zur Validierung des PSPI-Maßes genutzt wurde [PS08]. Abb. 2.8 zeigt die Zusammenhänge zwischen den Maßen. Alle Maße sind stark miteinander korreliert. Die Pearson-Korrelation der beiden Beobachtereinschätzungen OPR und PSPI (rechts) ist am höchsten. Beide basieren auf den sichtbaren Verhaltensreaktionen und haben daher die gleiche Basis. So bewerten beide Maße 85 von insgesamt 200 Videos mit der Schmerzintensität 0 (kein Schmerz), obwohl nach der Selbsteinschätzung die Patienten in lediglich 35 Videos keine Schmerzen empfunden haben (VAS = 0). Somit wurden von Beobachtern häufig keine Schmerzen erkannt, obwohl welche empfunden wurden. Im linken und mittleren Plot sieht man, dass dies (PSPI/OPR = 0) vor allem bei niedrigen VAS-Werten vorkommt, d. h. bei schwachen Schmerzintensitäten, die oftmals nicht zu Verhaltensreaktionen führen. Jedoch auch höhere Schmerzintensitäten (z. B. VAS $\geq$ 5) wurden zum Teil mit PSPI/OPR = 0 bewertet. Unter der Annahme, dass die Patientenäußerungen valide sind, legt das die Vermutung nahe, dass hier eine geringe Expressivität vorlag, d. h. die Schmerzen nicht gezeigt wurden. Auch wenn die Schmerzmaße insgesamt stark korreliert sind, zeigen sich hier Inkonsistenzen in den Grundwahrheiten, die auf Schwächen in der Validität hinweisen.

2.3.5. Leistungsfähigkeit des menschlichen Beobachters

Die subjektive Beurteilung jedes Samples, die für Teile der Datensätze BioVid und X-ITE vorgenommen wurde (vgl. Abschnitt 2.3.1), ermöglicht die Einschätzung der Leistungsfähigkeit von

[1] Kürzel der expressivsten Probanden in BioVid: „071313_m_41“, „073114_m_25“, „081014_w_27“, „081609_w_40“, „091809_w_43“, „100909_w_65“, „101814_m_58“

[2] Kürzel der expressivsten Probanden in X-ITE (elektrisch): „S001“, „S005“, „S006“, „S008“, „S011“, „S013“, „S017“, „S018“, „S021“, „S024“, „S025“, „S029“, „S031“, „S032“, „S034“, „S035“, „S036“, „S038“, „S040“, „S043“, „S044“, „S049“, „S051“, „S054“, „S057“, „S058“, „S060“, „S061“, „S066“, „S069“, „S070“, „S072“, „S073“, „S075“, „S078“, „S084“, „S090“, „S098“, „S104“, „S105“, „S106“, „S109“, „S110“, „S111“, „S115“, „S127“

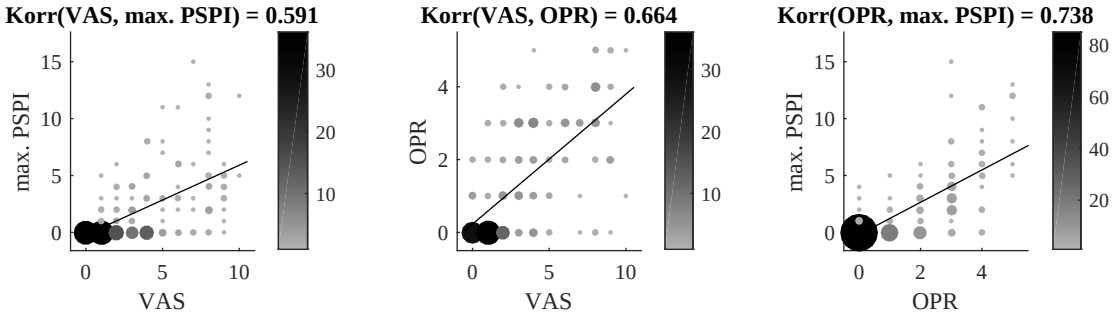


Abbildung 2.8.: Zusammenhang von Grundwahrheiten bei UNBC, zwischen VAS (Selbsteinschätzung), OPR und maximalem PSPI-Wert je Video. Die Größe und Grauabstufung der Punkte repräsentiert die Häufigkeit einer Kombination, die Linie ein lineares Modell, das mit der Methode der kleinsten Quadrate berechnet wurde. Über den Plots sind die Korrelationskoeffizienten angegeben. Alle Korrelationen sind statistisch signifikant mit $p < 0{,}001$.

menschlichen Beobachtern bei verschiedenen Schmerzerkennungsaufgaben. Sie dient später als Referenzpunkt für die Bewertung der Ergebnisse von automatisierten Erkennungssystemen, da aufgrund oftmals „fehlender“ mimischer Schmerzreaktionen eine fehlerfreie Erkennung auf Basis der Mimik nicht möglich sein wird.

Bei BioVid und X-ITE wird zur Berechnung der Leistungsfähigkeit die Übereinstimmung der vom Beobachter zugeordneten Klassen mit der tatsächlich angewendeten Schmerzstimulation (Reizintensität) berechnet. Die Übereinstimmung wird mit der Intra-Klassen-Korrelation (ICC) gemessen, die auch im Rest der Arbeit bevorzugt verwendet wird (siehe Anhang A). Der ICC-Wert wird dabei nicht wie in Abschnitt 2.3.3 für jeden Probanden, sondern direkt anhand aller Samples bestimmt. Abb. 2.9 stellt die Performance-Werte gegenüber. Gezeigt werden die Ergebnisse der vollständigen Datensätze (blau) und der Teildatensätzen der 7 expressivsten Probanden (rot). Wie zu erwarten war, verbessert sich die Performance deutlich durch die Fokussierung auf die Probanden, die ihre Schmerzen am deutlichsten in der Mimik zeigen. Ebenfalls verglichen werden die ICC-Werte, die mit den Einzelbildern (BioVid-S/S7, oben und mittig) und den Videos (BioVid-D/D7, unten) erreicht wurden. Hier zeigt sich, dass die einzelbildbasierte Erkennung der Schmerzen schwieriger ist als die videobasierte Erkennung, bei der höhere ICC-Werte erreicht wurden. Für Beobachter 1 (dem Autor dieser Dissertation) wurde eine größere Übereinstimmung zwischen Beurteilung und Schmerzreizen gemessen als für Beobachter 2, der weniger Erfahrung mit Schmerzmimik hatte. Dies zeigt, dass die Ausbildung und Erfahrung für die Leistungsfähigkeit der menschliche Beobachter eine Rolle spielt.

Für X-ITE wurden sowohl für den Teildatensatz mit Hitzereizen als auch den mit elektrischen Reizen jeweils 1.000 zufällig gewählte Videos beider Ansichten (frontal und seitlich) und aller Klassen betrachtet. Die Tabellen 2.3 (a) und (b) listen die Performance-Werte und Konfusionsmatrizen auf. Für die elektrischen Reize ist der ICC-Wert mit 0,451 leicht höher als für die Hitzereize. Die Konfusionsmatrizen zeigen, dass die Klassen `PH1`, `PH2`, `PE1` und `PE2` (niedrigere und mittlere Reizintensität) sehr häufig der Klasse `BLN` (kein Schmerz) zugeordnet wurden, da in den Videos oft keine Schmerzreaktionen zu beobachten waren. Bei den höchsten Schmerzintensitäten `PH3` und `PE3` wurden mehr Samples korrekt zugeordnet, da hier deutlich mehr Schmerzreaktionen auftreten, jedoch gibt es auch hier eine große Anzahl von Verwechslungen mit `BLN`, der mittleren und der niedrigen Schmerzintensität.

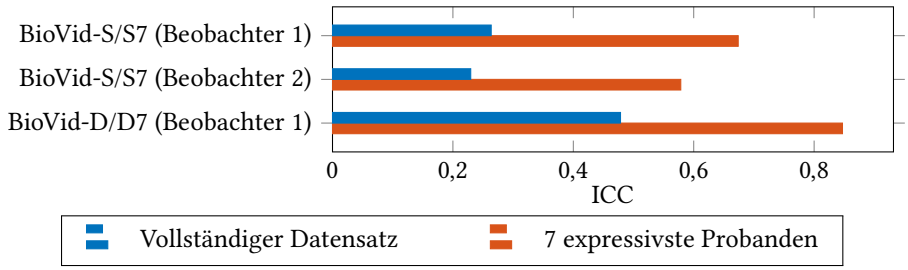


Abbildung 2.9.: Performance der menschlichen Beobachter auf BioVid (frontale Ansicht, Erkennung starker Schmerzen): Die Ergebnisse ausgehend von Einzelbildern (BioVid-S/S7, oben und mittig) sind schlechter als von Videos (BioVid-D/D7, unten). Die Reduzierung des Datensatzes auf die 7 expressivsten Probanden (rot) verbessert die Ergebnisse verglichen mit dem ganzen Datensatz (blau).

	Prädizierte Klasse			
	BLN	PH1	PH2	PH3
BLN	182	70	17	0
PH1	156	51	19	1
PH2	158	64	27	15
PH3	60	44	56	80

(a) Hitzereizung.
ICC = 0,435 / Accuracy = 0,340.

	Prädizierte Klasse			
	BLN	PE1	PE2	PE3
BLN	186	55	7	1
PE1	169	67	17	1
PE2	129	75	25	16
PE3	75	48	52	77

(b) Elektrische Reizung.
ICC = 0,451 / Accuracy = 0,355.

Tabelle 2.3.: Performance des menschlichen Beobachters auf X-ITE, veranschaulicht mit Konfusionsmatrizen. Diese geben an, wie viele Samples jeder Klasse (Zeilen) vom Beobachter welcher Klasse zugeordnet wurde (Spalten). Die Hintergrundfarbe ist proportional zu den angegegeben Werten gewählt.

Für den Datensatz UNBC gibt die Übereinstimmung der Grundwahrheiten VAS und OPR an, inwieweit menschliche Beobachter die empfundenen Schmerzen bzw. die Selbstbeurteilung der Patienten abschätzen können. Wie in Abb. 2.8 ersichtlich, liegt die Pearson-Korrelation bei 0,664. Werden die Wertebereiche von VAS und OPR durch Skalierung angeglichen, ergibt die Berechnung des ICC den Wert 0,663.

2.4. Diskussion

In diesem Kapitel wurde die Datengrundlage für das maschinelle Lernen in dieser Arbeit vorgestellt und untersucht. Verwendet werden die Mimikdatensätze BioVid, X-ITE, UNBC, BP4D und FERA 2017 sowie Bosphorus. Sie unterscheiden sich unter anderem hinsichtlich ihrer Größe, der vorkommenden Kopfposen sowie der verfügbaren Grundwahrheit. Von ihnen wird jeweils nur ein kleiner Teil des enorm großen Problemraumes abgebildet. Sehr begrenzt ist z. B. die Vielfalt der Probanden (hinsichtlich Ethnien, Alter und Anzahl der Individuen). Die Beleuchtungssituation ist nahezu immer ideal und Verdeckungen kommen nur selten vor, so dass zur Robustheit der Erkennung bezüglich dieser Faktoren keine Untersuchungen angestellt werden können.

Für die Erkennung von Gesichtsausdrücken, unabhängig von ihrer Deutung, stellt die objektive Beschreibung der Mimik anhand von Action Units (AUs) des Facial Action Coding Systems (FACS) [EFH02] die beste Grundlage dar. Die FACS-Codierung basiert auf der Auswertung von

Bildern oder Videos durch einen Experten anhand klar definierter Regeln, einem Prozess der mit einer hinreichenden Menge korrekt annotierter Daten gut in einem automatisierten Computer-Vision-System nachgebildet werden kann. Die Zuverlässigkeit der Codierung, d. h. die Übereinstimmung verschiedener FACS-Coder ist zumeist hoch [Luc+11b; Zha+14]. Somit sind die Herausforderungen vor allem technischer Natur, z. B. die begrenzte Verfügbarkeit von Daten bei einer großen Vielfalt möglicher Mimik und störender Faktoren (wie Kopfpose oder Identität).

Sollen anstatt klar definierter Muskelaktivitäten (FACS) Deutungskategorien wie *Schmerzmimik* klassifiziert werden, kommen die Unsicherheiten bzw. Unstimmigkeiten in der Definition dieser Kategorien hinzu. Die *Prkachin and Solomon Pain Intensity (PSPI)* [Luc+11b] ist ein Versuch, Schmerzmimik zu definieren und zu quantifizieren. Die mimischen Reaktionen auf Schmerzen sind jedoch verschieden und umfassen oft nur Teile des prototypischen Musters von PSPI oder andere AUs, die nicht Teil von PSPI sind [KL14]. Gleichzeitig sind die AUs von PSPI auch Teil anderer typischer Mimikmuster. Zu diesen ermöglicht PSPI keine Abgrenzung, so dass die Anwendung von PSPI außerhalb des Schmerzkontextes problematisch ist. Andererseits ist PSPI ein sehr objektives Maß, da es auf FACS basiert und klar definiert ist. Andere Beobachtereinschätzungen sind weniger klar definiert, deutlich subjektiver und stark vom Beobachter selbst und seinem Verhältnis zum Leidenden beeinflusst. Insofern wäre es vielversprechend, die Idee von PSPI weiterzuentwickeln, z. B. indem (1) AUs nicht-linear kombiniert werden, (2) der zeitliche Kontext einbezogen wird, und (3) AUs von Gesichtsausdrücken, die bei Schmerzen *nicht* vorkommen, „negativ“ in die Berechnung einbezogen werden (um Schmerzmimik besser von anderer Mimik abzugrenzen).

Alle Schmerzeinschätzungen, die sich nicht an objektiven, klar definierbaren Eigenschaften des Bildes bzw. Videos orientieren, fordern das maschinelle Lernen sehr heraus. Dieses sucht nach solchen gemeinsamen Eigenschaften, um eine deterministische Abbildung des Bildes bzw. Videos auf einen gewünschten Ausgabewert zu schaffen. Je weniger solche gemeinsame Eigenschaften zu finden sind, desto mehr ist die Optimierung zum Memorieren von Detailmustern gezwungen, z. B. zufälligen Gemeinsamkeiten bzw. Unterschieden von wenigen Trainingsbeispielen der gleichen bzw. verschiedenen Klassen, die so in den Testdaten nicht vorkommen. D. h. es kommt zu stärkerer Überanpassung und schlechteren Testergebnissen. Dieses Problem betrifft insbesondere die Grundwahrheiten Selbsteinschätzung und Schmerzreizung, die nicht auf der Beobachtung der Verhaltensreaktion beruhen. Wie in den vorherigen Abschnitten gezeigt und auch in der Literatur belegt [PC95; Wil95; KL14] führen empfundene Schmerzen nicht immer zu Verhaltensreaktionen. In diesen Fällen sind weder Menschen noch maschinelle Lernverfahren in der Lage, die empfundenen Schmerzen oder angewendete Schmerzstimulation zu erkennen, da sich die betrachteten Bilder oder Videos sich nicht systematisch von denen ohne Schmerzen bzw. Reizung unterscheiden. Dies ist insbesondere für niedrige Schmerzintensitäten der Fall, bei Personen mit geringer mimischer Expressivität jedoch auch für stärkere Schmerzen. Aus Sicht des Beobachters und des bildbasierten maschinellen Lernens wirken die Inkonsistenzen von Bild und Grundwahrheit wie falsche Label, wie *Label-Rauschen* (engl. label noise). Sie können einen negativen Einfluss auf das maschinelle Lernen haben und begrenzen die beste erreichbare Performance [FV14; AU20].

Um den Einfluss des Label-Rauschens zu untersuchen, werden in den folgenden Kapiteln zusätzlich zu den Gesamtdatensätzen BioVid-A, BioVid-S, BioVid-D und X-ITE auch Teildatensätze der expressivsten Probanden betrachtet (BioVid-A7, BioVid-S7, BioVid-D7 und X-ITE-E46). Die Ergebnisse mit Computer-Vision-Methoden werden dort auch mit den hier ermittelten Leistungsfähigkeiten von menschlichen Beobachtern verglichen. Letztere sind aufgrund der kleinen Anzahl von zwei Beobachtern nur begrenzt aussagekräftig bzw. nur für einen groben Vergleich

zwischen Mensch und Maschine geeignet. Zukünftige Arbeiten sollten für den Vergleich Beobachtereinschätzungen von mehreren medizinischen Fachkräften einholen, die in ihrer Arbeit mit Patienten regelmäßig Beobachterschmerzskalen anwenden.

Neben der Herausforderung, dass Schmerzreaktionen oft komplett fehlen, erschweren weitere Unterschiede in der Expressivität der Personen die Erkennung. Neben dem Schmerzschwellwert, der überschritten werden muss damit eine Reaktion gezeigt wird, unterscheiden sich die Personen auch hinsichtlich der Zunahme der Reaktion bei verstärkten Schmerzen. Kunz et al. [Kun+04] beschreiben diese zwei Aspekte als Schnittpunkt und Anstieg im Zusammenhang zwischen Reizintensität (bzw. Schmerzempfinden) und Schmerzreaktion. Personalisierung von Erkennungsmodellen kann bei der Problematik der verschiedenen Anstiege helfen [Wer+13; Wer+14b; LMRP17a], jedoch nicht das Problem fehlender mimischer Schmerzreaktionen kompensieren. Ein Ansatzpunkt für letztere Problematik ist die Hinzunahme weiterer Sensormodalitäten zur Erkennung. Biomedizinische Kontaktsensoren können auch bei niedrigeren Schmerzintensitäten oft Reaktionen messen, wo im Kamerabild keine Reaktion ersichtlich ist. Insbesondere der Hautleitwert und zum Teil auch EMG von mimischer Muskulatur liefern bei niedrigen Intensitäten bessere Ergebnisse [Wer+14b; Wer+19a]. Multimodale und personalisierte Ansätze zur Schmerzerkennung stehen jedoch nicht im Fokus der Arbeit und werden aus Platzgründen nicht weiter thematisiert.

Für die Erhebung neuer Datensätze wäre es sinnvoll, mehrere Grundwahrheiten zu erfassen, um deren Zusammenhänge und Validität zu untersuchen und eventuell neue zusammengesetzte Maße mit größerer Validität berechnen zu können. Label-Rauschen könnte in neuen experimentellen Datensätzen möglicherweise reduziert werden, indem andere Methoden der Schmerzreizung sowie verbesserte Methoden zur personenspezifischen Kalibrierung der Reizintensität angewendet werden. Am wichtigsten für die Weiterentwicklung der Schmerzerkennung ist jedoch die Erhebung von klinischen Schmerzdatensätzen. Wünschenswert wären hier Studien mit möglichst vielen Patienten, deren Schmerzen im klinischen Alltag erfasst werden, auch mit den dort anzutreffenden Verhältnissen bezüglich Beleuchtung und Verdeckungen, mit anderen Affektzuständen wie Angst oder Wut und mit Schmerzgrundwahrheiten gemäß der aktuell verbreiteten klinischen Messmethoden.

3. Einzelbildbasierte Erkennung

Dieses Kapitel beschäftigt sich mit der einzelbildbasierten Erkennung von Facial Action Units (AUs) und Schmerzmimik. Ausgehend von der Beschreibung der Grundlagen und verwandten Arbeiten (Abschnitt 3.1) werden Methoden zur einzelbildbasierten Erkennung von Mimik und zur Beantwortung der Forschungsfragen vorgeschlagen (Abschnitt 3.2). Zentrale Fragen sind: (1) Welche Computer-Vision- und Machine-Learnng-Methoden liefern die beste Performance? (2) Wie kann Kopfposeinvarianz erreicht werden, d. h. gute Erkennungsleistung unabhängig von der Kopfpose? (3) Kann die Erkennung mit niedrigen Hardwarekosten realisiert werden, d. h. ohne Spezialkameras, mit geringer Anzahl von Kameras und niedriger Auflösung? (4) Kann das entwickelte Erkennungssystem eine ähnlich gute Beurteilung der Schmerzen erreichen wie ein Mensch? In Abschnitt 3.3 werden die Methoden experimentell angewendet und evaluiert. Eine zusammenfassende Diskussion beschließt das Kapitel in Abschnitt 3.4.

3.1. Grundlagen und verwandte Arbeiten

In diesem Abschnitt werden zuerst, in Ergänzung zu den grundlegenden Ausführungen in Kapitel 1, detailliertere Einführungen zu den maschinellen Lernverfahren und den Evaluierungsmethoden gegeben, die in der restlichen Arbeit eine zentrale Rolle spielen. Anschließend werden verwandte Arbeiten vorgestellt, die Algorithmen zur Erkennung von Facial Action Units und Schmerzmimik vorschlagen und evaluieren, wobei insbesondere auf die verwendeten Merkmale und Lernverfahren eingegangen wird. Es folgt die Vorstellung verwandter Arbeiten zu zwei weiteren, abgrenzbaren Themen, zu denen die Dissertation Beiträge leistet: (1) zur Normierung des Gesichts und (2) zur Behandlung von Ungleichverteilungen der Klassenzugehörigkeit.

3.1.1. Maschinelle Lernverfahren

In Abschnitt 1.2.2 wurde bereits in die Thematik des maschinellen Lernen eingeführt. Hier werden nun die in der Arbeit verwendeten Lernverfahren vorgestellt. Für weitere Details wird der Leser auf die zitierte Literatur verwiesen.

Support Vector Machine (SVM) und Regression (SVR): Bei der SVM handelt es sich um ein maschinelles Lernverfahren, das zwei Klassen im Merkmalsraum durch eine spezielle Hyperebene trennt, durch genau die Hyperebene, die den größtmöglichen Abstand zu den am nächsten gelegenen Samples der beiden Klassen hat (engl. maximum margin) [Nob06]. Oft ist es in der Praxis nicht möglich eine Hyperebene zu finden, die alle Beispiele der zwei Klassen trennt, da sie sich im Merkmalsraum überschneiden, z. B. aufgrund von Messfehlern oder großer Ähnlichkeit der Klassen. Hier hilft die Idee des „soft margin“, die es einigen Datenpunkten („Ausreißern“) erlaubt auf der falschen Seite der trennenden Hyperebene zu liegen, ohne dass diese beeinflusst wird. Um steuern zu können, wie viele Ausreißer es geben darf und wie weit sie auf der falschen Seite

der Hyperebene liegen dürfen, wird der Hyperparameter C eingeführt. Für nicht-lineare Klassifikation wird die trennende Hyperebene in einem Raum gesucht, der höherdimensional als der eigentliche Merkmalsraum ist, und der implizit durch eine nicht-lineare Kernel-Funktion gegeben ist. Hierfür gibt es verschiedene Kernel-Funktionen, z. B. den RBF-Kernel (radiale Basisfunktion) oder polynomielle Kernel. Mit diesen Kerneln können bei nicht-linearen Klassifikationsproblemen oftmals bessere Ergebnisse erzielt werden als mit einer linearen SVM, jedoch kommen mit diesen ein oder mehrere Hyperparameter hinzu (z. B. γ bei RBF), die gut gewählt werden müssen, und das Training wird deutlich rechenaufwändiger. Die Hyperparameter können mit einer Gittersuche gewählt werden, bei der eine Evaluierung verschiedener Parameterwerte erfolgt [HCL03]. Diese kann mittels Kreuzvalidierung durchgeführt werden (siehe Abschnitt 3.1.2), jedoch nicht auf dem gesamten Datensatz sondern nur auf dem Trainingsdatensatz, da sonst auch die Testdaten zur Optimierung genutzt werden würden, was zu einem Bias in der Test-Performance führen würde. Wenn die Test-Performance mittels Kreuzvalidierung bestimmt werden soll, muss die Hyperparameterwahl somit für jeden fold in einer darin geschachtelten Kreuzvalidierung bestimmt werden.

Erkennungsprobleme mit mehr als zwei Klassen können abgebildet werden, indem mehrere SVMs trainiert werden. Die in dieser Dissertation genutzte Software LibSVM [CL11] nutzt den Ansatz, eine SVM für jede Klassenkombination zu trainieren, d. h. bei k Klassen $\frac{k(k-1)}{2}$ Modelle. Die Prädiktion wird durch Abstimmung aller Modelle entschieden. Es wird die Klasse zurückgegeben, die von den meisten Modellen prädiziert wurde.

Für Regression wird eine Variante der SVM eingesetzt, die Support Vector Regression (SVR) genannt wird [SS04]. Bei dieser wird ein Hyperparameter ε eingeführt, der angibt wie groß eine Abweichung der Modellprädiktion $f(\mathbf{x})$ vom Zielwert y maximal sein darf ohne das Modell zu beeinflussen. Es wird ein ähnliches quadratisches Optimierungsproblem gelöst wie bei der SVM, bei dem jedoch für jedes Trainingsbeispiel zwei Ungleichungen als Randbedingungen aufgestellt werden, vgl. [SS04]. Während bei der SVM ein Schwellwert angewendet wird, um die Klasse zu bestimmen, wird bei der SVR der Funktionswert direkt zurückgegeben.

Für SVM/SVR ist es wichtig, dass die Merkmale ähnliche Wertebereiche haben [HCL03]. Daher wird in dieser Arbeit jedes Merkmal standardisiert, bevor es in die SVM/SVR eingegeben wird, d. h. linear transformiert, so dass der Mittelwert über den Datensatz null und die Standardabweichung eins wird.

Random Forest (RF): Grundlage eines RF sind Entscheidungsbäume [Ste09]. Diese partitionieren den Merkmalsraum rekursiv, indem je Knoten ein Schwellwert auf eines der Merkmale angewendet wird[1]. Die Auswahl der Merkmale und Schwellwerte erfolgt beim Training typischerweise durch die Bewertung verschiedener Hypothesen anhand des Gini-Koeffizienten (Klassifikation) bzw. des mittleren quadratischen Fehlers (Regression), wobei die Hypothese ausgewählt wird, die innerhalb der beiden entstehenden Partitionen die „reinsten“ (ungleichsten) Klassenverteilungen bzw. niedrigsten Varianzen (Regression) erzeugt. Beim Testen wird jedes Sample anhand der Entscheidungsregeln Knoten für Knoten in eine finale Partition eingeordnet und der beim Training dafür hinterlegte Wert (Klasse bzw. Mittelwert bei Regression) zurückgegeben.

Ein RF ist ein Ensemble von Entscheidungsbäumen, bei dem für das Training eines jeden Baumes ein Vektor von Zufallszahlen gezogen und genutzt wird, um die Ähnlichkeit der konstruierten Bäume zu verringern [Bre01]. Einzelne Entscheidungsbäume neigen zur Überanpassung auf den Trainingsdatensatz, ein Problem das durch das Training und die Kombination einer großen Anzahl von möglichst unkorrelierten Entscheidungsbäumen vermieden bzw. reduziert werden kann.

[1] In dieser Arbeit werden nur reellwertige Merkmale verwendet. Für kategoriale Merkmale gibt es andere Regeln.

Die zentrale Idee des von Breiman vorgeschlagenen Random Forest betrifft die Suche nach der optimalen Partitionierung an den Entscheidungsknoten bei der Konstruktion der Bäume. Während hier bei klassischen Entscheidungsbäumen [Ste09] alle Merkmale in Betracht gezogen werden, was zu ähnlichen Bäumen führt, ziehen RF nur eine deutlich kleinere Teilmenge von F zufällig gewählten Merkmale in Betracht. F kann als Hyperparameter zur Optimierung der Performance variiert werden. Eine weitere Idee zur Reduzierung der Korrelation der Bäume ist es, die Trainingsdaten zu variieren. Breiman [Bre01] setzt hierzu *boostrapping* ein, bei dem für das Training eines jeden Baumes eine neue Stichprobe des originalen Trainingsdatensatzes *mit Zurücklegen* gezogen wird. Bei der Prädiktion werden die Einzelergebnisse der Bäume anhand eines Mehrheitsentscheides kombiniert (Klassifikation) bzw. wird der Mittelwert der Prädiktionen aller Bäume gebildet (Regression).

Convolutional Neural Network (CNN): Ein Convolutional Neural Network (CNN) [GBC16a] ist ein künstliches neuronales Netz, das Eingabedaten mit gitterartiger Struktur (hier Bilder) durch eine Abfolge von Faltungen (engl. convolution), Aktivierungsfunktionen (engl. activation function) und anderen Operationen verarbeitet. Die Ausgaben des Netzes können je nach Aufgabe sehr verschieden sein und reichen von Klassenzugehörigkeiten bzw. Regressionsergebnissen [Wer+19c; WSAH20], über fotorealistische oder künstlerische Bildern, die vom Netz generiert werden [Kar+20; GEB16] oder höherdimensionale Objekte (allgemein: Tensoren [Ji+19]), die beispielsweise zur Objektdetektion [BWL20; Sax+19] oder pixelweisen Segmentierung von Objekten [LSD15] genutzt werden. Die Verarbeitungsoperationen des Netzes werden Schichten (engl. layer) genannt und haben eine Reihe von Parametern, die ausgehend von einer Initialisierung beim Training des Netzes durch Minimierung einer Loss-Funktion angepasst werden. Die wichtigsten Parameter sind die Gewichte der künstlichen Neuronen, der Basiseinheiten des neuronalen Netzes. Diese berechnen jeweils eine gewichtete Summe von Eingabewerten, die aus der vorherigen Schicht (oder in der ersten Schicht aus dem Eingabebild) stammen. Meist folgt die Addition eines Skalars (Bias), das ebenfalls gelernt wird, und die Anwendung einer nicht-linearen Aktivierungsfunktion. Im Folgenden werden die wichtigsten Elemente von CNNs vorgestellt:

Convolution Layer: Die für CNNs namensgebende Operation ist die Faltung (engl. convolution), bei der für jeden Wert der Eingabedaten (z. B. jeden Pixel des Eingabebildes) eine gewichtete Summe der umgebenden Werte berechnet wird. Die Gewichte entstammen einem Filterkern (engl. filter kernel), der zumeist deutlich kleiner ist als die Eingabedaten (z. B. 3×3 oder 5×5 Pixel), und werden während des Trainings angepasst. In dieser Arbeit werden 2D-Faltungen angewendet, da die Eingabebilder eine 2-dimensionale räumliche Struktur (Höhe und Breite) haben, d. h. die gewichtete Summe wird entlang dieser zwei Dimensionen berechnet. Das Ergebnis jeder Faltung ist eine Merkmalskarte (engl. feature map), die angibt wo in den Eingabedaten gewisse Merkmale zu finden sind, die anhand des Filterkerns beschrieben werden.

Meist gibt es für jeden Punkt der räumlichen Struktur mehrere Werte, z. B. zur Codierung der Farbe im Eingabebild. Hierdurch ergibt sich ein 3-dimensionaler Tensor mit mehreren Kanälen (engl. channel), Höhe und Breite. Die Anzahl der Kanäle des Ausgabetensors m eines Convolution Layer, d. h. die Anzahl der resultierenden Merkmalskarten, lässt sich frei wählen und bestimmt die Anzahl der gelernten Filterkerne und Parameter. Bei einem Eingabetensor mit n Kanälen werden $n \times m$ Filterkerne gelernt, denn für jeden der m Ausgabekanäle werden die n Eingabekanäle mit jeweils einem eigenen Filterkern gefaltet und die Ergebnisse aufsummiert.

Convolution Layer haben einige Vorteile gegenüber den Schichten der klassischen neuronalen Netze, die aus vollständig verknüpften Neuronen bestehen (engl. fully connected layer / dense layer) [GBC16a]: (1) Sie nutzen die räumliche Struktur des Bildes aus und verknüpfen nur die lokale Information nahe beieinander liegender Pixel durch kleine Filterkerne. Das reduziert die Anzahl der Verknüpfungen, Parameter und Rechenoperationen. Komplexe, nicht lokale Interaktionen vieler Pixel lassen sich durch das Hintereinanderschalten mehrerer Convolution Layer modellieren. (2) Die Gewichte der Filterkerne werden für alle Eingabepixel gemeinsam genutzt (engl. parameter sharing), was wiederum die Anzahl der Parameter reduziert. (3) Die extrahierten Merkmale sind invariant bezüglich Translation, d. h. Verschiebung im Bild.

Aktivierungsfunktion: Auf die Ausgabe eines Convolution Layer (oder auch eines Fully Connected Layer) wird meist eine nicht-lineare Aktivierungsfunktion angewendet, um die Modellierung nicht-linearer Probleme zu ermöglichen. In der Ausgabeschicht wird im Kontext der Klassifikation meist die Sigmoid- oder die Softmax-Funktionen angewendet, deren Ausgaben Werte zwischen 0 und 1 annehmen und als Wahrscheinlichkeiten interpretiert werden können.

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

$$\text{softmax}(\mathbf{x})_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{3.2}$$

Die Sigmoid-Funktion (3.1) ist für binäre Klassifikation und binäre Multi-Label-Klassifikation (mit mehreren unabhängigen binären Labeln) geeignet. Die Softmax-Funktion eignet ich für Mehrklassenprobleme mit n Klassen und wird auf n Ausgabeneuronen angewendet, deren Ausgaben im Vektor $\mathbf{x}$ repräsentiert sind. (3.2) definiert sie für Klasse bzw. Neuron i. Die n Funktionswerte ergeben, wie bei einer Wahrscheinlichkeitsverteilung, in Summe eins. Für Regression wird in der Ausgabeschicht meist die lineare Aktivierung (3.3) angewendet, die beliebige Ausgabewerte erlaubt:

$$\text{linear}(x) = x \tag{3.3}$$

$$\text{relu}(x) = \max\{0, x\} \tag{3.4}$$

Die Aktivierungsfunktion ReLU (3.4) ist die am weitesten verbreitete Aktivierungsfunktion für die Zwischenschichten (engl. hidden layers), d. h. die Schichten vor der Ausgabeschicht. Sie ist eine nicht-lineare Funktion, ist jedoch abschnittsweise linear und verfügt über günstige Eigenschaften für das Training und die Generalisierung [GBC16b, S. 174].

Pooling: Bei einem CNN folgt auf die Aktivierung oft ein Pooling Layer, der jeden Wert der Merkmalskarte durch eine statistische Zusammenfassung der Nachbarschaft ersetzt, z. B. das Maximum oder den Mittelwert der 3×3 umgebenden Werte. Hierdurch wird die Invarianz des Netzes bezüglich lokaler Translation (kleiner Verschiebung der Eingabe) begünstigt. Pooling wird meist mit räumlicher Unterabtastung verbunden, um die Auflösung der Merkmalskarten für nachfolgende Schichten zu reduzieren. Vor der Ausgabeschicht wird oft ein globales Pooling eingesetzt, das n Merkmalskarten beliebiger räumlicher Ausdehnung in einem n-dimensionalen Merkmalsvektor zusammenfasst.

Fully Connected Layer: In CNNs wird bei der Klassifikation bzw. Regression als Ausgangsschicht oft eine klassische, voll verknüpfte Schicht von Neuronen eingesetzt, um die CNN-Merkmale auf die Ausgabewerte abzubilden, von denen je nach Erkennungsaufgabe unterschiedlich viele gebraucht werden. Ein Fully Connected Layer kann auch durch einen Convolution Layer mit Filterkernen der Größe 1×1 realisiert werden.

Eine CNN-Architektur ist eine genau festgelegte Abfolge dieser und gegebenenfalls weiterer Layer mit den zugehörigen Hyperparametern, wie z. B. Filterkerngrößen und Anzahl der Ausgabekanäle. Genauer gesagt muss es keine lineare Abfolge sein, sondern ist es ein Graph, durch den die Ein- und Ausgaben von Verarbeitungseinheiten miteinander verknüpft werden. Beispiele für CNN-Architekturen sind AlexNet [KSH12], VGG-19 [SZ15], NASNet [Zop+18] oder MobileNetV3 [How+19].

Das Training eines CNN erfolgt durch Optimierung einer Loss-Funktion, bei Klassifikation meist Cross-Entropy-Loss und bei Regression meist Mean-Squared-Error-Loss, wozu typischerweise das Gradientenabstiegsverfahren oder einer Variante davon eingesetzt wird. In einem Trainingsschritt werden dem Trainingsdatensatz jeweils zufällig n Beispiele entnommen, die jeweils einen sogenannten Batch (engl. für Stapel) bilden. Diese werden durch das Netz propagiert (engl. forward pass). Aus ausgehend von der Abweichung der Prädiktion von den Labeln in der letzten Schicht werden Gradienten zur Verringerung der Abweichung bestimmt und schichtweise bis zur Eingabeschicht zurück propagiert (engl. backward pass / backpropagation). Die Initialisierung der Gewichte zu Beginn des Trainings erfolgt meist zufällig nach Glorot und Bengio [GB10]. Eine Alternative ist das *Transferlernen*, bei dem das Training von Gewichten ausgeht, die zuvor beim Training der selben CNN-Architektur mit anderen Daten (und meist auch für eine andere Aufgabe) ermittelt wurden.

Transferlernen: Beim Transferlernen [WKW16] werden zwei Domänen betrachtet, zu denen jeweils verschiedene Daten und/oder Erkennungsaufgaben gehören. Die Zieldomäne umfasst die Daten und die Erkennungsaufgabe, die von Interesse ist. Die Quelldomäne zeichnet sich dadurch aus, dass mehr Daten zur Verfügung stehen bzw. die Daten leichter zu erheben sind. Ziel des Transferlernens ist es, die Erkennungsleistung des Modells, das in der Zieldomäne trainiert wird, zu verbessern, indem Informationen aus der Quelldomäne ausgenutzt werden [WKW16].

Im Zusammenhang mit CNNs ist eine typische Vorgehensweise, zuerst ein Modell auf dem ImageNet-Datensatz [Den+09] zur Objektklassifikation zu trainieren (Quelldomäne). Da der Datensatz mehrere Millionen Fotos verschiedenster Objekte umfasst, ist das resultierende Modell als Ausgangspunkt für viele andere Computer-Vision-Probleme geeignet. Beim Lernen der Objektklassifikation wird das CNN zu einem hierarchischem Merkmalsextraktor, der in den ersten Convolution Layern zunächst Primitive wie Kanten detektiert, aus denen in den folgenden Schichten Detektoren für immer komplexere geometrische Strukturen und Texturen zusammengesetzt werden, bis hin zu Detektoren für Objekte wie Köpfe oder Bäume. Insofern kann das in der Quelldomäne trainierte CNN genutzt werden, um in der Zieldomäne Merkmale zu extrahieren und diese anschließend mit klassischen Verfahren wie einer SVM zu klassifizieren. Zur Merkmalsextraktion werden hierbei die Bilder der Zieldomäne in das CNN eingegeben und die jeweils resultierenden Aktivierungen einer Zwischenschicht abgespeichert.

Eine andere Variante ist, das in der Quelldomäne vortrainierte CNN-Modell in der Zieldomäne weiter zu trainieren, um die Merkmalsextraktion für die andere Erkennungsaufgabe und/oder die anderen Daten zu optimieren. Hierfür werden die Architektur und die Gewichte aus der Quelldomäne so weit es geht übernommen. Wenn die andere Erkennungsaufgabe der Zieldomäne auch eine andere Ausgabeschicht erfordert (z. B. eine andere Anzahl von Neuronen), wird die letzte Schicht entsprechend ersetzt und zufällig initialisiert. Durch die Initialisierung der unteren Schichten mit Gewichten der Quelldomäne konvergiert das Training zum einen schneller und zum anderen oft auch zu einem besseren Modell als bei vollständig zufälliger Initialisierung. Insbesondere wenn CNNs mit einer großen Kapazität mit einem kleinen Datensatz trainiert werden, kann durch das Transferlernen eine Überanpassung vermieden (oder reduziert) und die Generalisierung verbessert werden.

Regularisierung bei CNN: Um Überanpassung von CNN an die Trainingsdaten entgegenzuwirken und die Test-Performance zu verbessern, werden verschiedenste Regularisierungsmethoden angewendet [KGC17]. Eine Methode ist *Datenaugmentierung*, bei der die vorhandenen Trainingsdaten transformiert werden um einen variantenreicheren und größeren (potentiell unendlich großen) Trainingsdatensatz zu erhalten. Die Parameter der Transformationen werden dabei im Allgemeinen zufällig gewählt. Bei Bildern sind typische Transformationen das Ausschneiden von Teilbildern (engl. cropping), horizontales Spiegeln (engl. flipping), Skalierung, Rotation oder Helligkeits-, Kontrast- oder Farbanpassungen. Weitere sehr verbreite Methoden mit regularisierender Wirkung sind *Dropout*, bei dem einzelne Neuronen einer Schicht zufällig (mit einer als Hyperparameter definierten Wahrscheinlichkeit) entfernt werden, oder *Batch Normalization*, bei welcher die Batches anhand ihres Mittelwert und ihrer Standardabweichung normiert werden. Alle Verfahren fügen den Trainingsprozess Zufallseinflüsse hinzu (nicht jedoch dem Testen) und zwingen das Netz robuster gegenüber Variationen in den Daten zu werden. Eine weitere Regularisierungsmethode ist *Weight Decay*, bei der die Loss-Funktion um einen Regularisierungsterm ergänzt wird, der die L2-Norm der Gewichte berechnet. Somit werden große Werte in den Gewichten bestraft und „einfache“ Lösungen gegenüber unnötige komplexen, überangepassten Lösungen bevorzugt. Ebenfalls als Regularisierung auffassen kann man das frühzeitige Beenden des Trainings, d. h. dass das Netz nicht so lange trainiert wird bis der Trainingsfehler minimal wird, sondern nach einer festgelegten Anzahl von Trainingsschritten oder in Abhängigkeit vom Fehler auf einem Validierungsdatensatz abgebrochen wird. Eine weitere Methode zur Regularisierung ist das gleichzeitige Lernen mehrerer Aufgaben (engl. *Multi-Task Learning*) [GBC16b, S. 244ff.], bei dem Beispiele mit unterschiedlicher Annotation (oft aus verschiedenen Datensätzen) zusammen genutzt werden, um ein Modell zu trainieren. Das Modell hat gemeinsam genutzte Parameter (bei einem CNN in den vorderen Schichten) und aufgabenspezifische Parameter (in den hinteren Schichten). Durch das gemeinsame Nutzen der Datensätze bzw. Annotationen stehen für das Training der gemeinsam genutzten Schichten mehr Beispiele bzw. Informationen zur Verfügung, was oft zu einer verbesserten Generalisierung führt. Wenn die Erkennungsaufgaben jedoch zu verschieden sind und/oder die Kapazität des Netzes nicht ausreicht, kann Multi-Task Learning auch zur Verschlechterung der Performance führen.

3.1.2. Evaluierung von Erkennungssystemen

Entscheidend für die Nützlichkeit eines Erkennungssystems ist, wie gut es generalisiert, d. h. wie gut es auf neuen Daten funktioniert. Um das zu testen, werden die Daten vor dem Training aufgeteilt und es wird nur ein Teil der Daten für das Training verwendet. Die Bewertung des trainierten Modells erfolgt anschließend, indem für den Rest der Daten (oder einem Teil davon) die Abweichung der Prädiktion des Modells von der Grundwahrheit bestimmt wird. Die Abweichung bzw. Übereinstimmung wird mithilfe eines Performance-Maßes beziffert.

Meist soll ein Erkennungssystem für Personen eingesetzt werden können, für die keine Trainingsdaten zur Verfügung stehen. Um die Eignung eines Systems für diese Anforderung zu bewerten, muss sichergestellt werden, dass *keine* der Personen *sowohl* im Trainings- *als auch* im Testdatensatz vorkommt. Da sich verschiedene Personen stark unterscheiden (vgl. Herausforderungen der Verarbeitungskette) sind die Daten einer Person meist korreliert und in Relation zur Gesamtheit der Daten ähnlich zueinander. Somit sind die Testdaten einer Person, die auch im Trainingsdatensatz vorkommt, nicht hinreichend neu und es entsteht infolge dessen ein Bias in der Performance. Weitere Ausführungen hierzu, inkl. Beispiel, sind in [Wer+19b, S. 17] zu finden.

Kreuzvalidierung: Wie sollte die Aufteilung in Trainings- und Testdaten erfolgen? Um eine möglichst genaue und wenig vom Zufall abhängende Schätzung der Generalisierungsfähigkeit zu erhalten, ist ein möglichst großer Testdatensatz von Vorteil. Gleichzeitig wirkt sich ein größerer Trainingsdatensatz meist positiv auf die Güte des Modells aus. Eine gute Handhabung dieses Zielkonflikt bietet, insbesondere bei kleinen Datensätzen, die Evaluierung mit k-facher Kreuzvalidierung (engl. k-fold cross validation). Dabei werden die Daten in k etwa gleichgroße Teile aufgeteilt und das Trainieren und Testen des Modells k Mal wiederholt, wobei jeder Teil einmal als Testdaten und alle übrigen Teile jeweils als Trainingsdaten verwendet werden. So kann ein großer Teil der Daten für das Training genutzt werden und dennoch – nach und nach – mit dem gesamten Datensatz getestet werden. In dieser Arbeit wird die Spezialform LSO-Kreuzvalidierung (LSO = leave-subjects-out, engl. für Personen auslassen) eingesetzt, um die Trennung der Personen zwischen Trainings- und Testdaten sicherzustellen. Hierfür wird die Aufteilung der Daten in k Teile anhand der Personenkennung durchgeführt. Für n Personen im Datensatz liefert eine n-fache LSO-Kreuzvalidierung die beste Schätzung des Erwartungswertes der Generalisierungsfähigkeit für ungesehene Personen, da das Bias mit steigendem k sinkt [Koh95]. Bei großen Datensätzen und aufwändigen Lernverfahren ist dies jedoch aufgrund der Laufzeit unpraktikabel. Um vergleichbare Ergebnisse zu erhalten, werden in der Arbeit Kreuzvalidierungen durchgängig mit $k = 5$ durchgeführt.

Performance-Maße: In dieser Arbeit wird ICC(3,1) [SF79] für alle Evaluationen als primäres Performance-Maß angewendet. Es basiert auf einer Varianzanalyse und ist gut für Intervallskalen geeignet, beispielsweise die Intensität von Schmerzen oder einer Action Unit. Aufgrund der günstigen Eigenschaften von ICC(3,1) und um die Interpretation der Ergebnisse zu vereinfachen wird das Maß auch für binäre Klassifikationsprobleme zwischen kein Schmerz und Schmerz angewendet. Zwar handelt es sich hierbei ursprünglich um kategoriale Merkmale, jedoch sind beide Zustände auf einer gemeinsamen Schmerzskala verortet. Die Kategorien sind durch Diskretisierung eines ursprünglich kontinuierlichen Phänomens entstanden und werden zur Berechnung des ICC als Werte 0 und 1 codiert. Eine uninformierte Prädiktion, sowohl eine rein zufällige wie auch eine triviale (immer Mehrheitsklasse), wird mit dem ICC-Wert 0 bewertet, eine fehlerfreie Prädiktion mit dem Wert 1. In Anhang A werden verschiedene Performance-Maße verglichen und die Wahl für diese Arbeit begründet.

Statistische Signifikanztests: Die Ergebnisse von Evaluationen variieren durch Zufallseinflüsse und können daher über oder dem echten Wert (Erwartungswert) liegen, was den Vergleich von Evaluationsergebnissen erschwert. Statistische Tests ermöglichen trotz dieser Problematik die Deutung der beobachteten Evaluationsergebnisse, indem sie nicht nur die Mittelwerte (oder die Korrelation) sondern auch die Anzahl und Verteilung der einzelnen Beobachtungen beachten. Hierfür wird zunächst auf Basis von Vorüberlegungen oder Unterschieden in den beobachteten Messwerden eine Hypothese aufgestellt, d. h. eine Vermutung wie z. B.: Die Performance von Modell A ist besser als die von Modell B. Vorläufig wird vom Gegenteil ausgegangen, d. h. dass die Nullhypothese wahr ist (im Beispiel Modell A ist gleich gut). Der statistische Test beantwortet nun die Frage, wie gut die gemachten Beobachtungen zu der Nullhypothese passen. Hierfür berechnet er einen p-Wert, die Wahrscheinlichkeit, dass die gemachten (oder noch extremere) Beobachtungen vorkommen wenn die Nullhypothese wahr ist. Liegt der p-Wert unter einem Signifikanzniveau (in dieser Arbeit $\alpha = 0{,}05$), wird die Nullhypothese verworfen und die Vermutung bestätigt. Im Beispiel wäre die Schlussfolgerung, dass Modell A statistisch signifikant besser ist als Modell B und der beobachtete Unterschied (mit hoher Wahrscheinlichkeit) nicht durch Zufall entstanden ist.

In der Regel werden in dieser Arbeit Permutationstests durchgeführt, die weniger Annahmen zugrunde legen und genauer sind als klassische Tests [Moo+03]. Dabei werden paarweise zusammengehörigen Beobachtungen betrachtet, z. B. ICCs für jeden Probanden, jeweils zum einen mit Modell A und zum anderen mit Modell B. Um zu überprüfen, ob Modell A besser ist als Modell B, werden die Zuordnungen der ICCs zu den Modellen 10.000 Mal permutiert. Für jede Permutation wird die Differenz des Mittelwertes der jetzt zu Modell A und der jetzt zu Modell B gehörigen Ergebnisse berechnet. Über die Permutationen ergibt sich die Verteilung der Performance-Differenzen unter der Annahme der Nullhypothese, dass beide Modelle gleich gut sind. Nun wird die Differenz der Mittelwerte von Modell A und B mit den tatsächlichen Beobachtungen berechnet. Der p-Wert ist die Wahrscheinlichkeit, dass die Differenzen (in der zuvor bestimmten Verteilung) größer oder gleich zu der tatsächlich beobachteten Differenz sind. Je geringer der Wert von p, desto mehr widersprechen die Beobachtungen der Nullhypothese und desto stärker bestätigen sie die Hypothese, dass Modell A besser ist als B.

3.1.3. Erkennung von Facial Action Units und Schmerzmimik

Hier wird ein Überblick zum Stand der Technik in der automatisierten Erkennung von Facial Action Units und von Schmerzmimik gegeben. Detailliertere Darstellungen zu den verwandten Arbeiten zu Mimik und Action Units (AUs) findet der interessierte Leser in [ZLZ20; LD20; Cor+16; SGC15]. Mit den verwandten Arbeiten zur Schmerzerkennung beschäftigt sich der Überblicksartikel von Werner et al. [Wer+19b] ausführlicher, auf dem Teile der Ausführungen dieses Abschnitts basieren.

Die Mimikerkennung folgt zumeist der in Abschnitt 1.2.1 vorgestellten Verarbeitungskette. Für den ersten Schritt, die Datenaufnahme, werden in der Regel Standardfarbkameras eingesetzt, deren Bilder auch im Fokus dieser Dissertation stehen. Einige andere Arbeiten nutzen z. B. Tiefenkameras oder Infrarotkameras, vgl. [Wer+19b; Cor+16]. Betrachtet man das Ziel der Schmerzerkennung in einem größeren Kontext, der über die Mimikerkennung hinausgeht, sind auch viele andere Sensoren relevant. Arbeiten, die diese Sensoren nutzen und zum Teil auch in multimodalen Erkennungssystemen mit der Mimikerkennung kombinieren, werden von Werner et al. [Wer+19b] ausführlich vorgestellt. Der folgende Abschnitt beschränkt sich auf Erkennung von Schmerzmimik und AUs anhand von Farbkameradaten.

Die Gesichtsdetektion und Landmarkenlokalisierung stehen ebenfalls nicht im Fokus dieser Arbeit, so dass der Leser für einen Überblick hierzu auf [ZZZ15] bzw. [WJ19] verwiesen wird. Die Gesichtsnormierung wird in Abschnitt 3.1.4 thematisiert. Im Folgenden werden die verwandten Arbeiten hinsichtlich der Merkmalsextraktion und der Klassifikation bzw. Regression, sowie der Evaluierung der Ansätze besprochen.

Merkmalsextraktion: Merkmale können unterschieden werden in menschengemachte Merkmale, die von Ingenieuren anhand von Vorwissen und Intuition entworfen wurden, und gelernte Merkmale, deren Extraktion anhand von Daten und einer Loss-Funktion optimiert wurde. Frühere Arbeiten haben sich oft mit dem manuellen Entwurf oder der Evaluierung von existierenden Merkmalsarten beschäftigt. Die Tendenz geht hin zu gelernten Merkmalen, da (1) zu vielen Fragestellungen immer mehr Daten zum Lernen zur Verfügung stehen, (2) die Methoden zum Lernen von Merkmalen, z. B. CNNs und Transferlernen, immer weiter verbessert werden, und (3) die nötige Rechenleistung immer leichter zugänglich wird. Verwendet wurden unter anderem:

1. Formmerkmale, zumeist *2D-Landmarkenkoordinaten* oder darauf basierende *generische geometrische Merkmale* [Ash+07; Ash+09; EVM17; Gha+14; KRP12; Luc+11a; Luc+12; MBB14;

RG15; RPP15; Rui+16; RV16; Wer+17; ZK14; ZCZ16; Aun+16; Yan+14; Min+15; Bin+14; WHAH17; Val+17; BRM16; Zha+19; Zha+18; BRM16],

2. Texturmerkmale, wie einfache *Pixelrepräsentationen* [Ash+07; Ash+09; Gha+14; Luc+11a; Luc+12; Wer+17; Zho+16; Bra+07; Adi+15; Che+12b], *Local Binary Pattern (LBP) Merkmale* [Che+13; KRP12; KTP16; Kha+13; RG16; RPP13b; San+14; Yan+16; Zen+14; ZCZ16; Aun+16; Mav+13; SZP13; WHAH17; BRM16], *Histogram of Oriented Gradients (HOG)* [CCF17; EVM17; Kha+13; RG16; Thi+16; TKS17; Che+12b; BRM16], *Gabor-Merkmale* [LL17; RG16; Roy+15; ZCZ16; Bar+14; LBL09; Sik+15; GCD15; MM14; Mav+13; SST12; Che+12b], *andere Filter* [HC12; INM15; Ira+15], *Scale Invariant Feature Transform (SIFT)* [NM15; SDB14; RVB14], *Discrete Cosine Transform (DCT)* [KRP12; Luc+08; Aun+16], und *andere Texturmerkmale* [FFV14; Flo+16; LL17; Yan+16; HZP16; Zam+17b],

3. manuell entworfene Merkmale, wie *Abstände im Gesicht in 2D* [RP+13; Käc+15b; Käc+17; Kes+17; Thi+16; TKS17; PIY06; Che+12a; MR09; Tsa+16] und *3D* [Nie+09], die oft *mit Maßen von Mimikfalten, -wülsten, und/oder -furchen kombiniert* werden um zusätzliche Änderungen im Erscheinugsbild zu erfassen [HK12; Wer+14b; Wer+13; WAHN12; WAHN14; Käc+15a; Wer+17; WAHW17],

4. *mit neuronalen Netzen gelernte Merkmale* [EVM17; KPM16; Ped15; Rod+17; Zho+16; Wan+17; Haq+18; ZPS17; Bat+17; Hua+20], und

5. *mit anderen Methoden gelernte Merkmale* [FFV14; Flo+16; RG15; RG16; Ami+17; Zha+19].

Transferlernen und gelernte Merkmale: Das Lernen der Merkmalsextraktion erfolgt oft unabhängig von der eigentlichen Lernaufgabe, mittels Transferlernen auf einem größeren Datensatz. Egede et al. [EVM17] trainieren CNNs zum Erkennen von AUs auf der Datenbank BP4D und wenden sie anschließend auf UNBC an, um Merkmale für die Schmerzerkennung zu extrahieren. Florea et al. [FFV14; Flo+16] lernen eine Merkmalstransformation auf der Emotionserkennungsdatenbank CK+ und transferieren die Datenrepäsenation auf den UNBC-Datensatz. Sie argumentieren, dass dadurch die Robustheit verbessert wird, da CK+ mehr Personen umfasst als UNBC. Zur Merkmalsextraktion wird von Kharghanian et al. [KPM16] ein Convolutional Deep Belief Network (CDBN) vorgeschlagen, das unüberwacht (ohne Nutzung von Label) trainiert wird. Die erste Schicht lernen sie auf Naturbildern, die zweite auf UNBC. Rodriguez et al. [Rod+17], Wang et al. [Wan+17], Haque et al. [Haq+18] und Huang et al. [Hua+20] nutzen und feinjustieren CNNs, die für die Identifikation von Gesichtern trainiert wurden, für die Erkennung von Schmerzmimik. Huang et al. [Hua+20] vergleichen zusätzlich mit einem Netz, das mit dem FER-2013 Datensatz zur Erkennung von Emotionskategorien vortrainiert wurde und finden, dass dieses eine bessere Performance auf UNBC ermöglicht als das für die Identifikation vortrainierte CNN. Andere Autoren nutzen kein Transferlernen, sondern trainieren neuronale Netze ausgehend von einer zufälligen Initialisierung: Pederson et al. [Ped15] schlägt einen teilüberwachten Autoencoder vor, um Merkmale für die Schmerzerkennung zu extrahieren, und vergleicht ihn mit einem klassischen unüberwachten Autoencoder. Verwandt mit unüberwachten Autoencodern ist *sparse coding* [Ami+17; Zha+19], bei dem durch Optimierung eine Menge von Basisvektoren gelernt wird, mit deren Hilfe die Samples durch dünn besetzte Merkmalsvektoren repräsentiert werden können. Ein rekurrentes CNN für die Schätzung der Schmerzintensität wird von Zhou et al. [Zho+16] vorgeschlagen. Im Gegensatz zu den anderen Autoren, extrahieren Zhou et al. [Zho+16] und Wang et al. [Wan+17] nicht explizit Merkmale, die anschließend in einem zumeist separaten Erkennungsmodell verarbeitet werden, sondern trainieren ein tiefes neuronales Netz, das die Merkmalsextraktion und das Erkennungsmodell in einer kaum separierbaren und gemeinsam optimierten Einheit integriert. Einige Arbeiten prädizieren AUs mit einer ersten Menge von

Modellen und nutzen die AUs als Merkmale für ein nachgeschaltetes Modell zur Schmerzerkennung [Bar+14; LBL09; Sik+15; Gha+14; Liu+18]. In diesem Kontext nutzen Bartlett, Littlewort, und Sikka [Bar+14; LBL09; Sik+15] Transferlernen, da die AU-Modelle auf anderen Datensätzen trainiert wurden. Viele Autoren [Bat+17; ZPS17; Wer+19c; FL20] adaptieren zur Erkennung und Intensitätsschätzung von AUs CNNs, die mit dem Datensatz ImageNet vortainiert wurden.

Modelle und Lernverfahren zur Klassifikation und Regression: In der Schmerz- und Mimikerkennung am meisten verwendet werden Support Vector Machines (SVMs) mit linearem Kernel [Ash+07; Ash+09; CCF17; HC12; KPM16; Luc+08; Luc+11a; Luc+12; NM15; Ped15; RG16; Roy+15; RV16; Wer+17; Yan+16; Zen+14; Käc+15a; Aun+16; Zam+17b; Adi+15; MR09; Tsa+16; GCD15; Min+15; Mav+13; BRM16] und RBF-Kernel [RG15; Wer+13; Bar+14; LBL09; Nie+09; WAHN12; WAHN14]. Für reellwertige Ausgaben werden Support Vector Regression (SVR) [FFV14; Flo+16; HZP16; San+14; Zen+14; ZCZ16; SST12; GCD15; Jen+13; SZP13; BRM16; Ami+17], Relevance Vector Regression [EVM17; KRP12; KTP16] und logistische Regression [Bin+14] eingesetzt. Ebenfalls weit verbreitet sind Random Forests (RF) [Käc+15a; Käc+15b; Käc+17; Wal+15; Wer+14b; WAHW17; Wer+17; WHAH17; Kes+17; Thi+16; TS17; TKS17; Aun+16; Zam+17b; Haq+18], Nächste-Nachbarn-Klassifikatoren [Kha+13; MBB14; ZK14; Zam+17b; Adi+15], Variationen von Conditional Random Fields und anderen probabilistischen grafischen Modellen [Gha+14; LMRP17a; RPP13b; RPP15; Rui+16; SZP13; Yan+14; MM14; Val+17], und verschiedene künstliche neuronale Netze. Zu den genutzten neuronalen Netzen gehören: Convolutional Neural Networks (CNN) [Rod+17; Wan+17; Zho+16; Haq+18; FL20; Hua+20; Bat+17; ZPS17], Long Short-Term Memory (LSTM) networks [LMRP17a; Rod+17; Haq+18], Radial Basis Function (RBF) networks [Käc+17], und klassische multi-layer perceptrons [Bra+07; Käc+17]. Ebenfalls genutzt werden Gaussian Mixture Models (GMM) [Liu+18] oder lineare Modelle mit selbst entwickelter Optimierung [Zha+18; Zha+19].

Multi-Task Learning wurde im Kontext der Schmerzerkennung bisher nur für das Training personenspezifischer Modelle genutzt [LMRP17b; RP+13]. In der AU-Erkennung wurden Multi-Task-CNNs trainiert, die zusätzlich die Kopfpose bzw. Kameraansicht [ZPS17] oder Emotionskategorien und Valenz-Arousal-Werte [KZ18; KZ19] prädizieren. Der Einfluss des Multi-Task Learning auf die Performance wurde jedoch nicht von allen Autoren untersucht bzw. er ist nicht in allen Fällen positiv. Einen positiven Einfluss des Multi-Task Learning haben Werner et al. [WSAH20] gefunden, als sie damit Daten mit manuellen AU-Annotationen und Daten mit automatisiert annotierten AUs kombiniert haben.

Evaluierung: Die Vergleichbarkeit von quantitativen Ergebnissen, die in verschiedenen Artikeln veröffentlicht wurde, kann trotz der Verwendung des gleichen Datensatzes begrenzt sein. Der Grund ist, dass sich die Experimente sehr oft in einem oder mehreren der folgenden Aspekte unterscheiden: (1) in den betrachteten Erkennungsaufgaben, (2) in der Evaluationsmethode, (3) bei den Performance-Maßen, (4) hinsichtlich des Grades der Automatisierung bzw. händischen Eingreifens sowie (5) bezüglich der genutzten Teilmenge der Daten [Wer+19b]. So kann die Performance z. B. allein durch die Änderung der Evaluationsmethode von 40-67% auf 91-96% steigen oder durch die Auswahl unterschiedlicher Teilmenge von Probanden zwischen 49% und 93% schwanken [Wer+19b]. Die Ergebnisse anderer Autoren zu reproduzieren ist ebenfalls schwer, da in vielen Artikeln wichtige Details ausgelassen werden (und einige Autoren nicht auf Anfragen reagieren). Die Problematik betrifft insbesondere die zahlreichen Arbeiten mit dem Datensatz UNBC [Wer+19b]. Gute Vergleichbarkeit ist bei FERA 2017 gegeben, da für die Challenge mit dem Datensatz das Evaluierungsprotokoll genau vorgegeben wurde.

3.1.4. Normierung des Gesichts

Die Normierung des Gesichts wurde im Artikel von Werner et al. [Wer+19c] adressiert, auf dem weite Teile dieses Abschnitts basieren. Die Gesichtsnormierung, die oft auch Gesichtsregistrierung (engl. face registration) genannt wird, hat sich in zahlreichen Arbeiten zur Mimikerkennung [RPP13a; Che+12b; Val15; SGC15] und auch zur Identifikation [Has+15; Zhu+15; Yim+15; Cao+20] als hilfreicher Schritt bewährt. In ihrer einfachsten Form kompensiert die Gesichtsnormierung Unterschiede der Position und Größe des Gesichts im Bild sowie der Rotation in der Bildebene. Komplexere Verfahren zielen darauf ab, zusätzlich die Variation durch weitere Faktoren zu entfernen, wie Rotationen aus der Bildebene (wenn der Kopf von der Kamera weggedreht wird – in diesem Zusammenhang wird oft von Frontalisierung des Gesichts gesprochen), verschiedene Gesichtsproportionen [Val15], Beleuchtung [Zhu+13; Yim+15], Verdeckungen [Sag+15], Mimik [Zhu+15] oder Hintergrund. Die grundlegende Idee ist es, Invarianz gegenüber Störfaktoren zu erlangen, indem ihr Einfluss auf die extrahierten Merkmale reduziert wird, und somit die Erkennungsaufgabe zu vereinfachen. In der Mimikerkennung sind sowohl Kopfpose als auch persönliche Unterschiede in der Gesichtsform und -textur eine Herausforderung [CTC13; SGC15]. Das Normieren bezüglich dieser Faktoren ist vorteilhaft, wenn dabei die Mimikinformation erhalten bleibt, weil dadurch Unterschiede innerhalb der Klassen reduziert werden.

Eine Vielzahl von Methoden wird für die Normierung genutzt. Die einfachste Form ist es, die bei der Gesichtsdetektion erhaltene Bounding Box auszuschneiden und auf eine einheitliche Größe zu skalieren [Jun+15; Bat+17]. Dies wird später als FaceDet (von engl. face detection) abgekürzt. Wenn Landmarken bekannt sind, ist eine weitere einfache Option das Bild lediglich zu skalieren [BQSM16], was z. B. ausreicht um lokale Merkmale in der Umgebung der Landmarken zu extrahieren. Eine weitere Möglichkeit ist es, das Gesicht anhand von Landmarken auszuschneiden und zu skalieren [Hua+20]. Typische weitergehende landmarkenbasierte Normierungsmethoden sind in Tabelle 3.1 zusammengefasst. Sie basieren auf verschiedenen Landmarken, wie nur den Augen, inneren Landmarken (ohne Kontur des Gesichts) oder Landmarken inklusive der Gesichtskontur (vgl. Spalte „Eingabe“). Nicht jede Software zur Landmarkenlokalisierung stellt auch Landmarken der Gesichtskontur bereit und diese werden oft weniger genau lokalisiert als die inneren Landmarken. Die meisten Methoden registrieren die Landmarken mit einem statischen Referenzgesicht (meist ein „mittleres“ Gesicht), vgl. Spalte „Ziel“. Sie unterscheiden sich jedoch hinsichtlich der genutzten Transformation: nicht-reflektierende Ähnlichkeitsabbildungen (Translation, Skalierung, Rotation) und affine Abbildungen (Translation, Skalierung, Rotation, Spiegelung, Scherung) werden am häufigsten eingesetzt.

Die ersten fünf Methoden in Tabelle 3.1 erzeugen ein normiertes Gesicht indem sie das Bild mit einer einzelnen Transformation verzerren (engl. warping), welche die Bilder zu einem gewissen Grad registriert, d. h. bezüglich gewisser Landmarken in gute räumliche Übereinstimmung bringt. Im Gegensatz dazu versuchen die anderen Methoden die Kopfposevarianz in den Bildern zu reduzieren, also zu einem beliebig gedrehtem Gesicht eine frontale Ansicht zu generieren, indem sie abschnittsweise affine Transformationen oder 3D-Rendering einsetzen. Die klassische Methode zur Frontalisierung ist abschnittsweise (engl. piecewise) affines Warping (abgekürzt PieceAff) hin zu der Punktverteilung eines Referenzgesichtes. Es ermöglicht eine genaue Registrierung für viele Kopfposen, hat aber folgende Nachteile: (1) Informationen zur Form der Gesichter werden entfernt, d. h. nicht nur Unterschiede in den Gesichtsproportionen sondern auch Verformungen aufgrund von Mimik, (2) das Warping kann auch relevante Texturinformationen verfälschen, verschwinden lassen oder große Bereiche anhand weniger Pixel füllen, und (3) die Methode hat keine Behandlung für Verdeckungen, d. h. zum Umgang mit bzw. Ersetzen von fehlenden Bildinformationen, was bei Frontalisierung von extremen Kopfposen zu Artefakten führt (Beispielbilder in Abb. 3.8 auf S. 82). Hassner et al. [Has+15] (abgekürzt 3dStatic) nutzen ein statisches

	Registrierung		Textur-Warping		
Abkürzung	**Eingabe**	**Ziel**	**Transformation**	**VB**	**Anwendung**
SimEye	Augenmittelpunkte	Referenzgesicht	NR Ähnlichkeitsabbildung	×	[VJM11; Liu+14; Sen+12; Zho+12; DBD15]
SimInner	innere LM	Referenzgesicht	NR Ähnlichkeitsabbildung.	×	[Zha+15; Din+13; ZCZ16; Mas+16]
SimStable [BMR15]	mimikunabhängige LM	Referenzgesicht	NR Ähnlichkeitsabbildung	×	[BMR15; BRM16; Rin+17]
AffInner	innere LM	Referenzgesicht	affine Abbildung	×	[WSAH15; RPP15; Ele+16; Ami+17; SWAH17]
AffStable	stable innere LM (Augen/Nase)	Referenzgesicht	affine Abbildung	×	[Val+15; KPP10; AMV15]
PieceAff	LM mit Gesichtskontur	Referenzgesicht	abschnittsweise affine Abb.	×	[Che+12b; Wan+13; KTP15; Wer+17; Rod+17]
3dStatic [Has+15]	innere LM	statisches 3D-Model	3D-Rendering	✓	[Has+15]
FaNC Abschn. 3.2.2	prädizierte Korrespondenzpunkte	prädizierte Korrespondenzpunkte	abschnittsweise affine Abb. + Alpha-Blending	✓	Abschnitt 3.3.2, [Wer+19c]

VB: Verdeckungen/Aufdeckungen werden behandelt LM: Landmarken des Gesichts
NR: Nicht-reflektierend (ohne Spiegelung)

Tabelle 3.1.: Landmarkenbasierte Methoden zur Gesichtsnormierung: Verfahren des Standes der Technik (oben) und das in Abschnitt 3.3.2 vorgeschlagene Verfahren „FaNC“. Die fett hervorgehobenen Verfahren „PieceAff“, „3dStatic“ und „FaNC“ reduzieren die Kopfposevarianz („Frontalisierung“).

3D-Modell mit korrespondierenden 3D-Landmarken. Sie nehmen an, dass die intrinsischen Kameraparameter bekannt sind und schätzen die extrinsischen Kameraparameter zur Bestimmung der Kopfpose. Anschließend texturieren sie das 3D-Modell mit dem Eingabebild und rendern es in frontaler Pose. Verdeckungen werden durch Alpha-Blending mit der gespiegelten Version des Modells behandelt.

Neben den landmarkenbasierten Methoden gibt es auch rein texturbasierte Ansätze zur Normierung von Gesichtern [Zhu+13; Yim+15; Yin+17; Cao+20], die jedoch hier nicht im Fokus stehen. Wenn es überhaupt möglich ist, sie mit hohen Bildwiederholraten anzuwenden, wird hierfür teure Hardware benötigt. Um gut zu generalisieren, werden außerdem sehr große Trainingsdatensätze mit hinreichenden Variantenreichtum in allen Freiheitsgraden (Identität, Mimik, Beleuchtung, Verdeckung etc.) benötigt. Eine weitere Alternative unter Nutzung von 3D-Informationen hat Niese [Nie10, S. 75ff.] vorgeschlagen. Basierend auf einer 3D-basierten Kopfposeschätzung mit dem Iterative Closest Point (ICP) Algorithmus rendert er das Gesicht in der frontalen Pose, ähnlich wie später Hassner et al. [Has+15]. Der Ansatz benötigt jedoch Tiefendaten, also mindestens eine kalibrierte Tiefenkamera oder ein kalibriertes Stereo-Kamerapaar.

3.1.5. Ungleichverteilung der Klassenzugehörigkeit

Die Arbeit von Werner et al. [WSAH15] adressiert die Ungleichverteilung der Klassenzugehörigkeit und ist Grundlage dieses Abschnitts. Standardmethoden des maschinellen Lernens sind voreingenommen (engl. biased) und bevorzugen die häufiger vorkommende Klasse, was zu vielen Falschklassifizierungen der Minderheitsklasse führt [Lop+13; HG09]. Typische Gegenmaßnahmen können in drei Gruppen eingeteilt werden: (1) Stichprobenentnahme (engl. sampling), (2) kostensensitives Lernen und (3) Ensemble-Techniken. Sampling-Methoden modifizieren die

Verteilung der Trainingsdaten, um die Klassenhäufigkeit auszubalancieren, und können mit beliebigen Klassifikatoren verwendet werden. Ein typischer Ansatz ist die zufällige Unterabtastung (engl. random under-sampling), bei der von der Mehrheitsklasse nur eine zufällig ausgewählte Teilmenge genutzt wird. Kostensensitives Lernen gewichtet die verschiedenen Fehlerarten (z. B. falsch positive und falsch negative Klassifikationen) in der beim Lernen optimierten Fehlerfunktion unterschiedlich. Ensemble-Techniken trainieren und kombinieren mehrere Klassifikatoren, wobei sie entweder Sampling oder kostensensitive Methoden verwenden. SMOTE [Cha+02] ist eine weit verbreitete Sampling-Methode zur Überabtastung der Minderheitsklasse, bei der neue Samples durch Interpolation im Merkmalsraum generiert werden. Für große Datensätze ist dies jedoch unpraktikabel, weil die Trainingszeit deutlich verlängert wird. EasyEnsemble [Liu09] ist eine Ensemble-Methode basierend auf Unterabtastung und hat sich in mehreren Evaluationen als zu SMOTE ebenbürtig erwiesen [Lop+13; Liu09; TKY12]. Die meisten Methoden für die Handhabung von Ungleichverteilung wurden für binäre Klassifikation entworfen. Mehrklassenprobleme werden meist auf mehrere Zweiklassenprobleme reduziert.

Jeni et al. [JCD13] haben den Einfluss hochgradig ungleich verteilter Daten auf verschiedene Performance-Maße für die Action-Unit-Detektion untersucht (binäre Klassifikation). Für Regression und Klassifikation der *Intensitäten* von Action-Unit haben Werner et al. [WSAH15] eine ähnliche Untersuchung durchgeführt, die im Anhang A zusammengefasst ist und zum Ergebnis kam, dass das hier verwendete ICC-Maß am besten geeignet ist. Andere Vorarbeiten zur Erkennung von Action Units und Schmerz sowie ihrer Intensitäten untersuchen das Thema nicht ausführlich, nutzen jedoch zum Teil Methoden zur Reduzierung der Ungleichverteilung. Girard et al. [GCD15], Rodriguez et al. [Rod+17] und Huang et al. [Hua+20] nutzen zufälligen Unterabtastung um die Daten in etwa auszubalancieren. Sandbach et al. [SZP13] wählen zufällig fünf Mal so viele Samples der Mehrheitsklasse, wie in allen anderen Klassen zusammen. Rudovic et al. [RPP15] und Yang et al. [Yan+14] betrachten nur einen Teil der Datenbank und gruppieren mehrere seltene Intensitäten zu einer Klasse um die Ungleichverteilung zu reduzieren. Andere Autoren [Bin+14; SST12] schließen bei der AU-Intensitätsschätzung die Klasse 0 (AU nicht aktiv) aus, was die Ungleichverteilung stark reduziert, jedoch von einer fehlerfreien Detektion der AUs (Klasse 0 vs. andere Intensitäten) ausgeht. Zhao et al. [Zha+16b] reduzieren die Ungleichverteilung, indem sie innerhalb der Videos aus langen Bildabfolgen ohne Label-Änderungen nur das erste Bild übernehmen. Viele Arbeiten ignorieren das Problem der Ungleichverteilung und trainieren entweder mit allen verfügbaren Daten [KRP12; Jen+13; MM14; Zha+19; FL20] oder nutzen Sampling-Methoden, die die Ungleichverteilung beibehalten [Mav+13; Min+15].

