ulm university universität ulm

Universität Ulm
Fakultät für Ingenieurwissenschaften,
Informatik und Psychologie

Institut für Neuroinformatik

Maschinelle Emotionserkennung in der Mensch-Maschine Interaktion

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat.
der Fakultät für Ingenieurwissenschaften, Informatik und Psychologie
an der Universität Ulm

vorgelegt von

Sascha Meudt
aus Wiesbaden

2018

Amtierender Dekan:
Prof. Dr.-Ing. Maurits Ortmanns

Gutachter:
PD Dr. Friedhelm Schwenker
Prof. Dr. Günther Palm

Tag der Promotion:
16. Mai 2019

ZUSAMMENFASSUNG

Zukünftige technische Systeme werden in Form von *Companion-Systemen* realisiert werden. Solche Companion-Systeme gehen in ihren technischen Möglichkeiten und ihrem Bedienkomfort weit über die Fähigkeiten gegenwärtiger smarter Technologien hinaus. Companion-Systeme werden in der Lage sein sich hoch adaptiv auf individuelle Bedürfnisse und Befindlichkeiten ihrer Nutzer anzupassen, und somit eher als Begleiter denn als technisches System wahrgenommen werden.

Um diese Adaptivität technisch zu realisieren werden Companion-Systeme, neben anderen Komponenten, mit einer Vielzahl von Sensoren ausgestattet sein, welche es erlauben ihre Umgebung und insbesondere den Nutzer multimodal zu erfassen (Sprache, Gestik, Mimik, Biophysiologie, etc.). Auf Basis dieser multimodalen Daten ist ein Companion-System in der Lage den emotionalen Zustand seiner Nutzer zu schätzen. Da menschliches emotionales Verhalten derart komplex ist, dass es durch Regeln nicht vollständig und ausreichend formal beschrieben werden kann, werden Verfahren der Mustererkennung und des maschinellen Lernens eingesetzt um multimodale Sensordaten auf emotionale Kategorien und Intensitäten abzubilden.

Diese Verfahren haben gemeinsam, dass sie aus Trainingsdaten lernen. Eine Vielzahl von Arbeiten hat gezeigt, dass für die Anwendung solcher Verfahren in der Emotionserkennung enorme Datenmengen notwendig sind, welche zudem möglichst genau der späteren Anwendungsdomäne entsprechen müssen. Im Falle der *Companion-Technologie* entspricht diese der teilweise emotionsbehafteten Mensch-Maschine Interaktion (MMI), welche sich deutlich von der emotionalen Mensch-Mensch Interaktion unterscheidet.

Diese Arbeit beschreibt zunächst notwendige Eigenschaften, welche Trainingsdaten erfüllen müssen, um optimal zum Training von Klassifikatoren geeignet zu sein, welche in der Companion-Technologie eingesetzt werden können. Da die identifizierten Eigenschaften von keinem bisherigen Korpus erfüllt werden, wird mit dem Ulm University Multimodal Affective Corpus (uulmMAC) ein Korpus entworfen und erhoben, welcher die notwendigen Eigenschaften erfüllt. Dabei verwendet der Korpus neben klassischen kontinuierlichen Emotionsmodellen zusätzliche MMI Kategorien anhand derer unter Anderem mentale Überforderung und Unterforderung untersucht werden kann.

Da solch umfangreiche Datenbasen nicht, oder nur unter enormem Aufwand, auf klassischem Wege manuell annotiert werden können, werden um Multimodalität erweiterte Ansätze des *aktiven Lernens* vorgeschlagen, um eine multimodale Annotation unter Verwendung von

Annotations-Assistenzsystemen zu ermöglichen. Der notwendige Aufwand der Annotation wird so um ein vielfaches reduziert, ohne Rückgang der erreichbaren Klassifikatorperformanz. Diese aktiven Lernverfahren sind Bestandteil des entwickelten Annotations- und Analysewerkzeuges Annotation Tool with a Learning Assistive System (ATLAS). ATLAS begleitet den Prozess der Entwicklung emotionserkennender Komponenten der Companion-Technologie nicht nur bei der Annotation, sondern durch entsprechende Analyse- und Visualisierungsfunktionen bereits vom explorativen Trainingsdatenentwurf über die Annotation, dem Klassifikatorentwurf mit entsprechender Evaluation hin zum prototypischen Systementwurf.

Aufbauend auf dem annotierten uulmMAC werden multimodale Klassifikationsverfahren und Architekturen evaluiert, welche es ermöglichen den mentalen Belastungszustand eines Nutzers zu erfassen. Dabei wird für jede Modalität, sowie für die fusionierte Schätzung, zwischen individueller und interindividueller Klassifikation unterschieden. Interindividuelle Klassifikatoren haben dabei keine Trainingsdaten des zu schätzenden Nutzers gesehen und performen naturgemäß schlechter gegenüber den individualisierten Varianten, welche gute Erkennungsraten aufweisen und den in Companion-Systemen eingesetzten nutzeradaptiven Szenarien entsprechen.

Abschließend werden die, auf Basis der erhobenen und annotierten Daten entworfenen Klassifikatoren, in prototypischen Companion-Systemen eingesetzt, und auf ihre reale Anwendungsfähigkeit hin bewertet. Dafür wurde das bereits zur sensorischen Datenerfassung entwickelte Modular Affect Recording and Recognition System (MAR^2S) Architektur erweitert um einen vollständig modularen Aufbau der Prototypen zu erlauben. Hierdurch wird eine flexible und modulare Systementwicklung unter Verwendung von Erkenntnissen aus der Softwaretechnik ermöglicht.

INHALTSVERZEICHNIS

1. Einleitung **5**

1.1. Motivation 5
1.2. Aufbau der Arbeit 9
1.3. Wesentliche Beiträge 10

2. Multimodale Datenbasen **15**

2.1. Psychologische Grundlagen 16
2.1.1. Kategorielle Emotionsmodelle 16
2.1.2. Dimensionale Emotionsmodelle 21
2.1.3. MMI Emotionsmodelle 23
2.2. Stand der Forschung - Datensätze 25
2.3. Ulm University Multimodal Affective Corpus - uulmMAC 29
2.3.1. Experimenteller Aufbau 29
2.3.2. Quantität und Qualität der erhobenen Daten 33
2.3.3. Technische Aufnahme Infrastruktur - MAR^{2}S 36
2.3.4. Annotation des uulmMAC 50
2.3.5. Merkmalsextraktion 55

3. Annotation multimodaler Datenbasen **57**

3.1. Stand der Forschung - Annotation 59
3.1.1. Verfügbare Annotationswerkzeuge 59
3.1.2. Defizite verfügbarer Annotationswerkzeuge 64
3.1.3. Aktive Lernverfahren für multimodale affektive Daten 65
3.2. Aufwand manueller Annotation 72
3.3. Methoden aktiver Lernverfahren für multimodale affektive Daten 73
3.3.1. Unimodale aktive Lernverfahren affektiver Daten 74
3.3.2. Multimodale aktive Lernverfahren affektiver Daten 91
3.3.3. Abbruchkriterium 98
3.3.4. Ausblick aktiver Lernverfahren 100
3.4. EmotionWheel 102
3.5. ATLAS 103
3.5.1. Aufbau und Funktionsweise von ATLAS 105

3.5.2. Aktive und passive Hilfsfunktionen 109
3.5.3. Aktive und passive Schnittstellen 114
3.5.4. Aktive Annotationsschnittstellen und -Funktionen 117
3.5.5. Verteilte Annotation 120
3.5.6. Technische Realisierung 123

4. Maschinelle Emotionserkennung 125
4.1. Stand der Forschung - Maschinelle Emotionserkennung 125
4.2. Informationsverdichtung 131
4.2.1. Selektion von Instanzen 131
4.2.2. Selektion von Merkmalen 133
4.3. Unimodale Emotionserkennung 137
4.3.1. Erkennung anhand von Prosodie 137
4.3.2. Erkennung anhand von Gesichtsexpressionen 144
4.3.3. Erkennung anhand von Körperposen 146
4.3.4. Erkennung anhand von Biophysiologie 152
4.4. Multimodale Emotionserkennung 155
4.4.1. Hierarchische audio-visuelle Emotionserkennung 156
4.4.2. Bimodale Erkennung von kognitiver Last 157
4.4.3. Multimodale Erkennung von kognitiver Last 161
4.4.4. Einbezug von situativem Kontextwissen 166

5. Technische Realisierung von Companion-Systemen 169
5.1. Stand der Forschung - Technische Realisierung von Companion-Systemen . . 170
5.2. Modulare dezentrale Architektur, MAR^{2}S 174
5.2.1. Modularer Aufbau und Kapselung der Funktionalitäten 174
5.2.2. Netzwerkbasierte Modulinteraktion 176
5.2.3. Organisation der Module 177
5.2.4. Systemrealisierung unter Verwendung von Softwaretechniken 179
5.3. Prototypische Realisierung von Companion-Systemen 180
5.3.1. Verteiltes Nutzertracking und dezentrale Nutzeridentifikation 181
5.3.2. Das Demonstrationsszenario 3 (DS3) des SFB-TRR 62 187
5.3.3. Adaptive Nutzer-Interaktion 194

6. Diskussion 205
6.1. Multimodale Datenbasen 205
6.2. Annotation multimodaler Daten 207
6.3. Maschinelle Emotionserkennung 208
6.4. Prototypische Companion-Systeme 209

6.5. Ganzheitliche und zusammenhängende Betrachtung der Teilbereiche 210
6.6. Fazit 213
6.7. Ausblick 214

A. Implementierte Komponenten **217**

B. Abstracts eigener Publikationen **223**

Abkürzungsverzeichnis **239**

Abbildungsverzeichnis **242**

Tabellenverzeichnis **245**

Bibliographie **248**

KAPITEL 1

EINLEITUNG

1.1. Motivation

Technische Systeme der Zukunft werden in Form von *Companion-Systemen* realisiert werden. Darunter werden kognitive technische Systeme verstanden, die ihre Funktionalität vollkommen individuell auf den jeweiligen Nutzer anpassen. Sie werden sich an Fähigkeiten, Vorlieben, Anforderungen und aktuellen Bedürfnissen ihrer Nutzer orientieren, und stellen sich dabei auf die Situation und emotionale Befindlichkeit ihres Nutzers ein. Companion-Systeme verhalten sich kooperativ und vertrauenswürdig, dabei treten sie dem menschlichen Nutzer als partnerschaftliche Dienstleister gegenüber.

Um solche Systeme realisieren zu können, bedarf es interdisziplinärer Kooperationen aus Informatikern, Ingenieuren, Neurobiologen und Psychologen. Die Forschergemeinschaft befasst sich dabei mit der Erforschung verschiedenster Fragestellungen zu kognitiven Fähigkeiten und deren Realisierung in technischen Systemen. Zentrale Eigenschaften von Companion-Systemen bilden die *Individualität*, *Anpassungsfähigkeit*, *Verfügbarkeit*, *Kooperativität* und *Vertrauenswürdigkeit*, welche als Companion-Eigenschaften bezeichnet werden [25, 27, 272]. Grundlage der Realisierung solcher Eigenschaften werden kognitive Prozesse in technischen Systemen sein. Diese werden die Grundlagen für eine Technologie schaffen, die menschlichen Nutzern eine völlig neue Ebene des Umgangs mit technischen Systemen eröffnen wird.

Die uns Menschen umgebende Geräte des Alltags – Haushaltsgeräte, Unterhaltungselektronik, Smartphones, Automaten oder Anwendungssysteme – werden bereits heutzutage, durch den Einsatz von Software-Technik (engl. Systems Engineering) Methoden, in immer kürzeren Innovationszyklen entwickelt und dabei scheinbar immer „intelligenter“. Dem menschlichen Wunsch diese innovativen Produkte umfangreich zu nutzen steht jedoch in vielen Fällen mangelnder Bedienkomfort und unzureichende Nutzerfreundlichkeit gegenüber. Ziel der Companion-Technologie ist es, neben der zunehmenden funktionalen Intelligenz technischer Systeme, ebenso eine emotinale Intelligenz zu realisieren. Dazu ist es notwendig Methoden und

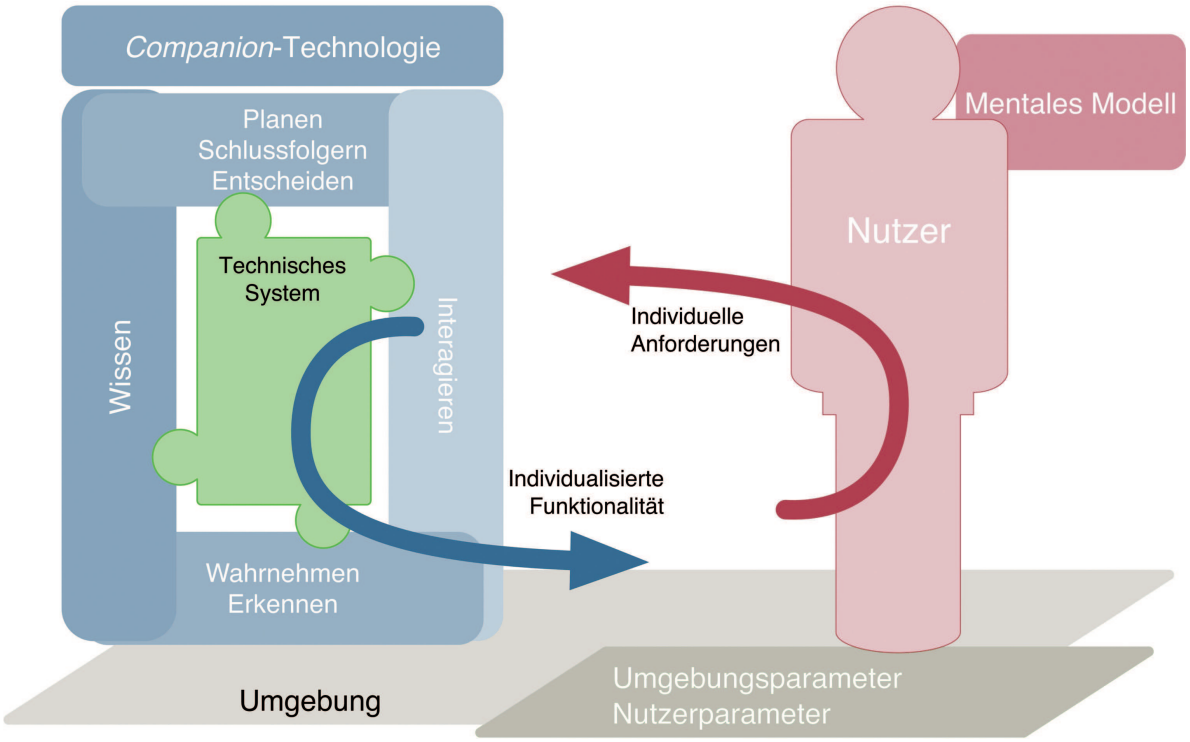

Abbildung 1.1.: Bezug der Companion-Technologie zu Nutzer und Umwelt. Bildquelle: Biundo et. Al. `https://forschung-sachsen-anhalt.de/news/cebit-news-companion-systeme-technik-730` (Stand 3. Juni 2019)

Modelle zu schaffen, sowie die Werkzeuge zur Realisierung von Companion-Eigenschaften in kognitiven technischen Systemen zu entwickeln. Weiter ist es notwendig Vorgehensmodelle und Referenzarchitekturen für Companion-Systeme zu entwerfen, und diese anhand exemplarischer Implementierungen für typische Anwendungsszenarien zu erproben.

Zur Erreichung dieser Ziele stehen die zentralen kognitiven Fähigkeiten der Planung und Schlussfolgerung, der Interaktion und Dialogführung, sowie der Wahrnehmung und Erkennung im Mittelpunkt der Forschung. Ein Schwerpunkt der Companion-Technologie Entwicklung liegt auf der Entwicklung von Konzepten, Modellen, Methoden und Experimentalparadigmen. Im Mittelpunkt stehen dabei die Eigenschaften der *Individualität*, der *Emotionsadaptivität* und der prototypischen *Methoden-* und *Modellintegration*. Die bidirektionale Verzahnung und Rückkopplung modularer Systemprozesse und Komponenten nehmen eine zentrale Stellung bei der Realisierung dieser Eigenschaften ein.

Das System fokussiert sich aus seiner Sicht dabei auf die Konstruktion und Integration von kognitiven Funktionseinheiten und die Realisierung von Companion-Eigenschaften durch diese Funktionseinheiten. Während der Nutzer aus seiner Perspektive ein mentales Modell des Systems entwickelt, welches wiederum sein Verhaltensmodelle beeinflusst. Der zunehmende

Einsatz immer komplexerer Systeme steigert auch die Anforderungen, die bei deren Bedienung an den Nutzer gestellt werden. Gleichzeitig bietet jedoch die technologische Entwicklung neue Möglichkeiten der technischen Unterstützung und digitalen Assistenz.

Abbildung 1.1 zeigt die grobe Gliederung eines solchen technischen Companion-Systems und die Interaktion mit seinem Nutzer, eingebettet in eine globale Umgebung. Das technische System besteht aus den Teilbereichen der *sensorischen Umgebungswahrnehmung*, der *Wissensmodellierung und Akquise*, des *Planens und Entscheidens* und der *Interaktionsgestaltung* mit dem Nutzer. Nutzer und System sind in die Umwelt eingebettet, wobei ein besonderes Augenmerk auf den Umgebungsparametern des Nutzers liegt, da eine vollständige Erfassung der globalen Umwelt wenig realistisch erscheint. Das System stellt seine individualisierte Funktionalität den individuellen Bedürfnissen des Nutzers bereit.

Der Sonderforschungsbereich Transregio 62 (SFB-TRR 62) „*Eine Companion-Technologie für kognitive technische Systeme*“[1][23, 24, 26] mit seinen beteiligten Partnern *Universität Ulm*, *Otto-von-Guericke-Universität Magdeburg* und dem *Leibniz-Institut für Neurobiologie Magdeburg* beschäftigt sich mit den Forschungsfragen rund um die Entwicklung von Companion-Technologien. Diese Arbeit wurde im Rahmen des Teilprojektes *C5* - „*Informationsfusion zur Emotions- und Dispositionserkennung*“ des SFB-TRR 62, welches sich mit der multimodalen Wahrnehmung und Erkennung von Nutzeremotionen beschäftigt, angefertigt.

Das Ziel der Informationsfusion in einem Companion-System besteht in der Erstellung eines umfassenden Modells zur Emotionsinterpretation für die Planungs- und Entscheidungsebene. Für die zuverlässige Erkennung der Nutzeremotion auf der Basis gesprochener Sprache, Gestik, Mimik und psychobiologischer Daten werden multimodale Informationsfusionsarchitekturen verschiedener Abstraktionsebenen entwickelt und evaluiert. Grundlage für die Erstellung des Emotionsmodells sind verschiedene Sensordaten (Kameras, Mikrofone, usw.). Desweiteren werden die sensorbasierten Architekturen und Ergebnisse so erweitert, dass sie die Integration anwendungsabhäniger symbolischer Kontextinformation erlauben.

Diese Arbeit erforscht und realisiert dabei die Module der multimodalen sensorischen Datenerfassung, Mustererkennung mittels maschineller Lernverfahren zur Emotionsschätzung und Fusion multimodaler Datenströme zu einer Einschätzung der emotionalen Gesamtsituation. Ferner werden technische Lösungen und Architekturen entwickelt, welche die modulare Konstruktion, als dezentrales System erlauben.

Indikatoren für den emotionalen Zustand des Nutzers können ihren Ursprung im Videokanal haben, ausgedrückt durch Mimik [196, 146, 72, 74], Körperhaltung [8, 60] oder Handgesten [40, 39, 159]. Darüber hinaus können Audiosignale Informationen über den emotionalen

[1] `http://www.sfb-trr-62.de/` (Stand: 3. Juni 2019)

Zustand des Benutzers [177, 185] beinhalten. Auch das Erkennen von Lachen liefert adäquate Informationen und gibt einige Einblicke in den emotionalen Zustand des Nutzers [207]. Eine weitere Quelle für emotionale Indikation ist der biophysiologische Kanal. Aus Herzfrequenz, Hautleitwert und EMG-Daten extrahierte Merkmale können verwendet werden um den emotionalen Zustand des Nutzers zu schätzen [176, 175, 174]. Biophysiologische Daten haben dabei den Vorteil, dass sie immer verfügbar sind und vom Nutzer nur sehr eingeschränkt willentlich beeinflusst werden können [280, 260].

Die Lösung eines solchen multimodalen Erkennungsproblems ist bei weitem nicht trivial. Eine regelbasierte formale Beschreibung von Emotionen ist nicht möglich. Dies zeigt sich schon an der Tatsache, dass wir Menschen wissen was wir meinen, wenn wir über Emotionen reden. Werden wir jedoch aufgefordert eine exakte, möglichst formale Beschreibung und Abgrenzung einer Emotion zu geben, ist uns dies kaum möglich. Wir haben in unserem Leben anhand von Beispielen seit frühester Kindheit gelernt was Emotionen sind, wie sich diese anfühlen und äußern und wie wir diese an unserem Gegenüber erkennen und wahrnehmen können. Ähnlich verhält es sich auch mit maschinellen Systemen. Da eine exakte formale regelbasierte Beschreibung zu komplex ist, werden Mustererkennungsverfahren des maschinellen Lernens eingesetzt, welche in der Lage sind anhand von Beispielen (Trainingsdaten mit und ohne Lehrersignalen) eine Abbildung von sensorischen Daten auf emotionale Kategorien und Intensitäten zu erlernen [163, 38, 259, 213]. Dazu müssen allerdings zunächst ausreichend Trainingsdaten verfügbar sein [190, 273, 199, 92].

Der Großteil verfügbarer Datensätze für das Training besteht derzeit aus schauspielerisch dargestellten Emotionen [121, 267], die oft ausdrucksvoller sind als Emotionen des realen Lebens. Werden Nutzer in einer natürlicheren Umgebung beobachtet, treten emotionale Ausdrücke seltener auf [62]. Darüber hinaus existieren bis heute sehr verschiedene psychologische Emotionsmodelle [71, 70, 197, 204], daran ist zu erkennen wie kontrovers diese Themen diskutiert werden. Es ist eine offene Frage, welches Modell am besten passt und ob dieses überhaupt für den Einsatz in den Mensch Maschine Interaktion (MMI) Komponenten der Companion-Technologie geeignet ist.