3.2. Vorgeschlagene Methodik

Im Folgenden werden Methoden für die Entwicklung eines möglichst gut generalisierenden einzelbildbasierten Mimikerkennungssystems vorgeschlagen. Hierfür werden systematisch verschiedene existierende Ansätze angewendet, adaptiert und mit neu entwickelten Ansätzen kombiniert. Eine große Rolle spielen hierbei (1) die begrenzte Verfügbarkeit von Trainingsdaten, (2) die angestrebte Unabhängigkeit der Erkennung von der Kofpose und (3) die Ungleichverteilung der Klassenzugehörigkeiten. Das letztere Thema wird im Abschnitt 3.2.5 behandelt. Die übrigen Unterabschnitte folgen der typischen Verarbeitungskette eines Mimikerkennungssystems, bestehend aus Gesichtsdetektion und Landmarkenlokalisierung (Abschnitt 3.2.1), Gesichtsnormierung (Abschnitt 3.2.2), Merkmalsextraktion (Abschnitt 3.2.3) und dem Lernen eines Erkennungsmodells (Abschnitt 3.2.4).

3.2.1. Vorverarbeitung

Da weder die Gesichtsdetektion noch die Landmarkenlokalisierung im Fokus dieser Arbeit stehen, werden hierzu frei verfügbare Implementierungen von Verfahren des Stand der Technik genutzt, konkret die Software-Bibliotheken dlib [Kin09] (`http://dlib.net/`), OpenFace [BRM16] (`https://github.com/TadasBaltrusaitis/OpenFace`) und IntraFace [Tor+15]. Die Modelle, die mit den Software-Paketen ausgeliefert werden, haben Schwierigkeiten bei stark aus der Kameraebene gedrehten Gesichtern (nicht-frontalen Kopfposen) die Landmarken korrekt zu detektieren. Zwar verfügen IntraFace und OpenFace beide über eine Tracking-Komponente, so dass bei frontaler Initialisierung des Trackings anschließende Bewegungen aus der Ebene gut verfolgt werden können. Dieser Ansatz ist jedoch nicht anwendbar für z. B. FERA 2017 oder die Seitenansichten von BioVid, bei denen die Köpfe zu keinem Zeitpunkt frontal zur Kamera gedreht sind. Bei dlib wird kein Tracking verwendet, jedoch wurde das dort mitgelieferte Modell zur Landmarkenlokalisierung ausschließlich mit dem Datensatz Helen [Le+12] trainiert, in dem nicht hinreichend viele weit aus der Ebene gedrehte Gesichter vorkommen.

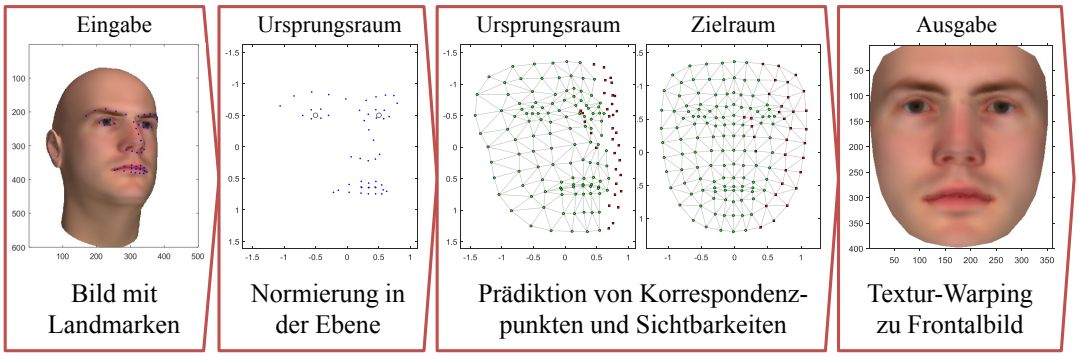
Für die erfolgreiche Lokalisierung der Landmarken in nicht-frontalen Datensätzen wird daher ein eigenes Modell trainiert. Hierfür wird das Verfahren von Kazemi und Sullivan [KS14] in der Implementierung von dlib eingesetzt. Trainiert wurde mit den Datensätzen Multi-PIE [Gro+10], AFW [ZR12], Helen [Le+12], IBUG, 300-W [Sag+16], 300-VW [Chr+15], und LFPW [Bel+11]. Die Punktannotationen für IBUG, AFW, Helen, 300-W und LFPW wurden von Sagonas et al. [Sag+13] zur Verfügung gestellt. Vom Videodatensatz 300-VW wurden die 10 schwierigsten Bilder eines jeden Videos ausgesucht, basierend auf dem Lokalisierungsfehler eines zuvor trainierten Modells. Vom Multi-PIE-Datensatz wurden alle vollständig annotierten Samples der Kameraansichten 080 und 190 verwendet. Das resultierende Modell lokalisiert die Landmarken deutlich genauer, als das mit dlib bereitgestellte Modell [Wer+19c].

Im Folgenden wird, wenn nicht anders erwähnt, das selbst trainierte dlib-Modell verwendet. Dieses lokalisiert wie auch OpenFace insgesamt 68 Punkte. Zu diesen gehören die 49 Punkte, die auch von IntraFace bereitgestellt werden und in Abb. 3.3a visualisiert sind. Zusätzlich gehören dazu 17 Punkte entlang der Kinn- und Wangenkontur des Gesichts und zwei zusätzliche Punkte an den Lippen. Im Folgenden werden zumeist nur die 49 Landmarken verwendet, da diese die für die Mimikerkennung wesentlichen geometrischen Informationen bereits erfassen und die zusätzlichen Punkte keinen signifikanten Mehrwert bringen.

3.2.2. Normierung des Gesichts[2]

Die Registrierung bzw. Normierung von Gesichtsbildern kann mit den in Abschnitt 3.1.4 beschriebenen Verfahren erfolgen. Diese haben jedoch folgende Probleme. Entweder es findet keine Frontalisierung statt, so dass Bilder mit gleicher Mimik bei unterschiedlichen Kopfposen große Varianz aufweisen, oder die Frontalisierung erfordert 3D-Daten, ist sehr rechenaufwändig oder erzeugt Artefakte, die sich negativ auf die Mimikerkennung auswirken können. Die meisten aktuellen Datensätze zur Mimikerkennung beinhalten vor allem frontale Gesichter. Werden Standardverfahren mit diesen Datensätzen trainiert, ist die resultierende Mimikerkennung nicht robust im Bezug auf nicht-frontale Kopfposen. Wenn jedoch nicht-frontale in frontale Gesichter überführt werden können, funktionieren auf frontalen Daten trainierte Modelle auch auf nicht-frontalen

[2]Dieser Abschnitt basiert auf dem Artikel Werner et al. [Wer+19c], in dem der Autor dieser Dissertation seine Arbeiten zur Gesichtsnormierung und Kopfposeinvarianz zuerst veröffentlicht hat.



(a) Verarbeitungskette des vorgeschlagenen Gesichtsnormierungsverfahrens FaNC.

(b) Synthetischer Datensatz SyLaFaN, die Datengrundlage für die Entwicklung des Gesichtsnormierungsverfahrens FaNC: Die Freiheitsgrade Identität, Mimik und Kopfpose werden systematisch varriert. Der Datensatz umfasst 73.800 Bilder, jeweils mit Korrespondenzpunkten, deren Sichtbarkeitsinformation und mit automatisch detektierten Landmarken.

Abbildung 3.1.: Überblick über das vorgeschlagene Verfahren zur Gesichtsnormierung (FaNC). Nach Werner et al. [Wer+19c], © 2019 IEEE.

Gesichtern deutlich besser. In dieser Dissertation wird daher ein neues Verfahren zur Frontalisierung vorgeschlagen, das später experimentell mit den in Abschnitt 3.1.4 beschriebenen Verfahren verglichen wird.

Das vorgeschlagene Verfahren wird **Fa**ce **N**ormalization based on learning **C**orrespondences genannt (**FaNC**, engl. für Gesichtsnormierung auf Basis des Lernens von Korrespondenzen). Es basiert ausschließlich auf Landmarkenkoordinaten und kann in Verbindung mit jeder Methode zur Landmarkenlokalisierung angewendet werden. Kernelement ist die Prädiktion der Koordinaten und Sichtbarkeit von Korrespondenzpunkten anhand der Landmarken. Diese Abbildung von Landmarken auf Korrespondenzpunkte zielt darauf ab, die Kopfpose und individualspezifische Gesichtsproportionen zu kompensieren, die für die Erkennung von Mimik als Störfaktoren angesehen werden können. Abb. 3.1a zeigt die wesentlichen Verarbeitungsschritte des Algorithmus [Wer+19c]. Die Eingaben sind ein beliebiges Bild $\mathbf{F}$ und die Landmarken $\mathbf{l}$ eines darauf befindlichen Gesichts. Im ersten Schritt werden die Landmarken normiert. Es folgt die Prädiktion der Korrespondenzpunkte, sowohl im Ursprungsraum (beliebiges Gesicht) als auch im Zielraum (frontales Gesicht), und die Prädiktion der Sichtbarkeit der Punkte. Zuletzt wird auf Basis der Prädiktionen das normierte Ausgabebild mittels abschnittsweise affinem Warping aus dem Eingabebild generiert, wobei Aufdeckungen von zuvor verdeckten Bereichen mittel Spiegelung aus der anderen Gesichtshälfte und Überblendung behandelt werden. Für das Training der Methode wurde ein synthetischer Datensatz generiert, der im folgenden Absatz vorgestellt wird.

Datensatz SyLaFaN: Zur Realisierung des Frontalisierungsverfahrens hat der Autor dieser Dissertation mit Methoden der Computergrafik einen Datensatz namens **SyLaFaN** erzeugt (engl.

Synthetic dataset for **La**ndmark based **Fa**ce **N**ormalization). Er umfasst 73.800 Bilder, die mittels des 3D Morphable Models (3D-MM) *FaceGen* synthetisiert wurden (`https://facegen.com/`, ähnlich zu [BV99]), wobei Identität, Mimik und Kopfposen systematisch variiert wurden (siehe Abb. 3.1b). 30 Gesichter verschiedener Ethnizitäten, Alter und Geschlechter wurden jeweils mit 30 Gesichtsausdrücken abgebildet, unter ihnen Basisemotionsmimik und Phoneme, so dass sich hier bereits 900 verschiedene texturierte Dreiecksnetze ergeben. Jedes Netz wird in 82 verschiedenen Kopfposen gerendert, unter ihnen die frontale Pose (0° Rotationswinkel) und 81 weitere Posen, die den Winkelbereich von ±45° abdecken, sowohl für den Gierwinkel (engl. yaw angle, nach links/recht drehen) als auch den Nickwinkel (engl. pitch, nach oben/unten drehen). Der Rollwinkel (engl. roll angle, Kopf nach links/rechts neigen) wird nicht variiert, da er durch Rotation in der Bildebene kompensiert werden kann. Für jedes Bild wird eine zuvor definierte Teilmenge der 3D-MM Knotenpunkte in das Bildkoordinatensystem projiziert, um 2D-Korrespondenzpunkte zu erhalten. Bei Kopfrotationen aus der Bildebene werden zum Teil Korrespondenzpunkte verdeckt. Daher wird für jeden Punkt zusätzlich zu den Koordinaten auch ein binärer Sichtbarkeitsindikator bereitgestellt.

Formal umfasst die Datenbank N Samples mit den Indices $i = \{1, 2, \ldots, N\}$, jedes mit einem Bild $\mathbf{F}_i \in \mathbb{R}^{a_1 \times a_2 \times c}$ mit $a_1 \times a_2$ Pixeln und c Kanälen. Für jedes Sample i sind M_p Korrespondenzpunkte $\mathbf{p}_{i,j} \in \mathbb{R}^2$ mit $j = 1, \ldots, M_\mathrm{p}$ gegeben, die in einem Vektor $\mathbf{p}_i \in \mathbb{R}^{2M_\mathrm{p}}$ zusammengefasst werden. Jeder Punkt j ist semantisch äquivalent für alle Samples i. Jedem Korrespondenzpunkt ist eine binäre Sichtbarkeit $v_{i,j} \in \{0, 1\}$ zugeordnet. Die Sichtbarkeiten des Samples i werden im Vektor $\mathbf{v}_i \in \{0, 1\}^{M_\mathrm{p}}$ zusammengefasst. Außerdem verfügbar sind M_l Gesichtslandmarken $\mathbf{l}_{i,j} \in \mathbb{R}^2$ mit $j = 1, \ldots, M_\mathrm{l}$, die in einem Vektor $\mathbf{l}_i \in \mathbb{R}^{2M_\mathrm{l}}$ zusammengefasst werden. Die Landmarken können mit einer beliebigen Methode automatisch lokalisiert werden. Im Folgenden werden die Landmarken verwendet, die mit der in Abschnitt 3.2.1 beschriebenen Methode lokalisiert wurden.

Normierung in der Bildebene: Die Gesichtslandmarken und Korrespondenzpunkte werden mit einer Ähnlichkeitsabbildung (ohne Spiegelung) registriert, um verschiedene Rotationen in der Bildebene, Translationen und Skalierungen auszugleichen. Die Abbildung $s(\mathbf{x})$ wird anhand der Augenmittelpunkte berechnet, die aus den Landmarken der Augenwinkel bestimmen lassen und anschließend auf alle Koordinaten $\mathbf{p}$ und $\mathbf{l}$ angewendet: $\hat{\mathbf{p}} = s(\mathbf{p})$ und $\hat{\mathbf{l}} = s(\mathbf{l})$. Die Prädiktion der Korrespondenzpunkte und der Sichtbarkeiten arbeitet ausschließlich in diesem normalisierten Koordinatensystem.

Prädiktion der Korrespondenzpunkte: Die Aufgabe beliebige Gesichter (Ursprungsraum in Abb. 3.1a) auf die gewünschten normierten Gesichter (Zielraum in Abb. 3.1a) abzubilden wird über eine Index-Zuordnungsfunktion $t(i) : \mathbb{N} \mapsto \mathbb{N}$ realisiert, die jedem Beispiel des SyLaFaN-Datensatzes ein zugehöriges frontales Ziel-Sample zuweist. Dem Bild $\mathbf{F}_i$ wird das frontale Bild $\mathbf{F}_{t(i)}$ zugeordnet und den Korrespondenzpunkten $\mathbf{p}_i$ die frontalen Korrespondenzpunkte $\mathbf{p}_{t(i)}$. Für die Aufgabe der Mimikerkennung wählt $t(i)$ das Sample mit frontaler Pose und selber Mimik wie das Sample i, jedoch von einer mittleren Identität. Auf diese Weise sollen neben der Kopfpose auch geometrische Unterschiede zwischen verschiedenen Personen, wie Gesichtsproportionen, normiert werden, um die Unterschiede zwischen verschiedenen Personen zu reduzieren, was für die Mimikanalyse von Vorteil sein kann.

Es wird ein Modell gelernt, das anhand der normierten Landmarken $\hat{\mathbf{l}}$ die Koordinaten der Korrespondenzpunkte $\hat{\mathbf{p}}$ prädiziert. Genauer gesagt wird die Grundwahrheit für den Ausgabevektor $\mathbf{y}_i$ des Sample i gebildet, indem die Korrespondenzpunkte des Ursprungsraumes $\hat{\mathbf{p}}_i$ (beliebige Pose) mit denen des Zielraumes $\hat{\mathbf{p}}_{t(i)}$ (zugeordnete frontale Pose) konkateniert werden, das

heißt $\mathbf{y}_i = [\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_{t(i)}]$. Die Abbildung wird über ein lineares Modell $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$ realisiert, um für die Prädiktion eine minimale Laufzeit zu erreichen und Überanpassung auf den verwendeten synthetischen Datensatz zu vermeiden. Da das modellierte Problem nicht linear ist, werden nicht-lineare Merkmale eingesetzt. Neben den normierten Landmarken $\hat{\mathbf{l}}$ werden auch Landmarken $\check{\mathbf{l}}$ genutzt, die anhand der Mundwinkel (anstelle der Augenmittelpunkte) ausgerichtet werden. Der Merkmalsvektor umfasst außerdem die elementweisen Quadrate $\hat{\mathbf{l}}^2$ und $\check{\mathbf{l}}^2$, das heißt $\mathbf{x}_i = [\hat{\mathbf{l}}_i, \check{\mathbf{l}}_i, \hat{\mathbf{l}}_i^2, \check{\mathbf{l}}_i^2]$. Zum Trainieren wird $\mathbf{W} \in \mathbb{R}^{4M_\mathrm{p} \times 8M_\mathrm{l}}$ und $\mathbf{b} \in \mathbb{R}^{4M_\mathrm{p}}$ in $4M_\mathrm{p}$ Modelle zerlegt (eines für jede Ausgabedimension). Die Wahl der Modellparameter erfolgt über die Minimierung des L2-regularisierten L2-Loss für Support Vector Regression mit der Bibliothek LIBLINEAR [Fan+08]. Vor dem Training werden die Merkmalsvektoren $\mathbf{x}$ standardisiert. Die Regressionsmodelle der Ursprungsraumkoordinaten werden nur mit den Samples trainiert, in denen der jeweilige Korrespondenzpunkt sichtbar ist, da das Warping nur die Koordinaten sichtbarer Punkte nutzt.

Prädiktion der Sichtbarkeiten: Ähnlich zu der Prädiktion der Punktkoordinaten werden Modelle gelernt, um die Sichtbarkeiten $\mathbf{v}$ der Korrespondenzpunkte zu prädizieren. Eingabe des linearen Modells sind auch hier die normierten Landmarken $\hat{\mathbf{l}}$, genauer gesagt die oben beschriebenen Merkmale $\mathbf{x}$. Da nur eine Ausgabe je Korrespondenzpunkt nötig ist, sind die Parametermatrizen in diesem Fall $\mathbf{W} \in \mathbb{R}^{M_\mathrm{p} \times 8M_\mathrm{l}}$ und $\mathbf{b} \in \mathbb{R}^{M_\mathrm{p}}$. Außerdem handelt es sich bei den Sichtbarkeiten um binäre Variablen, so dass Klassifikatoren anstelle von Regressoren gelernt werden, genauer gesagt M_p Support Vector Machines mit LIBLINEAR [Fan+08] mit L2-regularisiertem L2-Loss. Um Probleme durch ungleichmäßige Verteilung der Daten (engl. imbalanced data) [HG09; Lop+13] zu vermeiden, wird für die häufiger vorkommende Klasse nur eine zufällig gewählte Teilmenge für das Training genutzt, so dass die Klassenverteilung ausgewogen ist.

Textur-Warping: Das Warping (engl. für Verformung, Verzerrung) des Originalbildes zum Frontalbild basiert auf abschnittsweisen affinem Warping eines Dreiecksnetzes. Das Netz (siehe Zielraum in Abb. 3.1a) wurde mittels Delaunay-Triangulation der Korrespondenzpunkte eines frontalen Bildes der Datenbank SyLaFaN erzeugt. Im Gegensatz zum typischen abschnittsweisen affinem Warping, sind die Koordinaten der Knotenpunkte des Zielraumes nicht konstant, sondern sind wie die des Ursprungsraumes vom Eingabebild abhängig. Außerdem werden für das Warping nicht die Landmarken, sondern die prädizierten Korrespondenzpunkte genutzt. Aufdeckungen von im Originalbild verdeckten Bereichen werden behandelt, indem die betroffenen Regionen von der anderen Gesichtshälfte übernommen werden, wobei starke Kanten durch Überblendung mit weichen Übergängen vermieden werden.

Zum Warping eines Bildes werden zunächst die prädizierten Korrespondenzpunkte des Ursprungsraumes in den Raum des Eingabebildes zurück transformiert, indem die Normierung in der Bildebene rückgängig gemacht wird. Die Koordinaten im Zielraum werden in den Ausgabebildraum transformiert, wobei die Auflösung und die Position des Gesichtes frei gewählt werden kann. Die prädizierten binären Sichtbarkeiten werden wie folgt nachbearbeitet: (1) In Dreiecken mit einem oder zwei unsichtbaren Vertices ($v_{i,j} = 0$) werden alle Vertices auf unsichtbar gesetzt ($v_{i,j} := 0$). (2) In den benachbarten Dreiecken werden sichtbare Vertices ($v_{i,j} = 1$) auf halb sichtbar ($v_{i,j} := 0{,}5$) gesetzt. (3) In der Gesichtshälfte, die mehr sichtbare Vertices hat, werden alle Vertices auf sichtbar gesetzt. Anschließend wird das abschnittsweise affine Warping durchgeführt. In einem ersten Durchlauf werden die Eingabekoordinaten für jedes Dreieck, das mindestens eine Vertex-Sichtbarkeit $v_{i,j} < 1$ hat, auf die Koordinaten des korrespondierenden Dreiecks der anderen Gesichtshälfte gesetzt, das heißt die Textur wird für diese Dreiecke von der Seite gespiegelt, die der Kamera zugewandt ist. In einem zweiten Durchlauf wird mittels Alpha Blending verhindert, dass an den Rändern der gespiegelten Dreiecke starke Kanten entstehen. Dabei wird

jedes Dreieck mit mindestens einem Vertex $0 < v_{i,j} < 1$ unter Nutzung von Alpha Blending mit $\alpha_{i,j} = 1 - v_{i,j}$ mit dem Bild aus dem ersten Durchlauf vermischt. Zwischen den Vertices der Dreiecke wird α linear interpoliert, wodurch starke Kanten zwischen sichtbaren und gespiegelten Dreiecken verschwinden.

Beispielbilder für die Eingaben und Resultate des beschriebenen Normierungsverfahrens FaNC sind in Abb. 3.1a (S. 55) und Abb. 3.8 (S. 82) zu finden.

3.2.3. Merkmalsextraktion

Der Stand der Technik bietet eine reiche Auswahl an Methoden zur Merkmalsextraktion. Klassische Merkmalsarten sind menschengemacht, d. h. von Ingenieuren anhand von Vorwissen und Intuition entworfen worden. In Abgrenzung dazu gibt es auch gelernte Merkmale, deren Extraktion anhand von Daten und einer Loss-Funktion optimiert wurde. Die Folgenden Merkmalsarten werden häufig verwendet und daher später experimentell evaluiert:

Landmarken: Die lokalisierten Landmarken des Gesichts können als geometrische Merkmale für die Mimik- und Schmerzerkennung genutzt werden. In den Untersuchungen werden die 49 Punkte genutzt, die alle Landmarkendetektoren bereitstellen, vgl. Abb. 3.3a auf S. 65. Um Invarianz gegenüber Translation, Skalierung und Rotation in der Ebene zu erreichen, werden die Punkte jedoch zunächst mit den korrespondierenden Punkten eines immer gleichen Referenzgesichtes registriert. Dabei werden die Parameter einer Ähnlichkeitsabbildung (engl. similarity transformation) so gewählt, dass der mittlere quadratische Abstand zwischen korrespondierenden Punkten des beobachteten Gesichtes und des Referenzgesichtes minimiert wird. Die x- und y-Koordinaten der registrierten Punkte bilden den 98-dimensionalen Merkmalsvektor.

Uniform LBP: Als Erscheinungsbildmerkmale sind Histogramme von Local Binary Patterns (LBP) [AHP06] weit verbreitet. Für die Extraktion wird jeder Pixel anhand der Helligkeiten seiner Nachbarschaft einem Muster zugewiesen und die Häufigkeit der Muster wird in einem Histogramm gezählt. Jeder Punkt der Nachbarschaft (hier 8 Nachbarn im Abstand von 1 Pixel zum Zentrum) wird bezüglich seiner Helligkeit mit dem zentralen Pixel verglichen. Wenn ein Nachbar dunkler als der zentrale Pixel ist, wird eine 0 codiert, andernfalls eine 1. Die Kombination der Ergebnisse der 8 Nachbarn ergibt einen Vektor von 8 Bits, der als eine Ganzzahl im Wertebereich $[0, 255]$ aufgefasst wird, als die „Nummer“ des Musters. Beim ursprünglichen LBP wird hier somit ein 256-dimensionales Histogramm als Merkmalsvektor gebildet. Da viele der Muster selten vorkommen, werden meist „Uniform LBP“ extrahiert. Hierbei wird die Dimensionalität des Merkmalsvektors reduziert, indem nur Muster mit maximal zwei Übergängen zwischen 0 und 1 in separaten Histogrammeinträgen gezählt werden. Alle übrigen Muster werden gemeinsam in einem extra Eintrag gezählt, so dass sich bei 8-Nachbarschaft ein 59-dimensionales Histogramm ergibt. Um lokale Informationen zu erfassen, wird das normierte Bild des Gesichts meist mit einem regelmäßigen Gitter aufgeteilt, ein separates LBP-Histogramm für jede Gitterzelle erzeugt und die Histogramme konkateniert [KRP12; Mav+13; SZP13]. In dieser Dissertation werden Gitter mit 10×10 Zellen (5.900-dim. Merkmalsvektor) und 5×5 Zellen (1.475-dim.) eingesetzt.

CNN-Merkmale: Trainierte CNNs können zur Merkmalsextraktion eingesetzt werden (vgl. Transferlernen in Abschnitt 3.1.1 und z. B. [EVM17]). Dabei werden die Aktivierungen einer Zwischenschicht als Merkmalsvektor aufgefasst und dieser mit einem klassischen

Lernverfahren wie SVM verarbeitet. In den folgenden Untersuchungen werden die Merkmale aus der letzten Schicht vor den Ausgabeneuronen extrahiert (1.280-dim. Vektor bei MobileNetV3).

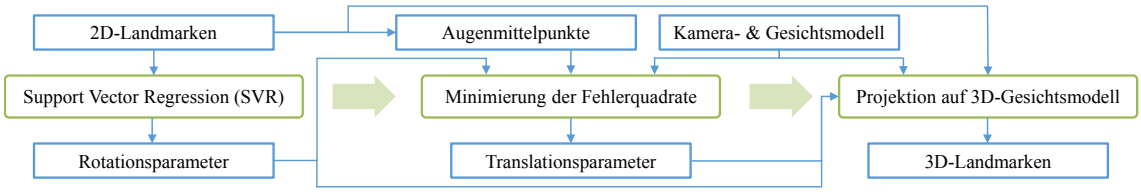
Diese Merkmale sind für verschiedenste Anwendungen nutzbar und im Allgemeinen nicht auf die Zielanwendung optimiert, d. h. sie enthalten auch Informationen, die irrelevant sind, was das Lernproblem erschweren und zu schlechterer Erkennungsleistung führen kann.

Eine für die jeweilige Zielanwendung optimierte Merkmalsextraktion lässt sich auf zwei Wegen erreichen. Zum einen kann ein Lernverfahren verwendet werden, das optimale Merkmale selektiert (wie z. B. spezielle Random Forests [Sho+13] oder AdaBoost [VJ01]) oder Filter zur Merkmalsextraktion optimiert (wie CNNs). Auf letztere Möglichkeit wird im Abschnitt 3.2.4 eingegangen. Zum anderen kann Vorwissen genutzt werden, um manuell einen Merkmalsraum zu definieren, der für die Anwendung geeignet ist. Diese Herangehensweise hat zwei Vorteile. Erstens sind die resultierenden Merkmale leicht zu interpretieren, was Rückschlüsse auf die Funktionsweise des gelernten Modells erlaubt. Zweitens ist der Merkmalsraum meist niedrigdimensionaler als der von generischen Merkmalen, was insbesondere bei Anwendungen günstig ist, bei denen nur relativ kleine Datensätze zur Verfügung stehen, wie auch in der Schmerzerkennung. In dieser Dissertation werden drei Arten von Merkmalen vorgeschlagen, die auf Vorwissen basieren: die Kopfpose, eine spezifische Auswahl von 3D-Abständen zwischen Landmarken des Gesichts und eine Auswahl von 3D-Koordinaten von Landmarken.

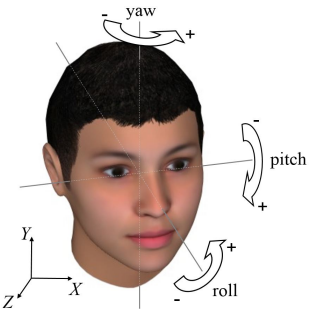
Kopfpose: Die Kopfpose beschreibt die Position und Orientierung des Kopfes im 3D-Raum, siehe Abb. 3.2b, in dieser Arbeit relativ zur Kamera. Die Orientierung des Kopfes im Raum ist für die reine Mimikerkennung zunächst ein Störfaktor, da verschiedene Kopfposen die Bilder des Gesichts trotz gleicher Mimik stark verändern. Da es eine systematische Störung ist, kann die Bezifferung dieses Störfaktors als Merkmal dem maschinell gelernten Modell jedoch auch helfen, den negativen Einfluss auf die Mimikerkennung zu reduzieren. Im Kontext der Erkennung affektiver Zustände kann die Kopfpose (wie auch die Mimik) als unabhängige Modalität betrachtet werden. Studien haben beispielsweise gezeigt, dass ein gesenkter Kopf auf Emotionen wie Scham, Verlegenheit oder Traurigkeit hindeuten kann [MC03; ACR09; Wer+18]. Auch im Zusammenhang mit Schmerzen ist oft ein gesenkter Kopf zu beobachten, ebenso wie eine Kopfbewegung in Richtung des Schmerzenden Körperteils oder Veränderungen in der Bewegungsgeschwindigkeit. Diese Effekte wurden statistisch erstmals in einer Studie des Autors dieser Dissertation nachgewiesen [Wer+18], worauf hier aus Platzgründen jedoch nicht detailliert eingegangen wird.

Die Kopfpose wird in dieser Arbeit anhand von 2D-Landmarken geschätzt. Das für diese Dissertation entwickelte Verfahren, das hierfür eingesetzt wird, ist in Abb. 3.2a schematisch zusammengefasst. Neben der Schätzung der Kopfpose ermöglicht es auch die Schätzung der 3D-Koordinaten der Landmarken, die anschließend für die Extraktion von Mimikmerkmalen genutzt werden.

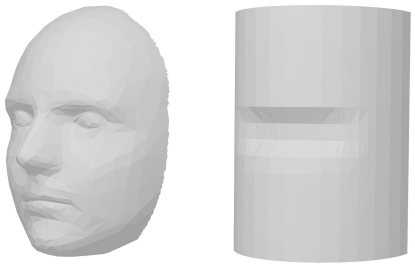
Zunächst wird die Orientierung des Kopfes (in Form der Rotationsparameter Nick-, Gier- und Rollwinkel, engl. pitch, yaw, roll, vgl. Abb. 3.2b) bestimmt, wozu das vom Autor dieser Dissertation bereits in Werner et al. [WSAH17] beschriebene Verfahren eingesetzt wird: Im ersten Schritt werden die Landmarken auf Basis der Augenmittelpunkte normiert, indem der Abstand auf eine Distanz von eins skaliert und der Mittelpunkt auf den Koordinatenursprung verschoben wird. Hierdurch werden die nachfolgenden Schritte unabhängig von der Skalierung und Translation. Diese normierten Koordinaten werden als Merkmale für drei Support Vector Regressionen (SVR) mit RBF-Kernel verwendet (eine Regression je Rotationswinkel). Zum Training kommt ein hierfür erzeugter synthetischer Datensatz zum Einsatz [WSAH17], ähnlich dem Datensatz SyLaFaN, der im vorherigen Abschnitt vorgestellt wurde. Die SVR-Parameter C, γ und ε werden mittels



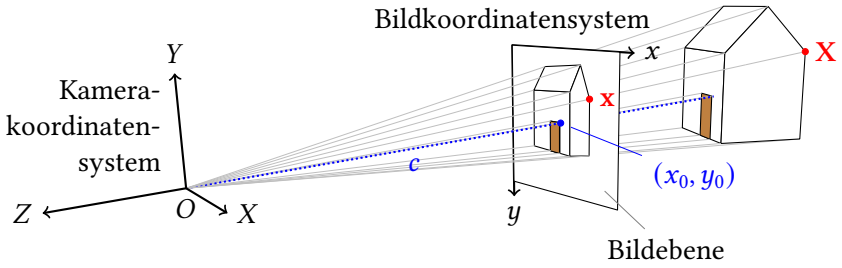
(a) Verfahren zur Schätzung der 3D-Kopfpose und 3D-Landmarken. Eingaben (oben), Ausgaben (unten) und Schritte des Schritte des Algorithmus (mittig, grün).



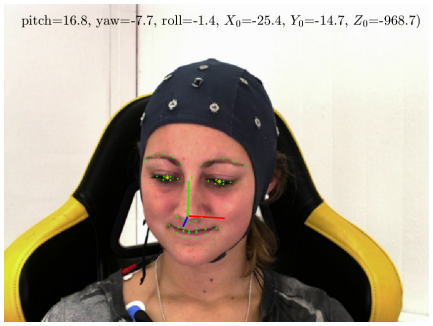
(b) Definition der Kopfpose durch drei Rotationswinkel (pitch, yaw, roll) und drei Translationsparameter (X_0, Y_0, Z_0).



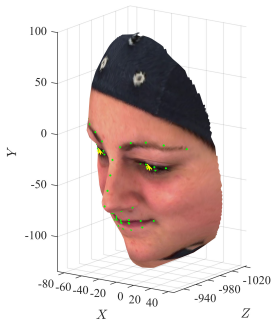
(c) Verwendete 3D-Gesichtsmodelle: realistisches Modell (links) und zylinderbasiertes Modell (rechts).



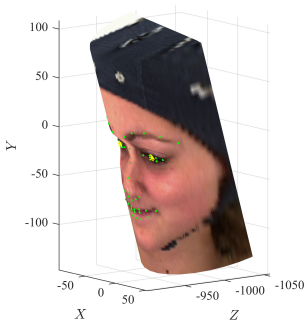
(d) Veranschaulichung des Lochkameramodells mit 3D-Kamerakoordinatensystem (in mm), 2D-Bildkoordinatensystem in der Bildebene (in Pixel), Projektionszentrum $O = (0,0,0)$, Bildhauptpunkt (x_0, y_0) (blau, Schnittpunkt der Z-Achse mit der Bildebene), Kamerakonstante c (Abstand zwischen O und (x_0, y_0) in mm), Sehstrahlen (grau), 3D-Punkt **X** und zugehörigem Bildpunkt **x**. Die Z-Achse und die Sehstrahlen bilden mit den Punkten **x** und **X** ähnliche Dreiecke, deren Seitenlängenverhältnisse ausgenutzt werden.



(e) Beispielbild der Datenbank BioVid-S mit 2D-Landmarken (grün), Augenmittelpunkten (gelb) und Kopfpose (Koordinatensystem auf Nase).



(f) Realistisches 3D-Gesichtsmodell zu (e) in Kamerakoordinatensystem mit projizierten Landmarken.



(g) Zylinderbasiertes 3D-Gesichtsmodell zu (e) in Kamerakoordinatensystem mit projizierten Landmarken.

Abbildung 3.2.: Überblick zur Schätzung der 3D-Kopfpose und 3D-Landmarken anhand von 2D-Landmarken.

einer Gittersuche auf dem Trainingsdatensatz ermittelt [HCL03]. Hierbei wird der Trainingsdatensatz zufällig in zwei Teile ohne Personenüberlappung aufgeteilt. Eine Teilmenge (zufällig reduziert auf 500 Samples) wird zum Training und eine (5.000 Samples) für die Validierung verwendet. Für jede Parameterkombination wird ein SVR-Training mit Validierung durchgeführt. Die Gittersuche wird dreimal wiederholt und die Winkelfehler gemittelt. Anschließend werden die SVR mit den am besten abschneidenden Parametern auf dem gesamten Trainingsdatensatz (mit mehr Samples) neu trainiert. Das resultierende Modell kann angewendet werden, um bei einem neuen Bild anhand der Landmarken die Orientierung des Kopfes zu schätzen [WSAH17].

Neben der Orientierung des Kopfes soll auch die Lage im 3D-Raum (inklusive der Entfernung zur Kamera) sowie 3D-Abstände zwischen Landmarken des Gesichts ermittelt werden. Da in vielen Datensätzen keine 3D-Daten vorliegen oder die Orts- und Tiefenauflösung der 3D-Daten gering ist, werden in dieser Arbeit die 3D-Koordinaten und -Abstände aus den 2D-Landmarken abgeschätzt, indem einige Annahmen getroffen und ausgenutzt werden:

1. Es wird angenommen, dass alle Personen die gleiche 3D-Gesichtsform haben. Um einzuschätzen, wie wichtig die geometrisch exakte Übereinstimmung zwischen dem realen beobachteten Gesicht und dem Modell ist, werden zwei 3D-Modelle genutzt und verglichen, siehe Abb. 3.2c. Bei einem Modell handelt es sich um das Standardkopfmodell der Software *FaceGen* (`https://facegen.com/`). Das zweite, einfachere Gesichtsmodell basiert auf einem elliptischen Zylinder, bei dem eine Vertiefung zur Annäherung der Augenhöhlen eingefügt wurde.

2. Es wird angenommen, dass der Augenabstand aller betrachteten Personen gleich ist. Es ist bekannt, dass der Augenabstand (gemessen als Abstand zwischen den Pupillen) variiert, insbesondere abhängig von Alter, Geschlecht und Ethnizität; im Mittel beträgt er etwa 63 mm [Dod04]. Er ist jedoch einer der wenigen Abstände im Gesicht, die robust anhand von Landmarken gemessen werden können und nicht mit der Mimik variieren. Daher wird der mittlere Augenabstand als konstanter und bekannter Maßstab im Bild angenommen, um die Entfernung des Kopfes von der Kamera schätzen zu können. Um dies umzusetzen, wurden die Augenmittelpunkte in den oben erwähnten 3D-Gesichtsmodellen markiert und die Modelle so skaliert, dass der gewünschte Augenabstand vorliegt.

3. Mithilfe eines Kameramodells und der zu einer konkreten Kamera gehörenden Kameraparameter kann ein mathematischer Zusammenhang zwischen dem aufgenommenen 3D-Raum und dem zugehörigen 2D-Bild hergestellt werden. Bei existierenden Datenbanken sind die Kameraparameter jedoch oft nicht bekannt, da keine Kalibrierung durchgeführt wurde. Als Ausgangspunkt für eine Lösung dieses Problems wird ein ideales Lochkameramodell angewendet, das die Projektion aus dem 3D-Raum in das 2D-Bildkoordinatensystem als zentralperspektive Abbildung modelliert [Luh10, S. 52f.], [Nie10, S. 27ff.], siehe Abb. 3.2d. Abweichungen hiervon, wie z. B. die radialsymmetrische Verzeichnung, werden nicht modelliert. Die Verwendung dieses einfachen Modells ermöglicht die Abschätzung der Kameraparameter auch ohne spezielle, für die Kalibrierung bestimmte Aufnahmen. Es wird die perspektivische Projektion

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \cdot w = \mathbf{K} \cdot \mathbf{X} = \begin{bmatrix} c & cs & x_0 \\ 0 & cm & y_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{3.5}$$

angewendet, die jeden 3D-Objektpunkt $\mathbf{X} = (X, Y, Z)$ (im Kamerakoordinatensystem) unter Verwendung von homogenen Koordinaten in den zugehörigen 2D-Bildpunkt $\mathbf{x} = (x, y)$ abbildet. Der Faktor $w \neq 0$ dient der Skalierung der homogenen Koordinaten des 2D-Bildpunktes (x, y). In (3.5) ergibt sich $w = Z$. Die Abbildung nutzt die Kameraparameter

der inneren Orientierung, die sich in der Kalibriermatrix **K** wiederfinden: die Kamerakonstante c, den Bildhauptpunkt (x_0, y_0), sowie die Parameter s und m, die Scherungen und Maßstabsunterschiede der Bildkoordinatenachsen beschreiben. Sie werden wie folgt abgeschätzt:

$$s = 0, \qquad m = -1, \tag{3.6}$$

$$x_0 = \frac{b}{2}, \qquad y_0 = \frac{h}{2}, \tag{3.7}$$

$$c = \frac{-0{,}5b}{\tan(0{,}5\alpha_b)}, \qquad \text{mit } \alpha_b = 30^\circ. \tag{3.8}$$

Es wird angenommen, dass keine Scherungen und Maßstabsunterschiede vorliegen (3.6). Der Faktor $m = -1$ kompensiert die unterschiedlichen Richtungen der y-Achsen von Bild- und Kamerakoordinatensystem, siehe Abb. 3.2d. Außerdem wird angenommen, dass der Bildhauptpunkt exakt in der Mitte des Bildes mit der Breite b und der Höhe h liegt (3.7) und dass die Kamera einen horizontalen Öffnungswinkel von $\alpha_w = 30^\circ$ hat (3.7). Nach qualitativen Betrachtungen mit den in dieser Arbeit verwendeten Daten bewertet der Autor dieser Dissertation diese einfache Kameramodellierung als hinreichend genau, da auch die anderen Bestandteile des Verfahrens (algorithmisch ermittelte Landmarkenkoordinaten, Kopfposeschätzung, 3D-Modell) mit Fehlern behaftet sind, die einen größeren Einfluss haben. Durch Vereinfachung der Gleichung (3.5) mit (3.6) ergibt sich:

$$x = c\frac{X}{Z} + x_0, \qquad y = -c\frac{Y}{Z} + y_0. \tag{3.9}$$

Die innere Orientierung der Kamera wird mit den unter Punkt 3 beschriebenen Annahmen anhand der Größe des Bildes berechnet. Die äußere Orientierung, d. h. die Translation und Rotation der Kamera wird relativ zum Kopf bestimmt und ist somit äquivalent zur Kopfpose (Rotation und Translation des Kopfes relativ zur Kamera). Die Abbildung vom lokalen Koordinatensystem des Gesichtsmodells in Kamerakoordinaten wird wie folgt modelliert:

$$\mathbf{X} = \mathbf{R} \cdot \mathbf{X_m} + \mathbf{X_0}. \tag{3.10}$$

Der Gesichtsmodellpunkt $\mathbf{X_m} \in \mathbb{R}^3$ wird unter Verwendung der Rotationsmatrix $\mathbf{R} \in \mathbb{R}^{3\times3}$ und des Verschiebungsvektors $\mathbf{X_0} = (X_0, Y_0, Z_0) \in \mathbb{R}^3$ in den Punkt $\mathbf{X} \in \mathbb{R}^3$ im Kamerakoordinatensystem abgebildet. Die Rotationsmatrix **R** ist wie in Vorarbeiten zur Kopfposeschätzung [XD13; BRM16] definiert:

$$\mathbf{R} = \mathbf{R_x}(\varphi) \cdot \mathbf{R_y}(\theta) \cdot \mathbf{R_z}(\psi), \tag{3.11}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{bmatrix} \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.12}$$

Die Matrix setzt sich aus Rotationen um die x-Achse, die y-Achse und die z-Achse zusammen, jeweils um die Winkel φ (pitch), θ (yaw) und ψ (roll). Diese drei Kopfrotationswinkel (und somit auch **R**) sind bereits über die SVR-basierte Kopfposeschätzung gegeben. Durch Einsetzen von (3.10) in (3.9) und Vereinfachung der Schreibweise durch $\mathbf{R} \cdot \mathbf{X_m} = \mathbf{X_m^*} = (X_m^*, Y_m^*, Z_m^*)$ erhält man:

$$x = c \cdot \frac{X_m^* + X_0}{Z_m^* + Z_0} + x_0, \qquad y = -c \cdot \frac{Y_m^* + Y_0}{Z_m^* + Z_0} + y_0. \tag{3.13}$$

Gesucht ist die Kopfposition im 3D-Raum $\mathbf{X_0} = (X_0, Y_0, Z_0)$. Daher werden die Gleichungen (3.13) umgestellt, um ein lineares Gleichungssystem der Form $\mathbf{A} \cdot \mathbf{X_0} = \mathbf{b}$ zu erhalten:

$$cX_0 + (x_0 - x)Z_0 = -cX_m^* + (x - x_0)Z_m^*, \qquad -cY_0 + (y_0 - y)Z_0 = cY_m^* + (y - y_0)Z_m^*, \tag{3.14}$$

$$\begin{bmatrix} c & 0 & x_0 - x \\ 0 & -c & y_0 - y \end{bmatrix} \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \end{bmatrix} = \begin{bmatrix} -cX_m^* + (x - x_0)Z_m^* \\ cY_m^* + (y - y_0)Z_m^* \end{bmatrix}. \tag{3.15}$$

Dieses Gleichungssystem ist unterbestimmt und damit nicht eindeutig lösbar. Es kann jedoch um weitere Gleichungen ergänzt werden, indem mehrere korrespondierende Paare von Bildpunkt $\mathbf{x}$ und zugehörigem Modellpunkt $\mathbf{X_m}$ (bzw. entsprechend der Kopfpose rotiertem Modellpunkt $\mathbf{X_m^*}$) betrachtet werden. Hier werden die beiden Augenmittelpunkte verwendet, die am 3D-Modell fest definiert sind und im Bild anhand der detektierten Landmarken bestimmt werden (für jedes Auge der Mittelpunkt zwischen den Augenwinkeln). So ergeben sich vier Gleichungen für die drei unbekannten Translationsparameter $\mathbf{X_0} = (X_0, Y_0, Z_0)$. Das nun überbestimmte Gleichungssystem kann im Allgemeinen nicht exakte gelöst werden. Es wird daher eine Näherungslösung $\mathbf{X_0}$ berechnet, die die Summe der Fehlerquadrate minimiert, formal $\min_{\mathbf{X_0}} (\mathbf{AX_0} - \mathbf{b})^T (\mathbf{AX_0} - \mathbf{b})$, indem die Gaußschen Normalengleichungen gelöst werden ($\mathbf{A}^T\mathbf{AX_0} = \mathbf{A}^T\mathbf{b}$).

Damit sind alle sechs Kopfposeparameter (pitch φ, yaw θ, roll ψ, X_0, Y_0, Z_0) bekannt. Für die anschließende Bestimmung von 3D-Abständen werden nun die 3D-Koordinaten der Landmarken geschätzt, indem die 2D-Landmarken unter Verwendung des Kameramodells auf das 3D-Gesichtsmodells projiziert werden, welches gemäß der berechneten Pose im 3D-Raum platziert ist. Hierfür wird wie bei Niese [Nie10, S. 87] ein Schnitttest mit dem Dreiecksnetz des Gesichtsmodells durchgeführt, d. h. es wird der Schnittpunkt zwischen dem Sehstrahl der Landmarke und dem Modell berechnet. Wenn mehrere Schnittpunkte gefunden werden, wird nur der Schnittpunkt genutzt, welcher der Kamera am nächsten ist. Die Projektion der Landmarken wird in Abb. 3.2e-3.2g veranschaulicht: Abb. 3.2e zeigt das Ausgangsbild mit den 2D-Landmarken und der geschätzten Kopfpose, Abb. 3.2f das entsprechend der Kopfpose im 3D-Raum platzierte Gesichtsmodell mit den projizierten 3D-Landmarken und Abb. 3.2g die Ergebnisse bei Verwendung eines einfacheren zylinderbasierten Gesichtsmodells. Die 3D-Landmarken können für die Berechnung von 3D-Merkmalen anhand der inversen Transformation von (3.10) vom Kamerakoordinatensystem $\mathbf{X} = (X, Y, Z)$ in das lokale Koordinatensystem des Gesichtsmodells (u, v, w) überführt werden:

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \mathbf{R}^{-1} \cdot (\mathbf{X} - \mathbf{X_0}). \tag{3.16}$$

Ein ähnliches Verfahren für die Kopfposeschätzung anhand von 2D-Bildern wurde von Niese vorgeschlagen [Nie10, S. 84ff.]. Niese schätzt die Rotation und Translation des Kopfes gemeinsam anhand von Ankerpunkten mit einer Minimierung der Fehlerquadrate, die hier lediglich für die Translationsparameter eingesetzt wird. Problematisch an Nieses Ansatz ist, dass personenspezifische Größenverhältnisse (verschieden lange Nasen oder „höhere“ Gesichter) nicht hinreichend berücksichtigt werden, was insbesondere die Schätzung des pitch-Winkels negativ beeinträchtigen kann. Im hier vorgeschlagenen Ansatz werden für die Bestimmung der Rotationswinkel SVR-Modelle eingesetzt, welche die Varianz in der Gesichtsgeometrie mit lernen und somit auch besser kompensieren können. Nieses Ansatz erfordert außerdem eine frontale Gesichtshaltung zur Initialisierung und Reinitialisierung, wodurch die Einsatzszenarien eingeschränkt werden.

Durch das eingesetzte Tracking wird Nieses Verfahren auch fehleranfälliger, insbesondere bei Rotationen aus der Bildebene. Niese berichtet, dass das Tracking lediglich im Winkelbereich von $[-25°, 25°]$ (yaw und pitch) stabil arbeitet.

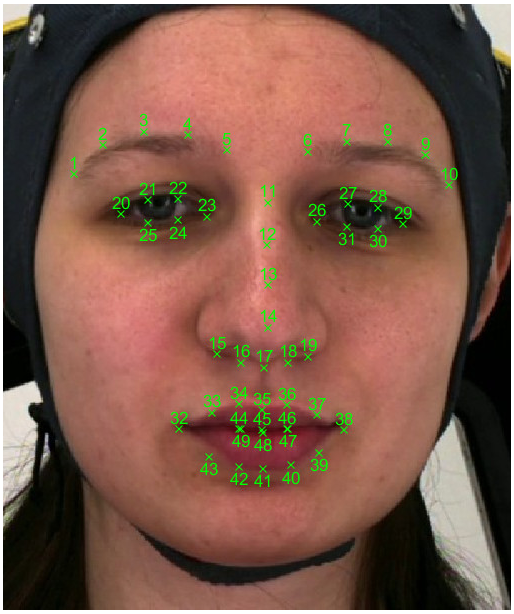
3D-Abstände: Inspiriert von der Arbeit von Niese [Nie10] werden 3D-Abstände als Merkmale für die Mimikerkennung evaluiert. Die Idee besteht darin, durch Messung im realen 3D-Raum (anstelle der Messung im 2D-Bild) Merkmale zu extrahieren, die von der Kopfpose unabhängig sind, d. h. nicht nur von der Position und Skalierung im Bild sondern auch von der Orientierung des Kopfes im 3D-Raum. Die konkreten Abstandsmerkmale, die vom Autor dieser Dissertation bereits in [Wer+14b; Wer+17] vorgestellt wurden, werden in Abb. 3.3a-c definiert und veranschaulicht. Sie wurden entworfen, um Komponenten der Schmerzmimik erfassen zu können, wie das Senken der Augenbrauen (AU 4), das Anheben der Oberlippe (AU 10), das Schließen der Augen (AU 43), oder auch das Öffnen der Lippen und des Mundes (AU 25/26) sowie das Auseinanderziehen der Lippen (AU 20). Die Berechnung erfolgt mittels des Euklidischen Abstandes jeweils zweier Landmarken im 3D-Raum, entweder in Kamerakoordinaten (X, Y, Z) oder im lokalen Gesichtskoordinatensystem (u, v, w).

3D-Koordinaten: Im lokalen Gesichtskoordinatensystem (u, v, w) wird die berechnete Kopfpose kompensiert, um ein kopfposeunabhängiges Bezugssystem zu schaffen, das sich zur Messung mimischer Bewegungen eignet. Entsprechend werden als Alternative zur Messung von 3D-Abständen, 3D-Merkmale auf Basis der *Koordinaten* der Landmarken in diesem Bezugssystem vorgeschlagen. Basierend auf der Beobachtung, dass die meisten Bewegungen im Gesicht entweder primär horizontal oder primär vertikal erfolgen, werden hier für die Erfassung verschiedener Bewegungen die entsprechenden horizontalen oder vertikalen Koordinaten ausgewählt. Die Definition des konkreten, 17-dimensionalen Merkmalsvektors kann Abb. 3.3d entnommen werden, aufbauend auf der Landmarkendefinition in Abb. 3.3a. Mit den koordinatenbasierten Merkmalen lässt sich die Mimik differenzierter erfassen, als mit den zuvor beschriebenen 3D-Abständen. So ermöglichen beispielsweise die Merkmale e_1 und e_2 bzw. e_4 und e_5 die unabhängige Erfassung des Hebens der inneren und äußeren Augenbrauen (AU 1 und 2). Die Merkmale e_6 bis e_9 messen separat die Bewegungen des unteren und des oberen Augenlids (AU 5-7 und 41-46). Das Heben der Mundwinkel (AU 12) kann vom Heben der Oberlippe (AU 10) gut differenziert werden, mithilfe von e_{14} bis e_{16}. Die Merkmale e_{10} bis e_{15} erfassen asymmetrische Bewegungen des Mundes.

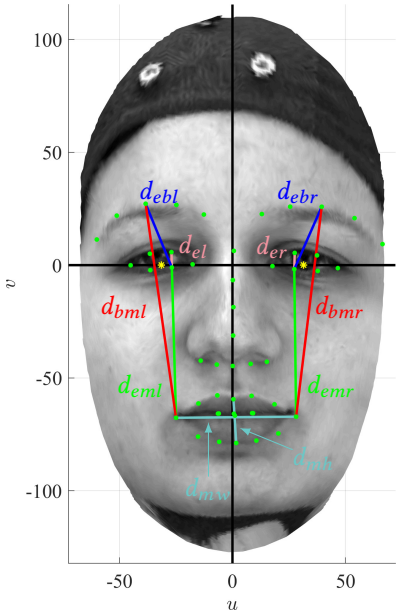
3.2.4. Lernverfahren

Im Folgenden werden Lernverfahren und Prädiktionsmodelle aus dem Stand der Technik für die Evaluierung in der Schmerz- und Mimikerkennung ausgewählt, diskutiert und adaptiert bzw. weiterentwickelt. Dabei spielen die Herausforderungen des maschinellen Lernens, insbesondere die begrenzte Verfügbarkeit von Schmerz- und Mimikdaten und die Eignung des Modells für die Erkennungsaufgabe eine wesentliche Rolle.

Regression: Die Schmerzintensitäten der Datensätze BioVid und X-ITE, die AU-Intensitäten und auch das PSPI-Maß liegen in Klassen vor, was die Verwendung von Klassifikationsmethoden nahe legt. Die Klassen sind jedoch geordnet, werden durch Ganzzahlen beschrieben und repräsentieren verschiedene Stufen eines in Wahrheit kontinuierlichen Phänomens, denn für zwei beliebige Schmerz- oder Mimikintensitäten ließe sich immer noch eine Zwischenstufe finden.



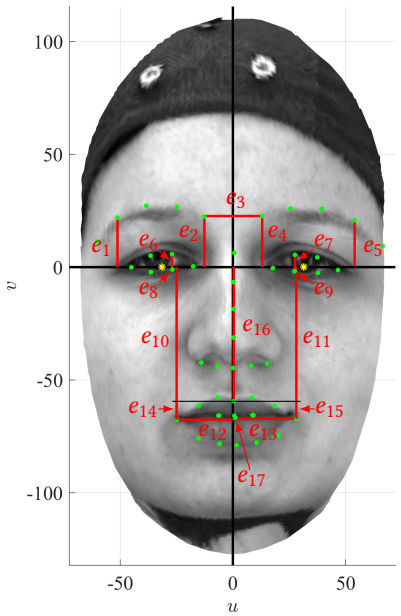
(a) Nummerierung der Landmarken für die Definition der 3D-Merkmale.



(b) Veranschaulichung der 3D-Abstände d. Formelle Definition unter (c).

Distanz	**Merkmalsdefinition**	
Augenbrauen / Mund	$d_{ebl} = \|\mathbf{p}_3 - \mathbf{p}_{32}\|$,	$d_{ebr} = \|\mathbf{p}_8 - \mathbf{p}_{38}\|$
Augen / Mund	$d_{eml} = \|\mathbf{p}_{24} - \mathbf{p}_{32}\|$,	$d_{emr} = \|\mathbf{p}_{31} - \mathbf{p}_{38}\|$
Augen / Augenbrauen	$d_{ebl} = \|\mathbf{p}_{24} - \mathbf{p}_3\|$,	$d_{ebr} = \|\mathbf{p}_{31} - \mathbf{p}_8\|$
Augenlider oben / unten	$d_{el} = \|\mathbf{p}_{24} - \mathbf{p}_{22}\|$,	$d_{er} = \|\mathbf{p}_{31} - \mathbf{p}_{27}\|$
Mundwinkel links / rechts	$d_{mw} = \|\mathbf{p}_{32} - \mathbf{p}_{38}\|$	
Lippe oben / unten	$d_{el} = \|\mathbf{p}_{35} - \mathbf{p}_{41}\|$	

(c) Definition der 3D-Abstandsmerkmale d anhand der 3D-Landmarkenpunkte $\mathbf{p_j} = (u_j, v_j, w_j)$.



Anatomische Basis	**Merkmalsdefinition**	
Augenbraue	$e_1 = v_1$,	$e_2 = v_5$,
	$e_3 = u_6 - u_5$,	
	$e_4 = v_6$,	$e_5 = v_9$
Oberes Augenlid	$e_6 = v_{22}$,	$e_7 = v_{27}$
Unteres Augenlid	$e_8 = v_{24}$,	$e_9 = v_{31}$
Mundwinkel	$e_{10} = v_{32}$,	$e_{11} = v_{38}$,
	$e_{12} = u_{32}$,	$e_{13} = u_{38}$
Lippen	$e_{14} = v_{35} - v_{32}$,	
	$e_{15} = v_{35} - v_{38}$,	
	$e_{16} = v_{35}$,	
	$e_{17} = v_{45} - v_{48}$,	

(d) Definition der 3D-Koordinatenmerkmale e_i anhand der Koordinaten (u_j, v_j) der Landmarken j.

Abbildung 3.3.: Definition der 3D-Merkmale unter Nutzung der auf das 3D-Modell projizierten Landmarken im lokalen Gesichtskoordinatensystem (u, v, w).

Die Ordnung der Klassen liefert zusätzliche Informationen, die beim Training von Regressionsmodellen ausgenutzt werden können. Während die Loss-Funktion der Klassifikation lediglich beachtet, ob die prädizierte und tatsächliche Intensitätsklasse übereinstimmen, nutzt die Regression die zusätzliche Information, wie sehr die prädizierte und tatsächliche Intensität voneinander abweichen. Davon lässt sich die Hypothese ableiten, dass die Regression der Klassifikation von Intensitäten überlegen ist, wenn mehr als zwei Intensitätsklassen betrachtet werden.

Ein weiterer Vorteil ist, dass die Regression auch Werte zwischen den Stufen der Grundwahrheit (Fließkommazahlen) ausgibt und somit die Verläufe der Intensität über die Zeit „glatter“ sind, während die Ausgabewerte der Klassifikation zwischen den Ganzzahlen springen. Die Regression liefert somit detailliertere Informationen als die Klassifikation. Aufgrund der Granularität der Grundwahrheit lässt sich dieser Vorteil jedoch für Einzelbilder nicht quantitativ messen. Wenn die Abfolge von Einzelbildern (Videos) betrachtet wird (siehe Kapitel 4), werden die Vorteile messbar, denn die genauer aufgelösten Intensitätszeitreihen der Regression erlauben die Extraktion und Ausnutzung von Dynamikinformationen für die Erkennung, wie z. B. die Geschwindigkeit oder Beschleunigung einer Bewegung.

Klassische Lernverfahren: Die klassische Vorgehensweise im Computer-Vision-Bereich trennt die Merkmalsextraktion von der Anwendung des maschinellen Lernverfahrens. Nach diesem Schema werden die im vorherigen Abschnitt behandelten Merkmalsarten mit den Lernverfahren Support Vector Machine (SVM, Klassifikation) bzw. Support Vector Regression (SVR) und Random Forest (RF, in den Varianten für Klassifikation und Regression), die bereits in Abschnitt 3.1.1 vorgestellt wurden, kombiniert und später experimentell evaluiert. Vor dem Aufkommen von Deep Learning gehörten SVM/SVR und RF zu den am meisten genutzten maschinellen Lernverfahren und sie wurden für zahlreiche Aufgabenstellungen sehr erfolgreich eingesetzt. Während Deep Learning seine Stärken besonders bei großen Datensätzen ausspielen kann, liegen die Stärken von SVM/SVR und RF eher darin, trotz geringer Datensatzgröße eine gute Generalisierung zu ermöglichen.

Merkmalsfusion: Durch die Kombination von verschiedenen Merkmalsarten werden dem Lernverfahren zusätzliche, oft komplementäre Informationen zugänglich gemacht, was zu Verbesserung der Performance führen kann. Andererseits können zusätzliche, für die Erkennungsaufgabe irrelevante Merkmale den Lernprozess erschweren und die Performance verschlechtern (vgl. „Fluch der Dimensionalität“ in Abschnitt 1.2.2). Um zu erforschen, ob sich die Merkmalsfusion im Kontext der Schmerz- und Mimikerkennung vorteilhaft auswirkt, werden daher verschiedene Merkmalsarten fusioniert und mit SVM/SVR und RF evaluiert. Hierbei wird Merkmalsfusion eingesetzt, d. h. die Merkmalsvektoren werden zu einem höherdimensionalen Vektor verkettet.

Convolutional Neural Networks (CNN): Anders als die klassischen Lernverfahren erfordern CNNs keine explizite Merkmalsextraktion, sondern sie erhalten Bilder des Gesichts als Eingabe und optimieren Filter zur problemspezifischen Merkmalsextraktion. Für die Experimente werden die sehr erfolgreichen CNN-Architekturen NASNet [Zop+18] und MobileNetV3 [How+19] verwendet. Die Arbeit mit den Netzen wurde kurz nach deren Veröffentlichung begonnen, als diese den Stand der Technik ihrer jeweiligen Kategorie neu definiert hatten. NASNet [Zop+18] war eines der ersten CNNs, deren *Architektur* (also Auswahl und Topologie der Schichten und nicht nur deren Gewichte) durch maschinelles Lernen optimiert wurde, mittels Network Architecture Search (NAS). Die Optimierung wurde auf dem Datensatz CIFAR-10 durchgeführt und die gelernte Architektur wurde anschließend auch auf den Datensätzen ImageNet

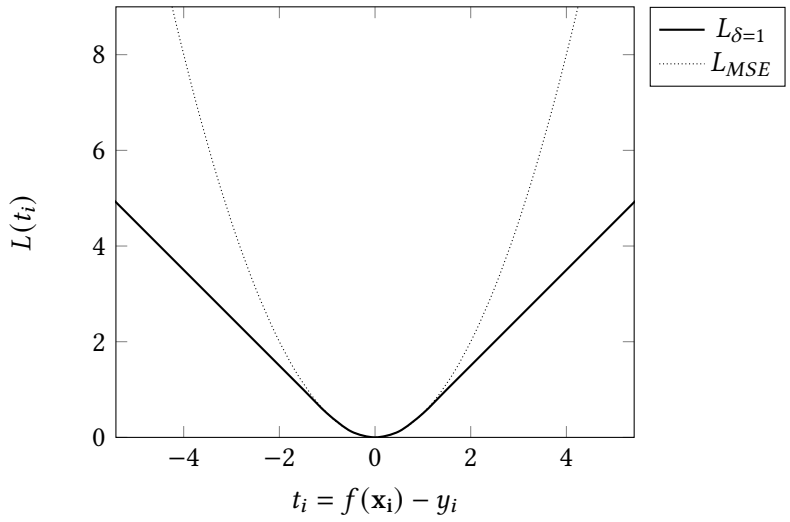


Abbildung 3.4.: Huber-Loss L_δ (3.17) mit $\delta = 1$ (fett) als Funktion der Abweichung zwischen Prädiktion $f(\mathbf{x_i})$ und Zielwert y_i sowie Mean-Squared-Error-Loss L_{MSE} zum Vergleich.

und COCO trainiert und getestet. Bei allen drei Datensätzen wurden die besten zuvor publizierten Ergebnisse deutlich übertroffen. Dies spricht für die Generalisierungsfähigkeit der NASNet-Architektur und lässt den Autor dieser Dissertation auch gute Ergebnisse in der Schmerz- und Mimikerkennung erwarten. MobileNetV3 [How+19] wurde etwa ein Jahr nach NASnet vorgestellt und basiert ebenfalls auf den Ergebnissen einer Netzwerkarchitektursuche. Der Fokus der Entwicklung von MobileNetV3 war jedoch, eine geringe Latenz bei der Ausführung auf mobilen Endgeräten zu erreichen, bei gleichzeitig hoher Erkennungsleistung. Das Ergebnis dieser Bemühungen ist unter anderem MobileNetV3-large, das mit etwas über 200 Millionen Multiply-Accumulate-Operationen (MACs) auf ImageNet eine Top-1-Accuracy von 75,2% erreicht, während die NASNet-mobile-Variante mit mehr als doppelt soviel Operationen (564 Millionen MACs) eine Accuracy von nur 74,0% erreicht. Zusätzlich zur Architektursuche wurden die MobileNet-CNNs jedoch auch manuell weiterentwickelt, z. B. wurde anstelle von ReLU eine andere Aktivierungsfunktion (swish) eingesetzt und für die schnellere Ausführung optimiert sowie die ersten und letzten Schichten manuell überarbeitet, ebenfalls für die schnellere Ausführung. Die Verbesserungen sind nicht nur für mobile Endgeräte von Vorteil, sondern auch zur Reduzierung der Trainingszeit mit GPUs.

Um die Architekturen für die Mimik- und Schmerzerkennung anzuwenden, wird die jeweils letzte Schicht angepasst. Wo für einen Datensatz mehrere Label zur Verfügung stehen, z. B. mehrere Action Units (AU), wird Multi-Label Learning eingesetzt, d. h. alle Label werden über *ein* CNN gelernt und prädiziert. Entsprechend müssen dem CNN zuvor Ausgabeneuronen für alle Label hinzugefügt werden. Bei binärer Klassifikation (zwei Klassen je Label) wird jeweils ein Neuron je Label verwendet, mit der Sigmoid-Funktion aktiviert und mit binärem Cross-Entropy-Loss optimiert. Sind mehr als zwei Klassen je Label vorhanden (z. B. sechs AU-Intensitäten), wird ein Neuron für jede Klasse dieses Labels hinzugefügt und diese gemeinsam mit der Softmax-Funktion aktiviert und mit Cross-Entropy-Loss optimiert. Für Regression wird ein Neuron je Label verwendet (ohne Aktivierung) und typischerweise mit dem Mean-Squared-Error-Loss (MSE) gelernt.

Bei Verwendung von MSE können einzelne Ausreißer, z. B. falsch annotierte Beispiele, durch das Quadrieren des Fehlers einen sehr großen Einfluss auf das Lernen haben. Auch der Beginn des Trainings, wenn die Prädiktionen noch sehr stark von den Grundwahrheiten abweichen, ist

problematisch. Um eine Divergenz durch zu große Gewichtsänderungen zu vermeiden, muss entweder eine sehr kleine Lernrate gewählt werden (die im späteren Verlauf das Lernen verlangsamt) oder es muss sichergestellt werden, dass die Gradienten nicht zu groß werden. Für die Experimente wurde die letztere Idee anhand von Gradient Clipping [GBC16b, S. 415ff.] umgesetzt, bei dem der Betrag des Gradienten (die „Länge“) berechnet und auf einen Maximalwert begrenzt wird, ohne die Richtung des Gradienten zu ändern. Die Berechnung der Gradientenlänge ist jedoch ein zusätzlicher Aufwand, der das Training verlangsamt. Daher wurde in später durchgeführten Experimenten das Problem zu großer Gradienten umgangen, indem anstelle des MSE der *Huber-Loss* [Hub64, S. 75] für robuste Regression verwendet wurde:

$$L_\delta(t_i) = \begin{cases} \frac{1}{2}t_i^2 & \text{für } |t_i| \leq \delta, \\ \delta|t_i| - \frac{1}{2}\delta^2 & \text{sonst,} \end{cases} \tag{3.17}$$

mit $t_i = f(\mathbf{x_i}) - y_i$. Der Huber-Loss (3.17) verhält sich quadratisch für kleine Abweichungen zwischen Prädiktion und Zielwert und linear für große Abweichungen, vgl. Abb. 3.4. Der Parameter δ bestimmt den Übergang zwischen quadratischem und linearen Verhalten. Da der Betrag des Gradienten für $|t_i| > \delta$ konstant ist und nicht von der Größe der Abweichung abhängt, ist zu Beginn des Training mit dem Huber-Loss weder Gradient Clipping noch eine sehr geringe Lernrate nötig, um Probleme durch zu große Gradienten (engl. exploding gradients) zu vermeiden.

Die Stärken von CNNs kommen insbesondere bei großen Datensätzen zum Tragen, mit denen sie nahezu beliebig komplexe nichtlineare Zusammenhänge lernen können. Mit Anwendung der richtigen Techniken können jedoch auch mit kleineren Datensätzen sehr gute Ergebnisse erzielt werden, insbesondere indem Überanpassung an die Trainingsdaten verhindert oder zumindest reduziert wird. Zwei weit verbreitete Möglichkeiten, auf die in den folgenden Absätzen eingegangen wird, sind Transferlernen und Multi-Task-Lernen. Eine weitere Möglichkeit, die auch mit den anderen kombiniert werden kann, ist die Reduzierung der Parameteranzahl. Zur Kombination mit Transferlernen kann z. B. ein größerer Teil der CNN-Architektur weggeschnitten werden, wie die hintere Hälfte des Netzes. Lediglich den vorderen Teil des CNN zu nutzen hilft nicht nur gegen Überanpassung, sondern beschleunigt auch das Training. Eine weitere, beliebig kombinierbare Möglichkeit ist der Einsatz von Regularisierungsmethoden (vgl. Abschnitt 3.1.1), wie z. B. Datenaugmentierung, Dropout, Batch Normalization, Weight Decay und frühes Beenden des Trainings. Eine weitere, erst in den späteren Versuchen eingesetzte Regularisierung ist Cutout [DT17], bei dem ein zufälliger rechteckiger Ausschnitt des Eingangsbildes entfernt wird.

Transferlernen mit CNN: Wie bereits in Abschnitt 3.1 erklärt, können in einer Quelldomäne vortrainierte CNN entweder zur Merkmalsextraktion angewendet und mit SVM/SVR oder RF kombiniert werden, oder sie können in der Zieldomäne weiter trainiert werden (feinjustiert, engl. fine-tuning). Beide Möglichkeiten werden in den Experimenten evaluiert. Dafür werden zum einen Netze verwendet, die auf ImageNet zur Objekterkennung vortrainiert wurden. Die Merkmalsextraktion, die diese Netze gelernt haben, ist gut für allgemeine fotorealistische Bilder geeignet, jedoch nicht für die Mimikanalyse von Gesichtern optimiert. Zum anderen wird vorgeschlagen, ein Netz auf einem umfangreichen Mimikdatensatz zu trainieren (Quelldomäne), um eine noch bessere Merkmalsextraktion bzw. Initialisierung für kleinere Schmerz- und Mimikdatensätze (Zieldomäne) zu erhalten.

Das Transferlernen wird mit den Architekturen NASNet und MobileNetV3 realisiert, für die im Internet Parameter zur Verfügung stehen, die durch Training mit ImageNet gelernt wurden. Zur Feinjustierung in der Zieldomäne wird die letzte Schicht des Netzwerkes abgeschnitten und durch eine zufällig initialisierte Schicht mit der benötigten Anzahl von Ausgabeneuronen ersetzt.

Anschließend wird das Netz für eine zuvor festgelegte (empirisch bestimmte) Anzahl von Iterationen trainiert.

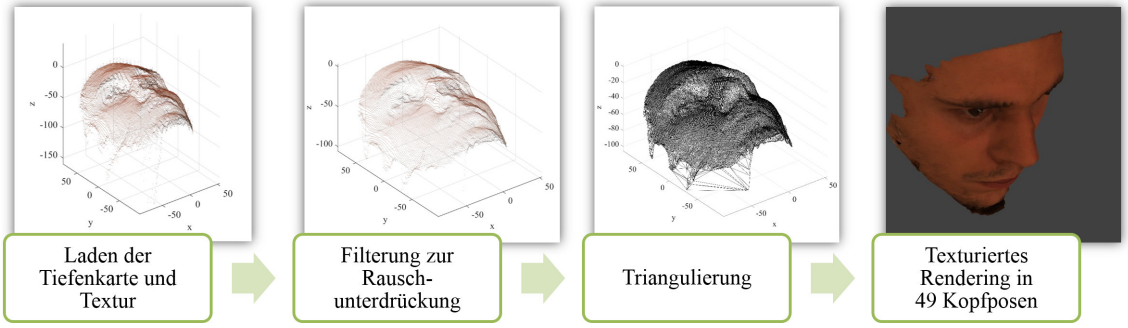
Für die zweite Variante des Transferlernens wird MobileNetV3, ebenfalls ausgehend von ImageNet-Gewichten, mit dem umfangreichen Mimikdatensatz Bosphorus3D trainiert, der später detailliert vorgestellt wird. Bei diesem Training ist die Objekterkennung auf ImageNet die Quelldomäne und die AU-Intensitätsschätzung auf Bosphorus3D die Zieldomäne. Das resultierende Netz wird anschließend weiter feinjustiert anhand verschiedener Datensätze und Erkennungsaufgaben (Zieldomäne), wobei Bosphorus3D die Quelldomäne darstellt.

Gleichzeitiges Lernen mehrerer Aufgaben (engl. Multi-Task Learning) mit CNN: Es wurde bereits über das Lernen mehrerer Label (engl. Multi-Label) mit einem CNN gesprochen, z. B. mehrerer AUs und der PSPI-Schmerzeinschätzung. Auch wenn das bereits als gleichzeitiges Lernen mehrerer Aufgaben aufgefasst werden kann, ist im Kontext dieser Arbeit der Begriff Multi-Task für etwas weiter gehendes reserviert: für das gemeinsame Training mit mehreren Datensätzen, die verschiedene Label haben. Im Idealfall ergänzen sich die Datensätze in einer Weise, dass die Merkmalsextraktion und die Generalisierungsfähigkeit verbessert werden können. Außerdem kann durch das Training mit mehreren Datensätzen die Anzahl und Varianz der Trainingsdaten erhöht werden, wodurch im Allgemeinen Überanpassung entgegengewirkt werden kann.

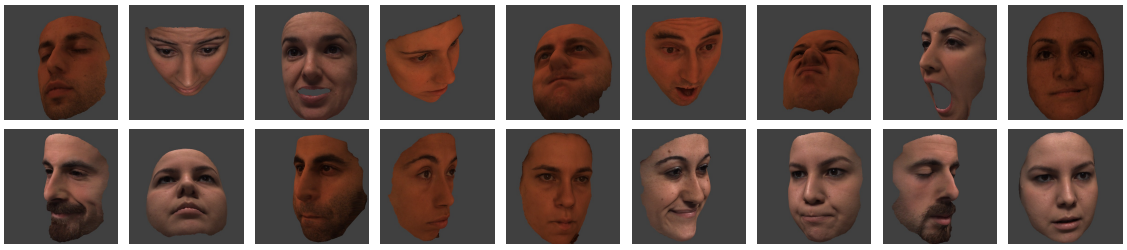
Die Evaluierung des Multi-Task-Lernen erfolgt in den Experimenten ebenfalls mit dem umfangreichen und variantenreichen Datensatz Bosphorus3D. Das gemeinsame Training mit dem anderen Datensatz, der primär von Interesse ist, erfolgt hier durch gleichverteiltes zufälliges Ziehen der Samples beider Datensätze, so dass beide Datensätze etwa gleich viel Einfluss auf das Modell ausüben. Die Label beider Datensätze werden konkateniert, wobei auch gleiche AUs verschiedener Datensätze als getrennte Labels betrachtet werden [WSAH20]. Für jedes Bild eines Datensatzes fehlen somit die Label des jeweils anderen Datensatzes. Dies wird behandelt, indem für unbekannte Label kein Loss und keine Gradienten berechnet werden, genauer gesagt, werden Loss und Gradienten für diese mittels Sample- und Label-spezifischer Gewichte auf null gesetzt.

Erzeugung des Datensatzes Bosphorus3D: Bosphorus3D wurde auf Basis des Datensatzes Bosphorus gerendert, um einen umfangreichen und variantenreichen Datensatz für das Transferlernen und Multi-Task-Lernen zu schaffen. Bosphorus eignet sich als Grundlage, weil er eine große Anzahl von Personen und Gesichtsausdrücken umfasst und mit vielen Label (Action-Unit-Intensitäten) annotiert ist. Die annotierten Bilder sind zwar ausschließlich frontal und ihre absolute Anzahl (2.902) ist vergleichsweise klein, jedoch liegen zu den Bildern auch Tiefenkarten vor. Dies ermöglicht das Rendering des Datensatzes aus verschiedenen Blickrichtungen, wodurch eine große Anzahl von Bildern und große Kopfposevarianz erreicht werden kann. Wenn auch in Bosphorus keine Schmerzmimik vorkommt, ist doch eine deutlich größere Vielfalt an anderen Gesichtsausdrücken abgedeckt als beispielsweise in mit FaceGen möglich (vgl. Datensatz SyLaFaN in Abschnitt 3.2.2). Diese Vielfalt geht weit die über die Basisemotionen hinaus und deckt auch die einzelne Ausführung von Action Units ab, deren Kombinationen typische Schmerzmimik ergeben. Bosphorus3D hat im Vergleich zum ebenfalls sehr umfangreichen Datensatz FERA 2017 weit mehr mit Intensitäten annotierte Action Units, mehr Personen, mehr Ansichten und weniger zeitliche Korrelation (FERA 2017 besteht aus Videos, Bosphorus aus unabhängigen Einzelbildern).

Anders als bei FaceGen, BP4D und BP4D+ (letztere sind die Grundlage für FERA 2017) ist die 3D-Geometrie bei Bosphorus nicht als Dreiecksnetz gegeben, was jedoch für das hochqualitative



(a) Verfahrensweise zur Generierung von Bosphorus3D. Die Schritte wurden für jedes der 2.902 annotierten Samples von Bosphorus durchgeführt. Es resultieren 142.198 Samples in Bosphorus3D.



(b) Beispielbilder aus Bosphorus3D (zufällig gewählt).

Abbildung 3.5.: Erzeugung und Beispielbilder des Datensatzes Bosphorus3D.

Rendering wichtig ist. Daher müssen die Dreiecksnetze zunächst aus den Tiefenbildern generiert werden, bevor sie für das Rendering aus verschiedenen Blickrichtungen genutzt werden können. Abb. 3.5a gibt einen Überblick über die Schritte bis zum Rendering, die für jedes der Samples im Bosphorus-Datensatz angewendet wurden. Nach dem Laden der Tiefenkarten fällt auf, dass die Tiefenwerte verrauscht sind. Störend für die Triangulierung in Dreiecksnetze sind insbesondere Ausreißer, die weit von der eigentlichen Gesichtsoberfläche entfernt sind und sich vor allem an den Rändern des Gesichtes befinden. Daher werden die Tiefenwerte in drei Schritten gefiltert. Zunächst werden Mittelwert und Standardabweichung aller z-Werte bestimmt und die Werte standardisiert, d. h. jeweils der Mittelwert abgezogen und durch die Standardabweichung geteilt. Der Betrag der standardisierten z-Werte wird anschließend mit einem Schwellwert verglichen (hier 4). Alle Punkte, für die dieser Schwellwert überschritten wird, sind extreme Ausreißer und werden entfernt. Dann wird für jeden Punkt $\mathbf{p}$ die Nachbarschaft der 900 nächstgelegenen Punkte betrachtet und von den z-Koordinaten der Nachbarschaft der Median und die mittlere absolute Abweichung vom Median (MAD) bestimmt. Von der z-Koordinate jedes Punkts $\mathbf{p}$ wird nun der Median der Nachbarschaft abgezogen und durch den MAD geteilt. Man erhält für jeden Punkt $\mathbf{p}$ eine normierte Bewertung die angibt, inwieweit der Tiefenwert des Punktes über, in oder unter der Verteilung der Nachbarschaft liegt. Wenn die normierte Bewertung größer oder gleich 5 ist, der Punkt also deutlich über der Oberfläche schwebt, wird der Punkt entfernt. Im nächsten Filterschritt werden Randpunkte entfernt, deren Tiefe meist fehlerhaft gemessen wird. Hierfür werden für jeden Punkt $\mathbf{p}$ die 25 nächstgelegenen Punkte betrachtet, jeweils der Median der Koordinaten bestimmt und gemessen, wie weit $\mathbf{p}$ vom den Mediankoordinaten der Nachbarschaft entfernt ist. Wird eine Schwelle (hier 0,5) überschritten, wird $\mathbf{p}$ entfernt. Die Schwellwerte und Nachbarschaften wurden in Voruntersuchungen empirisch bestimmt. Das Resultat der Filterung hat deutlich weniger Ausreißer und behält gleichzeitig alle wesentlichen geometrischen Strukturen bei, z. B. den Mundinnenraum, wenn dieser in der Tiefenkarte erfasst war. Es folgt das Rendering eines

jeden Samples in 49 verschiedenen Kopfposen, die gleichmäßig in Schritten von 15° über einen Wertebereich von −45° bis +45° (für pitch und yaw) verteilt sind. Der Rollwinkel wird dabei jeweils so gewählt, dass die vertikale Symmetrieachse der Gesichter etwa aufrecht ist. Abb. 3.5b zeigt einige Bilder der resultierenden Datenbank, die insgesamt 142.198 Samples umfasst.

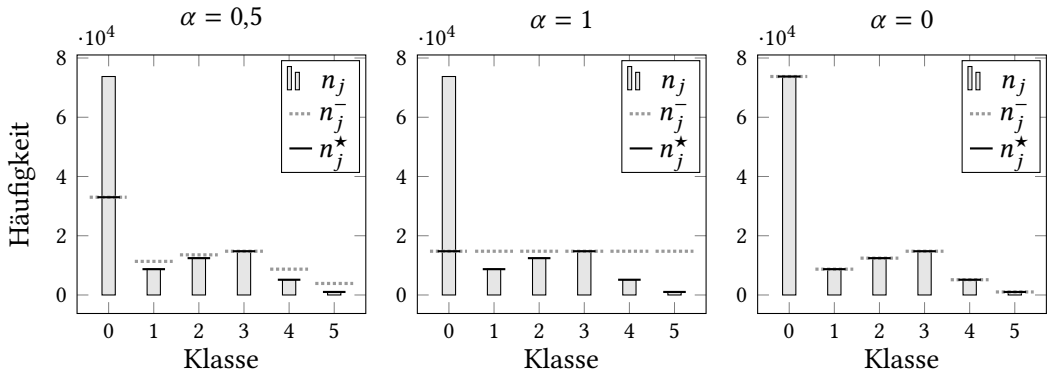
CNN-basierte Fusion mit menschengemachten Merkmalen: Wie bei der Merkmalsfusion mit klassischen Lernverfahren, kann auch die Performance von CNNs durch die Integration komplementärer Informationen verbessert werden, sie muss es jedoch nicht in jedem Fall. Zu dieser Fragestellung wird eine experimentelle Evaluierung mit MobileNetV3 durchgeführt, bei der neben den Bildern auch 2D-Landmarken, 3D-Koordinaten, Kopfpose bzw. Kombinationen davon als zusätzliche Merkmale in das CNN eingespeist werden. Hierfür wird das CNN um einen parallelen Zweig mit einem Fully Connected Layer ergänzt, der die zusätzlichen Merkmale auf einen 50-dimensionalen Vektor abbildet und mit der in MobilNetV3 üblichen Funktion aktiviert (Variante von swish). Abgesehen von der Anbindung dieses Zweiges durch Konkatenation wird das CNN unverändert gelassen. Auf der Seite des CNNs findet sich hinter der letzten Convolution ein Pooling, das einen 960-dimensionalen Vektor erzeugt. Dieser wird für die Fusion mit dem 50-dimensionalen Vektor verkettet und als Eingabe für die folgende Schicht (Fully Connected) genutzt, die wie in der Standardversion von MobileNetV3 einen 1280-dimensionalen Merkmalsvektor ausgibt, der wiederum wie in allen Trainings mittels Fully Connected Layer auf die Ausgabeneuronen abgebildet wird.

3.2.5. Ungleichverteilung der Klassenzugehörigkeit[3]

Mimik ist nicht gleichverteilt. Insbesondere die Abwesenheit von Mimik oder AUs kommt deutlich häufiger vor, da es sich hierbei um den Entspannungs- und damit Normalzustand der Gesichtsmuskulatur handelt. Insofern sind Mimikdatensätze bezüglich ihrer Klassen ungleich verteilt, wenn dem nicht (wie z. B. bei BioVid-S) durch eine vorherige Auswahl der Daten entgegengewirkt wird. Abb. 3.6 zeigt eine exemplarische Verteilung von AU-Intensitäten. Bei derartigen Trainingsdaten lernen viele Standardverfahren suboptimale Modelle, die ein Bias Richtung der häufig vorkommenden Klasse(n) aufweisen, was bei der oder den selten vorkommenden Klasse(n) zu hohen Falschklassifikationsraten führt [Lop+13; HG09].

Eine häufige Vorgehensweise in Zweiklassenproblemen ist die zufällige Unterabtastung (engl. random under-sampling) der Mehrheitsklasse, wobei von dieser eine zufällige Teilmenge gezogen wird (ohne Zurücklegen), so dass für beide Klassen die gleiche Anzahl an Samples (eine „ausgeglichene“ Verteilung) für das Training genutzt werden. Hierbei kann es jedoch sein, dass der häufigsten Klasse zu wenig Gewicht gegeben wird – sie ist auch in den Testdaten meist die häufigste Klasse – oder durch das Weglassen von Samples dem Training wichtige Informationen verloren gehen. Andererseits führt eine starke Ungleichverteilung zu Performance-Verlusten bei den seltenen Klassen. Zur Handhabung dieses Zielkonflikts wird ein Verfahren vorgeschlagen, bei dem die Reduzierung der Ungleichverteilung über einen Parameter α gesteuert und optimiert werden kann und das für beliebige Mehrklassenprobleme definiert ist. Die Reduzierung der Ungleichverteilung wird im Folgenden als „Dämpfung“ bezeichnet und das Verfahren als MID (Multiclass Imbalance Damping, engl. für Dämpfung der Ungleichverteilung bei Mehrklassenproblemen).

[3] Der Autor dieser Dissertation hat das Thema Ungleichverteilung in der Veröffentlichung Werner et al. [WSAH15] adressiert, auf der weite Teile dieses Abschnitts basieren.



(a) gedämpft **(b) „ausgeglichen“** **(c) Originalverteilung**

Abbildung 3.6.: Handhabung der Ungleichverteilung der Klassenzugehörigkeit mittels MID (Multiclass Imbalance Damping, engl. für Dämpfung der Ungleichverteilung bei Mehrklassenproblemen) anhand von Beispielen. (a) Mit $0 < \alpha < 1$, dämpft MID die Ungleichverteilung. (b) Mit $\alpha = 1$ (maximaler Dämpfung) wird die Verteilung der zwei häufigsten Klassen ausgeglichen. (c) Mit $\alpha = 0$ (keine Dämpfung) wird die Originalverteilung beibehalten. Nach Werner et al. [WSAH15].

Dämpfung der Ungleichverteilung bei Mehrklassenproblemen (MID): Wir betrachten ein einzelnes Label j mit C_j Klassen $c = 1, ..., C_j$, wobei n_{jc} die absolute Häufigkeit der Klasse c im Datensatz ist. Die Anzahl der Samples $n^\star_{jc}$, die aus der Klasse c gezogen werden sollen, berechnet sich wie folgt:

$$n^-_{jc} = \lceil s \cdot (n_{jc})^{1-\alpha} \rceil, \text{mit } s = \frac{n_{jf(k)}}{(n_{jf(k)})^{1-\alpha}}, \tag{3.18}$$

$$n^\star_{jc} = \min\{n_{jc}, n^-_{jc}\}. \tag{3.19}$$

In (3.18) ist $\alpha \in [0, 1]$ der Dämpfungsparameter. Er bestimmt, in welchem Maße die Ungleichverteilung reduziert wird: $\alpha = 0$ steht für die Beibehaltung der Originalverteilung, $\alpha = 1$ für eine „ausgeglichene“ Verteilung, und Werte dazwischen für eine Reduzierung der Ungleichverteilung bis zu einem gewissen Grad. Mit $\alpha > 0$ berechnet der Term $(n_{jc})^{1-\alpha}$ ausgeglichenere Klassenverhältnisse. Anschließend werden diese mit einem gemeinsamen Faktor s multipliziert, der die Gesamtzahl der zu wählenden Samples steuert. Er hängt ab vom Parameter $k \in \{2, ..., C_j\}$ und benutzt die Sortierfunktion $f(k)$, welche die k-t häufigste Klasse zurückgibt. Weil die Mimikerkennung jeweils nur eine klar dominante Mehrheitsklasse hat (das Fehlen von Mimik) wird in dieser Arbeit $k = 2$ genutzt. Somit orientiert sich die Dämpfung an der zweithäufigsten Klasse und es werden nur Samples der häufigsten Klasse verworfen. Abb. 3.6b veranschaulicht die Dämpfung (n^-_{jc}, gestrichelt) für $\alpha = 1$, bei welchem eine ausgeglichene Verteilung angestrebt wird, und Abb. 3.6a für $\alpha = 0{,}5$, welches eine mittlere Dämpfung hervorruft. Bei Abb. 3.6c mit $\alpha = 0$ wird die Verteilung beibehalten. Im nächsten Schritt (3.19) wird $n^\star_{jc}$ berechnet, indem das Minimum der vorhandenen Samples n_{jc} und der angestrebten Samples n^-_{jc} gewählt wird. Hierdurch wird verhindert, dass die Klassen mit wenigen Beispielen überabgetastet werden. Dieser Schritt hat in Versuchen zu einer Verbesserung der Performance geführt, weil sonst sehr wenige Samples, z. B. von Klasse 5 in Abb. 3.6, ein zu großes Gewicht im Training gegeben wird. In der Erstveröffentlichung [WSAH15] wurde das Verfahren allgemeiner (mit einem zusätzlichem hier aber unnötigen Parameter β) beschrieben.

Nutzung von MID bei SVM/SVR: MID berechnet anhand der Originalverteilung und α eine neue Verteilung. Mit SVM/SVR wird eine zufällige Unterabtastung (engl. under-sampling) klassenspezifisch angewendet um die neue Verteilung bei den Trainingsdaten herzustellen. Anschließend wird eine stratifizierte Zufallsstichprobe von 3.000 Samples genommen, bei der die Klassenverhältnisse erhalten bleiben. Eine Reduzierung der Sample-Anzahl n war für das Training der SVM/SVR unumgänglich, da die Laufzeit mit $O(n^3)$ ansteigt [BL07].

SVM/SVR Ensemble: Um die Performance bei Ungleichverteilung weiter zu verbessern, werden mehrere MID-SVM- bzw. MID-SVR-Modelle in einem Ensemble kombiniert. Hierfür werden T unabhängige MID-Stichproben des Trainingsdatensatzes gezogen und damit T verschiedene Modelle trainiert. Das Vorgehen ist ähnlich zu *Bagging (Bootstrap Aggregation)* [Bre96], jedoch erfolgt das Ziehen der Stichproben ohne Zurücklegen, da dies in Voruntersuchungen bessere Ergebnisse geliefert hat. Für die Aggregation der Modellausgaben wird mit einer neuen MID-Stichprobe ein Fusionsmodell gleichen Typs trainiert, das als Eingabevektor die Ausgabewerte der T zuvor trainierten Ensemble-Modelle bekommt.

Nutzung von MID bei Random Forest (RF): Bei RF handelt es sich auch Ensembles, jedoch von speziellen Entscheidungsbäumen. Für die Nutzung mit MID wird das gleiche Prinzip wie beim SVM/SVR-Ensemble angewendet. Jedoch wird zur Aggregation kein zusätzliches Modell trainiert, sondern der Gesamtprädiktionswert wird bestimmt wie bei RF üblich: bei Klassifikation mittels Mehrheitsentscheid und bei Regression mittels Mittelwertbildung.

Nutzung von MID bei Multi-Label CNN: Eine Unterabtastung ist insbesondere sinnvoll bei SVM und SVR, deren Laufzeit mit großen Datenmengen stark ansteigt und die bei geringer Sample-Anzahl gut funktionieren. Bei der Verwendung neuronaler Netze für das gleichzeitige Lernen mehrerer Label und Aufgaben ist die Unterabtastung von Trainingsdaten auf Basis einzelner Label mit dem Nachteil verbunden, dass damit auch die Klassenverteilung der anderen Label beeinflusst wird. Daher wird hier die Idee der Reduzierung der Ungleichverteilung über kostensensitives Lernen realisiert, bei dem die Multi-Label-Loss-Funktion L, die beim Lernen minimiert wird, mit Gewichten versehen wird:

$$L = \sum_{i=1}^{N} \sum_{j=1}^{M} w_j(y_{ij}) \cdot l(y_{ij}, \hat{y}_{ij}), \tag{3.20}$$

wobei y_{ij} das korrekte Label j des Samples i bezeichnet, $\hat{y}_{ij}$ den zugehörigen vom Netz prädizierten Wert und $l(\cdot,\cdot)$ den Loss, welcher aus diesen beiden Werten berechnet wird. N ist die Anzahl der Trainingssamples und M die Anzahl der Label. Der Term $w_j(y_{ij})$ ist eine Gewichtsfunktion, die den Einfluss der am häufigsten vorkommenden Klasse reduziert. Die Funktion ist für jedes Label j individuell anhand der Klassenverteilung definiert und gibt für jedes Sample i anhand seiner Klassenzugehörigkeit $c = y_{ij} \in \{1, ..., C_j\}$ das im Folgenden definierte Gewicht $w^\star_{jc}$ zurück:

$$w^\star_{jc} = \frac{n^\star_{jc}}{n_{jc}} \cdot \lambda_j. \tag{3.21}$$

Der Quotient aus $n^\star_{jc}$ und n_{jc} ist bei $\alpha > 0$ in dieser Arbeit (mit $k = 2$) nur für die häufigste Klasse kleiner als eins und für alle übrigen Klassen gleich eins. Somit reduziert er (wie die Unterabtastung bei der SVM) den Einfluss der häufigsten Klasse auf die Loss-Funktion und das gelernte

Modell. Das Skalar λ_j dient dem Zweck, eine Veränderung der zu erwartenden Gradientenlänge durch die Gewichtung mit $w_j(y_{ij})$ zu kompensieren. Wird die Gewichtsfunktion $w_j(c)$ im Kontext des Trainings als diskrete Zufallsvariable X aufgefasst und ihr Erwartungswert auf eins gesetzt, ergibt sich für λ_j:

$$\mathbb{E}(X) = \sum_{c=1}^{C_j} w_{jc}^{\star} \cdot P(X = w_{jc}^{\star}) = \sum_{c=1}^{C_j} \frac{n_{jc}^{\star}}{n_{jc}} \lambda_j \cdot \frac{n_{jc}}{\sum_{k=1}^{C_j} n_{jk}} = 1 \tag{3.22}$$

$$\lambda_j = \frac{\sum_{c=1}^{C_j} n_{jc}}{\sum_{c=1}^{C_j} n_{jc}^{\star}} \tag{3.23}$$

3.3. Experimente

Im Folgenden werden die vorgeschlagenen Methoden experimentell evaluiert. Hierzu werden zunächst Versuche zum Einfluss der Bildauflösung und der Landmarkendetektion durchgeführt (Abschnitt 3.3.1). Anschließend wird der Einfluss der Kopfpose und verschiedener Verfahren zur Gesichtsnormierung untersucht (Abschnitt 3.3.2). Zum Schluss folgen ausführliche Experimente zum Verglich verschiedener Merkmale und Lernverfahren (Abschnitt 3.3.3).

Vorverarbeitung der Daten: Wenn nicht im Einzelnen anders beschrieben, werden die Daten wie folgt vorverarbeitet: Die Gesichtsdetektion erfolgt mit dem CNN-Modell, das mit der Software-Bibliothek dlib [Kin09] Version 19.2 verfügbar ist, und die Landmarkenlokalisierung mit dem in Abschnitt 3.2.1 beschriebenen selbst trainierten Modell. Die Gesichtsbilder werden für die Nutzung von CNN und die Extraktion von LBP mittels SimStable (vgl. Abschnitt 3.1.4) auf ein Standardformat mit 240×240 Pixeln registriert. Fehlende Werte in den Merkmalsvektoren, z. B. wenn bei 3D-Merkmalen die Projektion einer Landmarke nicht auf das 3D-Modell trifft, werden auf den Mittelwert des Merkmals gesetzt (berechnet auf der Gesamtheit aller Daten).

Lernverfahren: Untersucht werden die Lernverfahren Support Vector Machine und Regression (SVM/SVR), Random Forest (RF) und Convolutional Neural Network (CNN). Die Experimente zur Schmerz- und Mimikerkennung beschränken sich auf die Anwendung linearer SVM und SVR, da Voruntersuchungen zur AU-Intensitätsschätzung [WSAH15] ergeben haben, dass nicht-lineare Kernel keine Verbesserungen bringen, sie jedoch deutlich mehr Rechenzeit für das Training und die Justierung der Hyperparameter erfordern. Voruntersuchungen haben ebenfalls ergeben, dass die Hyperparameter der linearen SVM/SVR nur einen geringen Einfluss haben und die Werte $C = 1$ und $\varepsilon = 0{,}1$ meist zu den optimalen Ergebnissen führen. Daher werden, wenn keine weiteren Angaben gemacht werden, im Folgenden diese Werte verwendet und es wird auf die Optimierung der Hyperparameter verzichtet. Auch beim RF wird wegen des Rechenaufwandes und geringen Einflusses auf die Ergebnisse (in Voruntersuchungen) auf die Optimierung von Hyperparametern verzichtet. Die Anzahl der Merkmale, die für die Partitionierung in Betracht gezogen werden, wird wie weithin üblich auf $F = \lfloor \sqrt{N} \rfloor$ bei Klassifikation mit N Merkmalen und $F = \lfloor \frac{N}{3} \rfloor$ bei Regression gesetzt. Als CNN wird in den Abschnitten 3.3.1 und 3.3.2 NASNet [Zop+18] eingesetzt und in 3.3.3 MobileNetV3 [How+19]. Die Netze werden im Allgemeinen für 50.000 Iterationen trainiert, für die kleineren BioVid-S Datensätze für 10.000 Iterationen.

Evaluierung: Die Datensätze BioVid-S, BioVid-S7, UNBC, BP4D und Bosphorus3D werden jeweils mit 5-facher LSO-Kreuzvalidierung evaluiert, d. h. ohne dass eine Person gleichzeitig in Trainingsdaten und Testdaten vorkommt. Für den Datensatz FERA 2017 wird die von den Datensatzerstellern vorgegebene Aufteilung in Training- und Validierungsdaten genutzt. Die untersuchten Erkennungsaufgaben unterscheiden sich zum Teil zwischen den Datensätzen und ergeben sich aus den jeweils verfügbaren Labeln. Für die Evaluierung wird das ICC-Maß ermittelt (vgl. Abschnitt 3.1.2 und Anhang A). Es wird jeweils der Gesamt-ICC berechnet, indem wie bei der FERA 2017 Challenge vorgeschlagen [Val+17] alle Prädiktionen konkateniert werden (auch über die folds einer Kreuzvalidierung hinweg). Bei Prädiktion von mehreren AUs wird aus Platzgründen oft nur der Mittelwert aller einzelnen AU-ICCs aufgeführt. Für die Signifikanztests werden die ICCs für jeden Probanden berechnet und als Einzelbeobachtungen genutzt. Bei mehreren AUs werden die ICCs je Proband und AU als separate Beobachtung angesehen. Zur Untersuchung der statistischen Signifikanz der Ergebnisse werden rechtsseitige Permutuationstest berechnet.

3.3.1. Einfluss der Auflösung und Landmarkendetektion

Im Folgenden wird explorativ untersucht, inwieweit die Performance der Mimikerkennung von der räumlichen Auflösung des Gesichts beeinflusst wird. In diesem Zusammenhang ist auch die Abhängigkeit vom Landmarkendetektionsverfahren interessant, da zu erwarten ist, dass die Genauigkeit der dieser Verfahren ebenfalls von der Auflösung abhängt. Der Autor dieser Dissertation erwartet, dass niedrigere Auflösungen zu schlechterer Mimikerkennung führen, da hierdurch weniger Messpunkte und somit weniger Detailinformationen zur Extraktion diskriminativer Merkmale zur Verfügung stehen. Andererseits wäre Mimikerkennung mit geringer aufgelösten Gesichtern von Vorteil, da hierdurch z. B. bei Nutzung der gleichen Kamera ein größerer Arbeitsabstand oder ein weitwinkligeres Objektiv verwendet werden könnte, um den Bewegungsspielraum der beobachteten Person zu erweitern.

Daten: Die Versuche werden mit zwei Datensätzen durchgeführt. Primär wird der Datensatz BP4D untersucht, der sich durch eine hohe Auflösung und eine große Anzahl von Samples auszeichnet, wobei die Aufgabe der Action-Unit-Intensitätsschätzung betrachtet wird. Zusätzlich wird der BioVid-S Datensatz genutzt, in dem Schmerzmimik deutlich stärker vertreten ist. Untersuchungsgegenstand ist hier das Zweiklassenproblem, zwischen „kein Schmerzstimulus (BLN)“ und „Schmerzstimulus der höchsten Intensität (PA4)“ zu unterscheiden.

Reduzierung der Auflösung: Um auf Basis der hochaufgelösten Bilder eine reduzierte Auflösung zu simulieren, werden alle Bilder des jeweiligen Datensatzes um einen konstanten Faktor herunter skaliert. Dabei werden die neuen Pixelwerte aus dem am nächsten liegenden Pixel des Originalbildes übernommen. Es wird bewusst auf die Nutzung eines Interpolationsverfahren verzichtet, da diese mit einer Glättung des Bildes und damit einer Reduzierung des Bildrauschens einhergehen. Dies ist hier nicht erwünscht, da bei der realen Reduzierung der Gesichtsauflösung durch Vergrößerung des Arbeitsabstandes oder Wechsel des Objektives (d. h. ohne Veränderung des Kamerasensors und der Belichtungszeit) keine Reduzierung des Bildrauschens stattfindet. Die Auflösung der Gesichter wird anhand des mittleren Augenabstandes in Pixeln gemessen. Die konstanten Skalierungsfaktoren zur Reduzierung der Bildauflösung werden anhand des originalen mittleren Augenabstandes (bei BP4D 252 Pixel und bei BioVid-S 170 Pixel) und des gewünschten mittleren Augenabstandes bestimmt. Für die Datenbank BP4D sind die untersuchten Auflösungen 252, 200, 150, 100, 50, 25 und 10 Pixel, bei der Datenbank BioVid-S sind es 170, 150, 100, 50, 25 und 10 Pixel.

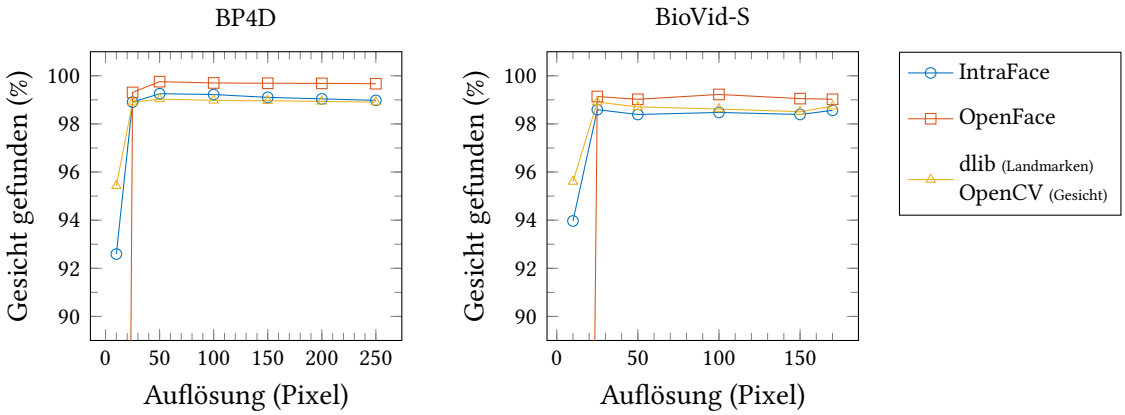
Gesichtsdetektion und Landmarkenlokalisierung: Für jede der Auflösungen werden die Software-Bibliotheken IntraFace [XD13], OpenFace [BRM16] und dlib [Kin09] angewendet, um Landmarken zu lokalisieren (siehe auch Abschnitt 3.2.1). Sie basieren auf unterschiedlichen Gesichtsdetektoren, die unterschiedlich oft angewendet werden. IntraFace nutzt den Detektor von Lienhart et al. [LKP03] aus der Bibliothek OpenCV, jedoch nur, wenn die im vorherigen Videobild detektierten Landmarken nicht erfolgreich verfolgt werden können (engl. tracking). Auch OpenFace wendet den Gesichtsdetektor nicht in jedem Videobild an, sondern verfolgt die Landmarken basierend auf dem vorherigen Zeitschritt. Zur Gesichtsdetektion kommt bei OpenFace ein auf HoG-Features und Support Vector Machine basierendes Verfahren [Kin15] zum Einsatz. Für den dlib-Landmarkendetektor, der keine Landmarkenverfolgung unterstützt, wurde der Detektor von Lienhart et al. [LKP03] aus der Bibliothek OpenCV für jedes Videobild unabhängig angewendet. Bei allen drei Verfahren wird hier davon ausgegangen, dass sich maximal ein Gesicht im Bild befindet und bei mehreren detektierten Gesichtern das Größte zurückgegeben.

Abb. 3.7a vergleicht für die zwei betrachteten Datensätze die verschiedenen Auflösungen und die Verfahren zur Gesichts- und Landmarkenlokalisierung, in wie vielen der Bilder ein Gesicht (und damit auch Landmarken) gefunden werden konnten. Für Augenabstände von 25 Pixeln und mehr schwankt die Erkennungsrate nur minimal. Deutlich geringer fällt sie für die niedrigste Auflösung aus (Augenabstand 10 Pixel). Bei dieser Auflösung funktioniert das Tracking der Landmarken (siehe IntraFace und OpenFace) schlechter, als die Anwendung des Detektors auf jedem Bild (dlib / OpenCV). Die Gesichtsdetektion von OpenFace schlägt bei 10 Pixeln komplett fehl (0% bei beiden Datensätzen), bei höheren Auflösungen liefert sie jedoch für einen größeren Anteil der Bilder Ergebnisse als die anderen Detektoren.

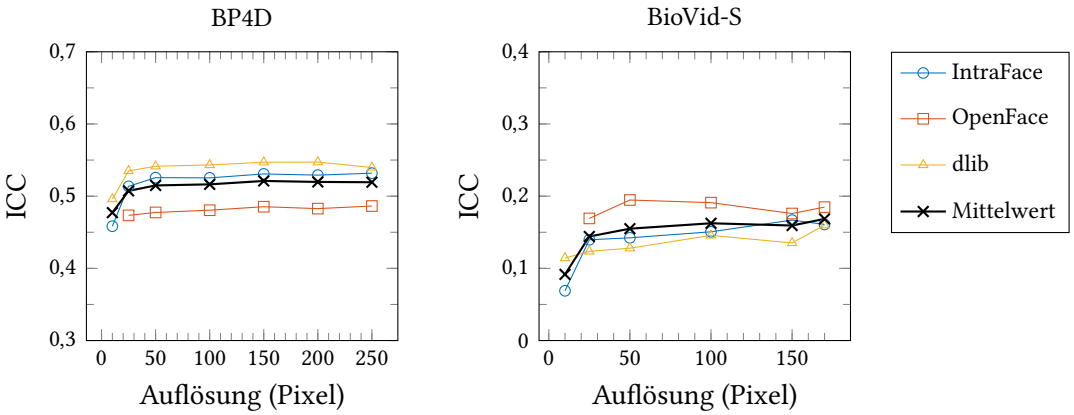
Damit die Ergebnisse der folgenden Versuche vergleichbar bleiben, werden jeweils nur Bilder weiter verwendet, bei denen Gesicht und Landmarken von allen drei Verfahren in allen Auflösungen gefunden worden sind (ausgenommen OpenFace bei 10 Pixeln). Zur Reduzierung der Datenmenge wird bei BP4D außerdem jedes zweite Videobild verworfen, da es durch die zeitliche Korrelation kaum Unterschiede bei aufeinander folgenden Bildern gibt. Bei BP4D reduziert sich somit die Zahl der Bilder von 146.847 auf 66.695 und bei BioVid-S von 3.480 auf 3.264. Mit diesen Daten werden im Folgenden zwei Lernverfahren zur Mimikerkennung evaluiert, indem für jedes Verfahren zur Landmarkenlokalisierung bei jeder Auflösung eine 5-fache LSO-Kreuzvalidierung durchgeführt wird.

Mimikerkennung mit Landmarkenmerkmalen und SVM: Das erste evaluierte Erkennungssystem basiert auf Landmarken (siehe Abschnitt 3.2.3), die als Merkmale für eine lineare SVM bzw. SVR genutzt werden. Es wird das in Abschnitt 3.2.5 beschriebene MID ($\alpha = 0{,}5$) eingesetzt, um bei ungleichen Klassenverteilungen ein Bias zu verhindern (insbesondere bei BP4D). Beim BP4D-Datensatz wird für jede Action Unit (AU) ein lineares SVR-Modell trainiert, für BioVid-S wird für das Zweiklassenproblem eine SVM trainiert. Bei BP4D wird die Zahl der Trainingsbeispiele außerdem auf maximal n =10.000 Samples begrenzt, da die Trainingszeit der SVM mit der Anzahl der Samples stark, mit $O(n^3)$ [BL07], anwächst.

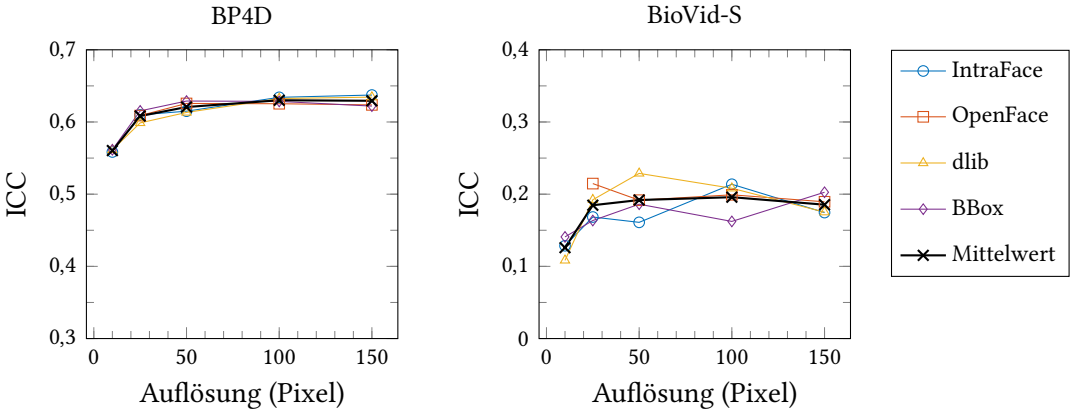
Abb. 3.7b zeigt die Performance der Mimikerkennung. Für BP4D wurden die erreichten Performances der 7 Action Units gemittelt. Da die Gesichts- und Landmarkenlokalisation von OpenFace beim Augenabstand von 10 Pixeln komplett fehlgeschlagen ist, fehlen hierfür die ICC-Werte. Die IntraFace-basierte Erkennung ist sowohl bei BP4D als auch bei BioVid-S bei 10 Pixeln signifikant schlechter als bei allen höheren Auflösungen (jeweils $p < 0{,}001$, Permutationstest). Die Unterschiede in den darüber liegenden Auflösungen sind geringer, jedoch sind 25 Pixel und 50 Pixel bei BP4D signifikant schlechter als Auflösungen von 100 oder mehr Pixeln ($p < 0{,}05$). Der Verlauf der dlib-Ergebnisse ist bei BP4D qualitativ ähnlich zu denen von IntraFace. Sie sind bei 10



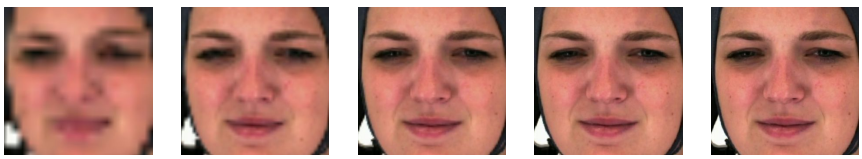
(a) Anteil der Bilder mit erfolgreicher Gesichts- bzw. Landmarkenlokalisation.



(b) Ergebnisse der Mimikerkennung basierend auf Landmarken und SVM.



(c) Ergebnisse der Mimikerkennung basierend auf Gesichtstextur und CNN (NASNet-A Large).



(d) Gesichtsbild mit unterschiedlichem Detailgrad (dlib-Eingabe für CNN) aufgrund unterschiedlichen Augenabstands: 10, 25, 50, 100, 150 Pixel (von links nach rechts).

Abbildung 3.7.: Einzelbildbasierte Mimikerkennung in Abhängigkeit von Auflösung und Landmarkenlokalisierungsverfahren für den Datensatz BP4D (links) und den Datensatz BioVid-S (rechts).

und 25 Pixeln Augenabstand signifikant schlechter als bei höheren Auflösungen ($p < 0{,}001$). Bei dlib unter BioVid-S gibt es keinen deutlichen Abfall der Performance von 25 zu 10 Pixeln. Aufgrund größerer Varianzen sind hier nur die Unterschiede zwischen 10 und 170 sowie zwischen 25 und 170 Pixeln statistisch signifikant ($p < 0{,}05$). Bei OpenFace gibt es keine signifikanten Unterschiede im betrachtbaren Auflösungsbereich.

Vergleicht man die Landmarkendetektoren untereinander, sind die Ergebnisse der beiden Datensätze widersprüchlich: bei BP4D liefert dlib die besten und OpenFace die schlechtesten Ergebnisse, bei BioVid-S ist es umgekehrt. Somit ist keines der drei Verfahren klar zu bevorzugen, da in Abhängigkeit von Daten und Lernaufgabe unterschiedliche Stärken und Schwächen zum Tragen kommen.

Der Mittelwert aller Verfahren (schwarz) zeigt bei beiden Datensätzen, dass die Erkennungsrate robust gegenüber der Verringerung der Auflösung ist und erst unter 25 Pixel Augenabstand eine deutliche Verschlechterung zu erwarten ist. Diese fällt jedoch deutlich geringer aus, als vom Autor dieser Dissertation erwartet, was den hohen Entwicklungsstand der Landmarkendetektoren zeigt, die Koordinaten mit Subpixelgenauigkeit ausgeben können.

Mimikerkennung mit CNN: Als zweiter, texturbasierter Mimikerkennungsansatz wird ein Convolutional Neural Network (CNN) verwendet. Neben dem Einfluss der Auflösung wird auch hier der Einfluss des Landmarkendetektors untersucht, der jedoch nicht direkt zur Merkmalsextraktion, sondern nur für die Registrierung der Eingabebilder (mittels Ähnlichkeitstransformation) genutzt wird. Die Ergebnisse mit landmarkenbasierter Registrierung werden außerdem mit denen verglichen, die ohne zusätzliche Registrierung mit direkter Nutzung des Bildausschnitts erzielbar sind, der vom Gesichtsdetektors [LKP03] ausgegebenen wurde (**BBox**, für engl. bounding box). Die Untersuchungen beschränken sich auf Augenabstände ≤ 150 Pixel, da größere Auflösungen aufgrund der zu hohen Rechenanforderungen zum Training und der schlechten Verfügbarkeit vortrainierter Netze mit solch großer Eingangsbildauflösung unpraktikabel sind. Zur besseren Vergleichbarkeit der Ergebnisse verschiedener Auflösungen wird für alle die selbe Netzarchitektur genutzt: „NASNet-A Large (6 @ 4032)“ [Zop+18] mit einer Eingabebildauflösung von 331×331 Pixeln. Alle Bilder werden auf 340×340 skaliert, das heißt Bilder mit Augenabständen unter 150 Pixeln werden größer skaliert, wobei bilinear interpoliert wird. Das Hochskalieren von zu niedrig aufgelösten Bildausschnitten zur weiteren Verarbeitung ist ein im Stand der Technik ein übliches Verfahren. Hierdurch erhöht sich zwar formal die Auflösung, nicht jedoch der Detailgrad der Bilder gegenüber der zugrunde liegenden niedrigeren Auflösung (vgl. Beispiel in Abb. 3.7d).

Das NASNet-Netz wird nach der zweiten Zelle abgeschnitten[4] und mit *Global Average Pooling Layer* (berechnet Mittelwert je Kanal) und *Fully Connected Layer* vervollständigt (für BioVid-S aktiviert mit der Sigmoid-Funktion). Trainiert wird mit einer Batch-Größe von 16 mit Transferlernen, ausgehend von auf ImageNet vortrainierten Gewichen. Dabei wird der MSE-Loss (BP4D) bzw. Cross-Entropy-Loss (BioVid-S) mittels stochastischem Gradientenabstieg optimiert. Begonnen wird mit einer Lernrate von 0,1 die mit einer cosinusbasierten Vorschrift mit einer Periode [Zop+18] bis 10^{-8} abgesenkt wird. Zur Regularisierung wird die *Drop Path Keep Probability* auf 0.9 und der *L2 Weight Decay* auf $4 \cdot 10^{-5}$ gesetzt. Die Trainingsdaten werden augmentiert, indem

[4] Mit tieferen Netzen lässt sich eine zuverlässigere Prädiktion erreichen, z.B. wurde in einem der vorläufigen Tests (BP4D) mit einem 8-Zellen-Netz (ca. 10 Millionen Parameter, ca. 14 Stunden Training) eine ICC von 0,661 erreicht, wohingegen das letztendlich benutzte 2-Zellen-Netz (ca. 1,1 Millionen Parameter, ca. 5 Stunden Training) ein ICC von 0,639 erreichte. Hier steht jedoch nicht das Erreichen der höchsten Erkennungsraten, sondern ein fairer Vergleich im Mittelpunkt, für den 190 Modelle trainiert werden mussten. Auch mit dieser reduzierten Netzgröße (2 Zellen) dauerte das Training aller 190 Modelle insgesamt bereits über 800 GPU-Stunden.

zufällig die Helligkeit, der Kontrast und die Sättigung geändert werden und das Bild zufällig horizontal gespiegelt und auf 331×331 zugeschnitten wird (engl. random cropping). Während der ersten 2.000 Iterationen wird *Gradient Clipping* angewendet, das die L2-Norm der Gradientenvektoren auf eine Maximallänge von 5 begrenzt. Hierdurch wird eine Divergenz durch zu große Gradienten, wie sie beim MSE insbesondere am Anfang des Trainings leicht auftreten können, vermieden.

Die Ergebnisse sind in Abb. 3.7c dargestellt. Für BP4D verlaufen die ICC-Kurven für IntraFace und dlib ähnlich: Bei 10 Pixeln sind die Werte signifikant schlechter als bei allen höheren Auflösungen und bei 25 und 50 Pixeln sind sie signifikant schlechter als bei 100 und 150 Pixeln ($p < 0{,}001$). Bei OpenFace ist der Unterschied zwischen 25 Pixeln und höheren Augenabständen signifikant ($p < 0{,}001$), bei BBox zwischen 10 und höheren Augenabständen ($p < 0{,}01$). Bei beiden Registrierungsmethoden (OpenFace und BBox) sind in den durchgeführten Experimenten bei Auflösungen über 50 keine Verbesserung zu beobachten. Die Ergebnisse mit der kleineren BioVid-S Datenbank haben deutlich mehr Varianz. Der Augenabstand 10 ist bei dlib signifikant schlechter als bei allen höheren Auflösungen ($p < 0{,}01$), bei IntraFace signifikant schlechter als 25, 100 und 150 ($p < 0{,}05$), und bei BBox ist 10 signifikant schlechter als 150 ($p < 0{,}05$). Bei beiden Datensätzen ist bei der CNN-Erkennung (wie auch bei der SVM oben) bei Auflösungen unter 25 Pixeln ein Einbruch der Performance zu beobachten. Die Auflösungen über 10 Pixel liefern im Mittel ähnliche ICC-Werte, wobei sich insbesondere bei BP4D bis 100 Pixel ein leichter Anstieg abzeichnet.

Schlussfolgerungen: Die Ergebnisse zeigen, dass die Reduzierung der Gesichtsauflösung bis 50 Pixel Augenabstand im Vergleich zu höheren Auflösungen nur mit geringen Performance-Einbußen bei der Mimikerkennung verbunden ist. Auflösungen von über 100 Pixeln Augenabstand brachten keine signifikanten Vorteile. Somit sind beispielsweise die Verwendung weitwinkligerer Objektive oder größerer Abstände zwischen Kamera und Gesicht denkbar, wodurch der Bewegungsspielraums der beobachteten Person ohne Einsatz zusätzlicher Kameras vergrößert werden könnte. Alternativ ließen sich die Hardware-Kosten durch den Einsatz niedriger aufgelöster Kameras reduzieren. Die Unterschiede zwischen den Landmarkenlokalsierungsverfahren sind gering bzw. erlauben keine konkreten Schlüsse. Bei den CNN zeigt sich in diesem Zusammenhang, dass die Registrierung der Gesichter bzw. das Weglassen des Registrierungsschritts (bei BBox) nur einen geringen Einfluss auf die Mimikerkennung mit überwiegend frontalen Gesichtern hat, wie sie in den hier verwendeten Datensätzen vorkommen. Im Folgenden wird der Einfluss der Registrierung bzw. komplexerer Normierungsverfahren bei größerer Kopfposevarianz untersucht.

3.3.2. Einfluss der Kopfpose und Gesichtsnormierung[5]

In den meisten Vorarbeiten zur Mimikerkennung und im vorherigen Abschnitt wurden Daten verwendet, die vor allem Gesichter in frontaler Ansicht zeigen. Bei natürlicher Interaktion generell und insbesondere auch beim Schmerzmonitoring muss jedoch damit gerechnet werden, dass das Gesicht der beobachteten Person nicht genau zur Kamera gerichtet ist und somit nicht-frontale, aus der Bildebene herausgedrehte Kopfposen vorkommen. Da die meisten Datensätze von frontalen Gesichtsbildern dominiert werden (oder ausschließlich frontale Gesichter enthalten), sind die damit trainierten Modelle für frontale Posen optimiert. Im Folgenden wird untersucht, inwieweit nicht-frontale Kopfposen die Performance der Mimikerkennung beeinträchtigen und inwieweit frontalisierende Gesichtsnormierungsverfahren (wie das in Abschnitt 3.2.2 vorgeschlagene

[5] Dieser Abschnitt basiert in weiten Teilen auf Werner et al. [Wer+19c].

FaNC) helfen können, einen Mangel an Posevarianz in den Trainingsdaten zu kompensieren (also auch mit überwiegend frontalen Bildern in den Trainingsdaten eine möglichst gute Mimikerkennung bei nicht-frontalen Gesichtern zu ermöglichen).

Daten: Die Untersuchung der Mimikerkennung erfolgt mit den Datensätzen FERA 2017 und BioVid-S. FERA 2017 hat den Vorteil, dass zusätzlich zu der frontalen Ansicht für jedes Bild acht weitere nicht frontale Ansichten vorliegen. Außerdem ist durch die FERA Challenge ein direkter Vergleich mit den Ergebnissen anderer Ansätze möglich. BioVid-S wurde gewählt, da hier Schmerzmimik stärker vertreten ist. Zusätzlich zu den primär frontalen Gesichtsbildern der mittleren Kamera von BioVid-S, die in den zuvor beschriebenen Untersuchungen verwendet wurden, werden außerdem die Bilder der linken und rechten Kamera eingesetzt, die das Gesicht im Vergleich zur mittleren Kamera mit einer Gierwinkelverschiebung (engl. yaw) von +45° bzw. −45° zeigen. Die evaluierten Erkennungsaufgaben sind die gleichen wie im vorherigen Abschnitt: Bei FERA 2017 werden sieben Action-Unit-Intensitäten geschätzt und bei BioVid-S wird das Zweiklassenproblem „kein Schmerzstimulus (BLN) versus Schmerzstimulus der höchsten Intensität (PA4)“ untersucht.

Zum qualitativen Vergleich der Ergebnisse der verschiedenen Gesichtsnormierungsverfahren werden Bilder beider Datensätze sowie zusätzlich Bilder des Datensatzes SyLaFaN genutzt. Das vorgeschlagene Verfahren FaNC wurde ausschließlich unter Nutzung des synthetischen Datensatzes SyLaFaN entwickelt und trainiert. Die Ergebnisse mit FERA 2017 und BioVid-S veranschaulichen somit, wie gut das Verfahren auf neuen Daten funktioniert und inwieweit ein Datensatz-Bias vorliegt. Qualitative Ergebnisse von SyLaFaN stammen aus den Testdaten einer LSO-Kreuzvalidierung, das heißt sie zeigen die Generalisierung bezüglich ungesehener Personen des Datensatzes. Der Konferenzbeitrag von Werner et al. [Wer+19c] zeigt weitere Ergebnisse mit den Datensätzen LFW [HLM14] und Multi-PIE [Gro+10], die hier aus Platzgründen und wegen geringer Relevanz für die Schmerzerkennung nicht aufgenommen wurden.

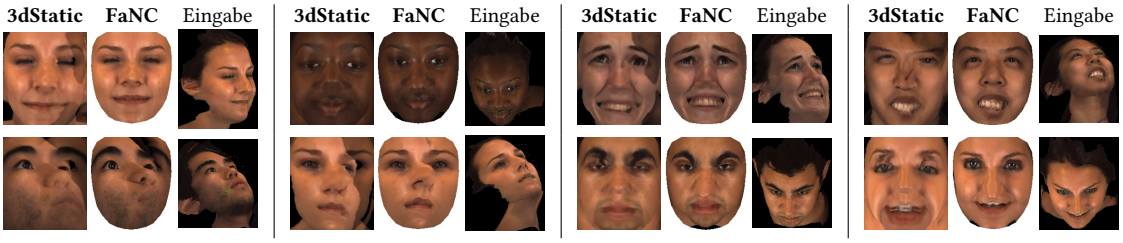
Vorverarbeitung und Training von FaNC: Zur Gesichtsdetektion wird ein mit der Bibliothek dlib [Kin09] bereitgestelltes CNN eingesetzt, da die in Abschnitt 3.3.1 verwendeten Ansätze bei nicht-frontalen Gesichtern oft fehlschlagen und das CNN deutlich zuverlässiger funktioniert. Zur Landmarkenlokalisierung wird ein selbst trainiertes dlib-Modell angewendet, siehe hierzu auch Abschnitt 3.2.1. Die von diesem Modell prädizierten Punkte sind insbesondere bei nicht-frontalen Kopfposen genauer, als bei dem Modell, das bei dlib mitgeliefert wird. Bei extremen Kopfposen und einigen Gesichtsausdrücken treten jedoch noch immer zum Teil große Fehler auf. Um einen negativen Einfluss dieser Fehler auf das FaNC-Gesichtsnormierungsmodell zu vermeiden, wird für das Training dieses Modells nur eine Teilmenge der SyLaFaN-Datenbank verwendet, die Samples mit großem Landmarkenfehler ausschließt. Diese Teilmenge wird bestimmt, indem zunächst für jedes Sample mit erfolgreicher Gesichtsdetektion der mittlere Abstand zwischen den detektierten Landmarkenpunkten und zugehörigen Korrespondenzpunkten (vom 3D Morphable Model) berechnet wird. Anschließend werden die Samples nach diesem mittleren Abstand sortiert und die 75% der Samples mit dem geringsten Fehler in die Teilmenge für das Training aufgenommen. Es ergibt sich ein reduzierter SyLaFaN Datensatz mit 54.175 Bildern. Für diese stehen $M_\mathrm{p} = 153$ Korrespondenzpunkte zur Verfügung, die für das Training von zwei Modellen eingesetzt werden, die sich bezüglich der Anzahl der Eingabelandmarken unterscheiden: Das primär genutzte Modell **FaNC** nutzt alle $M_\mathrm{l} = 68$ verfügbaren Landmarken. Zum qualitativen Vergleich wird in Abbildung 3.8 außerdem das Modell **FaNC51** gezeigt, das nur die 51 inneren Landmarken nutzt (ohne die Gesichtskonturpunkte am Kiefer und Kinn). Die Prädiktion der Koordinaten der

Korrespondenzpunkte wird mit den Parametern $\varepsilon = 0{,}005$ und $C = 0{,}25$ trainiert, die der Sichtbarkeit mit $C = 1$. Für die datensatzübergreifenden Experimente zur Mimikerkennung werden die SyLaFaN-Trainingsdaten augmentiert, indem Daten mit asymmetrischer Mimik gespiegelt werden. Das Warping erfolgt in eine Ausgabeauflösung 256×256 Pixeln (wie auch für die anderen Normierungsverfahren). Im Folgenden vergleichen wir das vom Autor dieser Dissertation entwickelte FaNC mit anderen Normierungsverfahren, siehe Abschnitt 3.1.4 und Tabelle 3.1. Die Methode 3dStatic [Has+15] wird mit den inneren 51 Landmarken angewendet, da dies bessere Ergebnisse liefert als mit 68 Landmarken.

Qualitativer Vergleich der Normierungsverfahren: Abb. 3.8 vergleicht FaNC und andere Normierungsverfahren anhand einiger Beispiele. Teil (a) zeigt Testbilder des SyLaFaN-Datensatzes. Die Registrierungsmethoden in den Spalten SimEye bis AffStable transformieren das Bild lediglich mit *einer* Ähnlichkeitsabbildung oder affinen Abbildung (mittels Verschiebung, Rotation, Skalierung und bei affiner Abbildung auch Scherung), um es bezüglich einiger Landmarken möglichst gut mit einem frontalen Referenzgesicht in Übereinstimmung zu bringen. Sie sind nicht in der Lage die Varianz reduzieren, die durch nicht-frontale Kopfposen entsteht, d. h. die Position und das Aussehen von Mund, Augen, Nase usw. unterscheiden sich bei verschiedenen Kopfposen stark, auch bei gleicher Mimik. SimEye ist anfällig für die perspektivische Verkürzung des Augenabstandes bei nicht-frontalen Posen, die wie in der vierten Zeile zu einer starken Veränderung der Skalierung führen kann. SimInner und SimStable liefern ähnliche Ergebnisse, wobei SimInner tendenziell eine bessere Registrierungsgenauigkeit an den Landmarken hat und SimStable tendenziell aufrechtere und zentriertere Gesichter liefert. AffInner hat mehr Potential, Unterschiede in den Gesichtsproportionen zu kompensieren, kann jedoch unrealistisch aussehende Scherungen des Bildes hervorrufen. Der letztere Effekt ist bei AffStable sogar noch stärker ausgeprägt, siehe z. B. Zeile 2 und 4. Die Methode PieceAff nutzt eine Reihe von abschnittsweise definierten affinen Transformationen und erreicht damit eine sehr genaue Registrierung, bei der jedoch der Großteil der durch die Mimik verursachten geometrischen Verformung verloren geht (siehe z. B. vorletzte Zeile). 3dStatic und FaNC erreichen eine genaue Registrierung und erhalten gleichzeitig die Mimikinformation. Bei PieceAff und 3dStatic tauchen bei der frontalisierten Ansicht in der zuvor teilverdeckten Gesichtshälfte starke Artefakte auf, z. B. Verzerrungen oder weiße Flecken. Das Verfahren PieceAff sieht keine Behandlung von Verdeckungen vor; wenige Pixel werden hier oft über einen großen Gesichtsbereich gestreckt und führen zu starken Verzerrungen im Bild, insbesondere bei ungenauer Landmarkenlokalisierung. 3dStatic nutzt ein statisches 3D-Modell; Artefakte kommen hier häufig zustande, wenn die 3D-Oberfläche des beobachteten Gesichts stark von der des statischen 3D-Modells abweicht. FaNC liefert die besten Normierungsergebnisse: die Posevarianz wird massiv reduziert, die Mimikinformation bleibt erhalten und es gibt nur wenig ungewollte Verzerrungen und Artefakte. Qualitativ ist kein Unterschied zu sehen zwischen FaNC51, das nur die 51 inneren Landmarken nutzt, und FaNC, das alle 68 Landmarken nutzt, inklusive Gesichtskontur.

Abb. 3.8 zeigt im Teil (b) einige Beispielbilder des Datensatzes FERA 2017 und im Teil (c) einige Beispielbilder des Datensatzes BioVid-S. Beide Datensätze wurden zur Entwicklung und zum Training von FaNC nicht eingesetzt, so dass die hierauf erzielten Ergebnisse veranschaulichen, wie gut das System auf realistischen neuen Daten funktioniert. Wenn die Landmarken gut lokalisiert werden, kann FaNC in dem meisten Fällen eine qualitativ hochwertige frontale Ansicht synthetisieren, siehe jeweils obere Zeile. Problemfälle sind jeweils in den unteren Zeilen dargestellt: Deutliche Artefakte treten auf, wenn die Landmarken ungenau lokalisiert werden oder komplett falsch liegen, siehe das erste bis dritte Beispiel von links bei FERA im Teil (b) und das erste und zweite Beispiel von links bei BioVid-S im Teil (c). Ein weiteres Problem ist, dass FaNC

(a) Zufällig gewählte Bilder aus den Testdaten der Datenbank SyLaFaN.



(b) Bilder der Datenbank FERA 2017. Die untere Zeile zeigt FaNC-Fehlerfälle, siehe Text.

(c) Bilder der Datenbank BioVid-S. Die untere Zeile zeigt FaNC-Fehlerfälle, siehe Text.

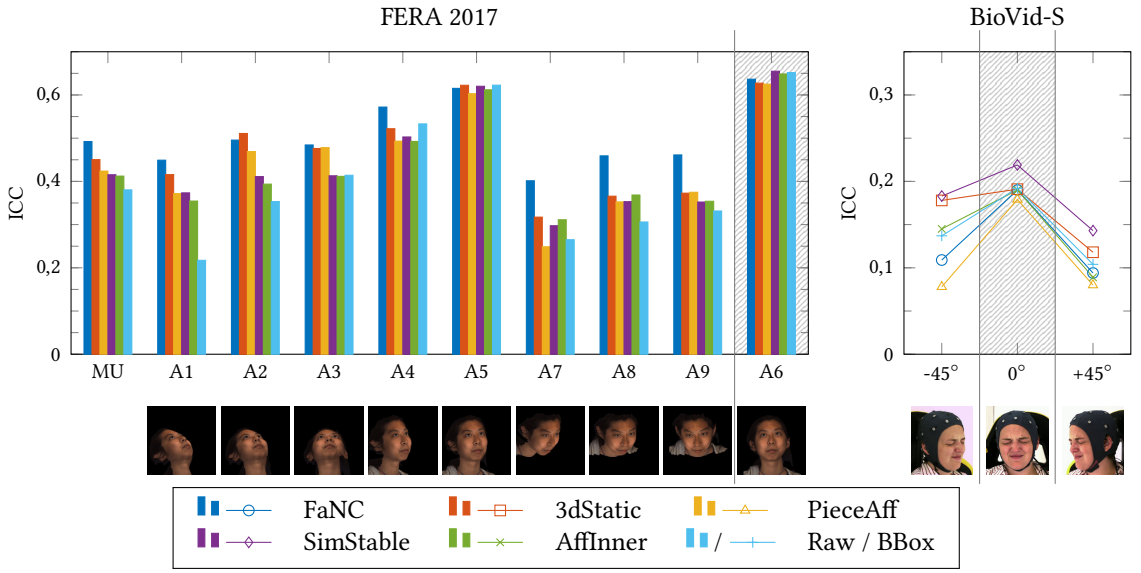
Abbildung 3.8.: Qualitativer Vergleich der Gesichtsnormierungsverfahren: Mit verschiedenen Verfahren normierte Gesichtsbilder (Spaltenüberschriften in fett) und zugehörige Eingabebilder (jeweils letzte Spalte).

Verdeckungen durch die Nase bei großen Nickwinkeln (engl. pitch) nicht behandeln kann und hierbei z. B. eine lange Nase und Deformationen an der Lippe entstehen können. Dies wird in der unteren Zeile von Teil (b), rechts deutlich, anhand der zwei Beispiele mit dem Blickwinkel von oben. Eine weitere Schwäche des Verfahrens FaNC zeigt sich in bei BioVid-S in den zwei Bildern rechts, in denen eine starke Schmerzmimik mit geöffnetem Mund bei nicht frontalen Kopfposen vorkommt. Hier erzeugt das Verfahren trotz recht genauer Landmarkenlokalisierung frontalisierte Bildern mit starken Artefakten, insbesondere in der Mundregion. Dies liegt wahrscheinlich daran, dass derartige Schmerzgesichtsausdrücke nicht in den Trainingsdaten für FaNC vorkommen (siehe auch unten, Abschnitt „Schlussfolgerungen“). Neben den Frontalisierungsergebnissen von FaNC zeigen (b) und (c) auch die mit 3dStatic erzielten Ergebnisse. Wie auch bei dem SyLaFaN-Datensatz treten bei 3dStatic häufiger Artefakte auf als bei FaNC, sogar obwohl verschiedene Anpassungen des 3dStatic-Quellcodes vorgenommen und getestet worden sind, um die Ergebnisse zu verbessern.

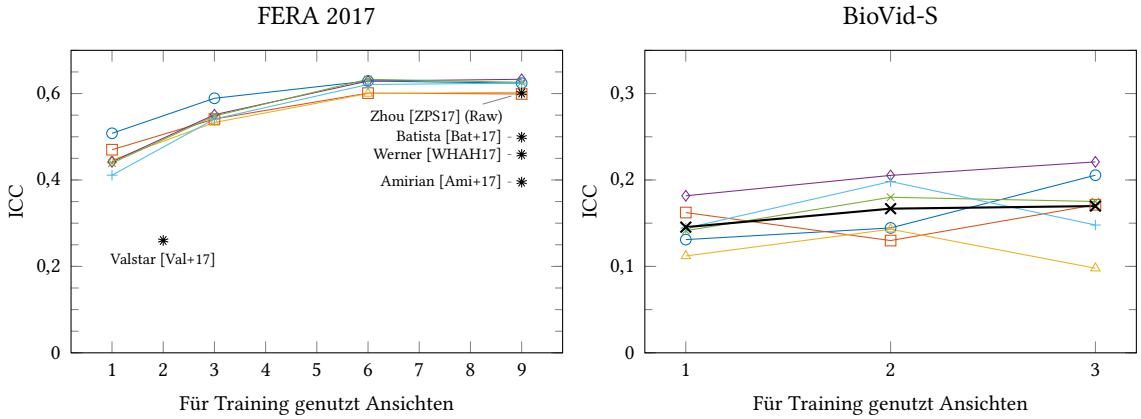
Qualitativ liefert FaNC auf SyLaFaN sehr gut Ergebnisse. Auf FERA 2017 und BioVid-S gibt es viele Bilder mit ähnlich guten Ergebnissen, jedoch ist der Anteil an Bildern mit ungewollten Verzerrungen und Artefakten, die sich negativ auf die Mimikerkennung auswirken könnten, verglichen mit SyLaFaN deutlich höher. Weitere qualitative Ergebnisse zeigen Werner et al. [Wer+19c].

Laufzeit der Normierungsverfahren: Eine geringe Laufzeit war ein wichtiges Entwurfskriterium für das Normierungsverfahren FaNC. Die aufwändigsten Schritte sind zwei Matrixmultiplikationen und das Textur-Warping, das sehr effizient mit jeder (auch einer sehr alten) GPU durchgeführt werden kann. Mit einer unoptimierten OpenGL 2.0 Implementierung dauert das Textur-Warping in ein 256×256 Bild mit einer Intel HD 4000 GPU (integriert in Intel i7-3770, veröffentlicht 2012) etwa 1,5 ms, inklusive Datentransfer zwischen CPU und GPU. Die Laufzeit ist damit ähnlich zu PieceAff und allen Verfahren, die lediglich eine Transformation einsetzen. Das Verfahren 3dStatic [Has+15] benötigt etwa 100 ms, um ein normiertes Bild gleicher Auflösung zu erzeugen. Texturbasierte Methoden des Standes der Technik nutzen tiefe neuronale Netze zur Bildgenerierung, erfordern umfangreiche GPU-Berechnungen und können hohe Bildwiederholraten (wenn überhaupt) nur mit teurer Hardware erreichen.

Mimikerkennung mit CNN: Der Einflusses von Kopfpose und Gesichtsnormierung auf die Mimikerkennung wird mit dem Convolutional Neural Network (CNN) „NASNet-A Mobile (4 @ 1056)“ [Zop+18] untersucht. Hierfür wird das mit dem ImageNet-Datensatz vortrainierte Modell „NASNet-A_Mobile_224“ weiter trainiert, das mit der Tensorflow-Implementierung von NASNet frei verfügbar ist. Aufgrund der begrenzten Variabilität der Trainingsdaten (im Vergleich zu ImageNet), wurde das Netzwerk hinter der sechsten von zwölf Zellen abgeschnitten, um das Training der zahlreichen evaluierten Modelle zu beschleunigen. Anschließend wurde ein linear aktivierter *Fully Connected Layer* mit sieben Neuronen angehängt. Außerdem wurden die Gewichte des sogenannten Stem, des ersten Teils von NASNet, unverändert gelassen um das Training weiter zu beschleunigen. Ähnlich wie Zhou et al. [ZPS17] wurde der Trainingsdatensatz jeder Ansicht von FERA 2017 zufällig unterabgetastet (engl. random under-sampling), wobei für jede der 7 AUs 3.000 Beispiele mit dem Intenstätslabel 0 und 3.000 Beispiele der Label 1 bis 5 gewählt wurden, so dass insgesamt 42.000 Beispiele je Ansicht für das Training genutzt wurden. Bei BioVid-S wurden alle Samples genutzt. Die Datenaugmentierung und das Training erfolgte wie in Abschnitt 3.3.1 beschrieben (lediglich abweichend mit einer Batch-Größe von 32 Samples). Jedes trainierte Modell wurde auf jedem Bild des gesamten Validierungsdatensatzes getestet, um die ICC-Performance je Ansicht und AU zu berechnen. Für jede Gesichtsnormierungsmethode wurden Training und Test fünf mal wiederholt und die Ergebnisse gemittelt.



(a) Training nur mit frontalen Bildern. Test-Performance (ICC) auf den beim Training ungesehenen Ansichten (weißer Hintergrund) und den für das Training genutzten Ansichten (schraffierter Hintergrund). MU = Mittelwert der ungesehenen Ansichten.



(b) Training mit unterschiedlicher Anzahl von Ansichten (x-Achse) und zugehörige Mittelwerte der Test-Performances aller Ansichten (ICC). Legende: siehe (a), ⁕ verwandte Arbeiten, —×— Mittelwert der Verfahren zur Gesichtsnormierung.

Abbildung 3.9.: Einzelbildbasierte Mimikerkennung in Abhängigkeit von Kopfpose und Gesichtsnormierung für den Datensatz FERA 2017 (links) und den Datensatz BioVid-S (rechts). Abb. zu FERA 2017 nach Werner et al. [Wer+19c], © 2019 IEEE.

Für jede Kameraansicht werden die Ergebnisse verglichen, die mit den Verfahren FaNC, 3dStatic, PieceAff, SimStable, AffInner, und Raw bzw. BBox erreicht wurden. Raw und BBox sind die einfachsten Verfahren und nutzen keine komplexe Transformation. Auf dem Datensatz FERA 2017 wird Raw evaluiert, bei dem das gesamte Bild als Eingabe für das CNN verwendet wird (Beispiele siehe Bilder in Abb. 3.9 (a), links). Auf dem Datensatz BioVid-S wird stattdessen BBox evaluiert, bei dem das begrenzende Rechteck (engl. Bounding Box) der Gesichtsdetektion als Eingabe für das CNN ausgeschnitten wird.

Ergebnisse bei Training mit frontalen Bildern: Zunächst wird die Mimikerkennung ausschließlich mit den Bildern der frontalen Kameraansicht trainiert, jedoch auf allen verfügbaren Ansichten getestet, um zu untersuchen, inwieweit die verschiedenen Gesichtsnormierungsverfahren auf zuvor ungesehenen Kopfposen generalisieren. Abb. 3.9 (a) zeigt die Ergebnisse für die Datensätze FERA 2017 (links, ICC-Mittelwerte aller AUs) und BioVid-S (rechts). Die Test-Performance der im Training ungesehenen Ansichten (weißer Hintergrund) ist in allen Fällen schlechter als die der Trainingsansicht (schraffierter Hintergrund). Die Differenz kommt durch die Unterschiede in den Verteilungen der Trainings- und Testdaten zustande und vergrößert sich mit der Verschiedenheit der Bilder, tendenziell insbesondere mit dem Winkelabstand der Kopfposen. Dem versucht die Frontalisierung bei FaNC, 3dStatic und PieceAff entgegenzuwirken, indem die Bildunterschiede zwischen den verschiedenen Kopfposen reduziert werden. Idealerweise sollen sie eliminiert werden, was jedoch mit aktuellen Normierungsverfahren nicht vollständig möglich ist, wie oben bei den qualitativen Betrachtungen veranschaulicht wurde.

Bei FERA 2017 können wir beobachten, dass alle Verfahren gut auf Ansicht 5 (A5) generalisieren, deren Gierwinkel (engl. yaw) sich nur 20° von der Trainingsansicht A6 unterscheidet. Jedoch fällt die Performance bei allen anderen Ansichten deutlich. Im Mittel (MU) und in den meisten Fällen generalisiert die vorgeschlagene Methode FaNC am besten auf im Training ungesehenen Posen. Bei Änderungen im Nickwinkel (±40°, A1-3 und A7-9) beobachten wir die geringste Performance, da sich hier das Erscheinungsbild in einer Weise ändert, die durch kein Normalisierungsverfahren vollständig kompensiert werden kann (z. B. Verdeckungen durch die Nase). Dabei übertrifft jedoch die Performance von FaNC die anderen Methoden in fünf von sechs Fällen, besonders deutlich in den Ansichten von oben (A7-9).

Bei BioVid-S ergibt sich ein anderes Bild: Hier liefert die einfache Transformation SimStable sowohl in der Trainingsansicht (0°), als auch in den ungesehenen Ansichten (±45° Gierwinkel) die besten Ergebnisse, gefolgt von dem Frontalisierungsverfahren 3dStatic. FaNC generalisiert hier schlechter als die nicht frontalisierenden Verfahren SimStable, BBox und AffInner, was darauf hindeutet, dass durch FaNC auf dem Datensatz BioVid-S viele ungewollte Verzerrungen und Artefakte entstehen und dass diese einen negativen Einfluss auf die Mimikerkennung haben, der stärker ist als der positive Einfluss durch die Reduzierung der Varianz zwischen den Posen.