Um multimodale Erkenner trainieren zu können, stellt sich also zunächst die Frage, wie geeignete Trainingsdaten beschaffen sein müssen im Bezug auf Umfang, Qualität, Modalitäten, Expressionsgrad und dem verwendeten psychologischen Model. Darauf aufbauend stellt sich die Frage nach der Annotation erfasster Daten, da die Qualität der Lehrersignale ebenfalls erheblichen Einfluss auf die erreichbaren Resultate hat [120, 99, 199]. Diese Arbeit widmet sich daher zunächst den Fragestellungen zur Datenerhebung und Annotation. Stehen geeignete annotierte Trainingsdaten als Grundlage zur Verfügung, können maschinelle Lernverfahren zur multimodal fusionierten Emotionserkennung trainiert werden. Anforderungen

an die Eigenschaften solcher Verfahren werden ebenfalls erörtert. Die trainierten Erkenner finden abschließend ihren Einsatz in prototypischen Companion-Systemen. Daraus resultiert die abschließend in der Arbeit bearbeitete Fragestellung nach einer möglichen Architektur der zu Beginn beschriebenen zukünftigen Companion-Systeme und ihren Fähigkeiten.

1.2. Aufbau der Arbeit

Die Gliederung der Arbeit erfolgt anhand der logisch aufeinander aufbauenden Reihenfolge der notwendigen Schritte zur Realisierung prototypischer Komponenten eines Companion-Systems. Solche prototypischen Module zur Erkennung der Nutzeremotion beinhalten als zentrale methodische Komponenten eine multimodale Fusionsarchitektur. Die Klassifikatoren der Fusionsarchitekturen werden auf zuvor annotierten emotionalen Datenbasen trainiert. Solche Datenbasen bilden also den Ausgangspunkt zur Realisierung von Wahrnehmungs- und Erkennungskomponenten einer Companion-Technologie. Die Arbeit gliedert sich in vier zentrale Bereiche - *Datenbasen*, *Annotation*, *Emotionserkennung* und *technische Realisierung von Companion-Systemen* - welche die Grundlage der Companion-Systeme bilden.

Im folgenden Kapitel werden zunächst die psychologischen Modellierungsgrundlagen beschrieben. Anhand derer werden bereits existierende Datenbasen im Hinblick auf ihre Eignung als Trainingsgrundlage bewertet. Als Resultat zeigt sich eine nur bedingte Einsetzbarkeit existierender Datenbasen zur Realisierung von prototypischen Erkennern. Als Lösung wurde der uulmMAC als alternativer Korpus erarbeitet, welcher in Kapitel 2 als zentrales Ergebnis präsentiert wird.

Kapitel 3 erörtert Herausforderungen bei der Annotation umfangreicher multimodaler Datenbasen, wie dem *Ulm University Multimodal Affective Corpus* (uulmMAC), welche eine vollständige, klassisch manuelle Annotation unmöglich machen. Zur Handhabbarkeit der Annotation solcher Korpora werden aktive Lernverfahren vorgeschlagen, welche derart erweitert werden, dass sie den Annotationsaufwand erheblich senken ohne dabei die Qualität der Annotation negativ zu beeinflussen. Die vorgeschlagenen Algorithmen sind Bestandteil des ebenfalls in Kapitel 3 vorgestellten Annotationswerkzeuges *Annotation Tool with a Learning Assistive System* (ATLAS), welches eine effiziente und umfangreiche multimodale kontinuierliche Annotation beliebiger Datenbasen ermöglicht.

Nachdem umfangreiche multimodale Datenbasen in annotierter Form zur Verfügung stehen, werden in Kapitel 4 Verfahren beschrieben um die Trainginsdateninformationen zunächst weiter zu verdichten. Auf Basis dieser Daten werden dann unimodale Merkmale und Klassifikations-

methoden beschrieben. Deren Resultate werden anschließend durch geeignete Fusionsarchitekturen unter Einbezug eventuell vorhandenem situativem Kontext zu einer Gesamtschätzung des emotionalen Zustandes des Nutzers fusioniert.

Nach Bereitstellung geeigneter methodischer Verfahren zur Emotionserkennung und Training solcher Klassifikatoren, werden in Kapitel 5 prototypisch realisierte Companion-Systeme vorgestellt. Diese Systeme sind als modulare Systeme realisiert. Kapitel 5 schlägt die entwickelte *Modular Affect Recording and Recognition System* (MAR^{2}S) Architektur vor, welche Entwurf und Realisierung von Companion-Systemen ermöglicht, und dabei in moderne Verfahren des Software Engineerings wie dem V-Modell eingebunden werden kann.

Eine Bewertung des aktuellen Standes der Forschung befindet sich jeweils am Anfang der vier Themengebiete. Eine abschließende gemeinsame Diskussion und Zusammenfassung der vier Bereiche bilden neben dem Fazit den Abschluss der Arbeit.

1.3. Wesentliche Beiträge

Die Resultate dieser Arbeit basieren auf vier Teilbereichen der maschinellen Emotionserkennung zum Einsatz in der Companion-Technologie. Diese umfassen die Erhebung von Trainingsdaten in realistischen schwach expressiven MMI Szenarien, eine effektive Annotation umfangreicher Datenbasen zur Gewinnung von Trainingsdaten, die Entwicklung geeigneter multimodaler Klassifikatorarchitekturen zur effizienten Emotionsprädiktion im MMI Kontext und die technische Umsetzung von Companion-Technologie Architekturen.

Konzeption und Entwicklung der multimodalem Datenbasis uulmMAC

Multimodale Datenbasen als Trainingsgrundlagen bilden die Basis der maschinellen Emotionserkennung. Um Klassifikatorarchitekturen erproben zu können, welche in realistischen MMI Szenarien außerhalb einer Laborumgebung eingesetzt werden können, müssen diese Datenbasen weitgehend ähnliches Nutzerverhalten beinhalten, entsprechend dem erwarteten Verhalten tatsächlicher Nutzer im späteren technischen Einsatz.

Diese Arbeit untersucht zunächst eine Vielzahl verfügbarer emotionaler Datenbasen im Bezug auf ihre Verwendbarkeit als Trainingsmaterial. Dabei wurde festgestellt, dass bisher in der Literatur beschriebene Korpora zwar geeignet sind eine Vielzahl von methodischen Arbeiten zur maschinellen Emotionserkennung zu ermöglichen, jedoch ist ihre Einsetzbarkeit mit dem Ziel real einsetzbare Systeme zu entwickeln nachweisbar nicht gegeben. Dies liegt zum Einen an

der zumeist geringen Anzahl von synchron erfassten Modalitäten und der Überexpressivität der darin enthaltenen emotionalen Ausdrücke. Zudem werden in der Regel klassische kategorielle Emotionsmodelle verwendet, welche nicht direkt auf Situationen der MMI abgebildet werden können.

Zur Lösung dieser Defizite wird der uulmMAC im Rahmen dieser Arbeit konzeptioniert und erhoben. Das Szenario des uulmMAC besteht aus einer abstrakten MMI in der Emotionen und Affekte derart auftreten, wie sie auch in realen Anwendungen zu erwarten sind. Der Korpus definiert die kognitiven Zustände *Überforderung*, *Unterforderung*, *Frustration* und *Neutralität* anhand einer Abbildung in das dimensionale Valenz-Arousal-Dominance (VAD) Emotionsmodell. Der Korpus ist einer der umfangreichsten Korpora die bisher erhoben wurden. Er umfasst 60 Probanden, aufgenommen in 100 Sessions zu je 45 Minuten. Die Aufzeichnung erfolgt unter Verwendung einer Vielzahl von Sensoren, resultierend in 15 synchron erfassten Sensordatenströmen (Modalitäten). Die technische Infrastruktur zur triggerfreien synchronen Erfassung wird in dieser Arbeit als Teil der MAR^{2}S Architektur vorgeschlagen und realisiert.

Erkenntnise der Analyse bestehender affektiver Korpora, sowie Informationen zur Erhebung des uulmMAC, wurden in [254, 126, 153, 226, 125, 202] publiziert.

Annotation multimodaler Datenbasen

Die Analyse der Annotation emotionaler Datenbasen zeigt, dass diese zumeist nur unzureichend annotiert sind. Die Gründe hierfür sind weitreichend. Teilweise werden emotionale Induktionshypothesen schlicht nicht überprüft, die Annotation erfolgt ausschließlich durch nicht überprüfte Selbsteinschätzungen des Probanden oder bei der Annotation wurden konzeptionelle und technische Fehler gemacht. Die Untersuchung der Annotationssituation lässt dabei zwei wesentliche Defizite als ursächlich erkennen. Zum Einen sind keine hochwertigen Annotationswerkzeuge verfügbar die einen Annotationsprozess adäquat unterstützen, zum anderen ist die Annotation umfangreicher Datenbasen zu zeitaufwendig und somit kostenintensiv. In dieser Arbeit identifizierte Nachteile bestehender Annoationsmethoden wurden unter Anderem in [125, 211, 126] publiziert.

Mit ATLAS wird ein Annotationswerkzeug geschaffen, welches die Annotation beliebiger Datenbasen ermöglicht. Die Größe einer Datenbasis, die Anzahl der darin enthalten Modalitäten und das zu verwendende Annotationsschema ist in ATLAS nicht beschränkt. ATLAS erlaubt kategorielle, kontinuierliche und hierarchische Annotationen. Durch seine zahlreichen Zusatzfunktionen kann ATLAS nicht nur zur Annotation eingesetzt werden, vielmehr begleitet es den gesamten Entwicklungsprozess von ersten Vorstudien eines Experimentes, über die

Annotation eines durchgeführten Experimentes, bis hin zur Begutachtung von Merkmalen und Klassifikatorresultaten, und schließlich der Erstellung von Prototypen oder Demonstratoren. Neben der flexiblen Gestaltung der Annotation stellen die aktiven Lernkomponenten einen wesentlichen Mehrwert von ATLAS dar. Diese beinhalten sowohl in ATLAS implementierte Lernverfahren, als auch eine interaktive Echtzeitschnittstelle zu Matlab zum Einsatz weiterer aktiver Lernverfahren und anderer Methoden. ATLAS wird der wissenschaftlichen Gemeinschaft frei zur Verfügung gestellt, es wurde in [151, 150, 156] veröffentlicht.

Unter Ausnutzung der Spärlichkeit des Auftretens emotionaler Ereignisse, welche in realistischen MMI Datensätzen zu beobachten sind, werden aktive Lernverfahren entwickelt, welche den Annotationsaufwand erheblich reduzieren. Diese affektiven aktiven Lernverfahren werden zunächst methodisch unimodal erprobt und anschließend zu multimodalen Verfahren erweitert. Resultate der Arbeit zeigen, dass die vorgeschlagenen Verfahren den Annotationsaufwand um 90% veringern können, ohne dass dabei die Qualität der daraus trainierten Klassifiaktoren abnimmt. Die vorgeschlagenen Methoden sind in [254, 255, 153, 253] beschrieben.

Der Einsatz dieser affektiven aktiven Lernverfahren unter Verwendung von ATLAS erlauben es aktuelle und zukünftige enorm große Datenbasen weiterhin beherrschbar zu halten. Der Erfolg von ATLAS zeigt sich auch durch seine mittlerweile verbreitete Verwendung in diversen internationalen Arbeitsgruppen.

Multimodale Emotionserkennung

In der Literatur sind eine Vielzahl von Arbeiten zur unimodalen Emotionserkennung bekannt. Auch multimodale Ansätze wurden in den vergangen Jahren umfangreich vorgeschlagen und evaluiert. Die meisten Arbeiten beschränken sich, aufgrund der wenigen verfügbaren hoch multimodalen Datensätze, jedoch auf die Analyse von Audio- und Videodaten. Erst in der jüngsten Vergangenheit finden sich Arbeiten, welche zusätzlich auch biophysiologische Daten hinzunehmen. Die Arbeiten im Feld der Emotions- und Affekterkennung erstrecken sich dabei über ein weites Anwendungsfeld aus den Bereichen der Medizin, Psychologie, Neuroprotetik, MMI, Entertainment, etc.

Diese Arbeit fokussiert sich auf die Vorstellung von multimodalen Klassifikationsarchitekturen zur Emotionsschätzung im Kontext der MMI. Dazu werden zunächst Verfahren der Informationsverdichtung (Clustering, Instanzselektion und Merkmalsselektion) auf ihrer Einsatzfähigkeit im gegebenen Kontext untersucht (Veröffentlicht in [229, 154, 285, 158]). Darauf aufbauend werden zunächst unimodale Erkenntnisse gewonnen [106, 105, 129, 128, 155, 83], welche abschließend zu multimodalen Architekturen erweitert wurden [101, 131, 153, 226, 123, 227,

201, 202]. Die Modalitäten umfassen dabei einen einmaligen Umfang, hervorzuheben ist auch der erfolgreiche Einbezug von situativem Kontext zur Verbesserung der Emotionsschätzung.

Die Resultate zeigen, dass ein Einsatz in realistischen MMI Szenarien bisher nur im Falle individualisierter Klassifikatoren machbar ist, während innerhalb der Grenzen eines expressiven Benchmarkkorpuses durch Fusion durchaus akzeptable interindividuelle Ergebnisse erreicht werden können. Die für Demonstratoren erarbeiteten Architekturen werden auf Grund der Erkenntnisse dieser Arbeit in individualisierter Form realisiert.

Architektur von Companion-Systemen

Bisher existieren Companion-Systeme lediglich als nicht vollumfänglich funktionierende Prototypen und Demonstratoren. Im Rahmen dieser Arbeit wurde an solchen Demonstratoren mitgewirkt. Es werden eigene Anwendungsbeispiele und prototypische Szenarien vorgestellt und in unterschiedlichem Umfang realisiert. Dabei zeigt sich, dass Companion-Systeme zukünftig als modulare dezentrale Systeme realisiert werden müssen um ihre Komplexität beherrschbar zu halten.

Mit MAR^{2}S schlägt diese Arbeit einen modularen netzwerkbasierten Architekturansatz vor, welcher eine flexible Realisierung von Companion-Systemen erlaubt. Der MAR^{2}S Ansatz ist sowohl bei der Erhebung von Trainingsdaten oder verteilten Annotationen einsetzbar, als auch zur Realisierung tatsächlicher Companion-Systeme. Ein Architekturparadigmenwechsel ist somit während der gesamten Entwicklungszeit nicht notwendig. Die modulare MAR^{2}S Struktur erlaubt zudem die Entwicklung in größeren Umgebungen, bei der sich nicht zwingend alle beteiligten Projektpartner auf einheitlichen Entwicklungsstufen befinden müssen. Es ist hervorzuheben, dass MAR^{2}S so den Einsatz von Softwareentwicklungsmethoden wie dem V-Modell oder SCRUM erlaubt. Die MAR^{2}S Architektur wird über diese Arbeit hinaus bereits von weiteren Arbeitsgruppen zur Datenerfassung und Systemrealisierung eingesetzt.

Die entworfenen Prototypen und Prototypen beim deren Entwurf maßgeblich mitgewirkt wurde, sind in [153, 157, 226, 152] publiziert. Abschließend ist noch die Publikation [153] hervorzuheben. Diese setzt alle Teilbereiche der Arbeit in eine gemeinsame Beziehung und zeigt zudem Querbezüge zu angrenzenden Forschungsgebieten der MMI im Feld der Companion-Technologie auf.

KAPITEL 2

MULTIMODALE DATENBASEN

Maschinelle Lernverfahren lernen aus Trainingsbeispielen. Hierfür sind große Datenmengen in guter Qualität notwendig. Zum Training von Emotionserkennern ist die Erzeugung geeigneter Trainingsdaten nicht trivial. Der Entstehungskontext einer Emotion hat Einfluss auf die Art wie Menschen Emotionen äußern. Daher ist es nur schwer möglich Trainingsbeispiele aufzunehmen um einen Erkenner zu trainieren, welcher später in einem veränderten Umfeld eingesetzt werden soll. Es ist unbedingt erforderlich, dass die auftretenden Emotionen möglichst natürlichen Ursprungs sind und den Mustern entsprechen, welche in einer späteren realen Anwendung vorkommen. Transferlernen wird in diesem Kontext noch nicht erfolgreich eingesetzt. Emotionen werden zudem zumeist in mehreren Modalitäten ausgedrückt. Um diese Modalitäten erfassen zu können sind verschiedenste Sensoren notwendig. Um einen zeitlichen Zusammenhang zwischen Modalitäten herstellen zu können, ist es notwendig alle Sensoren in einer zeitsynchronen Umgebung aufnehmen zu können. Neben den inhaltlichen Anforderungen sind also auch die technischen Synchronitätsanforderung sowie Anforderungen an die sensorische Qualität von großer Bedeutung.

In den folgenden Abschnitten werden zunächst die emotionstheoretischen Grundlagen beschrieben und deren Anwendbarkeit auf die MMI diskutiert. Anschließend werden die wesentlichen öffentlich verfügbaren emotionalen Korpora vorgestellt. Da diese Korpora diverse Nachteile aufweisen, welche bei der Entwicklung von Emotionserkennern für die MMI problematisch sind, wird weiter der für diese Arbeit erhobene uulmMAC vorgestellt. Dabei wird neben einem auf die Bedürfnisse der MMI angepassten Versuchsablaufes auch die technische MAR^{2}S Umgebung vorgestellt, welche es ermöglicht eine unbeschränkte Anzahl von Sensoren in einer verteilten Infrastruktur aufzunehmen.

2.1. Psychologische Grundlagen

Die psychologische Forschung des menschlichen emotionalen Verhaltens und dem Ausdruck von Empfindungen hat eine lange Vergangenheit. Erste wissenschaftliche Veröffentlichungen finden sich bereits im 19. und 20. Jahrhundert [118, 102, 82, 276, 41]. Die Verarbeitung menschlicher emotionaler Eigenschaften in der MMI ist hingegen eine relativ junge Disziplin der Informatik. Erste erfolgversprechende Ansätze der Mustererkennung hierzu gehen auf den Anfang des 21. Jahrhunderts zurück [173, 17, 143, 115]. Seit Beginn der emotionalen Mustererkennung wurde in der Literatur auf eine Vielzahl verschiedener psychologischer Emotionsmodelle zurückgegriffen. Diese unterscheiden sich dabei teilweise erheblich in ihren Eigenschaften. Grundsätzlich kann zwischen kategoriellen und dimensionalen Modellen unterschieden werden. Da beide Varianten aus Untersuchungen des menschlichen Verhaltens ohne MMI Bezug hergeleitet wurden, sind diese in der MMI in ihrer psychologisch fundierten Form nur bedingt einsetzbar. Dies hat dazu geführt, dass in der MMI weitere spezielle Modelle entwickelt wurden. Auch sind bei der Emotionsforschung und Emotionsverarbeitung individuelle und kulturelle Unterschiede zu beachten. Im folgenden werden die gebräuchlichsten Modelle erläutert.

2.1.1. Kategorielle Emotionsmodelle

Paul Ekman beschreibt in seinen Arbeiten sechs Basis Emotionen [71, 70], bestehend aus:

- Freude
- Wut
- Ekel
- Furcht
- Trauer
- Überraschung

Nach seiner Auffassung setzt sich der menschliche Gemütszustand aus einer Kombination aus diesen Kategorien zusammen. Nach Ekman gilt dies, aufgrund der phylogenetischen Entstehung, für alle Menschen verschiedener Kulturen und Herkunft gleichermaßen. Dementsprechend wird ein kulturübergreifendes, universelles, mimisches Ausdrucksverhalten (z. B. Lachen oder Weinen) als Beleg für die Existenz der Basisemotion angesehen [73]. Der Ausdruck dieser

Abbildung 2.1.: Mimischer Ausdruck der Basisemotionen nach Ekman (Bildquelle: Aronson et. Al. Seite 86 [7])

Emotionen kann sich im sozialen Kontext jedoch unterscheiden. Begründet ist dies unter anderem durch die Erwünschtheit oder Unerwünschtheit von Emotionen in unterschiedlichen sozialen Kontexten. Diese Unterschiede sind beim Entwurf emotionsklassifizierender Systeme zu berücksichtigen, da sich bereits das notwendige Trainingsdatenmaterial unterscheiden kann. Auch bezieht sich keine Arbeit von Ekman auf einen MMI Kontext, vielmehr basieren seine Untersuchungen auf Mensch-Mensch Interaktionen oder stimulusbasierten Untersuchungen von Individuen ohne konkrete Interaktion. Das kategorielle Modell der Basisemotionen kann daher im MMI Kontext nicht voll umfänglich übernommen werden. Da die passive Induktion durch Präsentation von International Affective Picture System (IAPS) Bildmaterial mit geringem Aufwand realisierbar ist, basieren viele frühe Arbeiten der maschinellen Mustererkennung für Emotionen (ohne MMI Bezug) auf Datensätzen welche dieses Basisemotionsmodell verwenden.

Weitere Arbeiten von Ekman beschäftigen sich mit der mimischen Expression dieser Basisemotionen wie in Abbildung 2.1 dargestellt [72]. Das Facial Action Coding System (FACS) ordnet jeder Bewegung der mimischen Muskulatur eine Action Unit (AU) zu, welche die einzelnen oder mehrere Muskelbewegungen zusammenfassen. Diese Kodierung ermöglicht es Gesichtsausdrücke symbolisch zu erfassen. Im FACS existieren insgesamt 44 AU (siehe Tabelle 2.1), 12 für das obere Gesicht, 32 für das untere Gesicht. AU im Untergesicht werden hinsichtlich ihrer Richtung der Bewegungen eingeteilt, wodurch sich horizontale, vertikale, schräge, kreisförmige und gemischte Aktionen unterscheiden lassen.

AU	Beschreibung
1	Heben der inneren Augenbrauen
2	Heben der äußeren Augenbrauen
4	Zusammenziehen beider Augenbrauen
5	Heben des oberen Augenlides
6	Zusammenziehen des äußeren Ringmuskels um die Augen
7	Zusammenziehen des inneren Ringmuskels um die Augen
9	Rümpfen der Nase
10	Anheben der oberen Lippe
11	Äußeren Teil der Oberlippe schräg nach oben ziehen
12	Anheben der Mundwinkel
13	Spitzes Anheben der Mundwinkel
14	Einziehen der Mundwinkel
15	Herabziehen der Mundwinkel
16	Unterlippe nach unten ziehen
17	Hinaufschieben des Kinns
18	Kussmund
22	Lippen trichterförmig nach außen wölben
23	Spannen der Lippen
24	Zusammenpressen der Lippen
25	Öffnen der Lippen
26	Öffnen des Mundes durch Entspannung der Unterkiefermuskulatur
27	Weites Öffnen des Mundes
28	Lippen zwischen die Zähne saugen
43	Absenken des oberen Augenlides
45	Blinzeln (beide Augen)
46	Zwinkern (ein Auge)

Tabelle 2.1.: Beschreibung der AU gemäß FACS.