Ergebnisse bei Training mit Bildern verschiedener Ansichten: Im Folgenden wird untersucht, wie sich das Hinzufügen weiterer Ansichten zu den Trainingsdaten auf die Mimikerkennung auswirkt. Dabei wird die Zahl der Trainingsiterationen konstant gehalten, um den Effekt der Varianz in den Trainingsdaten unabhängig von der Dauer des Trainings zu untersuchen. Abb. 3.9b zeigt die Ergebnisse (mittlere Test-ICCs aller Ansichten) in Abhängigkeit von der Anzahl der für das Training genutzten Ansichten. Der linke Teil vergleicht die Ergebnisse, die hier mit NASNet und den verschiedenen Normierungsverfahren erzielt wurden, und die Ergebnisse anderer Arbeiten, die zur AU-Intensitätsschätzung auf FERA 2017 berichtet wurden. Valstar et al. [Val+17] haben als einzige andere Autoren nur eine Teilmenge der Trainingsdaten genutzt (A5 und 6), um die Generalisierung auf ungesehenen Ansichten zu untersuchen. Ihr einfaches Challenge-Baseline-System liefert jedoch die mit Abstand schlechtesten Ergebnisse. Amirian et al. [Ami+17] und Werner et al. [WHAH17] trainieren beide mit allen Ansichten und übertreffen mit großem Abstand die Baseline. Die tiefen neuronalen Netze, die von Batista et al. [Bat+17] und Zhou et al. [ZPS17] trainiert wurden, schneiden jedoch deutlich besser ab. Die FERA 2017 Challenge-Gewinner Zhou et al. [ZPS17] nutzen die Originalbilder (Normierungsverfahren „Raw“) und verfeinern für jede AU ein separates VGG16-Netzwerk. Das hier genutzte NASNet liefert ähnliche Ergebnisse mit den Verfahren 3dStatic und PieceAff, übertrifft jedoch mit den anderen Normierungsverfahren alle verwandten Arbeiten. Selbst bei gleicher Performance

hat NASNet den Vorteil, mit weniger Modellparametern und Speicher auszukommen: beide anderen Netzwerke [Bat+17; ZPS17] haben 300-mal so viele Parameter wie das genutzte NASNet [Wer+19c]. Neben den Ergebnissen des Trainings mit allen 9 Ansichten zeigt die Grafik auch die Ergebnisse mit einer Ansicht (A6), drei Ansichten (A3, 6, 9) und sechs Ansichten (A1, 3, 4, 6, 7, 9). Das vorgeschlagene Verfahren FaNC, trainiert mit NASNet auf nur den frontalen Bildern, erzielt bessere Ergebnisse als Batista et al. [Bat+17], die auf allen Ansichten trainiert und den zweiten Platz bei der Challenge erreicht haben. Weiterhin beobachten wir, dass sich die Performance zwischen einer und sechs Trainingsansichten in allen Fällen verbessert, nicht aber zwischen sechs und neun Ansichten (wobei die drei Ansichten mit Gierwinkel $-20°$ hinzukommen). Um zu untersuchen, ob die Ergebnisse sich durch längeres Training verbessern, wurde 100.000 statt 50.000 Iterationen trainiert, wobei jedoch kein signifikanter Unterschied festgestellt wurde. Daher schlussfolgert der Autor dieser Dissertation, dass die zusätzlichen Ansichten mit dazwischenliegenden Gierwinkeln keinen signifikanten Einfluss auf die Performance haben und das Modell bereits gut auf dazwischenliegenden Kopfposen generalisiert.

Die mit dem BioVid-S Datensatz erzielten Ergebnisse im rechten Teil von Abb. 3.9 (b) unterscheiden sich wie in (a) deutlich von FERA 2017. Wieder funktioniert die mimikbasierte Erkennung der Schmerzstimuli mit SimStable am besten, wobei die mittlere Test-Performance aller Ansichten durch das hinzunehmen weiterer Trainingsansichten leicht ansteigt. Auch bei FaNC und dem Mittelwert aller Normierungsverfahren zeigt sich ein leichter Anstieg der Performance bei Nutzung von einer ($0°$), zwei ($0°$ und $-45°$) und allen drei Ansichten beim Training. Die Performance mit den anderen Normierungsverfahren hat keinen klaren Trend. Wie auch bei den Ergebnissen in Abb. 3.7 haben die Ergebnisse auf BioVid-S eine deutlich größere Varianz als der Vergleichsdatensatz, wahrscheinlich primär aufgrund der geringeren Datensatzgröße.

Schlussfolgerungen: Nicht-frontale Kopfposen sind für die Mimikerkennung noch immer eine Herausforderung, insbesondere da der überwiegende Teil der verfügbaren Daten frontale Gesichter zeigt. Dies wird deutlich, wenn mit frontalen Gesichtern trainiert und mit nicht-frontalen Gesichtern getestet wird, siehe Abb. 3.9a. Es wurden drei Ansätze untersucht, um die Kopfposeinvarianz zu verbessern: (1) Frontalisierung des Gesichts, insbesondere mit dem Verfahren FaNC, (2) Nutzung mehrerer Ansichten des Gesichts beim Training, und (3) die Kombination von (1) und (2). Die Frontalisierung mit dem Verfahren FaNC zeigt großes Potential bei der Generalisierung über Kopfposen hinweg, d. h. dem Training mit frontalen Gesichtern und dem Test mit nicht-frontalen Gesichtern. Im Vergleich mit anderen Normierungsverfahren des Standes der Technik erreicht FaNC hier die besten Ergebnisse bei der Schätzung von AU-Intensitäten auf dem Datensatz FERA 2017 (siehe oben) sowie bei der Klassifizierung verschiedener Mimiken (Neutral, Lächeln, Überraschung, Ekel, Schrei, zugekniffene Augen) auf dem Multi-PIE Datensatz (siehe [Wer+19c]). Bei der Erkennung der Schmerzstimuli auf dem reinen Schmerzerkennungsdatensatz BioVid-S überzeugt FaNC jedoch nicht, und andere, auch nicht-frontalisierende Verfahren generalisieren besser. Der Grund ist, dass frontalisierende Verfahren komplexe Änderungen am Bild vornehmen, bei denen es zu ungewollten Verzerrungen und anderen Artefakten kommen kann, die die Mimikerkennung negativ beeinträchtigen können. Bei BioVid-S überwiegen derartige negative Effekte gegenüber den positiven, d. h. der beabsichtigten Senkung der Texturvarianz innerhalb der Klassen. Mögliche Ursachen für die ungewollten Verzerrungen und Artefakte sind (1) ungenaue Landmarken, (2) starke Abweichung der 3D-Kopfform von den beim Training vorkommenden Kopfformen, und (3) starke Abweichung der Mimik von den im Trainingsdatensatz vorkommenden Mimiken. Was BioVid-S eindeutig von den Datensätzen FERA 2017 und Multi-PIE unterscheidet, sind die darin vorkommenden Gesichtsausdrücke. Neben neutralen Gesichtern taucht in BioVid-S nahezu ausschließlich Schmerzmimik auf, die im Trainingsdatensatz

für das FaNC-Gesichtsnormierungsmodell nicht vorkommt. Dieses Datensatzbias des Modells bezüglich Mimik lässt sich jedoch im Rahmen dieser Arbeit nicht beheben, da kein 3D-Morphable-Model verfügbar ist, das Schmerzmimik abbildet, und da die verfügbaren Daten aufgrund fehlender 3D-Informationen bzw. mangelhafter Genauigkeit nicht geeignet sind, ein entsprechendes 3D-Morphable-Model zu erstellen. Der Autor dieser Dissertation schlägt jedoch vor, das Verfahren FaNC in weiterführenden Arbeiten zu verbessern, insbesondere indem Schmerzmimik und andere nicht repräsentierte Gesichtsausdrücke sowie eine größere Vielfalt an Identitäten (Personen) in den Trainingsdatensatz integriert werden.

Eine Alternative für die Verbesserung der Kopfposeinvarianz ist das Training mit mehreren Ansichten. Hierfür müssen Kameraaufnahmen des Gesichts aus verschiedenen Richtungen verfügbar sein, oder 3D-Daten, aus denen solche Bilder synthetisiert werden können. Problematisch ist, dass dies bei nur wenigen existierenden Datensätzen der Fall ist und dass die Aufnahme neuer derartiger Datensätze aufwändig und teuer ist. Die Ergebnisse in Abb. 3.9b und Werner et al. [Wer+19c] zeigen jedoch, dass beim Training mit verschiedenen Ansichten (in den meisten Fällen und im Mittel) eine bessere Kopfposeinvarianz erzielt wird als beim Training mit ausschließlich einer frontalen Ansicht. Dies gilt auch für einfache Registrierungsmethoden auf Basis einer einzelnen geometrischen Transformation, wie z. B. SimStable. Die Kombination beider Optionen, d. h. sowohl ein Frontalisierungsverfahren wie FaNC einzusetzen als auch mit mehreren Ansichten zu trainieren, bringt bei hinreichender Abdeckung des Kopfposeraums keinen Vorteil. Bei FERA 2017 ist die Performance mit den Frontalisierungsverfahren (FaNC, 3dStatic, PieceAff) bei sechs oder mehr Trainingsansichten ähnlich oder schlechter als mit einfachen Registrierungsmethoden. Bei BioVid-S führt Frontalisierung selbst beim Training mit nur einer Ansicht zu keinen konsistenten Verbesserungen für die Poseinvarianz; auf diesem Datensatz ist z. B. SimStable verglichen mit FaNC klar überlegen. Diese Ergebnisse deuten darauf hin, dass die Nutzung einer einfachen Registrierungsmethode (ohne potentielle unerwünschte Verzerrungen und Artefakte) gemeinsam mit dem Training auf mehreren Ansichten im aktuellen Stand der Technik die beste Option für die Mimikerkennung und insbesondere für die Schmerzerkennung darstellt. Die Ergebnisse auf FERA 2017 (und auch auf Multi-PIE [Wer+19c]) legen nahe, dass bei der Verwendung von CNNs auch ein Trainingsdatensatz mit einer groben Abdeckung des Kopfposeraumes in Winkelschritten von etwa 40° hinreichend ist, um auch auf dazwischen liegende Posen zu generalisieren [Wer+19c]. Die mit BioVid-S durchführbaren Untersuchungen lassen zu dieser Fragestellung keine Aussage zu, da lediglich drei Ansichten mit zu großen Winkelabständen vorliegen. Auch bei Verwendung von mehreren Ansichten bleibt jedoch die Problematik erhalten, dass für Kopfposen, die im Trainingsdatensatz unterrepräsentierte sind (z. B. große Nickwinkel in BioVid-S), mit einer schlechteren Performance zu rechnen ist als bei stark vertretenen Kopfposen.

3.3.3. Einfluss von Merkmalen und Lernverfahren

Im Folgenden werden verschiedene menschengemachte Merkmale und gelernte Merkmalsextraktoren (CNNs), verschiedene Lernverfahren bzw. Prädiktionsmodelle, sowie Klassifikation und Regression verglichen. Außerdem untersucht werden verschiedene Merkmalsfusionen sowie der Einfluss der Ungleichverteilung der Klassenzugehörigkeiten und der vorgeschlagenen Verfahren zur Handhabung der Ungleichverteilung.

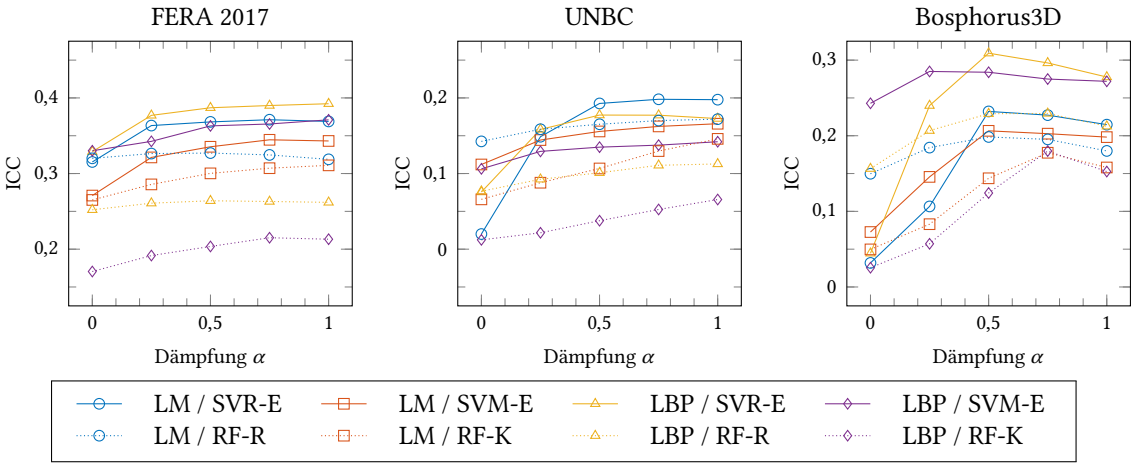
Die Vorverarbeitung, das Lernen und die Evaluierung erfolgen wie oben beschrieben, mit folgenden Abweichungen: Für einen fairen Vergleich von Regression und Klassifikation werden die Prädiktionen der Regressionen diskretisiert, so dass das ICC-Maß immer mit ganzzahligen

Prädiktionswerten berechnet wird [WSAH15]. Die Ungleichverteilung wird mit dem vorgeschlagenen Verfahren MID behandelt, mit $\alpha = 0{,}5$ wenn nicht extra anders angegeben. Als CNN wird hier MobileNetV3 eingesetzt und mit Batch-Größe 16 und konstanter Lernrate 0,01 trainiert. Regularisiert wird das Training mit Dropout 20%, Weight Decay $4 \cdot 10^{-6}$ und Cutout [DT17] mit Kantenlänge bis zu 40% des Eingabebildes, sowie der bereits oben beschriebenen Datenaugmentierung. Bei der Regression wird der Huber Loss (mit $\delta = 1$) minimiert, wodurch kein Gradient Clipping nötig ist (vgl. Abschnitt 3.2.4)

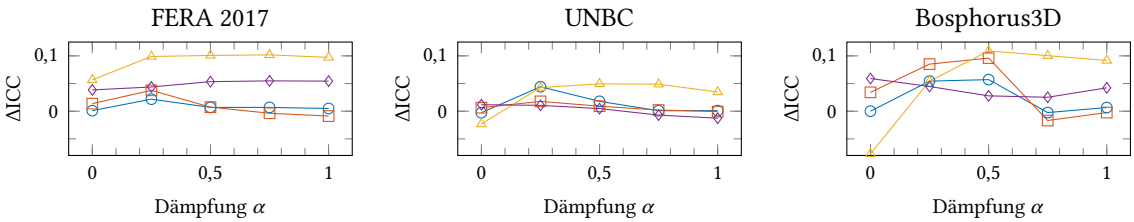
Daten: Die Untersuchung des Einflusses von Merkmalen und Lernverfahren erfolgt mit den Datensätzen FERA 2017, UNBC, BioVid-S, BioVid-S7 und Bosphorus3D. Bei den Datensätzen FERA 2017, UNBC und Bosphorus3D werden die annotierten AU-Intensitäten prädiziert und die Mittelwerte der ICCs aller AUs des jeweiligen Datensatzes betrachtet (bei FERA 7 AUs, bei UNBC 10 AUs, bei Bosphorus3D 26 AUs). AU-spezifische Ergebnisse für einige Modelle sind im Anhang B.1 zu finden. Der UNBC-Datensatz bietet neben den AUs mit dem PSPI-Wert auch eine einzelbildbasierte Schmerzeinschätzung, die ebenfalls prädiziert wird. Diese wird auf 6 Klassen reduziert, da die meisten PSPI-Werte über 5 nur unter 100 Mal vorkommen und einige gar nicht. Daher werden alle PSPI-Werte über 5 auf die Klasse 5 abgebildet. Bei BioVid-S und BioVid-S7 wird das Zweiklassenproblem „kein Schmerzstimulus (BLN) versus Schmerzstimulus der höchsten Intensität (PA4)“ untersucht. FERA 2017, BioVid-S und BioVid-S7 verfügen über mehrere Kameraansichten. Die meisten Experimente werden mit zwei Varianten dieser Datensätze durchgeführt, mit jeweils dem gesamten Datensatz mit allen Ansichten und mit dem Teildatensatz der Kamera, welche die Gesichter überwiegend frontal zeigt. Bei allen Versuchen mit mehreren Ansichten werden die Bilder unterschiedlicher Ansichten als unabhängige Samples aufgefasst, d. h. es findet *keine* Datenfusion über verschiedene Kameras hinweg statt. Die Datensätze unterscheiden sich sehr stark hinsichtlich ihrer Sample-Anzahl. Bei FERA 2017 (alle Ansichten) sind es beispielsweise insgesamt ca. 2 Millionen Bilder, bei Bosphorus3D 142.200, bei UNBC 48.400 und bei BioVid-S7 (frontale Ansicht) nur 280 (weitere Zahlen in Tabelle 3.2).

Handhabung der Ungleichverteilung der Klassenzugehörigkeit: Die Datensätze FERA 2017, UNBC und Bosphorus3D sind mit AU-Intensitäten annotiert und weisen eine Ungleichverteilung der Klassenzugehörigkeit auf. Zur Handhabung dieser Problematik wurde in Abschnitt 3.2.5 das Verfahren MID (Multiclass Imbalance Damping) vorgeschlagen, um die Ungleichverteilung bei Mehrklassenproblemen zu reduzieren (zu „dämpfen“). Über den Parameter α lässt sich einstellen, in welchem Maße die Ungleichverteilung reduziert wird: $\alpha = 0$ steht für die Beibehaltung der Originalverteilung, $\alpha = 1$ für eine „ausgeglichene“ Verteilung (in der die zwei häufigsten Klassen gleich oft vorkommen) und Werte dazwischen für einen Kompromiss.

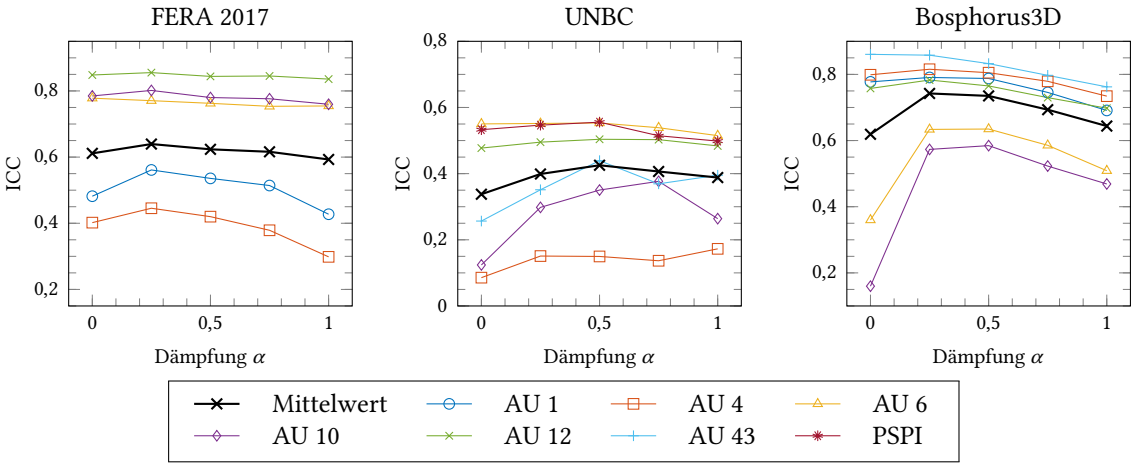
Abb. 3.10a zeigt Ergebnisse der Anwendung von MID mit zwei Merkmalsarten (LM = 2D-Landmarken, LBP = Local Binary Pattern Merkmale aus 5×5-Gitter) in Kombination mit vier Lernmodellen (SVR-E = Support Vector Regression Ensemble, SVM-E = Support Vector Machine Ensemble, RF-R = Random Forest Regression, RF-K = Random Forest Klassifikation) auf den drei oben genannten Datensätzen. Bei keiner der Merkmal-Modell-Kombinationen liefert $\alpha = 0$, was auch dem Ignorieren der Ungleichverteilung ohne Anwendung von MID entspricht, die besten Ergebnisse. Durch die vorgeschlagene Dämpfung der Ungleichverteilung lassen sich in allen Fällen bessere Ergebnisse erreichen. Der optimale Wert des Parameters α ist abhängig von Datensatz, Merkmalen und Lernverfahren und variiert zwischen 0,25 und 1. Insofern ist α ein Tuning-Parameter, der wie andere Modellparameter beim maschinellen Lernen systematisch gewählt werden kann, z. B. mittels Gittersuche. Der Vergleich der Ergebnisse von RF und Support-Vektor-Verfahren bei $\alpha = 0$ und $\alpha = 0{,}25$ zeigt, dass der RF für Ungleichverteilung weniger anfällig ist.



(a) Test-Performances verschiedener Kombinationen von Merkmalen und Klassifikatoren / Regressoren: Mittelwerte über alle Labels. Merkmale: 2D-Landmarken (LM, 98-dimensional), Local Binary Pattern Histogramme (LBP, 1475-dim.). Klassifikatoren: Support Vector Machine Ensemble (SVM-E), Random Forest (RF-K). Regressoren: Support Vector Regression Ensemble (SVR-E), Random Forest (RF-R).



(b) Verbesserung der Test-Performances durch Ensemble (SVR-E / SVM-E) im relativ zu einfacher SVR / SVM (Differenz ΔICC, positive Werte sind Verbesserungen). Legende: siehe (a).



(c) Test-Performances des Multi-Label CNN MobileNetV3 (trainiert mit gemäß MID gewichtetem Loss): Mittelwerte über alle Label (schwarz / fett) und Performances einiger ausgewählter Label.

Abbildung 3.10.: Handhabung der Ungleichverteilung der Klassenzugehörigkeit mittels MID für den Datensatz FERA 2017 (links, 7 AUs, alle Ansichten), den Datensatz UNBC (mittig, 10 AUs und PSPI) und den Datensatz Bosphorus3D (rechts, 26 AUs). Die Wahl des Dämpfungsparameters α bestimmt die Reduzierung der Ungleichverteilung in den Trainingsdaten: $\alpha = 0$ entspricht dem Beibehalten der Originalverteilung, $\alpha = 1$ dem vollständigen „Angleichen“ der beiden häufigsten Klassen, und $0 < \alpha < 1$ einem Kompromiss dazwischen.

Insbesondere die RF-Regression erreicht auch ohne Dämpfung ($\alpha = 0$) bereits recht gute Ergebnisse. Jedoch sind die Ergebnisse der Support-Vektor-Verfahren mit hinreichend starker Dämpfung $\alpha \geq 0{,}5$ in allen Fällen besser als die eines jeden RF. Auch der Vergleich von Regression und Klassifikation zeigt, dass die Regressionsverfahren spätestens bei $\alpha \geq 0{,}5$ bessere Ergebnisse liefern – oft auch bei kleineren α. Die besten Ergebnisse werden sowohl auf dem FERA 2017 als auch auf dem Bosphorus3D-Datensatz mit der Kombination LBP-Merkmale und SVR-Ensemble-Modell erreicht. Auf dem UNBC-Datensatz erzielt die Kombination Landmarken-Merkmale und SVR-Ensemble-Modell die besten Ergebnisse. UNBC hat weniger Varianz in den Trainingsdaten als die anderen beiden Datensätze, da die Anzahl der Probanden und Samples sowie der verschiedenen Kopfposen und Gesichtsausdrücke geringer ist. Vermutlich kann die Performance bei UNBC daher weniger von den höherdimensionalen LBP-Merkmalen profitieren, als bei Bosphorus3D und FERA.

Abb. 3.10b fasst zusammen, inwiefern die Verwendung von Ensembles die Performance der klassischen SVR und SVM verbessert. Dargestellt sind die ICC-Differenzen von SVR/SVM-Ensemble und einfacher SVR/SVM. Entgegen der Erwartung des Autors dieser Dissertation sind die Ergebnisse mit Ensemble nicht in allen Fällen besser. Größere Performance-Vorteile ergeben sich tendenziell bei Datensätzen mit höherer Sample-Anzahl und -Varianz (FERA und Bosphorus3D), bei Verwendung höherdimensionaler Merkmale (LBP) und bei Dämpfung der Ungleichverteilung ($\alpha > 0$), insbesondere für die Kombination LBP mit SVR-E. Insgesamt betrachtet sind die Ergebnisse der SVR/SVM-Ensemble jedoch in allen drei Datensätzen statistisch signifikant besser (Permutationstest; FERA: Mittelwert 0,313 vs. 0,352, $p < 0{,}001$; UNBC: 0,133 vs. 0,145, $p < 0{,}01$; Bosphorus3D: 0,172 vs. 0,208, $p < 0{,}01$). Daher werden im Folgenden nur noch Ergebnisse aufgeführt, die mit SVR/SVM-Ensemble erzielt wurden.

Abb. 3.10c veranschaulicht den Einfluss der Dämpfung auf die Performance des CNN MobileNetV3, das jeweils mittels Transferlernen (ausgehend von ImageNet-Gewichten) mit dem Huber-Loss (Regression) trainiert wurde. Der Mittelwert der Performances aller Label (schwarz) wird durch die Dämpfung der Ungleichverteilung ($\alpha > 0$) verbessert. Für die einzelnen Label ist der Einfluss von α sehr verschieden, wie auch die jeweilige Ungleichverteilung. Die Sample-Anzahl der Mehrheitsklasse im Verhältnis zu der Summe aller anderen Klassen ist z. B. bei AU 10 in FERA 1,03 zu 1, in UNBC 91 zu 1, und in Bosphrous 12,3 zu 1. Bei FERA beobachtet man entsprechend einen geringen und bei den anderen beiden Datensätzen einen größeren Einfluss von α. Auch zeigt sich noch deutlicher als in Abb. 3.10a, dass zu starke Dämpfung der Performance auch schaden kann und der optimale Wert von α als Tuning-Parameter gewählt werden sollte. Das vorgeschlagene MID-Verfahren ermöglicht systematisch nach einer optimalen Dämpfung der Ungleichverteilung zu suchen und diese anzuwenden. Für die folgenden Untersuchungen wird dies wird jedoch *nicht* gemacht, da der dafür nötige Rechenaufwand mit den Ressourcen, die dem Autor der Dissertation zur Verfügung stehen, nicht zu bewältigen ist.

Untersuchung der 3D-Merkmale und Kopfpose: In Abschnitt 3.2.3 wurden Algorithmen zur Extraktion von 3D-Merkmalen und der Schätzung der Kopfpose vorgeschlagen. Die Extraktion der 3D-Merkmale nutzt die Kopfposeschätzung und hängt insofern vom Fehler dieser Schätzung ab. Außerdem können auch die Kopfposeparameter selbst als Merkmal genutzt werden. Zur Evaluierung des Verfahrens zur Kopfposeschätzung wird hier ein quantitativer Vergleich mit dem in [Wer+13; Wer+14b; Wer+17; WAHW17] verwendeten Verfahren durchgeführt, dessen Schätzung auf Tiefenkarten und dem *Iterative Closest Point (ICP)* Algorithmus basiert und in [Wer+13; NWAH13] validiert wurde. Die 3D-basierte Schätzung wurde mithilfe der Daten der Kinect-Kamera des BioVid-Datensatzes durchgeführt, die unmittelbar über der Kamera für die Frontalansicht angebracht war. Trotz der frontalen Platzierung der Kamera gibt es aufgrund der

Bewegungsfreiheit der Probanden eine deutliche Kopfposevarianz (Wertebereich pitch ca. ±30°, yaw ca. ±40°, roll ca. ±20° und ±15 cm zum Mittelwert für X, Y, und Z; Standardabweichungen unter 15° bzw. 5 cm). Abb. 3.11a zeigt für jeden Kopfposeparameter zwei Maße zur Beurteilung der Übereinstimmung der Ergebnisse der beiden Kopfposeschätzverfahren: Die Korrelationen sind alle größer 0,7 und damit als hoch zu bewerten. Die Mittelwerte der Absolutwerte der Differenzen der beiden Schätzergebnisse sind gering, was für die Validität der Ergebnisse spricht. Eine Ausnahme bildet die Differenz bei dem Z-Parameter, d. h. der Entfernung zur Kamera, die mit 16 cm auffällig groß ist. Dieser Fehler des in Abschnitt 3.2.3 vorgeschlagenen Verfahrens entsteht insbesondere aufgrund der Annahmen, die bezüglich des Augenabstandes und des Öffnungswinkels der Kamera getroffen wurden, damit das Verfahren auf beliebigen 2D-Bilddaten, auch ohne 3D-Kamera anwendbar ist. Abweichungen von diesen Annahmen erzeugen je nach Person und Kamera einen systematischen Fehler, der sich insbesondere auf die Z-Koordinate auswirkt. Ein weiterer wesentlicher Bestandteil des Fehlers kommt durch Ungenauigkeiten bei der Landmarkenlokalisation zustande. Auf die 3D-Merkmale, die auf Basis der Kopfposeschätzung extrahiert werden, hat der Fehler der Z-Koordinate jedoch einen sehr geringen Einfluss.

Die in Abschnitt 3.2.3 beschriebenen 3D-Abstands- und 3D-Koordinatenmerkmale lassen sich jeweils mit dem detaillierteren 3D-Gesichtsmodell und dem einfacheren zylinderbasierten 3D-Modell extrahieren. Im Folgenden werden die vier Kombinationen untereinander und mit den zugrunde liegenden 2D-Landmarken verglichen. Betrachtet werden die Mittelwerte der Erkennungsergebnisse, die jeweils mit den Lernverfahren SVM-E, SVR-E, RF-K und RF-R erzielt wurden. Abb. 3.11b zeigt als Tabelle die ICC-Werte für die Merkmale (Zeilen) und die Datensätze FERA 2017, BioVid-S und BioVid-S7 (Spalten), jeweils mit ausschließlich frontaler Ansicht und mit allen verfügbaren Ansichten, sowie Mittelwerte (M). Im Mittel über alle Datensätze (letzte Spalte) sind die Ergebnisse mit 3D-Koordinaten und zylinderbasiertem Modell am besten, gefolgt von den 2D-Landmarken. Bei den Frontalansichten ist die Rangfolge je nach Datensatz verschieden, im Mittel sind jedoch auch hier 3D-Koordinaten und zylinderbasiertes Modell überlegen. Bei mehreren Ansichten, mit größerer Kopfposevarianz, liefern jedoch in allen Fällen die 2D-Landmarken bessere Ergebnisse als die 3D-Merkmale. Relativ betrachtet schneiden somit die 3D-Koordinaten bei den Frontalansichten besser ab als bei den Datensätzen mit mehreren Ansichten. Dies ist insofern überraschend, dass es der Motivation der 3D-Merkmale widerspricht, verbesserte Kopfposeinvarianz zu erreichen. Hierbei ist jedoch zu bedenken, dass die Frontalansichten nicht ausschließlich frontale Gesichter zeigen, sondern auch über eine mittlere Kopfposevarianz verfügen. Daher deuten bereits die Ergebnisse der Frontalansicht darauf hin, dass die 3D-Koordinaten eine bessere Kopfposeinvarianz ermöglichen. Diesem positiven Effekt steht eine Merkmalsreduktion gegenüber (98-dim. bei 2D-Landmarken auf 17-dim. bei 3D-Koordinaten), durch die relevante Informationen verloren gehen können, was möglicherweise der Grund für die schlechteren Ergebnisse bei FERA ist. Die Datensätze mit mehreren Ansichten bestehen überwiegend aus nichtfrontalen Bilder mit Kopfposewinkeln von im Mittel etwa ±45°. Hier kommt die Modellierung an ihre Grenzen. Während das im Methodenabschnitt gezeigte Beispiel bei frontaler Kopfpose (Abb. 3.3) den Idealfall der Punktprojektion und Merkmalsextraktion zeigt, treten bei den Beispielen in Abb. 3.11 Probleme auf. Bereits einer leichten yaw-Drehung von 24,3° (Abb. 3.11c) gibt es in den frontalisierten Punkten (im u-v-Koordinatensystem) Ungenauigkeiten auf der Gesichtshälfte, die der Kamera abgewandt ist, insbesondere erkennbar an der Asymmetrie des Mundes und der Nase. Außerdem fehlen zwei Landmarken an der im Bild rechten Augenbraue, da die Sehstrahlen der zugehörigen 2D-Landmarken das 3D-Modell nicht schneiden sondern knapp verfehlen und somit keine Projektion auf das Modell möglich ist. Ursache der beiden Phänomene sind kleine Ungenauigkeiten der Landmarkenlokalisierung, der Kopfposeschätzung und kleine Abweichungen der realen 3D-Geometrie vom verwendeten 3D-Modell, die jeweils einen Teil zum Gesamtfehler beitragen oder sich durch Fehlerfortpflanzung verstärken. Noch deutlicher zeigen sich die

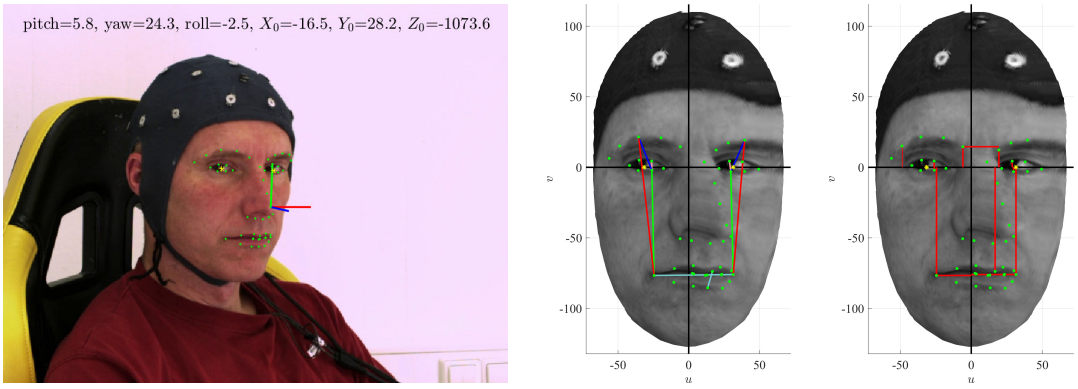
	Pitch	Yaw	Roll	X	Y	Z
Pearson-Korrelation	0,83	0,76	0,87	0,98	0,96	0,75
Mittlere absolute Differenz	5,3°	5,3°	2,2°	3,6 cm	1,4 cm	16,2 cm

(a) Evaluierung der genutzten, auf 2D-Landmarken basierenden Kopfposeschätzung durch Vergleich mit einem 3D-Referenzverfahren [Wer+13; NWAH13].

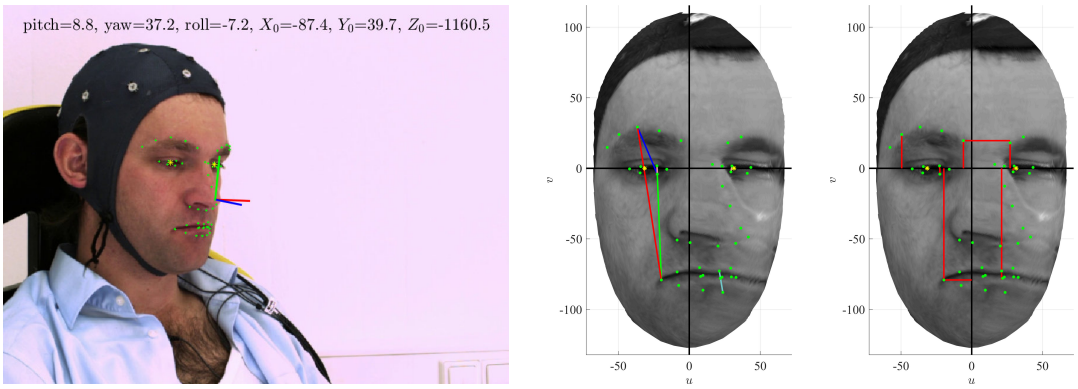
	Frontale Ansicht				Mehrere Ansichten				
Merkmale / 3D-Modell	**FERA17**	**BioVid-S**	**BioVid-S7**	*M*	**FERA17[a]**	**BioVid-S[b]**	**BioVid-S7[b]**	*M*	*M*
2D-Landmarken	**0,424**	0,150	0,411	0,328	**0,346**	**0,146**	**0,444**	**0,312**	0,320
3D-Abstände / Kopf	0,388	0,150	0,433	0,324	0,303	0,111	0,409	0,274	0,299
3D-Abstände / Zylinder	0,387	0,136	0,380	0,301	0,316	0,126	0,438	0,293	0,297
3D-Koordinaten / Kopf	0,419	**0,166**	0,520	0,368	0,337	0,116	0,359	0,270	0,319
3D-Koordinaten / Zylinder	0,417	0,165	**0,593**	**0,391**	0,336	0,111	0,360	0,269	**0,330**

[a] 9 Ansichten (sehr große Kopfposevarianz in pitch und yaw) [b] 3 Ansichten (sehr große Kopfposevarianz in yaw)

(b) Mit landmarkenbasierten Merkmalen erreichte Performance für verschiedene Datensätze (jeweils Mittelwert von SVM-E, SVR-E, RF-K und RF-R; bei FERA17 zusätzlich Mittelwert aus 7 AUs). Gegeben sind außerdem Mittelwerte (*M*) der Ergebnisse von (1) den 3 frontalen Datensätzen, (2) den 3 Datensätze mit mehreren Ansichten (d. h. mehr nicht-frontalen Kopfposen) und (3) von allen 6 Datensätzen.



(c) Beispiel für leichte Drehung aus der Bildebene: Eingabebild mit Kopfpose und Landmarken (links), Veranschaulichung der 3D-Abstandsmerkmale (mittig) und der 3D-Koordinatenmerkmale (rechts).



(d) Beispiel für stärkere Drehung aus der Bildebene: Eingabebild mit Kopfpose und Landmarken (links), Veranschaulichung der 3D-Abstandsmerkmale (mittig) und der 3D-Koordinatenmerkmale (rechts).

Abbildung 3.11.: Untersuchung der 3D-Merkmale und der Kopfposeschätzung. Beispiele aus BioVid-S unter Verwendung des detaillierten 3D-Kopfmodells.

Probleme bei größeren Winkeln (Abb. 3.11d, yaw 37,2°). Hier sind die falschen Asymmetrien der frontalisierten Landmarkenkoordinaten noch stärker ausgeprägt und es fehlen noch mehr Landmarken in der im Bild rechten Gesichtshälfte. Die Probleme bei Verwendung des zylinderbasierten 3D-Modells sind qualitativ ähnlich. Der geringere Detailgrad des Zylindermodells führt in den Experimenten im Allgemeinen nicht zu schlechteren Ergebnissen.

Folge der mit den Kopfposewinkeln zunehmenden Fehler sind poseabhängige Unterschiede in den Merkmalen bzw. das Fehlen von Merkmalsinformationen im Falle von fehlenden frontalisierten Punkten. Insofern ist die Invarianz gegenüber der Kopfpose, die mit den vorgestellten 3D-Merkmalen erreicht wird, begrenzt. Dennoch können mit den 3D-Merkmalen, die sich durch Nutzung von Vorwissen durch eine sehr niedrige Dimensionalität auszeichnen, bei einigen Erkennungsaufgaben sehr gute Ergebnisse erzielt werden, wie im Folgenden gezeigt wird. Dabei werden die 3D-Abstände in Kombination mit dem realitätsnahen Gesichtsmodell und die 3D-Koordinaten mit dem zylinderbasierten Modell verwendet.

Vergleich der Merkmale und Lernverfahren: In vorangegangenen Abschnitten wurden im Zuge anderer Betrachtungen bereits einige Vergleiche von Merkmalsarten und Lernverfahren angestellt. Hier erfolgt nun eine umfassende und systematische Gegenüberstellung von zahlreichen Kombinationen von Merkmalsarten und Lernverfahren. Die Ansätze werde auf den Datensätzen FERA 2017, UNBC, BioVid-S und BioVid-S7 untersucht, wobei FERA 2017 und die beiden BioVid-Datensätze jeweils mit nur der Frontalansicht und mit allen Ansichten betrachtet werden. Für UNBC werden die Ergebnisse für die zehn annotierten AUs (Mittelwert der ICCs) sowie für die Prädiktion des PSPI-Wertes aufgeführt. Tabelle 3.2 stellt die Ergebnisse der verschiedenen Kombinationen von Merkmalsarten und Lernverfahren (Zeilen) auf den verschiedenen Datensätzen (Spalten) gegenüber. Die Wahl der Hintergrundfarben basiert auf der Verteilung der Performances der jeweiligen Spalte. ICC-Werte unterhalb des 30%-Quantil sind weiß hinterlegt und die farbliche Sättigung darüber ist proportional zum ICC-Wert bis zum maximalen ICC-Wert der Spalte. Vergleicht man die drei Hauptabschnitte der Tabelle fällt farblich sofort ins Auge, dass für fast alle Datenbanken die besten Ergebnisse beim Transferlernen mit CNN zu finden sind, genauer gesagt bei der Anwendung des CNN (MobileNetV3 = MN) als Lernverfahren (Zeilen #29-32 und #37-40), d. h. einer Feinjustierung der CNN-Gewichte. Die Performances bei Verwendung des auf Bosphorus3D vortrainierten CNN als Ausgangspunkt (Zeilen #37-40, vorgeschlagen in Abschnitt 3.2.4) liegen dabei (mit im Mittel 0,463) statistisch signifikant ($p = 0{,}002$, Permutationstest) über den Performances bei Verwendung des auf ImageNet vortainierten CNN (Zeilen #29-32, im Mittel 0,434). Die drittbeste Kategorie in dieser Evaluierung ist das CNN-Training mit zufälliger Initialisierung (Zeilen #41-44), dessen Ergebnisse mit im Mittel 0,326 statistisch signifikant schlechter sind als die CNN-Transferlernverfahren ($p < 0{,}0001$). Insbesondere bei den Datensätzen BioVid-S und BioVid-S7, die nur relativ wenige Samples umfassen, sind die Ergebnisse deutlich schlechter. Bei BioVid-S, das aufgrund des Label-Rauscheffektes schwerer zu klassifizieren ist, konvergiert das Netz zum Teil zu einer Lösung, die nicht besser generalisiert als Raten (ICC nahe 0). Das Transferlernen mit Support-Vector- und Random-Forest-Methoden unter Nutzung der Merkmale des auf Bosporus3D vortrainierten CNN (Zeilen #33-36) funktioniert im Mittel ähnlich gut wie das CNN ohne Transferlernen (Mittelwert 0,315, Unterschied nicht signifikant). In den Zeilen #25-28 (Mittelwert 0,175) zeigt sich, dass die CNN-Merkmale für die ImageNet-Klassifikation ohne weitere Feinjustierung nur relativ wenige zur Mimikerkennung nützliche Informationen bereitstellen, da alle anderen Lernverfahren mit Nutzung von CNN-Merkmalen deutlich bessere Ergebnisse erzielen.

Bei den menschengemachten Merkmalen (Zeilen #1-24) ist keine der Merkmalskategorien klar überlegen. Einzig die Kopfpose sticht heraus, die mit einer mittleren Performance von 0,032 den

#	Merkmale / Ausgangspunkt (Dim.)	Lernverfahren / Modell	FERA 2017 Frontal 7 AUs	FERA 2017 9 Ans. 7 AUs	UNBC 10 AUs	UNBC PSPI	BioVid-S Frontal Stim.	BioVid-S 3 Ans. Stim.	BioVid-S7 Frontal Stim.	BioVid-S7 3 Ans. Stim.
	Menschengemachte Merkmale									
1	2D-Landmarken	SVM-E	0,419	0,340	0,135	0,345	0,170	0,140	0,384	0,535
2	(98-dim.)	SVR-E	0,463	0,370	0,173	0,387	0,140	0,152	0,440	0,503
3		RF-K	0,398	0,329	0,087	0,258	0,144	0,145	0,434	0,380
4		RF-R	0,417	0,344	0,154	0,396	0,144	0,146	0,385	0,358
5	LBP 5×5	SVM-E	0,424	0,356	0,114	0,348	0,110	0,075	0,588	0,500
6	(1475-dim.)	SVR-E	0,450	0,384	0,155	0,401	0,070	0,102	0,567	0,440
7		RF-K	0,327	0,256	0,013	0,021	0,172	0,136	0,664	0,483
8		RF-R	0,371	0,281	0,093	0,280	0,182	0,157	0,626	0,439
9	LBP 10×10	SVM-E	0,372	0,366	0,109	0,382	0,161	0,135	0,664	0,530
10	(5900-dim.)	SVR-E	0,456	0,408	0,178	0,473	0,137	0,123	0,667	0,520
11		RF-K	0,322	0,237	0,012	0,011	0,183	0,111	0,542	0,498
12		RF-R	0,334	0,279	0,125	0,330	0,166	0,134	0,502	0,474
13	3D-Abstände	SVM-E	0,376	0,304	0,176	0,363	0,183	0,105	0,483	0,429
14	(10-dim.)	SVR-E	0,401	0,318	0,184	0,395	0,163	0,101	0,464	0,390
15		RF-K	0,386	0,297	0,143	0,371	0,127	0,121	0,371	0,404
16		RF-R	0,389	0,292	0,182	0,450	0,129	0,117	0,412	0,413
17	3D-Koordinaten	SVM-E	0,412	0,338	0,159	0,392	0,191	0,120	0,578	0,344
18	(17-dim.)	SVR-E	0,439	0,341	0,181	0,414	0,197	0,112	**0,669**	0,299
19		RF-K	0,410	0,335	0,099	0,255	0,138	0,103	0,569	0,412
20		RF-R	0,405	0,332	0,167	0,408	0,133	0,107	0,554	0,385
21	Kopfpose	SVM-E	0,041	0,004	0,007	-0,008	0,072	0,018	0,054	-0,010
22	(6-dim.)	SVR-E	0,015	0,002	0,012	0,022	0,061	0,010	0,000	0,003
23		RF-K	0,068	0,068	0,004	0,014	0,016	0,023	0,117	0,023
24		RF-R	0,068	0,072	0,016	0,053	0,005	0,022	0,112	0,033
	Transferlernen mit CNN									
25	MN-K ImageNet	SVM-E	0,380	0,340	0,038	0,131	0,072	0,080	0,279	0,206
26	(1280-dim.)	SVR-E	0,430	0,377	0,064	0,175	0,045	0,102	0,256	0,107
27		RF-K	0,333	0,266	0,001	0,001	0,085	0,070	0,122	0,261
28		RF-R	0,356	0,267	0,029	0,094	0,069	0,087	0,234	0,248
29		MN-K	0,614	0,601	0,187	0,480	0,188	0,228	0,572	0,530
30		MN-R	0,651	0,630	0,319	0,535	0,188	0,204	0,536	0,525
31		MN-K^{MT}	0,630	0,598	0,261	0,522	0,192	0,215	0,397	0,542
32		MN-R^{MT}	0,633	0,622	0,348	0,548	0,207	0,227	0,412	0,532
33	MN-R Bosphorus3D	SVM-E	0,437	0,424	0,138	0,385	0,162	0,143	0,289	0,393
34	(1280-dim.)	SVR-E	0,472	0,464	0,161	0,369	0,067	0,127	0,332	0,237
35		RF-K	0,435	0,416	0,065	0,201	0,205	0,206	0,517	0,513
36		RF-R	0,470	0,441	0,179	0,423	0,189	0,202	0,508	0,491
37		MN-K	0,606	0,623	0,298	0,474	0,183	0,208	0,488	0,582
38		MN-R	0,651	**0,638**	0,383	**0,599**	0,200	0,214	0,645	0,586
39		MN-K^{MT}	**0,661**	0,610	0,329	0,486	0,208	**0,238**	0,586	0,585
40		MN-R^{MT}	0,653	0,630	**0,385**	0,572	**0,244**	0,235	0,406	**0,601**
	CNN ohne Transferlernen									
41	MN	MN-K	0,575	0,539	0,142	0,308	0,137	-0,004	0,279	0,336
42	(1280-dim.)	MN-R	0,605	0,555	0,262	0,417	0,065	0,000	0,324	0,361
43		MN-K^{MT}	0,545	0,492	0,288	0,418	0,128	0,060	0,390	0,474
44		MN-R^{MT}	0,562	0,498	0,238	0,430	0,072	0,054	0,475	0,415
	Größe Trainingsset(s)		147k	1.322k	38.720		2.784	8.352	224	672
	Größe Testset(s)		76k	681k	9680 (×5)		696 (×5)	2.088 (×5)	56 (×5)	168 (×5)

^{MT} Multi-Task-Lernen mit Bosphorus3D SVM-E: Support Vector Machine Ensemble (Klassifikation)
SVR-E: Support Vector Regression Ensemble RF-K: Random Forest Klassifikation RF-R: Random Forest Regression
MN-K: MobileNetV3-large (CNN) Klassifikation MN-R: MobileNetV3-large (CNN) Regression

Tabelle 3.2.: Einzelbildbasierte Mimikerkennung in Abhängigkeit von Merkmalen und Lernverfahren: Test-Performances (ICC) mit verschiedenen Kombinationen von Merkmalsarten und Lernverfahren (Zeilen) für die Datensätze (Spalten): FERA17 (Frontalansicht und alle 9 Ansichten [große Kopfposevarianz], mittlerer ICC der 7 AUs), UNBC (mittlerer ICC von 10 AUs und ICC von PSPI), BioVid-S (Frontalansicht und alle 3 Ansichten) und BioVid-S7 (Teilmenge von BioVid-S mit den 7 expressivsten Probanden). Angegeben sind auch die Dimensionalität des Merkmalsraums und die Größe der Trainings- und Testsets.

anderen Merkmalen deutlich unterlegen ist. Von diesen schneiden die 3D-Koordinaten mit im Mittel 0,313 am besten ab, gefolgt von LBP 10×10 (0,311), 2D-Landmarken (0,301), LBP 5×5 (0,300), und den 3D-Abständen (0,295). Die Unterschiede dieser fünf Merkmalsarten sind jedoch nicht statistisch signifikant (nach Varianzanalyse), so dass keine datensatzübergreifende Schlussfolgerung möglich ist.

Vergleicht man die besten CNN-Ergebnisse mit den besten Ergebnissen mit menschengemachten Merkmalen, fällt auf, dass insbesondere die Datensätze mit großen Trainingssets wie FERA (etwa 147.000 bzw. 1,3 Millionen Beispiele) und UNBC (etwa 39.000 Samples) vom CNN-Training profitieren. Die Abstände werden jedoch mit der Anzahl der verfügbaren Trainingsbeispiele geringer. Bei BioVid-S7 (frontal), dem kleinsten Datensatz mit im Mittel 224 Trainingsbeispielen (die Anzahl variiert zwischen den folds der 5-fachen Kreuzvalidierung), schneiden sogar mehrere Kombinationen von klassischen Merkmalen und Lernverfahren (#7, #9-10, #18) besser ab als die besten Ergebnisse mit CNN. Die besten Merkmale für diesen Datensatz sind die vom Autor dieser Dissertation vorgeschlagenen 3D-Koordinaten sowie die weit verbreiteten LBP-Texturmerkmale. Interessant ist, dass die LBP-Merkmale so gute Ergebnisse liefern, die in Relation zur Größe des Datensatzes sehr hochdimensional sind. In der LBP-Variante mit 5×5 Gitterzellen (ca. 1.500 Dimensionen) sind die nicht-linearen Random Forests den linearen Support Vector Methoden überlegen; in der 10×10-Variante mit 5.900 Dimensionen ist es umgekehrt – der größere Merkmalsraum ermöglicht hier eine bessere lineare Trennung. Beim BioVid-S7-Datensatz mit drei Ansichten auf den Kopf, der auch dreimal so viele Samples umfasst, kann die Erkennung mit LBP vom höherdimensionalen Merkmalsraum der 10×10-Variante profitieren (gegenüber 5×5). Trotzdem sind die Ergebnisse mit drei Ansichten verglichen mit denen bei einer Ansicht in den meisten Fällen schlechter, da durch die größere Kopfposevarianz auch die Erkennungsaufgabe schwieriger ist. Interessante Ausnahmen hiervon sind die Erkennung basierend auf 2D-Landmarken und SVM/SVR (Zeilen #1 und #2) sowie einige Ergebnisse unter Nutzung von CNNs. Bei BioVid-S (mit allen 87 Probanden) zeigen sich ähnliche Phänomene. Überwiegend sind die Ergebnisse mit drei Ansichten dort schlechter als nur mit der Frontalansicht; lediglich die 2D-Landmarken schneiden zum Teil etwas besser ab, ebenso wie die Mehrheit der CNN-Ergebnisse. Deutlich klarer ist das Bild bei FERA, wo fast alle Ergebnisse im Datensatz mit großer Kopfposevarianz (9 Ansichten) deutlich schlechter abschneiden als in dem mit nur einer Ansicht. Hier überwiegt der Nachteil, dass die Erkennungsaufgabe deutlich schwieriger wird gegenüber dem positiven Effekt durch mehr Varianz in den Trainingsdaten. Wahrscheinlich kommt letzterer Effekt nur wenig zum Tragen, da der Frontaldatensatz mit 147.000 Trainingsbeispielen bereits sehr groß ist.

Die Tabelle vergleicht in aufeinanderfolgenden Zeilen jeweils die Ergebnisse von Klassifikation und Regression. Betrachten wir zunächst die Datensätze, bei denen Intensitätslabel mit mehr als zwei Abstufungen vorliegen: FERA und UNBC. Hier liefert die Regression signifikant bessere Ergebnisse als die Klassifikation (im Mittel 0,345 vs. 0,299, $p < 0{,}0001$ mit Permutationstest). Anders verhält es sich bei den Datensätzen BioVid-S und BioVid-S7, bei denen nur zwei Klassen betrachtet werden. Die Mittelwerte aller Regressionsergebnisse (0,267) und Klassifikationsergebnisse (0,275) unterscheiden sich hier nicht so deutlich, der Permutationstest erreicht jedoch knapp das Signifikanzniveau ($p = 0{,}046$). Betrachtet man bei diesen Datensätzen jedoch nur die reinen CNN-Transferlernmodelle (Zeilen #29-32 und 37-40), die meist die besten Ergebnisse liefern, so findet man keinen signifikanten Unterschied (Regression im Mittel 0,373, Klassifikation 0,371).

Beim Vergleich von Support-Vector- und Random-Forest-Methoden schneiden die Support-Vector-Methoden signifikant besser ab (im Mittel 0,265 vs. 0,245, $p = 0{,}007$). Das Multi-Task-Lernen mit dem Bosphorus3D-Datensatz (Zeilen #31-32, #39-40; im Mittel 0,449) bringt bei den CNN-Transferlernmodellen im Allgemeinen keinen signifikanten Vorteil gegenüber dem klassischen Transferlernen mit nur einem Datensatz (Zeilen #29-30, #37-38; im Mittel 0,447). Bei den

kleineren BioVid-S und BioVid-S7 Datensätzen zeigt sich in den Mittelwerten ein kleiner Vorteil durch die Multi-Task-Modelle (0,380 vs. 0,364). Auch dieser Unterschied erreicht jedoch nicht das Signifikanzniveau.

Fusion von Merkmalen: In Tabelle 3.2 werden acht Merkmalsarten verglichen, die sich miteinander kombinieren lassen. Da eine vollständige Untersuchung der Fusion aller Kombinationen den Rahmen dieser Arbeit sprengen würde, werden nur einige ausgewählte Kombinationen betrachtet. Die Merkmale lassen sich in drei Kategorien einteilen, die komplementäre Informationen bereitstellen: geometrische Merkmale, texturbasierte Merkmale und die Kopfpose. Im Folgenden werden Ergebnisse der Fusion von Merkmalen gezeigt, die jeweils aus verschiedenen Kategorien kommen, d. h. verschiedene geometrische Merkmale werden nicht miteinander kombiniert, ebenso wenig verschiedene texturbasierte Merkmale. Als geometrische Merkmale werden die 2D-Landmarken und 3D-Koordinaten betrachtet. Die untersuchten texturbasierten Merkmale sind LBP 10×10 und die vom MobileNetV3 extrahierten Merkmale (Regression mit Bosphorus3D, Zeile #34 in Tabelle 3.2).

Tabelle 3.3a listet die Fusionsergebnisse mit SVR Ensembles auf. Die Färbung der Zellen ist hier relativ zur besten Merkmalsart, die bei der Fusion beteiligt ist: grüne Fusionsergebnisse sind besser, rote schlechter (Details siehe Tabellenbeschriftung). Im Mittel über alle Datensätze (Spalte *M*) liefert die Kombination von 2D-Landmarken und LBP die besten Ergebnisse, gefolgt von 2D-Landmarken mit LBP und Kopfpose und 3D-Koordinaten mit LBP. Keine dieser Kombinationen ist jedoch datensatzübergreifend signifikant besser als die beteiligten Merkmalsarten für sich genommen (Permutationstest). Den größten Performance-Gewinn durch Fusion beobachten wir bei BioVid-S7 mit drei Ansichten: 5,3% bei 3D-Koordinaten mit Kopfpose und 3,4% bei 2D-Landmarken mit LBP. Diese sind jedoch nicht statistisch signifikant (Permutationstest), aufgrund der geringen Datensatzgröße und der großen zufälligen Varianz der zugrundeliegenden Ergebnisse. Statistisch signifikante Verbesserungen ergeben sich bei FERA und UNBC, insbesondere bei den Kombinationen 2D-Landmarken, LBP und Kopfpose. Bei FERA fällt die Kombination 3D-Koordinaten, MobilNetV3 und Kopfpose auf, die sowohl bei der Frontalansicht als auch bei dem Gesamtdatensatz die besten Ergebnisse erzielt.

Ergebnisse, die mit SVM-E, RF-R und RF-K ermittelt wurden, können Anhang B.2 entnommen werden. Qualitativ sind die Ergebnisse ähnlich, wobei ein deutlich Unterschied darin liegt, dass die Support-Vector-Methoden bei Beteiligung von LBP-Merkmalen bessere Ergebnisse liefern, wohingegen Random Forest Methoden bei Beteiligung der MobileNetV3-Merkmale besser sind.

In Tabelle 3.3b sind Ergebnisse aufgeführt, die durch CNN-basierte Fusion erzielt wurden, d. h. durch die Integration von geometrischen Merkmalen und Kopfpose in MobileNetV3 (Regression, vortrainiert auf Bosphorus3D). Abgesehen von einigen wenigen Ergebnissen bei BioVid sind alle Ergebnisse signifikant besser als in Tabelle 3.3a (im Mittel 0,484 vs. 0,311). Datensatzübergreifend betrachtet (Spalte *M*) verbessern sich die Ergebnisse durch die Fusion nur bei der Kombination von MobileNetV3 mit 2D-Landmarken, jedoch nicht statistisch signifikant (Permutationstest). Die größte Verbesserung bringt diese Kombination für BioVid-S7; für die Variante mit drei Ansichten ist die Verbesserung um 5% statistisch signifikant. Weitere statistisch signifikante Verbesserungen finden sich bei FERA unter anderem bei der Fusion mit der Kopfpose bzw. mit der Kopfpose und 3D-Koordinaten, die für die frontale bzw. alle Ansichten jeweils die besten Ergebnisse aller bisherigen Untersuchungen erzielen. Während die Kopfpose als einzelne Merkmalsart ungeeignet für die Mimikerkennung ist, kann sie wie sich hier zeigt jedoch in Fusion mit anderen Merkmalen einen positiven Einfluss auf die Ergebnisse haben.

Merkmale	FERA 2017		UNBC		BioVid-S		BioVid-S7		
	Frontal 7 AUs	9 Ans. 7 AUs	10 AUs	PSPI	Frontal Stim.	3 Ans. Stim.	Frontal Stim.	3 Ans. Stim.	*M*
2D-Landmarken	0,463	0,370	0,173	0,387	0,140	**0,152**	0,440	0,503	0,329
3D-Koordinaten	0,439	0,341	0,181	0,414	**0,197**	0,112	**0,669**	0,299	0,332
LBP	0,456	0,408	0,178	0,473	0,137	0,123	0,667	0,520	0,370
Kopfpose	0,015	0,002	0,012	0,022	0,061	0,010	0,000	0,003	0,016
MobileNetV3	0,472	0,464	0,161	0,369	0,067	0,127	0,332	0,237	0,279
2D-L. + LBP	0,473	0,417**	0,185	0,479	0,135	0,131	0,656	**0,554**	**0,379**
2D-L. + Kopfp.	0,470	0,370	0,169	0,388	0,158	0,146	0,444	0,512	0,332
2D-L. + MobileN.	0,483*	0,470**	0,171	0,386	0,083	0,121	0,323	0,258	0,287
3D-K. + LBP	0,466*	0,415**	0,183	0,480	0,134	0,129	0,659	0,543	0,376
3D-K. + Kopfp.	0,445	0,351	0,171	0,406	0,180	0,115	0,578	0,352	0,325
3D-K. + MobileN.	**0,485****	0,468*	0,169	0,388	0,056	0,130	0,308	0,241	0,281
LBP + Kopfp.	0,455	0,409	0,178	0,474	0,136	0,129	0,664	0,520	0,371
MobileN. + Kopfp.	0,476	0,470**	0,161	0,361	0,079	0,131	0,325	0,232	0,279
2D-L. + LBP + Kopfp.	0,474*	0,419**	**0,186***	**0,481***	0,135	0,130	0,650	0,553	0,378
2D-L. + Mob. + Kopfp.	0,481	0,466	0,168	0,388	0,085	0,126	0,331	0,291	0,292
3D-K. + LBP + Kopfp.	0,464	0,415**	0,182	0,477	0,132	0,131	0,656	0,543	0,375
3D-K. + Mob. + Kopfp.	**0,485****	**0,471****	0,168	0,389	0,076	0,128	0,329	0,240	0,286

* signifikant besser als bestes Einzelmerkmal der Fusion, $p < 0{,}05$
** signifikant besser als bestes Einzelmerkmal der Fusion, $p < 0{,}001$

(a) Merkmalsfusion mit SVR-E.

Merkmale	FERA 2017		UNBC		BioVid-S		BioVid-S7		
	Frontal 7 AUs	9 Ans. 7 AUs	10 AUs	PSPI	Frontal Stim.	3 Ans. Stim.	Frontal Stim.	3 Ans. Stim.	*M*
MN-R	0,651	0,638	**0,383**	**0,599**	0,200	0,214	0,645	0,586	0,490
MN-R + 2D-Landmarken	0,644	0,643*	0,367	0,572	0,190	**0,216**	**0,679**	**0,636***	**0,493**
MN-R + 3D-Koordinaten	0,650	0,645**	0,358	0,570	0,201	0,209	0,623	0,600	0,482
MN-R + Kopfpose	**0,670***	0,643	0,356	0,585	0,177	0,210	0,622	0,597	0,483
MN-R + 2D-L.+Kopfp.	0,631	0,640	0,357	0,578	0,187	0,206	0,601	0,587	0,473
MN-R + 3D-K.+Kopfp.	0,644	**0,652****	0,365	0,575	**0,215**	0,210	0,614	0,589	0,483

* signifikant besser als MN-R, $p < 0{,}05$ ** signifikant besser als MN-R, $p < 0{,}01$

(b) CNN-basierte Fusion: Regression mit auf Bosphorus3D vortrainierten MobileNetV3 (MN-R).

Tabelle 3.3.: Einzelbildbasierte Mimikerkennung mit Fusion von Texturmerkmalen, punktbasierten Merkmalen und Kopfposemerkmalen: Test-Performances (ICC) mit verschiedenen Kombinationen von Merkmalsarten (Zeilen) für verschiedene Datensätze (Spalten, Beschreibung siehe Tabelle 3.2) und Mittelwert aller Datensätze (*M*). Die Färbung der Zellen ist relativ zu der maximalen Performance der besten Merkmalsart, die bei der Fusion beteiligt war, und erstreckt sich von der Differenz -0,05 (gesättigtes rot) über ±0 (weiß) bis +0,05 (gesättigtes grün).

Die Hypothese, dass die Fusion von Merkmalen die Ergebnisse der Mimikerkennung verbessert, wird von den Experimenten nicht allgemein bestätigt. Auffallend ist jedoch, dass die Fusionsergebnisse für den Datensatz FERA 2017, insbesondere für die neun Ansichten, sehr positiv ausfallen. Da es sich bei FERA um den größten Datensatz handelt, liegt die Vermutung nahe, dass das schlechtere Abschneiden der anderen Datensätze mit dem Fluch der Dimensionalität in Verbindung steht, d. h. dass dort tendenziell zu wenige Samples zur Verfügung stehen um beim Lernprozess die Relevanz der Merkmale, deren Anzahl sich durch die Fusion vergrößert, für die Erkennungsaufgabe richtig zu bewerten.

Schlussfolgerungen: Die Untersuchungen haben gezeigt, dass mit Transferlernen mit CNNs zumeist bessere Ergebnisse erzielt werden können, als mit der Extraktion von menschengemachten Merkmalen und anschließendem Lernen eines Prädiktionsmodells. Das Transferlernen ausgehend vom Mimikdatensatz Bosphorus3D hat sich als besonders erfolgreich herausgestellt. Anders als beim Lernen mit menschengemachten Merkmalen und SVM oder Random Forest ermöglichen CNNs die Optimierung der Merkmalsextraktion, die implizit im neuronalen Netz passiert, für das jeweilige Lernproblem. CNNs verfügen über eine große Anzahl an Parametern und damit eine große Kapazität für die Anpassung des Modells an die Trainingsdaten. Sie können daher auch mit größerer Varianz in den Trainingsdaten, z. B. Kopfposevarianz durch mehrere Ansichten, sehr gut umgehen. Bei kleineren Datensätzen kommt es dadurch jedoch auch leicht zu Overfitting – trotz guter Initialisierung durch Transferlernen – und es muss durch Regularisierung gegengesteuert werden, z. B. durch Datenaugmentierung, frühes Beenden des Trainings und Drop-Out.

Bei dem sehr kleinen frontalen BioVid-S7 Datensatz (mit 224 Trainingsbeispielen) sind die besten Ergebnisse überwiegend bei den menschengemachten Merkmalen zu finden, bei den vorgeschlagenen niedrigdimensionalen 3D-Koordinaten und bei den hochdimensionalen LBP-Merkmalen. Dies zeigt, dass es sich für kleine Datensätzen noch immer lohnen kann, sich mit dem Entwickeln von Merkmalen zu beschäftigen. Das beste Ergebnis wurde jedoch auch hier mit einem CNN-Modell in Fusion mit den 2D-Landmarkenmerkmalen erreicht. Insbesondere für Daten mit großer Kopfposevarianz, wie hier bei den Datensätzen mit mehreren Ansichten, ist es sehr schwierig Merkmale zu entwerfen, die eine ähnlich gute Performance liefern wie mit einem CNN extrahierte und optimierte Merkmale. Hier ist dies nicht gelungen.

Die Fusion verschiedener Merkmalsarten kann Performance-Vorteile bringen. Beim FERA-Datensatz mit allen 9 Ansichten war beispielsweise keine Fusion schlechter als die beteiligten Merkmale und die meisten waren signifikant besser. Datensatzübergreifend gab es bei den Fusionsexperimenten jedoch keine signifikante Verbesserung. Auch reicht die Fusion von klassischen Merkmalen in keinem der Fälle aus um das Performance-Niveau des CNN zu erreichen. Die Nutzung von menschengemachten Merkmalen als zusätzliche Eingabe für ein CNN ist insofern vielversprechender.

Das vorgeschlagene MID-Verfahren bietet einen systematischen Ansatz zur Handhabung ungleich verteilter Klassenzugehörigkeiten und kann, wie die Experimente gezeigt haben, zu deutlichen Verbesserungen der Ergebnisse führen. Durch Variation des Tuning-Parameters α und Modellselektion kann die Performance für jeden Datensatz optimiert werden.

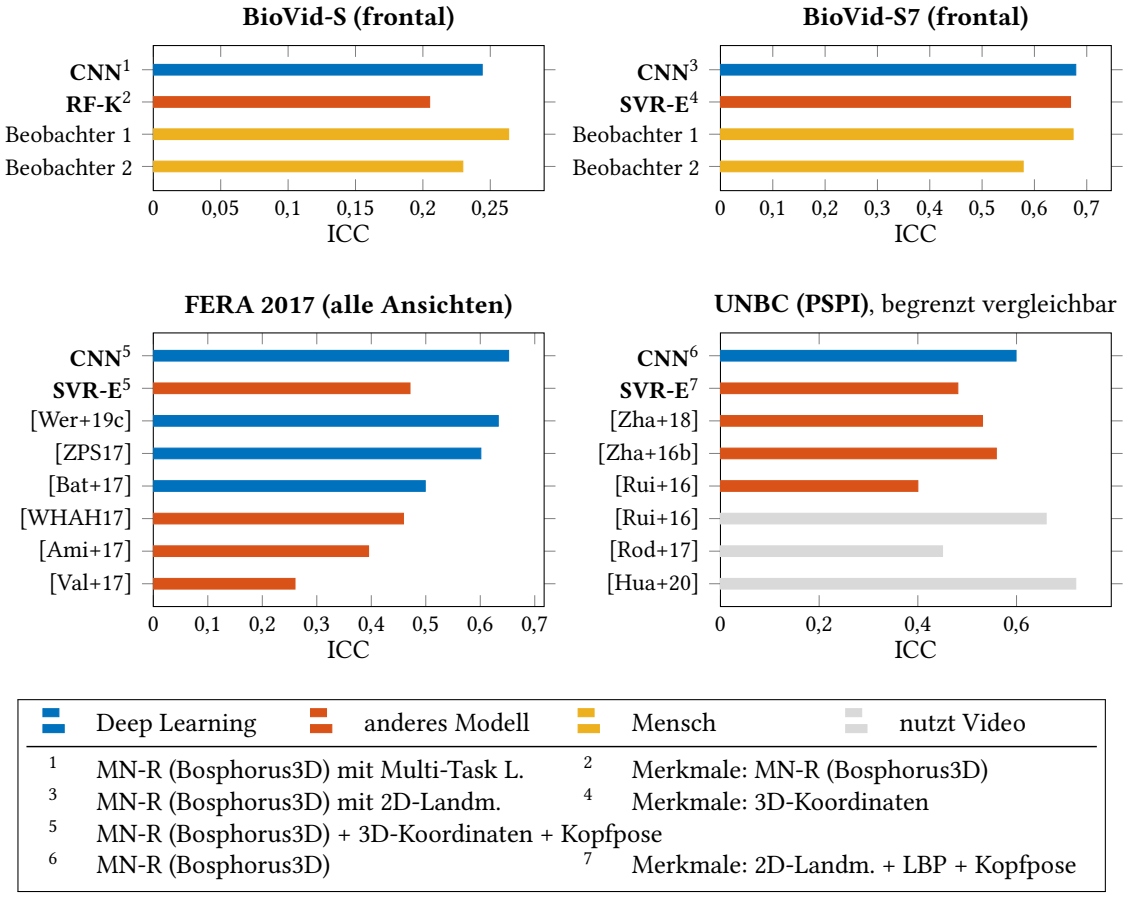


Abbildung 3.12.: Vergleich mit verwandten Arbeiten und Performance des Menschen mit den Datensätzen BioVid-S, BioVid-S7, FERA 2017 und UNBC. Verglichen werden die besten Ergebnisse mit und ohne CNN als Lernverfahren. Die Ergebnisse auf UNBC sind nur begrenzt vergleichbar, siehe Diskussion in [Wer+19b].

3.4. Diskussion

In diesem Kapitel wurden ausgehend von den Grundlagen und dem Stand der Technik verschiedene Methoden zur einzelbildbasierten Erkennung von Schmerzmimik und Action Units vorgeschlagen und evaluiert. Diese betreffen Lernverfahren und Prädiktionsmodelle, Merkmalsextraktion, Gesichtsnormierung, Landmarkenlokalisierung und Bildaufnahme.

Abb. 3.12 stellt die Performances, die mit den vorgeschlagenen Methoden in den Experimenten erreicht wurden, gegenüber mit der Performance menschlicher Beobachter und verwandter Arbeiten. Betrachtet werden (1) die fontalen Datensätze BioVidS und BioVidS-7, für die in Abschnitt 2.3.5 die Performance zweier menschlicher Beobachter bestimmt wurde, (2) FERA 2017, für das durch die Challenge [Val+17] und das dort vorgegebenen Evaluierungsprotokoll sehr gut vergleichbare Ergebnisse anderer Arbeiten vorliegen, sowie (3) die PSPI-Prädiktion auf UNBC, die oft verwendet wird, deren Ergebnisse jedoch aufgrund unterschiedlicher Evaluierungsprotokolle nur begrenzt vergleichbar sind (vgl. Abschnitt 3.1.3 und [Wer+19b]).

Das beste einzelbildbasierte gelernte Modell ist auf allen vier Datensätzen das in Abschnitt 3.2.4 vorgeschlagene Transferlernen mit MobileNetV3, vortrainiert auf dem Datensatz Bosphorus3D, der für diese Dissertation erzeugt wurde. Auf BioVid-S ist die Performance mit 0,244 besser als

Beobachter 2 (0,230), schlechter als Beobachter 1 (0,264) und sehr ähnlich zum Mittelwert beider Beobachter (0,247). Auf BioVid-S7 ist die Performance des CNN (0,679) ähnlich zu der von Beobachter 1 (0,674), jedoch besser als Beobachter 2 (0,579) und der Mittelwert der Beobachter (0,626). Insofern kann geschlussfolgert werden, dass die Performance des vorgeschlagenen CNN bei der einzelbildbasierten Unterscheidung zwischen hitzestimulierten Schmerzen an der Toleranzschwelle und keinen Schmerzen zumindest ähnlich zu der eines menschlichen Beobachters ist. Auch wenn die absoluten Performance-Werte vielleicht niedrig erscheinen mögen, zeigt dieses Ergebnis, dass Erkennungssysteme mit hohem praktischem Nutzen entwickelt wurden, da sie einen menschlichen Beobachter bei dieser Aufgabe ersetzen oder unterstützen könnten. Für FERA 2017 und UNBC ist kein Vergleich mit der menschlichen Performance möglich. Die Gegenüberstellung mit den Performances anderer Arbeiten des Standes der Technik zeigt jedoch, dass das vorgeschlagene CNN-basierte Erkennungssystem sehr gute Ergebnisse erzielt, nach dem Kenntnisstand des Autors dieser Dissertation die besten bisher publizierten Ergebnisse. Lediglich für Verfahren, die auch den zeitlichen Kontext nutzen (grau, auf UNBC), wurden bessere Ergebnisse berichtet. Auf die Nutzung des zeitlichen Kontextes wird im folgenden Kapitel 4 ausführlich eingegangen.

Die besten Ergebnisse auf BioVid-S7 und FERA 2017 wurden mit CNN-Modellen erzielt, die auch menschengemachte Merkmale ausnutzen (2D-Landmarken bei BioVid-S7 bzw. 3D-Koordinaten und Kopfpose bei FERA). Insbesondere bei FERA, aber auch mit anderen Datenbanken, gab es signifikante Verbesserungen durch die Fusion verschiedener Merkmalsarten. Jedoch waren die Verbesserungen mit keiner der Merkmalsarten datensatzübergreifend konsistent. Bei allen Datensätzen, abgesehen von BioVid-S7, waren die Ergebnisse mit feinjustierten CNNs, insbesondere ausgehend von Bosphorus3D, besser als mit Support-Vector-Verfahren und Random Forests. BioVid-S7 profitiert mit nur 280 Samples sehr wenig von den Vorteilen von CNNs. Für derart kleine Datensätze und auch zur Fusion mit CNNs ist die Entwicklung und Evaluierung von Merkmalsextraktionsmethoden noch immer relevant, trotz des Erfolges von CNNs und ihrer optimierten Merkmalsextraktion. Sehr gute Ergebnisse konnten beispielsweise mit den vom Autor dieser Dissertation entwickelten 3D-Koordinaten erzielt werden. Im Vergleich von Support-Vector-Ensembles und Random Forests schnitten die Support-Vector-Verfahren besser ab. Das Multi-Task-Lernen mit Bosphorus3D hat im Mittel einen kleinen, nicht signifikanten Vorteil in der Performance gebracht. Es eignet sich in vielen Fällen als ein (zusätzliches) Verfahren zur Regularisierung, hat jedoch einen geringeren Effekt als das Transferlernen mit Bosphorus3D (Feinjustierung des CNN). Regression war bei Datensätzen mit mehr als zwei Intensitätsstufen der Klassifikation überlegen. Bei binären Probleme waren die Ergebnisse ähnlich. Abgesehen von der hier ermittelten quantitativen Performance, die lediglich die Granularität der Grundwahrheit abbilden kann, ist die Regression im Hinblick auf die videobasierte Erkennung im nächsten Kapitel vielversprechend. Die Ausgabewerte der Regression sind feingranularer aufgelöst als die Klassifikationsergebnisse, entsprechen damit eher der kontinuierlichen Natur der modellierten Mimik und ermöglichen die Berechnung von Dynamikinformationen wie der Geschwindigkeit einer Bewegung.

Zu Behandlung der Ungleichverteilung der Klassenzugehörigkeiten wurde das Verfahren MID vorgeschlagen und evaluiert. Es ermöglicht über den Hyperparameter α die Optimierung der Performance für einen spezifischen Datensatz. Wie der Vergleich von $\alpha = 0$ (ohne MID) und $\alpha > 0$ (mit MID) in Abb. 3.10 zeigt, hat das Verfahren einen positiven Einfluss auf die erreichten Performances. Dies gilt auch für die übrigen Ergebnisse, bei denen MID mit $\alpha = 0{,}5$ eingesetzt wurde und die zum Teil durch das mittels MID verbesserte, auf Bosphorus3D vortrainierte Modell profitieren konnten.

Zur Gesichtsnormierung wurde die Methode FaNC entwickelt und evaluiert. Die Motivation war, eine möglichst gute Mimikerkennung auch bei nicht-frontalen Kopfposen zu erreichen, die in den meisten Datensätzen unterrepräsentiert sind oder zum Teil gar nicht vorkommen. Anders gesagt sollte eine gute Generalisierung auf im Training nicht verwendeten Kopfposen erreicht werden. Für die Erkennung von AUs und verschiedene Emotionskategorien konnte dieses Ziel mit FaNC erreicht werden, denn hier wurden bessere Ergebnisse erzielt als mit anderen Methoden zur Gesichtsnormierung, jedoch *nicht* für Schmerzen. Da Schmerzmimik in den Trainingsdaten von FaNC nicht vorkam und (mit realisierbarem Aufwand) nicht hinzugefügt werden konnte, kam es bei der Frontalisierung von Schmerzmimik häufiger zu starken ungewollten Verzerrungen des Bildes, welche die Mimikerkennung negativ beeinflusst haben. Als beste Option, um Kopfposeinvarianz für beliebige Mimik zu erreichen, hat sich die Verwendung von Datensätzen mit mehreren Ansichten auf das Gesicht (engl. multi-view) bewährt. Beim Training mit diesen Datensätzen sind auch einfache Gesichtsnormierungsverfahren hinreichend, insbesondere im Zusammenspiel mit CNNs, die mit ihrer hohen Kapazität auch innerhalb von Klassen einen großen Variantenreichtum modellieren können. Für die Entwicklung eines Systems, das später auch mit wenigen Kameras gut funktioniert, ist es insofern optimal, Trainingsdaten mit vielen Kameras (verschiedenen Ansichten auf das Gesicht) aufzuzeichnen, um kopfposeinvariante Erkennungsmodelle trainieren zu können. Die experimentellen Ergebnisse legen nahe, das eine Abdeckung des Kopfposeraumes in Winkelschritten von etwas 40° hinreichend ist, um auch auf dazwischen liegende Posen zu generalisieren. Eine Alternative, die mit Bosphorus3D realisiert wurde und auch dem Datensatz FERA 2017 zugrunde liegt, ist die Verwendung von 3D-Daten zum Synthetisieren von verschiedenen Ansichten. Die resultierenden Bilder enthalten jedoch aufgrund von Messungenauigkeiten in den 3D-Daten Artefakte (unrealistische Verzerrungen), welche die Performance eines Erkennungssystems negativ beeinflussen können. Insofern ist eine echte multi-view-Datenbank der aus 3D-Daten gerenderten Variante vorzuziehen.

Bezüglich der Datenaufnahme konnte gezeigt werden, dass normale 2D-Kameras für den Testfall, d. h. die Anwendung des Systems, ausreichen. Es sind weder 3D-Informationen noch eine sehr hohe Auflösung nötig. Die Untersuchungen haben gezeigt, dass die Mimikerkennung mit Gesichtsbildern mit einem Augenabstand von 50 oder mehr Pixeln ähnlich gut funktioniert. Starke Performance-Einbußen gab es erst bei unter 25 Pixeln Augenabstand. Durch die Verbesserung der Poseinvarianz könnte so z. B., wie in bei der X-ITE Database mit der seitlichen Kamera und dem Spiegel umgesetzt, eine Schmerzüberwachung mit nur *einer Kamera* gelingen. Diese hat, wenn die Sicht nicht durch eine andere Person verdeckt ist, einen direkten oder durch den Spiegel indirekten Blick auf das Gesicht des Probanden bzw. Patienten, unabhängig davon wie dieser seinen Kopf dreht.

Nicht-frontale Kopfposen stellen für Modelle zur Landmarkenlokalisierung eine Herausforderung dar. Mit dem auf mehreren Datensätzen trainierten Modell (vgl. Abschnitt 3.2.1) konnten gute Ergebnisse erzielt werden, sowohl bezüglich der Gesichtsnormierung auf Basis der Landmarken, als auch bezüglich der landmarkenbasierten Merkmale. Die Ergebnisse ohne Nutzung von Landmarken (BBox und Raw in Abb. 3.7 und 3.9) deuten jedoch darauf hin, dass mit CNN und einfacher Gesichtsnormierung anhand der Gesichtsdetektion auch ohne die Lokalisierung von Landmarken ähnlich gute Ergebnisse erzielt werden können.

4. Videobasierte Erkennung

In diesem Kapitel wird, aufbauend auf der einzelbildbasierten Schmerzerkennung des vorherigen Kapitels, eine videobasierte Schmerzerkennung realisiert. Diese nutzt mehrere nacheinander aufgenommener Bilder und kann daher Informationen über Bewegungen extrahieren, wie die Stärke einer Veränderung, die Dauer oder die Geschwindigkeit. Durch diese Zusatzinformationen wird eine bessere Erkennungsleistung als bei einzelbildbasierten Systemen erwartet.

Ausgehend von der Vorstellung der verwandten Arbeiten in Abschnitt 4.1 werden verschiedene Ansätze zur zeitlichen Integration, d. h. dem Zusammenführen von Bildinformationen über die Zeit, vorgeschlagen (siehe Abschnitt 4.2). Es folgen Experimente zur Erkennung starker Schmerzen und der automatisierten Messung der Schmerzintensität, um die vorgeschlagenen Methoden zu evaluieren (Abschnitt 4.3). Folgende Forschungsfragen stehen im Vordergrund: (1) Welche Computer-Vision- und Machine-Learning-Methoden liefern im Kontext der Videoerkennung die beste Performance? (2) Lässt sich mit videobasierter Erkennung eine bessere Performance erreichen als mit einzelbildbasierter Erkennung? (3) Kann das entwickelte Erkennungssystem eine ähnlich gute Beurteilung der Schmerzen erreichen wie ein Mensch? Das Kapitel endet mit einer zusammenfassenden Diskussion in Abschnitt 4.4.

4.1. Verwandte Arbeiten

Videobasierte Methoden nutzen mehrere in zeitlicher Abfolge aufgenommene Bilder zur Erkennung. Sie können unterschiedlich große Zeitspannen in die Prädiktion einbeziehen und die Prädiktionen mit verschiedener Häufigkeit abgeben. Diese Aspekte werden in den meisten Forschungsarbeiten durch die zeitliche Granularität der Grundwahrheit bestimmt, d. h. dadurch, wie viele Label-Werte je Zeit bzw. Video vorhanden sind und wie groß die Zeitspannen sind, die sie jeweils beurteilen. Abschnitt 4.1.1 geht genauer auf die zeitliche Granularität ein.

Videobasierte Erkennung baut zumeist auf der einzelbildbasierten Erkennung auf und erweitert sie um die zeitliche Integration von Informationen, die auf drei Arten erfolgen kann: (1) auf Merkmalsebene mithilfe zeitlicher Deskriptoren, die unabhängig vom Lernverfahren bzw. Prädiktionsmodell extrahiert werden, (2) durch spezielle Lernverfahren und Modelle, die für Videos oder Zeitreihen entwickelt wurden, und (3) durch die Nachverarbeitung der Prädiktionsergebnisse. Diese drei Arten werden in den Abschnitten 4.1.2 bis 4.1.4 thematisiert. Sie schließen sich nicht gegenseitig aus, sondern können auch in Kombination angewendet werden. Detailliertere Ausführungen zu Vorarbeiten hat der Autor im Artikel Werner et al. [Wer+19b] veröffentlicht, auf dem die folgenden Unterabschnitte basieren.

4.1.1. Zeitliche Granularität der Grundwahrheit

Grundwahrheiten haben verschiedene zeitliche Granularität. Die in den meisten Veröffentlichungen genutzte Grundwahrheit ist PSPI, eine einzelbildbasierte Grundwahrheit, die den Schmerz zum jeweiligen Zeitpunkt des Bildes anhand der Mimik abschätzt. Neben PSPI stellt der viel

verwendete UNBC-Datensatz auch noch Grundwahrheiten zur Verfügung, die anstelle eines Einzelbildes den Schmerz während des gesamten Videos bewerten: eine Selbstbeurteilung des Patienten mit Visueller Analogskala (VAS) sowie eine Beobachtereinschätzung (engl. oberver pain rating, OPR). In der klinischen Praxis sind Grundwahrheiten mit einer solchen gröberen zeitliche Granularität, die Zeiträume oder Schmerzereignisse beurteilen, deutlich mehr anzutreffen, als einzelbildbasierte Grundwahrheiten. Letztere sind ohne automatisierte Messsysteme schwer zu erheben und sind entsprechend weniger gut validiert. Bei BioVid und X-ITE stammt die Grundwahrheit von der angewendeten Schmerzstimulation. In den bisher veröffentlichten Arbeiten und auch hier wird sie als eine Annotation je Video (oder Zeitfenster) aufgefasst. Zwar liegt die Grundwahrheit in zeitlich sehr hoher Auflösung vor, eine feingranulare Nutzung ist jedoch schwierig, da die Schmerzreaktionen zeitlich verzögert auf den Reiz folgen und einer sehr großen Varianz unterliegen.

Die meisten Arbeiten nutzen *eine* Art und Granularität von Grundwahrheit für Training und Evaluation. Typisch ist die Nutzung von PSPI als Grundwahrheit für jedes Einzelbild beim Trainieren und Testen, wobei vorangegangene und/oder nachfolgende Einzelbilder ausgenutzt werden, um die Erkennungsleistung zu verbessern, z. B. bei [EVM17; Flo+16; Rod+17; RPP15; ZPS17]. Ebenfalls in vielen Arbeiten anzutreffen (auch in dieser Dissertation) ist die Prädiktion einer Schmerzeinschätzung für ein ganzes Video, z. B. [Ash+09; Gha+14; LMRP17a; Rui+16]. Es gibt jedoch auch Arbeiten, die gezeigt haben, dass ein Modell zur Schmerzeinschätzung jedes Einzelbildes auch mit ausschließlich videobasierten Grundwahrheiten (wie OPR) gelernt werden kann. Sikka et al. [SDB14] haben einen Klassifikator mit binarisierten OPR-Videolabels trainiert und seine Prädiktion auf Einzelbildern mit binarisiertem PSPI verglichen. In ihrer Arbeit fanden sie auch eine Korrelation zwischen den Entscheidungswerten des Klassifikators und PSPI, die darauf hindeutet, dass auch die Schmerzintensität zu einem gewissen Grad geschätzt werden kann. Die zugrunde liegende Idee (Multiple Instance Learning) wurde von Ruiz et al. [Rui+16] mit Ordinal Regression kombiniert, um die Schmerzintensität auf direktem Wege zu modellieren. So konnten sie auf Einzelbildebene prädizierte Intensitäten mit PSPI validieren, obwohl lediglich die videobasierte Grundwahrheit OPR für das Training genutzt worden war.

Im Allgemeinen muss ein Erkennungsmodell nicht zwingend mit der gleichen zeitlichen Granularität angewendet werden, mit der gelernt wurde. Jedes videobasierte Modell kann mit einem gleitenden Zeitfenster (engl. sliding window) angewendet werden, um eine zeitlich kontinuierliche Intensitätsschätzung mit einer beliebigen gewünschten Wiederholrate zu erhalten, auch wenn es mit einer Grundwahrheit mit gröberer zeitlicher Granularität angelernt wurde. Dies wurde mit der BioVid-Datenbank gezeigt [Käc+15b; Käc+16; Käc+17; AKS16]. Z. B. haben Amirian et al. [AKS16] mit Zeitfenstern trainiert, die zeitlich an den Schmerzreizen ausgerichtet waren, und anschließend mit einem gleitenden Zeitfenster über das gesamte Schmerzexperiment hinweg prädiziert.

4.1.2. Zeitliche Deskriptoren

Eine Möglichkeit zur zeitlichen Integration ist die Berechnung von zeitlichen Deskriptoren. Dabei handelt es sich um Merkmalsvektoren, die Informationen mehrerer Bilder, eines Zeitfensters oder eines ganzen Videos zusammenfassen, und mit einem beliebigen Klassifikations- oder Regressionsverfahren kombiniert werden können.

Ausgangspunkt von zeitlichen Deskriptoren ist typischerweise die Merkmalsextraktion beim Einzelbild. Jeder Texturdeskriptor, wie LBP [AHP06], kann in die Zeitdomäne erweitert werden, indem das Prinzip der drei orthogonalen Ebenen (engl. Three Orthogonal Planes, TOP) angewendet

wird. Anstatt nur Merkmale von der räumlichen x-y-Ebene zu extrahieren wird die selbe Methode in den raum-zeitlichen Ebenen x-t und y-t angewendet. Die resultierenden drei Merkmalsvektoren werden verkettet. Verschiedene Arbeiten zur Schmerzerkennung nutzen LBP-TOP [Käc+15b; Yan+16; Käc+17; Thi+16; TS17; TKS17; KTP16], HOG-TOP [CCF17] und LGBP-TOP (Local Gabor Binary Pattern TOP) [TKS17].

Ein alternativer Ansatz ist es, jedes einzelne Einzelbildmerkmal als Zeitreihe aufzufassen. Anschließend kann ein Deskriptor berechnet werden, indem für jede Zeitreihe Statistikparameter extrahiert werden. Die einfachste Variante ist es, einen Parameter je Zeitreihe zu extrahieren, z. B. das Maximum [TKS17; Thi+16; Liu+18]. Dies entspricht *Max. Pooling*, wie es in CNN häufig eingesetzt wird, in der zeitlichen Dimension. Meist werden mehrere statistische Maße extrahiert, z. B. Mittelwert und Quartile [Sik+15]. Littlewort et al. [LBL09] extrahieren 5 Maße, Xu et al. [Xu+19] 9 Maße, Kächele et al. [Käc+15b; Käc+17] und Liu et al. [Liu+17] 10 Maße, und Tsai et al. [Tsa+16] 15 Maße je Einzelbildmerkmal. In einer Vorarbeit dieser Dissertation hat ihr Autor bereits 2013 vorgeschlagen, Statistiken auch von der ersten und zweiten Ableitung der Zeitreihe zu extrahieren [Wer+13], was später auch für andere Arbeiten übernommen wurde [Wer+14b; Käc+15a; Thi+16; TS17; TKS17; Kes+17; Oth+19b; Oth+21].

Weitere Möglichkeiten für zeitliche Deskriptoren basieren auf Histogrammen [Gha+14; Tsa+16; Bar+14]. Tsai et al. [Tsa+16] nutzen einen Bag-of-Words-Deskriptor, d. h. sie berechnen beim Training k Cluster im Merkmalsraum, die k „Wörter“ (Zustände) repräsentieren, und codieren jedes Video als k-dimensionales Histogramm der Häufigkeiten der Zustände. Bartlett et al. [Bar+14] haben den *Bag of Temporal Features (BoTF) Deskriptor* vorgeschlagen. Zur Berechnung wird ein Satz von zeitlichen Gabor-Filtern auf die Merkmalszeitreihen angewendet. Für jede Filterantwort werden die Flächen von positiven und negativen Segmenten (Abschnitte der Kurve über bzw. unter der Nulllinie) berechnet. Die Flächengrößen werden in 6 Histogrammklassen (engl. bins) eingeteilt sowie die Häufigkeiten der Flächengrößen der positiven und negativen Segmente in gesonderten Histogrammen gezählt. Der BoTF-Deskriptor ergibt sich durch die Verkettung der positiven und negativen Histogramme aller Gabor-Filter und Merkmalszeitreihen. Als Merkmale wurden von Bartlett et al. [Bar+14] Action Units (AUs) eingesetzt, die von zuvor trainierten Modellen extrahiert wurden. Auch in anderen Arbeiten werden AUs als Merkmale genutzt, um in einem nachfolgend angewendeten Modell Schmerzen zu erkennen oder zu bewerten [LBL09; Bar+14; Sik+15; Gha+14; Liu+18].

4.1.3. Spezialisierte Lernverfahren

Im Folgenden werden Lernverfahren vorgestellt, die speziell zur zeitlichen Integration entwickelt oder angepasst wurden. Die meisten dieser Verfahren basieren entweder auf neuronalen Netzen oder auf probabilistischen graphischen Modellen. Ein Ansatz ist es, mehrere Bilder des Videos als Eingabe für ein CNN zu verwenden. Othman et al. [Oth+19b; Oth+21] nutzen drei Bilder, die in Graustufen konvertiert und zu einem RGB-Bild zusammengesetzt werden. Dies ermöglicht Transferlernen mit CNN-Modellen, die für die Objekterkennung mit ImageNet vortrainiert wurden, für die videobasierte Schmerzerkennung. Egede et al. [EVM17] extrahieren Bildmerkmale mit Dynamikinformationen für die Schätzung von PSPI, indem zusätzlich zum eigentlichen Bild zwei vorangegangene und zwei nachfolgende Bilder in das CNN eingegeben werden.

Neben CNN werden oft auch rekurrente Netze (engl. Recurrent Neural Networks, RNNs) eingesetzt, meist mit Long Short-Term Memory Unit (LSTM) [HS97], bidirektionalen LSTM oder Gated Recurrent Unit (GRU) [Chu+14]. Die RNNs werden zumeist nach CNNs oder anderen neuronalen Netzen angewendet, welche die Dimensionalität der Ursprungsdaten deutlich reduzieren. Liu et al. [Liu+17] nutzen ein LSTM-RNN als Baseline für die Schmerzerkennung, um eine

Zeitreihe schwacher Prädiktionen zu einer Prädiktion für das gesamte Video zu kombinieren. Ein CNN zur Merkmalsextraktion aus 16 Einzelbildern wird von Rodriguez et al. [Rod+17] mit LSTM kombiniert, um den zeitlichen Kontext für eine verbesserte PSPI-Schätzung auszunutzen. Lopez-Martinez et al. [LMRP17a] verarbeiten die Landmarken aus Zeitfenstern von 15 Bildern mit einem bidirektionalen LSTM-RNN, ebenfalls um PSPI zu schätzen. Die PSPI-Werte werden anschließend mithilfe eines probabilistischen graphischen Models verrechnet, um die Schmerzintensität des Videos in der VAS zu schätzen. Thiam et al. [TKS20] schlagen eine komplexe Architektur vor, bestehend aus zwei parallelen Zweigen zur Videoklassifikation, deren Entscheidung am Ende fusioniert wird. Beide Zweige haben den gleichen Aufbau aus CNNs zur Extraktion von Bildmerkmalen, bidirektionalem LSTM zum zeitlichen Austausch von Informationen, einer Aufmerksamkeitsschicht, die Ausgaben der LSTMs gewichtet und aufaddiert, sowie einem neuronalen Netz zur Klassifikation. Die Eingabe der beiden Zweige sind jedoch nicht die Originalbilder des Videos, sondern Bildrepräsentationen der Veränderung (Motion History Images und Optical Flow Images). Die Aufmerksamkeitsschicht berechnet die Gewichte für die LSTM-Ausgabemerkmale mithilfe eines Fully Connected Layer, der die LSTM-Ausgabemerkmale als Eingabe erhält. Auch Huang et al. [Hua+20] nutzen eine zeitliche Aufmerksamkeitsschicht, nach einem RNN (hier mit GRU anstelle von LSTM) und einem CNN. Eingabe für das CNN sind hier jedoch die registrierten Farbbilder des Gesichts. Außerdem ist in das CNN auch ein räumliches Aufmerksamkeitsmodul integriert. Zhou et al. [Zho+16] verwenden als einzige Arbeit ein rekurrentes CNN ohne LSTM oder GRU, bei dem die rekurrenten Verbindungen zum vorherigen Zeitschritt bei den Convolutions angewendet werden. Ein großer Vorteil von LSTM und GRU ist, dass mit ihrer Hilfe auch Abhängigkeiten über viele Zeitschritte hinweg in der Praxis gut modelliert werden können.

Neben neuronalen Netzen werden in der Schmerzerkennung oft probabilistische graphische Modelle eingesetzt, um zeitliche Informationen zu kombinieren. Neben Hidden Markov Models [MBB14] sind das vor allem verschiedene Varianten von Conditional Random Fields [Gha+14; RPP13b; RPP15; Rui+16; Liu+17; LMRP17a]. Rudovic et al. [RPP13b; RPP15] und Ruiz et al. [Rui+16] konnten die Klassifikation von Schmerzintensitäten verbessern, indem sie die Ordnungsrelation zwischen den Intensitätsklassen in verschiedenen Varianten von Conditional Ordinal Random Field (CORF) Modellen berücksichtigt haben.

Hammal und Kunz [HK12] realisieren die zeitliche Fusion mit dem regelbasierten Transferable Belief Model, Sikka et al. [SDB14] und Ruiz et al. [Rui+16] mit Multiple Instance Learning. Szczapa et al. [Szc+21] nutzen Support Vector Regression (SVR) mit einem speziellen Kernel, dem Global Alignment Kernel, der die Ähnlichkeit zweier Trajektorien berechnet. Zuvor werden Koordinaten und Geschwindigkeiten der Landmarken über Gram-Matrizen auf Trajektorien in einer Riemannian Manifold abgebildet.

4.1.4. Nachverarbeitung der Prädiktion

Eine weitere Möglichkeit zur zeitlichen Integration ist die Nachverarbeitung der Prädiktion eines Modells. Genauer gesagt, werden hierbei Zeitreihen von Ausgaben verarbeitet, die ein Modell für mehrere Einzelbilder oder Zeitfenster gemacht hat.

Eine Variante ist, was hier im Folgenden *Entscheidungsfusion* genannt wird: Für jedes der Einzelbilder eines Videos liefert ein Modell einen Entscheidungswert. Über die Zeitreihe der Entscheidungswerte wird nun ein Mittelwert gebildet, um *eine* Gesamtentscheidung für das Video zu treffen. Ashraf et al. [Ash+09] haben diesen Ansatz in einer sehr frühen Arbeit zur Schmerzerkennung mit SVM angewendet. Bei Liu et al. [Liu+17] dient der Ansatz in Kombination mit einem neuronalen Netz als Baseline.

Die Nachverarbeitung von Prädiktionen wurde auch angewendet, um die Erkennungsleistung auf Einzelbildebene zu steigern: Florea et al. [Flo+16] haben die Schmerzintensität, die von ihrem Modell ausgegeben wurde, zeitlich gefiltert, um sie zu glätten und Artefakte zu entfernen, die durch Lidschlag ausgelöst wurden. Um personenspezifische Biases zu reduzieren haben Egede et al. [EVM17] vorgeschlagen, den Modalwert aller prädizierten Schmerzintensitäten von den Prädiktionen abzuziehen. Der Methode liegt die Annahme zugrunde, dass das Gesicht einer Person die meiste Zeit über neutral ist und so die Prädiktion des entspannten Gesichtes ohne Schmerzmimik abgezogen wird.

4.2. Vorgeschlagene Methodik

Im Folgenden werden verschiedene Methoden zur zeitlichen Integration von Einzelbildinformationen für die Schmerzerkennung anhand von Videos vorgeschlagen. Sie lassen sich in zwei Kategorien einteilen: (1) Methoden, bei denen die Merkmalsextraktion feststehend und unabhängig von der Klassifikation bzw. Regression ist, und (2) CNN-Methoden mit Ende-zu-Ende-Training, bei denen die Merkmalsextraktion gemeinsam mit der Klassifikation bzw. Regression anhand der Trainingsdaten optimiert wird.

In diesem Kapitel werden videobasierte Grundwahrheiten verwendet, d. h. die Schmerzreizung, VAS und OPR (vgl. Abschnitt 2.1). Wie schon mehrfach erwähnt ist die relativ geringe Größe der verfügbaren Schmerzerkennungsdatensätze eine Herausforderung für das maschinelle Lernen. Diese verschärft sich bei Betrachtung von Videos anstelle von Einzelbildern noch weiter, da höherdimensionale Informationen und gleichzeitig weniger Samples zur Verfügung stehen (z. B. bei UNBC nur 200 Videos und VAS/OPR-Label). Zur Handhabung dieser Problematik wird auf den Ergebnissen der Einzelbilderkennung aufgebaut: Zum einen werden bewährte niedrigdimensionale Einzelbildmerkmale als Basis für die zeitliche Integration mit unabhängiger Klassifikation bzw. Regression (mit Support-Vector-Methoden bzw. Random Forests) genutzt. Zum anderen wird Transferlernen eingesetzt, um CNNs zur videobasierten Erkennung ausgehend von Einzelbild-CNNs zu lernen, insbesondere ausgehend vom Vortraining mit Bosphorus3D zur Regression der Intensitäten von Action Units (AUs).

Abb. 4.1 gibt einen Überblick über die wesentlichen Varianten, die vorgeschlagen und untersucht werden. Bei (a) und (b) sind die Extraktion der Einzelbildmerkmale, die zeitliche Integration sowie die Klassifikation bzw. Regression unabhängige Komponenten. Bei (c)-(e) werden die Merkmalsextraktion, die zeitliche Integration und die Prädiktion gemeinsam (Ende-zu-Ende) in einem CNN trainiert. In Abschnitt 4.2.1 wird auf die hier verwendeten Einzelbildmerkmale eingegangen, die im Videokontext Zeitreihen bilden. Abschnitt 4.2.2 schlägt einen zeitlichen Deskriptor für die Variante (a) vor, der Statistikdeskriptor genannt wird. Die Entscheidungsfusion der Varianten (b) und (c) wird in Abschnitt 4.2.3 detailliert besprochen. In Abschnitt 4.2.4 bzw. 4.2.5 werden die CNN-Varianten (d) bzw. (e) vorgeschlagen. Von (b), (c) und (d) werden jeweils zwei Varianten der zeitlichen Integration betrachtet, zum einen anhand des Mittelwertes, zum anderen anhand des Maximalwertes. Für die Schätzung der Schmerzintensität wird in Abschnitt 4.2.6 außerdem eine spezielle Gewichtung der Intensitätsklassen präsentiert.

4.2.1. Betrachtete Zeitreihen

Jedes Einzelbild kann mit Methoden, die in Kapitel 3 beschrieben wurden, durch einen Punkt in einem d-dimensionalen Merkmalsraum repräsentiert werden. Wird nun ein Video mit n Einzelbildern betrachtet, ergibt sich eine Trajektorie von n Punkten in diesem d-dimensionalen Raum.

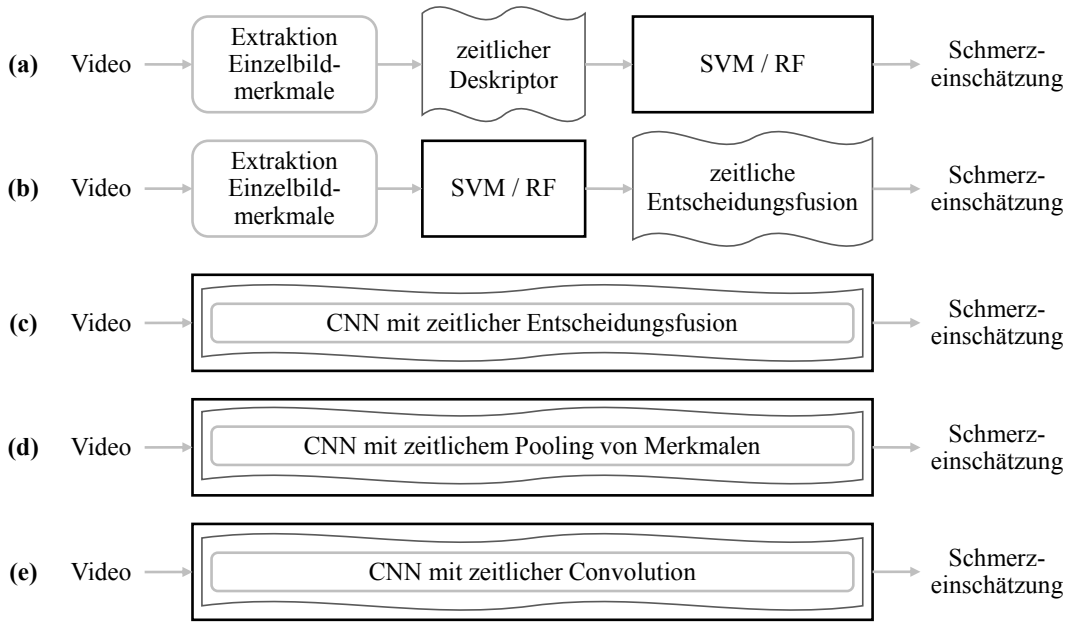


Abbildung 4.1.: Methoden zur videobasierten Schmerzerkennung.

Diese kann auch durch eine Matrix mit $d \times n$ Einträgen geschrieben werden. Jede der d Zeilen in der Matrix ist ein n-dimensionaler Vektor, der die zeitliche Entwicklung eines konkreten skalaren Merkmals des Merkmalsraums beschreibt. Im Folgenden werden diese Vektoren Zeitreihen genannt. Abb. 4.2 veranschaulicht einige Zeitreihen für eine beispielhafte Schmerzstimulation der Datenbank BioVid. Die abgebildeten Zeitreihen stammen aus drei verschiedenen Merkmalsräumen bzw. von drei Verfahren zur Merkmalsextraktion, deren Namen links an den vertikalen Achsen stehen. Auf diese und einen weiteren hier verwendeten Merkmalsraum wird im Folgenden genauer eingegangen:

MN-R Bos3D CNN AU: Das CNN MobileNetV3-large wurde im vorherigen Kapitel mit dem Datensatz Bosphorus3D zur Regression von AU-Intensitäten trainiert, vgl. die Abschnitte 3.2.4 und 3.3.3. Es prädiziert 26 AUs, unter anderem alle AUs, für die in Studien Zusammenhänge mit Schmerzen gefunden wurde (AU 4, 6, 7, 9, 10, 43, 1, 2, 12, 20, 25, 26, 27). Die vollständige Liste der prädizierten AUs sowie deren Test-Performance sind im Anhang in Tabelle B.4 zu finden. Abb. 4.2 zeigt oben die prädizierten Zeitreihen für vier AUs. Neben den konkreten Werten der einzelnen Bilder zeigen die Plots Dynamikaspekte, wie die Geschwindigkeit oder Dauer einer Bewegung. Im Zusammenhang mit den zugehörigen Bildern des Gesichts ermöglichen sie auch die qualitative Prüfung der AU-Prädiktionen. Insgesamt werden das Schließen der Augenlider (AU 43, gelb) und die starke Schmerzmimik in den Sekunden 2,5 bis 3,5 gut erkannt. Es zeigen sich jedoch auch Schwächen im AU-Modell. Z. B. sinkt AU 43 um Sekunde 3 ab, obwohl die Augen voll geschlossen bleiben, vermutlich weil in den Trainingsdaten (Bosphorus3D) keine stark zusammengekniffenen Augen vorkommen. Auch das Absinken von AU 6 bei voll geschlossenen Augen um 0,5 s, 3,5 s und 4,5 s hängt sicherlich damit zusammen, dass in den Trainingsdaten die Kombination von AU 43 und AU 6 nicht vorkommt. Hier zeigt das Modell Verbesserungspotential für zukünftige Arbeiten. Vielversprechend wäre, den Trainingsdatensatz Bosphorus3D um AU-annotierte Schmerzdaten zu ergänzen.

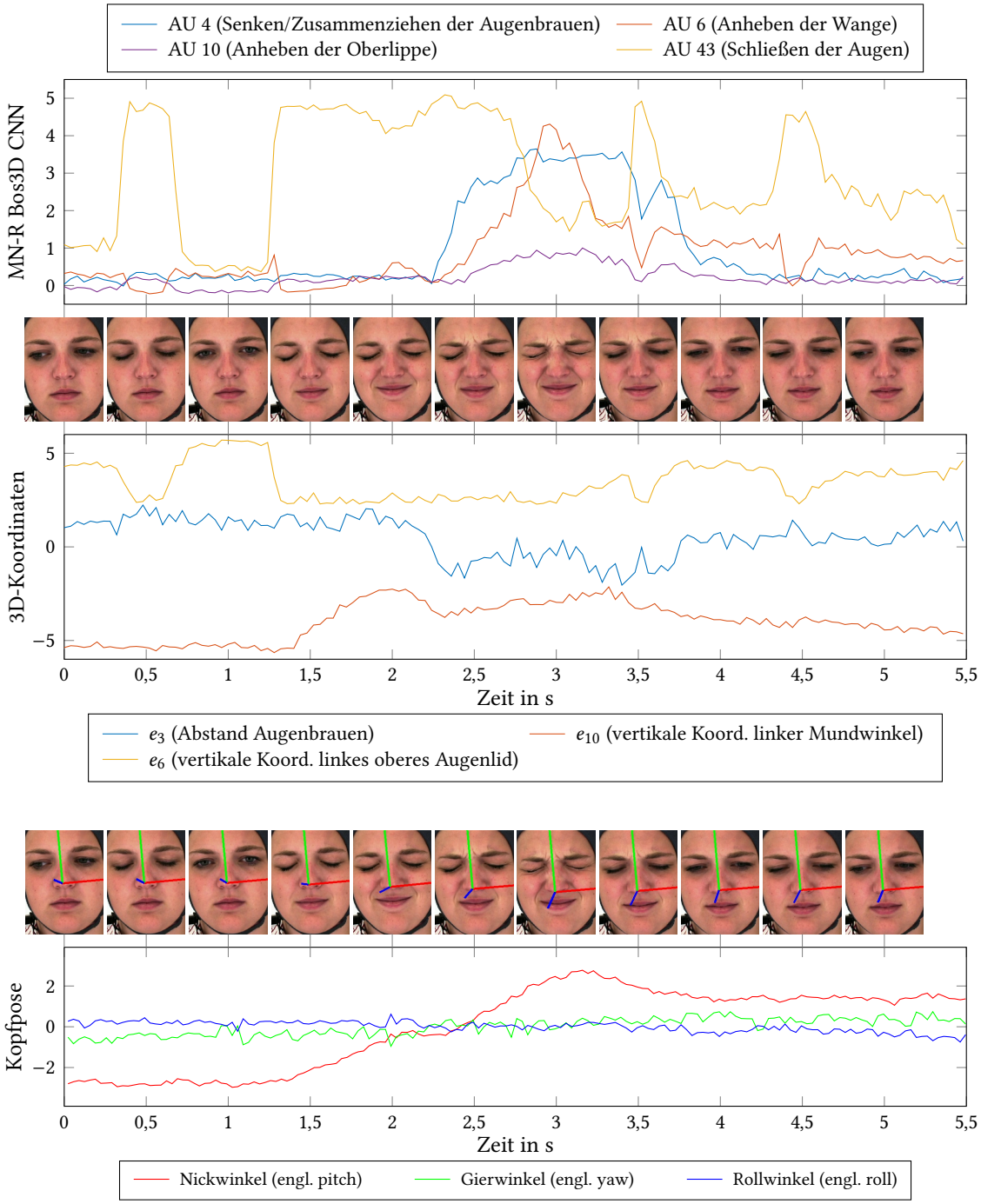


Abbildung 4.2.: Veranschaulichung einiger Zeitreihen von Einzelbildmerkmalen mit zugehörigen Bildern für ein Beispiel des Datensatzes BioVid-A mit starken Schmerzen (PA4): Im oberen Plot werden 4 der 26 AU-Intensitäten dargestellt, die mit dem CNN MN-R Bos3D prädiziert wurden. Der mittlere Plot zeigt 3 der 17 3D-Koordinatenmerkmale, die mithilfe eines 3D-Modells als Merkmale extrahiert wurden, vgl. Abb. 3.3d. Zur Verbesserung der Darstellung wurden die Zeitreihen standardisiert und in y-Richtung verschoben. Unten wird die Kopfrotation veranschaulicht (Winkel in °, vgl. Abb. 3.2b).

OpenFace AU: Zum experimentellen Vergleich mit dem vorgeschlagenen Modell MN-R Bos3D wird die AU-Erkennung der sehr oft verwendeten Open Source Bibliothek OpenFace [BRM16] herangezogen. Sie prädiziert die Intensität von 14 AUs, die alle als Einzelbildmerkmale verwendet werden: AU 1, 2, 4, 5, 6, 9, 10, 12, 14, 15, 17, 20, 25 und 26.

3D-Koordinaten: Als Vertreter der menschengemachten Merkmale werden die in Abschnitt 3.2.3 bzw. Abb. 3.3d vorgeschlagenen 3D-Koordinaten untersucht. Diese haben bei der einzelbildbasierten Mimikerkennung (vgl. Tabelle 3.2) im Mittel am besten abgeschnitten und haben mit nur 17 Dimensionen einen für die videobasierte Erkennung nicht zu großen Merkmalsraum (im Vergleich zu z. B. LBP mit 5.900 Dimensionen). Abb. 4.2 zeigt mittig drei der 17 Zeitreihen. Der Abstand der Augenbrauen e_3 (blau) eignet sich um das Senken/Zusammenziehen der Augenbrauen (AU 4) zu erfassen. Die vertikale Koordinate des linken oberen Augenlides e_6 (gelb) bildet das Schließen der Augen ab, ebenso wie AU 43. Das Anheben der Wange (AU 6) korreliert der vertikalen Koordinate des linken Mundwinkels e_{10} (orange).

Kopfpose: Die 3D-Kopfrotation und 3D-Kopfposition (insgesamt 6-dimensional, vgl. Abb. 3.2b) werden als ergänzende Merkmale untersucht. Wie der Autor dieser Dissertation in einer Studie mit drei Datensätzen gezeigt hat [Wer+18], unterscheiden sich Kopfbewegungen und -posen bei Schmerzen und ohne Schmerzen. Die Unterschiede treten jedoch weniger konsistent auf und sind weniger stark als bei der Mimik, so dass die Kopfpose insbesondere in Fusion mit Mimikmerkmalen Vorteile bringt, da sie zusätzliche Informationen liefert.

4.2.2. Erkennung mit zeitlichem Statistikdeskriptor

Eine Möglichkeit zur zeitlichen Integration ist es, jede der Zeitreihen von Einzelbildmerkmalen durch einen zeitlichen Deskriptor zusammenzufassen. Das Resultat ist ein Merkmalsvektor zur Repräsentation des gesamten Videos, der mit beliebigen maschinellen Lernverfahren weiterverarbeitet werden kann. Der hier vorgeschlagene sogenannte Statistikdeskriptor wurde vom Autor dieser Dissertation erstmals 2017 in Werner et al. [Wer+17] veröffentlicht. Er ergänzt eine bereits 2013 vom Autor vorgeschlagene Variante [Wer+13; Wer+14b] um weitere Statistiken.

Zur Berechnung des Statistikdeskriptors einer Zeitreihe wird diese zunächst zeitlich geglättet, indem ein Butterworth-Filter erster Ordnung mit einer Grenzfrequenz von 1 Hz angewendet wird. Abb. 4.3 veranschaulicht diesen Schritt oben für eine beispielhafte Zeitreihe $f_1, f_2, \ldots, f_n$. Anschließend werden die erste und zweite zeitliche Ableitung der geglätteten Zeitreihe anhand von Differenzquotienten abgeschätzt (Beispiel siehe Abb. 4.3). In Analogie zur Position s, Geschwindigkeit v und Beschleunigung a in der Kinematik wird die geglättete Zeitreihe $s(f)$ genannt, ihre erste Ableitung $v(f)$ und ihre zweite Ableitung $a(f)$. Im Folgenden wird $s(f)$ auch als Zustandszeitreihe, $v(f)$ als Geschwindigkeitszeitreihe und $a(f)$ als Beschleunigungszeitreihe des Merkmals f bezeichnet.

Im nächsten Schritt werden von jeder der drei Zeitreihen (Zustands-, Geschwindigkeits- und Beschleunigungszeitreihe) 16 Statistiken berechnet. Die Statistiken kodieren verschiedene Aspekte der Zeitreihen, insbesondere bezüglich zentraler Tendenz und Extremwerten, Streuung (Variabilität), Dauer und Häufigkeit. Tabelle 4.1 listet die Statistiken im Detail auf und Abb. 4.3 veranschaulicht unten einige der Statistiken am Beispiel. Die Auswahl der Statistiken ist zum Teil durch FACS-Parameter zur Beschreibung von Mimik inspiriert. Z. B. bilden bei der Zustandszeitreihe $s(f)$ die Statistiken `max` und `range` die Intensität der Mimik ab. `TGM` schätzt die Dauer einer Mimik ab und `tmax` den Zeitpunkt des Apex (bei dem die Mimik am stärksten ausgeprägt ist). Weitere

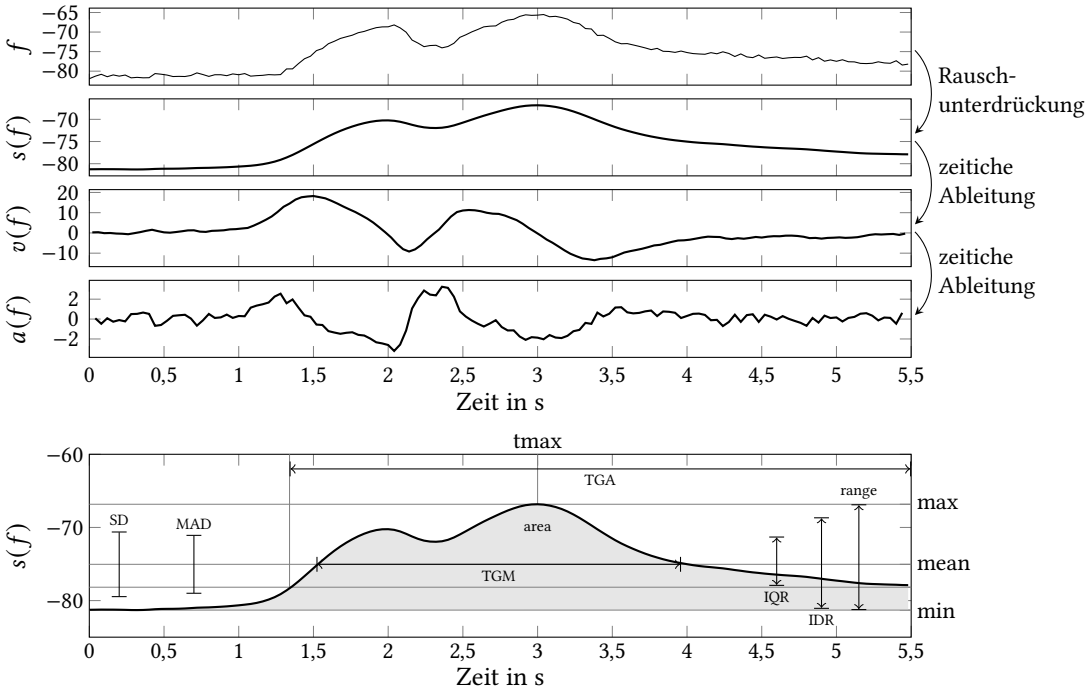


Abbildung 4.3.: Berechnung des Statistikdeskriptors für eine Zeitreihe am Beispiel: Die Zeitreihe f (oben) wird geglättet zu $s(f)$ (2. Plot) sowie die 1. Ableitung $v(f)$ und 2. Ableitung $a(f)$ bestimmt (3. und 4. Plot). Anschließend werden 16 Statistiken (vgl. Tabelle 4.1) berechnet, jeweils für $s(f)$, $v(f)$ und $a(f)$. Der unterste Plot veranschaulicht einige der Statistiken. Abb. nach [Wer+17] © 2017 IEEE.

Name	Bedeutung
`mean`	Mittelwert der Zeitreihe y
`median`	Medianwert der Zeitreihe y
`min`	Kleinster Wert der Zeitreihe y
`max`	Größter Wert der Zeitreihe y
`range`	Spannweite der Zeitreihe y, `range` = `max` − `min`
`SD`	Standardabweichung der Zeitreihe y
`IQR`	Interquartilsabstand von y (Abstand zwischen 25%- und 75%-Quantil)
`IDR`	Interdezilsabstand von y (Abstand zwischen 10%- und 90%-Quantil)
`MAD`	Median der absoluten Abweichungen (engl. median absolute deviation von y, `MAD` = median($\lvert y_i - \text{median}(y)\rvert$)
`tmax`	Zeitpunkt an dem die Zeitreihe y ihr Maximum erreicht
`TGM`	Dauer, während der die Zeitreihe größer ist als `mean`
`TGA`	Dauer, während der die Zeitreihe größer ist als der Mittelwert zwischen `mean` und `min`
`SGM`	Anzahl der Segmente (zusammenhängende Abschnitte), bei denen y größer ist als `mean`
`SGA`	Anzahl der Segmente, bei denen y größer ist als der Mittelwert zwischen `mean` und `min`
`area`	Fläche zwischen der Zeitreihe y und `min`
`areaR`	Quotient aus `area` und Fläche zwischen `max` und `min`

Tabelle 4.1.: Statistiken zur Beschreibung einer Zeitreihe. Für jede Zeitreihe $f_1, f_2, ..., f_n$ werden alle Statistiken dreimal berechnet: für das geglättete Zustandszeitreihe $y = s(f)$, für dessen 1. Ableitung $y = v(f)$ und 2. Ableitung $y = a(f)$.

Statistiken erfassen zusätzliche Informationen wie die maximale Geschwindigkeit und wann sie auftritt (`max` und `tmax` der Geschwindigkeitszeitreihe $v(f)$).

Die berechneten Statistiken der drei Zeitreihen $s(f)$, $v(f)$ und $a(f)$ aller d Einzelbildmerkmale werden verkettet, um den vollständigen Merkmalsvektor zur Beschreibung des Videos zu erhalten. Es ergibt sich ein Merkmalsraum mit $d \cdot 3 \cdot 16 = 48d$ Dimensionen. Der kleinste bzw. größte Merkmalsraum findet sich bei den Kopfpose- bzw. MN-R Bos3D AU-Merkmalen mit 288 bzw. 1.248 Dimensionen.

Neben dem selbst entwickelten Statistikdeskriptor wird auch der *Bag of Temporal Words Deskriptor* von Bartlett et al. [Bar+14] experimentell evaluiert, in Kombination mit den hier vorgeschlagenen Einzelbildmerkmalen.

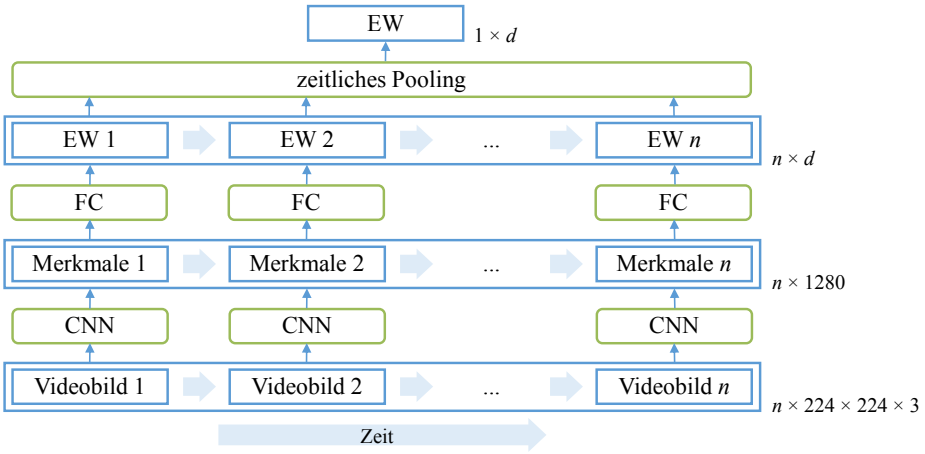
Zur Klassifikation bzw. Regression mit den Deskriptoren werde Support Vector Machine (SVM), SVM Ensemble (SVM-E), Support Vector Regression (SVR) und SVR Ensemble (SVR-E) sowie Random Forest Klassifikation bzw. Regression (RF-K bzw. RF-R) eingesetzt, wie bereits in den Abschnitten 3.2.4, 3.2.5 und 3.3 beschrieben.

4.2.3. Erkennung mit zeitlicher Entscheidungsfusion

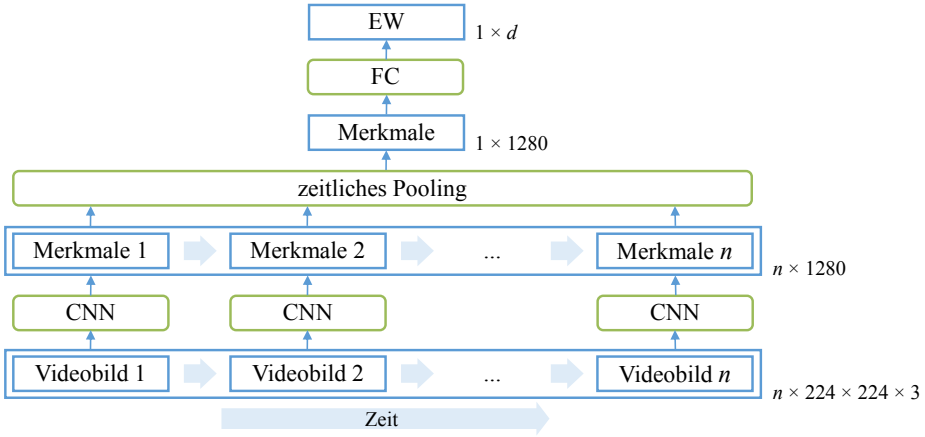
Die zeitliche Integration über mehrere Einzelbilder kann mittels Entscheidungsfusion erfolgen. Hierfür wird ein Klassifikator auf jedes Einzelbild angewendet und die Entscheidungswerte (engl. decision scores) aller Einzelbilder zusammengefasst. Im Kontext der videobasierten Schmerzerkennung haben Ashraf et al. [Ash+09] die Nutzung des SVM-Klassifikators und die Fusion anhand des Mittelwertes der Entscheidungswerte vorgeschlagen. Bei SVM ist der Entscheidungswert der vorzeichenbehaftete Abstand des Beispiels zur trennenden Hyperebene. Da das Training mit *allen* n Einzelbildern zu zeitaufwändig wäre – die Laufzeitkomplexität der SVM ist $O(n^3)$ [BL07] – wird bei Ashraf et al. die Anzahl der Beispiele je Video auf k reduziert, indem k-Means Clustering auf dem Merkmalsraum eingesetzt wird (mit $k = 20$). Beim Testen werden für jedes Video jedoch die Entscheidungswerte *aller* Einzelbilder berechnet und zusammengefasst.

Die originale Idee von Ashraf et al. wird in den Experimenten mit verschiedenen Einzelbildmerkmalen evaluiert und als „Entscheidungsfusion mit Mittelwert“ bezeichnet. Zusätzlich werden in dieser Dissertation folgende Varianten vorgeschlagen:

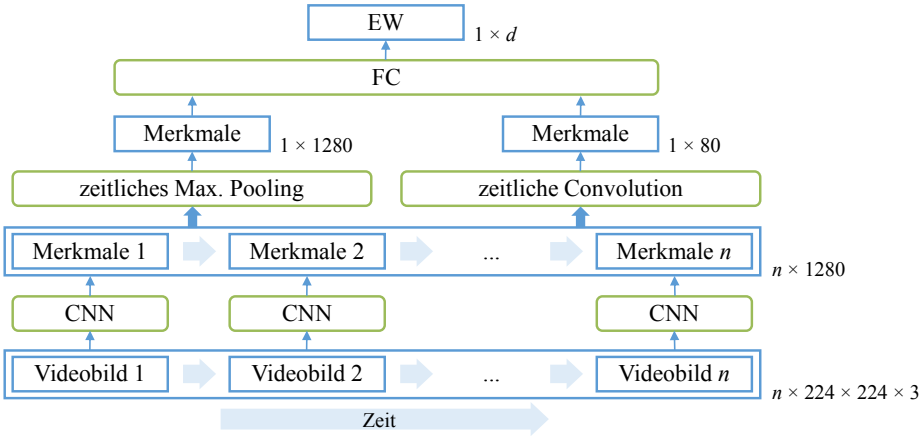
1. **Klassifikation mit SVM Ensemble:** Statt einer einzelnen SVM werden die Einzelbilder mit einem SVM Ensemble (SVM-E, siehe Abschnitt 3.2.5) klassifiziert. Der Entscheidungswert des Ensemble-Klassifikators entspricht dem Entscheidungswert des Aggregationsmodells.
2. **Klassifikation mit Random Forest (RF):** Anstelle der SVM wird ein RF mit Mehrheitsentscheid eingesetzt. Als Entscheidungswert dient der Prozentsatz der Bäume, die für die positive Klasse abgestimmt haben.
3. **Klassifikation / Regression mit CNN:** Ein Einzelbild-CNN wird für jedes Einzelbild angewendet, genauer gesagt wird es durch Parameter-Sharing und zeitliches Pooling der Einzelbild-Logits (bzw. Regressionswerte) zum Video-CNN erweitert. Abb. 4.4a veranschaulicht die Architektur. Als Basis dient wie in Abschnitt 3.3.3 ein vortrainiertes MobileNetV3-large [How+19], es können jedoch auch andere CNNs genutzt werden. MobileNetV3-large extrahiert für jedes der n Einzelbilder 1.280 Merkmale, d. h. für das Video liegen 1.280 Merkmalszeitreihen mit jeweils n Elementen vor. Bei der Entscheidungsfusion wird nun für jedes Einzelbild mittels Fully Connected (FC) Layer ein Entscheidungswert berechnet (Schmerzintensität bei Regression) bzw. d Entscheidungswerte (logits) bei



(a) Zeitliche Entscheidungsfusion mit CNN. Siehe Abschnitt 4.2.3.



(b) Zeitliches Pooling von Merkmalen. Siehe Abschnitt 4.2.4.



(c) Variante mit zeitlicher Convolution. Details siehe Abschnitt 4.2.5 und Abb. 4.5a.

Abbildung 4.4.: Vorgeschlagene Methoden zur zeitlichen Integration mit CNNs. Die n Bilder des Videos werden durch das gleiche CNN propagiert, um Merkmale zu extrahieren. Diese werden in drei verschiedenen Varianten zusammengeführt und weiterverarbeitet, um einen oder d Entscheidungswerte (EW) für das gesamte Video zu berechnen, z. B. d Logits zur Klassifikation von d Mimikintensitäten oder einen Fließkommawert zur Regression. FC = Fully Connected Layer.

der Klassifikation von d Klassen. Die Entscheidungswerte werden über die Zeit mittels Pooling kombiniert, um einen bzw. d Entscheidungswerte für das gesamte Video zu erhalten. Die Parameter von CNN und FC-Schicht werden für alle Bilder gemeinsam genutzt, d. h. die Anzahl der Parameter ist unabhängig von der Anzahl der Bilder n.

Für ein schnelles Ende-zu-Ende-Training werden die Videos unterabgetastet und für einen Datensatz immer die gleiche Anzahl Bilder verwendet (um eine feste Batch-Größe zu erhalten). Bei BioVid-A und X-ITE werden 5 Bilder je Sekunde verwendet, so dass sich bei BioVid-A $n = 26$ Bilder je Video ergibt, und bei X-ITE $n = 36$ Bilder. Für den Datensatz UNBC, bei dem die Videolänge variiert, werden $n = 32$ Bilder verwendet (festgelegt anhand der mittleren Videolänge). Beim Training werden 4 hintereinander gehängte Videos jeweils zu einem Batch zusammengesetzt. Diese Vorgehensweise wird auch für die anderen zeitlichen Integrationsmethoden mit CNN (Abschnitt 4.2.4 und 4.2.5) angewendet.

Wie bereits mehrfach diskutiert, ist Mimik ungleich verteilt. Insgesamt ist das Auftreten von Mimik über die Zeit eher selten und die Mimik ist oft nur kurz präsent, so dass sie im zeitlichen Mittelwert zum Teil schwach repräsentiert ist. Wie auch von Sikka et al. [SDB14] im Kontext von Multiple Instance Learning vorgeschlagen, ist es daher sinnvoll, zeitliche Maximalwerte zu betrachten. Diese fallen zeitlich typischerweise mit der extremsten Mimik des Videos zusammen.

In dieser Dissertation wird daher die **Entscheidungsfusion anhand des maximalen Entscheidungswertes** vorgeschlagen. Diese lässt sich auf alle oben genannten Varianten von Klassifikation und Regression anwenden, indem die Berechnung des zeitlichen Mittelwertes durch die Bestimmung des zeitlichen Maximalwertes ersetzt wird. Bei CNNs wird anstelle von Mean Pooling das Maximum (Max.) Pooling eingesetzt und mit dieser geänderten Architektur trainiert. Bei den Varianten von SVM und RF wird jedoch mit unabhängig betrachteten Einzelbildern trainiert. Die Bilder sind auf Ebene der Videos annotiert und somit enthalten Videos mit Schmerzmimik auch immer Bilder mit neutraler Mimik (entspannter Gesichtsmuskulatur), die als Schmerz annotiert sind. Dies verschiebt die Entscheidungsgrenze und Entscheidungswerte der trainierten Klassifikatoren. Es wird daher vorgeschlagen und im Kontext der Entscheidungsfusion mit Maximalwert evaluiert, den Schwellwert für die finale Entscheidung nach dem einzelbildbasierten Training von SVM und RF neu zu bestimmen. Hierfür wird das trainierte Modell auf die Trainingsdaten angewendet, um die Entscheidungswerte für jedes Einzelbild zu erhalten. Für jedes Video wird nun der maximale Entscheidungswert bestimmt. Diese Entscheidungswerte sowie die Grundwahrheiten der Videos werden anschließend genutzt, um den optimalen Schwellwert für die Entscheidungswerte zu wählen. Eine Möglichkeit hierfür ist es, alle Entscheidungswerte der Trainingsdaten als Schwellwert zu testen, jeweils die Performance der resultierenden Klassenzuordnung zu bestimmen und den Schwellwert mit der besten Performance zu wählen.

4.2.4. Erkennung mit zeitlichem Pooling von Merkmalen

Als Alternative zur Entscheidungsfusion mit CNNs schlägt der Autor dieser Dissertation die zeitliche Integration mittels Pooling von Merkmalen vor, siehe Abb. 4.4b. Die Architektur und Idee ist in weiten Teilen gleich zur Entscheidungsfusion mit CNN, die im vorherigen Abschnitt und in Abb. 4.4a vorgestellt wurde. Es wird ein vortrainiertes MobileNetV3-large für jedes der n Einzelbilder angewendet, um 1.280 Zeitreihen von Merkmalen zu extrahieren, und mit gemeinsam genutzten Parametern weiter trainiert. Jedoch wird das zeitliche Pooling hier nicht für die Entscheidungswerte, sondern bereits für die Zeitreihen von Einzelbildmerkmalen angewendet. Der Fully Connected Layer folgt nach dem Pooling, d. h. er verarbeitet Merkmale des gesamten Videos (und nicht jeweils eines Einzelbildes wie bei der Entscheidungsfusion). So werden deutlich

mehr Informationen der Einzelbilder zusammengeführt (1.280 Dimensionen), um die Videoentscheidung zu treffen, als bei der Entscheidungsfusion (1 bis 6 Dimensionen).

In der experimentellen Evaluierung werden zwei Varianten untersucht, eine mit Mean Pooling (d. h. der zeitliche Mittelwert jedes Merkmals wird genutzt) und eine mit Maximum Pooling (d. h. das zeitliche Maximum jedes Merkmals wird genutzt).

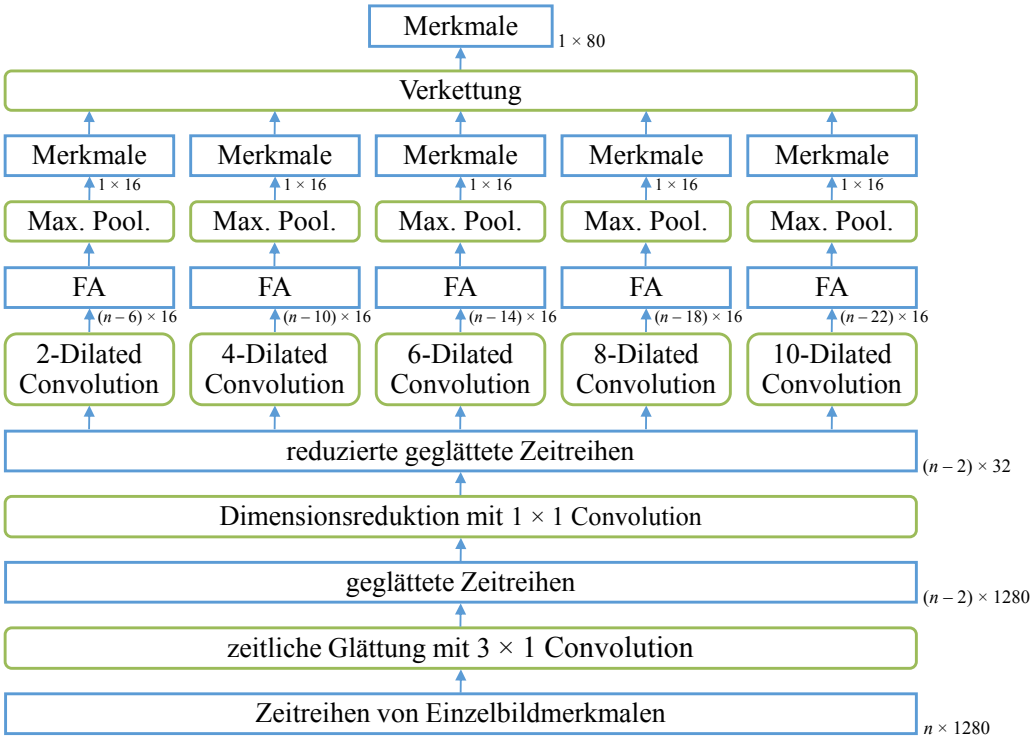
4.2.5. Erkennung mit zeitlicher Convolution

Zeitliches Pooling kann Merkmale verschiedener Einzelbilder kombinieren, jedoch keine Dynamikinformationen wie die Geschwindigkeit oder Dauer einer Bewegung extrahieren. Um letzteres zu ermöglichen, wird hier eine CNN-Variante mit 1-dimensionaler Convolution vorgeschlagen, die auf die Zeitreihen der Einzelbildmerkmale angewendet wird. Die auf diese Weise extrahierten Dynamikmerkmale werden mit Merkmalen kombiniert, die mit zeitlichem Max. Pooling extrahiert werden, wie im vorherigen Abschnitt beschrieben. In Vorversuchen hat sich gezeigt, dass durch die Kombination deutlich bessere Ergebnisse erzielt werden können, als wenn ausschließlich die zeitliche Convolution genutzt wird. Abb. 4.4c gibt einen Überblick über die Architektur. Grundlage ist wie in den vorherigen Abschnitten das MobileNetV3-large CNN, das für jedes der n Einzelbilder des Videos einen 1.280-dimensionalen Merkmalsvektor generiert, d. h. für das Video liegen 1.280 Merkmalszeitreihen mit jeweils n Elementen vor. Diese werden durch zeitliches Max. Pooling auf 1.280 Merkmale reduziert und parallel von einem CNN-Modul zur zeitlichen Convolution verarbeitet, dass 80 Merkmale extrahiert. Die Merkmalen werden verkettet und mit einer Fully-Connected-Schicht auf die Entscheidungswerte abgebildet.

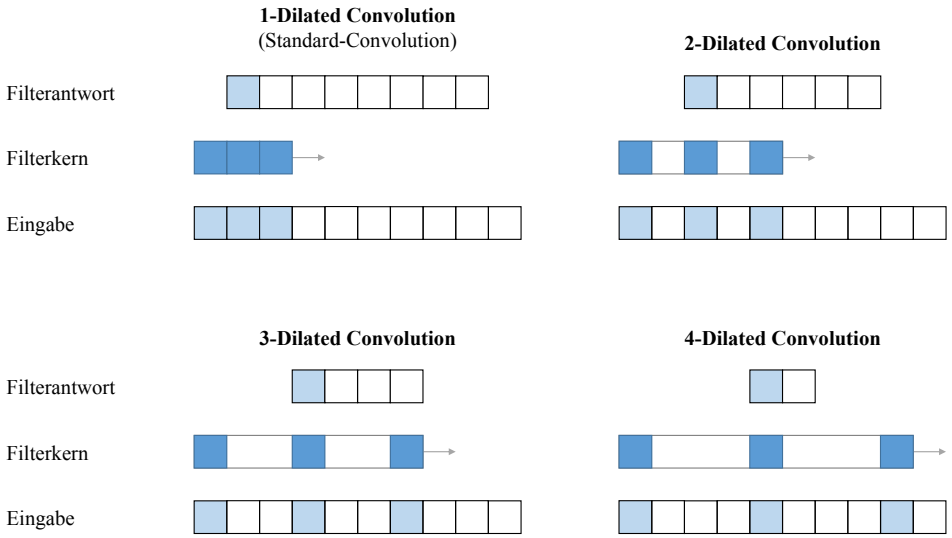
In Vorversuchen wurden einige Möglichkeiten für die Nutzung zeitlicher Convolution evaluiert. Hierbei hat sich gezeigt, dass die Überanpassung der Modelle ein großes Problem darstellt und die Zahl der neu zu lernenden Parameter gering gehalten werden muss, um gute Ergebnisse zu erzielen. Auf Basis dieser Anforderung wurde das in Abb. 4.5a dargestellte CNN-Modul für die zeitliche Convolution entwickelt. Es umfasst etwa 49.000 Parameter, eine im Vergleich zur restlichen Architektur (etwa 4,2 Millionen Parameter) relativ geringe Anzahl.

Ausgangspunkt sind die 1.280 Zeitreihen von Merkmalen, die extrahiert werden, indem das MobileNetV3-CNN auf jedem der n Einzelbilder des Videos angewendet wird. Diese Zeitreihen werden zunächst mit einer 1-dimensionalen Convolution geglättet, indem ein fester, nicht gelernter Filterkern mit den Parametern $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ angewendet wird. Wie bei der Deskriptorberechnung in Abschnitt 4.2.2 dient die Glättung der zeitlichen Rauschunterdrückung. Bei allen zeitlichen Convolutions wird hier kein Padding eingesetzt, d. h. die Ränder der Eingabe werden nicht erweitert, so dass die Ausgabe in der Filter-Dimension weniger Elemente hat (hier $n - 2$). Anschließend wird die Anzahl der Zeitreihen auf 32 reduziert, um die Parameteranzahl für die folgenden zeitlichen Convolutions gering zu halten. Hierfür wird eine 1×1 Convolution eingesetzt, die wie ein elementweise angewendeter Fully Connected Layer wirkt. Diese und alle folgenden Convolutions werden wie die anderen Convolutions im MobileNetV3 ohne Bias realisiert, mit einer anschließenden Batch Normalization ergänzt und einer darauf folgenden hard-swish-Funktion aktiviert.

Anschließend folgen die eigentlichen zeitlichen Convolutions, die als parallele Zweige im CNN-Graphen und als *Dilated Convolutions* realisiert sind. Dilated Convolutions wurden durch Yu und Koltun [YK16] bekannt, die sie erfolgreich für die semantische Segmentierung von Bildern eingesetzt haben. Abb. 4.5b veranschaulicht die Funktionsweise von 1-dimensionalen Dilated Convolutions. Sie verfügen über einen Parameter l, der bestimmt, wie der Filterkern angewendet wird. Standard-Convolutions entsprechen 1-Dilated Convolutions. Bei 2-Dilated Convolutions



(a) Realisierung der zeitlichen Convolution. Die Einbettung dieses Moduls in das Gesamt-CNN wird Abb. 4.4c gezeigt. FA = Filterantworten der zeitlichen l-Dilated Convolution.



(b) l-Dilated Convolution einer Zeitreihe. Mit $l = 1$ entspricht sie der Standard-Convolution. Mit $l > 1$ wird ohne zusätzliche Filterparameter ein größerer zeitlicher Kontext einbezogen.

Abbildung 4.5.: Zeitliche Integration mit Dilated Convolutions. Auch bei geringer Parameteranzahl ermöglichen Dilated Convolutions die Berechnung von Dynamikmerkmalen über verschiedene Zeitskalen.

wird der Filterkern „geweitet“, indem die Filterparameter mit jedem zweiten Eingabewert multipliziert werden. Allgemein werden bei l-Dilated Convolutions zwei benachbarte Filterparameter mit Eingabewerten multipliziert, die l Elemente auseinander liegen. Mit verschiedenen l werden Merkmale von zeitlich verschieden weit entfernten Bildern kombiniert. Durch die parallele Nutzung von verschiedenen l-Dilated Convolutions (siehe Abb. 4.5a) wird das CNN in die Lage versetzt, Zusammenhänge zu lernen, die eine unterschiedliche zeitliche Ausdehnung haben, wie z. B. verschieden lang andauernde Mimik.

Im vorgeschlagenen CNN-Modul (Abb. 4.5a) werden parallel fünf l-Dilated Convolutions mit $l \in \{2, 4, 6, 8, 10\}$ auf die reduzierten und geglätteten Zeitreihen angewendet, wobei die Dimensionalität weiter von 32 auf 16 Zeitreihen reduziert wird. Die Filterantworten einer jeden Dilated Convolution werden anschließend durch zeitliches Max. Pooling in jeweils 16 skalare Merkmale zusammengefasst. Alle Merkmale werden verkettet und bilden die Ausgabe des Moduls. Sie werden in der Gesamtarchitektur (Abb. 4.4c) mit den Merkmalen verkettet, die durch Maximum Pooling der originalen Merkmalszeitreihen berechnet wurden, und danach mittels Fully Connected Layer auf die Entscheidungswerte abgebildet.

4.2.6. Gewichtung der Intensitäten

Wie in Abschnitt 2.3 besprochen, gibt es bei Schmerzgrundwahrheiten, die nicht von einem Beobachter stammen, die Herausforderung, dass die Label oft nicht zur sichtbaren Schmerzreaktion passen. Für das maschinelle Lernen wirkt dies wie Label-Rauschen. Bei niedriger und mittlerer Schmerzintensität ist das Phänomen besonders ausgeprägt, da empfundene Schmerzen sich nur in der Mimik zeigen, wenn sie eine personenspezifische Schwelle überschreiten [PC95; Kun+04]. Diese wird bei den Datensätzen X-ITE und BioVid zum Teil erst bei der höchsten Schmerzintensität überschritten, bei einigen Probanden selbst bei dieser nicht (vgl. Abschnitt 2.3). Die Label der Klassen ohne Schmerzen und mit der höchsten Schmerzintensität sind am zuverlässigsten (im Sinne der Übereinstimmung mit der sichtbaren Schmerzmimik). Ohne Schmerzstimulation taucht nur in sehr wenigen Fällen Schmerzmimik auf (und dort mit niedriger Intensität). Die höchste Schmerzintensität ist die Klasse mit den meisten und stärksten mimischen Schmerzreaktionen. Daher wird vorgeschlagen, die Beispiele dieser beiden Klassen beim Training stärker zu gewichen, so dass sie einen größeren Einfluss auf das Modell haben als die Beispiele mit stärkerem Label-Rauschen.

Die Idee wird bei CNNs evaluiert und über kostensensitives Lernen realisiert. Hierzu wird die Loss-Funktion L, die beim Lernen des CNNs minimiert wird, mit Gewichten versehen:

$$L = \sum_{i=1}^{N} w(y_i) \cdot l(y_i, \hat{y}_i), \tag{4.1}$$

wobei N die Anzahl der Trainingsbeispiele bezeichnet, y_i das korrekte Label des Samples i, $\hat{y}_i$ den zugehörigen vom Netz prädizierten Wert und $l(\cdot, \cdot)$ den Loss, welcher aus diesen beiden Werten berechnet wird. Der Term $w(y_i)$ ist eine Funktion, die das Gewicht für die jeweilige Klasse $y_i \in \{1, 2, ..., C\}$ zurückgibt (bei C Intensitätsklassen):

$$w(y_i) = \begin{cases} \frac{C}{2}\lambda, & \text{wenn } y_i \in \{1, C\} \\ \frac{C}{C-2}(1-\lambda), & \text{sonst.} \end{cases} \tag{4.2}$$

Die Gewichtung wird über den Parameter $\lambda \in \mathbb{R}$ mit $0 \leq \lambda \leq 1$ gesteuert. λ gibt an, welchen Anteil die zwei Klassen Baseline und höchste Schmerzintensität zusammen am Loss haben sollen.