Die Ausprägung einer Bewegung wird im FACS durch fünf Stufen repräsentiert, dabei gilt:

- Stärke A: an der Wahrnehmungsgrenze
- Stärke B: gut sichtbar
- Stärke C: deutlich sichtbar
- Stärke D: ausgeprägt
- Stärke E: im physiologischen Höchstmaß

Nach Ekman existiert eine Abbildung der Gesichtsmuskelaktivität (oder Kombinationen daraus) auf die sieben Basisemotionen. So kann das zunächst emotionsunabhängige FACS auf den Kontext der Emotionserkennung abgebildet werden. Diese Emotionen sind nach Ekman

Basisemotion	Beteiligte AU
Wut	4, 5, 7, 23
Freude	6, 12
Trauer	1, 4, 15
Ekel	9, 15, 16
Verachtung	12, 14
Angst	1, 2, 4, 5, 7, 20, 26
Überraschung	1, 2, 5, 26

Tabelle 2.2.: Basisemotionen nach Ekman und den daran beteiligten AUs.

kulturübergreifend, denn ihre zugehörigen Gesichtsausdrücke sind universal. Tabelle 2.2 stellt die Abbildung dar.

Das FACS beschreibt lediglich den mimischen visuellen Kanal. Ähnliche Transkriptionssysteme für verbale Modalitäten wie Gesprächsanalytisches Transkriptionssystem (GAT) oder Halbinterpretative Arbeitstranskriptionen (HIAT) sind in der Literatur wenig ausgeprägt und nicht etabliert [232, 233, 214]. Kodierungen für Körperpose, biophysiologische Parameter oder weitere Modalitäten existieren in dieser Form überhaupt nicht, wodurch Kodierungssysteme in multimodalem Kontext nur geringfügige Anwendung finden. Lediglich in frühen Verarbeitungsphasen (Merkmalsextraktion) findet FACS im Bereich der *Computervision* Anwendung [140, 49]. In einigen Arbeiten finden sich erfolgsversprechende Klassifikationsmethoden, welche (semi-)automatisch erkannte AU als Eingangsmerkmale verwenden [54, 140]. In den frühen Arbeiten zur maschinellen Emotionserkennung finden sich jedoch zahlreiche Arbeiten, welche anhand von uni- und bimodalen Datensätzen die Emotionskategorien von Ekman klassifizieren [54, 37, 231, 79].

Eine weiterer kategorieller Ansatz wird durch Robert Plutchik beschrieben [179, 180]. Plutchiks Emotionstheorie nimmt an, dass Emotionen eine genetische Grundlage haben, und so unser Verhalten steuern. Ähnlich wie Ekman benennt er ebenfalls Basisemotionen. Er fasst diese allerdings in einem dreidimensionalen Modell zusammen. Die acht Emotionen sind auf einem Kegel angeordnet und sind hinsichtlich ihrer Intensität abstufbar. Zur Spitze hin schwächen sich die Emotionen ab, und sind daher nicht mehr gut unterscheidbar. Im Modell gegenüberliegende Emotionen sind dabei gegensätzliche Emotionen, nebeneinanderliegende Emotionen entsprechend ähnliche Emotionen. Die Spitze kann als neutraler Zustand angenommen werden. Weitere Emotionen sind Kombinationen der Primäremotionen. Abbildung 2.2 zeigt eine Abwicklung des Emotionskegels nach Plutchik.

Da dieser kategorielle Ansatz neben den reinen Kategorien auch Intensitäten der Empfindung beschreiben, ist dieser Ansatz für die maschinelle Emotionserkennung von größerem Interesse. Es existieren zahlreiche Erweiterungen des ursprünglichen Modells, welche auch

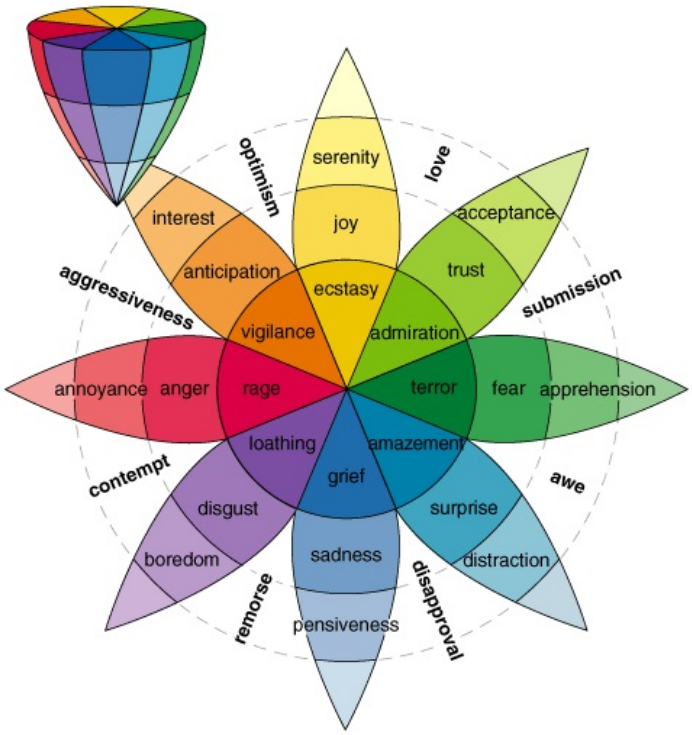

Abbildung 2.2.: Abwicklung des Emotionskegelmodels nach Plutchik. Ähnliche Emotionen sind darin benachbart, die Intensität der Emotionen nimmt nach oben hin zu (Nach innen in der Abwicklung) Bildquelle: `https://www.6seconds.org/2017/04/27/plutchiks-model-of-emotions/` (Stand 3. Juni 2019)

gemischte Emotionen enthalten. Abbildung 2.3 zeigt ein in der Annotation von emotionalen Datensätzen zur maschinellen Emotionserkennung gebräuchliches, erweitertes Modell. Hierbei werden 20 Emotionen (Basis- und zusammengesetzte Emotionen) in jeweils vier Intensitäten unterschieden, als weiteren 21. Zustand ist Neutral ohne Intensität gegeben[1].

Ähnlich dem Modell von Ekman ist auch das Modell von Plutchik nur bedingt zur Gewinnung von Daten im MMI Kontext geeignet, da bei seinem Entwurf eine Interaktion mit Maschinen und die dabei auftretenden Emotionen nicht berücksichtigt wurden. Auch ist eine Unterscheidung von bis zu 21 Klassen nach heutigem Stand der Forschung unrealistisch. Durch die Definition von Intensitäten und einer neutralen Klasse ist es jedoch besser geeignet als das weiter verbreitete Modell von Ekman.

[1] diese Modell wurde im Rahmen dieser Arbeit als Annotationswerkzeug für Audio und Video Daten implementiert (siehe Kapitel 3.4)

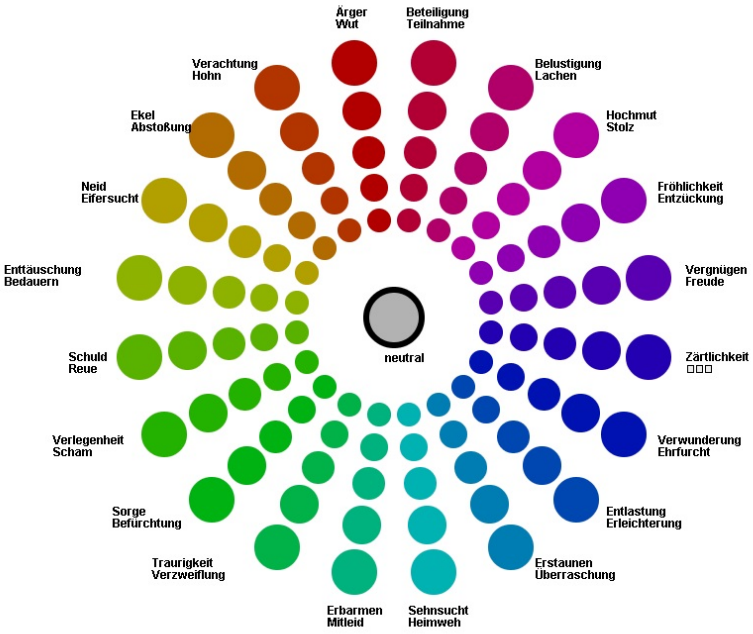

Abbildung 2.3.: Typische Darstellung des Modells nach Plutchik in der Annotation

2.1.2. Dimensionale Emotionsmodelle

Einen alternativen Ansatz verfolgen Mehrabian und Russel bei der Beschreibung eines vollständig kontinuierlichen, dreidimensionalen Valenz-Arousal-Dominance (VAD) Modells [197]. Das Modell beschreibt dabei alle denkbaren emotionalen Zustände eines Menschen anhand der drei Dimensionen *Valenz*, *Erregung* und *Dominanz* [195]. Abbildung 2.4 beschreibt eine Darstellung des Modells nach Russel.

Die Dimension der Valenz beschreibt den Grad an Wohlgefühl eines Menschen. Ärger und Angst sind beispielsweise Gefühle welche durch geringes Wohlgefühl und somit einen geringen Valenzwert beschrieben werden. Freude hingegen entspricht einem hohen Valenzwert.

Erregung repräsentiert den Intensitätsgrad einer Empfindung. Ärger und Freude werden hier durch einen hohen Erregungszustand gekennzeichnet, wärend Langeweile eher durch geringe Erregung charakterisiert ist.

Die Achse der Dominanz beschreibt den Grad der Kontrolle über eine Empfindung. Ärger wird hier beispielsweise durch eine hohe Dominanz gekennzeichnet, während Angst typischerweise durch geringe Kontrolle und somit geringe Dominanz beschrieben wird.

Da alle kategoriellen Modelle auf das dimensionale VAD Modell abgebildet werden können,

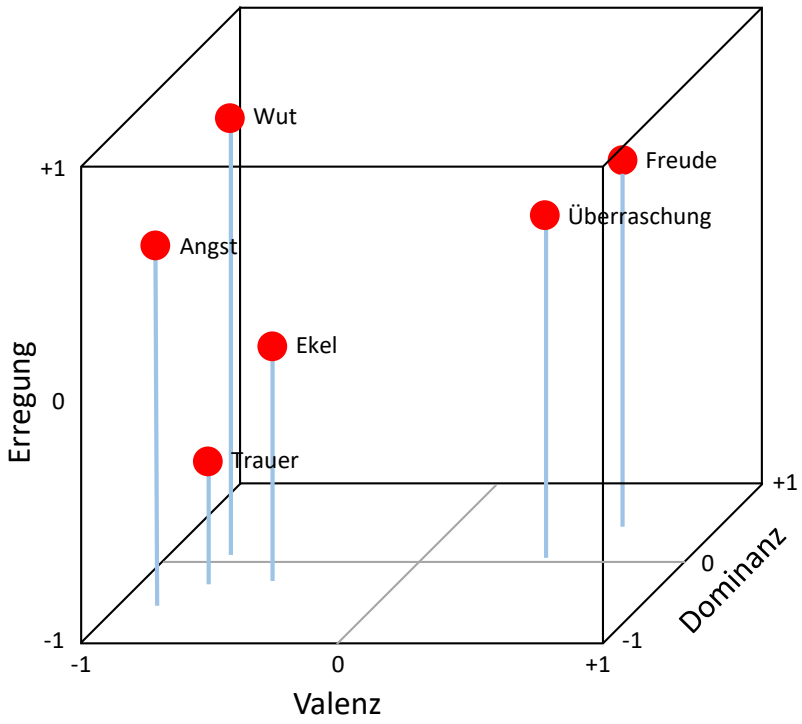

Abbildung 2.4.: Dimensionales kontinuierliches VAD Model nach Russel mit Einordnung der Basisemotionen nach Ekman

ist dieses als mächtiger anzusehen. Allerdings ist die Einordnung eines Zustandes durchaus nicht trivial. Abbildung 2.4 zeigt eine mögliche Anordnung von Basisemotionen von Ekman und Russel. Ist für eine Annotation einer Emotion im Falle der kategoriellen Modelle lediglich die Auswahl einer bestimmten Kategorie notwendig (mit eventueller diskreter Abstufung), so muss im VAD Modell drei mal eine unabhängige Einschätzung eines kontinuierlichen Wertes erfolgen. Auch ist die Bezeichnung der diskreten Kategorien meist umgangssprachlich geläufig, und somit auch von Laien eine (Selbst-)Annotation möglich. Die Bezeichnungen im VAD Modell sind hingegen weniger intuitiv und insbesondere die Einschätzung der jeweiligen Intensität ist problematisch. Um dennoch Einschätzungen von Laien (meist Selbsteinschätzungen von Versuchsteilnehmern) zu ermöglichen kann das Self Assessment Manikin (SAM) eingesetzt werden [30]. Hierbei werden die drei Dimensionen des VAD Modells durch die in Abbildung 2.5 gezeigten standardisierten Piktogramme gekennzeichnet. Die Einschätzung erfolgt dabei jedoch zunächst nur in fünf diskreten Abstufungen, einige Anwendungen (wie der uulmMAC) erweitern diesen Ansatz um Zwischenstufen um eine bessere Auflösung zu erreichen.

Im Bereich der Mustererkennung im MMI Kontext setzt sich das VAD Modell zunehmend durch, da hier auch typische vorkommende Emotionen und affektive Zustände in den meist

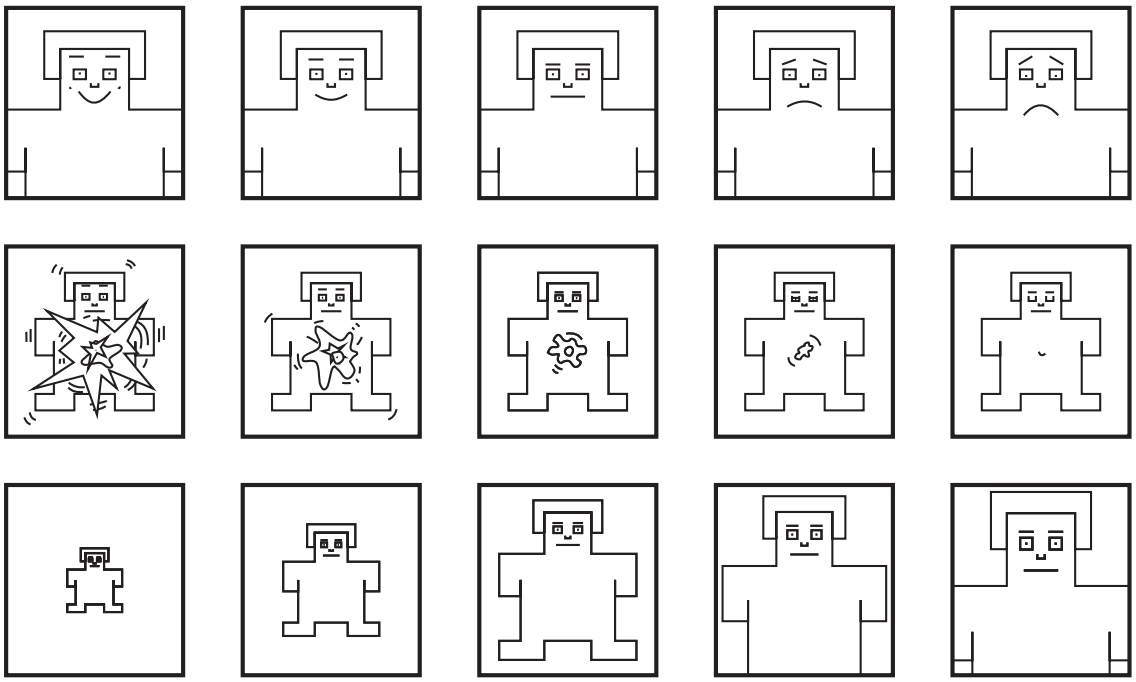

Abbildung 2.5.: Standardisierte Piktogramme des SAM zur Erfassung der drei Dimensionen des VAD Raums. Erste Reihe: Valenz von positiv (links) nach negativ (rechts). Zweite Reihe: Erregung von hoch (links) nach niedrig (rechts). Dritte Reihe: Dominanz von gering (links) nach hoch (rechts).

schwächer ausgeprägten Intensitäten darstellbar sind. Zudem kann dasselbe Modell bei methodisch eng verwandten Ansätzen zur Schätzung emotionaler Zustände außerhalb des MMI Kontextes verwendet werden (z.B. der Prognose von mentalen Störungen wie PTSD [270, 111] oder der Analyse von Mensch-Mensch Interaktionen [135, 172]).

2.1.3. MMI Emotionsmodelle

Der Forschungsbereich der Emotionen im Kontext der MMI ist bisher nur wenig untersucht, daher fehlen tiefgreifende psychologische Grundlagen. Dies liegt im wesentlich an der noch sehr jungen Disziplin der MMI, welche in der Lage ist emotionale Zustände des Menschen zu berücksichtigen und adäquat darauf zu reagieren. Sicher ist jedoch, dass klassische psychologische Emotionsmodelle nicht unverändert anwendbar sind. Emotionen nach Ekman treten typischerweise nicht auf, auch aus dem Modell von Plutchik treten nur wenige gemischte Emotionen auf. Kategorielle Modelle existieren bisher in der Literatur nicht in ausgeprägter Form. Allerdings definieren einige Arbeiten eigene Kategorien wie *positiv* und *negativ* [10], *passiv* und *aktiv* [95] oder *erfolgreicher* und *gestörter Diskursverlauf* [191]. Diese Arbeiten stehen jedoch nur in losen Zusammenhang und die definierten Kategorien sind pragmatisch

empirisch bestimmt. In dieser Arbeit wird in Abschnitt 2.3 ein kategorielles Modell vorgeschlagen, welches den kognitiven Belastungszustand eines Menschen bei der MMI beschreibt. Dieses definiert die Kategorien *kognitive Überforderung*, *neutral* und *kognitive Unterforderung*. Weiter wird eine Definition von *Frustration* (gegenüber einem technischen System) vorgeschlagen. Alle vorgeschlagene Kategorien sind ebenfalls auf die Dimensionen des VAD Modells abgebildet.

In einigen Arbeiten, welche nicht unmittelbar in der Mustererkennung Anwendung finden, finden sich einige Begriffsdefinitionen, welche nicht als strikte Emotionskategorien definiert werden, aber dennoch zum emotionalen Zustand gegenüber einem System beitragen können. Diese sind zum Beispiel, *Vertrauen* (gegenüber dem System), *Akzeptanz* (des Systems), *Zuverlässigkeit* (des Systems) oder *Verständnis* (des Systemverhaltens) [19, 219, 116].

Neben den veränderten Kategorien ist zu beobachten, dass die Intensität auftretender emotionaler Empfindungen meist schwach ausgeprägt ist. Dies wird gestützt durch die Analyse der Annotation verschiedener Korpora, welche im MMI Kontext erstellt wurden [126]. Hier ist zu beobachten, dass die maximalen Werte der jeweiligen Emotionen bei der Annotation nur selten auftreten. Kontinuierliche oder dimensionale Modelle können dies abbilden. Im Falle diskreter Modelle würde ein Großteil der relevanten Emotionen nicht abgebildet, da ihre Ausprägung unter den typischen Schwellwerten liegt. Es ist weiter zu beobachten, dass ein großer Teil einer Interaktion keine Emotionen hervorruft, sich der Nutzer also im wesentlichen in einer neutralen Stimmung bezogen auf die Interaktion befindet. So wurde in einer Vorstudie des uulmMAC und dem uulmMAC selber gezeigt, dass das Verhältnis von Emotionen zu Neutralität bei ca. 1:14 liegt (siehe Tabelle 3.3) und diese schwach expressiv sind. MMI Emotionsmodelle müssen daher zwingend kontinuierlichen Charakter haben und eine neutrale Klasse aufweisen.

Neben der Definition anwendungsabhäniger Kategorien ist der am meisten etablierte Ansatz zur Beschreibung emotionaler Zustände im MMI Kontext eine geringfügig angepasste Interpretation des VAD Modells. Hierbei werden die Dimensionen übernommen, diese befinden sich allerdings immer Beziehung zwischen dem System und dem Nutzer. Speziell die Dominanz bezieht sich darauf, ob der Nutzer oder das System Kontrolle über die Interaktion besitzt. Hat der Nutzer das Gefühl Herr der Lage zu sein, hat er die dominante Stellung. Ist der Nutzer gezwungen passiv auf das Systemverhalten zu reagieren, ohne das Gefühl zu haben den Verlauf kontrollieren zu können, hat das System die Kontrolle. Die Dimensionen der Erregtheit und der Valenz können weitgehend intuitiv übernommen werden. Sie beziehen sich aber immer auf den durch das Systemverhalten und den Inhalt der Interaktion ausgelösten Zustand des Nutzers. Eventuell überlagerte Zustände, welche außerhalb des Einflussbereichs eines Systems liegen, können zusätzlich durch klassische Modelle beschrieben werden. In der Literatur wurde ein solcher Versuch bisher jedoch noch nicht beschrieben. Auch der in dieser Arbeit erhobene Kor-

pus verwendet im wesentlichen das angepasste dimensionale VAD Modell. Ebenfalls beziehen sich die meisten der präsentierten Resultate, neben den definierten Kategorien der Über- und Unterforderung, auf dieses Modell.

2.2. Stand der Forschung - Datensätze

Datensätze bilden die Arbeitsgrundlage für die Erforschung von Methoden zur Erkennung emotionaler Muster. Erste Arbeiten hierzu gehen dabei auf den Anfang des Jahrtausends zurück. Tabelle 2.3 zeigt eine Übersicht bekannter Datensätze aus dem Zeitraum der letzten 17 Jahren.

Cohn und Kanade haben im Jahr 2000 einen Datensatz vorgestellt welcher Graustufen Videomaterial beinhaltet[121]. Ziel des Datensatzes ist die Bereitstellung von Daten zur Entwicklung und Evaluation automatischer AU Erkenner. Der Datensatz beinhaltet kurze (wenige Sekunden lange) Videosequenzen in denen Schauspieler verschiedenste Kombinationen von AUs stark expressiv darstellen. Eine Abbildung auf Emotionen ist nur indirekt durch die Zuordnung von AUs zu Basisemotionen nach Ekman gegeben. Der Datensatz beinhaltet ca. eine Stunde Datenmaterial verteilt auf 97 Personen und ist vollständig FACS codiert. Ein vergleichbarer Datensatz für Basisemotionen nach Ekman in Form akustischer Daten wurde 2005 von Sendlmeier mit der Berlin Database of Emotional Speech (EmoDB) vorgestellt [36]. Die Daten wurden ebenfalls durch Schauspieler expressiv eingesprochen und sind von eher geringem Umfang verglichen mit heutigen Datensätzen. Beide Datensätze wurden verwendet um erste Methoden der unimodalen Mustererkennung zu evaluieren. Eine Anwendung in realen MMI Anwendungen war zu diesem Zeitpunkt eine ferne Vision.

Der Airplane Behaviour Corpus (ABC), 2007 vorgestellt von Schuller et. al., beschreibt einen ersten Ansatz bei dem Audio- und Videomodalitäten gemeinsam erhoben wurden [222]. MMI relevante Emotionskategorien waren zu diesem Zeitpunkt noch nicht von Interesse. Die wesentliche Neuerung bestand darin, die Emotionen nicht durch Schauspieler darstellen zu lassen, sondern Probanden entsprechendes Stimulusmaterial zu präsentieren um so Emotionen zu induzieren. Der Induktionserfolg wurde anschließend durch Wahrnehmungstests überprüft. Dabei wurden unabhängigen Personen die aufgenommen Daten präsentiert und diese sollten die gezeigte Emotion klassifizieren. Wurde die Emotion von einer Mehrheit erkannt, wurde die Induktion als erfolgreich angesehen. Da eine Unterscheidung aller Basisemotionen kaum möglich ist, wurden die verwendeten Kategorien auf die vier Kategorien *angst*, *neutral*, *positiv* oder *negativ* beschränkt.

Da die Erstellung Probanden gestützter Korpora enorm aufwendig ist haben Grimm et. al.

2008 mit dem Vera am Mittag (VaM) Corpus eine gänzlich alternative Herangehensweise vorgeschlagen [94]. Ihr Ansatz schlägt vor TV-Aufzeichnungen als Datenmaterial heran zu ziehen. Der Aufwand reduziert sich dadurch enorm, andererseits besteht keine Kontrolle mehr über technische Eigenschaften wie Beleuchtung, Bildausschnitt, Audiopegel oder ähnliches. Auch ist unklar, ob das gewonnen Datenmaterial schauspielerischen oder natürlichen Ursprungs ist. Eine Erweiterung auf zusätzliche Modalitäten ist nicht möglich. Der wesentliche Aufwand bei der Erstellung dieser Korpora besteht in der notwendigen Annotation. Um eine Annotation leichter zu gestalten wurden in diesem Korpus expressive Sequenzen einer damals beliebten deutschen Talk Show „*Vera am Mittag*“ausgewählt. Die meist kurzen Videosequenzen wurden durch Studenten und Mitarbeiter der Arbeitsgruppe nach Ekman annotiert.

Einen ähnlichen Ansatz verfolgen Göcke et al. bei der Erstellung der Benchmark Datensätze der EmotiW Challenges 2013 bis 2017 [67, 66, 68, 65, 64]. Die Datensätze basieren auf Filmausschnitten bekannter Kinofilme. Auch hier wurden expressive Sequenzen nach Ekman annotiert. Die Auswahl der relevanten Passagen wurde durch den Handlungskontext des Filmes und den Inhalt der transkribierten Untertitelspur gegeben. Dem Annoteur bleibt lediglich die Aufgabe, die ausgewählten und kontextannotierten Sequenzen hinsichtlich Korrektheit zu prüfen. Kinofilme beinhalten häufig Hintergrundgeräusche und Musik, sowie Verdeckungen, schlechte Beleuchtungsverhältnisse oder kritische Bildausschnitte. Die Daten sind daher als sehr herausfordernd einzustufen, es ist jedoch fraglich ob diese Herausforderungen auch in realistischen MMI Sensorsystemen so auftreten [126]. Zudem ist der schauspielerische Ursprung und die begrenzte Modalitätsanzahl problematisch.

Eine weitere bekannte Datensatzfamilie bilden die Audi Visual Emotion recognition Challenge (AVEC) Zusammenstellungen [223, 224, 263, 262, 189, 261]. Diese verwenden seit ihrer ersten Vorstellung im Jahr 2011 das VAD Modell als Annotationsschema. Der erste AVEC Datensatz war dabei noch auf eine diskrete Variante beschränkt, wurde aber bereits im Folgejahr durch kontinuierliche, dimensionale Varianten erweitert. Die Annotation erfolgte dabei durch mindestens zwei Experten, wobei das verwendete Annotationswerkzeug mittels Schiebereglern für jede Dimension, wie in [127, 125] berichtet, einige Schwächen aufweist, welche das Annotationsergebnis negativ beeinflussen. Schuller et. al. verwenden ein möglichst natürliches Induktionsparadigma. In allen Datensätzen berichten die Probanden in einer Mensch-Mensch Interaktion oder Mensch-Avatar Interaktion über eigene Erlebnisse oder lesen emotional geprägte Geschichten vor. Kritisch zu betrachten ist der gänzliche Verzicht auf MMI Komponenten, allerdings steht die Entwicklung von MMI Technologien auch nicht im Fokus der Challenge. Die AVEC Datensätze wurden kontinuierlich weiter entwickelt, wodurch in den neuen Versionen zum ersten mal neben Audio und Video auch Biophysiologie als Modalität bereitgestellt werden. Auch die Gesamtmenge des Materials erreicht mit 20 Stunden eine

annehmbare Größe. Neben dem Einsatz in der Emotionserkennung finden diese Datensätze auch Beachtung in der Erkennung psychologischer Störungen wie der Depressionserkennung, für die im Jahr 2014 eine eigen Subchallenge eingeführt wurde.

Einen ersten Ansatz unter Verwendung von MMI Komponenten bei der Induktion schlagen Rösner et. al. im Last Minute Corpus (LMC) und Walter et. al. im Emotion Recognition Corpus (EmoRec) Korpus vor [191, 194]. Beide verwenden Wizzard of Oz (WoZ) Elemente, welche einem Probanden ein vollständig funktionierendes System simulieren, welches in Wahrheit vom Versuchsleiter gesteuert wird. Im LMC bekommt ein Proband die Aufgabe mittels einem sprachgesteuerten System einen Koffer für eine soeben gewonnene Reise zu packen. Nach der Hälfte des Experimentes erfährt der Proband, dass die Reise, nicht wie zunächst angenommen, auf eine Südseeinsel geht, sondern in eine kalte Region. Er muss daraufhin unter Zeitdruck seinen Koffer umpacken. Induzierende Elemente sind hier nicht in der Interaktion sondern in der gestellten Aufgabe zu finden. Auf eine Annotation nach klassischen Modellen wurde zugunsten der Definition von vier kritischen Zeitpunkten im Dialog, sogenannten Diskursbarrieren, verzichtet. Der Korpus beinhaltet multiple akustische und visuelle Sensoren, sowie physiologische Sensoren, welche allerdings qualitative erhebliche Probleme aufweisen. Der EmoRec II Korpus besteht ebenfalls aus einem WoZ basierten System, hier bekommt der Nutzer die Aufgabe Memory zu spielen. Der Spielverlauf wird neben variabler Komplexität der Aufgabe, durch akustisches ermutigendes oder entmutigendes Feedback durch den Versuchsleiter beeinflusst. Die Annotation erfolgt aufgrund des experimentellen Ablaufs. Es wird also angenommen, dass sich der Nutzer zu bestimmten Zeiten an diskreten Punkten des VAD Modells befindet. Beide Korpora sind bezüglich der Anzahl der Modalitäten und ihrem Umfang an aufgezeichneten Material zum Zeitpunkt ihrer Veröffentlichung einzigartig.

Die Entwicklung der vorgestellten Datensätzen ging von kleinen unimodalen überexpressiven Datensätzen hin zu umfangreichen multimodalen natürlicheren Datensätzen. Erst die Datensätze EmoRec, LMC und AVEC 2016 stellen eine minimal ausreichende Anzahl von Modalitäten zur Entwicklung multimodaler Klassifikatoren bereit. Der AVEC Datensatz verzichtet gänzlich auf MMI basierte Interaktionselemente. Es ist daher zu befürchten, dass Klassifikatoren, welche auf diesen Daten trainiert wurden, in einem MMI Szenario nicht einsetzbar sind. Der LMC Korpus ist nicht hinsichtlich einem gängigen Emotionsmodell oder MMI typischen Kategorien annotiert. Das verwendet Induktionsparadigma unterscheidet sich zudem von MMI Fragestellungen. Der LMC Korpus verwendet weder ein dimensionales, kontinuierliches Emotionsmodell, noch ein MMI typisches kategorielles Modell, auch wurde der Induktionserfolg nicht überprüft. Es ist davon auszugehen, dass ein Transfer auf Basis EmoRec, LMC oder AVEC erlernter Klassifikatoren in realen Anwendungen nicht erfolgreich ist. Diese Annahme hat sich auch bei den ersten Entwürfen der in Kapitel 5 vorgestellten Demonstratoren bestätigt. Dort wurden

Datensatz	Jahr	Ind.paradigma	Emotionsmodell	Annotation	Modalitäten	Personenzahl	Länge
CK [121]	2000	Schauspiel	AU	FACS	1 x Video	97	1h
EmoDB [36]	2005	Schauspiel	Kategoriell (nach Ekman)	Wahrnehmungstest	1 x Audio	10	1,5h
ABC [222]	2007	Passiv induziert	Kategoriell (Angst, Neutral, Positiv, Negativ)	Wahrnehmungstest	1 x Audio 1 x Video	8	11,5h
VAM [94]	2008	TV Show	VAD (nach Ekman + Neutral)	Experten	1 x Audio 1 x Video	104	12h
AVEC [223]	2011	Natürlich	Kategoriell (Valenz, Erregung, Erwartung, Kraft)	2 Experten	1 x Audio 1 x Video	20	12,5h
EmotiW [67]	2013	Kino Film	Kategoriell (nach Ekman + Neutral)	Film Kontext	1 x Audio 1 x Video	330	2h
LMC [191]	2012	WoZ	Diskursbarrieren	Protokoll	1 x Audio 4 x Video 5 x Physiologie	130	17h
EmoRec II [194]	2013	WoZ	Diskretes VAD	Versuchsablauf	1 x Audio 4 x Video 6 x Physiologie	50	30h
AVEC [262]	2014	Natürlich	VAD	Wahrnehmungstest	1 x Audio 1 x Video	84	20h
AVEC [261]	2016	Natürlich	VA Depression	Wahrnehmungstest	1 x Audio 4 x Video 2 x Physiologie	84	20h
EmotiW [65]	2016	Kino Film	Kategoriell (Ekman + Neutral)	Film Kontext	1 x Audio 1 x Video	465	3h
uulmMAC	2017	MMI	VAD und Kategoriell (Neutral, Frustration,) Über-, Unterforderung)	Protokoll und SAM, USF	3 x Audio 3 x Video 6 x Physiologie 1 x Kinect	40 + 20x3	75h

Tabelle 2.3.: Übersicht bestehender emotionaler Datensätze.

Klassifikatoren auf Basis des LMC und EmoRec trainiert, welche im praktischen Test nicht erfolgreich eingesetzt werden konnten. Kurze probandenspezifische Datensätze, welche im Anwendungkontext erhoben wurden, konnten jedoch unter Verwendung identischer Methoden erfolgreich eingesetzt werden.

Da keiner der verfügbaren Datensätze die erforderlichen Eigenschaften aufweist, wurde im Rahmen dieser Arbeit der uulmMAC Datensatz erhoben, welcher in ausreichender Modalitätenzahl und Umfang in einem entworfenen MMI Kontext erstellt wurde und sowohl durch ein eigens entwickeltes MMI Emotionsmodell, sowie durch das anerkannte VAD Modell annotiert ist. Der im folgenden Abschnitt detailliert beschriebene uulmMAC ist geeignet um Klassifikatoren in MMI Kontext für reale Anwendungen zu entwerfen.

2.3. Ulm University Multimodal Affective Corpus - uulmMAC

Die Zielsetzung der Erhebung des uulmMAC ist die Gewinnung von emotional gefärbten Sensor- und Nutzerdaten im Kontext der MMI. Die im Korpus erfassten Emotionen werden nicht durch Schauspieler dargestellt oder durch passive Induktion (Präsentation von Stimuli) erzeugt. Die emotionale Induktion erfolgt durch Veränderung einer MMI Schnittstelle während der Nutzerinteraktion. Hierdurch werden emotionale Reaktionen erzeugt, welche typischen Nutzungszenarien entsprechen. Auch ist der Proband durch Einsatz von mobiler Sensorik, nicht wie in anderen Studien, fest an einen Ort gebunden oder durch Anbringung der Sensorik erheblich in seiner Bewegungsfreiheit eingeschränkt. Aufgrund des hohen Freiheitsgrades der Studie wird die Induktion auf vier emotionale Zustände beschränkt, wobei im speziellen eine Veränderung der kognitiven Belastung, hin zu Über- und Unterforderung, des Probanden im Vordergrund steht. Der Induktionserfolg wird durch Selbstevaluation der Probanden im kontinuierlichen VAD Raum überprüft, und eine Einordnung der induzierten Kategorien in diesen ermöglicht. Um ausreichendes Trainingsmaterial zu gewinnen und signifikante Resultate sicherstellen zu können, wurde der Umfang der Studie auf 100 Aufnahmen festgelegt.

2.3.1. Experimenteller Aufbau

Zielsetzung der Emotionserkennung im Feld der Companion-Technologie ist die zuverlässige Erfassung emotionaler Nutzerzustände im Kontext der MMI. Daher ist es sinnvoll bereits bei der Erzeugung von Trainingsdaten ein Szenario zu wählen, welches auf eine MMI aufbaut.

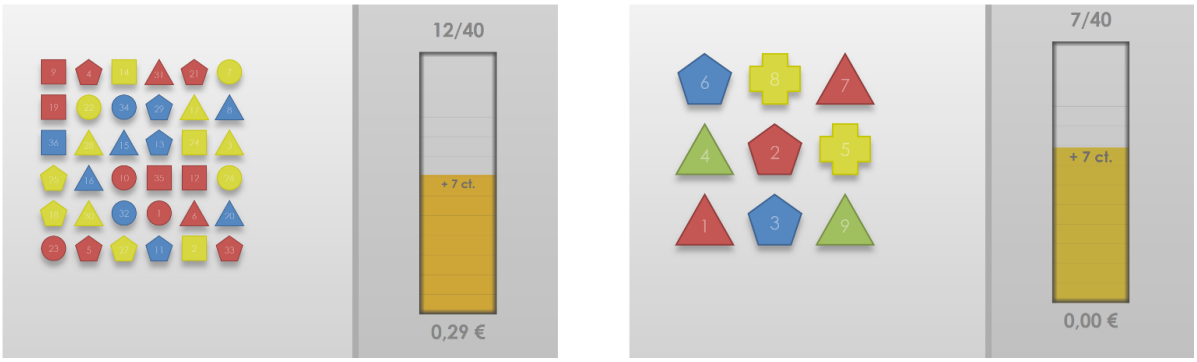

Abbildung 2.6.: Screenshot der uulmMAC Nutzerschnittstelle. Links: Höchste Schwierigkeit $6x6$ Matrix, 3 Farben und 6 Sekunden Countdown (Zielelement 36). Rechts: Geringe Schwierigkeit mit $3x3$ Matrix, 4 Farben und 100 Sekunden Countdown (Zielelement 2).

Schüssel et al. haben ein abstraktes Szenario entworfen, welche typische Elemente einer MMI beinhaltet, ohne dabei eine konkrete Anwendung umzusetzen [218]. Die vorgeschlagene Elemente beinhalten typische Interaktionsmuster, wie das Suchen eines Menüpunktes oder Icons mit anschließender Interaktion. In abstrakter Form wird dies durch das Suchen, Auffinden und Benennen eines Zielelementes realisiert. Das Zielelement ist in einer Matrix aus farbigen geometrischen Elementen in Form und Farbe einmalig, während die Distraktoren in identischer Ausprägung mindestens doppelt enthalten sind. Abbildung 2.6 zeigt zwei solche Matrizen mit den Zielelementen 36 und 2. Schüssel et. al. beschränken sich dabei nicht auf eine Interaktionsmodalität, sondern lassen es dem Nutzer freigestellt mit dem aufgefunden Element mittels Sprach-, Touch- oder Mauseingabe zu interagieren. Durch den gameifizierten Charakter der vorgeschlagen Interaktion wird die Aufmerksamkeit und Motivation des Probanden über einen längeren Zeitraum gesichert. Schüssel et. al. haben diese Szenario zur Untersuchung des Eingabeverhaltens von Nutzern herangezogen [220]. Emotionale Fragestellungen wurden dabei nicht berücksichtigt.

Im Rahmen der Vorarbeiten zum uulmMAC konnte durch Meudt und Thiam in Vorstudien gezeigt werden, dass in der Studie von Schüssel typische, zu erwartende, emotionale Zustände enthalten sind, und diese zudem auf einem realistischen Expresionsniveau liegen [253]. Diese unterscheiden sich dabei erheblich von Emotionen in geschauspielerten oder passiv induzierten Korpora. Basierend auf der Vorstudie konnte gezeigt werden, dass die Gewinnung von Trainingsdaten basierend auf abstrakten Nutzer Interaktionen vielversprechend ist.

Die Anwendung von Schüssel et. al. wurde im Rahmen der Erhebung der uulmMAC Daten dahingehend abgewandelt, dass die Eingabemodalität auf Spracheingabe beschränkt wurde um sicherzustellen, dass in der akustischen Modalität ausreichend viel Datenmaterial erzeugt wird. Weiter wurde die Komplexität des Spiels variabler gestaltet um mittels Anzahl von

Abbildung 2.7.: Screenshot der Nutzerschnittstelle zur Selbsteinschätzung durch den Probanden. Links: umgangssprachliche Fragen (USF), Rechts SAM

Farben und Formen der Interaktionselemente eine dynamische Schwierigkeit ausprägen zu können. Als weiteres beschränkendes Element wurde die maximale Zeit (Count Down) für eine Nutzerantwort variabel gestaltet. Eine weitere Vorstudie hat gezeigt, dass diese Anpassungen optimal zur Induktion emotionaler und kognitiver Zustände in der MMI geeignet sind. Auch lässt sich der modulare Aufbau der Anwendungskomponenten ideal mit dem in dieser Arbeit vorgeschlagenen MAR^{2}S System kombinieren. Abbildung 2.6 zeigt eine typische Ansicht der Benutzer Schnittstelle (engl. User Interface) (UI) für eine leichte und schwere Interaktion.

Als Ziel der Studie wurde die Induktion der emotionalen Zustände *Überforderung*, *Unterforderung*, *Frustration* und *Neutral* gewählt. Eine überfordernde Situation wird induziert durch eine komplexe Aufgabe, welche in sehr kurzer Zeit gelöst werden muss. Unterforderung wird durch wiederholtes Lösen einfacher Aufgaben mit unwesentlicher Zeitbegrenzung erreicht. Da die Vergütung des Probanden erfolgsabhängig ist kann verhindert werden, dass durch die Unterforderung eine desinteressierte und langweilende Emotion überlagert wird. Zur Induktion von Frustration dienen Elemente welche System-Fehlverhalten simulieren (Erfassung falscher Eingaben oder systemseitige Reaktionsverzögerungen). Eine neutrale Situation wird geschaffen durch Stellen von Aufgaben mit angemessener Schwierigkeit und Zeitbeschränkung. Um bei den Probanden eine ausreichende Motivation sicherzustellen hängt ein Teil der Probandenvergütung vom erreichten Spielerfolg ab. Obwohl die erfolgsabhängige Vergütung sich lediglich im Cent Bereich bewegt und den Gesamtertrag nur unwesentlich beeinflusst, reicht dies aus um ein vorzeitiges innerliches Abbrechen der Studie (mentale Aufgabe) durch die Probanden zu verhindern.