Bei gleichverteilten Klassenhäufigkeiten entspricht $\lambda = \frac{2}{C}$ dem Standardtraining ohne Wichtung der Intensitäten. Für den Datensatz BioVid mit $C = 5$ entspricht somit $\lambda = 0{,}4$ dem Standardtraining, bei X-ITE mit $C = 4$ ist es $\lambda = 0{,}5$. Die Idee, die Baseline und die höchste Intensität stärker zu gewichten, wird in Abschnitt 4.3.2 experimentell untersucht, indem größere λ gewählt und die Erkennungsergebnisse mit dem Standardtraining verglichen werden.

4.3. Experimente

Im Folgenden werden die vorgeschlagenen Methoden zur videobasierten Erkennung experimentell evaluiert. Zunächst werden Versuche zur Erkennung starker Schmerzen mit binärer Klassifikation durchgeführt (Abschnitt 4.3.1), wobei alle vorgeschlagenen zeitlichen Integrationsmethoden in Kombination mit verschiedenen Einzelbildmerkmalen bzw. Ausgangspunkten für CNN-Transferlernen getestet werden. Abschnitt 4.3.1 befasst sich mit der Messung der Schmerzintensität, wobei zusätzlich der Vergleich der Ergebnisse von Klassifikation und Regression sowie die Gewichtung der Intensitäten untersucht werden.

Erkennungsmethoden: Die videobasierte Erkennung baut auf den Arbeiten zur einzelbildbasierten Erkennung (Kapitel 3) auf und nutzt die gleichen Methoden zur Vorverarbeitung, Merkmalsextraktion und zum maschinellen Lernen. Insofern wurden, wenn nicht explizit anders beschrieben, auch die gleichen Parameter verwendet (vgl. Abschnitt 3.3). Für das Training der MobileNetV3-large (MN) CNNs wurde bei Regression mit Huber-Loss mit der Lernrate 0,001 begonnen und bei Klassifikation mit der Lernrate 0,003. Die Lernrate wurde nach 75% der Trainingsiterationen auf $\frac{1}{10}$ der Ausgangslernrate reduziert. Mit den Datensätzen BioVid-A/D und X-ITE wurde für 10.000 Iterationen trainiert, bei den kleineren Datensätze BioVid-A7/D7 und UNBC nur für 3.000 Iterationen um Überanpassung (engl. overfitting) zu reduzieren. Wenn nicht explizit die Gewichtung der Intensitäten nach Abschnitt 4.2.6 angewendet wurde (gekennzeichnet durch Angabe des Parameters λ), wurde die Klassengewichtung nach dem in Abschnitt 3.2.5 vorgeschlagenen Verfahren MID mit $\alpha = 0{,}5$ genutzt.

Daten: In Abschnitt 4.3.1 werden die Datensätze BioVid-D und BioVid-D7 (mit zwei Klassen) eingesetzt, in Abschnitt 4.3.2 BioVid-A und BioVid-A7 (mit allen 5 Klassen) sowie X-ITE und UNBC. Bei BioVid werden wie bei der einzelbildbasierten Erkennung jeweils zwei Varianten betrachtet, eine mit ausschließlich den Videos der frontalen Kamera und eine mit den Videos aller drei Kameraansichten, bei denen die Varianz der Kopfposen sehr viel größer ist. Bei X-ITE werden alle Modelle mit Videos beider Kameraansichten trainiert und getestet, d. h. mit sehr großer Kopfposevarianz. In allen Versuchen mit mehreren Ansichten werden die Videos unterschiedlicher Ansichten als unabhängige Samples aufgefasst, d. h. es findet *keine* Datenfusion über verschiedene Kameras hinweg statt.

Evaluierung: Bei allen hier verwendeten Datensätzen wird mit 5-facher LSO-Kreuzvalidierung evaluiert, d. h. ohne dass eine Person gleichzeitig in Trainingsdaten und Testdaten vorkommt. Zur Performance-Messung werden ICC-Werte berechnet, wie bei der einzelbildbasierten Erkennung nach Verkettung aller Prädiktionen und Grundwahrheiten. Die statistische Signifikanz von Unterschieden wird über Datenbanken und Methoden hinweg (auf Basis der abgedruckten Tabellenwerte) mittels rechtsseitigem Permutationstest berechnet.

4.3.1. Erkennung starker Schmerzen

Tabelle 4.2 fasst die Ergebnisse der Erkennung starker Schmerzen zusammen. Die Versuche wurden mit vier Datensätzen durchgeführt (Spalten): BioVid-D und BioVid-D7, jeweils in den zwei Varianten mit Videos von ausschließlich der frontalen Kamera und von allen drei Kameras. Die Datensätze umfassen die Samples der höchsten Schmerzintensität (an der Schmerztoleranz des Probanden, PA4) sowie die Baseline-Samples ohne Schmerzen (BLN). Evaluiert wurden: (1) die Klassifikation mit Random Forest (RF-K), Support Vector Machine (SVM) bzw. SVM Ensemble (SVM-E) basierend auf verschiedenen Einzelbildmerkmalen und Möglichkeiten der zeitlichen Integration (Zeilen #1-26) sowie (2) die Klassifikation mit Convolutional Neural Networks (CNN), die ausgehend von verschiedenen Vortrainings und mit verschiedenen zeitlichen Integrationsvarianten feinjustiert wurden (Zeilen #27-36). Im Interesse der Übersichtlichkeit zeigt Tabelle 4.2 nicht alle Ergebnisse mit RF-K, SVM-E und Einzel-SVM, sondern jeweils nur die bessere Klassifikation, entweder RF-K (normal gedruckt) oder SVM-E (kursiv gedruckt). Eine vollständige Auflistung aller Ergebnisse findet der interessierte Leser im Anhang in Tabelle C.1. Im Mittel liefert RF-K die besten Ergebnisse, gefolgt von SVM-E und Einzel-SVM (Mittelwerte der Entscheidungsfusion aus Tabelle C.1: 0,251 vs. 0,225 vs. 0,217). Beim Permutationstest verfehlen die Unterschiede zwischen RF-K und SVM-E sowie zwischen SVM-E und Einzel-SVM knapp das Signifikanzniveau. Der Unterschied zwischen Einzel-SVM und RF-K ist jedoch signifikant mit $p = 0{,}01$.

Die Wahl der Hintergrundfarben in Tabelle 4.2 basiert auf der Verteilung der Performances der jeweiligen Spalte. ICC-Werte unterhalb des 30%-Quantil sind weiß hinterlegt und die farbliche Sättigung darüber ist proportional zum ICC-Wert bis zum maximalen ICC-Wert der Spalte (wie in Tabelle 3.2). Wie auch bei vorherigen Untersuchungen sind die ICC-Werte mit der Teilmenge der 7 expressivsten Personen von BioVid (BioVid-D7) deutlich höher als mit dem gesamten, 87 Personen umfassenden Datensatz (BioVid-D), vgl. Label-Rauscheffekt in Abschnitt 2.3.

Einzelbildmerkmale / Ausgangspunkt für CNN-Feinjustierung: Bei Betrachtung der Merkmalsarten (Zeilen #1-26, Details siehe Abschnitt 4.2.1) stechen die Ergebnisse mit MobileNetV3-Bosphorus3D Action Units (Zeilen #19-26) positiv hervor, da sie bei allen Datensätzen sehr gut abschneiden (ICC im Mittel 0,604). Die Varianten mit 3D-Koordinaten, Kopfpose und/oder OpenFace [BRM16] Action Units (Zeilen #1-18, mittlerer ICC 0,375) zeigen insbesondere Schwächen bei den Datensätzen mit 3 Ansichten, die deutlich mehr nicht-frontale Kopfposen umfassen. Die Ergebnisse der feinjustierten CNN-Klassifikation (Zeilen #27-36, im Mittel 0,463) sind abhängig von der Anzahl der Trainingsbeispiele: Bei dem frontalen Teildatensatz von BioVid-D7 (mit nur 224 Trainingsbeispielen) sind die Ergebnisse mit CNN-Feinjustierung deutlich schlechter als bei Nutzung des vortainierten CNN zur Action-Unit-Schätzung mit anschließender Deskriptorberechnung und RF-K/SVM-E-Klassifikation (Zeilen #19-26). Die Differenz der mittleren ICC-Werte beträgt hier 0,231. Bei BioVid-D7 mit 3 Ansichten (672 Trainingsbeispiele) ist der Abstand schon geringer (0,183). Bei BioVid-D mit 2.784 (frontal) bzw. 8.352 Traingsbeispielen ist die Differenz der Mittelwerte noch geringer (0,077 bzw. 0,073) und in Zeile #31 werden für die beiden Datensätze die jeweils besten Ergebnisse erzielt. Das bessere Abschneiden der reinen CNN-Klassifikation bei steigender Anzahl von Trainingsbeispielen legt die Vermutung nahe, dass die CNN-Performance insgesamt an einer zu geringen Datenmenge leidet und von noch größeren Trainingsdatensätzen profitieren könnte.

Die MobileNetV3-Bosphorus3D Action Units (Zeilen #19-26) wurden in zwei Varianten verwendet: zum einen auf Basis von Regression (abgekürzt MN-R Bos3D), d. h. die AU-Intensitäten wurden als Fließkommawerte zur Berechnung der zeitlichen Deskriptoren verwendet, zum anderen

#	Einzelbildmerkmale / Ausgangspunkt	Zeitliche Integration	BioVid-D Frontal	BioVid-D 3 Ans.	BioVid-D7 Frontal	BioVid-D7 3 Ans.	MW
	Klassifikation mit RF-K (bzw. SVM-E wenn kursiv)						
1	3D-Koordinaten	Entscheidungsfusion (MW)	*0,187*	0,134	0,411	*0,463*	0,299
2		Entscheidungsfusion (Max.)	*0,260*	0,146	0,554	0,415	0,344
3		BoTF Deskriptor [Bar+14]	0,320	0,250	0,768	0,605	0,486
4		Statistikdeskriptor	0,347	0,278	*0,835*	0,613	0,518
5	Kopfpose	Entscheidungsfusion (MW)	0,039	0,032	*0,059*	0,004	0,034
6		Entscheidungsfusion (Max.)	*0,103*	0,067	*0,159*	0,134	0,116
7		BoTF Deskriptor [Bar+14]	0,248	0,207	0,495	0,459	0,352
8		Statistikdeskriptor	0,289	0,235	0,497	0,462	0,371
9	3D-Koord. + Kopfp.	Entscheidungsfusion (MW)	*0,190*	0,142	*0,389*	*0,449*	0,292
10		Entscheidungsfusion (Max.)	*0,265*	0,156	0,690	0,486	0,399
11		BoTF Deskriptor [Bar+14]	0,320	0,258	*0,798*	0,623	0,500
12		Statistikdeskriptor	0,359	0,282	*0,859*	0,609	0,527
13	OpenFace AU [BRM16]	Entscheidungsfusion (MW)	0,255	0,165	0,552	0,331	0,326
14		Entscheidungsfusion (Max.)	*0,278*	0,153	0,664	0,425	0,380
15		BoTF Deskriptor [Bar+14]	0,351	0,181	0,804	0,340	0,419
16		Statistikdeskriptor	0,363	0,215	*0,834*	0,498	0,478
17	OpenFace AU [BRM16] + Kopfp.	BoTF Deskriptor [Bar+14]	0,352	0,183	0,809	0,343	0,422
18		Statistikdeskriptor	0,371	0,229	*0,833*	0,498	0,483
19	MN-R Bos3D AU	BoTF Deskriptor [Bar+14]	0,367	0,359	0,848	0,831	0,601
20		Statistikdeskriptor	0,373	0,372	*0,854*	**0,870**	0,617
21	MN-R Bos3D AU + Kopfp.	BoTF Deskriptor [Bar+14]	0,376	0,357	0,840	0,828	0,600
22		Statistikdeskriptor	0,384	0,373	0,864	0,859	**0,620**
23	MN-K Bos3D AU	BoTF Deskriptor [Bar+14]	0,341	0,339	*0,824*	0,828	0,583
24		Statistikdeskriptor	0,345	0,347	0,867	0,859	0,605
25	MN-K Bos3D AU + Kopfp.	BoTF Deskriptor [Bar+14]	0,362	0,350	0,820	0,827	0,590
26		Statistikdeskriptor	0,364	0,363	**0,874**	0,862	0,616
	Klassifikation mit CNN nach Feinjustierung						
27	MN-R Bos3D	Entscheidungsfusion (MW)	0,254	0,248	0,606	0,665	0,443
28		Entscheidungsfusion (Max.)	0,300	0,309	0,664	0,729	0,500
29		Pooling von Merkmalen (MW)	0,268	0,221	0,625	0,660	0,444
30		Pooling von Merkmalen (Max.)	0,361	0,364	0,676	0,753	0,538
31		Pooling mit zeitlicher Conv.	**0,386**	**0,403**	0,672	0,740	0,551
32	MN-K ImageNet	Entscheidungsfusion (MW)	0,182	0,190	0,556	0,523	0,363
33		Entscheidungsfusion (Max.)	0,270	0,271	0,494	0,674	0,427
34		Pooling von Merkmalen (MW)	0,210	0,190	0,583	0,611	0,399
35		Pooling von Merkmalen (Max.)	0,294	0,295	0,683	0,653	0,481
36		Pooling mit zeitlicher Conv.	0,345	0,355	0,625	0,619	0,486

AU: Action Unit BoTF: Bag of Temporal Features
MN-R Bos3D: MobileNetV3-large CNN, vortrainiert zur Regression von AU mit Bosphorus3D Datensatz
MN-K Bos3D: MobileNetV3-large CNN, vortrainiert zur Klassifikation von AU mit Bosphorus3D Datensatz
MN-K ImageNet: MobileNetV3-large CNN, vortrainiert zur Klassifikation von Objekten mit ImageNet

Tabelle 4.2.: Videobasierte Erkennung starker Schmerzen: Test-Performances (ICC) mit verschiedenen Kombinationen von Merkmalsarten, zeitlichen Integrationsmethoden und Lernverfahren (Zeilen) für die Datensätze (Spalten): BioVid-D (Frontalansicht und alle 3 Ansichten) und BioVid-D7 (Teilmenge von BioVid-D mit den 7 expressivsten Probanden). Angegeben ist auch der Mittelwert (MW) aller 4 Datensätze (letzte Spalte).

auf Basis von Klassifikation (abgekürzt MN-K Bos3D), d. h. die AUs-Intensitäten waren diskretisiert wie im FACS definiert (ganzzahlige Werte von 0 bis 5). Die Ergebnisse mit Regression (Zeilen #19-22, im Mittel 0,610) sind zwar nur geringfügig besser als die Ergebnisse mit Klassifikation (Zeilen #23-26, Mittelwert 0,598), jedoch ist der Unterschied statistisch signifikant ($p = 0{,}0036$). Dies bestätigt die Hypothese aus Abschnitt 3.2.4, dass die Regression von AU-Intensitäten gegenüber der Klassifikation Vorteile bringt, die im Videokontext messbar werden.

Die Ergebnisse mit 3D-Koordinaten, Kopfpose und OpenFace Action Units unterscheiden sich deutlich. Die Mittelwerte der Ergebnisse mit zeitlichem Deskriptor sind mit 3D-Koordinaten 0,502 (Zeilen #3-4), mit Kopfpose 0,362 (#7-8) und mit OpenFace AU 0,448 (#15-16). Somit führen die vorgeschlagenen 3D-Koordinaten als Einzelbildmerkmale zu einer besseren Performance als die Action Units von OpenFace [BRM16], insbesondere aufgrund ihres besseren Abschneidens bei nicht-frontalen Kopfposen (Daten mit 3 Ansichten). Für sich genommen ist die Kopfpose das schwächste Merkmal, in Fusion mit Mimikmerkmalen verbessert sie jedoch die Ergebnisse in fast allen Versuchen, sowohl in Kombination mit 3D-Koordinaten (0,513, Zeilen #11-12) und OpenFace AU (0,452, #17-18), als auch mit den MobileNetV3-Bosphorus3D AU (im Mittel 0,609 vs. 0,610 und 0,594 vs. 0,603 für #19-20 vs. #21-22 und #23-24 vs. #25-26). Der paarweise Permutationstest der Ergebnisse mit und ohne Kopfpose zeigt trotz geringem Unterschied in den Mittelwerten (0,495 ohne Kopfpose, 0,505 mit Kopfpose) eine statisch signifikante Verbesserung durch die Fusion mit der Kopfpose ($p = 0{,}009$).

Bei der feinjustierten CNN-Klassifikation (Zeilen #27-36) wurde zum einen das mit Bosphorus3D vortrainierte CNN als Ausgangspunkt genutzt (#27-31), zum anderen das ausschließlich zur Objektklassifikation mit ImageNet trainierte CNN (#32-36). Die Performances nach Vortrainieren mit Bosphorus3D sind signifikant besser als mit ImageNet (im Mittel 0,495 vs. 0,431, $p = 0{,}0002$), wie auch schon in den Experimenten der einzelbildbasierten Erkennung (Abschnitt 3.3.3, Tabelle 3.2).

Zeitliche Integrationsmethoden: Folgende zeitliche Integrationsmethoden wurden untersucht und sind in Tabelle 4.2 zu finden:

Entscheidungsfusion (MW): Zur zeitlichen Integration über mehrere Einzelbilder wird der Klassifikator auf jedes Einzelbild angewendet und der *Mittelwert* der Entscheidungswerte (engl. decision scores) aller Einzelbilder gebildet. Das Verfahren wurde Ashraf et al. [Ash+09] mit SVM-Klassifikation vorgeschlagen. Die originale Idee von Ashraf et al. wurde mit verschiedenen Einzelbildmerkmalen evaluiert (Ergebnisse mit SVM im Anhang C.1) und mit Ensemble-Klassifikatoren kombiniert (SVM-E und RF-K). Die Idee wurde auch auf das MobileNetV3-CNN angewendet, das mit Parameter Sharing vom Einzelbild- zum Video-CNN erweitert wurde (vgl. Abschnitt 4.2.3).

Entscheidungsfusion (Max.): Der Klassifikator wird auf jedes Einzelbild angewendet und der *Maximalwert* der Entscheidungswerte aller Einzelbilder gebildet. Details dieser vom Autor dieser Dissertation vorgeschlagenen Methode sind in Abschnitt 4.2.3 zu finden, auch zur Variante mit MobileNetV3-CNN.

BoTF Deskriptor: Der Bag of Temporal Features (BoTF) Deskriptor wurde von Bartlett et al. [Bar+14] zur Unterscheidung von echten und vorgespielten Schmerzen vorgeschlagen. Zur Berechnung werden die Merkmalszeitreihen mit Gabor Filtern gefiltert und die Ergebnisse in Histogrammen zusammengefasst. Um die Idee mit den hier verfügbaren Merkmalen optimal auszunutzen, wird von jeder Zeitreihe zuvor ihr Mittelwert abgezogen.

Statistikdeskriptor: Wie in Abschnitt 4.2.2 dieser Dissertation vorgeschlagen, werden für jede Merkmalszeitreihe und ihre Ableitungen statistische Maße extrahiert.

Pooling von Merkmalen (MW / Max.): Das zeitliche Pooling von CNN-Merkmalen wurde in Abschnitt 4.2.4 vorgeschlagen und wird in zwei Varianten evaluiert: mit Mean Pooling (Mittelwert = MW) und mit Max. Pooling.

Pooling mit zeitlicher Convolution: Hierbei handelt es sich um den in Abschnitt 4.2.5 vorgeschlagenen Ansatz, der zeitliche Dilated Convolutions und Max. Pooling kombiniert.

Die Evaluierung der Entscheidungsfusion mit SVM-E und RF-K ist extrem rechen- und zeitaufwändig, insbesondere beim Testen (aller Einzelbilder). Da sich bei den ersten untersuchten Merkmalsarten bereits abzeichnete, dass die Entscheidungsfusion mit SVM-E/RF-K deutlich schlechter funktioniert als die Deskriptoren (im Mittel 0,274 vs. 0,456 bei gleichen Einzelbildmerkmalen), wurde bei einem Teil der Merkmale auf die Untersuchung der Entscheidungsfusion verzichtet. Der Vorteil der Deskriptoren ist, dass im Gegensatz zur Entscheidungsfusion Dynamikinformationen zur Klassifikation ausgenutzt werden können, wie z. B. Geschwindigkeit, Beschleunigung oder die Dauer einer Bewegung.

Bei der Entscheidungsfusion ist die Nutzung des Maximalwertes im Mittel besser als die Nutzung des Mittelwertes, sowohl bei den Ergebnissen mit SVM-E/RF-K (0,310 vs. 0,238), als auch bei Klassifikation mit CNN (0,464 vs. 0,403). Dies ist insofern schlüssig, dass das Einzelbild mit der am stärksten ausgeprägten Reaktion (Mimik / Kopfpose) mehr über die Schmerzen aussagt, als ein Mittelwert über alle Bilder (inklusive einem potentiell hohem Anteil ohne Schmerzreaktion).

Vergleicht man die mittleren Ergebnisse der zwei untersuchten Deskriptoren, findet sich eine statistisch signifikante Überlegenheit des vorgeschlagenen Statistikdeskriptors (Mittelwert 0,537) gegenüber dem BoTF Deskriptor von Bartlett et al. [Bar+14] (0,506) mit $p < 0{,}0001$. Ein möglicher Grund ist, dass der vorgeschlagene Statistikdeskriptor zusätzliche Informationen codiert. So werden im BoTF Deskriptor zwar Häufigkeiten und Frequenzen sehr detailliert erfasst, Amplitudeninformationen sind jedoch nur wenig repräsentiert.

Beim feinjustiertem CNN-Klassifikator funktioniert die zeitliche Fusion von Merkmalen (Zeilen #29-31 und #34-36, mittlerer ICC 0,483) besser als die Entscheidungsfusion mit CNN (#27-28 und #32-33 ICC 0,433). Unter den zeitlichen Integrationsmethoden mit CNN erreicht Pooling mit zeitlicher Convolution im Mittel die besten ICC-Werte, sowohl für MobileNetV3 mit Bosphorus3D (#31, ICC 0,551) als auch mit ImageNet (#36, 0,486), gefolgt von Max. Pooling von Merkmalen (#30 mit 0,538 und #35 mit 0,481) und Entscheidungsfusion mit Maximalwert (#28 mit 0,500 und #33 mit 0,427). Die mittelwertbasierten Fusionen schneiden deutlich schlechter ab.

Die insgesamt besten Ergebnisse je Datensatz (fett in der Tabelle) basieren in allen 4 Fällen auf dem MobileNetV3, das mit Bosporus3D vortrainiert wurde. Für die größeren Datensätze BioVid-D (frontal und mit 3 Ansichten) ist das Max. Pooling mit zeitlicher Convolution (#31) die beste Integrationsmethode. Für die kleineren BioVid-D7-Datensätze (mit wenigen Hundert Trainingsbespielen) ist die Verwendung des Statistikdeskriptors mit RF-K die beste Option, da die CNNs bei der Feinjustierung overfitten (sich überanpassen).

Kopfposeinvarianz: Die Performances mit menschengemachten Merkmalen und OpenFace Action Units (Zeilen #1-18 in Tabelle 4.2) sinken, wenn zu den frontalen Daten die zwei Seitenansichten (±45°) hinzugenommen werden (vgl. linke und rechte Spalte bei BioVid-D und BioVid-D7). Dies zeigt die Schwäche von menschengemachten Merkmalen bei großer Posevarianz. Die AU-Prädiktionen, die MobileNetV3 nach dem Training mit Bosphorus3D extrahiert, sind unabhängig von der Kopfpose bzw. ermöglichen das Training eines von der Kopfpose unabhängigen Erkennungsmodells. Dies zeigen die Ergebnisse in den Zeilen #19-26, die mit Hinzunahme der Seitensichten nicht oder nur geringfügig schlechter werden.

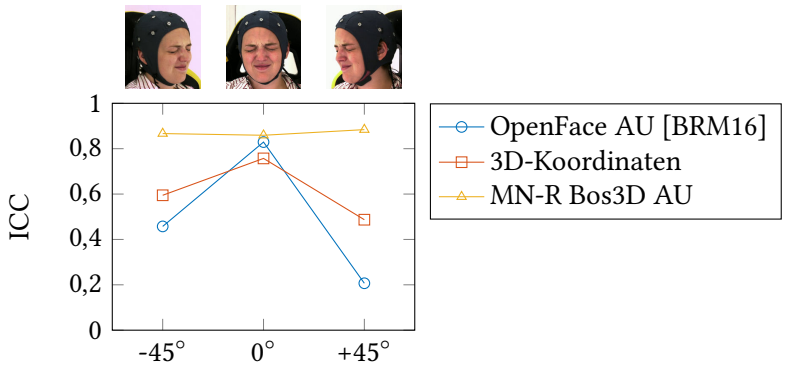


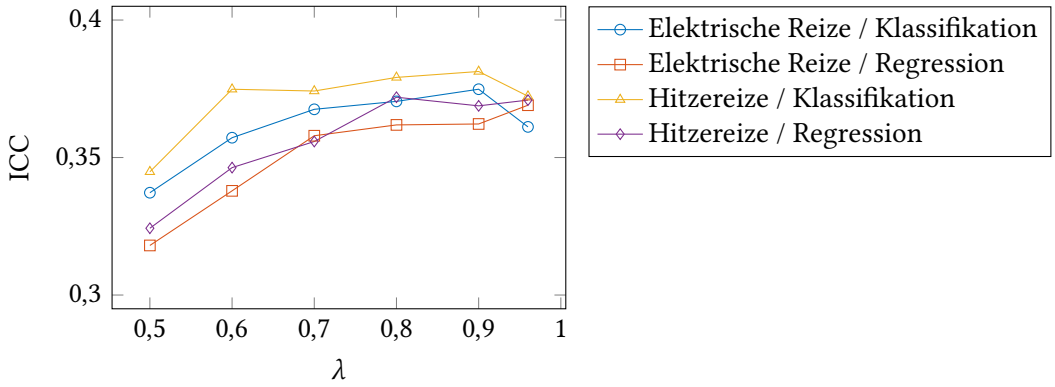
Abbildung 4.6.: Videobasierte Erkennung starker Schmerzen in Abhängigkeit der Kameraansicht bzw. mittleren Kopfpose für den Datensatz BioVid-D7 mit verschiedenen Einzelbildmerkmalen bei Klassifikation mit RF-K und zeitlicher Integration mit Statistikdeskriptor.

Abb. 4.6 visualisiert einige Performances auf BioVid-D7 mit 3 Kameraansichten aufgeschlüsselt nach Ansicht. Verglichen werden die Erkennungsleistungen mit den Merkmalen OpenFace AU (Zeile #16), 3D-Koordinaten (#4) und MobileNetV3 Bosphorus3D AU (#20). OpenFace schneidet bei der frontalen Ansicht auf das Gesicht sehr gut ab (ICC 0,829), jedoch fällt die Performance bei den Seitenansichten auf 0,457 bzw. 0,207. Die 3D-Koordinatenmerkmale sind bei der frontalen Ansicht leicht unterlegen (ICC 0,757), liefern jedoch bei den Seitenansichten bessere Ergebnisse als OpenFace. Die AU-Merkmale aus dem mit Bosphorus3D trainierten CNN liefern bei allen Ansichten etwa die gleiche Performance, was für eine sehr gute Kopfposeinvarianz spricht.

4.3.2. Messung der Schmerzintensität

Die Experimente zur automatisierten Messung der Schmerzintensität mit Klassifikation und Regression werden mit den Datensätzen BioVid-A, BioVid-A7, X-ITE und UNBC durchgeführt. BioVid-A und BioVid-A7 haben fünf Klassen, vier Intensitäten von Schmerzstimuli (PA1 bis PA4) und „kein Schmerzstimulus“ (BLN, Intensität 0). Diese werden als geordnete Klassen durch die Werte 0 bis 4 repräsentiert und können somit sowohl für Klassifikation als auch für Regression genutzt werden. Bei X-ITE werden zwei Aufgaben betrachtet: die Schmerzmessung bei elektrischen Reizen und die Schmerzmessung bei Hitzereizen. Es gibt jeweils drei Intensitäten von Schmerzstimuli und somit (ergänzt um „kein Schmerzstimulus“) vier geordnete Klassen, die durch die Werte 0 bis 3 repräsentiert werden. Die Klassen des Teildatensatzes der elektrischen Reize werden auch BLN, PE1, PE2 und PE3 genannt, die des Hitze-Teildatensatzes BLN, PH1, PH2 und PH3. Beim Datensatz UNBC werden die videobasierte Beobachterbeurteilung (engl. Observer Pain Rating, OPR, ganzzahlige Werte im Intervall $[0, 5]$) sowie die Selbsteinschätzung der Patienten anhand der Visuellen Analogskala (VAS, verfügbar als ganzzahlige Werte im Intervall $[0, 10]$) betrachtet. Es werden somit unabhängige Experimente mit den gleichen Videos, aber zwei verschiedenen Grundwahrheiten durchgeführt.

Gewichtung der Intensitäten: In Abschnitt 4.2.6 wurde vorgeschlagen, die Klassen „kein Schmerz“ (Baseline, BLN) und „höchste Schmerzintensität“ (bei X-ITE mit elektrischen Reizen PE3 und mit Hitzereizen PH3, bei BioVid PA4) stärker zu gewichten. Diese Klassen werden im



(a) Erkennungsleistung (ICC) für verschiedene λ.

	Prädizierte Klasse			
	BLN	PE1	PE2	PE3
BLN	4817	2114	508	255
PE1	4593	2139	615	430
PE2	4235	1933	758	863
PE3	2697	1225	804	3057

(b) Klassifikation ohne Gewichtung ($\lambda = 0{,}5$). ICC = 0,334 / Accuracy = 0,347.

	Prädizierte Klasse			
	BLN	PE1	PE2	PE3
BLN	7210	0	4	480
PE1	6974	0	3	800
PE2	6393	1	0	1395
PE3	3934	0	1	3848

(c) Klassifikation mit Gewichtung ($\lambda = 0{,}9$). ICC = 0,375 / Accuracy = 0,356.

	Prädizierte Klasse			
	BLN	PE1	PE2	PE3
BLN	3410	3991	287	6
PE1	3127	4249	393	8
PE2	2792	4213	677	107
PE3	1447	3326	1970	1040

(d) Regression ohne Gewichtung ($\lambda = 0{,}5$). ICC = 0,318 / Accuracy = 0,302.

	Prädizierte Klasse			
	BLN	PE1	PE2	PE3
BLN	6190	1231	236	37
PE1	5841	1456	417	63
PE2	5296	1614	691	188
PE3	3079	1619	1657	1428

(e) Regression mit Gewichtung ($\lambda = 0{,}96$). ICC = 0,369 / Accuracy = 0,315.

Abbildung 4.7.: Einfluss der Gewichtung der Schmerzintensitätsklassen auf die videobasierte Erkennung beim Datensatz X-ITE mit dem CNN MN-R Bos3D mit zeitlicher Convolution. Der Parameter λ gibt an, wie stark die zwei Klassen Baseline (BLN) und höchste Schmerzintensität (PE3 bzw. PH3) im Loss gewichtet sind (vgl. Abschnitt 4.2.6). $\lambda = 0{,}5$ entspricht dem Standardtraining ohne Gewichtung, größere λ erhöhen den Einfluss von BLN und PE3 (bzw. PH3) auf das Training und reduzieren den Einfluss der mittleren Klassen. **(b)-(e) Konfusionsmatrizen** mit elektrischen Reizen. Für die Regression wurden die Klassen als Intensität 0-3 codiert und anschließend zur Berechnung der Konfusionsmatrix diskretisiert.

Folgenden Randklassen genannt, um sie von den dazwischen liegenden Intensitäten abzugrenzen. Wie stark die Randklassen gewichtet werden, wird über den Parameter λ gesteuert (vgl. Abschnitt 4.2.6). Abb. 4.7a zeigt die Performance, die ein MobilNetV3-large CNN (vortrainiert auf Bosphorus3D, mit zeitlicher Convolution) auf dem Datensatz X-ITE erreicht, wenn der Parameter variiert wird. Für alle vier Varianten (elektrische Reize und Hitzereize jeweils mit Klassifikation und Regression) verbessern sich die ICC-Werte deutlich, wenn die Gewichtung ausgehend von $\lambda = 0{,}5$, was dem Standardtraining ohne Gewichtung entspricht, erhöht wird. Insofern hat die Gewichtung einen messbaren positiven Einfluss auf die Performance.

Für weitergehende Betrachtungen zeigen Abb. 4.7 (b)-(e) einige Konfusionsmatrizen der Modelle. In Abb. 4.7b (Klassifikation ohne Gewichtung) wird ersichtlich, dass (1) die Randklassen am zuverlässigsten klassifiziert werden und (2) viele Beispiele, insbesondere von PE1 und PE2, aber auch von PE3, fälschlicherweise als BLN klassifiziert werden. Diese Beobachtungen unterstützten (1) den Ausgangspunkt der Idee, dass die Randklassen am wenigsten von Label-Rauschen betroffen sind und (2) die Aussage, dass in vielen Videos trotz angewendeter Schmerzstimulation keine Schmerzreaktion zu beobachten ist. Wird die Gewichtung auf $\lambda = 0{,}9$ gesetzt, erhöhen sich der ICC-Wert von 0,334 auf 0,375 und die Korrektklassifikationsrate (Accuracy) von 0,347 auf 0,356. Wie in Abb. 4.7c ersichtlich wird der Intensitätsklassifikator hierbei jedoch effektiv zum binären Klassifikator, der fast ausschließlich BLN und PE3 ausgibt, was der eigentlichen Aufgabe widerspricht. Dies zeigt, dass die Wahl einer aussagekräftigen Performance-Metrik eine große Herausforderung darstellt (vgl. auch Anhang A) und auch die Betrachtung der konkreten Prädiktionsergebnisse wichtig ist.

Bei der Regression verhält es sich anders. Zunächst veranschaulicht hier Abb. 4.7d (Regression ohne Gewichtung) ein Problem, dem die stärkere Gewichtung der Randklassen entgegenwirkt. Der größte Teil der Beispiele wird hier als PE1 klassifiziert. Das Training der Regression basiert auf Loss-Funktionen, welche die Differenz zwischen Grundwahrheit und Prädiktion bilden. Bei schlechter Unterscheidbarkeit tendiert ein Regressionsmodell infolgedessen zum Mittelwert, da dieser beim Training mit einen relativ niedrigen Loss einhergeht. Eine stärkere Gewichtung der Randklassen wirkt diesem Problem entgegen, wie in Abb. 4.7e ($\lambda = 0{,}96$) deutlich wird. Hier werden BLN, PE3 und auch PE2 deutlich häufiger richtig erkannt. Die resultierende ICC-Performance liegt mit 0,369 zwar zahlenmäßig unter den besten Ergebnissen der *Klassifikation* (Abb. 4.7c), jedoch ist dieses Regressionsmodell zur Intensitätsschätzung deutlich nützlicher, da es nicht nur zwei Klassen prädiziert, und auch nicht nur vier Klassen wie von der Konfusionsmatrix suggeriert, sondern aufgrund des Fließkommaausgabewertes sehr viele Zwischenstufen.

Vergleich der Erkennungsmethoden: Tabelle 4.3 vergleicht die ICC-Werte verschiedener Methoden auf den Datensätzen BioVid-A und BioVid-A7 (beide jeweils mit ausschließlich Frontalansicht sowie allen 3 Ansichten), X-ITE mit 2 Ansichten jeweils für elektrische Stimulation und Hitzestimulation, sowie UNBC mit den zwei Grundwahrheiten OPR (Observer Pain Rating) und VAS (Visualle Analogskala). Die letzte Spalte enthält den Mittelwert (MW) der 8 Datensätze. Die Werte in den Spalten sind wie in Tabelle 4.2 eingefärbt, wobei jedoch der oberste Teil (Mittelwerte der deskriptorbasierten Ansätze) aufgrund des anderen Wertebereiches unabhängig von den darunter liegenden Teilen eingefärbt wurde. Die vollständigen deskriptorbasierten Ergebnisse sind im Anhang in Tabelle C.2 zu finden. In Tabelle 4.3 wurden sie anhand von Mittelwerten (Zeilen #1-15) sowie den besten Ergebnissen (#16-30) bezüglich Lernverfahren, zeitlichem Deskriptor und Einzelbildmerkmalen zusammengefasst. Unten werden die Performances der Feinjustierung ausgehend von MN-R Bos3D mit CNN-Klassifikation (#31-37) und -Regression (#38-44), jeweils mit verschiedenen zeitlichen Integrationsmethoden, gegenübergestellt.

#	Methodik	BioVid-A Front.	BioVid-A 3 Ans.	BioVid-A7 Front.	BioVid-A7 3 Ans.	X-ITE (2 Ans.) Elektr.	X-ITE (2 Ans.) Hitze	UNBC OPR	UNBC VAS	MW
	Deskriptorbasiert: Mittelwerte									
1	SVM-E	0,153	0,126	0,449	0,343	0,224	0,223	0,602	0,420	0,318
2	SVR-E	0,159	0,106	0,271	0,199	0,196	0,196	0,586	0,417	0,266
3	RF-K	0,235	0,188	0,574	0,485	0,264	0,282	0,639	0,463	0,391
4	RF-R	0,144	0,098	0,490	0,411	0,236	0,249	0,601	0,451	0,335
5	BoTF Deskriptor [Bar+14]	0,155	0,113	0,451	0,324	0,204	0,217	0,618	0,450	0,317
6	Statistikdeskriptor	0,190	0,146	0,441	0,395	0,256	0,258	0,595	0,426	0,338
7	3D-Koordinaten	0,165	0,107	0,449	0,359	0,229	0,233	0,635	0,481	0,332
8	Kopfpose	0,131	0,086	0,236	0,223	0,203	0,224	0,439	0,312	0,232
9	3D-Koord. + Kopfp.	0,170	0,110	0,454	0,340	0,225	0,231	0,661	0,498	0,336
10	OpenFace AU [BRM16]	0,176	0,081	0,452	0,224	0,207	0,214	0,560	0,360	0,284
11	OpenFace AU [BRM16] + Kopfp.	0,175	0,087	0,450	0,228	0,205	0,211	0,567	0,379	0,288
12	MN-R Bos3D AU	0,188	0,176	0,502	0,468	0,252	0,259	0,648	0,474	0,371
13	MN-R Bos3D AU + Kopfp.	0,191	0,177	0,498	0,467	0,249	0,255	0,653	0,478	0,371
14	MN-K Bos3D AU	0,174	0,170	0,473	0,466	0,252	0,258	0,647	0,475	0,364
15	MN-K Bos3D AU + Kopfp.	0,185	0,173	0,498	0,461	0,249	0,253	0,653	0,482	0,369
	Deskriptorbasiert: beste Ergebnisse									
16	SVM-E	0,188	0,190	0,548	0,491	0,264	0,251	0,669	0,508	0,389
17	SVR-E	0,219	0,189	0,440	0,364	0,263	0,256	**0,712**	0,559	0,375
18	RF-K	0,264	0,249	**0,649**	**0,647**	0,298	0,308	0,698	**0,567**	**0,460**
19	RF-R	0,174	0,157	0,585	0,568	0,285	0,291	0,661	0,506	0,404
20	BoTF Deskriptor [Bar+14]	0,253	0,244	0,613	0,622	0,280	0,308	**0,712**	0,559	0,449
21	Statistikdeskriptor	0,264	0,249	**0,649**	**0,647**	0,298	0,308	0,678	**0,567**	0,458
22	3D-Koordinaten	0,237	0,171	0,576	0,480	0,276	0,284	0,690	0,541	0,407
23	Kopfpose	0,192	0,148	0,366	0,313	0,251	0,263	0,565	0,434	0,316
24	3D-Koord. + Kopfp.	0,240	0,170	0,573	0,494	0,279	0,286	**0,712**	0,559	0,414
25	OpenFace AU [BRM16]	0,248	0,137	0,620	0,362	0,257	0,265	0,602	0,417	0,363
26	OpenFace AU [BRM16] + Kopfp.	0,246	0,143	0,610	0,373	0,258	0,273	0,599	0,421	0,365
27	MN-R Bos3D AU	0,260	0,249	**0,649**	**0,647**	0,294	0,306	0,689	0,539	0,454
28	MN-R Bos3D AU + Kopfp.	0,264	0,249	0,627	0,646	0,298	0,308	0,698	0,543	0,454
29	MN-K Bos3D AU	0,240	0,238	0,625	0,636	0,294	0,308	0,689	0,534	0,446
30	MN-K Bos3D AU + Kopfp.	0,262	0,247	0,625	0,644	0,292	0,305	0,698	**0,567**	0,455
	CNN-Klassifikation mit MN-R Bos3D									
31	Entscheidungsfusion (MW)	0,147	0,151	0,281	0,376	0,229	0,201	0,388	0,254	0,254
32	Entscheidungsfusion (Max.)	0,190	0,213	0,430	0,457	0,263	0,251	0,549	0,241	0,324
33	Pooling von Merkmalen (MW)	0,161	0,154	0,265	0,380	0,221	0,205	0,412	0,293	0,261
34	Pooling von Merkmalen (Max.)	0,252	0,267	0,391	0,415	0,334	0,334	0,528	0,400	0,365
35	Pooling von Merkmalen (Max.), $\lambda = 0{,}9$	0,257	0,263	0,451	0,412	0,356	0,371			
36	Pooling mit zeitlicher Conv.	0,277	0,278	0,377	0,461	0,337	0,345	0,531	0,389	0,375
37	Pooling mit zeitlicher Conv., $\lambda = 0{,}9$	0,277	0,301	0,440	0,466	**0,375**	**0,381**			
	CNN-Regression mit MN-R Bos3D									
38	Entscheidungsfusion (MW)	0,131	0,116	0,263	0,388	0,228	0,181	0,519	0,342	0,271
39	Entscheidungsfusion (Max.)	0,226	0,215	0,423	0,432	0,292	0,272	0,651	0,372	0,360
40	Pooling von Merkmalen (MW)	0,139	0,112	0,290	0,374	0,228	0,184	0,494	0,388	0,276
41	Pooling von Merkmalen (Max.)	0,268	0,243	0,393	0,458	0,320	0,306	0,578	0,385	0,369
42	Pooling von Merkmalen (Max.), $\lambda = 0{,}9$	0,287	0,294	0,460	0,555	0,354	0,330			
43	Pooling mit zeitlicher Conv.	0,282	0,266	0,401	0,472	0,318	0,324	0,597	0,421	0,385
44	Pooling mit zeitlicher Conv., $\lambda = 0{,}9$	**0,322**	**0,312**	0,522	0,551	0,362	0,369			

Tabelle 4.3.: Videobasierte Messung der Schmerzintensität: Test-Performances (ICC) mit verschiedenen Methoden (Zeilen) für die Datensätze (Spalten): BioVid-A (Frontalansicht und alle 3 Ansichten), BioVid-A7 (Teilmenge von BioVid-A mit den 7 expressivsten Probanden), X-ITE mit 2 Ansichten (für elektrische Stimulation und Hitzestimulation) und UNBC (mit Grundwahrheiten Oberver Pain Rating und Visueller Analogskala). Angegeben ist auch der Mittelwert (MW) aller 8 Datensätze (letzte Spalte). Der oberste Teil zeigt die Mittelwerte deskriptorbasierter Ergebnisse für die Lernmethoden, Deskriptoren und Einzelbildmerkmale (mit vom unteren Teil unabhängiger Einfärbung), Details vgl. Tabelle C.2. Im unteren Teil werden die besten deskriptorbasierten Ergebnisse sowie die Performances mit CNN-Klassifikation und -Regression aufgelistet.

Unter den Lernverfahren für die deskriptorbasierte Erkennung ist die Random Forest Klassifikation (RF-K) klar das Beste (Zeile #3, MW 0,391), gefolgt von Random Forest Regression (#4, MW 0,335), SVM Ensemble Klassifikation (#1, MW 0,318) und SVR Ensemble Regression (#2, MW 0,266). Die Unterschiede sind alle statistisch signifikant mit $p < 0{,}0001$ (Permutationstest auf Werten aus Tabelle C.2). Bei den besten mit den Lernverfahren erzielten Ergebnissen (#16-19) ist im Mittel über alle Datensätze die Rangfolge gleich. Eine interessante Ausnahme ist der Datensatz UNBC, bei dem SVR-E (#17) ähnlich gute Ergebnisse liefert wie RF-K (#18).

Auch bei den Deskriptoren findet sich ein statisch hochgradig signifikanter Unterschied ($p < 0{,}0001$): Der vorgeschlagene Statistikdeskriptor übertrifft mit einem MW von 0,338 (#6) den BoTF-Deskriptor [Bar+14] (0,317, #5). Abweichendes Verhalten zeigt sich auch hier beim Datensatz UNBC, bei dem der BoTF-Deskriptor im Mittel besser ist (vgl. #5) und mit OPR auch das insgesamt beste Ergebnis erzielt (#20). Dieses beste Ergebnis wird in Kombination mit SVR-E und 3D-Koordinaten+Kopfpose erreicht. Bei der Prädiktion der Selbstbeurteilung (VAS) findet sich die insgesamt beste Performance bei zeitlicher Integration mit Statistikdeskriptor, Nutzung von MN-K Bosphorus3D AUs und Kopfpose als Einzelbildmerkmale und RF-K zur Prädiktion.

Die im Mittel besten Einzelbildmerkmale sind die AUs, die mit dem vorgeschlagenen und auf Bosphorus3D trainierten MobileNetV3-large CNN extrahierten wurden (Regression mit und ohne Kopfpose #12-13, dicht gefolgt von Klassifikation mit Kopfpose #15). Sowohl die Regression als auch die Klassifikation von AUs mit der vorgeschlagenen Methode führt zu signifikant besseren Ergebnissen als die Verwendung der OpenFace AUs [BMR15] ($p < 0{,}0001$, mit und ohne Kopfpose). Auch die vorgeschlagenen 3D-Koordinaten als Merkmale liefern signifikant bessere Performance als OpenFace AUs ($p < 0{,}0001$, mit und ohne Kopfpose), sind jedoch signifikant schlechter als die CNN-basierten Merkmale ($p < 0{,}0001$). Bei den besten Ergebnissen (#22-30) zeigt sich klar die überlegene Kopfposeinvarianz der CNN-basierten Merkmale (#27-30) gegenüber 3D-Koordinaten und OpenFace AUs (#22, 24-26), insbesondere bei BioVid-A und BioVid-A7 im Vergleich zwischen Frontalansicht und allen 3 Ansichten.

In den Zeilen #31-44 wurde die zeitliche Integration nicht über einen Deskriptor realisiert, sondern die Merkmalsextraktion, zeitliche Integration und Klassifikation bzw. Regression zusammen (Ende-zu-Ende) in einem CNN trainiert, ausgehend von dem auf Bosphorus3D vortrainierten MobileNetV3-large (MN-R Bos3D). Die Ergebnisse sind im Mittel deutlich schlechter als mit zeitlichen Deskriptoren (#27-30 bzw. #12-15), was insbesondere an dem schlechten Abschneiden mit den kleineren Datensätzen UNBC (200 Videos) und BioVid-A7 liegt (700 Videos bei Frontalansicht und 2.100 Videos bei 3 Ansichten). Bei den größeren Datensätzen BioVid-A (8.700 bzw. 26.100 Videos) und X-ITE (30.800 Videos bei Hitzestimulation und 31.041 Videos bei elektrischer Stimulation) kann das Ende-zu-Ende-Training seine Stärken besser ausspielen, wohingegen es bei den kleineren Datensätzen zur Überanpassung (engl. overfitting) kommt. Den Vergleich zwischen Klassifikation (#31-37, MW 0,327) und Regression (#38-44, MW 0,346) gewinnt die Regression mit einem statistisch signifikanten Vorsprung ($p = 0{,}0024$).

Bei den zeitlichen Integrationsmethoden ist die Nutzung von Maximalwerten (#32, 34, 39, 41: MW 0,355) besser als von Mittelwerten (#31, 33, 38, 40: MW 0,265) mit $p < 0{,}0001$. Pooling von Merkmalen (#33-34 und 40-41: MW 0,318) ist der Entscheidungsfusion (#31-32 und 38-39: MW 0,302) weniger deutlich überlegen, jedoch statistisch signifikant mit $p = 0{,}029$. Das Max. Pooling mit zeitlicher Convolution (#36-37 und 43-44: MW 0,384) ist wiederum statistisch signifikant besser ($p = 0{,}0002$) als reines Max. Pooling von Merkmalen (#34-35 und 41-42: MW 0,367).

Die Gewichtung der Randklassen mit $\lambda = 0{,}9$ wurde für UNBC nicht angewendet, da es hier ein großes Ungleichgewicht der Klassen gibt (stattdessen wurde MID angewendet, siehe Abschnitt 3.2.5). Für die übrigen Datensätze erhöht die Gewichtung die ICC-Performance signifikant (#35, 37, 42, 44: MW 0,378 vs. #34, 36, 41, 43: MW 0,338) mit $p < 0{,}0001$. Wie oben diskutiert

sind jedoch die quantitativen Verbesserungen bei der *Klassifikation* kritisch zu betrachten (z. B. mithilfe von Konfusionsmatrizen), da die Gewichtung hier zu einer massiven Bevorzugung der Randklassen führen kann. Im Extremfall werden die mittleren Klassen gar nicht mehr prädiziert, was der Idee der Intensitätsschätzung widerspricht. Bei der Regression tritt diese Problematik nicht auf, so dass hier eine Bevorzugung der Regression auch bei niedrigeren ICC-Werten sinnvoll sein kann. Daher bewertet der Autor dieser Dissertation die Ergebnisse bei X-ITE mit zeitlichem Pooling mit Regression (#44) als nützlicher und besser als die mit Klassifikation (#37), vgl. auch Abb. 4.7. Die insgesamt besten Performances bei BioVid-A werden mit CNN-Regression, zeitlicher Convolution und Gewichtung $\lambda = 0{,}9$ erreicht (#44).

Vergleich mit expressiver Teilmenge bei X-ITE: Wie schon bei vorherigen Ergebnissen mit dem BioVid-Datensatz zeigt sich auch in Tabelle 4.3, dass die Reduzierung des Gesamtdatensatzes auf die 7 expressivsten Probanden eine deutliche Verbesserung der Performance mit sich bringt. Im Folgenden wird für den Datensatz X-ITE überprüft, ob auch hier eine Reduzierung der Daten auf die expressivsten Probanden zu Verbesserungen führt. Hierfür wurde eine subjektive Kategorisierung der Probanden vorgenommen (vgl. Abschnitt 2.3.1), bei der zufällig ausgewählte Videos der höchsten Schmerzintensität der elektrischen Reizung betrachtet wurden. Das Ergebnis war eine Expressivitätsbewertung je Proband im Wertebereich 1 (keine Mimik beobachtet) bis 5 (starke Schmerzmimik in fast allen betrachteten Videos). 46 Probanden wurden hierbei mit der höchsten Expressivität bewertet und für das Experiment ausgewählt. Evaluiert wurden die CNN-Klassifikation und -Regression mit MN-R Bos3D mit zeitlicher Convolution, jeweils ohne und mit Gewichtung der Randklassen, sowie die Kombination von MN-R Bos3D AU Einzelbildmerkmalen, Statistikdeskriptor und RF-K. Abb. 4.8 stellt die Test-Performances gegenüber. Mit dem vollständigen Datensatz (blau) liefern die CNN-Klassifikation und -Regression ähnliche Ergebnisse im Bereich 0,318 bis 0,370, mit Statistikdeskriptor etwas schlechtere (0,294). Durch die Beschränkung auf die 46 expressivsten Probanden (rot) verbessern sich alle Ergebnisse deutlich auf 0,554 bis 0,595. Hierbei profitiert die Regression noch stärker als die CNN-Klassifikation, so dass die Regression mit Gewichtung ($\lambda = 0{,}8$) beim reduzierten Datensatz die auch zahlenmäßig beste Performance (0,595) erreicht. Die größte Steigerung ist bei der RF-Klassifikation mit Statistikdeskriptor zu beobachten. Hier wird mit einem ICC von 0,587 ein besseres Ergebnis erreicht, als bei der CNN-Klassifikation. In allen Fällen mit CNN (wie auch in den vorangegangen Experimenten) verbessert durch die Gewichtung der Randklassen (hier $\lambda = 0{,}8$) die ICC-Performance.

Vergleich mit verwandten Arbeiten auf UNBC: Der Datensatz UNBC wurde bereits in zahlreichen Vorarbeiten verwendet, von denen sich einige auch der videobasierten Messung der Schmerzintensität mit OPR mit 6 Intensitätensstufen und VAS mit 11 Stufen gewidmet haben. Diese werden hier mit den vorgeschlagenen deskriptorbasierten Erkennungsverfahren verglichen. Bei OPR haben Ruiz et al. [Rui+16] die Nutzung einer Teilmenge von UNBC (157 Videos) vorgeschlagen. Um eine möglichst gute Vergleichbarkeit zu erreichen, wurden daher die besten Modelle aus Tabelle C.2 für diese Teilmenge neu evaluiert. Für VAS wurden in den verwandten Arbeiten alle 200 Videos verwendet. Alle Ergebnisse basieren auf 5-facher Kreuzvalidierung (basierend auf den Probanden), abgesehen von Lopez-Martinez et al. [LMRP17a], deine eine einmalige Aufteilung der Probanden für Training und Test nutzen. Tabelle 4.4 listet die Ergebnisse auf. Neben dem in dieser Arbeit bevorzugten Maß ICC wurde im Kontext von VAS auch MAE (Mean Absolute Error, engl. für mittlerer absoluter Fehler) genutzt, bei dem niedrigere Werte für bessere Ergebnisse sprechen. Sowohl bei OPR (links) als auch bei VAS als Grundwahrheit (rechts) schneiden alle verwandten Arbeiten (oben) deutlich schlechter ab als die vom Autor dieser Dissertation vorgeschlagenen Erkennungsverfahren unten. Auch die Ergebnisse von Lopez-Martinez

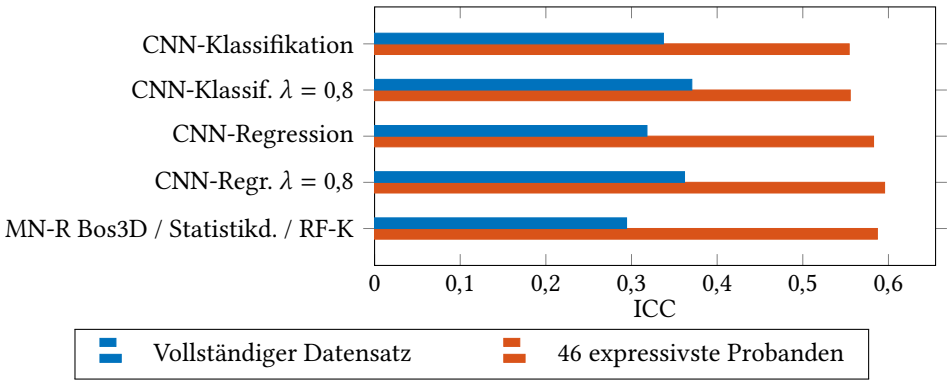


Abbildung 4.8.: Vergleich der Performance mit vollständigem und reduziertem X-ITE Datensatz (elektrische Reize): CNN-Klassifikation und -Regression mit MN-R Bos3D mit zeitlicher Convolution, ohne und mit Gewichtung ($\lambda = 0{,}8$) der Klassen, sowie Random Forest-Klassifikation (RF-K) mit Statistikdeskriptor von MN-R Bos3D AU-Merkmalen.

Grundwahrheit →	**OPR (157 Videos)**	**VAS (200 Videos)**		
Perform.-Maß →	**ICC**	**ICC**	**MAE**	
SIL-OR [Rui+16]	0,56	0,19	2,80	[LMRP17a]
MI-OR [Rui+16]	0,63	0,36	2,46	[LMRP17a]*
HCORF [KP10]	0,30	0,27	2,24	[Liu+17]
HCRF [Qua+07]	0,26	0,35	2,18	[Liu+17]*
MIR [HLC14]	0,63	0,43	1,95	[Xu+19]
MILBoost [SDB13]	0,38	-	2,34	[Ere+20]
MI-HCRF [Rui+16]	0,26	-	2,44	[Szc+21]
MI-DORF [Rui+16]	0,66			
3D-Koordinaten + Kopfpose / BoTF Deskriptor. / SVR-E				
	0,774	0,559	**1,791**	
MN-K Bos3D AU + Kopfpose / Statistikdeskriptor / RF-K				
	0,765	**0,567**	1,889	

* Nutzung von Personeneigenschaften als Merkmale

Tabelle 4.4.: Vergleich mit verwandten Arbeiten auf Datensatz UNBC: Test-Performances für Prädiktion des Observer Pain Rating (OPR) mit 157 Videos wie von [Rui+16] vorgeschlagen (links) sowie der Visuellen Analogskala (VAS) mit dem gesamten Datensatz (rechts). Es werden die Maße ICC (höher ist besser) und MAE (Mean Absolute Error, niedriger ist besser) angegeben. Je Spalte sind die besten Ergebnisse sattgrün und die schlechtesten weiß hinterlegt (mit linearer Abstufung dazwischen).

et al. [LMP17] und Liu et al. [Liu+17], bei denen die Modelle eine *personenspezifische* Expressivitätsbewertung bzw. Eingruppierung nach Alter, Geschlecht und Hautfarbe nutzen, werden von den vorgeschlagenen *nicht personalisierten* Modellen übertroffen.

4.4. Diskussion

In diesem Kapitel wurden ausgehend vom Stand der Technik Methoden zur videobasierten Erkennung von Schmerzen vorgeschlagen und evaluiert. Diese bauen auf Ergebnissen des Kapitels 3 zur einzelbildbasierten Erkennung auf, insbesondere auf dem dort mit Bosphorus3D vortrainierten CNN MobileNetV3-large (MN-R Bos3D) und den dort vorgeschlagen Verfahren zur Merkmalsextraktion. Beides wird mithilfe von zeitlichen Integrationsverfahren auf die Videoerkennung angewendet.

Eine Hypothese dieser Dissertation war, dass sich mit der videobasierten Erkennung eine bessere Performance erreichen lässt als mit der einzelbildbasierten Erkennung. Abb. 4.9 stellt vergleichbare Ergebnisse der einzelbild- und videobasierten Erkennung starker Schmerzen gegenüber. Diese stammen vom Datensatz BioVid, der sowohl in der einzelbildbasierten Erkennung (Variante BioVid-S) als auch der videobasierten Erkennung (Variante BioVid-D) evaluiert wurde. Die Performances sind den Tabellen 3.2, 4.2 und C.1 entnommen. Für alle angestellten Vergleiche, sowohl für den größten BioVid-Datensatz (links, 10.440 Samples) als auch für den kleinsten Datensatz (rechts, 280 Samples), sowohl für verschiedene Merkmale mit SVM-E bzw. RF-K als auch für CNN-basierte Klassifikation (MN-K), zeigt sich eine deutliche Verbesserung durch die Verwendung der Videos (rot) gegenüber der Verwendung einzelner Bilder (blau). Auch bei den Datensatzvarianten mit 87 Probanden und Frontalansicht sowie mit 7 Probanden und 3 Ansichten (hier nicht dargestellt) gibt es nicht einen Fall, bei dem die videobasierte Erkennung schlechter abschneidet. Über alle vier Datensätze findet sich ein statisch höchst signifikanter Unterschied ($p < 0{,}0001$) mit Mittelwerten von 0,255 (einzelbildbasiert) und 0,446 (videobasiert), der die Hypothese klar bestätigt. Von den Dynamikinforationen der Videos profitiert am meisten die Performance mit ausschließlich Kopfposemerkmalen, die mit Einzelbildern nahe 0 (Zufallsentscheidung) liegt. Die besten videobasierten Ergebnisse werden ausgehend vom MN-R Bos3D CNN erzielt, wobei beim größeren Datensatz (links) die Feinjustierung mit zeitlicher Convolution und beim kleineren Datensatz (rechts) die Kombination mit Statistikdeskriptor und RF-K überlegen ist.

Die Abhängigkeit von der Datensatzgröße hat sich auch bei automatisierten Messung der Schmerzintensität gezeigt. Bei kleineren Datensätzen (BioVid-A7 und UNBC) ist die klassische Herangehensweise mit unabhängiger Prädiktion (hier insbesondere RF-K) und Merkmalsextraktion (hier Extraktion von Einzelbildmerkmalen und anschließender Berechnung von zeitlichen Deskriptoren) überlegen. Zur Extraktion von Einzelbildmerkmalen hat sich insbesondere das CNN MobileNetV3-large als performant erwiesen, das auf Bosphorus3D für die AU-Regression trainiert wurde (MN-R Bos3D). Die so extrahierten AUs sowie auch die vorgeschlagenen 3D-Koordinaten liefern im Mittel besser Ergebnisse als die mit OpenFace [BRM16] berechneten AUs. Insbesondere ist MN-R Bos3D bei nicht-frontalen Kopfposen klar überlegen. Bei seitlichen Ansichten von ±45° wird eine ähnliche Performance erreicht wie bei Frontalansichten. Die AU-Regression hat sich als Basis für die videobasierte Erkennung als besser herausgestellt als die AU-Klassifikation, auch wenn AU-Grundwahrheiten nur in geordneten Klassen vorliegen. Dies bestätigt die Hypothese aus Abschnitt 3.2.4, dass die Regression von AU-Intensitäten gegenüber der Klassifikation Vorteile bringt, die im Videokontext messbar werden. Die Kopfpose ist für sich genommen das schwächste Merkmal, liefert jedoch ergänzende Zusatzinformationen und verbessert in Fusion mit anderen Merkmalen oft die Ergebnisse.

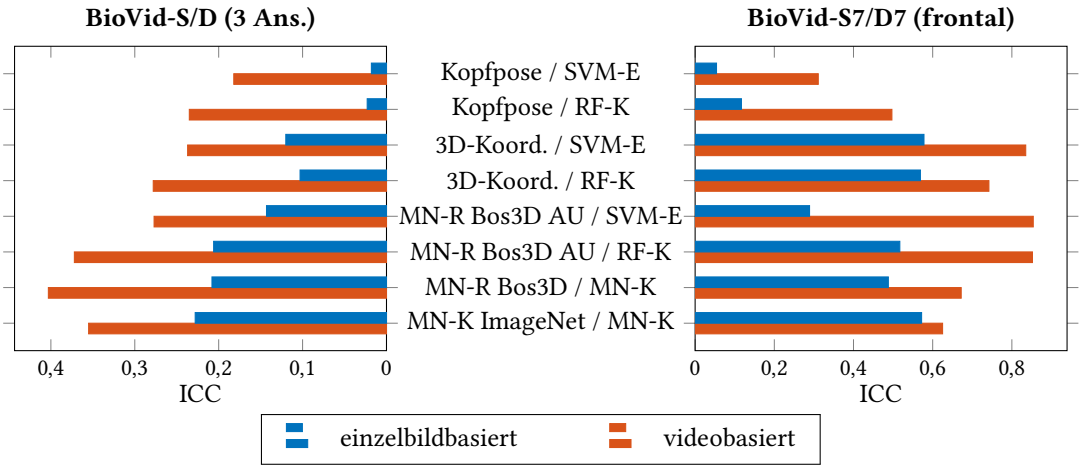


Abbildung 4.9.: Vergleich einzelbild- und videobasierter Erkennung anhand des vollständigen BioVid-Datensatzes mit allen 3 Ansichten (links) und dem Teildatensatz der 7 expressivsten Probanden mit nur Frontalansicht (rechts). Erkennung starker Schmerzen mit verschiedenen Einzelbildmerkmalen / Klassifikatoren (oben) bzw. Ausgangspunkten für die feinjustierte CNN-Klassifikation mit MN-K (unten). Bei den videobasierten Ergebnissen wurde der Statistikdeskriptor bzw. zeitliche Convolution angewendet. Beim linken Plot ist die x-Achse gespiegelt.

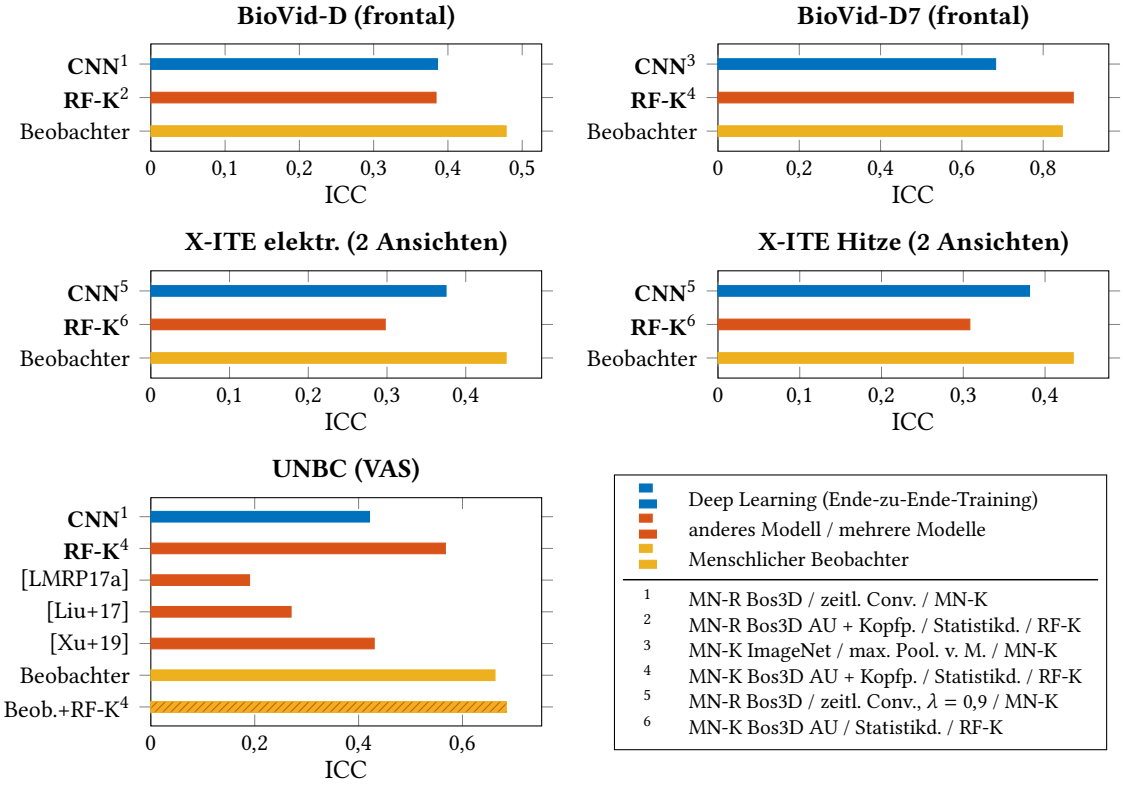


Abbildung 4.10.: Vergleich der videobasierten Erkennung mit der Performance des Menschen und verwandten Arbeiten mit den Datensätzen BioVid-D, BioVid-D7, X-ITE (elektrische Reize und Hitzereize) sowie UNBC. Verglichen werden die besten Ergebnisse mit und ohne CNN als Lernverfahren.

Bei den größeren Datensätzen (BioVid-A und X-ITE) ist die Feinjustierung des MN-R Bos3D CNNs in der Variante mit zeitlicher Convolution die beste Option. Sowohl die zeitlichen Deskriptoren und als auch die zeitliche Convolution können Dynamikinformationen ausnutzen, wie Geschwindigkeit, Beschleunigung, Tendenz oder zeitliche Varianz, was ihnen im Vergleich zur Entscheidungsfusion oder dem Pooling von Merkmalen im Mittel einen Performance-Vorteil verschafft. Bei Entscheidungsfusion und Pooling von Merkmalen ist die vorgeschlagene Nutzung von Maximalwerten besser als von Mittelwerten. Mit dem vorgeschlagene Statistikdeskriptor werden bessere Ergebnisse erzielt als mit dem BoTF Deskriptor von Bartlett et al. [Bar+14]. Ausnahmen finden sich bei UNBC, dem einzigen Datensatz mit Videos unterschiedlicher Länge. Dies wirft die Frage auf, ob BoTF besser für die Kodierung von Informationen aus Videos verschiedener Länge geeignet ist als der vorgeschlagene Statistikdeskriptor. BoTF erfasst Häufigkeiten und Frequenzen sehr detailliert, Amplitudeninformationen werden jedoch weniger stark repräsentiert als im Statistikdeskriptor.

Der Vergleich zwischen den kompletten Datensätzen BioVid und X-ITE mit den Teilmengen der expressivsten Probanden zeigt, dass die Performance sich durch die Reduzierung des Label-Rauschens (vgl. Abschnitt 2.3) deutlich verbessert. Hierbei verkleinert sich jedoch auch der Trainingsdatensatz, was wiederum einen negativen Einfluss auf die Performance von Modellen mit hoher Kapazität haben kann. Von der drastischen Reduzierung auf 7 Probanden bei BioVid profitiert die Erkennung mit Statistikdeskriptor deutlich mehr als die CNN-Feinjustierung, bei der es leicht zu Überanpassung (engl. overfitting) kommen kann. Bei X-ITE führt die Reduzierung auf die 46 expressivsten Probanden zu einer größeren Verbesserung bei der CNN-Feinjustierung, so dass bei der CNN-Regression mit Gewichtung eine bessere Erkennung erreicht wird als mit dem Statistikdeskriptor. Hier, bei reduzierten Label-Rauschen liefert die CNN-Regression bei X-ITE eine bessere ICC-Performance als die CNN-Klassifikation, wie auch bei den anderen Datensätzen. Die Ergebnisse mit BioVid und X-ITE zeigen, dass eine Reduzierung des Label-Rauschens sinnvoll ist, der Datensatz jedoch dabei nicht zu stark verkleinert werden sollte, um das große Potential der CNN-Feinjustierung ausnutzen zu können. Für die Verbesserung der Erkennungsleistung scheint es daher langfristig eine sinnvolle Strategie zu sein, noch größere Datensätze aufzuzeichnen, wenn dabei oder anschließend für eine Reduzierung des Label-Rauschens gesorgt wird.

Die Gewichtung der Randklassen bei der Messung der Schmerzintensitäten hat sich auf BioVid und X-ITE als sehr vorteilhaft erwiesen. Diese Idee hilft zum Umgang mit dem Label-Rauschen in den niedrigen Schmerzintensitäten und auch zur Kompensation der Tendenz zum Mittelwert bei der Regression. Mit Vorsicht sollte sie jedoch bei der *Klassifikation* von Intensitäten eingesetzt werden, da sie im Extremfall dazu führen kann, dass der Klassifikator ausschließlich die Randklassen prädiziert. Ein weiteres Argument für die Bevorzugung der CNN-Regression gegenüber der Klassifikation ist, dass sie die Intensität als Fließkommazahl ausgibt und so mehr Abstufungen prädiziert. Somit kann eine Prädiktion erreicht werden, die im Wertebereich feingranularer ist als die für das Training genutzte Grundwahrheit, da maschinelle Lernmodelle zwischen Trainingsbeispielen interpolieren und generalisieren können.

Wie in den Konfusionsmatrizen in Abb. 4.7 sowie in den Tabellen C.3 und C.4 zu sehen ist, werden insbesondere die niedrigen Schmerzintensitäten häufig mit „kein Schmerz“ verwechselt, da zum einen oft keine sichtbare Schmerzreaktion auftritt und zum anderen auftretende schwach ausgeprägte Mimik schwer von „keiner Mimik“ abzugrenzen ist. Stärkere Schmerzen, bei BioVid und X-ITE insbesondere bei der höchsten Schmerzintensität, gehen mit stärkeren und „zuverlässiger auftretenden“ Reaktionen einher, die dank eines größeren Signal-Rausch-Abstandes auch zuverlässiger zu erkennen sind. Auch bei den hohen Intensitäten kommen Verwechslungen mit „kein

Schmerz“ vor, jedoch deutlich seltener, wenn eine Teilmenge der expressivsten Probanden betrachtet wird, vgl. Tabelle C.3 (b) und (c). Dies veranschaulicht die Einfluss des Label-Rauschens und die Vorteile seiner Reduzierung.

Insgesamt zeigen die Ergebnisse des Kapitels die Leistungsfähigkeit des vorgeschlagenen MobileNet-large Bosphorus3D CNN zur Merkmalsextraktion und Prädiktion in Kombination mit zeitlichen Deskriptoren oder zeitlicher Convolution. Die Konzepte können jedoch auch in anderer Weise kombiniert werden. Beispielsweise kann eine andere CNN-Architektur mit dem Datensatz Bosphorus3D trainiert werden, um die Extraktion von Einzelbildmerkmalen zu realisieren oder einen Ausgangspunkt für videobasiertes Transferlernen mit zeitlicher Convolution zu erhalten. Der vorgeschlagene Statistikdeskriptor kann mit verschiedensten Einzelbildmerkmalen angewendet werden und hat sich nach der ersten Veröffentlichung [Wer+17] und mehrfacher Anwendung in der Schmerzerkennung [WAHW17; Wer+19a; Oth+19b; Wal+20; Oth+21] auch für andere videobasierte Erkennungsaufgaben bewährt [SWAH17; Oth+19a].