Der Aufbau des Experimentes ist in sechs Phasen unterteilt. Die erste Phase soll beim Probanden Interesse wecken und dient dem kennenlernen der Situation. Die Spielkomplexität ist dabei moderat gehalten. Die hier gewonnen Daten sind nur bedingt zum Klassifikatortraining vorgesehen, da hier häufig noch emotionale Zustände überlagert sind, welche nicht in der Interaktion sondern der ungewohnten und neuen Gesamtsituation geschuldet sind. In der zweiten Phase

wird die Schwierigkeit unmittelbar und unerwartet gesteigert, der Nutzer gerät hierdurch in eine fast unlösbare und stark überfordernde Situation. In der dritten Phase wird die Anwendung auf ein moderates und lösbares Komplexitätsniveau gesenkt wodurch eine weitgehend neutraler mentaler Zustand erreicht wird. In der vierten Phase wird die Komplexität weiter gesenkt. Es ist hier aufgrund des zuvor erlebten und der noch nicht lange andauernden leichten Situation nicht zwingend damit zu rechnen, dass hier bereits eine Unterforderung induziert wird. Erst wiederholtes Interagieren auf geringem Schwierigkeitsniveau in Phase fünf induziert sicher eine Unterforderung. Zu Analysezwecken wird daher nur die fünfte Phase als Unterforderung bezeichnet, wenngleich eine Unterforderung am Ende der vierten Phase nicht auszuschließen ist. Die abschließende sechste Phase erzeugt durch eingestreutes fehlerhaftes oder verzögertes Systemverhalten eine frustrierende Situation beim Probanden, da ihm unverschuldet Gewinne verloren gehen. Jede der sechs Phasen umfasst dabei 40 als *Trails* bezeichnete Einzelspiele. Die experimentellen Phasen sind jeweils durch Selbstevaluationen der Probanden und eine kurze Ruhephase zur Beruhigung zeitverzögerter, biophysiologischer Parameter unterbrochen. Zur Validierung des Induktionserfolges wird das standardisierte SAM verwendet. Hierbei beschreibt der Proband seinen emotionalen Zustand anhand eine piktographischen Skala von 1 bis 9 in den Dimensionen Valenz, Arousal und Dominanz. Anschließen werden dem Nutzer vier Fragen zu seiner Wahrnehmung der letzten Sequenz gestellt, welche dazu dienen die SAM Antworten zu verifizieren. Abbildung 2.7 zeigt die Eingabemasken der SAM und USF. Die Fragen lauten:

- **Schwierigkeit:** „Empfanden Sie die letzte Aufgabe als einfach (1) oder schwer (10)“
- **Interesse:** „Wie schätzen Sie ihr Interesse von gering (1) bis sehr groß (10)“
- **Stress:** „Wie fühlten Sie sich von entspannt/zufrieden (1) bis gestresst/frustriert(10)“
- **Motivation:** „Wie motiviert waren Sie die Aufgabe zu lösen, von garnicht (1) bis sehr (10)“

Abschließend wird dem Probanden die Gelegenheit gegeben sich in Form von verbalem Freitext zu seinem Befinden zu äußern. Dies dient der Kontrolle durch den Experimentalleiter und der Erfassung von eventuell relevantem Feedback. Ist die Selbstbewertung abgeschlossen folgt eine Atemübung bevor die nächste Phase beginnt. Abbildung 2.8 zeigt den experimentellen sequenziellen Ablauf der Studie, Tabelle 2.4 beschreibt die Schwierigkeitseinstellungen der UI. Die Gesamtdauer des Experimentes wird so gewählt, dass eine Stunde nicht überschritten wird, wobei im Mittel mit einer Dauer von ca. 45 Minuten für den Durchlauf aller sechs Phasen und Selbstevaluationen zu rechnen ist.

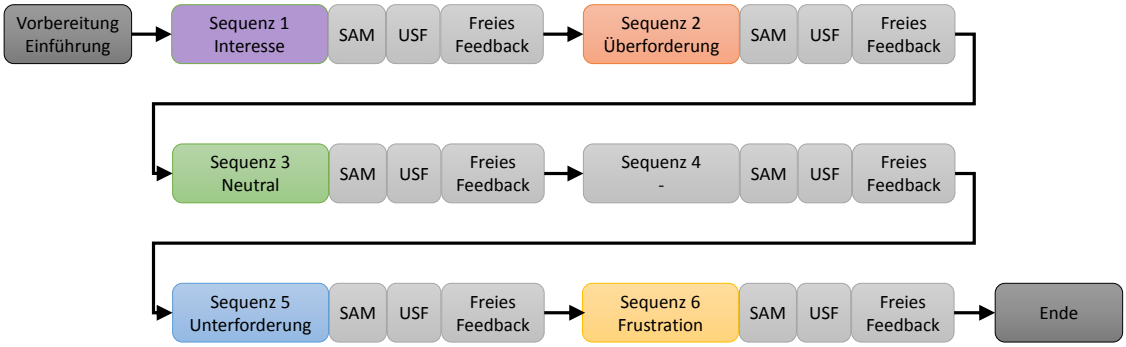

Abbildung 2.8.: Sequenzieller Ablauf des uulmMAC Experimentes. Nach jeder der sechs Experimentalsequenzen erfolgt eine Bewertung des Nutzerzustandes mittels SAM, USF und freiem verbalen Feedback. Sequenz 4 ist keiner Emotion zugeordnet, sie dient dem kontrollierten Übergang in die Unterforderung, da diese erst nach einiger Spielzeit auftritt.

Sequenz Nr.	Induzierte Emotion	Anz. Trials	Matrix Größe	Max. Antwortzeit
1	Interesse	40	3x3, 4x4	10 Sekunden
2	Überforderung	40	6x6	6 Sekunden
3	Neutral	40	4x4	10 Sekunden
4	-	40	3x3	100 Sekunden
5	Unterforderung	40	3x3	100 Sekunden
6	Frustration	40	3x3, 4x4	10 Sekunden

Tabelle 2.4.: Eigenschaften der uulmMAC Experimentalsequenzen

2.3.2. Quantität und Qualität der erhobenen Daten

Der erhoben Datensatz umfasst 100 Aufnahmen von 60 Probanden. Die Aufnahmen sind dabei in zwei Gruppen zu unterscheiden. Eine Gruppe umfasst 40 Einzel-Aufnamen von 40 Personen, dies wird im folgenden als *40Singles* bezeichnet. Bei dieser Gruppe wurde zugunsten der visuellen Auswertung der Gesichtsmerkmale auf die Erfassung der Elektromyographie (EMG) Signale des *Musculus Zygomaticus* (*Jochbeinmuskel*) und *Musculus Corrugator Supercilii* (*Stirnrunzler*) am Auge verzichtet. Eine weiter Gruppe umfasst 60 Aufnahmen von 20 Probanden, diese wird im folgenden als *20trippels* bezeichnet. Hierbei haben die Probanden das Experiment jeweils drei mal mit einem zeitlichen Abstand von ca. einer Woche durchlaufen. Die 20trippels Aufnahmen dienen der Untersuchung zeitlicher Stabilität emotionaler Reaktionen, speziell der Biophysiologie[2]. Daher wurde hier auch die Muskelaktivität des Musculus Zygomaticus und Musculus Corrugator Superciliir via dermaler Klebeelektroden erfasst. Aufgrund der minimalen Unterschiede der beiden Gruppen kann eine Kombination der Daten abhängig von der zu untersuchenden Fragestellung häufig sinnvoll sein.

[2]Untersuchungen der temporalen Stabilität sind nicht Gegenstand dieser Arbeit

Als Probanden wurden eine Kohorte bestehend aus jungen Studierenden gewählt. Voruntersuchungen haben gezeigt, dass Studierende aus dem Fachbereich der Informatik und Elektrotechnik zumeist gut über aktuelle technische Möglichkeiten informiert sind und daher WoZ Experimente zumeist schnell durchschauen. Da in diesen Fällen mit verändertem Verhalten zu rechnen ist, wurden Studierende dieser Fachbereich von der Datenerhebung ausgeschlossen. Der Altersdurchschnitt aller Probanden beträgt 21,7 Jahre mit einer Varianz von 2,7 Jahren. Die Geschlechter wurden in beiden Gruppen jeweils gleichmäßig auf männliche und weibliche Probanden aufgeteilt. Neben der Erhebung sensorischer Daten umfasst der Datensatz auch die Persönlichkeitsfragebogen *Emotion Regulation Questionnaire* (ERQ) [85], *Trait Emotional Intelligence Questionnaire Short Form* (TEIQue-SF) [50] und *Ten Item Personality Measure* (TIPI) [186] zu allen Probanden.

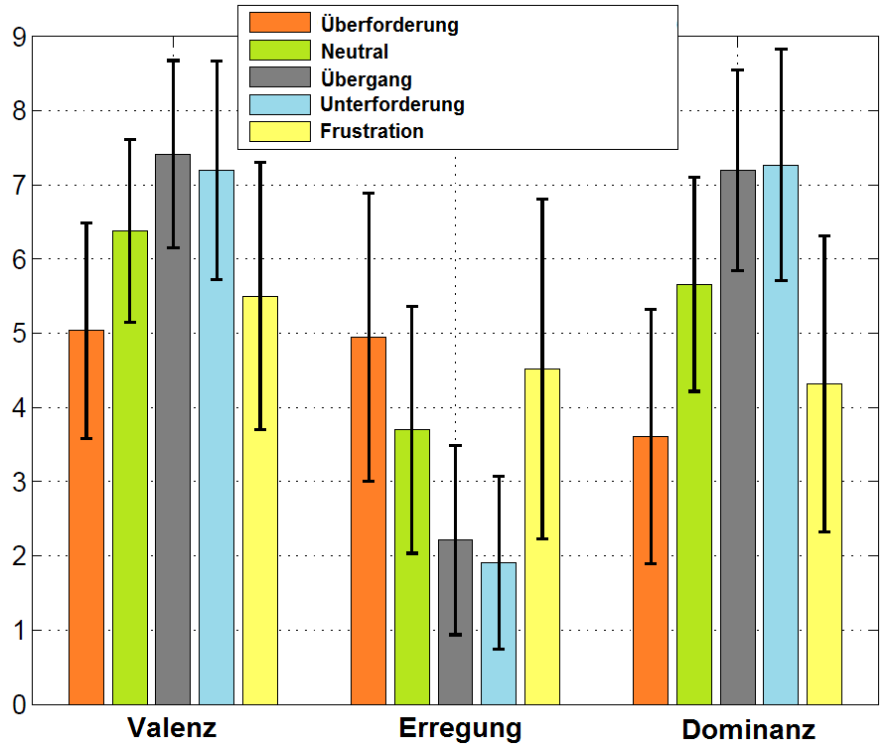

Abbildung 2.9.: SAM Einschätzungen der uulmMAC Experimentalsequenzen und Einordnung im VAD Raum. Deutlich erkennbar sind die Unterschiede der Extrempunkte Überforderung und Unterforderung.

Wie zuvor beschrieben wurde nach Durchlauf einer Experimentalphase unverzüglich eine Selbsteinschätzung des Probanden zu seinem emotionalen Zustand abgefragt. Die Analyse der SAM Antworten zeigt, dass die Induktion der mentalen Zustände den Erwartungen entspricht und von einem Induktionserfolg ausgegangen werden kann. Abbildung 2.9 zeigt die Intensitätswerte der VAD Selbsteinschätzungen. Wie zu erwarten ist, wird die Überforderungsphase durch hohe Erregung (engl. Arousal) und niedrige Valenz und Dominanz charakterisiert. Die

Unterforderung ist entsprechend gegensätzlich angeordnet. Insbesondere besteht bei allen Probanden ein signifikanter Abstand zwischen den Einschätzungen der entsprechenden Phasen. Ein Wilcoxon-Signed-Rank Test mit einen Signifikanzlevel von 5% zeigt bezüglich der Hypothesen

H_1: Differenz von Paaren folgen keiner symmetrischen Verteilung um Null
H_0: Differenz von Paaren folgen einer symmetrischen Verteilung um Null

für die Analyse von Valenz und Erregung ein Signifikanzlevel von $p \approx 7 \cdot 10^{-7}$, H_1 ist somit bestätigt. Für Dominanz beträgt das Signifikanzlevel $p \approx 7 \cdot 10^{-8}$, H_1 ist somit ebenfalls anzunehmen.

Ein weiterer Wilcoxon-Signed-Rank-Test untersucht die Richtung der Einschätzungsunterschiede. Dieser Test bestätigte die erwarteten Richtungen. Valenz ist im Falle von Unterforderung höher, Erregung ist im Falle Überforderung höher und Dominanz im Falle Unterforderung höher. Entsprechend umgekehrte Richtungen zeigen keine Signifikanz. Ausgehend von diesen Resultaten wird folgende Abbildung mentaler Über- und Unterforderung in die Oktanten des kontinuierlichen VAD Raums als gesichert angesehen:

Überforderung: (V- , A+ , D-)
Unterforderung: (V+ , A- , D+)

Um eventuellen Missverständnissen bei der Erklärung der SAM Piktogrammen und gezielten Falschangaben der Probanden vorzubeugen wurden zusätzliche USF durch die Probanden beantwortet. Diese, die Bereiche *Motivation*, *Stress*, *Performanz* und *Schwierigkeit* umfassenden Fragen, sind eng mit den SAM Einschätzungen verknüpft. Es ist beispielsweise für die Überforderungsphase ein starker Zusammenhang zwischen Erregung und dem persönlichen Stressempfinden zu erwarten. Für den Fall der Unterforderung sollte sich hier ebenfalls eine entsprechende Korrelation zeigen. Tabelle 2.5 und 2.6 gibt einen Überblick über die Korrelationen zwischen SAM und umgangssprachlicher Einschätzung der Probanden. Es ist festzustellen, dass alle gefunden signifikanten Korrelationen plausibel erklärt werden können. Insbesondere die Zusammenhänge zwischen Arousal, Stress und Schwierigkeitsempfinden lassen zusätzlich zu den signifikanten SAM Differenzen von einem Induktionserfolg ausgehen.

Die erfassten Freitextantworten dienen einer subjektiven Einschätzung. Die Überprüfung dieser Antworten hinsichtlich Induktionserfolg erfolgte durch einen Experten, es gab dabei keine Hinweise auf Diskrepanzen zu den objektivierbaren Antworten der SAM und USF Antworten. Nach Auswertung aller drei Selbsteinschätzungselemente und einer Expertenanalyse des Experimentalleiters kann abschließend von einem vollumfänglichen Induktionserfolg der emotionalen Zustände Überforderung und Unterforderung ausgegangen werden.

Frage	Korrelation mit Valenz	Korrelation mit Arousal	Korrelation mit Dominanz
Schwierigkeit	$\rho = -0.28, p = 0.09$	$\boldsymbol{\rho = 0.45, p = 0.004}$	$\rho = -0.50, p = 0.001$
Performanz	$\rho = 0.21, p = 0.21$	$\rho = -0.025, p = 0.88$	$\rho = 0.19, p = 0.252$
Stress	$\boldsymbol{\rho = -0.41, p = 0.01}$	$\boldsymbol{\rho = 0.65, p = 3 \cdot 10^{-6}}$	$\boldsymbol{\rho = -0.69, p = 3 \cdot 10^{-6}}$
Motivation	$\rho = -0.05, p = 0.77$	$\rho = 0.12, p = 0.48$	$\rho = -0.1, p = 0.53$

Tabelle 2.5.: Korrelation zwischen SAM und USF für die Überforderungssequenz

Frage	Korrelation mit Valenz	Korrelation mit Arousal	Korrelation mit Dominanz
Schwierigkeit	$\rho = 0.17, p = 0.307$	$\rho = -0.07, p = 0.68$	$\rho = 0.12, p = 0.47$
Performanz	$\rho = 0.03, p = 0.82$	$\rho = -0.22, p = 0.18$	$\rho = 0.18, p = 0.29$
Stress	$\boldsymbol{\rho = -0.46, p = 0.004}$	$\boldsymbol{\rho = 0.59, p = 1 \cdot 10^{-4}}$	$\rho = -0.08, p = 61$
Motivation	$\boldsymbol{\rho = 0.74, p = 1 \cdot 10^{-7}}$	$\rho = -0.22, p = 0.19$	$\rho = 0.13, p = 0.41$

Tabelle 2.6.: Korrelation zwischen SAM und USF für die Unterforderungssequenz

2.3.3. Technische Aufnahme Infrastruktur - MAR²S

Elementare Eigenschaft einer Sensordatenerfassung ist die Möglichkeit einer zeitsynchronen Erfassung aller experimentellen Ereignisse und sensorischen Daten. Bisherige Aufnahme-Paradigmen zur Erhebung einfacher uni- oder bimodaler emotionaler Datensätze verwenden zumeist einen einzelnen und somit zentralen PC zur simultanen Aufzeichnung sensorischer Daten. Auch die Steuerung des experimentellen Ablaufes und des Nutzerinterfaces ist zumeist von geringem Aufwand und wird ebenfalls auf dem selben PC ausgeführt. Eine solche Infrastruktur wurde zum Beispiel zur Erstellung der Datensätze AVEC 2011-2017 [223, 261] so wie EmoRec I und II [194] verwendet. Kommerzielle Hersteller einiger Highend-Sensoren bieten teilweise Lösungen an, bei denen das gesamte Experiment nicht nur auf einem PC abläuft, sondern sogar eine zentrale Softwareumgebung alle Funktionen zentral erfüllt[3]. Im Falle mittels Schauspielern erhobenen Datensätze wie EmoDB [36] oder Kohn Canade [121] vereinfacht sich die infrastrukturelle Anforderung weiter, da hier keine emotionsinduzierende Komponente benötigt wird. Bei allen Infrastrukturen, welche auf einer zentralen Hardware und Software Einheit ausgeführt werden können, entfällt eine aufwendige Synchronisation aller beteiligten Komponenten, da diese sich alle auf die interne Uhr und Taktgeber des PC's beziehen können.

Sollen Datensätze mit einer hohen Anzahl an Modalitäten erzeugt werden, reicht die Performanz eines einzelnen PC's meist nicht aus um die gesamte eingesetzte Sensorik simultan aufzuzeichnen. Hierbei entstehen konfigurationsabhängig typischerweise Engpässe bei den Durchsätzen der beteiligten Busssysteme und Festplatten. Ebenfalls werden Resourcenlimits der CPU durch Kompressionsalgorithmen wie z.B. der MP4-Videokodierung schnell erreicht. Dies wird verschärft, da diese Kompressionsverfahren typischerweise hinsichtlich schneller

[3] Beispielsweise durch Firma gTec mit dem g.MOBIlab+ Sensor, die Firma MindMedia durch den Nexus32 Sensor oder die Firma BrainProducts durch den Enobio Sensor.

Dekomprimierung (abspielen von Videos) optimiert sind, zulasten der Komprimierung (aufzeichnen von Videos). Dies resultiert aus typischen multimedialen Anwendungen. Neben den sensorischen Anforderungen haben sich die Induktionsparadigmen weiterentwickelt von *induktionsfrei* (Schauspiel), über einfache Induktion durch passive Reizpräsentation wie IAPS Bildern, hin zu Vortäuschung komplexer interaktiver Systeme gegenüber dem Nutzer WoZ. Die Technische Umsetzung solcher WoZ Umgebungen erfordert meist ebenfalls eine verteilte Architektur um neben gestiegener Performanzanforderungen auch der räumlichen Trennung von Proband und Versuchsleiter Rechnung zu tragen.

Erste Ansätze solcher verteilten Aufnahme-Infrastrukturen verwenden ein zentral erzeugtes Triggersignal welches, in meist festen Frequenzen, Rechtecksignale an alle beteiligten Komponenten sendet [191, 247, 112, 164]. Der Beginn des Signals (erste steigende Flanke) kennzeichnet dabei den Beginn des Experimentes und somit einen globalen Nullpunkt der Zeitachse. Die Übertragung des Triggersignals erfolgt über technisch aufwendig zu realisierende, separate Leitungen, welche nicht Bestandteil einer standard PC Architektur sind. Allerdings wird dadurch eine hohe Synchronität erreicht. Ein wesentlicher Nachteil besteht allerdings darin, dass alle beteiligten Hardware und Software Komponenten eine Schnittstelle für ein entsprechendes gemeinsames Triggersignal besitzen müssen. Aufgrund umfangreicher elektronischer Übertragungsverfahren, Signalpegeln, Frequenzen und Normen ist jedoch meist eine Anpassung an jeden spezifischen Sensor notwendig, sofern dieser überhaupt eine Triggerschnittstelle besitzt. Ist dies nicht gegeben kann der Sensor nicht verwendet werden. Auch Nutzer und WoZ Schnittstellen können nur bedingt, unter erheblichen Synchronisationseinbußen, an ein solches triggersynchrones System angebunden werden. Ein solche Architektur wurde zur Erhebung des EmoRec II und Last-Minute Korpus verwendet [194, 191].

Die hier vorgeschlagene MAR^{2}S Architektur überwindet die Beschränkungen der Einzel-PC und triggersynchronen Systeme. Die gesamte Synchronisation und Kommunikation aller beteiligten Komponenten erfolgt via Local Area Network (LAN). Dies ermöglicht neben dem Austausch reiner Synchronisationszeitstempel auch den Austausch weiterer Informationen wie z.B. Zustand einzelner Sensoren, Nutzerinformationen und anderer genereller Metainformationen. Zudem ist eine lockere und flexible Kopplung einzelner Hard- und Softwaremodule möglich. Die Synchronisation erfolgt dezentral mittels Trigger Nachrichten auf dem Transmission Control Protocol / Internet Protocol (TCP / IP) Protokoll, die maximal erreichbare Synchronität ist dabei durch die Latenz des LAN beschränkt und erreicht nicht ganz die Genauigkeit hardwaregetriggerter Lösungen. Für den Anwendungskontext des Affective Computing ist die erreichte Synchronität der Modalitäten jedoch ausreichend. Tabelle 2.8 stellt die Latenzen welche im uulmMAC erreicht wurden dar, und lässt erkennen, dass der erwartete mittlere Synchronisationsfehler von 1,11 Millisekunden weit unter der Länge typischer

Modalität	Label	Annotation	Länge (Sekunden)	Std. Abw. (Sekunden)	Minimum (Millisek.)
Video	Alg. Reaktion	Manuell	0,33	0,09	200
Video	Lächeln	Automatisch (openSMILE)	0,85	0,13	240
Video	AU	Automatisch (CERT)	1,87	0,54	240
Video	AU	Manuell (FACS)	0,6	0,10	160
Audio	Alg. Reaktion	Manuell	0,57	0,07	218
Audio	Wortsegmente	Automatisch (Energie)	0,84	0,15	310
Audio	Diskurspartikel	Semiautomatisch	0,26	0,03	100
Pose	Statische Posen	Manuell	7,35	1,88	840
Pose	Statische Posen	Automatisch	7,56	1,89	800
Pose	Dynamische Posen	Manuell	1,35	0,25	480
Pose	Dynamische Posen	Automatisch	1,28	0,26	400
EKG	Resultat Extrema	Klassifikation	15,34	3,78	400
SCL	Signalflanken	Manuell	1,37	0,11	378
SCL	Signalflanken	Automatisch	0,95	0,08	220

Tabelle 2.7.: Durchschnittliche Längen der Labels in verschieden Modalitäten. Die manuelle FACS Annotation erfolgte auf drei Probanden. Alle anderen manuellen Annotationen erfolgten auf neun Probanden. Für die automatisierten Annotationen wurden alle Probanden verwendet. In allen Fällen wurde eine statistische Aussagekraft sichergestellt.

emotionaler Labels liegt (siehe Tabelle 2.7). Der geringfügig erhöhte Synchronisationsfehler ist somit vernachlässigbar. Sollte zwischen zwei im Verbund arbeitenden Sensoren (z.B. zwei Einzelkameras im Stereoverbund) eine geringerer Synchronisationsfehler notwendig sein, kann diese Reduzierung durch Einsatz physikalischer Zusatztrigger zwischen genau diesem Sensorpaar erreicht werden. Die vorgeschlagene MAR^{2}S Architektur optimiert die Synchronisation zusätzlich durch eine Messung der aktuellen Netzwerklatenz zum Zeitpunkt des Aufnahmebeginns. Es werden alle Zeitstempel um die Differenz der gemessenen mittleren Latenz aller Module korrigiert. Der Fehler reduziert sich so von der erwarteten tatsächlichen Latenz auf die tatsächliche Varianz der Latenzen. Durch Erfassung der Latenzen und ihrer Varianzen kann zudem ein Gütemaß für die aktuelle Synchronität gegeben werden. Dieses kann beim Training von Klassifikatoren oder im Einsatz in Companion-Systemen berücksichtigt werden. Im Falle eines Livesystems sollte die Latenzmessung periodisch wiederholt werden um Änderungen im dezentralen Modulgefüge zu erfassen.