Eine zentrale Forschungsfrage ist, ob mit der automatisierten Erkennung eine ähnlich gute Schmerzbeurteilung erreicht werden kann, wie von einem Menschen. Abb. 4.10 vergleicht die ICC-Performance der besten Erkennungsmethoden mit der eines menschlichen Beobachters (gelb). Zur Bestimmung der Performance des Menschen siehe Abschnitt 2.3.5. Für jeden Datensatz wird das jeweils beste Modell mit CNN-Feinjustierung (blau) sowie das beste andere Modell (rot, hier immer RF-K mit Statistikdeskriptor) gezeigt. Für die Prädiktion der VAS (Visuellen Analogskala) auf dem Datensatz UNBC werden außerdem Ergebnisse verwandter Arbeiten (ohne Nutzung personenspezifischer Kontextinformation) aufgeführt. Keine dieser Arbeiten nutzt ein Ende-zu-Ende-Training (blau), sondern es werden mehrere Modelle unabhängig voneinander trainiert und nacheinander angewendet (rot). Bei der Erkennung starker Schmerzen mit den frontalen Daten BioVid-D und BioVid-D7 (oben) gibt es qualitative Unterschiede. Mit BioVid-D, dem größeren Datensatz, ist die Performance mit CNN und RF-K ähnlich (0,386 und 0,384), jedoch deutlich schlechter als die des menschlichen Beobachters (0,478). Im reduzierten Datensatz BioVid-D7 ist die Performance des RF-K (0,874) etwas besser als die des Menschen (0,847). Vermutlich hilft hier die Reduzierung des Label-Rauschens, die mit einer Vereinfachung des Lernproblems einhergeht, das Ergebnis des RF-K im Vergleich zu BioVid-D stärker zu steigern als das des Menschen. Die CNN-Feinjustierung ist hier aufgrund der geringen Datensatzgröße weit abgeschlagen (0,683). Beide X-ITE-Datensätze (elektrische bzw. Hitzereize) zeigen qualitativ ähnliche Ergebnisse bei der Einschätzung der Schmerzintensität: Die CNN-Feinjustierung (0,375 bzw. 0,381) schneidet besser ab als RF-K (0,298 bzw. 0,308), jedoch schlechter als der menschliche Beobachter (0,451 bzw. 0,435). Auch bei UNBC liegt die beste Performance der vorgeschlagenen Modelle (0,567 mit RF-K) deutlich unter der des Menschen (0,663), jedoch gleichzeitig deutlich über der Performance verwandter Arbeiten (0,19 bis 0,43).

Zwar wird die menschliche Leistungsfähigkeit bei der videobasierten Schmerzeinschätzung im Allgemeinen noch nicht erreicht, dennoch wurden hier Erkennungssysteme vorgeschlagen, die bereits einen großen Nutzen für die klinische Praxis haben können. Im Gegensatz zu menschlichen Beobachtern lässt sich die automatisierte Schmerzeinschätzung permanent einsetzen und kann somit ein zeitlich lückenloses Schmerz-Monitoring ermöglichen, was insbesondere für Patienten, die sich nicht selbst zu ihren Schmerzen äußern können, einen großen Mehrwert bringen kann. Für diese Gruppen soll die Schmerzbeurteilung regelmäßig erfolgen [Her+11; Deu09]. Im realen klinischen Alltag herrschen jedoch Zeit- und Kostendruck sowie Personalmangel, wodurch die regelmäßige Beurteilung erschwert bzw. deren Qualität beeinträchtigt werden kann. Hier könnten die vorgeschlagenen Systeme helfen. Neben dem permanenten Monitoring ist auch ein gemeinsamer Einsatz von menschlicher und automatisierter Beurteilung denkbar und vielversprechend. Wird beispielsweise bei UNBC der Mittelwert der Einschätzung des menschlichen

Beobachters und des RF-K gebildet, so verbessert sich die Übereinstimmung mit der Selbsteinschätzung des Patienten (VAS) auf den ICC-Wert 0,684 (im Vergleich zu 0,663 beim menschlichen Beobachter). Für diese ICC-Unterschiede lässt sich bei UNBC mit nur 200 Videos kein aussagekräftiger Signifikanztest berechnen, da die ICC-Werte je Patient bei so kleiner Datenbasis (im Mittel 8 Videos je Patient, zum Teil nur 2 oder 3) instabil werden. Betrachtet man als Maß den mittleren absoluten Fehler (MAE), für den dieses Problem nicht existiert, so findet man dort eine signifikante Verbesserung ($p = 0{,}007$), wenn der menschliche Beobachter durch die Prädiktion des RF-K unterstützt wird, von einem Fehler von 1,765 auf 1,570. Insofern ist auch die gemeinsame Beurteilung der Schmerzen durch Mensch und Maschine eine vielversprechende Möglichkeit.

5. Schlussbetrachtungen

5.1. Zusammenfassung

Die vorliegende Arbeit befasst sich mit der bild- und videobasierten Gesichtsanalyse für eine objektive und automatisierte Messung von akuten Schmerzen. Forschungsgegenstand ist vor allem die Erkennung von Mimik, zum einen von Schmerzmimik, zum anderen von Action Units (AUs) nach dem Facial Action Coding System (FACS) [EFH02], die zur Beschreibung von beliebigen Gesichtsausdrücken dienen.

Ausgehend von der Einführung in die Thematik mit ihren Herausforderungen (Kapitel 1) wurden zunächst die zur Verfügung stehenden Datensätze und Grundwahrheiten vorgestellt und untersucht (Kapitel 2). Anschließend wurden Methoden zur *einzelbildbasierten* Erkennung von Action Units und Schmerzen entwickelt und evaluiert (Kapitel 3). Darauf aufbauend wurden *videobasierte* Methoden zur Erkennung von Schmerzen vorgeschlagen und evaluiert (Kapitel 4). Im Rahmen der Ausführungen wurden die Forschungsfragen aus Abschnitt 1.3 beantwortet:

Inwiefern sind die verfügbaren Schmerzerkennungsdatensätze von Unsicherheiten der Grundwahrheit betroffen? Was bewirken sie? Wie kann mir ihnen umgegangen werden? Alle bekannten Maße für Schmerzen haben ihre Schwächen und sind nur begrenzt zuverlässig und valide. Insofern sind alle Datensätze in gewissem Maße von Unsicherheiten der Grundwahrheit betroffen. Die Selbsteinschätzung wird als der direkteste Zugang zu den subjektiv empfundenen Schmerzen gesehen. Ihre Validität und Reliabilität hängt jedoch von den kognitiven Fähigkeiten des Patienten, seinen bewussten und unbewussten Zielen, dem sozialen Kontext, sowie von seinen Denkweisen und Bewältigungsstrategien ab [PC95; HC04; Cra09; CPG11]. Die Datensätze BioVid und X-ITE nutzen die Selbsteinschätzung zur personenspezifischen Anpassung der Schmerzreizintensitäten, die wiederum als Grundwahrheit dienen. Ob tatsächlich der laut Versuchsdesign beabsichtigte Schmerz empfunden wurde, könnte hier von Zielen der Probanden, wie z. B. wenig leiden zu wollen, sowie von der Begrenzung der maximalen Reizintensität (zur Vermeidung von Gewebeschädigungen) beeinflusst worden sein. Eine beobachtbare Schmerzreaktion setzt erst beim Überschreiten einer gewissen Schmerzintensität ein [PC95; Kun+04]. Bei sehr vielen, auch starken Schmerzreizungen in den Datensätzen BioVid und X-ITE wurde diese offensichtlich nicht überschritten. Zum einen zeigt das Fehlen von Schmerzreaktionen trotz Schmerzempfinden eine Limitierung der beobachterbasierten Schmerzmaße, denn diese können solche Schmerzen nicht vom schmerzfreien Zustand unterscheiden. Zum anderen ergibt sich durch die Nutzung des Schmerzreizes als Grundwahrheit und der (fehlenden) Schmerzreaktion als Modelleingabe eine Inkonsistenz, die maschinelle Lernverfahren herausfordert wie fehlerhafte Label. Eine Alternative ist es, die Schmerzmessung auf die Messung sichtbarer Schmerzreaktionen zu reduzieren, indem Beobachtermaße als Grundwahrheit verwendet werden. Jedoch haben auch diese Unsicherheiten. In der Klinik verwendete Maße werden beeinflusst von persönlichen Erfahrungen und Einstellungen des Beobachters, seinem Wissen und der Beziehung zum Leidenden. FACS-basierte Maße zeichnen sich demgegenüber durch Objektivität aus, jedoch ist dem

Autor dieser Dissertation kein solches Maß bekannt, dass für einen klinisches Schmerzmonitoring hinreichend sensitiv und spezifisch ist. Das weit verbreitete Maß PSPI ist insbesondere außerhalb des Schmerzkontextes problematisch, da es Schmerzmimik nicht von anderer Mimik mit gemeinsamen AUs unterscheiden kann, wie sie z. B. bei Ekel, Angst, Trauer, Freude und Schlaf auftritt.

Die Unsicherheiten in den Grundwahrheiten wirken wie Label-Rauschen und sind eine Herausforderung für die maschinellen Lernverfahren – insbesondere im Zusammenhang mit der begrenzten Verfügbarkeit von Daten. Da die Unsicherheiten auch die Testdaten betreffen, begrenzen sie die beste erreichbare Test-Performance. Um die Performance-Werte dennoch einordnen zu können, wurde für einige Erkennungsaufgaben die Performance von menschlichen Beobachtern ermittelt und mit denen der automatisierten Methoden verglichen. Um den Einfluss des Label-Rauschens auf das maschinelle Lernen zu zeigen, wurden Teildatensätze mit geringerem Label-Rauschen erzeugt, für die lediglich die expressivsten Probanden einbezogen wurden. Mit den expressivsten Probanden wurden in allen Fällen deutlich bessere Performances erreicht als mit allen Probanden. Da das Label-Rauschen einen dominanten Einfluss auf das Lernen ausübt aber auch andere Fragestellungen betrachtet werden sollten, wurden neben der Intensitätsmessung die Zweiklassenprobleme mit dem stärksten Schmerzreiz und ohne Schmerzreiz betrachtet, die weniger von Label-Rauschen betroffen sind als die niedrigen Schmerzintensitäten.

Welche Computer-Vision- und Machine-Learning-Methoden erreichen die beste Performance? Die Verarbeitungskette für die Mimikerkennung umfasst Bildaufnahme, Gesichtsdetektion, Landmarkenlokalisierung, und Gesichtsnormierung, die unten gesondert besprochen werden, sowie Merkmalsextraktion und Prädiktion. Für die letzteren beiden Schritte wurden verschiedene neue Verfahren vorgeschlagen, evaluiert und mit verwandten Arbeiten verglichen. Dies geschah im Kontext vergleichsweise kleiner Datensätze, gewisser Unsicherheit in der Grundwahrheit und zum Teil starker Ungleichverteilungen der Klassenzugehörigkeiten. Betrachtet wurden die klassische Herangehensweise mit Merkmalsextraktion und anschließendem unabhängig gelernten Prädiktionsmodell sowie Ende-zu-Ende-Lernen mit einem CNN, das das Prädiktionsmodell mit einer für die Aufgabe optimierten Merkmalsextraktion vereint. Besonders erfolgreich war das Transferlernen mit CNN ausgehend von einem Vortraining mit dem Datensatz Bosphorus3D. Dieser wurde vom Autor dieser Dissertation auf Basis des gut annotierten und sehr vielfältigen Datensatzes Bosphorus erzeugt, indem letzterer aus 49 verschiedenen Blickwinkeln gerendert wurde, um auch nicht-frontale Kopfposen abzudecken. Als Ausgangspunkt für verschiedene Ansätze wurde das CNN MobilNetV3-large mit Bosphorus3D zur Regression von Facial Action Units vortrainiert („Bosphorus-CNN“). Darauf aufbauende Ergebnisse waren signifikant besser als die Nutzung von ImageNet als Quelldomäne für das Transferlernen, einer weit verbreiteten Herangehensweise für die Nutzung von CNNs mit kleinen Datensätzen. Dabei waren beide Varianten des Transferlernens signifikant besser als das Training ausgehend von einer zufälliger Initialisierung. Für die einzelbildbasierte Erkennung wurden bei allen acht Datensätzen die besten Performances mit Feinjustierung des Bosphorus3D-CNN erreicht, zum Teil mit zusätzlichem Multi-Task-Lernen oder der Fusion mit menschengemachten Merkmalen. Welche Variante die besten Ergebnisse erzielte, war datensatzabhängig. Kleinere Datensätze konnten weniger vom CNN profitieren, insbesondere BioVid-S7 (frontal) mit lediglich 224 Trainingsbeispielen, bei dem SVM-E/SVR-E und RF mit menschengemachten Merkmalen in verschiedenen Varianten besser abschnitten als die CNN-basierte Erkennung. Hier schnitten unter anderem die vorgeschlagenen niedrigdimensionalen 3D-Koordinatenmerkmale sehr gut ab, die auch in Fusion mit anderen Merkmalen oft zu Verbesserungen führten. Auch die Kopfposemerkmale, die mit einem neu vorgeschlagenen Verfahren berechnet wurden, führten in Fusion mit anderen Merkmalen zu Verbesserungen. Das Multi-Task-Lernen hat sich für den Umgang mit kleinen Datensätzen als weniger

hilfreich erwiesen als erwartet. Es eignet sich in vielen Fällen als ein (zusätzliches) Verfahren zur Regularisierung, hat jedoch einen deutlich geringeren Einfluss als das Transferlernen.

Die Ausnutzung von zeitlichen Informationen hat sich wie erwartet positiv auf die Erkennungsleistung ausgewirkt. Im direkten Vergleich mit BioVid war die videobasierte Erkennung starker Schmerzen statistisch höchst signifikant besser als die einzelbildbasierte Erkennung. Für die videobasierte Erkennung wurden verschiedene zeitliche Integrationsmethoden vorgeschlagen, evaluiert und mit verwandten Arbeiten verglichen. Vorgeschlagen wurden ein Statistikdeskriptor sowie verschiedene Varianten von Video-CNN, die mit Transferlernen auf dem Einzelbild-CNN aufbauen. Die besten Ergebnisse wurden jeweils auf Basis des Bosphorus3D-CNN aus dem Einzelbildkapitel erreicht, wobei es jedoch auf verschiedene Weise angewendet wurde.

Bei den kleineren Datensätzen (BioVid-D7, BioVid-A7 und UNBC) war die klassische Herangehensweise mit fester Merkmalsextraktion und anschließendem unabhängig gelernten Prädiktionsmodell erfolgreicher als Ende-zu-Ende-Lernen. Zur Merkmalsextraktion wurden hier Einzelbildmerkmale extrahiert und mit zeitlichen Deskriptoren zusammengefasst. Als Einzelbildmerkmale schnitten die mit dem Bosphorus3D-CNN extrahierten AU-Intensitäten sehr gut ab, insbesondere waren sie anderen Merkmalen bei nicht-frontalen Kopfposen klar überlegen. Bei seitlichen Ansichten von ±45° wird mit den AUs des Bosphorus3D-CNN eine ähnliche Performance erreicht wie bei Frontalansichten, wohingegen die Ergebnisse mit anderen Merkmalen stark abfallen. Bei den größeren Datensätzen (BioVid-D, BioVid-A und X-ITE) zeigt sich die Stärke des Ende-zu-Ende-Lernens. Hier ist die Feinjustierung des Bosphorus-CNN in der Variante mit zeitlicher Convolution die beste Option. Sowohl die zeitlichen Deskriptoren und als auch die zeitliche Convolution können Dynamikinformationen ausnutzen, wie Geschwindigkeit, Beschleunigung, Tendenz oder zeitliche Varianz, was ihnen im Vergleich zur Entscheidungsfusion oder dem Pooling von Merkmalen, die ebenfalls untersucht wurden, im Mittel einen Performance-Vorteil verschafft. Als Ausgangspunkt für Transferlernen war das Bosphorus3D-CNN besser als das ImageNet-basierte CNN, wie schon bei der Einzelbilderkennung. Eine höhere Gewichtung der Randklassen bei der Messung von Schmerzintensitäten hat sich auf BioVid und X-ITE als vorteilhaft erwiesen, insbesondere für die Regression. Die Regression von Schmerz- und AU-Intensitäten liefert in Gegensatz zur Klassifikation Fließkommazahlen, die Dank der Fähigkeit des maschinellem Lernens zur Interpolation und Generalisierung im Wertbereich feingranularer sind als die Grundwahrheit. Bei mehr als zwei Intensitätsstufen in der einzelbildbasierten Erkennung war die Regression der Klassifikation auch in der ICC-Performance überlegen. Als Ausgangspunkt für zeitliche Deskriptoren liefert die Regression von AU-Intensitäten detailliertere Dynamikinformationen als die Klassifikation, was sich positiv auf die videobasierten Ergebnisse mit AU-Regression ausgewirkt hat. Für das Prädiktionmodell der videobasierten Messung der Schmerzintensität sind die quantitativen Ergebnisse weniger klar. Bei deskriptorbasierten Ansätzen hat die Klassifikation (mit SVM-E und RF) in der Regel besser funktioniert. Die Performances mit CNN-Regression sind in den meisten Fällen besser als mit CNN-Klassifikation, wobei eine Reduzierung des Label-Rauschens sich bei der Regression noch stärker positiv auswirkt, als bei der Klassifikation. Bei ähnlicher quantitativer Performance empfiehlt der Autor dieser Dissertation die Verwendung der Regression aufgrund der qualitativen Vorteile der Prädiktion.

Verschiedene Vergleiche mit Verfahren des Standes der Technik zeigen die Konkurrenzfähigkeit bzw. Überlegenheit der vorgeschlagenen Ansätze. So konnten bei der Messung der AU-Intensität auf FERA 2017 sowie bei der Messung der Schmerzintensität mit VAS und OPR auf UNBC die zuvor publizierten Ergebnisse anderer Autoren deutlich übertroffen werden. Die videobasierte Erkennung starker Schmerzen und Messung von Schmerzintensitäten hat gezeigt, dass das vorgeschlagene Bosphorus3D-CNN dem weit verwendeten OpenFace [BRM16] bei der Extraktion von AU-Merkmalen klar überlegen ist, insbesondere bei nicht-frontalen Kopfposen. Auch hat

der vorgeschlagene Statistikdeskriptor zur zeitlichen Integration in den meisten Untersuchungen (Ausnahme bei UNBC) bessere Performances geliefert als der BoTF-Deskriptor von Bartlett et al. [Bar+14]. Außerdem wurde die oft angewendete zeitliche Entscheidungsfusion mit mehreren vorgeschlagenen Ansätzen übertroffen.

Wie kann erreicht werden, dass das System bei möglichst vielen Kopfposen gut funktioniert? Bei der Mimikerkennung bereiten nicht-frontale Kopfposen meist Probleme, da diese in den meisten Trainingsdatensätzen unterrepräsentiert sind. Zur Verbesserung der Kopfposeinvarianz wurden zwei Ansätze vorgeschlagen und untersucht, zum einen die Frontalisierung des Gesichts als Vorverarbeitungsschritt (algorithmischer Ansatz), zum anderen das Training der Mimikerkennung mit Datensätzen mit mehreren Ansichten auf das Gesicht (Verbesserung der Trainingsdaten). Zur Frontalisierung, d. h. um eine frontale Ansicht auf ein beliebig gedrehtes Gesicht zu generieren, wurde das Gesichtsnormierungsverfahren FaNC entwickelt. Das Ziel war es, eine größere Kopfposeinvarianz zu erreichen, indem die durch die Kopfpose hervorgerufenen Bildunterschiede reduziert werden. So sollte auch beim Training mit ausschließlich oder überwiegend frontalen Kopfposen eine bessere Generalisierung auf nicht-frontalen Kopfposen möglich werden, um die existierenden überwiegend frontalen Datensätze gut ausnutzen zu können. Für die Datensätze FERA 2017 (Messung von AU-Intensitäten) und Multi-PIE (Klassifizierung verschiedener Mimikkategorien) konnte mit FaNC eine bessere Generalisierung erreicht als mit anderen Normierungsverfahren, jedoch nicht bei der Erkennung von Schmerzmimik. Bei letzterer treten in den frontalisierten Bildern oft Artefakte (unrealistische Verzerrungen) auf, da Schmerzmimik für das *Training von FaNC* nicht berücksichtigt werden konnte. Als bessere Option, um Kopfposeinvarianz für beliebige Mimik zu erreichen, hat sich das Training mit Datensätzen mit mehreren Ansichten auf das Gesicht bewährt, die jeden Moment der Aufzeichnung aus verschiedenen Blickwinkeln und damit mit verschiedenen Kopfposen zeigen. Mit diesem Ansatz ließ sich mit CNNs und nicht-frontalisierenden Normierungsverfahren eine bessere Performance erzielen als mit frontalisierenden Verfahren. Infolge dessen wurde die Idee entwickelt und umgesetzt, den vielfältigen und gut annotierten Datensatz Bosphorus aus verschiedenen Ansichten zu rendern und den resultierenden Bosphorus3D-Datensatz als Ausgangspunkt für Transferlernen und Multi-Task-Lernen mit CNN zu nutzen. Das Ziel hierbei war es, über die Kopfposevielfalt im Datensatz Kopfposeinvarianz bei der Merkmalsextraktion innerhalb des CNN zu erreichen. CNNs verfügen über eine große Anzahl an Parametern und eine große Kapazität, so dass sie die große Varianz bei verschiedenen Kopfposen gut lernen können. Bei Verwendung des Bosphorus3D-CNN als Merkmalsextraktor waren die Performances mit nur frontaler und mit allen Ansichten ähnlich, was belegt, dass eine gute Poseinvarianz erreicht wurde.

Kann die Erkennung mit niedrigen Hardwarekosten realisiert werden, d. h. ohne Spezialkameras, mit geringer Anzahl von Kameras, mit niedriger Auflösung? Diese Arbeit setzt für den Testfall, d. h. die Anwendung des Systems bewusst auf Standardfarbkameras, die z. B. in Smartphones oder als Webcams allgegenwärtig sind, um einfach und vielfältig einsetzbare, kostengünstige Erkennungssysteme zu entwickeln. Durch die Verbesserung der Kopfposeinvarianz in dieser Arbeit ist die Mimikerkennung auch bei Kopfdrehungen bis ±45° ohne oder mit nur geringen Performance-Einbußen möglich. Hierdurch kann mithilfe einer seitlichen Kamera und einem Spiegel, wie bei der X-ITE Database umgesetzt, mit nur *einer Kamera* ein Bewegungsspielraum von ±90° abgedeckt werden. Somit hätte die Schmerzüberwachung für ein typisches Krankenbett einen direkten oder durch den Spiegel einen indirekten Blick auf das Gesicht des Patienten, unabhängig davon wie dieser seinen Kopf dreht. Weitere Kameras wären nur nötig,

um bei Verdeckungen (z. B. durch eine andere Person) oder dem Verlassen des Sichtbereichs keinen Informationsverlust hinnehmen zu müssen. Eine sehr hohe räumliche Auflösung des Bildes bzw. Gesichtes ist nicht nötig, weder für die Gesichtsdetektion, noch für die Landmarkenlokalisierung oder folgende Schritte. Die Untersuchungen haben gezeigt, dass die Mimikerkennung mit Gesichtsbildern mit einem Augenabstand von 50 oder mehr Pixeln ähnlich gut funktioniert. Starke Performance-Einbußen gab es erst bei unter 25 Pixeln Augenabstand. Somit können auch niedriger aufgelöste, preisgünstigere Kameras eingesetzt werden oder alternativ über einen größeren Öffnungswinkel ein größerer Bewegungsspielraum für den Patienten toleriert werden.

Kann das entwickelte Erkennungssystem eine ähnlich gute Beurteilung der Schmerzen erreichen wie ein Mensch? Untersucht wurde die einzelbildbasierte Unterscheidung zwischen hitzestimulierten Schmerzen an der Toleranzschwelle und keinen Schmerzen mit den Datensätzen BioVid-S und BioVid-S7 (beide nur frontale Kameraansicht) und zwei menschlichen Beobachtern. Die Performance des vorgeschlagenen CNN war bei den vier Vergleichen in zwei Fällen besser als der Mensch, in einem Fall ähnlich gut und in einem Fall schlechter, so dass hier zumindest von einer ähnlichen Leistungsfähigkeit gesprochen werden kann. Bei der videobasierten Unterscheidung mit BioVid-D und BioVid-D7 (beide frontal) waren die Ergebnisse ebenfalls nicht eindeutig. Für BioVid-D7 mit reduziertem Label-Rauschen konnte das automatisierte Modell eine leicht bessere Performance erreichen als der menschliche Beobachter. Beim vollständigen Datensatz BioVid-D wurde mit den automatisierten Methoden eine niedrigere Performance erreicht als vom Menschen. Ähnlich war es bei der Messung der Schmerzintensität bei X-ITE (elektrische und Hitzereize, jeweils mit zwei Kameraansichten) sowie bei UNBC. Im Allgemeinen wurde die menschliche Leistungsfähigkeit bei der videobasierten Schmerzeinschätzung somit noch nicht erreicht. Dennoch wurden hier Erkennungssysteme vorgeschlagen, die bereits einen großen Nutzen für die klinische Praxis haben können. Im Gegensatz zu menschlichen Beobachtern lässt sich die automatisierte Schmerzeinschätzung permanent einsetzen und kann somit ein zeitlich lückenloses Schmerz-Monitoring ermöglichen, was insbesondere für Patienten, die sich nicht selbst zu ihren Schmerzen äußern können, einen großen Mehrwert bringen kann. Auch die gemeinsame Beurteilung der Schmerzen durch Mensch und Maschine ist eine vielversprechende Möglichkeit, wie anhand von UNBC gezeigt werden konnte. Insofern wäre die Technologie gut geeignet um menschliche Beobachter zu entlasten und zu unterstützen, auch wenn sie ihnen im Bezug auf die Übereinstimmung mit den Grundwahrheiten noch nicht ebenbürtig ist.

5.2. Ausblick

Hier werden Richtungen für weiterführende Arbeiten vorgeschlagen, die vom Autor dieser Dissertation zum Teil bereits in Werner et al. [Wer+19b] veröffentlicht wurden.

Weitere Verbesserung der Schmerzmessung Wie sich in den Untersuchungen mehrfach gezeigt hat, ist Label-Rauschen ein schwerwiegendes Problem für die Schmerzmessung und seine Reduzierung vielversprechend für die Verbesserung der Performance. Ansatzpunkte hierfür liegen in der Weiterentwicklung von Versuchsdesigns für neue Datensätze und von Methoden zur Erkennung von Label-Rauschen, sowie in der Evaluierung von Lernmethoden, die eine bessere Performance bei Label-Rauschen versprechen, wie z. B. Co-Teaching [AU20]. Auch die Entwicklung besserer Grundwahrheiten ist sinnvoll, z. B. weiterer FACS-basierter (und dadurch objektiver) Grundwahrheiten für Schmerzmimik, bei denen (1) AUs nicht-linear kombiniert werden, (2) der zeitliche Kontext einbezogen wird, und (3) AUs von Gesichtsausdrücken, die bei Schmerzen

nicht vorkommen, „negativ“ in die Berechnung einbezogen werden (um Schmerzmimik besser von anderer Mimik abzugrenzen). Der Datensatz Bosphorus3D hat sich als sehr nützlich erwiesen, ist jedoch nicht optimal für die Schmerzerkennung, da in ihm keine Schmerzmimik vorkommt. Insofern wäre es vielversprechend, Bosphorus3D um AU-annotierte Schmerzdaten zu ergänzen. Weiterhin gibt es für die Schmerzmessung noch viel Raum für die Untersuchung, den Vergleich und die Weiterentwicklung von Computer-Vision-Methoden, die hier nicht betrachtet wurden, wie z. B. Generative Adversarial Networks, rekurrente neuronale Netze oder probabilistische grafische Modelle. Auch die vollständige Optimierung der Tuning-Parameter der verwendeten Algorithmen, die aufgrund des Rechenaufwandes im Rahmen dieser Arbeit nicht möglich war, könnte die Performance weiter verbessern. Ein weiterer Ansatzpunkt für Verbesserungen ist die Personalisierung, d. h. die Anpassung der Erkennung an den jeweiligen Patienten.

Weitere Verbesserung der Kopfposeinvarianz Die Unabhängigkeit der Mimikerkennung von der Kopfpose konnte mit Datensätzen, die aus 3D-Daten gerenderten wurden (Bosphorus3D und FERA 2017), deutlich verbessert werden. Die gerenderten Bilder enthalten jedoch aufgrund von Messungenauigkeiten in den 3D-Daten Artefakte (unrealistische Verzerrungen), welche die Performance negativ beeinflussen können. Dieser Artefakte könnten vermieden werden, indem neue Schmerzdatensätze mit vielen Kameras aus verschiedenen Ansichten aufgezeichnet würden. Hierdurch würden realistische Trainingsdaten für viele Kopfposen zur Verfügung stehen, mit denen eine bessere Performance für ein Erkennungssystem erreicht werden könnte, das am Ende mit nur einer Kamera auskommt. Mit den passenden Trainingsdaten wäre es aus Sicht des Autors möglich, eine zuverlässige Schmerzerkennung auch bei Profilansichten des Gesichtes zu ermöglichen. Ein alternativer, jedoch deutlich herausfordernder Ansatz ist die Verbesserung des Gesichtsnormierungsverfahrens FaNC, insbesondere indem Schmerzmimik und andere nicht repräsentierte Gesichtsausdrücke sowie eine größere Vielfalt an Identitäten (Personen) in den Trainingsdatensatz integriert werden. Außerdem wäre eine Weiterentwicklung des Ansatzes nötig, um Aufdeckungen bei großen Nickwinkeln (nach oben/unten gerichteter Kopf) besser zu behandeln, denn hier gibt es mit FaNC insbesondere an Nasen und Augenhöhlen Artefakte.

Multi-Modale Erkennung Die Idee der Einbeziehung weiterer Sensorik für Biosignale und Verhaltensreaktionen (Audio und Körperbewegungen) wurde in dieser Arbeit nicht betrachtet, um im vorgegebenen Umfang eine detaillierte Untersuchung der mimikbasierten Schmerzerkennung möglich zu machen. Der Autor dieser Dissertation hat sich jedoch in anderen Publikationen mit anderen Sensormodalitäten auseinandergesetzt [Wer+19a; Wer+14b] und konnte z. B. mithilfe der BioVid Heat Pain Database zeigen, dass die Korrektklassifikationsrate durch die Fusion von Mimik, EDA, EKG, und EMG verbessert werden kann [Wer+14b; Käc+15a; Käc+15b]. Hier ergibt sich großes Forschungspotential zu bisher nicht betrachteten Modalitäten, wie der Thermografie und der Körperbewegung bei X-ITE, sowie zum Einsatz komplexerer Fusionsverfahren. Des Weiteren gibt es Forschungen zur kontaktfreien Messung von Vitalparametern, so auch vom Autor dieser Dissertation zur Messung des Herzschlages anhand von Videobildern [Wer+14a; RWAH16; RWAH19], die auch für die Schmerzerkennung eingesetzt werden könnten.

Erforschung verwandter Erkennungsprobleme Hier wurde die Messung von akuten Schmerzen betrachtet. Gesellschaftlich noch deutlich relevanter sind chronische Schmerzen. Damit sind sie auch ein wichtiges Forschungsthema, ihre Messung ist jedoch noch deutlich herausfordernder als bei akuten Schmerzen. Weitere Themen für zukünftige Arbeiten sind z. B. neuropathische und viszerale Schmerzen, die Unterdrückung oder Verstärkung von Schmerzen, die Qualität und Lokalität von Schmerzen, die Unterscheidung von echter und gespielter Schmerzmimik oder die Differenzierung von Schmerzen und Emotionen, die oft auch zusammen auftreten.

Schritte in Richtung klinischer Anwendung Für die klinische Anwendung der Technologie sind Verbesserungen bezüglich der Robustheit der Computer-Vision-Verfahren nötig, insbesondere in Bezug auf Beleuchtung und Verdeckungen. Ein Ansatz für Beleuchtungsinvarianz ist die Nutzung einer nicht sichtbaren Beleuchtung im Nahinfrarotbereich (NIR) in Zusammenhang mit Videokameras, die das Licht in diesem Spektralbereich aufnehmen. Hierdurch würde ein permanentes Schmerzmonitoring auch nachts möglich, ohne für Patienten störende Beleuchtung. Für das Patientenmonitoring müssen die Verfahren für Echtzeitdatenauswertung angepasst werden. Zur Anpassung der Schmerzerkennungsmodelle auf die klinischen Gegebenheiten, Anforderungen und Schmerzmodalitäten ist Transferlernen mit klinischen Datensätzen vielversprechend. Weitere entscheidende Schritte sind die Validierung der Erkennungs- und Messsysteme mit den Patientengruppen, für die die Technologie eingesetzt werden soll, sowie Untersuchungen zur Akzeptanz der Technologie und zur Kosten-Wirksamkeitsanalyse. Bei der Validierung sollten auch Beobachtereinschätzungen von mehreren medizinischen Fachkräften eingeholt werden, die in ihrer Arbeit mit den Patienten regelmäßig Beobachterschmerzskalen anwenden, um aussagekräftige Vergleiche zwischen der neuen automatisierten Schmerzmessung und bereits etablierten Beobachterwerkzeugen anstellen zu können. So wird sich langfristig zeigen, wie viel die entwickelten Technologien zur Verbesserung der Schmerzmessung und -behandlung der Patienten beitragen können.

Anhang A. Performance-Maße

Im Folgenden werden einige weit verbreitete Maße zur Beurteilung der Güte (engl. Performance) von Klassifikatoren bzw. Regressoren verglichen, um das am Besten geeignete Maß auszuwählen. Die Ausführungen basieren zum Teil auf Werner et al. [WSAH15]. Folgende Maße werden betrachtet:

- Die **Korrektklassifikationsrate (engl. Accuracy)** berechnet sich als Quotient aus der Anzahl aller *korrekt* klassifizierten Fälle und der Anzahl *aller* klassifizierten Fälle und entspricht der Abschätzung der Wahrscheinlichkeit der korrekten Klassifikation.
- Das **F1-Maß** wird als harmonisches Mittel der Genauigkeit (engl. Precision) und Trefferquote (engl. Recall) berechnet [Yan99]. Ein Problem mit dem F1-Maß ist, dass die Anzahl der korrekt negativ klassifizierten Fälle nicht eingeht und damit beliebig variieren kann ohne das Maß zu beeinflussen [Pow11]. Dieses Problem lässt sich mittels Macro-Averaging [Yan99] beheben, wobei das **MF1-Maß** entsteht. Dabei wird jede der Klassen einmal als positive Klasse betrachtet und ein F1-Wert für diese Klasse berechnet. Anschließend werden die F1-Werte aller Klassen gemittelt um den MF1-Wert zu erhalten. Dieses Vorgehen ermöglicht auch die Nutzung des MF1-Maßes für Probleme mit beliebiger Anzahl von Klassen.
- Die **Pearson-Korrelation (engl. Pearson Correlation Coefficient)** schätzt den Grad des linearen Zusammenhangs zwischen Grundwahrheiten Y und Modellprädiktionen $\hat{Y}$ bei intervallskalierte Daten, wie Schmerzintensitäten. Sie wird berechnet aus der Kovarianz von Y und $\hat{Y}$, geteilt durch die Wurzel des Produktes der Varianzen von Y und $\hat{Y}$.
- Die **Intra-Klassen-Korrelation (engl. Intra-Class-Correlation, ICC)** quantifiziert die Übereinstimmung von k Beurteilern bei der Bewertung von n Samples. Es gibt verschiedene Varianten der ICC. Wie auch in Vorarbeiten zur Mimikintensitätsschätzung [Val+17; WSAH15] wird hier ICC(3,1) nach Shrout und Fleiß [SF79] verwendet. Dabei werden die Grundwahrheiten Y als eine Beurteilung aufgefasst und die die Modellprädiktionen $\hat{Y}$ als zweite Beurteilung, das heißt hier ist immer $k = 2$. Zur Berechnung werden die Werte von Y und $\hat{Y}$ zu einer Matrix mit den Elementen $y_{i,j}$ zusammengefasst, mit dem Target- bzw. Sample-Index $i \in \{1, ..., n\}$ und dem Judges- bzw. Beurteiler-Index $j \in \{1, 2\}$. Basierend auf einer Varianzanalyse wird der $ICC(3,1)$ wie folgt berechnet [SF79]:

$$ICC(3,1) = \frac{BMS - EMS}{BMS + (k-1) \cdot EMS}, \tag{A.1}$$

  wobei *BMS (Between target Mean Squares)* und *EMS (Residual Mean Squares)* sich wie folgt aus den empirischen Daten ergeben:

$$BMS = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (T_i - M)^2 \cdot k, \tag{A.2}$$

$$EMS = \frac{1}{(k-1) \cdot (n-1)} \cdot (WSS - JSS). \tag{A.3}$$

Hierfür sind wiederum *WSS (Within target Sum of Squares)*, *JSS (between Judges Sum of Squares)*, T_i (mean per Target), J_j (mean per Judge), *und* M (total Mean) zu berechnen:

$$WSS = \sum_{i=1}^{n}\sum_{j=1}^{k}(y_{i,j} - T_i)^2, \qquad JSS = \sum_{j=1}^{k}(J_j - M)^2, \tag{A.4}$$

$$T_i = \frac{1}{k}\sum_{j=1}^{k} y_{i,j}, \qquad J_j = \frac{1}{n}\sum_{i=1}^{n} y_{i,j}, \tag{A.5}$$

$$M = \frac{1}{kn}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{i,j}. \tag{A.6}$$

- Der **mittlere quadratische Fehler (engl. Mean Squared Error, MSE)** berechnet sich wie folgt aus den Grundwahrheiten y_i und Modellprädiktionen $\hat{y}_i$ (auf Intervallskalen):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{A.7}$$

- Der **mittlere absolute Fehler (engl. Mean Absolute Error, MAE)** wird wie der MSE berechnet, jedoch wird das Quadrat durch den Absolutbetrag $|\cdot|$ ersetzt.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{A.8}$$

Die Pearson-Korrelation liefert oft ähnliche Werte wie die Intra-Klassen-Korrelation (ICC), hat jedoch folgende hier ungünstige Eigenschaften. Erstens beschreibt sie lediglich den Grad des linearen Zusammenhangs zwischen Grundwahrheit und Modellprädiktion, nicht jedoch den Grad der absolute Übereinstimmung der Werte. Hat beispielsweise ein Regressionsmodell ein Bias bei dem niedrige Prädiktionen überschätzt und hohe Prädiktionen unterschätzt werden, wie es bei Random-Forest-basierter Regression oft auftritt [Son15], hat dieses Problem keinen Einfluss auf den Korrelationswert, da dieser invariant gegenüber linearer Transformation der Prädiktionen ist. Zweitens ist die Pearson-Korrelation nicht definiert, wenn in den Prädiktionen denn keine Varianz auftritt, da in diesem Fall durch 0 geteilt wird. Dieses Problem tritt insbesondere auf, wenn Klassenhäufigkeiten stark ungleich verteilt sind und dies beim maschinellen Lernens nicht beachtet wird. Dies führt oft dazu, dass ein Modell gelernt wird, welches immer die häufigste Klasse ausgibt und somit keine Varianz in der Prädiktion hat.

Stark ungleiche Verteilungen sind für die meisten der betrachteten Performance-Maße problematisch, wie im Folgenden gezeigt wird. Wir betrachten vier Klassifikationsprobleme, deren Klassen sich auf Intervallskalen einordnen lassen, z. B. Schmerzintensitäten: zwei mit binärer Klassifikation und zwei Dreiklassenprobleme. Für jedes Klassifikationsproblem werden verschiedene mögliche Prädiktionsergebnisse angenommen und die zugehörigen Performance-Kennzahlen verglichen. Folgende Prädiktionsergebnisse werden betrachtet und in Tabelle A.1 in Form von Konfusionsmatrizen angegeben:

Optimum: Alle Testfälle werden korrekt klassifiziert. Dies ist das bestmögliche Ergebnis.

Modell A: Ausgehend vom Optimum wurde ein normalverteiltes Rauschen mit einer Standardabweichung von 0,5 auf die Prädiktionen addiert, um eine Konfusionsmatrix zu erhalten, wie sie realen Klassifikationsproblemen oft auftritt.

Modell B: Ausgehend vom Optimum wurde ein normalverteiltes Rauschen mit einer Standardabweichung von 0,9 auf die Prädiktionen addiert. Dies simuliert ebenfalls ein typisches Klassifikationsergebnis, jedoch mit schlechterer Performance als Modell A.

Zufällig: Die Prädiktionen sind gleichverteilt, das heißt beim Training konnten keine Muster identifiziert werden, die korrekte Prädiktionen ermöglichen.

Immer 0: Das Modell gibt immer die am häufigsten vorkommende Klasse (hier 0) zurück. Dieses Modell wird häufig als trivialer Klassifikator bezeichnet.

Immer 1: Das Modell gibt immer die mittlere Klasse zurück. Dieses Modell illustriert Probleme von Fehlermaßen (MSE und MAE) bei Mehrklassenproblemen.

Bei *Zufällig*, *Immer 0* und *Immer 1* handelt es sich um Modelle, die keinen praktischen Nutzen haben. Die diesen Modellen zugeordneten Performance-Werte sollten als untere Schranke für die Bewertung der Nützlichkeit von Modellen verwendet werden[1]. Bei starker Ungleichverteilung der Klassen ist dies ist jedoch bei den Meisten der betrachteten Maße problematisch, da nützliche Modelle zum Teil schlechter bewertet werden als *Immer 0*.

Tabelle A.1 vergleicht die Performance-Kennzahlen. In Teil (a) und (b) werden binäre Probleme und in (c) und (d) Dreiklassenprobleme betrachtet. Teil (a) und (c) zeigen jeweils die Performance-Kennzahlen bei gleichverteilten Klassenhäufigkeiten. Hier erfüllen alle Maße die Anforderung, dass die nützlichen Modelle besser bewertet werden als *Zufällig*, *Immer 0* und *Immer 1*. In Teil (b) und (d), mit ungleichen Klassenhäufigkeiten, ist das anders. Hier wird das Modell *Immer 0* von den Maßen Acc, MSE und MAE besser bewertet als Modell A und B (siehe fett hervorgehobene Zahlen). Von diesem Problem sind ICC und MF1 *nicht* betroffen. Ein großer Vorteil von ICC gegenüber allen anderen Maßen ist, dass allen nicht nützlichen Modellen unabhängig von der Klassenanzahl und -häufigkeitsverteilung die gleiche Performance (0) zugewiesen wird. Für ICC ergibt sich ein klarer Wertebereich von 0 (das Modell kann keine nützlichen Prädiktionen treffen) bis 1 (das Modell liegt mit seinen Prädiktionen immer richtig).

Aufgrund der günstigen Eigenschaften wird in der Arbeit für alle Evaluationen, bei denen eine Intensität prädiziert wird, das ICC-Maß angewendet. Hierzu zählt der Autor dieser Dissertation auch binäre Klassifikationsprobleme zwischen kein Schmerz und Schmerz, da beide Zustände auf einer gemeinsamen Schmerzskala verortet sind und es einen fließenden Übergang zwischen diesen Zuständen gibt.

Weitere Diskussionen zu Performance-Maßen sind in Werner et al. [WSAH15] zu finden.

[1] In Praxis werden diese unteren Schranken jedoch zum Teil unterschritten. Dies liegt an der Streuung der empirisch ermittelten Performance um den tatsächlichen Erwartungswert, das heißt an der Varianz aufgrund der begrenzten Teststichprobe.

	Prädiktion 0	Prädiktion 1	Prädiktion 0	Prädiktion 1	Prädiktion 0	Prädiktion 1	Prädiktion 0	Prädiktion 1	Prädiktion 0	Prädiktion 1
0	10000	0	0	8392	1608	0	7079	2921	0	5000	5000	0	10000	0
1	0	10000	1	1633	8367	1	2911	7089	1	5000	5000	1	10000	0

Maß	*Optimum*	*Modell A*	*Modell B*	*Zufällig*	*Immer 0*
Acc	1,00	0,84	0,71	0,50	0,50
MF1	1,00	0,84	0,71	0,50	0,33
ICC	1,00	0,68	0,42	0,00	0,00
MSE	0,00	0,16	0,29	0,50	0,50
MAE	0,00	0,16	0,29	0,50	0,50

(a) Binäre Klassifikation mit gleichen Klassenhäufigkeiten.

	Prädiktion 0	Prädiktion 1	Prädiktion 0	Prädiktion 1	Prädiktion 0	Prädiktion 1	Prädiktion 0	Prädiktion 1	Prädiktion 0	Prädiktion 1
0	10000	0	0	8361	1639	0	7102	2898	0	5000	5000	0	10000	0
1	0	1000	1	156	844	1	286	714	1	500	500	1	1000	0

Maß	*Optimum*	*Modell A*	*Modell B*	*Zufällig*	*Immer 0*
Acc	1,00	0,84	0,71	0,50	**0,91**
MF1	1,00	0,69	0,56	0,40	0,48
ICC	1,00	0,44	0,23	0,00	0,00
MSE	0,00	0,16	0,29	0,50	**0,09**
MAE	0,00	0,16	0,29	0,50	**0,09**

(b) Binäre Klassifikation mit ungleichen Klassenhäufigkeiten.

	Prädiktion 0	Prädiktion 1	Prädiktion 2	Prädiktion 0	Prädiktion 1	Prädiktion 2	Prädiktion 0	Prädiktion 1	Prädiktion 2	Prädiktion 0	Prädiktion 1	Prädiktion 2	Prädiktion 0	Prädiktion 1	Prädiktion 2	Prädiktion 0	Prädiktion 1	Prädiktion 2
0	10000	0	0	8433	1551	16	7109	2432	459	3334	3333	3333	10000	0	0	0	10000	0
1	0	10000	0	1574	6883	1543	2969	4188	2843	3334	3333	3333	10000	0	0	1	10000	0
2	0	0	10000	17	1570	8413	482	2436	7082	3334	3333	3333	10000	0	0	2	10000	0

Maß	*Optimum*	*Modell A*	*Modell B*	*Zufällig*	*Immer 0*	*Immer 1*
Acc	1,00	0,79	0,61	0,33	0,33	0,33
MF1	1,00	0,79	0,61	0,33	0,17	0,17
ICC	1,00	0,84	0,65	0,00	0,00	0,00
MSE	0,00	0,21	0,48	1,33	1,67	0,67
MAE	0,00	0,21	0,42	0,89	1,00	0,67

(c) Klassifikation von drei Klassen mit gleichen Häufigkeiten.

	Prädiktion 0	Prädiktion 1	Prädiktion 2	Prädiktion 0	Prädiktion 1	Prädiktion 2	Prädiktion 0	Prädiktion 1	Prädiktion 2	Prädiktion 0	Prädiktion 1	Prädiktion 2	Prädiktion 0	Prädiktion 1	Prädiktion 2	Prädiktion 0	Prädiktion 1	Prädiktion 2
0	10000	0	0	8388	1603	9	7098	2419	483	3334	3333	3333	10000	0	0	0	10000	0
1	0	1000	0	170	668	162	278	424	298	334	333	333	1000	0	0	1	1000	0
2	0	0	100	0	13	87	5	26	69	34	33	33	100	0	0	2	100	0

Maß	*Optimum*	*Modell A*	*Modell B*	*Zufällig*	*Immer 0*	*Immer 1*
Acc	1,00	0,82	0,68	0,33	**0,90**	0,09
MF1	1,00	0,60	0,39	0,22	0,32	0,06
ICC	1,00	0,55	0,30	0,00	0,00	0,00
MSE	0,00	0,18	0,45	1,58	**0,13**	0,91
MAE	0,00	0,18	0,36	0,97	**0,11**	0,91

(d) Klassifikation von drei Klassen mit ungleichen Häufigkeiten.

Tabelle A.1.: Vergleich von Performance-Werten verschiedener Maße zur Bewertung von Klassifikationsergebnissen auf Intervallskalen (hier insbesondere Schmerzintensitäten). Acc: Accuracy/Korrektklassifikationsrate, MF1: Macro-Averaged F1, ICC: Intra-Klassen-Korrelation, MSE: mittlerer quadratischer Fehler, MAE: mittlerer absoluter Fehler.

Anhang B. Weitere Ergebnisse der einzelbildbasierten Mimikerkennung

B.1. Ergebnisse je AU

Als Ergänzung zu den Ergebnissen in Abschnitt 3.3.3, wo für die Datensätze FERA 2017 und UNBC zumeist nur Mittelwerte aufgeführt sind, werden einige Ergebnisse nach den Facial Action Units (AU) aufgeschlüsselt. Für FERA 2017 in den Varianten mit nur der frontalen Ansicht 6 (Tabelle B.1) und mit allen Ansichten (Tabelle B.2) sowie für UNBC (Tabelle B.3) wurden ausgewählt: (1) das beste Ergebnis mit menschengemachten Merkmalen und klassischem Lernverfahren aus Tabelle 3.2, (2) das beste Ergebnis mit CNN-Merkmalen und klassischem Lernverfahren aus Tabelle 3.2, (3) das beste Ergebnis mit CNN als Lernverfahren aus Tabelle 3.2, (4) zu allen zuvor genannten bei Regression jeweils die zugehörige Klassifikationsvariante und bei Klassifikation die zugehörige Regressionsvariante, (5) das beste Fusionsergebnis mit SVR-E aus Tabelle 3.3a, sowie (6) das beste Fusionsergebnis mit CNN aus Tabelle 3.3b. Zusätzlich listet Tabelle B.4 AU-weise Ergebnisse der Evaluierung mit Bosphorus3D auf, unter anderem einige zu Abb. 3.10 (alle Ergebnisse mit $\alpha = 0{,}5$). Das CNN in der letzten Zeile der Tabelle war der Ausgangspunkt für das Transferlernen ausgehend von Bosphorus3D.

In den Tabellen B.1-B.4 repräsentieren die hinterlegten Graustufen (ähnlich wie in Tabelle 3.2) für jede AU bzw. den Mittelwert (Spalten), welche Modelle (Zeilen) die besten Ergebnisse liefern. Die Helligkeit ist skaliert vom Maximum der Performance („dunkelstes“ grau) bis zum Minimum (weiß).

Merkmale	Modell	MW	AU 1	AU 4	AU 6	AU 10	AU 12	AU 14	AU 17
2D-Landmarken	SVM-E	0,419	0,194	0,158	0,592	0,646	0,738	0,343	0,261
	SVR-E	0,463	0,181	0,136	0,645	0,690	0,761	0,539	0,293
MN-R Bosphorus3D	SVM-E	0,437	0,203	0,178	0,663	0,622	0,753	0,338	0,302
	SVR-E	0,472	0,202	0,220	0,677	0,696	0,763	0,424	0,325
	MN-K^{MT}	0,661	0,624	0,619	0,784	0,792	0,855	0,472	0,479
	MN-R^{MT}	0,653	0,475	0,466	0,769	0,814	0,864	0,612	0,569
MN-R + 3D-Koord.	SVR-E	0,485	0,246	0,218	0,690	0,704	0,766	0,436	0,338
MN-R + Kopfpose	MN-R	0,670	0,631	0,487	0,784	0,791	0,855	0,602	0,539

^{MT} Multi-Task-Lernen mit Bosphorus3D MW: Mittelwert aller AU
SVM-E: Support Vector Machine Ensemble (Klassifikation) SVR-E: Support Vector Regression Ensemble
MN-K: MobileNetV3-large (CNN) Klassifikation (vortrainiert mit Bosphorus3D)
MN-R: MobileNetV3-large (CNN) Regression (vortrainiert mit Bosphorus3D)

Tabelle B.1.: Ergebnisse auf FERA 2017 (frontale Ansicht) je AU

Merkmale	Modell	MW	AU 1	AU 4	AU 6	AU 10	AU 12	AU 14	AU 17
LBP 10×10	SVM-E	0,366	0,163	0,093	0,575	0,596	0,668	0,332	0,137
	SVR-E	0,408	0,170	0,120	0,617	0,643	0,696	0,444	0,169
MN-R Bosphorus3D	SVM-E	0,424	0,178	0,164	0,623	0,643	0,718	0,374	0,272
	SVR-E	0,464	0,171	0,167	0,669	0,701	0,759	0,477	0,306
	MN-K	0,623	0,599	0,531	0,752	0,765	0,840	0,408	0,467
	MN-R	0,638	0,562	0,487	0,764	0,779	0,847	0,554	0,472
MN-R + 3D-K. + Kopfp.	SVR-E	0,471	0,187	0,179	0,666	0,707	0,760	0,496	0,304
MN-R + 3D-K. + Kopfp.	MN-R	0,652	0,577	0,498	0,784	0,788	0,845	0,572	0,499

MW: Mittelwert aller AU
SVM-E: Support Vector Machine Ensemble (Klassifikation) SVR-E: Support Vector Regression Ensemble
MN-K: MobileNetV3-large (CNN) Klassifikation (vortrainiert mit Bosphorus3D)
MN-R: MobileNetV3-large (CNN) Regression (vortrainiert mit Bosphorus3D)

Tabelle B.2.: Ergebnisse auf FERA 2017 (alle Ansichten) je AU

Merkmale	Modell	MW	AU 4	AU 6	AU 7	AU 9	AU 10	AU 12	AU 20	AU 25	AU 26	AU 43
3D-Abstände	SVM-E	0,176	0,109	0,401	0,174	0,111	0,090	0,371	0,034	0,132	0,006	0,328
	SVR-E	0,184	0,064	0,427	0,190	0,100	0,085	0,424	0,072	0,208	0,016	0,260
MN-R Bosphorus3D	RF-K	0,065	0,000	0,240	0,024	0,000	0,000	0,205	0,000	0,044	0,000	0,137
	RF-R	0,179	0,123	0,419	0,216	0,103	0,041	0,372	0,060	0,193	0,092	0,173
	MN-K^{MT}	0,329	0,251	0,512	0,444	0,165	0,376	0,506	0,046	0,296	0,226	0,465
	MN-R^{MT}	0,385	0,286	0,604	0,533	0,224	0,346	0,554	0,169	0,411	0,238	0,482
2D-L. + LBP + Kopfp.	SVR-E	0,186	0,070	0,416	0,240	0,106	0,065	0,431	0,050	0,154	0,060	0,271
MN-R + 2D-Landm.	MN-R	0,367	0,295	0,567	0,435	0,280	0,335	0,538	0,186	0,378	0,235	0,421

^{MT} Multi-Task-Lernen mit Bosphorus3D MW: Mittelwert aller AU
SVM-E: Support Vector Machine Ensemble (Klassifikation) SVR-E: Support Vector Regression Ensemble
RF-K: Random Forest Klassifikation RF-R: Random Forest Regression
MN-K: MobileNetV3-large (CNN) Klassifikation (vortrainiert mit Bosphorus3D)
MN-R: MobileNetV3-large (CNN) Regression (vortrainiert mit Bosphorus3D)

Tabelle B.3.: Ergebnisse auf UNBC je AU

Merkmale	Modell	MW	AU01	AU02	AU04	AU05	AU06	AU07	AU09	AU10	AU11
2D-Landmarken	SVM-E	0,206	0,206	0,266	0,230	0,194	0,167	0,336	0,259	0,180	0,003
	SVR-E	0,232	0,251	0,303	0,281	0,225	0,196	0,356	0,297	0,209	0,006
LBP 5×5	SVM-E	0,284	0,480	0,532	0,314	0,327	0,162	0,382	0,442	0,210	0,008
	SVR-E	0,309	0,522	0,539	0,363	0,353	0,195	0,419	0,442	0,249	0,013
LBP 10×10	SVM-E	0,321	0,525	0,557	0,330	0,356	0,192	0,427	0,495	0,228	0,006
	SVR-E	0,347	0,555	0,574	0,400	0,395	0,237	0,484	0,485	0,287	0,014
MN-K ImageNet	MN-R	0,678	0,788	0,755	0,805	0,702	0,635	0,706	0,818	0,585	0,141

Merkmale	Modell	...	AU12	AU14	AU15	AU16	AU17	AU18	AU20	AU22	AU23
2D-Landmarken	SVM-E		0,513	0,122	0,043	0,110	0,125	0,270	0,068	0,125	0,082
	SVR-E		0,546	0,146	0,076	0,162	0,183	0,301	0,100	0,152	0,116
LBP 5×5	SVM-E		0,531	0,127	0,086	0,104	0,202	0,374	0,160	0,193	0,072
	SVR-E		0,586	0,148	0,117	0,149	0,234	0,401	0,166	0,215	0,102
LBP 10×10	SVM-E		0,583	0,140	0,102	0,115	0,240	0,428	0,187	0,222	0,070
	SVR-E		0,626	0,168	0,143	0,167	0,289	0,452	0,199	0,238	0,110
MN-K ImageNet	MN-R		0,765	0,528	0,589	0,584	0,659	0,792	0,647	0,789	0,446

Merkmale	Modell	...	AU24	AU25	AU26	AU27	AU28	AU34	AU38	AU43
2D-Landmarken	SVM-E		0,195	0,553	0,163	0,455	0,217	0,181	0,088	0,217
	SVR-E		0,234	0,540	0,222	0,450	0,213	0,215	0,111	0,144
LBP 5×5	SVM-E		0,232	0,505	0,225	0,567	0,365	0,355	0,119	0,309
	SVR-E		0,278	0,535	0,259	0,587	0,323	0,343	0,159	0,340
LBP 10×10	SVM-E		0,281	0,555	0,263	0,609	0,444	0,504	0,145	0,339
	SVR-E		0,339	0,598	0,301	0,621	0,357	0,414	0,195	0,373
MN-K ImageNet	MN-R		0,736	0,899	0,630	0,782	0,798	0,743	0,464	0,833

MW: Mittelwert aller AU
SVM-E: Support Vector Machine Ensemble (Klassifikation) SVR-E: Support Vector Regression Ensemble
MN-K: MobileNetV3-large (CNN) Klassifikation MN-R: MobileNetV3-large (CNN) Regression

Tabelle B.4.: Ergebnisse auf Boshporus3D je AU

B.2. Merkmalsfusion

In den Tabellen B.5-B.8 werden ICC-Performances aufgelistet, die bei der Verwendung von punktbasierten Merkmalen, Texturmerkmalen, Kopfposemerkmalen und deren Merkmalsfusion mit den Lernverfahren Support Vector Machine Ensemble (SVM-E), Support Vector Regression Ensemble (SVR-E), Random Forest Regression (RF-R) und Random Forest Klassifikation (RF-K) ermittelt wurden. Es handelt sich hierbei um eine Ergänzung zu den Ergebnissen in Abschnitt 3.3.3. Tabelle B.5 taucht zahlenmäßig bereits in Tabelle 3.3a auf, wo jedoch die Einfärbung so gestaltet war, dass sie den Performance-Gewinn bzw.-Verlust durch die Fusion verglichen mit den besten bei der Fusion beteiligten Merkmalen hervorhebt. In den Tabellen B.5-B.8 repräsentieren die hinterlegten Graustufen (wie in Tabelle 3.2) für jeden Datensatz (Spalte), welche Modelle (Zeilen) die besten Ergebnisse liefern. Die Helligkeit ist skaliert vom Maximum („dunkelstes“ grau) bis zum 30%-Quantil (weiß).

	FERA17		**UNBC**		**BioVid-S**		**BioVid-S7**		
Merkmale	**Frontal 7 AUs**	**9 Ans. 7 AUs**	**10 AUs**	**PSPI**	**Frontal Stim.**	**3 Ans. Stim.**	**Frontal Stim.**	**3 Ans. Stim.**	*M*
2D-Landmarken	0,463	0,370	0,173	0,387	0,140	0,152	0,440	0,503	0,329
3D-Koordinaten	0,439	0,341	0,181	0,414	0,197	0,112	0,669	0,299	0,332
LBP	0,456	0,408	0,178	0,473	0,137	0,123	0,667	0,520	0,370
Kopfpose	0,015	0,002	0,012	0,022	0,061	0,010	0,000	0,003	0,016
MobileNetV3	0,472	0,464	0,161	0,369	0,067	0,127	0,332	0,237	0,279
2D-L. + LBP	0,473	0,417	0,185	0,479	0,135	0,131	0,656	0,554	0,379
2D-L. + Kopfp.	0,470	0,370	0,169	0,388	0,158	0,146	0,444	0,512	0,332
2D-L. + MobileN.	0,483	0,470	0,171	0,386	0,083	0,121	0,323	0,258	0,287
3D-K. + LBP	0,466	0,415	0,183	0,480	0,134	0,129	0,659	0,543	0,376
3D-K. + Kopfp.	0,445	0,351	0,171	0,406	0,180	0,115	0,578	0,352	0,325
3D-K. + MobileN.	0,485	0,468	0,169	0,388	0,056	0,130	0,308	0,241	0,281
LBP + Kopfp.	0,455	0,409	0,178	0,474	0,136	0,129	0,664	0,520	0,371
MobileN. + Kopfp.	0,476	0,470	0,161	0,361	0,079	0,131	0,325	0,232	0,279
2D-L. + LBP + Kopfp.	0,474	0,419	0,186	0,481	0,135	0,130	0,650	0,553	0,378
2D-L. + Mob. + Kopfp.	0,481	0,466	0,168	0,388	0,085	0,126	0,331	0,291	0,292
3D-K. + LBP + Kopfp.	0,464	0,415	0,182	0,477	0,132	0,131	0,656	0,543	0,375
3D-K. + Mob. + Kopfp.	0,485	0,471	0,168	0,389	0,076	0,128	0,329	0,240	0,286

Tabelle B.5.: Ergebnisse der Merkmalsfusion mit SVR-E

	FERA17		UNBC		BioVid-S		BioVid-S7		
Merkmale	**Frontal 7 AUs**	**9 Ans. 7 AUs**	**10 AUs**	**PSPI**	**Frontal Stim.**	**3 Ans. Stim.**	**Frontal Stim.**	**3 Ans. Stim.**	*M*
2D-Landmarken	0,419	0,340	0,135	0,345	0,170	0,140	0,384	0,535	0,308
3D-Koordinaten	0,412	0,338	0,159	0,392	0,191	0,120	0,578	0,344	0,317
LBP	0,372	0,366	0,109	0,382	0,161	0,135	0,664	0,530	0,340
Kopfpose	0,041	0,004	0,007	-0,008	0,072	0,018	0,054	-0,010	0,022
MobileNetV3	0,437	0,424	0,138	0,385	0,162	0,143	0,289	0,393	0,297
2D-L. + LBP	0,368	0,377	0,114	0,398	0,161	0,142	0,673	0,556	0,348
2D-L. + Kopfp.	0,422	0,340	0,133	0,330	0,171	0,141	0,493	0,518	0,318
2D-L. + MobileN.	0,458	0,426	0,153	0,397	0,149	0,134	0,354	0,409	0,310
3D-K. + LBP	0,369	0,369	0,109	0,388	0,162	0,130	0,664	0,551	0,343
3D-K. + Kopfp.	0,402	0,344	0,142	0,367	0,201	0,124	0,561	0,420	0,320
3D-K. + MobileN.	0,451	0,425	0,144	0,375	0,133	0,140	0,349	0,391	0,301
LBP + Kopfp.	0,367	0,366	0,112	0,385	0,162	0,130	0,664	0,528	0,339
MobileN. + Kopfp.	0,438	0,420	0,140	0,377	0,138	0,141	0,296	0,389	0,292
2D-L. + LBP + Kopfp.	0,386	0,377	0,114	0,382	0,162	0,140	0,661	0,559	0,348
2D-L. + Mob. + Kopfp.	0,456	0,425	0,153	0,399	0,147	0,138	0,361	0,407	0,311
3D-K. + LBP + Kopfp.	0,369	0,367	0,112	0,389	0,163	0,136	0,664	0,555	0,344
3D-K. + Mob. + Kopfp.	0,452	0,428	0,144	0,387	0,135	0,137	0,341	0,385	0,301

Tabelle B.6.: Ergebnisse der Merkmalsfusion mit SVM-E

	FERA17		UNBC		BioVid-S		BioVid-S7		
Merkmale	**Frontal 7 AUs**	**9 Ans. 7 AUs**	**10 AUs**	**PSPI**	**Frontal Stim.**	**3 Ans. Stim.**	**Frontal Stim.**	**3 Ans. Stim.**	*M*
2D-Landmarken	0,417	0,344	0,154	0,396	0,144	0,146	0,385	0,358	0,293
3D-Koordinaten	0,405	0,332	0,167	0,408	0,133	0,107	0,554	0,385	0,311
LBP	0,334	0,279	0,125	0,330	0,166	0,134	0,502	0,474	0,293
Kopfpose	0,068	0,072	0,016	0,053	0,005	0,022	0,112	0,033	0,047
MobileNetV3	0,470	0,441	0,179	0,423	0,189	0,202	0,508	0,491	0,363
2D-L. + LBP	0,388	0,329	0,153	0,399	0,168	0,159	0,493	0,456	0,318
2D-L. + Kopfp.	0,421	0,347	0,151	0,396	0,145	0,146	0,459	0,361	0,303
2D-L. + MobileN.	0,489	0,443	0,190	0,439	0,205	0,197	0,587	0,480	0,379
3D-K. + LBP	0,379	0,331	0,138	0,361	0,172	0,147	0,492	0,512	0,316
3D-K. + Kopfp.	0,414	0,340	0,148	0,403	0,138	0,114	0,594	0,401	0,319
3D-K. + MobileN.	0,484	0,445	0,184	0,427	0,202	0,199	0,524	0,506	0,371
LBP + Kopfp.	0,338	0,280	0,125	0,332	0,178	0,136	0,493	0,483	0,296
MobileN. + Kopfp.	0,471	0,440	0,177	0,422	0,197	0,204	0,487	0,497	0,362
2D-L. + LBP + Kopfp.	0,392	0,330	0,153	0,399	0,175	0,156	0,480	0,449	0,317
2D-L. + Mob. + Kopfp.	0,491	0,443	0,188	0,447	0,205	0,200	0,557	0,515	0,381
3D-K. + LBP + Kopfp.	0,381	0,331	0,138	0,360	0,173	0,144	0,503	0,509	0,317
3D-K. + Mob. + Kopfp.	0,485	0,445	0,184	0,425	0,204	0,201	0,495	0,508	0,368

Tabelle B.7.: Ergebnisse der Merkmalsfusion mit RF-R

	FERA17		UNBC		BioVid-S		BioVid-S7		
Merkmale	**Frontal 7 AUs**	**9 Ans. 7 AUs**	**10 AUs**	**PSPI**	**Frontal Stim.**	**3 Ans. Stim.**	**Frontal Stim.**	**3 Ans. Stim.**	*M*
2D-Landmarken	0,398	0,329	0,087	0,258	0,144	0,145	0,434	0,380	0,272
3D-Koordinaten	0,410	0,335	0,099	0,255	0,138	0,103	0,569	0,412	0,290
LBP	0,322	0,237	0,012	0,011	0,183	0,111	0,542	0,498	0,239
Kopfpose	0,068	0,068	0,004	0,014	0,016	0,023	0,117	0,023	0,042
MobileNetV3	0,435	0,416	0,065	0,201	0,205	0,206	0,517	0,513	0,320
2D-L. + LBP	0,359	0,285	0,027	0,058	0,174	0,134	0,566	0,528	0,267
2D-L. + Kopfp.	0,399	0,334	0,082	0,261	0,152	0,141	0,455	0,378	0,275
2D-L. + MobileN.	0,445	0,418	0,071	0,214	0,201	0,204	0,550	0,503	0,326
3D-K. + LBP	0,331	0,254	0,015	0,024	0,188	0,119	0,595	0,516	0,255
3D-K. + Kopfp.	0,404	0,330	0,072	0,220	0,155	0,115	0,585	0,417	0,287
3D-K. + MobileN.	0,438	0,418	0,066	0,194	0,205	0,204	0,503	0,501	0,316
LBP + Kopfp.	0,319	0,237	0,011	0,008	0,171	0,120	0,557	0,481	0,238
MobileN. + Kopfp.	0,435	0,417	0,065	0,206	0,204	0,202	0,529	0,518	0,322
2D-L. + LBP + Kopfp.	0,358	0,286	0,027	0,068	0,179	0,145	0,543	0,509	0,264
2D-L. + Mob. + Kopfp.	0,444	0,417	0,070	0,212	0,206	0,199	0,570	0,504	0,328
3D-K. + LBP + Kopfp.	0,332	0,253	0,014	0,015	0,179	0,120	0,562	0,512	0,249
3D-K. + Mob. + Kopfp.	0,439	0,419	0,066	0,199	0,205	0,204	0,520	0,512	0,320

Tabelle B.8.: Ergebnisse der Merkmalsfusion mit RF-K

Anhang C. Weitere Ergebnisse der videobasierten Schmerzerkennung

C.1. Erkennung starker Schmerzen

Als Ergänzung zu den zusammengefassten Ergebnissen in Abschnitt 4.3.1 sind in Tabelle C.1 die vollständigen Ergebnisse der Erkennung starker Schmerzen mit den Klassifikatoren SVM, SVM Ensemble (SVM-E) und Random Forest (RF-K) aufgeführt. Dabei repräsentieren die hinterlegten Graustufen für jeden Datensatz bzw. den Mittelwert (Spalten), welche Kombinationen von Einzelbildmerkmal, zeitlicher Integration und Klassifikation (Zeilen) die besten Ergebnisse liefern. Die Helligkeit ist skaliert vom Maximum der Performance („dunkelstes“ grau) bis zum Minimum (weiß).

C.2. Messung der Schmerzintensität

Als Ergänzung zu den zusammengefassten Ergebnissen in Abschnitt 4.3.2 sind in Tabelle C.2 die vollständigen Ergebnisse der deskriptorbasierten Messung der Schmerzintensität mit Support-Vector-Ensembles und Random Forests, jeweils Klassifikation und Regression, aufgeführt. Dabei repräsentieren die hinterlegten Graustufen für jeden Datensatz bzw. den Mittelwert (Spalten), welche Kombinationen von Einzelbildmerkmal, zeitlicher Integration und Klassifikation (Zeilen) die besten Ergebnisse liefern. Die Helligkeit ist skaliert vom Maximum der Performance („dunkelstes“ grau) bis zum Minimum (weiß).

Tabelle C.3 und C.4 zeigen Konfusionsmatrizen der Messung der Schmerzintensität für die Datensätze BioVid und UNBC.

Einzelbildmerkmale	Zeitliche Integration	Klassif.	BioVid-A Frontal	BioVid-A 3 Ans.	BioVid-A7 Frontal	BioVid-A7 3 Ans.	MW
3D-Koordinaten	Entscheidungsfusion (MW)	SVM	0,190	0,083	0,381	0,464	0,280
		SVM-E	0,187	0,080	0,390	0,463	0,280
		RF-K	0,152	0,134	0,411	0,428	0,281
	Entscheidungsfusion (Max.)	SVM	0,251	0,120	0,545	0,391	0,327
		SVM-E	0,260	0,105	0,539	0,363	0,316
		RF-K	0,203	0,146	0,554	0,415	0,330
	BoTF Deskriptor [Bar+14]	SVM-E	0,107	0,114	0,733	0,472	0,357
		RF-K	0,320	0,250	0,768	0,605	0,486
	Statistikdeskriptor	SVM-E	0,255	0,237	0,835	0,511	0,460
		RF-K	0,347	0,278	0,742	0,613	0,495
Kopfpose	Entscheidungsfusion (MW)	SVM	0,047	0,013	-0,032	-0,105	-0,019
		SVM-E	0,037	0,008	0,059	-0,064	0,010
		RF-K	0,039	0,032	-0,095	0,004	-0,005
	Entscheidungsfusion (Max.)	SVM	0,086	0,040	0,055	0,025	0,052
		SVM-E	0,103	0,040	0,159	0,070	0,093
		RF-K	0,099	0,067	0,019	0,134	0,080
	BoTF Deskriptor [Bar+14]	SVM-E	0,192	0,149	0,446	0,243	0,257
		RF-K	0,248	0,207	0,495	0,459	0,352
	Statistikdeskriptor	SVM-E	0,254	0,182	0,311	0,243	0,247
		RF-K	0,289	0,235	0,497	0,462	0,371
3K-Koord. + Kopfpose	Entscheidungsfusion (MW)	SVM	0,183	0,073	0,393	0,464	0,278
		SVM-E	0,190	0,085	0,389	0,449	0,278
		RF-K	0,156	0,142	0,297	0,374	0,242
	Entscheidungsfusion (Max.)	SVM	0,247	0,112	0,564	0,332	0,314
		SVM-E	0,265	0,108	0,530	0,348	0,313
		RF-K	0,207	0,156	0,690	0,486	0,385
	BoTF Deskriptor [Bar+14]	SVM-E	0,144	0,127	0,798	0,566	0,409
		RF-K	0,320	0,258	0,765	0,623	0,491
	Statistikdeskriptor	SVM-E	0,241	0,232	0,859	0,548	0,470
		RF-K	0,359	0,282	0,751	0,609	0,500
OpenFace AU [BRM16]	Entscheidungsfusion (MW)	SVM	0,182	0,096	0,325	0,157	0,190
		SVM-E	0,187	0,095	0,319	0,168	0,192
		RF-K	0,255	0,165	0,552	0,331	0,326
	Entscheidungsfusion (Max.)	SVM	0,271	0,146	0,511	0,348	0,319
		SVM-E	0,278	0,136	0,515	0,357	0,321
		RF-K	0,250	0,153	0,664	0,425	0,373
	BoTF Deskriptor [Bar+14]	SVM-E	0,213	0,102	0,752	0,313	0,345
		RF-K	0,351	0,181	0,804	0,340	0,419
	Statistikdeskriptor	SVM-E	0,289	0,174	0,834	0,424	0,430
		RF-K	0,363	0,215	0,828	0,498	0,476
OpenFace AU [BRM16] + Kopfpose	BoTF Deskriptor [Bar+14]	SVM-E	0,194	0,120	0,796	0,327	0,359
		RF-K	0,352	0,183	0,809	0,343	0,422
	Statistikdeskriptor	SVM-E	0,271	0,191	0,833	0,394	0,422
		RF-K	0,371	0,229	0,823	0,498	0,480
MN-R Bos3D CNN AU	BoTF Deskriptor [Bar+14]	SVM-E	0,233	0,227	0,784	0,803	0,512
		RF-K	0,367	0,359	0,848	0,831	0,601
	Statistikdeskriptor	SVM-E	0,217	0,277	0,854	0,794	0,536
		RF-K	0,373	0,372	0,852	0,870	0,617
MN-R Bos3D CNN AU + Kopfpose	BoTF Deskriptor [Bar+14]	SVM-E	0,197	0,240	0,794	0,815	0,512
		RF-K	0,376	0,357	0,840	0,828	0,600
	Statistikdeskriptor	SVM-E	0,217	0,268	0,855	0,804	0,536
		RF-K	0,384	0,373	0,864	0,859	0,620
MN-K Bos3D CNN AU	BoTF Deskriptor [Bar+14]	SVM-E	0,208	0,222	0,824	0,785	0,510
		RF-K	0,341	0,339	0,821	0,828	0,582
	Statistikdeskriptor	SVM-E	0,238	0,285	0,744	0,754	0,505
		RF-K	0,345	0,347	0,867	0,859	0,605
MN-K Bos3D CNN AU + Kopfpose	BoTF Deskriptor [Bar+14]	SVM-E	0,191	0,231	0,786	0,800	0,502
		RF-K	0,362	0,350	0,820	0,827	0,590
	Statistikdeskriptor	SVM-E	0,237	0,280	0,770	0,774	0,515
		RF-K	0,364	0,363	0,874	0,862	0,616

AU: Action Unit BoTF: Bag of Temporal Features MW: Mittelwert
SVM-E: Support Vector Machine Ensemble (Klassifikation) RF-K: Random Forest Klassifikation

Tabelle C.1.: Erkennung starker Schmerzen mit SVM, SVM-E und RF-K.