Durch den Verzicht auf Hardwaresynchronisation entfällt die Beschränkung auf Sensoren mit Triggeranschluss und es können zusätzlich eine Vielzahl von *off the shelf* Sensoren verwendet werden. Auch die Problematik der drahtlosen und latenzfreien Übertragung eines Hardware-

PC (Sensor)	**Mittlere Latenz** (Millisekunden)	**Mittlerer Synchronisationsfehler** (Millisekunden)
Kinect PC (MS Kinect v2)	5,61	1,35
Video PC (Frontale Webcam) (Headset)	4,95	1,27
Sensor PC (Rückkamera) (MobiLab+)	5,37	0,83
UI PC (-)	5,17	0,95
WoZ PC (-)	5,42	1,17
Durchschnitt:	5,28	**1,11**

Tabelle 2.8.: uulmMAC multimodaler mittlerer Synchronisationsfehler. Die Messung erfolgte durch die Delay Module der MAR^{2}S. Im Mittel weisen Modalitäten einen Synchronisationsfehler von +- 1,11 Millisekunden bezogen auf die gemeinsame Zeitbasis auf. 3 der 100 Messungen wiesen eine Fehler von 4-10 Millisekunden auf. Bei einer Messung wurde ein maximaler Fehler von +13 Millisekunden für den Kinect Sensor beobachtet.

triggersignals im Falle mobiler Sensoren [164, 258] entfällt durch die Verwendung von Wireless LAN Komponenten.

Ein weiterer Vorteil der MAR^{2}S Architektur besteht darin, dass die Aufnahmearchitektur dem selben Aufbau der in Kapitel 5 vorgeschlagen Architektur für interaktive Companion-Systeme und der in in Kapitel 3 beschriebenen verteilten Annotation entspricht. Es erfolgt somit kein Architekturparadigmen Wechsel von der initialen Datenerhebung über die Datenannotation und Klassifikatorentwicklung bis hin zur Implementation voll funktionsfähiger Systeme. Während des Entwicklungszyklusses ist es somit möglich einzelne Komponenten Schritt für Schritt auszutauschen und zu optimieren, ohne Änderungen am gesamten System vornehmen zu müssen. Somit wird ein Einsatz der hier vorgeschlagenen Architektur in gängigen Software- oder System-Entwicklungsmodellen wie dem V-Modell ermöglicht [34, 35].

Die Kommunikation der zum Aufnahmezeitpunkt beteiligten Module erfolgt über die Middleware, welche im Rahmen des SEMAIN Projektes entwickelt wurde [217]. Eine Beschränkung auf diese Middelware ist jedoch nicht gegeben. Abbildung 2.10 zeigt einen Überblick über den infrastrukturellen Aufbau, welcher zur Aufnahme des uulmMAC verwendet wurde. Die Infrastruktur besteht aus fünf PC's welche via LAN gekoppelt sind. Der WoZ PC beinhaltet

Sensor	Einstellung (Kompression)	Roh Datenrate	Komprimierte Datenrate	Datenvol. pro Stunde
GTec	8 Kanal Analog 16Bit@256Hz (nicht notwendig)	32,77 kbit/s	-	60 MB
HD Video	1920x1080@30fps (H264 18Mbit/s)	2,98 Gbit/s	max. 18 Mbit/s	9 GB
SD Video	1280x720@30fps (WMV 3Mbit/s)	1,33 Gbit/s	max. 3 Mbit/s	2,2 GB
SD Video	1280x720@30fps (H264 18Mbit/s)	1,33 Gbit/s	max. 18 Mbit/s	2,2 GB
CD Audio	16bit@44,1kHz Stereo (Wave)	1,41 Mbit/s	1,44 Mbit/s	600 MB
DVD Audio	24bit@48kHz Stereo (MP3@192kbit/s)	2,30 Mbit/s	192 kbit/s	90 MB
Kinect HD Video	1920x1080@30fps (nicht möglich)	5,96 Gbit/s	-	440 GB
Kinect Infrarot	512x484@30fps (nicht möglich)	118,9 Mbit/s	-	60 GB
Kinect Tiefe	512x484@30fps (nicht möglich)	118,9 Mbit/s	-	60 GB
Kinect Audio	24bit@48kHz gerichtetes mono (nicht möglich)	1,54 Mbit/s	-	3,2 GB
Kinect Pose	6 Körper a 25 Knoten + Gesicht bei 30fps (nicht möglich)	100,80 kbit/s	-	362 MB
Kinect Summe	properietär (nicht möglich)	6,19 Gbit/s	-	564 GB
Kinect Nachbearbeitet	- (H264, MP3, XML, CSV)	-	-	28,5 GB
uulmMAC gesamt	-	-	-	**37,5 GB**

Tabelle 2.9.: Datendurchsatz und Speichervolumen der verwendeten Sensoren und Aufnahmeformate

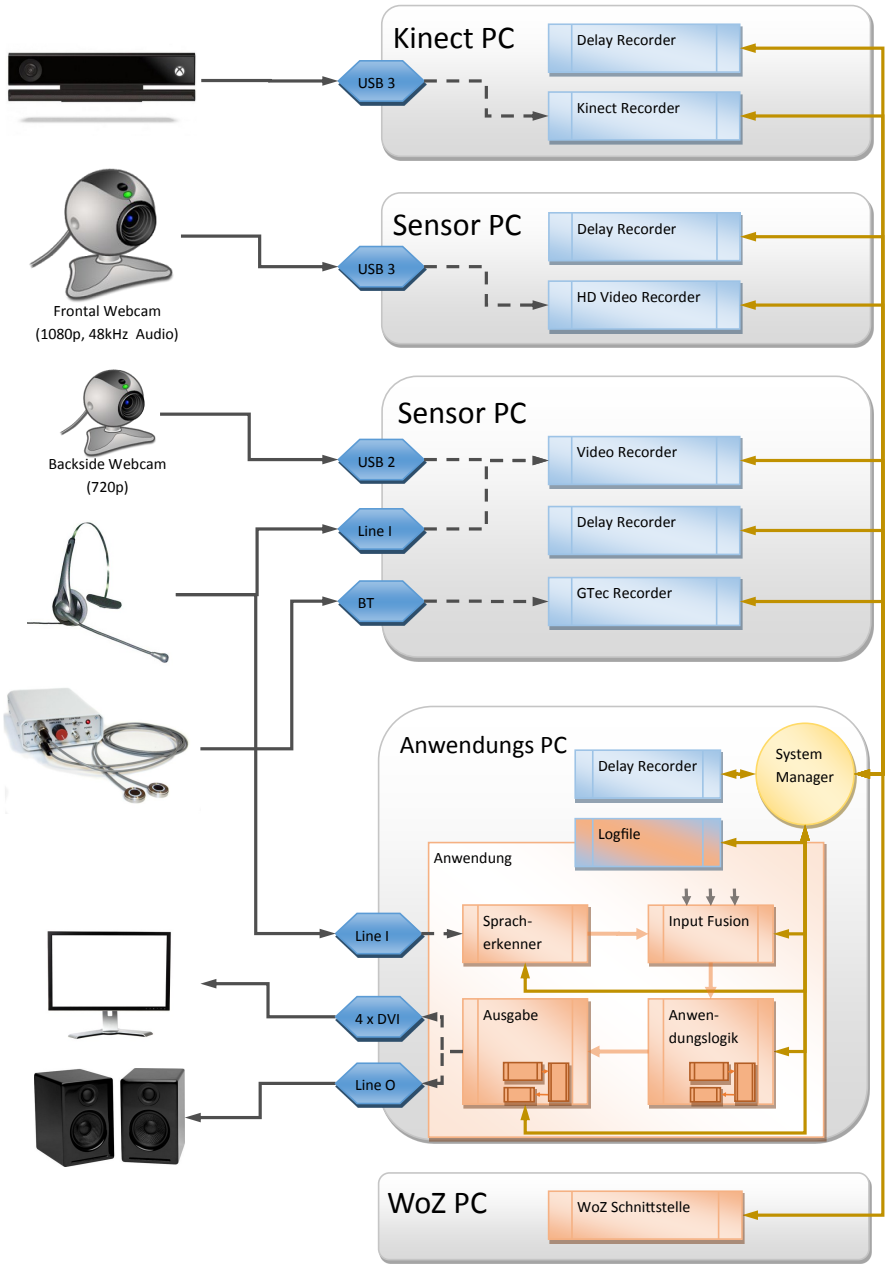

Abbildung 2.10.: Aufnahme Architektur des uulmMAC. Links oben dargestellt die eingesetzten Sensoren mit ihrer jeweiligen Anschlussart an den PC (blaues Sechseck). Grau hinterlegt die eingesetzten fünf PCs, blaue Rechtecke stellen die MAR^{2}S Module der Sensorik dar, orange hinterlegt die Anwendungs- und WoZ- Komponenten angepasst an die MAR^{2}S Struktur. Die zentrale Kommunikation erfolgt über den SEMAINE System Manager (gelb). Links unten dargestellt die Ein/Ausgabegeräte der Anwendungen mit ihren Hardwareanschlüssen.

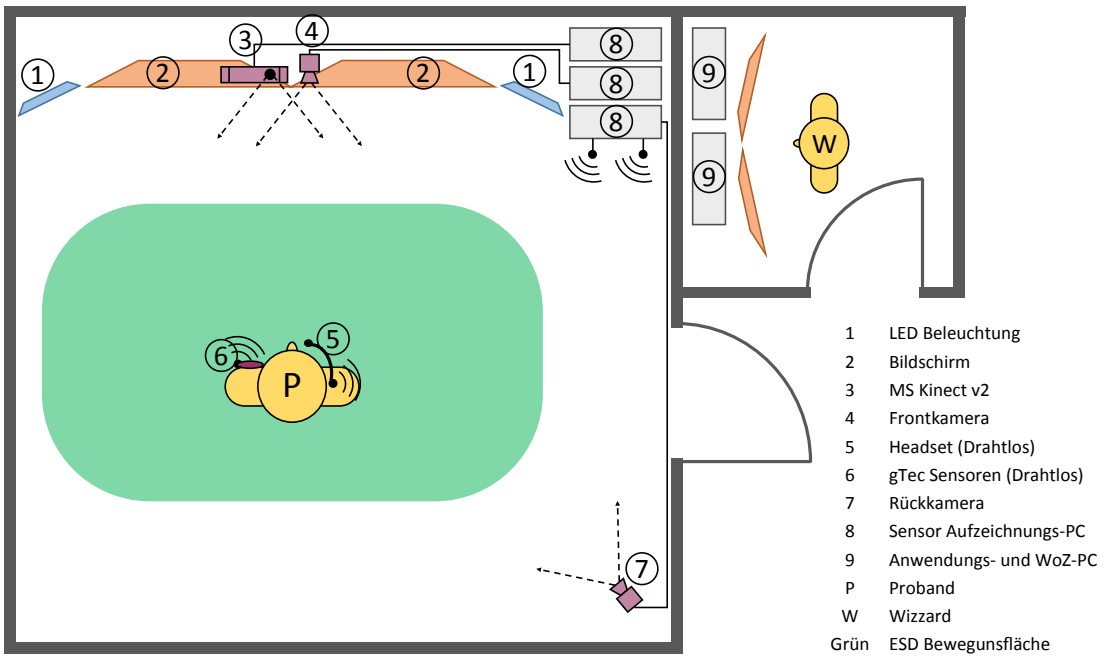

Abbildung 2.11.: Bauliche Anordnung des zur Durchführung des uulmMAC Experimentes eingesetzten Equipments. Der Probandenraum ist baulich vom Wizzard Raum getrennt, die Türen sind während der Aufzeichnung geschlossen. Der Proband bewegt sich während der Interaktion, durch Verwendung drahtloser Sensoren, frei vor der Monitorwand.

lediglich die WoZ Schnittstelle zur Überwachung des experimentellen Ablaufs und eventuell notwendigen manuellen Eingriffen durch den Versuchsleiter. Dies erfolgte räumlich getrennt von der verbleibenden Versuchsanordnung. Der Anwendungs-PC beinhaltet alle notwendigen Komponenten des UI, sowie den Systemmanager der Middleware, welcher als Nachrichtenverteiler fungiert. Die Sensorik ist aus Performanzgründen auf drei weitere PCs verteilt. Die Verteilung der Komponenten und Sensoren auf die PCs erfolgte primär auf Basis der benötigten Datendurchsätze und sekundär aufgrund baulicher Beschränkungen. Tabelle 2.9 gibt einen Überblick über typische Datenraten verschiedener Sensoren. Die räumliche und bauliche Anordnung ist in Abbildung 2.11 dargestellt.

Microsoft Kinect v2

Zur Erfassung der Körperpose und Mimik wird ein Microsoft Kinect v2 Sensor verwendet. Hierbei handelt es sich um einen Kombinationssensor, welcher mit einer internen Taktrate von 30Hz arbeitet. Dieser erfasst simultan ein HD Farb-Video sowie ein Infrarot-Video. Zusätzlich kommt ein integrierter Tiefensensor zum Einsatz, welcher nach dem *time of flight* Prinzip arbeitet. Aus diesen drei Basismodalitäten wir mittels integrierter Algorithmen die Körperpose von bis zu sechs Personen geschätzt. Die Körperpose beinhaltet dabei einen Graphen (Skelett) mit 25 Gelenkknoten, deren Position in einem kartesischen (X, Y, Z) Koordinatensystem relativ zum Sensor angegeben sind. Die Güte der Schätzung jedes Gelenkes wird zudem durch einen Konfidenzwert ergänzt. Abbildung 2.13 zeigt den Bildausschnitt des Sensors (links), eine Tiefendarstellung mit Skelett (mitte) und die Montageposition des Sensors über dem Bildschirm (rechts). Weiter wird die Kopforientierung geschätzt, die Winkelgenauigkeit ist jedoch auf fünf Grad begrenzt. Ebenfalls wird die Position der Augen, Nase und Mundwinkel durch den Kinect Sensor erfasst. Aus diesen Informationen werden wiederum einfache kategoriale Nutzerinformationen abgeleitet. (Augen offen oder geschlossen, Mund offen oder geschlossen, Brillenträger, Blickrichtung, Basisemotionen) Die Qualität dieser abgeleiteten Nutzerinformationen ist als akzeptabel einzustufen, hängt jedoch deutlich hinter modernen Methoden der Wissenschaft hinterher. Für genauere Schätzungen ist es unerlässlich auf Basis der sensorischen Abtastwerte (Farb- und IR-Video sowie Tiefenkarte) durch verbesserte Algorithmen verbesserte Vorhersagen zu berechnen. Neben den visuellen Kanälen wird ein, mittels Mikrofonarray erfasstes, Audiosignal erzeugt, welches neben der akustischen Abtastung zusätzlich eine Konfidenz behaftete Richtungsangabe der Haupt-Audioquelle angibt.

Abbildung 2.12.: Screenshot des MAR^{2}S Kinect Aufzeichnungsmoduls (KinectRecorder)

Die Hardware der Kinect besitzt keine Möglichkeit zum Anschluss eines Hardwaretriggersignals, wodurch eine Anbindung nur mittels der MAR^{2}S Architektur möglich ist. Aufgrund der enormen Datenraten[4] ist es zu empfehlen den Kinect Sensor exklusiv auf einem PC zur

[4]Typischerweise 100 bis 150 MB/Sekunde

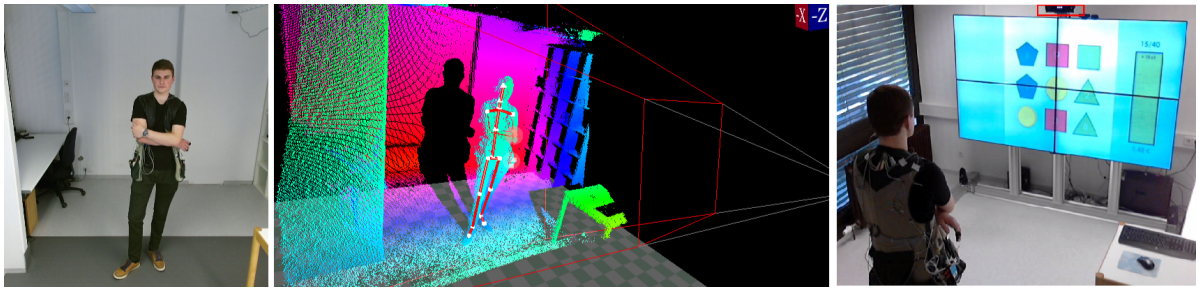

Abbildung 2.13.: Links: Kinect Videobild. Mitte: 3D Darstellung der erfassten Tiefenkarte. Rechts: Die Rückansicht der Szene zeigt die Position der Kinect oberhalb der Monitorwand.

Aufzeichnung einzusetzen. Weiter ist es mit gängigen CPUs nicht möglich die entstehenden Datenströme in Echtzeit zu komprimieren. Das hierfür in C# entwickelte MAR^{2}S Modul zeichnet daher die anfallenden Datenströme zunächst unkomprimiert auf, wobei auf eine hoch effiziente Implementierung zu achten ist, da die erforderlichen Datendurchsätze an der Grenze der Systemressourcen moderne highend Hardware liegen, wodurch kaum Ressourcen für nebenläufige Prozesse oder Threads vorhanden sind.[5] Hierbei wird das von Microsoft bereitgestellte XEF Format[6] verwendet. Abbildung 2.12 zeigt einen Screenshot des im Einsatz befindlichen Moduls. Erst im Nachgang einer Aufzeichnung werden die sensorischen Rohdaten sowie in erster und zweiter Instanz abgeleiteten Informationen unter Einsatz eines, in Kooperation mit Rau et. al. entwickelten, Konvertierungswerkzeuges zu gängigen und standartoffenen Formaten weiterverarbeitet. Dabei handelt es sich um die ATLAS kompatiblen Formate MPEG-4 Part 14 (MP4)(H264), WAVE, XML, CSV und Plain Text. Die erzeugten XEF Rohdaten können zudem unter Verwendung der Microsoft Kinect Studio v2 Umgebung zur Simulation eines physikalischen Kinect v2 Sensors verwendet werden.

Biophysiologische Daten

Zur Erfassung der biophysiologischen Messwerte kommt der portable Signalverstärker g.MOBIlab+ der Firma gTec[7] zum Einsatz. Dieser bietet die Möglichkeit bis zu acht analoge und ein acht Bit breites digitales Signal zu erfassen und drahtlos via Bluetooth 2.0 an einen PC zu übertragen. Die anlogen Messwerte werden mit 16 Bit Auflösung abgetastet, die Abtastrate ist bei allen Eingängen einheitlich vom Hersteller auf 256 Hz festgelegt. Die elektrische Konfiguration der analogen Eingänge wird ebenfalls vom Hersteller einmalig fest eingestellt, eine Änderung der Konfiguration ist nicht möglich. Prinzipiell bietet der digitale

[5] Das sensorische Modul wird in unveränderter Form bei der Erhebung des SFB-TRR 62 Datensatzes Ganganalyse und Fitness (GF) und weiterer Pilotstudien eingesetzt.
[6] Proprietäres Format, vom Hersteller werden keine Informationen hierzu veröffentlicht.
[7] http://www.gtec.at/Products/Hardware-and-Accessories/g.MOBIlab-Specs-Features

Eingang die Möglichkeit ein Hardwaretriggersignal zu empfangen, allerdings bietet der Hersteller keine Lösung zur drahtlosen Übertragung digitaler Signale zum Messverstärker an, weiter ist es auch nicht vorgesehen die digitalen Eingänge zur Steuerung des Sensors einzusetzen. Da im Rahmen der MAR^{2}S Architektur ein Trigger nicht vorgesehen ist wurde darauf verzichtet eine, als störanfällig zu erwartende, technisch aufwendige Lösung, wie von Tornow et. al. beschrieben [257], zu erarbeiten. Stattdessen wird der Sensor so konfiguriert, dass er nach einschalten und verbinden zum PC (via Bluetooth) zu einem, eigens entwickelten, MAR^{2}S Modul beginnt Messdaten zu übertragen. So ist auch die Möglichkeit für den Versuchsleiter gegeben die Qualität der Messdaten bereits zum Zeitpunkt der Anbringung der Elektroden und Sensoren am Probanden zu validieren. Abbildung 2.14 zeigt einen Screenshot des Moduls mit Livedarstellung der Messdaten.

Die vom Hersteller bereitgestellte Software ist nicht quelloffen und bietet keine Möglichkeit eine Aufnahme extern zu steuern. Das Konzept der Herstellersoftware entspricht dem technischen Stand einer Einzel-PC Aufnahmelösung, gesteuert mittels einer proprietären und zentralen Softwarekomponente. Es besteht dort lediglich die Möglichkeit des Einsatzes einer Webcam als weiteren Sensor. Es wurde daher auf Basis der bereitgestellten *low level* C/C++ Application Programming Interface (API), welche unmittelbar auf den Bluetooth Stack aufsetzt, zunächst eine Wrapper Bibliothek entwickelt, welche die technischen Details des Sensors steuert (Initialisierung, Timing, Ringspeicher, Bitshifting, etc.) und eine Anbindung an C# Umgebungen auf abstrakterem Level erlaubt (*Aufnahme Start / Stop / Reset*). Dieser Wrapper ist in das eigentliche MAR^{2}S Aufnahmemodul gekapselt, welches die Anforderungen der MAR^{2}S Schnittstellen realisiert und die empfangenen Daten entsprechend ihrer synchronen Zeitstempel in eine CSV Datei schreibt[8]. Um spätere

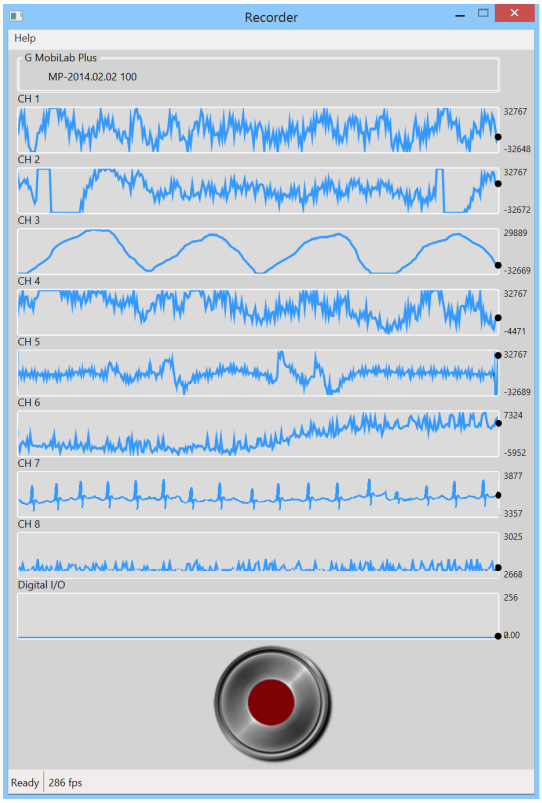

Abbildung 2.14.: Screenshot des MAR^{2}S biophysiologischen Aufzeichnungsmoduls (gRecorder)

[8] Wrapper und sensorisches Modul wurden in unveränderter Form bei der Erhebung des SFB-TRR 62 Datensatzes GF und in angepasster Form bei der Aufzeichnung des Schmerz-Datensatzes des BMBF geförderten Projektes SENSE-Emotion verwendet. Darüber hinaus kommt das Modul bei einigen weiteren Studien durch

Einsatzmöglichkeiten nicht zu beschränken, erfolgt zur Aufnahmezeit keine Signalanpassung bzw. Signalfilterung.

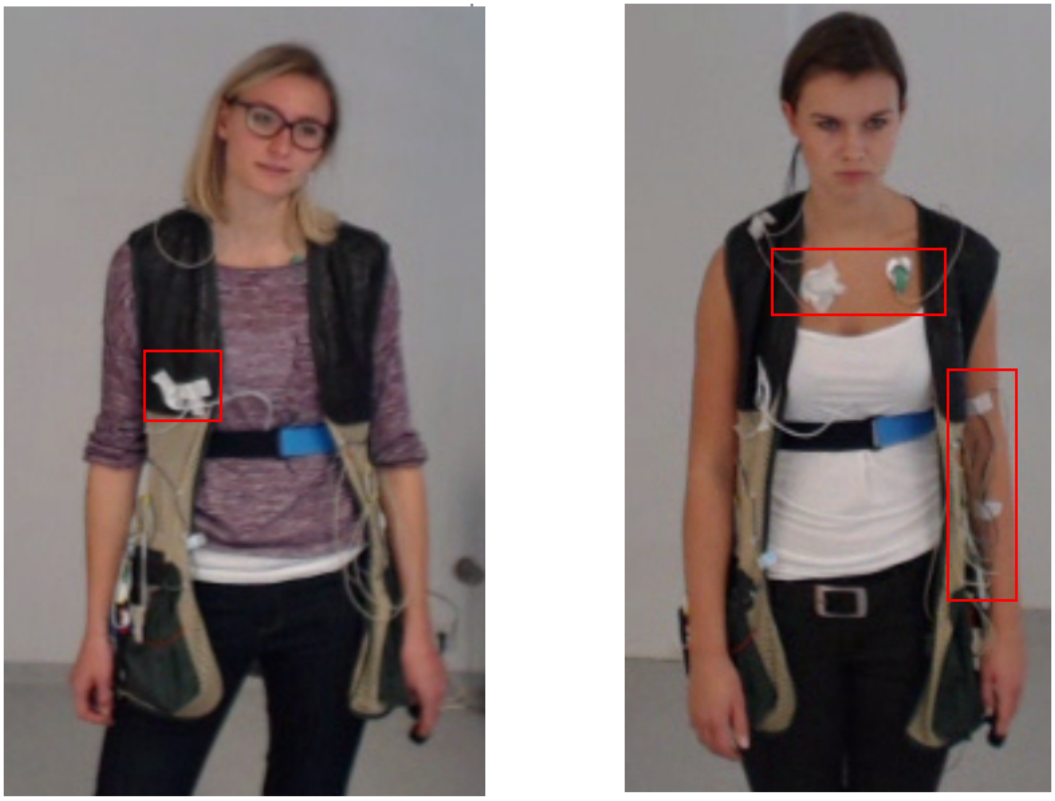

Abbildung 2.15.: Positionierung und Befestigung der drahtlosen biophysiologischen Sensoren des uulmMAC. Unter der Kleidung getragen Sensoren sind teilweise verdeckt.

Von den Probanden werden folgende Messwerte sensorisch erfasst:

- EKG, Bipolar, erfasst via dermaler Brustelektrode
- Atmung, aktiv verstärkt, erfasst via Brustdehnungsgurt
- Hautleitwert, aktive verstärkt, via zwei Fingersensoren an nicht dominanter Hand
- Temperatur, aktiv verstärkt, Fingersensor an nicht dominanter Hand
- EMG, unipolar, Musculus Zygomaticus
- EMG, unipolar, Musculus Corrugator Supercilii

Arbeitsgruppen der Universität Ulm zum Einsatz. Dem EU geförderte Projekt BRISC wurde eine, an den Enobio Sensor des Hersteller Neuroelectrics, angepasste Variante des Moduls zur Verfügung gestellt.

Akustische Daten

Zur Erfassung der Raum- und Probandenakustik kommen drei Mikrofone zum Einsatz. Mittels eines im Kinect Sensor enthaltenen Mikrofonarrays wird durch *virtual beam forming* ein Richtmikrofon in Richtung des Probanden simuliert [18]. Die Aufzeichnung dieser Audiodaten erfolgt im Kinect Modul.

Ein weiteres ungerichtetes Mikrofon ist in der rückwärtigen Kamera integriert, es erfasst die Umgebungsgeräusche des Raumes. Diese Atmosphärenakustik wird zusammen mit dem rückwärtigen Video im entsprechenden Videomodul aufgezeichnet. Als primäres Mikrofon ist am Probanden ein drahtloses Headset angebracht (Frequenzbereich 25 Hz - 22 kHz, Nierencharakteristik, via digitalem UHF Funk übertragen). Dieser akustische Kanal wird mittels Mischpult individuell eingepegelt, das Signal wird zum einen an den Eingang des Spracherkenners angelegt, zum anderen an den akustischen Eingang das Aufzeichnungsmoduls des frontalen Videos und dort erfasst. Alle akustischen Kanäle werden einheitlich in DVD Qualität aufgezeichnet (48 kHz Abtastrate mit 2x16 Bit Stereo Auflösung). Durch die Aufzeichnung der akustischen Kanäle mittels der im nächsten Abschnitt beschriebenen Videomodule, entfällt die Entwicklung eines separaten Audioaufzeichnungsmoduls. Die Videomodule erlauben es im Spezialfall reiner akustischer Anforderungen den Videokanal unbelegt zu lassen, somit ist eine modulare Anwendbarkeit auch in rein akustischen Anwendungsszenarien gewährleistet. Im Falle kombinierter Aufnahmen können visuelle und akustische Spur einer Aufnahme im Nachgang getrennt werden ohne dabei an Synchronität zu verlieren.

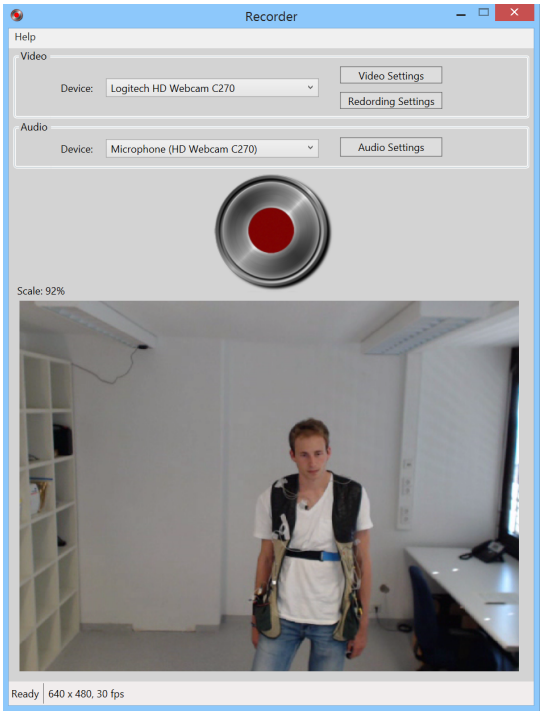

Abbildung 2.16.: Screenshot des MAR^{2}S Audio-Video Aufzeichnungsmoduls.

Visuelle Daten

Visuelle Daten können neben klassischen Farbvideoaufzeichnungen auch weitere, durch Bildfolgen darstellbare, sensorische Informationen beinhalten. Beispiele hierfür sind Infrarot (IR)-, Wärmebild- oder Tiefenkameras. Auch ist die Aufzeichnung nicht auf eine Signalquelle begrenzt, so können z.B. die beiden Bilder einer Stereokamera simultan nebeneinander in einen Videostrom kodiert werden. Das hierfür entwickelte Modul ist nicht an einen speziellen Sensor angepasst, sofern vom jeweiligen Hersteller ein entsprechender Videotreiber für Microsoft Windows bereitgestellt wird. Das Modul zeichnet zunächst aus Performanzgründen im WMV Codec auf, obwohl dieser aufgrund seiner Eigenschaften für einen spätere bildgenaue Positionierung des Bildzählers nicht geeignet ist. Im Nachgang einer Aufnahme werden diese Daten dann ohne Synchronisationsdefizit in den besser geeigneten, jedoch aufwendiger zu kodierenden, MP4 Codec gewandelt.

Für die Aufzeichnung des uulmMAC kommen zwei Videomodule[9] Instanzen zum Einsatz. Eines zeichnet eine rückwärtige Kamera auf, welche dem Probanden über die Schulter schaut und die gesamte Szenerie erfasst. Diese Ansicht wird zusätzlich dem Versuchsleiter zur Verfügung gestellt, da sie eine guten Überblick über den gesamten Versuch liefert. Zusätzlich wird in diesem Video eine Atmosphärenakustik der Umgebung aufgezeichnet. Die Videoauflösung liegt bei 1280x720 RGB-Farbpixeln. Die zweite Instanz zeichnet eine frontal platzierte Kamera auf, welche Gesicht und Körpervorderseite des Probanden beobachtet. Dieser Ansicht ist das akustische Signal des Headsetmikrophons zugeordnet. Die Videoauflösung liegt bei 1920x1080 RGB-Farbpixeln. Bei beiden Kameras kommt eine Modell der Firma Logitech vom Typ C9100 zum Einsatz, dessen Bildrate an die Bildrate des Kinect Sensors angepasst wurde, somit laufen alle bilderzeugenden Sensoren mit einheitlichen 30 Bildern pro Sekunde. Die verwendeten handelsüblichen *low cost* Webcams besitzen keine Anschlussmöglichkeit an Hardwaretriggersignale und können somit nur mit der MAR^{2}S Lösung betrieben werden. Durch den Einsatz handelsüblicher Sensoren entspricht das erzeugte Datenmaterial in Bezug auf Qualität und Bildausschnitt den typischen Kameras, welche in Smart-Devices, Laptops oder ähnlichen Consumer Elektronik Geräten eingesetzter Bilderzeugungssensoren. In Bezug auf Training von Klassifikatoren welche in realen Umgebungen eingesetzt werden können, ist daher mit geringeren Problemen zu rechnen.

[9] Entwickelt von F. Schüssel und F. Honold

Anwendungs- und Systemdaten

Neben den sensorischen Daten werden während des experimentellen Ablaufs auch eine Vielzahl von symbolischen Informationen generiert. Diese symbolischen Daten können zum verbesserten Training von Klassifikatoren eingesetzt werden. Darüber hinaus enthalten diese eine Vielzahl von Annotationsinformationen, weitere Informationen hierzu sind im Abschnitt 2.3.4 beschrieben. Aufgrund der modularen Struktur werden alle, das Systemverhalten beschreibenden Informationen, über die Middleware der MAR^{2}S Architektur versendet, ein entsprechendes Modul zeichnet diesen Datenstrom vollständig auf. Diese Daten umfassen unter anderem:

- Zeitstempel und Inhalt aller Nutzereingaben
- Zeitstempel, Konfidenz und Inhalt aller vom Spracherkenner erkannten Eingaben
- Zeitstempel und Inhalt aller Systemausgaben
- Zeitstempel und Inhalt aller WoZ Eingaben und Systemeingriffe
- Nutzerstatus bezüglich Suchphasen, Reaktionsphasen und Selbstbewertungsphasen
- Angaben zur Selbsteinschätzung
- Informationen, welche im Nutzermodell erfasst sind

Hieraus kann im Nachgang das gesamte Experiment im Detail nachvollzogen werden. Mittels des von Schüssel und Honold entwickelten System Logger Wiedergabe Moduls können alle Inhalte des Logfiles zeitsynchron wiedergegeben werden. Zusammen mit den Wiedergabemodulen der sensorischen Daten kann so das gesamte Experiment zu einem späteren Zeitpunkt simuliert werden. Dies ist besondere vorteilhaft bei der Entwicklung und Erprobung von Echtzeitklassifaktionsmodulen und vollständigen Companion-Architekturen, da solch komplexe Systeme nur unter Einsatz von Vorgehensmodellen wie dem V-Modell, Wasserfallmodell oder Scrum realisierbar sind, welche alle eine modulare Struktur voraussetzen [34, 35, 225].

Netzwerk-Latenz-Monitor

Zur Erfassung der Netzwerklatenzen und der damit verbunden Abschätzung des erwarteten Synchronisationsfehlers wurde ein C# Modul entwickelt, welches beim Erhalt einer *Start* Nachricht einige Impulsnachrichten zum Middleware Server (System Manager) schickt, welche dieser wiederum umgehend zurück schickt (ähnlich einer TCP / IP Ping Nachricht). Die mittlere Laufzeit dieser Nachrichten entspricht der erwartenden Verzögerung, bestehend aus

Netzwerklatenz und Verarbeitungsdauer am Server[10]. Da auch das startauslösende Modul erst mit der Aufnahme beginnt wenn es seine eigene Start Nachricht zurück erhalten hat, entspricht der mittlere Synchronisationsfehler lediglich den Differenzen der mittleren Latenz. Durch Nachverarbeitung der Zeitstempel kann die Synchronität im Falle von Offlineverarbeitung der Daten so weiter erhöht werden. Da dies im Onlinebetrieb realer Systeme allerdings kaum möglich ist und der erreichte Synchronisationsfehler (siehe Tabelle 2.8) bereits ausreichend gering ist, wurde hier darauf verzichtet. Durch entsprechend hoch effiziente Entwicklung und vorab Initialisierung aller MAR^{2}S Module kann die Berechnungszeit zwischen Eintreffen einer Startnachricht und dem tatsächlichen Beginn der Aufnahme im jeweiligen Modul vernachlässigt werden.

2.3.4. Annotation des uulmMAC

Der gesamte Datenbestand des uulmMAC steht in Form von aufbereiteten und voll umfänglichen ATLAS Projekten zur Verfügung. Abbildung 2.17 zeigt eine beispielhafte Abbildung. Neben der Darstellung aller erhoben multimodalen Sensordaten beinhalten diese Projekte auch diverse Labelspuren welche automatisch, halb automatisch und manuell erzeugt wurden. Im folgenden werden alle, zum Zeitpunkt der Erstellung dieser Arbeit, vorhanden Labelspuren beschrieben, wobei durch weitere Fragestellungen und Untersuchungen kontinuierlich neue Informationen annotiert werden. Nach entsprechender Qualitätsprüfung werden diese Erweiterungen den ATLAS Projekten des uulmMAC hinzugefügt. Abbildung 2.18 zeigt eine Auswahl der zur Verfügung stehenden Labels.

Experimentalphasen

Aus den Aufzeichnungen der Anwendungs- und Systemdaten können die relevanten Zeitpunkte der Experimentalphasen automatisch extrahiert werden. Diese umfassen neben den Start- und Endzeitpunkten der sechs Experimentalphasen auch die Zeitpunkte der zwischen den Experimentalphasen gelegenen Nutzerselbsteinschätzungen (SAM, USF und Freitext Einschätzungen). Die Labels der Selbsteinschätzung beinhalten zusätzlich die vom Nutzer angegeben Einschätzungswerte. Diese Labels stehen für alle Aufnahmen zur Verfügung. Neben den emotionalen Labels einer Phase welche zum Klassifikatortraining verwendet werden, und einer Orientierungshilfe für den Experten, stellen diese Labels eine wichtige Kontextinfor-

[10]Das Verfahren ist an das Simple Network Time Protocol (SNTP) Verfahren angelehnt. `https://de.wikipedia.org/wiki/Network_Time_Protocol` (Stand 3. Juni 2019)

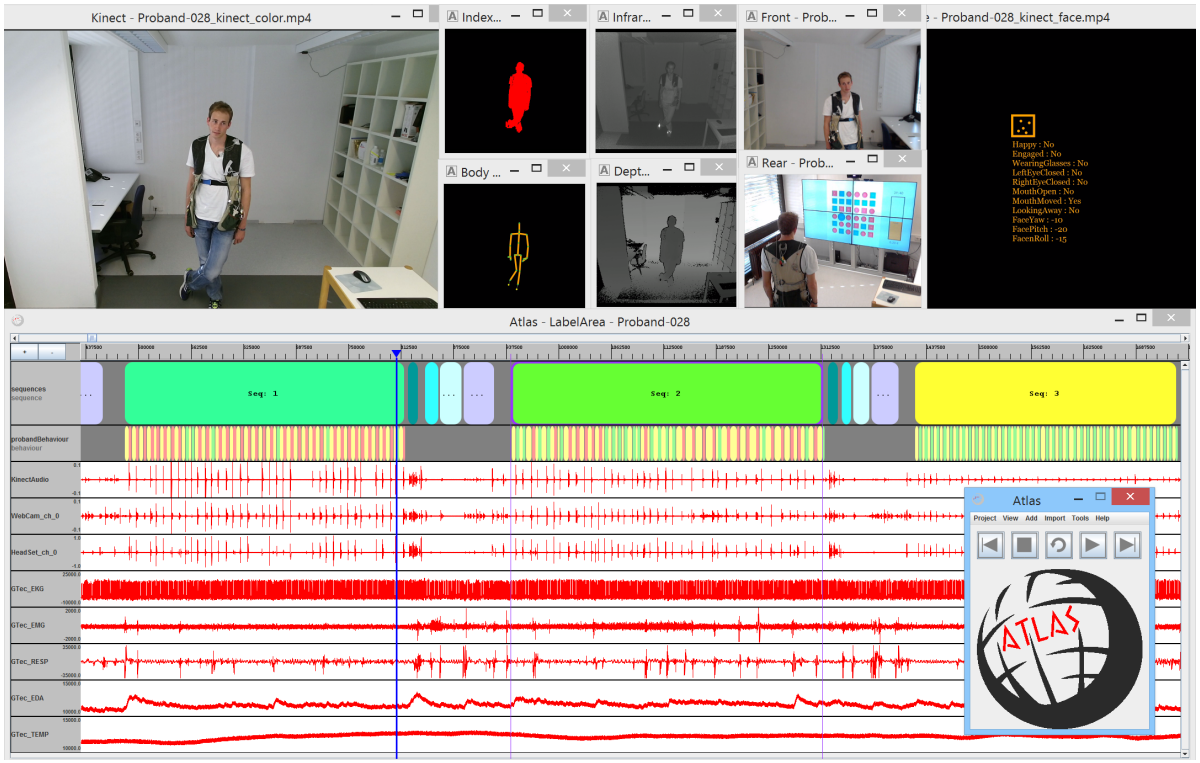

Abbildung 2.17.: Screenshot des uulmMAC in ATLAS. Oben links ist das frontale Videobild groß dargestellt. Daneben die Segmentierung der Körpersilhouette, der Körperposengraph, Infrarotvideo, Tiefenkarte, sekundäres Frontalvideo, Rückansicht und Gesichtsposition mit Labels ganz rechts. Darunter die Labelspuren der Sequenzgrenzen und die Labels der Such- und Antwortphasen. Anschließend folgen die sensorischen Daten, von oben nach unten: gerichtetes Audio, Atmosphären Audio, Headset Audio, Elektrokardiographie (EKG), EMG (Trapetius), Atmung, Hautleitwert (engl. Skin Conductance Level) (SCL) und Temperatur.

mation zur Klassifikation dar. So ist a priori im Falle komplexer Phasen mit einer erhöhten Wahrscheinlichkeit einer Überforderung zu rechnen und vice versa.