Merkmale	Deskriptor	Modell	BioVid-A Frontal	BioVid-A 3 Ans.	BioVid-A7 Frontal	BioVid-A7 3 Ans.	X-ITE (2 Ans.) Elektr.	X-ITE (2 Ans.) Hitze	UNBC OPR	UNBC VAS	MW
3D-Koordinaten	BoTF	SVM-E	0,117	0,082	0,435	0,307	0,198	0,206	0,628	0,453	0,303
		SVR-E	0,127	0,070	0,420	0,113	0,166	0,157	0,690	0,541	0,285
		RF-K	0,213	0,154	0,565	0,462	0,251	0,272	0,672	0,491	0,385
		RF-R	0,117	0,050	0,464	0,377	0,201	0,217	0,652	0,487	0,321
	Statistikd.	SVM-E	0,162	0,124	0,490	0,330	0,243	0,235	0,587	0,471	0,330
		SVR-E	0,199	0,127	0,131	0,364	0,255	0,252	0,545	0,397	0,284
		RF-K	0,237	0,171	0,576	0,480	0,276	0,284	0,674	0,535	0,404
		RF-R	0,149	0,074	0,511	0,436	0,240	0,241	0,628	0,474	0,344
Kopfpose	BoTF	SVM-E	0,084	0,060	0,203	0,154	0,166	0,187	0,454	0,302	0,201
		SVR-E	0,130	0,071	0,124	0,154	0,182	0,206	0,368	0,314	0,194
		RF-K	0,169	0,126	0,363	0,296	0,218	0,248	0,565	0,401	0,298
		RF-R	0,070	0,034	0,240	0,208	0,166	0,199	0,474	0,434	0,228
	Statistikd.	SVM-E	0,147	0,101	0,166	0,210	0,223	0,237	0,434	0,332	0,231
		SVR-E	0,153	0,088	0,159	0,244	0,215	0,232	0,209	0,000	0,162
		RF-K	0,192	0,148	0,366	0,313	0,251	0,263	0,559	0,345	0,305
		RF-R	0,104	0,058	0,270	0,207	0,205	0,224	0,451	0,371	0,236
3D-Koordinaten + Kopfpose	BoTF	SVM-E	0,144	0,104	0,502	0,282	0,210	0,212	0,639	0,464	0,320
		SVR-E	0,122	0,058	0,405	0,129	0,139	0,131	0,712	0,559	0,282
		RF-K	0,217	0,160	0,565	0,440	0,246	0,277	0,697	0,525	0,391
		RF-R	0,116	0,053	0,463	0,371	0,200	0,222	0,649	0,499	0,322
	Statistikd.	SVM-E	0,171	0,126	0,471	0,321	0,240	0,233	0,669	0,508	0,342
		SVR-E	0,195	0,132	0,161	0,270	0,251	0,242	0,638	0,423	0,289
		RF-K	0,240	0,170	0,573	0,494	0,279	0,286	0,678	0,535	0,407
		RF-R	0,151	0,077	0,489	0,415	0,238	0,242	0,604	0,474	0,336
OpenFace AU	BoTF	SVM-E	0,124	0,058	0,434	0,186	0,174	0,167	0,530	0,302	0,247
		SVR-E	0,130	0,029	0,235	0,054	0,140	0,144	0,577	0,359	0,208
		RF-K	0,228	0,118	0,564	0,236	0,217	0,233	0,579	0,323	0,312
		RF-R	0,122	0,048	0,484	0,189	0,175	0,195	0,522	0,366	0,263
	Statistikd.	SVM-E	0,175	0,086	0,472	0,224	0,226	0,229	0,601	0,417	0,304
		SVR-E	0,219	0,106	0,241	0,239	0,233	0,240	0,512	0,339	0,266
		RF-K	0,248	0,137	0,620	0,362	0,257	0,265	0,602	0,389	0,360
		RF-R	0,162	0,067	0,566	0,308	0,232	0,236	0,556	0,381	0,313
OpenFace AU + Kopfpose	BoTF	SVM-E	0,136	0,070	0,441	0,172	0,172	0,166	0,590	0,379	0,266
		SVR-E	0,118	0,025	0,232	0,069	0,118	0,121	0,599	0,421	0,213
		RF-K	0,228	0,120	0,571	0,236	0,215	0,234	0,589	0,328	0,315
		RF-R	0,120	0,050	0,486	0,189	0,177	0,195	0,521	0,372	0,264
	Statistikd.	SVM-E	0,179	0,107	0,466	0,228	0,229	0,227	0,589	0,420	0,306
		SVR-E	0,212	0,109	0,243	0,249	0,239	0,235	0,518	0,380	0,273
		RF-K	0,246	0,143	0,610	0,373	0,258	0,273	0,583	0,372	0,357
		RF-R	0,164	0,068	0,551	0,306	0,236	0,235	0,546	0,363	0,309
MN-R Bos3D CNN AU	BoTF	SVM-E	0,163	0,160	0,462	0,462	0,225	0,226	0,638	0,463	0,350
		SVR-E	0,138	0,099	0,356	0,108	0,157	0,160	0,655	0,539	0,276
		RF-K	0,253	0,242	0,611	0,618	0,280	0,306	0,689	0,468	0,433
		RF-R	0,158	0,142	0,561	0,543	0,258	0,279	0,661	0,496	0,387
	Statistikd.	SVM-E	0,162	0,173	0,548	0,476	0,256	0,247	0,645	0,418	0,366
		SVR-E	0,200	0,189	0,242	0,322	0,262	0,255	0,611	0,406	0,311
		RF-K	0,260	0,249	0,649	0,647	0,294	0,306	0,632	0,532	0,446
		RF-R	0,168	0,156	0,585	0,568	0,285	0,290	0,651	0,470	0,397
MN-R Bos3D CNN AU + Kopfpose	BoTF	SVM-E	0,173	0,170	0,489	0,480	0,236	0,234	0,626	0,442	0,356
		SVR-E	0,121	0,084	0,379	0,159	0,121	0,125	0,664	0,535	0,274
		RF-K	0,253	0,244	0,613	0,613	0,275	0,308	0,698	0,479	0,435
		RF-R	0,161	0,142	0,545	0,534	0,256	0,280	0,644	0,506	0,383
	Statistikd.	SVM-E	0,188	0,190	0,546	0,491	0,263	0,251	0,648	0,425	0,375
		SVR-E	0,192	0,184	0,206	0,261	0,256	0,245	0,670	0,411	0,303
		RF-K	0,264	0,249	0,627	0,646	0,298	0,305	0,630	0,543	0,445
		RF-R	0,171	0,157	0,582	0,555	0,285	0,291	0,642	0,486	0,396
MN-K Bos3D CNN AU	BoTF	SVM-E	0,137	0,148	0,445	0,463	0,221	0,227	0,637	0,461	0,342
		SVR-E	0,120	0,101	0,393	0,114	0,157	0,164	0,656	0,534	0,280
		RF-K	0,238	0,236	0,593	0,618	0,279	0,300	0,689	0,498	0,431
		RF-R	0,163	0,142	0,510	0,535	0,259	0,279	0,661	0,498	0,381
	Statistikd.	SVM-E	0,160	0,167	0,501	0,485	0,256	0,244	0,640	0,415	0,358
		SVR-E	0,182	0,179	0,244	0,309	0,263	0,256	0,611	0,402	0,306
		RF-K	0,240	0,238	0,625	0,636	0,294	0,308	0,628	0,521	0,436
		RF-R	0,154	0,145	0,472	0,568	0,285	0,290	0,655	0,473	0,380
MN-K Bos3D CNN AU + Kopfpose	BoTF	SVM-E	0,159	0,164	0,534	0,444	0,237	0,232	0,627	0,451	0,356
		SVR-E	0,114	0,084	0,440	0,209	0,122	0,125	0,660	0,533	0,286
		RF-K	0,242	0,234	0,608	0,622	0,279	0,304	0,698	0,486	0,434
		RF-R	0,163	0,142	0,491	0,528	0,260	0,279	0,652	0,496	0,376
	Statistikd.	SVM-E	0,179	0,182	0,470	0,463	0,264	0,248	0,647	0,436	0,361
		SVR-E	0,184	0,177	0,264	0,223	0,252	0,240	0,659	0,408	0,301
		RF-K	0,262	0,247	0,625	0,644	0,292	0,305	0,638	0,567	0,447
		RF-R	0,174	0,151	0,550	0,555	0,285	0,290	0,643	0,480	0,391

AU: Action Unit BoTF: Bag of Temporal Features Deskriptor [Bar+14] MW: Mittelwert
SVM-E: Support Vector Machine Ensemble (Klassifikation) SVR-E: Support Vector Regression Ensemble RF-K: Random Forest Klassifikation
RF-R: Random Forest Regression

Tabelle C.2.: Erkennung von Schmerzintensitäten mit Deskriptoren.

	Prädizierte Klasse					
	BLN	PA1	PA2	PA3	PA4	Recall
BLN	1157	393	149	32	9	0,665
PA1	1116	368	175	59	22	0,211
PA2	1073	360	191	81	35	0,110
PA3	818	400	246	152	124	0,087
PA4	598	344	259	234	305	0,175
Precision	0,243	0,197	0,187	0,272	0,616	

(a) BioVid-A (frontal): ICC = 0,322 / Accuracy = 0,250 / MAE = 1,466.

	Prädizierte Klasse					
	BLN	PA1	PA2	PA3	PA4	Recall
BLN	2941	1495	579	169	36	0,563
PA1	2818	1464	622	205	111	0,280
PA2	2660	1501	617	270	172	0,118
PA3	2079	1379	829	532	401	0,102
PA4	1434	1191	836	726	1033	0,198
Precision	0,246	0,208	0,177	0,280	0,589	

(b) BioVid-A (3 Ansichten): ICC = 0,312 / Accuracy = 0,252 / MAE = 1,412.

	Prädizierte Klasse					
	BLN	PA1	PA2	PA3	PA4	Recall
BLN	273	93	49	5	0	0,650
PA1	223	43	74	69	11	0,102
PA2	182	69	76	86	7	0,181
PA3	89	74	60	159	38	0,379
PA4	13	29	105	199	74	0,176
Precision	0,350	0,140	0,209	0,307	0,569	

(c) BioVid-A7 (3 Ansichten): ICC = 0,551 / Accuracy = 0,298 / MAE = 1,080.

Tabelle C.3.: Konfusionsmatrizen für Schmerzintensität bei BioVid mit CNN-Regression, zeitlicher Convolution und Gewichtung $\lambda = 0{,}9$.

	Prädiktion											
	0	1	2	3	4	5	6	7	8	9	10	Recall
0	99	61	15	0	0	0	0	0	0	0	0	0,57
1	92	68	23	2	13	0	1	0	11	0	0	0,32
2	25	50	15	8	14	1	3	0	4	0	0	0,13
3	8	28	13	5	28	0	6	0	10	2	0	0,05
4	8	39	24	13	11	0	7	0	3	0	0	0,11
5	5	22	5	4	2	1	2	0	14	0	0	0,02
6	5	7	12	12	11	1	1	0	3	3	0	0,02
7	0	4	5	1	8	1	0	0	11	0	0	0,00
8	14	7	5	2	10	1	2	0	45	4	0	0,50
9	1	3	12	4	8	0	0	0	22	0	0	0,00
10	0	0	0	0	4	0	0	0	2	4	0	0,00
Precision	0,39	0,24	0,12	0,10	0,10	0,20	0,05	-	0,36	0,00	-	

Tabelle C.4.: Konfusionsmatrizen für Schmerzintensität (VAS) bei UNBC mit MN-K Bos3D AU + Kopfpose / Statistikdeskriptor / RF-K. Die Kreuzvalidierung mit den 200 Samples von UNBC wurde 5 Mal wiederholt. ICC = 0,567 / Accuracy = 0,245 / MAE = 1,889.

Literatur

[ACR09] Zara Ambadar, Jeffrey F Cohn und Lawrence Ian Reed. „All Smiles are Not Created Equal: Morphology and Timing of Smiles Perceived as Amused, Polite, and Embarrassed/Nervous“. In: *Journal of Nonverbal Behavior* 33.1 (2009), S. 17–34. DOI: 10.1007\s10919-008-0059-5.

[Adi+15] Mohammad Adibuzzaman u. a. „Assessment of Pain Using Facial Pictures Taken with a Smartphone“. In: *Annual Computer Software and Applications Conference*. IEEE, 2015, S. 726–731. DOI: 10.1109/COMPSAC.2015.150.

[AHP06] Timo Ahonen, Abdenour Hadid und Matti Pietikainen. „Face description with local binary patterns: Application to face recognition“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.12 (2006), S. 2037–2041.

[AKS16] Mohammadreza Amirian, Markus Kächele und Friedhelm Schwenker. „Using Radial Basis Function Neural Networks for Continuous and Discrete Pain Estimation from Bio-physiological Signals“. In: *Artificial Neural Networks in Pattern Recognition*. 2016, S. 269–284. DOI: 10.1007/978-3-319-46182-3.

[Amb+92] Bruce Ambuel u. a. „Assessing distress in pediatric intensive care environments: the COMFORT scale“. In: *Journal of Pediatric Psychology* 17.1 (1992), S. 95–109.

[Ami+17] M Amirian u. a. „Support Vector Regression of Sparse Dictionary-Based Features for View-Independent Action Unit Intensity Estimation“. In: *IEEE International Conference on Automatic Face Gesture Recognition (FG)*. 2017, S. 854–859. DOI: 10.1109/FG.2017.109.

[AMV15] Timur Almaev, Brais Martinez und Michel Valstar. „Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection“. In: *IEEE International Conference on Computer Vision (ICCV)*. 2015, S. 3774–3782.

[Ash+07] A.B. Ashraf u. a. „The painful face - Pain expression recognition using active appearance models“. In: *International Conference on Multimodal Interfaces (ICMI)*. 2007, S. 9–14. DOI: 10.1145/1322192.1322197.

[Ash+09] A B Ashraf u. a. „The Painful Face - Pain Expression Recognition Using Active Appearance Models“. In: *Image & Vision Computing* 27.12 (2009), S. 1788–1796. DOI: 10.1016/j.imavis.2009.05.007.

[ATT12] Hillel Aviezer, Yaacov Trope und Alexander Todorov. „Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions“. In: *Science* 338.6111 (2012), S. 1225–1229. DOI: 10.1126/science.1224313.

[AU20] Görkem Algan und İlkay Ulusoy. *Label Noise Types and Their Effects on Deep Learning*. 2020. arXiv: 2003.10471. (Besucht am 18. 06. 2021).

[Aun+16] Min S.H. Aung u. a. „The Automatic Detection of Chronic Pain-Related Expression: Requirements, Challenges and the Multimodal EmoPain Dataset“. In: *IEEE Transactions on Affective Computing* 7.4 (2016), S. 435–451. DOI: 10.1109/TAFFC.2015.2462830.

[Bar+14] Marian Stewart Bartlett u. a. „Automatic decoding of facial movements reveals deceptive pain expressions“. In: *Current Biology* 24.7 (2014), S. 738–743. DOI: 10.1016/j.cub.2014.02.009.

[Bat+17] J Batista u. a. „AUMPNet: simultaneous Action Units detection and intensity estimation on multipose facial images using a single convolutional neural network“. In: *International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2017, S. 866–871.

[Bel+11] P N Belhumeur u. a. „Localizing parts of faces using a consensus of exemplars“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011, S. 545–552. DOI: 10.1109/CVPR.2011.5995602.

[Bin+14] Deniz Bingöl u. a. „Facial action unit intensity estimation using rotation invariant features and regression analysis“. In: *International Conference on Image Processing (ICIP)*. IEEE, 2014, S. 1381–1385.

[BL07] Léon Bottou und Chih-jen Lin. „Support Vector Machine Solvers“. In: *Large-Scale Kernel Machines*. Hrsg. von Léon Bottou u. a. MIT Press, 2007. DOI: 10.7551/mitpress/7496.003.0003.

[BMR15] Tadas Baltrusaitis, Marwa Mahmoud und Peter Robinson. „Cross-dataset learning and person-specific normalisation for automatic action unit detection“. In: *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 2015.

[BQSM16] C F Benitez-Quiroz, R Srinivasan und A M Martinez. „EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, S. 5562–5570. DOI: 10.1109/CVPR.2016.600.

[Bra+07] Sheryl Brahnam u. a. „Machine assessment of neonatal facial expressions of acute pain“. In: *Decision Support Systems* 43.4 (2007), S. 1242–1254. DOI: 16\j.dss.2006.02.004.

[Bre01] Leo Breiman. „Random Forests“. In: *Machine Learning* 45.1 (2001), S. 5–32. DOI: 10.1023/A:1010933404324.

[Bre96] Leo Breiman. „Bagging predictors“. In: *Machine Learning* 24.2 (1996), S. 123–140. DOI: 10.1007/BF00058655.

[BRM16] Tadas Baltrusaitis, Peter Robinson und Louis Philippe Morency. „OpenFace: An open source facial behavior analysis toolkit“. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2016. DOI: 10.1109/WACV.2016.7477553.

[BV99] Volker Blanz und Thomas Vetter. „A Morphable Model for the Synthesis of 3D Faces“. In: *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 1999, S. 187–194. DOI: 10.1145/311535.311556.

[BWL20] Alexey Bochkovskiy, Chien-Yao Wang und Hong-Yuan Mark Liao. *YOLOv4: Optimal Speed and Accuracy of Object Detection*. 2020. arXiv: 2004.10934 [cs.CV].

[Cao+18] Qiong Cao u. a. „VGGFace2: A dataset for recognising faces across pose and age“. In: *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. 2018, S. 67–74. DOI: 10.1109/FG.2018.00020. arXiv: 1710.08092.

[Cao+20] Jie Cao u. a. „Towards High Fidelity Face Frontalization in the Wild“. In: *International Journal of Computer Vision* 128.5 (2020), S. 1485–1504. DOI: 10.1007/s11263-019-01229-6.

[CCF17] Junkai Chen, Zheru Chi und Hong Fu. „A new framework with multiple tasks for detecting and locating pain events in video“. In: *Computer Vision and Image Understanding* 155 (2017), S. 113–123. DOI: 10.1016/j.cviu.2016.11.003.

[Cha+02] Nitesh V. Chawla u. a. „SMOTE: synthetic minority over-sampling technique“. In: *Journal of Artificial Intelligent Research* 16 (2002), S. 321–357.

[Che+12a] Jixu Chen u. a. „Person-specific expression recognition with transfer learning“. In: *IEEE International Conference on Image Processing (ICIP)*. 2012, S. 2621–2624.

[Che+12b] Sien W Chew u. a. „In the pursuit of effective affective computing: The relationship between features and registration“. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 42.4 (2012), S. 1006–1016.

[Che+13] Dong Chen u. a. „Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification“. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2013, S. 3025–3032.

[Chr+15] G. G. Chrysos u. a. „Offline Deformable Face Tracking in Arbitrary Videos“. In: *International Conference on Computer Vision Workshops (ICCVW)*. 2015.

[Chu+14] Junyoung Chung u. a. „Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling“. In: *CoRR* 1412.3555 (2014). arXiv: 1412.3555.

[CL11] Chih-Chung Chang und Chih-Jen Lin. „LIBSVM – A Library for Support Vector Machines“. In: *ACM Transactions on Intelligent Systems and Technology* 2.3 (2011), 27:1–27:27.

[Cor+02] William H. Cordell u. a. „The high prevalence of pain in emergency medical care“. In: *The American Journal of Emergency Medicine* 20.3 (2002), S. 165–169. DOI: 10.1053/AJEM.2002.32643.

[Cor+16] C A Corneanu u. a. „Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.8 (2016), S. 1548–1568. DOI: 10.1109/TPAMI.2016.2515606.

[CPG11] Kenneth D Craig, Kenneth M Prkachin und Ruth Eckstein Grunau. „The facial expression of pain“. In: *Handbook of Pain Assessment*. Hrsg. von Dennis C Turk und Ronald Melzack. Guilford Press, 2011.

[Cra09] Kenneth D Craig. „The social communication model of pain.“ In: *Canadian Psychology* 50.1 (2009), S. 22–32.

[Cra92] Kenneth D Craig. „The facial expression of pain - Better than a thousand words?“ In: *APS Journal* 1.3 (1992), S. 153–162. DOI: 10.1016/1058-9139(92)90001-S.

[CTC13] Wen-Sheng Chu, Fernando De La Torre und Jeffery F Cohn. „Selective transfer machine for personalized facial action unit detection“. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2013, S. 3515–3522.

[DBD15] Arnaud Dapogny, Kevin Bailly und S Dubuisson. „Pairwise conditional random forests for facial expression recognition“. In: *IEEE International Conference on Computer Vision (ICCV)*. 2015, S. 3783–3791.

[Den+09] Jia Deng u. a. „ImageNet: A Large-Scale Hierarchical Image Database“. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2009.

[Deu09] Deutsche Interdisziplinäre Vereinigung für Schmerztherapie (DIVS). *Behandlung akuter perioperativer und posttraumatischer Schmerzen (S3-Leitlinie)*. 2009.

[Din+13] Xiaoyu Ding u. a. „Facial action unit event detection by cascade of tasks“. In: *IEEE International Conference on Computer Vision (ICCV)*. 2013, S. 2400–2407. DOI: 10.1109/ICCV.2013.298.

[Dod04] Neil A. Dodgson. „Variation and extrema of human interpupillary distance“. In: *Proc. SPIE Vol. 5291, Stereoscopic Displays and Virtual Reality Systems XI* 5291.January (2004), S. 36–46. DOI: 10.1117/12.529999.

[DT17] Terrance DeVries und Graham W. Taylor. *Improved Regularization of Convolutional Neural Networks with Cutout*. 2017. arXiv: 1708.04552. (Besucht am 17. 09. 2020).

[EFH02] P Ekman, W Friesen und J Hager. *Facial Action Coding System: The Manual on CD ROM*. A Human Face, 2002.

[Ele+16] Stefanos Eleftheriadis u. a. „Gaussian Process Domain Experts for Model Adaptation in Facial Behavior Analysis“. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2016.

[Ere+20] Diyala Erekat u. a. „Enforcing multilabel consistency for automatic spatio-temporal assessment of shoulder pain intensity“. In: *ICMI 2020 Companion - Companion Publication of the 2020 International Conference on Multimodal Interaction* (2020), S. 156–164. DOI: 10.1145/3395035.3425190.

[EVM17] Joy Egede, Michel Valstar und Brais Martinez. „Fusing Deep Learned and Hand-Crafted Features of Appearance, Shape, and Dynamics for Automatic Pain Estimation“. In: *IEEE Conference on Face and Gesture Recognition (FG)*. 2017. arXiv: 1701.04540.

[Fan+08] Rong-En Fan u. a. „LIBLINEAR: A library for large linear classification“. In: *Journal of Machine Learning Research* 9.Aug (2008), S. 1871–1874.

[Fan+13] Gabriele Fanelli u. a. „Random Forests for Real Time 3D Face Analysis“. In: *International Journal of Computer Vision* 101.3 (2013), S. 437–458. DOI: 10.1007/s11263-012-0549-0.

[FFV14] Corneliu Florea, Laura Florea und Constantin Vertan. „Learning Pain from Emotion: Transferred HoT Data Representation for Pain Intensity Estimation“. In: *European Conference on Computer Vision (ECCV) Workshops*. 2014.

[FL20] Yingruo Fan und Zhaojiang Lin. „G2RL: Geometry-Guided Representation Learning for Facial Action Unit Intensity Estimation“. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. 2020, S. 731–737. DOI: 10.24963/ijcai.2020/102.

[Flo+16] Corneliu Florea u. a. „Pain intensity estimation by a self-taught selection of histograms of topographical features“. In: *Image and Vision Computing* 56 (2016), S. 13–27. DOI: 10.1016/J.IMAVIS.2016.08.014.

[FV14] Benoît Frénay und Michel Verleysen. „Classification in the presence of label noise: A survey“. In: *IEEE Transactions on Neural Networks and Learning Systems* 25.5 (2014), S. 845–869. DOI: 10.1109/TNNLS.2013.2292894.

[GB10] Xavier Glorot und Yoshua Bengio. „Understanding the difficulty of training deep feedforward neural networks“. In: *International Conference on Artificial Intelligence and Statistics*. Hrsg. von Yee Whye Teh und Mike Titterington. Bd. 9. Proceedings of Machine Learning Research. 2010, S. 249–256.

[GBC16a] Ian Goodfellow, Yoshua Bengio und Aaron Courville. „Convolutional Networks“. In: *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[GBC16b] Ian Goodfellow, Yoshua Bengio und Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[GCD15] Jeffrey M Girard, Jeffrey F Cohn und Fernando De la Torre. „Estimating smile intensity: A better way“. In: *Pattern Recognition Letters* 66 (2015), S. 13–21. DOI: 10.1016/j.patrec.2014.10.004.

[GEB16] Leon A Gatys, Alexander S Ecker und Matthias Bethge. „Image Style Transfer Using Convolutional Neural Networks“. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016, S. 2414–2423.

[Gha+14] Afsane Ghasemi u. a. „Social signal processing for pain monitoring using a hidden conditional random field“. In: *IEEE Workshop on Statistical Signal Processing (SSP)*. IEEE, 2014, S. 61–64. DOI: 10.1109/SSP.2014.6884575.

[GM16] Julie Gregory und Linda Mcgowan. „An examination of the prevalence of acute pain for hospitalised adult patients: A systematic review“. In: *Journal of Clinical Nursing* 25.5-6 (2016). DOI: 10.1111/jocn.13094.

[Gro+10] Ralph Gross u. a. „Multi-PIE“. In: *Image Vision Comput.* 28.5 (Mai 2010), S. 807–813. DOI: 10.1016/j.imavis.2009.08.002.

[Gru+19] Sascha Gruss u. a. „Multi-Modal Signals for Analyzing Pain Responses to Thermal and Electrical Stimuli“. In: *Journal of Visualized Experiments* 146 (2019). DOI: 10.3791/59057.

[Had+02] Thomas Hadjistavropoulos u. a. „Using facial expressions to assess musculoskeletal pain in older persons“. In: *European Journal of Pain* 6.3 (2002), S. 179–187. DOI: 10.1053/eujp.2001.0327.

[Haq+18] Mohammad A. Haque u. a. „Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities“. In: *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. 2018, S. 250–257. DOI: 10.1109/FG.2018.00044.

[Has+15] Tal Hassner u. a. „Effective Face Frontalization in Unconstrained Images“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, S. 4295–4304.

[HC04] Thomas Hadjistavropoulos und Kenneth D Craig. *Pain: Psychological Perspectives*. Bd. 60. Lawrence Erlbaum Associates, Incorporated, 2004, S. 400. ISBN: 978-0-8058-4299-9.

[HC12] Zakia Hammal und Jeffrey F. Cohn. „Automatic detection of pain intensity“. In: *International Conference on Multimodal Interaction (ICMI)*. ACM, 2012, S. 47. DOI: 10.1145/2388676.2388688.

[HCL03] Chih-Wei Hsu, Chih-Chung Chang und Chih-Jen Lin. *A Practical Guide to Support Vector Classification*. Techn. Ber. 2003.

[He+17] Kaiming He u. a. „Mask R-CNN“. In: *IEEE International Conference on Computer Vision (ICCV)*. 2017, S. 2961–2969.

[Her+11] Keela Herr u. a. „Pain assessment in the patient unable to self-report: position statement with clinical practice recommendations.“ In: *Pain management nursing: official journal of the American Society of Pain Management Nurses* 12.4 (2011), S. 230–50. DOI: 10.1016/j.pmn.2011.10.002.

[HG09] Haibo He und Edwardo A. Garcia. „Learning from imbalanced data“. In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), S. 1263–1284. DOI: 10.1109/TKDE.2008.239.

[Him07] Alexander Himme. „Gütekriterien der Messung: Reliabilität, Validität und Generalisierbarkeit“. In: *Methodik der empirischen Forschung*. Hrsg. von Sönke Albers u. a. Wiesbaden: Gabler, 2007, S. 375–390. ISBN: 978-3-8349-9121-8. DOI: 10.1007/978-3-8349-9121-8_25.

[HK12] Zakia Hammal und Miriam Kunz. „Pain monitoring: A dynamic and context-sensitive system“. In: *Pattern Recognition* 45.4 (2012), S. 1265–1280. DOI: 10.1016/j.patcog.2011.09.014.

[HLC14] Kuang Jui Hsu, Yen Yu Lin und Yung Yu Chuang. „Augmented multiple instance regression for inferring object contours in bounding boxes“. In: *IEEE Transactions on Image Processing* 23.4 (2014), S. 1722–1736. DOI: 10.1109/TIP.2014.2307436.

[HLM14] Gary B Huang und Erik Learned-Miller. *Labeled faces in the wild: Updates and new reporting procedures*. Techn. Ber. 2014, S. 1–5.

[How+19] Andrew Howard u. a. „Searching for MobileNetV3“. In: *IEEE International Conference on Computer Vision (ICCV)*. 2019, S. 1314–1324. arXiv: 1905.02244.

[HS97] Sepp Hochreiter und Jürgen Schmidhuber. „Long Short-Term Memory“. In: *Neural Computation* 9.8 (1997), S. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[Hua+20] Dong Huang u. a. „Pain-attentive network: a deep spatio-temporal attention model for pain estimation“. In: *Multimedia Tools and Applications* (2020), S. 1–26. DOI: 10.1007/s11042-020-09397-1.

[Hub64] Peter J. Huber. „Robust Estimation of a Location Parameter“. In: *The Annals of Mathematical Statistics* 35.1 (1964), S. 73–101. DOI: 10.1214/aoms/1177703732.

[Hug68] Gordon F. Hughes. „On the Mean Accuracy of Statistical Pattern Recognizers“. In: *IEEE Transactions on Information Theory* 14.1 (1968), S. 55–63.

[HZP16] Xiaopeng Hong, Stefanos Zafeiriou und Maja Pantic. „Capturing correlations of local features for image representation“. In: *Neurocomputing* 184 (2016), S. 99–106. DOI: 10.1016/J.NEUCOM.2015.07.134.

[INM15] Ramin Irani, Kamal Nasrollahi und Thomas B Moeslund. „Pain Recognition using Spatiotemporal Oriented Energy of Facial Muscles“. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2015, S. 679–692.

[Int21] International Association for the Study of Pain. *IASP Terminology*. https://www.iasp-pain.org/Education/Content.aspx?ItemNumber=1698#Pain, aufgerufen am 01.06.2021. 2021.

[Ira+15] Ramin Irani u. a. „Spatiotemporal Analysis of RGB-D-T Facial Images for Multi-Modal Pain Level Recognition“. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2015.

[JCD13] Laszlo A Jeni, Jeffrey F Cohn und Fernando De La Torre. „Facing Imbalanced Data: Recommendations for the Use of Performance Metrics“. In: *Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2013, S. 245–251.

[Jen+13] L A Jeni u. a. „Continuous AU intensity estimation using localized, sparse facial feature space“. In: *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 2013, S. 1–7. DOI: 10.1109/FG.2013.6553808.

[Ji+19] Yuwang Ji u. a. „A Survey on Tensor Techniques and Applications in Machine Learning“. In: *IEEE Access* 7 (2019), S. 162950–162990. DOI: 10.1109/ACCESS.2019.2949814.

[JO05] Girish P. Joshi und Babatunde O. Ogunnaike. *Consequences of inadequate postoperative pain relief and chronic persistent postoperative pain*. 2005. DOI: 10.1016/j.atc.2004.11.013.

[Jun+15] Heechul Jung u. a. „Joint fine-tuning in deep neural networks for facial expression recognition“. In: *IEEE International Conference on Computer Vision (ICCV)*. 2015, S. 2983–2991.

[Käc+15a] Markus Kächele u. a. „Bio-Visual Fusion for Person-Independent Recognition of Pain Intensity“. In: *Multiple Classifier Systems*. Hrsg. von Friedhelm Schwenker, Fabio Roli und Josef Kittler. Lecture Notes in Computer Science. Springer International Publishing, 2015, S. 220–230.

[Käc+15b] Markus Kächele u. a. „Multimodal Data Fusion for Person-Independent, Continuous Estimation of Pain Intensity“. In: *Engineering Applications of Neural Networks*. Hrsg. von Lazaros Iliadis und Chrisina Jayne. Communications in Computer and Information Science. Springer International Publishing, 2015, S. 275–285.

[Käc+16] Markus Kächele u. a. „Methods for Person-Centered Continuous Pain Intensity Assessment From Bio-Physiological Channels“. In: *IEEE Journal of Selected Topics in Signal Processing* 10.5 (2016), S. 854–864. DOI: 10.1109/JSTSP.2016.2535962.

[Käc+17] Markus Kächele u. a. „Adaptive confidence learning for the personalization of pain intensity estimation systems“. In: *Evolving Systems* 8.1 (2017), S. 71–83. DOI: 10.1007/s12530-016-9158-4.

[Kar+20] Tero Karras u. a. „Analyzing and Improving the Image Quality of StyleGAN“. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[KE00] Dacher Keltner und Paul Ekman. „Facial expression of emotion“. In: *Handbook of Emotions*. Hrsg. von M. Lewis und J. M. Haviland-Jones. Second Edi. New York: Guilford, 2000, S. 236–249.

[Kes+17] Viktor Kessler u. a. „Multimodal fusion including camera photoplethysmography for pain recognition“. In: *International Conference on Companion Technology (ICCT)*. 2017. ISBN: 978-1-5386-1160-9. DOI: 10.1109/COMPANION.2017.8287083.

[KGC17] Jan Kukačka, Vladimir Golkov und Daniel Cremers. *Regularization for Deep Learning: A Taxonomy*. 2017. arXiv: 1710.10686. (Besucht am 23. 01. 2020).

[KGN09] Anne Marie Kabes, Janet K. Graves und Joan Norris. „Further validation of the nonverbal pain scale in intensive care patients“. In: *Critical Care Nurse* 29.1 (2009), S. 59–66. DOI: 10.4037/ccn2009992.

[Kha+13] Rizwan Ahmed Khan u. a. „Pain detection through shape and appearance features“. In: *IEEE International Conference on Multimedia and Expo*. 2013, S. 1–6. DOI: 10.1109/ICME.2013.6607608.

[Kin09] Davis E. King. „Dlib-ml: A Machine Learning Toolkit“. In: *Journal of Machine Learning Research* 10 (2009), S. 1755–1758.

[Kin15] Davis E. King. *Max-Margin Object Detection*. 2015. arXiv: 1502.00046. (Besucht am 05. 06. 2020).

[KL14] M Kunz und S Lautenbacher. „The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain“. In: *European Journal of Pain* 18.6 (2014), S. 813–823. DOI: 10.1002/j.1532-2149.2013.00421.x.

[KMM10] Z Kalal, K Mikolajczyk und J Matas. „Face-TLD: Tracking-Learning-Detection Applied to Faces“. In: *International Conference on Image Processing* (2010).

[Koh95] Ron Kohavi. „A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection“. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. 1995, S. 1137–1143. DOI: 10.5555/1643031.1643047.

[KP10] Minyoung Kim und Vladimir Pavlovic. „Hidden conditional ordinal random fields for sequence classification“. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Bd. 6322 LNAI. PART 2. Springer, Berlin, Heidelberg, 2010, S. 51–65. ISBN: 364215882X. DOI: 10.1007/978-3-642-15883-4_4.

[KPM16] Reza Kharghanian, Ali Peiravi und Farshad Moradi. „Pain detection from facial images using unsupervised feature learning approach“. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2016, S. 419–422. ISBN: 978-1-4577-0220-4. DOI: 10.1109/EMBC.2016.7590729.

[KPP10] Sander Koelstra, Maja Pantic und Ioannis Patras. „A dynamic texture-based approach to recognition of facial actions and their temporal models“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.11 (2010), S. 1940–1954. DOI: 10.1109/TPAMI.2010.50.

[KRP12] Sebastian Kaltwang, Ognjen Rudovic und Maja Pantic. „Continuous pain intensity estimation from facial expressions“. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Bd. 7432 LNCS. Lecture Notes in Computer Science PART 2. Springer Berlin Heidelberg, 2012, S. 368–377. ISBN: 9783642331909. DOI: 10.1007/978-3-642-33191-6_36.

[KS14] Vahdat Kazemi und Josephine Sullivan. „One millisecond face alignment with an ensemble of regression trees“. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2014, S. 1867–1874.

[KSH12] Alex Krizhevsky, Ilya Sutskever und Geoffrey E Hinton. „ImageNet Classification with Deep Convolutional Neural Networks“. In: *Neural Information Processing Systems (NIPS)*. 2012, S. 1097–1105.

[KTP15] Sebastian Kaltwang, Sinisa Todorovic und Maja Pantic. „Latent Trees for Estimating Intensity of Facial Action Units“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[KTP16] Sebastian Kaltwang, Sinisa Todorovic und Maja Pantic. „Doubly Sparse Relevance Vector Machine for Continuous Facial Behavior Estimation“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.9 (2016), S. 1748–1761. DOI: 10.1109/TPAMI.2015.2501824.

[Kun+04] Miriam Kunz u. a. „On the relationship between self-report and facial expression of pain“. In: *Journal of Pain* 5.7 (2004), S. 368–376. DOI: 10.1016/j.jpain.2004.06.002.

[Kun+07] M. Kunz u. a. „The facial expression of pain in patients with dementia“. In: *Pain* 133.1 (2007), S. 221–228. DOI: 10.1016/j.pain.2007.09.007.

[KZ18] Dimitrios Kollias und Stefanos Zafeiriou. *A Multi-Task Learning & Generation Framework: Valence-Arousal, Action Units & Primary Expressions*. 2018. arXiv: 1811.07771.

[KZ19] Dimitrios Kollias und Stefanos Zafeiriou. „Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace". In: *British Machine Vision Conference (BMVC)*. 2019, S. 1–15.

[LBL09] Gwen C Littlewort, Marian Stewart Bartlett und Kang Lee. „Automatic coding of facial expressions displayed during posed and genuine pain". In: *Image and Vision Computing* 27.12 (2009), S. 1797–1803. DOI: 10.1016/j.imavis.2008.12.010.

[LD20] Shan Li und Weihong Deng. „Deep Facial Expression Recognition: A Survey". In: *IEEE Transactions on Affective Computing* 3045.c (2020), S. 1–20. DOI: 10.1109/TAFFC.2020.2981446. arXiv: 1804.08348.

[Le+12] Vuong Le u. a. „Interactive Facial Feature Localization". In: *European Conference on Computer Vision*. 2012.

[LE+16] K Limbrecht-Ecklundt u. a. „Mimische Aktivität differenzierter Schmerzintensitäten". In: *Der Schmerz* 30.3 (2016), S. 248–256. DOI: 10.1007/s00482-016-0105-x.

[Lin+11] Yuanqing Lin u. a. „Large-scale image classification: Fast feature extraction and SVM training". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011, S. 1689–1696. DOI: 10.1109/CVPR.2011.5995477.

[Liu09] Tie-Yan Liu. „Learning to rank for information retrieval". In: *Foundations and Trends in Information Retrieval* 3.3 (2009), S. 225–331.

[Liu+14] Mengyi Liu u. a. „Learning Expressionlets on Spatio-temporal Manifold for Dynamic Facial Expression Recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, S. 1749–1756. DOI: 10.1109/CVPR.2014.226.

[Liu+17] Dianbo Liu u. a. „DeepFaceLIFT: Interpretable Personalized Models for Automatic Estimation of Self-Reported Pain". In: *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing* (2017). arXiv: 1708.04670.

[Liu+18] Peng Liu u. a. „Clinical valid pain database with biomarker and visual information for pain level analysis". In: *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. 2018, S. 525–529. DOI: 10.1109/FG.2018.00084.

[LKP03] Rainer Lienhart, Alexander Kuranov und Vadim Pisarevsky. „Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection". In: *Joint Pattern Recognition Symposium*. 2003, S. 297–304. DOI: 10.1007/978-3-540-45243-0_39.

[LL17] Liliana Lo Presti und Marco La Cascia. „Boosting Hankel matrices for face emotion recognition and pain detection". In: *Computer Vision and Image Understanding* 156 (2017), S. 19–33. DOI: 10.1016/j.cviu.2016.10.007. arXiv: 1506.05001.

[LMP17] Daniel Lopez-Martinez und Rosalind Picard. „Multi-task neural networks for personalized pain recognition from physiological signals". In: *International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2017, S. 181–184. DOI: 10.1109/ACIIW.2017.8272611.

[LMRP17a] Daniel Lopez-Martinez, Ognjen Rudovic und Rosalind Picard. „Personalized Automatic Estimation of Self-Reported Pain Intensity from Facial Expressions“. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, S. 2318–2327. DOI: 10.1109/CVPRW.2017.286. arXiv: 1706.07154.

[LMRP17b] Daniel Lopez-Martinez, Ognjen Rudovic und Rosalind Picard. „Physiological and Behavioral Profiling for Nociceptive Pain Estimation Using Personalized Multitask Learning.“ In: *NIPS Workshop on Machine Learning for Health*. 2017.

[Lop+13] Victoria Lopez u. a. „An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics“. In: *Information Sciences* 250 (2013), S. 113–141.

[LSD15] Jonathan Long, Evan Shelhamer und Trevor Darrell. „Fully Convolutional Networks for Semantic Segmentation“. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[Luc+08] Patrick Lucey u. a. „Improving Pain Recognition Through Better Utilisation of Temporal Information.“ In: *International Conference on Auditory-Visual Speech Processing*. 2008, S. 167–172.

[Luc+11a] Patrick Lucey u. a. „Automatically Detecting Pain in Video Through Facial Action Units“. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 41.3 (2011), S. 664–674. DOI: 10.1109/TSMCB.2010.2082525.

[Luc+11b] Patrick Lucey u. a. „Painful data: The UNBC-McMaster shoulder pain expression archive database“. In: *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG)*. 2011, S. 57–64. DOI: 10.1109/FG.2011.5771462.

[Luc+12] Patrick Lucey u. a. „Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database“. In: *Image and Vision Computing* 30.3 (2012), S. 197–205. DOI: 10.1016/j.imavis.2011.12.003.

[Luh10] Thomas Luhmann. „Erweiterte Verfahren zur geometrischen Kamerakalibrierung in der Nahbereichsphotogrammetrie“. Diss. 2010. ISBN: 9783769650570.

[Lyn11] Mary E Lynch. „The need for a Canadian pain strategy.“ In: *Pain research & management* 16.2 (2011), S. 77–80.

[Män+01] P Mäntyselkä u. a. „Pain as a reason to visit the doctor: a study in Finnish primary health care.“ In: *Pain* 89.2-3 (2001).

[Mas+16] Iacopo Masi u. a. „Pose-aware face recognition in the wild“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, S. 4838–4846.

[Mav+13] S. Mohammad Mavadati u. a. „DISFA: A spontaneous facial action intensity database“. In: *IEEE Transactions on Affective Computing* 4.2 (2013), S. 151–160. DOI: 10.1109/T-AFFC.2013.4.

[MBB14] Hongying Meng und Nadia Bianchi-Berthouze. „Affective state level recognition in naturalistic facial and vocal expressions“. In: *IEEE Transactions on Cybernetics* 44.3 (2014), S. 315–318. DOI: 10.1109/TCYB.2013.2253768.

[MC03] Alain Mignault und Avi Chaudhuri. „The Many Faces of a Neutral Face: Head Tilt and Perception of Dominance and Emotion“. In: *Journal of Nonverbal Behavior* 27.2 (2003), S. 111–132. DOI: 10.1023\A:1023914509763.

[MC68] R Melzack und KL Casey. „Sensory, motivational and central control determinants of pain: a new conceptual model“. In: *The skin senses*. Hrsg. von D Kenshalo. Springfield, IL, 1968, S. 423–443.

[Mer79] H Merskey. „Pain terms: a list with definitions and notes on usage“. In: *PAIN* 6.3 (1979), S. 249–252.

[Min+15] Zuheng Ming u. a. „Facial Action Units Intensity Estimation by the Fusion of Features with Multi-kernel Support Vector Machine“. In: *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 2015.

[MM14] S M Mavadati und M H Mahoor. „Temporal Facial Expression Modeling for Automated Action Unit Intensity Measurement“. In: *International Conference on Pattern Recognition (ICPR)*. 2014, S. 4648–4653. DOI: 10.1109/ICPR.2014.795.

[MMJ97] Henry McQuay, Andrew Moore und Douglas Justins. „Treating acute pain in hospital.“ In: *BMJ: British Medical Journal* 314.7093 (1997), S. 1531.

[Moo+03] David S Moore u. a. „Bootstrap Methods and Permutation Tests“. In: *The practice of business statistics: using data for decisions*. Wh Freeman, 2003.

[MR09] Md Maruf Monwar und Siamak Rezaei. „Support vector machine for automatic pain recognition“. In: *Computational Imaging VII*. 2009, S. 724613–724618. DOI: 10.1117/12.806143.

[Nie+09] Robert Niese u. a. „Towards Pain Recognition in Post-Operative Phases Using 3D-based Features From Video and Support Vector Machines“. In: *International Journal of Digital Content Technology and its Applications* 3.4 (2009), S. 21–33. DOI: 10.4156/jdcta.vol3.issue4.2.

[Nie10] Robert Niese. „Verbesserung der Störsicherheit bei der Mimikanalyse in mono- und binokularen Farbbildsequenzen durch Auswertung geometrischer und dynamischer Merkmale“. Diss. Otto-von-Guericke-Universität Magdeburg, 2010.

[Nil19] Paul Nilges. „Klinische Schmerzmessung“. In: *Praktische Schmerzmedizin: Interdisziplinäre Diagnostik - Multimodale Therapie*. Hrsg. von Ralf Baron u. a. Springer Berlin Heidelberg, 2019, S. 97–104. ISBN: 978-3-662-57487-4. DOI: 10.1007/978-3-662-57487-4_8.

[NM15] Nikolay Neshov und Agata Manolova. „Pain detection from facial characteristics using Supervised Descent Method“. In: *International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*. 2015, S. 213–218. DOI: 10.1109/IDAACS.2015.7340738.

[Nob06] William S. Noble. *What is a support vector machine?* 2006. DOI: 10.1038/nbt1206-1565.

[NWAH13] Robert Niese, Philipp Werner und Ayoub Al-Hamadi. „Accurate, Fast and Robust Realtime Face Pose Estimation Using Kinect Camera“. In: *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2013, S. 487–490. DOI: 10.1109/SMC.2013.89.

[NZ05] W Niesert und M Zenz. „Prophylaxe chronischer Schmerzen“. In: *Deutsches Ärzteblatt* 102.22 (2005).

[Ole+12] Anne Estrup Olesen u. a. „Human Experimental Pain Models for Assessing the Therapeutic Efficacy of Analgesic Drugs“. In: *Pharmacological Reviews* 64.3 (2012), S. 722–779. DOI: 10.1124/pr.111.005447.

[Oth+19a] E. Othman u. a. „Predicting group contribution behaviour in a public goods game from face-to-face communication“. In: *Sensors* 19.12 (2019). DOI: 10.3390/s19122786.

[Oth+19b] Ehsan Othman u. a. „Cross-database evaluation of pain recognition from facial video“. In: *International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 2019, S. 181–186. DOI: 10.1109/ISPA.2019.8868562.

[Oth+21] Ehsan Othman u. a. „Automatic vs. Human Recognition of Pain Intensity from Facial Expression on the X-ITE Pain Database“. In: *Sensors* 21.9 (2021), S. 3273. DOI: 10.3390/s21093273.

[PC95] Kenneth M Prkachin und Kenneth D Craig. „Expressing pain: The communication and interpretation of facial pain signals“. In: *Journal of Nonverbal Behavior* 19.4 (1995), S. 191–205. DOI: 10.1007/BF02173080.

[Ped15] Henrik Pedersen. „Learning Appearance Features for Pain Detection Using the UNBC-McMaster Shoulder Pain Expression Archive Database“. In: *International Conference on Computer Vision Systems*. Springer, Cham, 2015, S. 128–136. DOI: 10.1007/978-3-319-20904-3_12.

[PIY06] P. Pal, A.N. Iyer und R.E. Yantorno. „Emotion Detection From Infant Facial Expressions And Cries“. In: *IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*. Bd. 2. IEEE, 2006, S. II–721–II–724. DOI: 10.1109/ICASSP.2006.1660444.

[Pow11] D M W Powers. „Evaluation: From Precision, Recall and F-Measure to ROC., Informedness, Markedness & Correlation“. In: *Journal of Machine Learning Technologies* 2.1 (2011), S. 37–63.

[Prk92] Kenneth M Prkachin. „The consistency of facial expressions of pain: a comparison across modalities“. In: *Pain* 51.3 (1992), S. 297–306. DOI: 10.1016/0304-3959(92)90213-U.

[PS08] Kenneth M. Prkachin und Patricia E. Solomon. „The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain“. In: *Pain* 139.2 (2008), S. 267–274. DOI: 10.1016/j.pain.2008.04.010.

[Qua+07] Ariadna Quattoni u. a. „Hidden conditional random fields“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.10 (2007), S. 1848–1853. DOI: 10.1109/TPAMI.2007.1124.

[Raj+20] Srinivasa N. Raja u. a. „The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises“. In: *Pain* 161.9 (2020), S. 1976–1982. DOI: 10.1097/j.pain.0000000000001939.

[Reg13] Registered Nurses' Association of Ontario. „Practice Recommendations“. In: *Assessment and Management of Pain*. 3rd ed. Toronto: Registered Nurses' Association of Ontario, 2013, S. 21.

[RG15] Neeru Rathee und Dinesh Ganotra. „A novel approach for pain intensity detection based on facial feature deformations“. In: *Journal of Visual Communication and Image Representation* 33 (2015), S. 247–254. DOI: 10.1016/J.JVCIR.2015.09.007.

[RG16] Neeru Rathee und Dinesh Ganotra. „Multiview Distance Metric Learning on facial feature descriptors for automatic pain intensity detection“. In: *Computer Vision and Image Understanding* 147 (2016), S. 77–86. DOI: 10.1016/J.CVIU.2015.12.004.

[Rin+17] Fabien Ringeval u. a. „AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge“. In: *Proceedings of the Workshop on Audio/Visual Emotion Challenge (AVEC '17)*. New York, New York, USA: ACM Press, 2017, S. 3–9. ISBN: 9781450355025. DOI: 10.1145/3133944.3133953.

[Rod+17] Pau Rodriguez u. a. „Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification“. In: *IEEE Transactions on Cybernetics* (2017). DOI: 10.1109/TCYB.2017.2662199.

[Roy+15] Sourav Dey Roy u. a. „An Approach for Automatic Pain Detection through Facial Expression“. In: *International conference on Intelligent Human Computer Interaction*. Elsevier, 2015. DOI: 10.1016/J.PROCS.2016.04.072.

[RP+13] Bernardino Romera-Paredes u. a. „Transfer learning to account for idiosyncrasy in face and body expressions“. In: *International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, S. 1–6. DOI: 10.1109/FG.2013.6553779.

[RPP13a] Ognjen Rudovic, Maja Pantic und Ioannis Y Patras. „Coupled Gaussian Processes for Pose-Invariant Facial Expression Recognition“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.6 (2013), S. 1357–1369. DOI: 10.1109/TPAMI.2012.233.

[RPP13b] Ognjen Rudovic, Vladimir Pavlovic und Maja Pantic. „Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields“. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Bd. 8034 LNCS. Lecture Notes in Computer Science PART 2. Springer, 2013, S. 234–243. ISBN: 9783642419386. DOI: 10.1007/978-3-642-41939-3_23. arXiv: 1301.5063.

[RPP15] Ognjen Rudovic, Vladimir Pavlovic und Maja Pantic. „Context-sensitive dynamic ordinal regression for intensity estimation of facial action units“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.5 (2015), S. 944–958. DOI: 10.1109/TPAMI.2014.2356192.

[Rui+16] Adria Ruiz u. a. „Multi-Instance Dynamic Ordinal Random Fields for Weakly-Supervised Pain Intensity Estimation“. In: *Asian Conference on Computer Vision*. Lecture Notes in Computer Science. Springer International Publishing, 2016, S. 171–186. DOI: 10.1007/978-3-319-54184-6.

[RV16] Moses Rupenga und Hima B Vadapalli. „Automatic spontaneous pain recognition using supervised classification learning algorithms“. In: *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. 2016. DOI: 10.1109/RoboMech.2016.7813150.

[RVB14] Adria Ruiz, Joost Van de Weijer und Xavier Binefa. „Regularized Multi-Concept MIL for weakly-supervised facial behavior categorization“. In: *British Machine Vision Conference (BMVC)*. 2014. DOI: 10.5244/C.28.13.

[RWAH16] Michal Rapczynski, Philipp Werner und Ayoub Al-Hamadi. „Continuous low latency heart rate estimation from painful faces in real time“. In: *International Conference on Pattern Recognition*. IEEE, 2016, S. 1165–1170. DOI: 10.1109/ICPR.2016.7899794.

[RWAH19] M. Rapczynski, P. Werner und A. Al-Hamadi. „Effects of video encoding on camera-based heart rate estimation“. In: *IEEE Transactions on Biomedical Engineering* 66.12 (2019). DOI: 10.1109/TBME.2019.2904326.

[Sag+13] C. Sagonas u. a. „A semi-automatic methodology for facial landmark annotation“. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2013.

[Sag+15] Christos Sagonas u. a. „Robust statistical face frontalization“. In: *IEEE International Conference on Computer Vision (ICCV)*. 2015, S. 3871–3879.

[Sag+16] Christos Sagonas u. a. „300 Faces In-The-Wild Challenge: database and results“. In: *Image and Vision Computing* 47 (2016), S. 3–18.

[San+14] Enver Sangineto u. a. „We are not All Equal: Personalizing Models for Facial Expression Analysis with Transductive Parameter Transfer“. In: *ACM International Conference on Multimedia*. 2014, S. 357–366. DOI: 10.1145/2647868.2654916.

[Sav+08] Arman Savran u. a. „Bosphorus Database for 3D Face Analysis“. In: *Biometrics and Identity Management*. Hrsg. von Ben Schouten u. a. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008, S. 47–56.

[Sax+19] Frerk Saxen u. a. „Detecting Arbitrarily Rotated Faces for Face Analysis“. In: *International Conference on Image Processing (ICIP)*. 2019, S. 3945–3949. DOI: 10.1109/ICIP.2019.8803631.

[Sch+07] Theodor Heinrich Schiebler u. a. „Kopf und Hals“. In: *Anatomie*. Steinkopff, 2007, S. 581–679. DOI: 10.1007/978-3-7985-1771-4_13.

[SDB13] Karan Sikka, Abhinav Dhall und Marian Bartlett. „Weakly supervised pain localization using multiple instance learning“. In: *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 2013. DOI: 10.1109/FG.2013.6553762.

[SDB14] Karan Sikka, Abhinav Dhall und Marian Stewart Bartlett. „Classification and weakly supervised pain localization using multiple segment representation“. In: *Image and Vision Computing* 32.10 (2014), S. 659–670. DOI: 10.1016/j.imavis.2014.02.008.

[Sen+12] Thibaud Senechal u. a. „Facial action recognition combining heterogeneous features via multikernel learning“. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 42.4 (2012), S. 993–1005. DOI: 10.1109/TSMCB.2012.2193567.

[SF79] Patrick E. Shrout und Joseph L. Fleiss. „Intraclass correlations: Uses in assessing rater reliability“. In: *Psychological Bulletin* 86.2 (1979), S. 420–428. DOI: 10.1037/0033-2909.86.2.420.

[SGC15] E Sariyanidi, H Gunes und A Cavallaro. „Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.6 (2015), S. 1113–1133. DOI: 10.1109/TPAMI.2014.2366127.

[Sho+13] Jamie Shotton u. a. „Real-time human pose recognition in parts from single depth images“. In: *Communications of the ACM* 56.1 (2013), S. 116–124.

[Sik+15] Karan Sikka u. a. „Automated Assessment of Children's Postoperative Pain Using Computer Vision“. In: *Pediatrics* 136.1 (2015), e124–31. DOI: 10.1542/peds.2015-0029.

[Sim+08] Daniela Simon u. a. „Recognition and discrimination of prototypical dynamic expressions of pain and emotions“. In: *Pain* 135.1 (2008), S. 55–64. DOI: 10.1016/j.pain.2007.05.008.

[Son15] Jongwoo Song. „Bias corrections for Random Forest in regression using residual rotation“. In: *Journal of the Korean Statistical Society* 44.2 (2015), S. 321–326. DOI: 10.1016/J.JKSS.2015.01.003.

[SS04] Alex J. Smola und Bernhard Schölkopf. „A tutorial on support vector regression“. In: *Statistics and Computing* 14.3 (2004), S. 199–222. DOI: 10.1023/B:STCO.0000035301.49549.88.

[SST12] Arman Savran, Bulent Sankur und M Taha Bilge. „Regression-based intensity estimation of facial action units“. In: *Image and Vision Computing*. 3D Facial Behaviour Analysis and Understanding 30.10 (2012), S. 774–784. DOI: 10.1016/j.imavis.2011.11.008.

[Ste09] Dan Steinberg. „CART: classification and regression trees“. In: *The top ten algorithms in data mining*. Chapman & Hall / CRC Press, 2009.

[SWAH17] Frerk Saxen, Philipp Werner und Ayoub Al-Hamadi. „Real vs. Fake Emotion Challenge: Learning to Rank Authenticity from Facial Activity Descriptors“. In: *IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2017, S. 3073–3078. DOI: 10.1109/ICCVW.2017.363.

[SZ15] Karen Simonyan und Andrew Zisserman. „Very Deep Convolutional Networks for Large-Scale Image Recognition“. In: *International Conference on Learning Representations*. 2015.

[Szc+21] Benjamin Szczapa u. a. „Automatic Estimation of Self-Reported Pain by Interpretable Representations of Motion Dynamics“. In: *International Conference on Pattern Recognition*. 2021.

[SZP13] Georgia Sandbach, Stefanos Zafeiriou und Maja Pantic. „Markov random field structures for facial action unit intensity estimation“. In: *IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2013, S. 738–745.

[Thi+16] Patrick Thiam u. a. „Audio-Visual Recognition of Pain Intensity“. In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction Workshop*. 2016, S. 110–126. DOI: 10.1007/978-3-319-59259-6_10.

[TKS17] Patrick Thiam, Viktor Kessler und Friedhelm Schwenker. „Hierarchical Combination of Video Features for Personalised Pain Level Recognition“. In: *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 2017, S. 465–470.

[TKS20] Patrick Thiam, Hans A. Kestler und Friedhelm Schwenker. „Two-Stream Attention Network for Pain Recognition from Video Sequences“. In: *Sensors* 20.3 (2020), S. 839. DOI: 10.3390/s20030839.

[TKY12] Muhammad Atif Tahir, Josef Kittler und Fei Yan. „Inverse random undersampling for class imbalance problem and its application to multi-label classification“. In: *Pattern Recognition* 45 (2012), S. 3738–3750.

[TM11a] Dennis C Turk und Ronald Melzack. „Preface“. In: *Handbook of Pain Assessment*. Guilford Press, 2011.

[TM11b] Dennis C Turk und Ronald Melzack. „The Measurement of Pain and the Assessment of People Experiencing Pain“. In: *Handbook of Pain Assessment*. Guilford Press, 2011.

[Tor+15] Fernando De la Torre u. a. „IntraFace“. In: *IEEE International Conference on Automatic Face Gesture Recognition (FG)*. 2015.

[Tru79] G. V. Trunk. „A Problem of Dimensionality: A Simple Example“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.3 (1979), S. 306–307. DOI: 10.1109/TPAMI.1979.4766926.

[TS17] Patrick Thiam und Friedhelm Schwenker. „Multi-modal data fusion for pain intensity assessment and classification“. In: *International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2017, S. 1–6. DOI: 10.1109/IPTA.2017.8310115.

[Tsa+16] Fu-Sheng Tsai u. a. „Toward Development and Evaluation of Pain Level-Rating Scale for Emergency Triage based on Vocal Characteristics and Facial Expressions“. In: *Interspeech*. 2016, S. 92–96. DOI: 10.21437/Interspeech.2016-408.

[Val+15] M Valstar u. a. „Fera 2015-second facial expression recognition and analysis challenge“. In: *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. 2015.

[Val15] Michel Valstar. „Automatic Facial Expression Analysis“. In: *Understanding Facial Expressions in Communication*. Hrsg. von Manas K Mandal und Avinash Awasthi. Springer India, 2015, S. 143–172.

[Val+17] Michel F Valstar u. a. „FERA 2017 - Addressing Head Pose in the Third Facial Expression Recognition and Analysis Challenge“. In: *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. 2017. DOI: 10.1109/FG.2017.107.

[VJ01] Paul Viola und Michael Jones. „Robust real-time face detection“. In: *Proceedings of the IEEE International Conference on Computer Vision* 2 (2001), S. 747. DOI: 10.1109/ICCV.2001.937709.

[VJM11] MF Valstar, Bihan Jiang und Marc Mehu. „The first facial expression recognition and analysis challenge“. In: *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. 2011, S. 921–926. DOI: 10.1109/FG.2011.5771374.

[WAHN12] Philipp Werner, Ayoub Al-Hamadi und Robert Niese. „Pain recognition and intensity rating based on Comparative Learning“. In: *International Conference on Image Processing (ICIP)*. 2012. DOI: 10.1109/ICIP.2012.6467359.

[WAHN14] Philipp Werner, Ayoub Al-Hamadi und Robert Niese. „Comparative learning applied to intensity rating of facial expressions of pain“. In: *International Journal of Pattern Recognition and Artificial Intelligence* 28.05 (2014), S. 1451008. DOI: 10.1142/S0218001414510082.

[WAHW17] Philipp Werner, Ayoub Al-Hamadi und Steffen Walter. „Analysis of facial expressiveness during experimentally induced heat pain“. In: *International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 2017, S. 176–180. DOI: 10.1109/ACIIW.2017.8272610.

[Wal+13] S. Walter u. a. „The BioVid Heat Pain Database: Data for the advancement and systematic validation of an automated pain recognition“. In: *IEEE International Conference on Cybernetics (CYBCONF)*. 2013. DOI: 10.1109/CYBConf.2013.6617456.

[Wal+15] Steffen Walter u. a. „Data fusion for automated pain recognition“. In: *International Conference on Pervasive Computing Technologies for Healthcare*. 2015, S. 261–264. DOI: 10.4108/icst.pervasivehealth.2015.259166.

[Wal+20] S. Walter u. a. „Multimodal recognition of pain intensity and pain modality with machine learning“. In: *Schmerz* 34.5 (2020), S. 400–409. DOI: 10.1007/s00482-020-00468-8.

[Wan+13] Ziheng Wang u. a. „Capturing global semantic relationships for facial action unit recognition“. In: *IEEE International Conference on Computer Vision (ICCV)*. 2013, S. 3304–3311.

[Wan+17] Feng Wang u. a. „Regularizing face verification nets for pain intensity regression“. In: *IEEE International Conference on Image Processing*. 2017. DOI: 10.1109/ICIP.2017.8296449. arXiv: 1702.06925.

[Wer+13] Philipp Werner u. a. „Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges“. In: *British Machine Vision Conference (BMVC)*. 2013. DOI: 10.5244/C.27.119.

[Wer+14a] Philipp Werner u. a. „Automatic heart rate estimation from painful faces“. In: *International Conference on Image Processing (ICIP)*. 2014. DOI: 10.1109/ICIP.2014.7025390.

[Wer+14b] Philipp Werner u. a. „Automatic pain recognition from video and biomedical signals“. In: *International Conference on Pattern Recognition (ICPR)*. IEEE, 2014. DOI: 10.1109/ICPR.2014.784.

[Wer+17] Philipp Werner u. a. „Automatic Pain Assessment with Facial Activity Descriptors“. In: *IEEE Transactions on Affective Computing* 8.3 (2017), S. 286–299. DOI: 10.1109/TAFFC.2016.2537327.

[Wer+18] Philipp Werner u. a. „Head movements and postures as pain behavior“. In: *PLOS ONE* 13.2 (2018), e0192767. DOI: 10.1371/journal.pone.0192767.

[Wer+19a] P. Werner u. a. „Twofold-Multimodal Pain Recognition with the X-ITE Pain Database“. In: *International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 2019. DOI: 10.1109/ACIIW.2019.8925061.

[Wer+19b] Philipp Werner u. a. „Automatic Recognition Methods Supporting Pain Assessment: A Survey“. In: *IEEE Transactions on Affective Computing* (2019). DOI: 10.1109/TAFFC.2019.2946774.

[Wer+19c] Philipp Werner u. a. „Generalizing to Unseen Head Poses in Facial Expression Recognition and Action Unit Intensity Estimation“. In: *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. 2019.

[WHAH17] P. Werner, S. Handrich und A. Al-Hamadi. „Facial action unit intensity estimation and feature relevance visualization with random regression forests“. In: *International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2017. DOI: 10.1109/ACII.2017.8273631.

[Wil02] Amanda C De C Williams. „Facial expression of pain: an evolutionary account.“ In: *The Behavioral and brain sciences* 25.4 (2002), S. 439–455. DOI: 10.1017/S0140525X02000080.

[Wil95] Diana J. Wilkie. „Facial Expressions of Pain in Lung Cancer“. In: *Analgesia* 1.2 (1995), S. 91–99. DOI: 10.3727/107156995819564301.

[WJ19] Yue Wu und Qiang Ji. „Facial Landmark Detection: A Literature Survey“. In: *International Journal of Computer Vision* 127.2 (2019), S. 115–142. DOI: 10.1007/s11263-018-1097-z. arXiv: 1805.05563.

[WKW16] Karl Weiss, Taghi M. Khoshgoftaar und Ding Ding Wang. „A survey of transfer learning“. In: *Journal of Big Data* 3.1 (2016), S. 1–40. DOI: 10.1186/s40537-016-0043-6.

[WSAH15] Philipp Werner, Frerk Saxen und Ayoub Al-Hamadi. „Handling Data Imbalance in Automatic Facial Action Intensity Estimation“. In: *British Machine Vision Conference (BMVC)*. 2015, S. 124.1–124.12. DOI: 10.5244/C.29.124.

[WSAH17] Philipp Werner, Frerk Saxen und Ayoub Al-Hamadi. „Landmark based head pose estimation benchmark and method“. In: *International Conference on Image Processing (ICIP)*. 2017. DOI: 10.1109/ICIP.2017.8297015.

[WSAH20] Philipp Werner, Frerk Saxen und Ayoub Al-Hamadi. „Facial Action Unit Recognition in the Wild With Multi-Task CNN Self-Training for the EmotioNet Challenge“. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2020.

[XD13] Xuehan Xiong und Fernando De la Torre. „Supervised Descent Method and its Applications to Face Alignment“. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013, S. 532–539. DOI: 10.1109/CVPR.2013.75.

[Xu+19] Xiaojing Xu u. a. „Pain Evaluation in Video using Extended Multitask Learning from Multidimensional Measurements“. In: *Proceedings of Machine Learning Research* 116 (2019), S. 141–154.

[Yan+14] Shuang Yang u. a. „Personalized Modeling of Facial Action Unit Intensity“. In: *Advances in Visual Computing*. Hrsg. von George Bebis u. a. Lecture Notes in Computer Science. Springer International Publishing, 2014, S. 269–281.

[Yan+16] Ruijing Yang u. a. „On pain assessment from facial videos using spatio-temporal local descriptors“. In: *International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2016, S. 1–6. DOI: 10.1109/IPTA.2016.7820930.

[Yan99] Yiming Yang. „An evaluation of statistical approaches to text categorization“. In: *Information retrieval* 1.1-2 (1999), S. 69–90.

[YB97] Philip Yancey und Paul Brand. „The gift of pain“. In: *Grand Rapids, MI: Zondervan* (1997).

[Yim+15] Junho Yim u. a. „Rotating Your Face Using Multi-Task Deep Neural Network“. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2015, S. 676–684.

[Yin+17] Xi Yin u. a. „Towards Large-Pose Face Frontalization in the Wild“. In: *International Conference on Computer Vision (ICCV)*. 2017.

[YK16] Fisher Yu und Vladlen Koltun. „Multi-Scale Context Aggregation by Dilated Convolutions“. In: *International Conference on Learning Representations (ICLR)*. 2016. arXiv: 1511.07122.

[Zam+17a] Ghada Zamzmi u. a. „A Review of Automated Pain Assessment in Infants: Features, Classification Tasks, and Databases“. In: *IEEE Reviews in Biomedical Engineering* (2017), S. 1–1. DOI: 10.1109/RBME.2017.2777907.

[Zam+17b] Ghada Zamzmi u. a. „Automated Pain Assessment in Neonates“. In: *Scandinavian Conference on Image Analysis*. Springer, Cham, 2017, S. 350–361. DOI: 10.1007/978-3-319-59129-2_30.

[ZCZ16] Kaili Zhao, Wen-Sheng Chu und Honggang Zhang. „Deep Region and Multi-Label Learning for Facial Action Unit Detection“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, S. 3391–3399.

[Zen+14] Gloria Zen u. a. „Unsupervised Domain Adaptation for Personalized Facial Emotion Recognition“. In: *International Conference on Multimodal Interaction (ICMI)*. ACM Press, 2014, S. 128–135. DOI: 10.1145/2663204.2663247.

[Zha+14] Xing Zhang u. a. „BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database“. In: *Image and Vision Computing* 32.10 (2014), S. 692–706. DOI: 10.1016/j.imavis.2014.06.002.

[Zha+15] Kaili Zhao u. a. „Joint Patch and Multi-Label Learning for Facial Action Unit Detection“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, S. 2207–2216.

[Zha+16a] Zheng Zhang u. a. „Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis“. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016, S. 3438–3446.

[Zha+16b] Rui Zhao u. a. „Facial Expression Intensity Estimation Using Ordinal Information“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, S. 3466–3474. DOI: 10.1109/CVPR.2016.377.

[Zha+18] Yong Zhang u. a. „Bilateral Ordinal Relevance Multi-instance Regression for Facial Action Unit Intensity Estimation“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, S. 7034–7043. DOI: 10.1109/CVPR.2018.00735.

[Zha+19] Yong Zhang u. a. „Joint representation and estimator learning for facial action unit intensity estimation“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, S. 3452–3461. DOI: 10.1109/CVPR.2019.00357.

[Zho+12] Lin Zhong u. a. „Learning active facial patches for expression analysis“. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[Zho+16] Jing Zhou u. a. „Recurrent Convolutional Neural Network Regression for Continuous Pain Intensity Estimation in Video“. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2016. DOI: 10.1109/CVPRW.2016.191. arXiv: 1605.00894.

[Zhu+13] Zhenyao Zhu u. a. „Deep learning identity-preserving face space“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013, S. 113–120.

[Zhu+15] Xiangyu Zhu u. a. „High-Fidelity Pose and Expression Normalization for Face Recognition in the Wild“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, S. 787–796.

[ZK14] Zuhair Zafar und Nadeem Ahmad Khan. „Pain Intensity Evaluation through Facial Action Units“. In: *International Conference on Pattern Recognition*. IEEE, 2014, S. 4696–4701. DOI: 10.1109/ICPR.2014.803.

[ZLZ20] Ruicong Zhi, Mengyi Liu und Dezheng Zhang. „A comprehensive survey on automatic facial action unit analysis“. In: *Visual Computer* 36.5 (2020), S. 1067–1093. DOI: 10.1007/s00371-019-01707-5.

[Zoë+15] Sigridur Zoëga u. a. „Quality Pain Management in the Hospital Setting from the Patient's Perspective“. In: *Pain Practice* 15.3 (2015), S. 236–246. DOI: 10.1111/papr.12166.

[Zop+18] B Zoph u. a. „Learning Transferable Architectures for Scalable Image Recognition“. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018.

[ZPS17] Y Zhou, J Pi und B E Shi. „Pose-Independent Facial Action Unit Intensity Regression Based on Multi-Task Deep Transfer Learning“. In: *IEEE International Conference on Automatic Face Gesture Recognition (FG)*. 2017, S. 872–877. DOI: 10.1109/FG.2017.112.

[ZR12] X. Zhu und D. Ramanan. „Face detection, pose estimation and landmark localization in the wild“. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[Zwa+06] Sandra M G Zwakhalen u. a. „Pain in elderly people with severe dementia: A systematic review of behavioural pain assessment tools“. In: *BMC Geriatrics* 6.1 (2006), S. 3. DOI: 10.1186/1471-2318-6-3.

[ZZZ15] Stefanos Zafeiriou, Cha Zhang und Zhengyou Zhang. „A survey on face detection in the wild: Past, present and future“. In: *Computer Vision and Image Understanding* 138 (2015). DOI: 10.1016/j.cviu.2015.03.015.

Ehrenerklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die Hilfe eines kommerziellen Promotionsberaters habe ich nicht in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Verwendete fremde und eigene Quellen sind als solche kenntlich gemacht.

Ich habe insbesondere nicht wissentlich:

- Ergebnisse erfunden oder widersprüchliche Ergebnisse verschwiegen,
- statistische Verfahren absichtlich missbraucht, um Daten in ungerechtfertigter Weise zu interpretieren,
- fremde Ergebnisse oder Veröffentlichungen plagiiert,
- fremde Forschungsergebnisse verzerrt wiedergegeben

Mir ist bekannt, dass Verstöße gegen das Urheberrecht Unterlassungs- und Schadensersatzansprüche des Urhebers sowie eine strafrechtliche Ahndung durch die Strafverfolgungsbehörden begründen kann.

Ich erkläre mich damit einverstanden, dass die Dissertation ggf. mit Mitteln der elektronischen Datenverarbeitung auf Plagiate überprüft werden kann.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, 22. Oktober 2021

_____

Philipp Werner