System Ein-/Ausgabe Zeitpunkte

Innerhalb der experimentellen Phasen stehen die Zeitpunkte der Systemausgaben, Zeitpunkte der Präsentation der Suchmatrix, Zeitpunkte des Countdown Beginns, sowie die Zeitpunkte der Nutzereingabe zur Verfügung. Anhand dieser Informationen wird jeweils zwischen Suchphase und Reaktionsphase des Nutzers unterschieden. Die Reaktionsphase wird dabei weiter in Reaktion auf richtiges, falsches oder nicht Lösen der Aufgabe unterschieden. Diese automatisch erzeugten Labels stehen für alle Aufnahmen zur Verfügung. Diese Labels sind als relevanter Kontext für Klassifikations- und Fusionsarchitekturen einzustufen. So ist während des Suchens der Lösung beim Nutzer mit erhöhter Anspannung zu rechnen, während im Falle von falscher

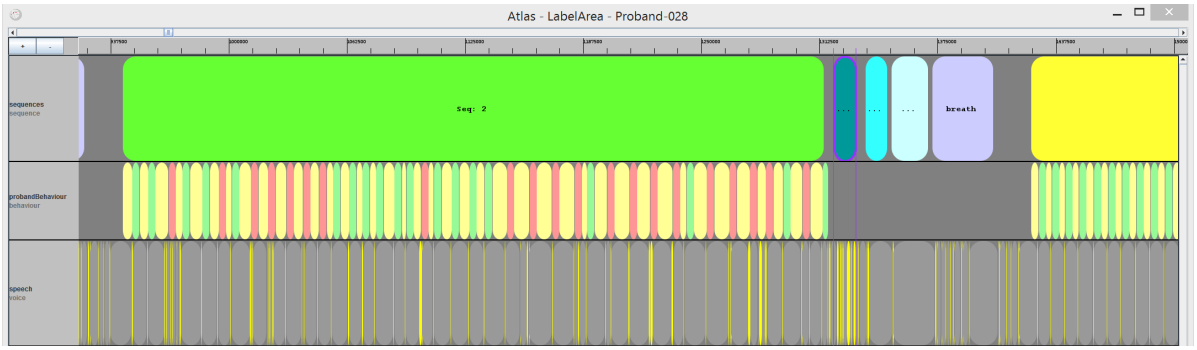

Abbildung 2.18.: Screenshot der uulmMAC Labels in ATLAS (Auszug aus Sequenz 2). Spur 1: Sequenzgrenzen (grün), SAM, USF, Freitext Feedback und Atemübung. Spur 2: Suchphasen (gelb), richtige Antwort (grün), falsche oder verspätete Antwort (rot)). Spur 3: Spracheingaben (gelb).

oder richtiger Nutzerantwort mit einer erhöhten Wahrscheinlichkeit für positives oder negatives emotionales Nutzerverhalten zu rechnen ist. Diese Kontext Information wurde bei der Erstellung des SFB-TRR 62 Cluster 8 Demonstrators einbezogen und demonstriert.

Automatische Spracherkennung

Die verbalen Eingaben des Nutzers werden durch den, auf der *Cortana Engine* basierenden, von Microsoft Windows bereitgestellten, Spracherkenner verarbeitet. Dieser stellt neben Zeitstempeln der Eingabe auch einen Konfidenzwert zur Qualität der Erkennung bereit. Zeitpunkt, Inhalt und Konfidenz aller verbalen Äußerungen stehen als Label-Spur zur Verfügung. Diese weist neben den erwarteten Nutzerantworten in Spielphasen auch Informationen zu gelegentlichen spontanen Nutzeräuserungen (off Talk) auf. Diese Labels stehen ebenfalls für alle Nutzer zur Verfügung.

Automatische akustische Segmentierung

Basierend auf der Signalenergie sind in dieser Labelspur alle akustischen Eingaben, unabhängig vom Spracherkenner, segmentiert. Die hier enthaltenen Labels enthalten die mittlere Signalenergie des jeweiligen Segmentes als zusätzliche Information. Da der Spracherkenner nicht alle verbalen Äußerungen verarbeitet, bzw. nicht ausgibt, enthält diese Spur neben den verbalen Eingaben und spontanen Äußerungen zusätzliche verbale akustische Reaktionen wie z.B. Räuspern, Husten oder tiefe Atemgeräusche. Diese Spur steht für alle Aufnahmen zur Verfügung.

Manuelle akustische emotionale Labels

Ausgehend von den akustischen Segmenten wurden diese manuell bezüglich emotionalem Inhalt annotiert. Dabei wurden den Segmenten die Eigenschaften *emotional behaftet* oder *emotional nicht behaftet* zugewiesen. Eine Unterscheidung bezüglich Inhalt der auftretender Emotion erfolgt nicht. Eine Betrachtung der emotionalen Anteile zeigt, dass nur ein geringer Anteil aller auftretenden Äußerungen emotional behaftet ist (siehe Tabelle 3.3). Die Schwelle der Unterscheidung zwischen emotionaler Behaftung und neutraler Stimmlage wird dabei durch einen erfahrenen Annoteur festgelegt. Zusätzlich zur manuellen inhaltlichen Annotation wurden eventuell falsch erfasste Segmentgrenzen korrigiert oder fälschlich verbundener Segmente getrennt. Eine solche manuelle Annotation bedarf, wie in Tabelle 3.1 dargestellt, einem enormen Zeitaufwand. Diese Labels stehen nur bei einem Subset bestehend aus neun Probanden zur Verfügung. Diese dienen u.A. als Grundlage zur Entwicklung und Evaluation von aktiven Lernverfahren.

Manuelle visuelle emotionale Labels

Analog der manuellen akustischen Annotation wurde auch der visuelle Kanal hinsichtlich emotionaler Expresionen im Gesicht manuell untersucht [253, 254]. Da hier keine Vorsegmentierung erfolgt, wird die Segmentierungsaufgabe vollständig vom Annoteur übernommen. Annotiert wurden emotionale Ausdrücke ohne Betrachtung der zu Grunde liegenden Emotionskategorie. Die Schwelle der Unterscheidung zwischen emotional behafteter und neutraler Ausdrücke wird dabei durch einen erfahrenen Annoteur festgelegt. Auch die visuelle manuelle Annotation bedarf, wie in Tabelle 3.1 dargestellt, einem enormen Zeitaufwand. Diese Labels stehen analog zu den manuellen akustischen Labels nur bei einem (identischen) Subset bestehend aus neun Probanden zur Verfügung. Diese dienen ebenfalls als Grundlage zur Entwicklung und Evaluierung von aktiven Lernverfahren.

Manuelle Körperpose Labels

Für das neun Probanden umfassend Subset wurde die Körperpose ebenfalls manuell annotiert. Auch hier wurden diejenigen Segmente manuell annotiert, welche gemäß entsprechender Vorüberlegungen potentiell mit emotionalen Zuständen in Verbindung stehen [106, 105]. Dabei wurde zwischen ausgewählten statischen Körperposen (Arme gekreuzt, Hände hinter dem Rücken, Hände in der Hüfte, gekreuzte Beine, Beine in Schrittstellung) und dynamischen

Körperposen (seitwärts Bewegung der Hände, Selbstberührungen im Gesicht, tänzeln der Beine) unterschieden.

Halbautomatische Labels

Durch Einsatz der in Abschnitt 3.3 beschriebenen aktiven Lernverfahren wurden auf Basis der manuellen Labels Klassifikatoren für den akustischen und visuellen Kanal trainiert. Da diese aktiven Lernverfahren nur auf individual spezifischen Klassifikatoren schnell gute Ergebnisse liefern, ist die Anwendung dieser Klassifikatoren zwar hilfreich bei der initialen Annotation weiterer Probanden, bedarf allerdings einer manuellen Kontrolle der Ergebnisse. Der Annoteur startet im Falle halbautomatischer Annotation somit mit automatisch generierten Labelvorschlägen, diese werden vom Annotuer bestätigt oder als falsch abgelehnt. Die initialen Klassifikatoren werden so stückweise auf das Individuum angepasst und die Vorschläge optimiert (Methodische Details finden sich in Kapitel 3). Nach erreichen einer Qualitätsschranke wird der Annotationsprozess abschließend mit den individualisierten Klassifikatoren unüberwacht beendet. Mittels dieses Prozesses wurden akustische und visuelle unimodale Labels für die verbleibenden 31 Probanden des 40Singles Subset erzeugt.

Automatische Labels

Eine völlig unüberwachte akustische oder visuelle Annotation ist selbst unter Zuhilfenahme bereits im experimentellen Kontext vortrainierter Klassifikatoren derzeit nicht in akzeptablen Qualitätsmaß möglich. Auf die Erzeugung automatischer Labels wurde im visuellen und akustischen Kanal daher verzichtet.

Aus den manuell generierten Körperpose Labels lassen sich einfachste Statistiken berechnen, mittels derer die vom Kinect Sensor gelieferten Pose Informationen automatisch annotiert werden können. Für statische Posen sind dies im wesentlichen distanzbasierte Regeln wie z.B.

$$HaendeHinterRuecken = dist(Hand_R, Hand_L) < a \wedge posZ(Hand_R) > posZ(Huefte)$$

Wobei der Parameter a aus den manuellen Labels empirisch bestimmt wurde. Dynamische Labels benötigen zusätzlich zu diesen Regeln noch Informationen zur mittleren Geschwindigkeit und Beschleunigungen der relevanten Knoten (erste und zweite Ableitung der Knoten). Mittels dieser Information können ebenfalls manuell Regeln definiert werden, welche zur automatischen Annotation ausreichen. Die erreichte Qualität dieser Labels ist ausreichend hoch, wodurch kein halbüberwachter Zwischenschritt notwendig ist. Allerdings betrachten diese

Labels nicht direkt einen emotionale Zustand, sie können daher auch als Merkmal höherer Ordnung angesehen werden. Die Labels stehen für alle Probanden zur Verfügung.

2.3.5. Merkmalsextraktion

Bestandteil des uulmMAC ist neben den Sensorischen- und Annotationsdaten auch eine umfangreiche Merkmalsbibliothek. Diese umfasst die wesentlichen, in der *Affektive Computing Community* etablierten Merkmale. Methodische Details der Merkmale finden sich in Kapitel 4. Es stehen folgende modalitätstypische Merkmale für das *40Singles* Subset zur Verfügung[11].

- Visueller Kanal

  LBP-TOP (30 Frames Analysefenster)

  PHOG (30 Frames Analysefenster)

  Geometrischer Gesichtsgraph (64 Knoten, Einzelbilder)

  OpenSmile (default settings)

- Akustischer Kanal

  MFCC (20 Dimensionen, 32ms Analysefenster)

  LPC (8 Dimensionen 32ms Analysefenster)

  RastaPLP (8 Dimensionen, 40ms Analysefenster)

  ModSpec (8 Dimnsionen, 1000ms äußeres mit 200ms innerem Analysefenster)

  EAC (256 Dimensionen, 32ms Analysefenster)

  Signal Energie (32ms. und 1000ms Analysefenster)

  Statistische Merkmale (64ms Analysefenster)

  OpenEar (default settings)

- Biophysiologie

  EKG wavelet und PQRST Deskriptor (5 Sekunden Analysefenster)

[11] Der Umfang der zur Verfügung stehenden Merkmale wird regelmäßig erweitert und somit auf dem aktuellen Stand der Forschung gehalten.

SCL LedaLab (default settings)

Alle Kanäle, statistische Merkmale, variable Analysefenster

- Pose

  Mittlere Beschleunigungen (120 Frames Analysefenster)

  Mittlere Geschwindigkeiten (120 Frames Analysefenster)

  Statistische Merkmale (30 und 120 Frames Analysefenster)

KAPITEL 3

ANNOTATION MULTIMODALER DATENBASEN

Neben der Gewinnung von multimodalem Datenmaterial spielt die Gewinnung von Grundwahrheiten, welche als Lehrersignal in maschinelle Lernprozesse eingehen, eine wesentliche Rolle für die Modellierung emotionaler Zustände [22, 245]. Neben der Datenqualität ist auch die Qualität der Grundwahrheiten eine wesentliche Größe für einen späteren Klassifikationserfolg [247, 205]. Die Grundwahrheiten können im Falle von affektiven Korpora auf drei grundsätzlich verschiedene Arten gebildet werden [126]. Zum einen können emotionale Zustände geschauspielert werden, die Grundwahrheiten sind dabei leicht zu ermitteln und bedürfen lediglich einer geringfügigen Überprüfung der schauspielerischen Qualität. Datensätze dieses Charakters wurden im vorangegangen Kapitel 2 erläutert und ihre Schwächen aufgezeigt. Alternativ kann die Grundwahrheit durch einen experimentellen Ablauf erzeugt werden. Hierbei ist ein Experiment zur Datenerzeugung so gestaltet, dass der Proband gezielt in die gewünschten emotionalen Zustände versetzt wird. Die Grundwahrheit ist dabei durch den experimentellen Abschnitt und dessen Design gegeben. Datensätze, welche dieses Paradigma zur Gewinnung von Grundwahrheiten verwenden, wurden ebenfalls im vorangegangenen Kapitel erläutert. Bei beiden Varianten ist die Annotation zumeist von geringem Aufwand, da lediglich überprüft werden sollte, ob die Induktion oder das Schauspiel erfolgreich und überzeugend war. Dies geschieht zumeist durch einfache Wahrnehmungstest, bei denen mehrere Experten das Datenmaterial betrachten, und im Falle übereinstimmender Meinung eines Erfolges, das vorgegebene Label als erfolgreich induziert oder geschauspielert bestätigen. Hierfür sind zumeist keine aufwendigen Werkzeuge notwendig, da Segmentgrenzen, Intensitäten (sofern vorhanden) und Grundwahrheiten bereits vorgeben sind. Zu beachten ist allerdings die Tendenz menschlicher Annoteure, systemseitige Vorschläge eher zu akzeptieren als abzulehnen. Auch erfolgt die Überprüfung im Falle von multimodalen Daten nur anhand einzelner Modalitäten, zumeist Einkanal Video und Audio. Dabei wird impliziert, dass andere Modalitäten ebenfalls erfolgreich induziertes Verhalten zeigen, ohne dies zu überprüfen.

Anders sind die Voraussetzungen bei frei erzeugten Datensätzen, bei denen die Probanden zwar emotional beeinflusst werden, aber nicht davon ausgegangen wird, dass alle Probanden dieselben emotionalen Zustände erreichen. Diese Datensätze sind aufgrund ihrer freien Struktur und der Möglichkeit einen MMI Kontext im Experiment zu verwenden wesentlich wertvoller für die Entwicklung qualitativ hochwertiger Emotionsklassifikatoren in der Companion-Technologie. Der Annotationsaufwand ist dabei allerdings erheblich höher. Hier ist es erforderlich, dass mindestens ein, besser mehrere Experten das gesamte Datenmaterial durchsehen und vollständig manuell annotieren. Das experimentelle Design sorgt für eine erhöhte Wahrscheinlichkeit des Auftretens von bestimmten emotionalen Zuständen zu bestimmten Experimentalzeitpunkten (Kontextwissen). Auch ein mögliches Selbst-Rating des Probanden ist als Hilfe für den Annoteur vorhanden, ersetzt dessen Einschätzung jedoch nicht. Auch sollte die Annotation auf allen Modalitäten separat erfolgen, da Probanden nicht immer in allen Modalitäten Expressionen zeigen (Zum Beispiel zeigt sich Lächeln nur im Video Kanal, während eine geänderte Stimmlage nur im Audiokanal zu finden ist.). Die manuelle Annotation erlaubt zudem eine Annotation auf verschiedenen Abstraktionsniveaus. So können zum Beispiel abstraktere Labels wie *allgemein erregt* oder *zeitweilig frustriert* auf längeren Zeitskalen annotiert werden, während kurze mimische Expresionen wie Lächeln bzw. Disskurspartikel wie „hä“ oder „hmm“ auf wesentlich kürzeren Segmenten annotiert werden können.

Im optimalen Fall lässt sich so die bestmögliche Gewinnung von Trainingsdaten mit einer optimalen Grundwahrheit kombinieren, wodurch ideale Ergebnisse mit Hinblick auf einen Einsatz in realen Umgebungen zu erwarten sind. Problematisch bei der manuellen Annotation ist jedoch der enorme Zeitaufwand, verglichen mit Grundwahrheiten, entnommen aus Schauspiel oder Experimentaldesign. Ein weiteres Problem ergibt sich dadurch, dass der eingesetzte Annotationsprozess Einfluss auf das Resultat der Annotation haben kann, und so zu elementar falschen oder veränderten Grundwahrheiten führen kann [125]. In Abschnitt 3.1 wird vertieft auf die Problematik manueller Annotation eingegangen. Um den Zeit- und Kostenaufwand zu senken, können aktive Lernverfahren aus der Familie der teilüberwachten Lernverfahren eingesetzt werden, bei denen der Annoteur durch Unterbreitung guter systemseitiger Vorschläge unterstützt wird. Weiter können aktive Lernverfahren unter Verwendung der Spärlichkeitseigenschaft Daten so selektieren, dass nicht der gesamte Datensatz annotiert werden muss um einen optimalen Klassifikator zu trainieren. Abschnitt 3.3 beschreibt die Methodik uni- und multimodaler Lernverfahren eingesetzt im Kontext affektiver MMI Daten.

Diese aktiven Lernverfahren müssen dabei in Annotationswerkzeuge eingebettet sein, welche den zugrundeliegenden methodischen Prozess gegenüber dem Annoteur graphisch repräsentieren. Der Entwicklung eines Annotaionswerkzeuges kommt dabei eine zentrale Bedeutung zu, da es im optimalen Fall ein wesentliches Bindeglied zwischen Datenaufnahme, Datenana-

lyse, Datenannotation, Merkmalsdarstellung, Interpretation von Resultaten und Bewertung von prototypischen Versuchsanordnungen darstellt. Da in der Literatur bisher kein solches Werkzeug bekannt ist, wurde im Rahmen dieser Arbeit das Annotationswerkzeug ATLAS realisiert, welches im letzten Abschnitt 3.5 dieses Kapitels beschrieben ist. In diesem Werkzeug wurden auch die entwickelten aktiven Lernverfahren eingebettet.

3.1. Stand der Forschung - Annotation

3.1.1. Verfügbare Annotationswerkzeuge

Zu Beginn dieser Arbeit, im Sommer 2011, wurde der Problemstellung der Annotation emotionaler Daten in der wissenschaftlichen Gemeinschaft eine eher geringe Bedeutung zuteil. Die verwendeten Datensätze beschränkten sich weitgehend auf geschauspielerte Inhalte. Erste Datensätze deren Grundwahrheit durch ein spezielles experimentelles Design erhoben wurden, waren in frühen Entstehungsphasen. Sofern der Induktionserfolg überhaupt in Frage gestellt und überprüft wurde, geschah dies unter Verwendung von einfachen Medienplayer zur Wiedergabe von Videos, spezielle Werkzeuge wurden nicht verwendet. Mit Aufkommen erster Datensätze deren Inhalt freier gestaltet wurde, wurde das Thema der Annotation diskutiert. Daraus entstanden erste Annotationswerkzeuge, welche zumeist auf spezielle Annoatationsaufgaben zugeschnitten waren. Auch wurde zu diesem Zeitpunkt nicht berücksichtigt, dass der Prozess der Annotation und die Gestaltung des Werkzeuges Einfluss auf die Annotationsresultate haben kann. Im Folgenden werden Werkzeuge vorgestellt, welche unter anderem bei der Annotation der Datensätze, welche in Kapitel 2 vorgestellt wurden, Verwendung fanden. Diese unterscheiden sich dabei in Werkzeuge welche auf genau eine Annotationsaufgabe eines Datensatzes zurecht geschnitten wurden, und Werkzeuge die den Anspruch haben verallgemeinerte Annotation zuzulassen. Die Liste der untersuchten Werkzeuge ist dabei aus Lizenzierungsgründen auf frei verfügbare Werkzeuge beschränkt. Es ist davon auszugehen, dass Firmen, welche an Produkten mit emotionalen Erkennungskomponenten arbeiten (wie Google, Microsoft, IBM, Facebook oder Amazon) über eigene Werkzeuge verfügen, welche der wissenschaftlichen Öffentlichkeit nicht zugänglich sind. So wurde im Rahmen der ICMI 2013 Einblick in ein Microsoft internes Werkzeug ermöglicht. Dieses befand sich in Entwicklung. Ideen, welche innerhalb der wissenschaftlichen Gemeinschaft diskutiert wurden, waren darin bereits teilweise umgesetzt. Der Einsatz von aktiven Lernverfahren war in der Planung dieses Werkzeuges als Vision vorgesehen jedoch nicht realisiert. Eine öffentliche Zugänglichkeit wurde aus unternehmensstrategischen Gründen verweigert.

Feeltrace

Bei diesem Werkzeug handelt es sich um ein kontinuierliches Annotationswerkzeug zur simultanen Annotation in zwei Dimensionen, bezeichnet als *Aktivierung* und *Evaluation* [53, 198]. Diese Dimensionen sind dabei als Achsenkreuz auf dem Bildschirm dargestellt. Die Evaluation ist dabei horizontal dargestellt und wird als positive oder negative Richtung bezeichnet. Die Aktivierung wird orthogonal dazu dargestellt und wird mit aktiv und passiv bezeichnet. Das verwendete Emotionsmodell ist dabei ähnlich zur VA Darstellung des um Dominanz reduzierten VAD Modells. Der Annoteur bewegt nun mittels Maus oder Joystick einen Punkt in diesem 2D Koodinatensystem, während er ein zu annotierendes Video- oder Audiosegment betrachtet. Hieraus resultiert eine kontinuierliche Trajektorie der Expertenbewertung des emotionalen Zustandes des Probanden.

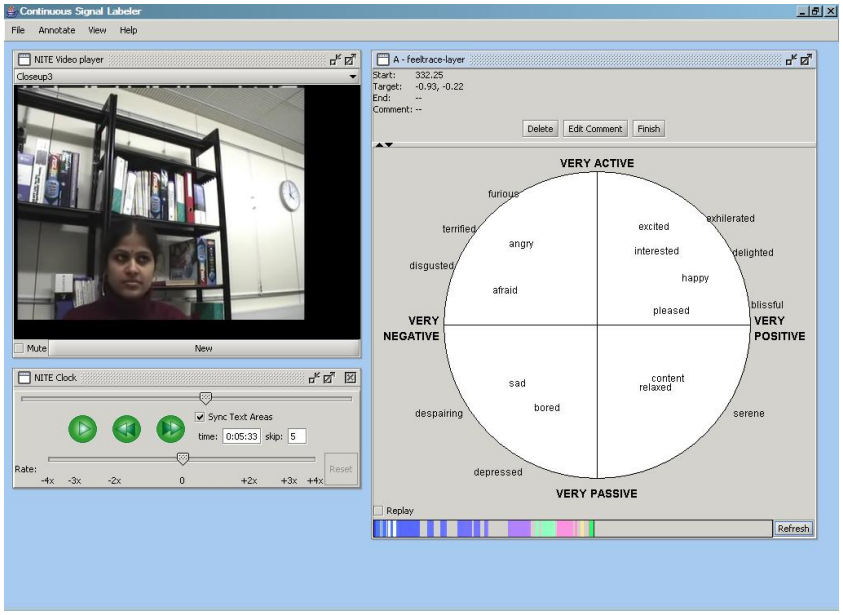

Abbildung 3.1.: Feeltrace Screenshot. Links oben: Videowiedergabe, Links unten: Abspielkontrolle, Rechts: VA Annotationsfläche.

Vorteilhaft an diesem Werkzeug ist eine schnelle Annotation mit abspielgeschwindigkeit des Videos. Fraglich ist dabei ob ein Annoteur kognitiv in der Lage ist zwei Dimension zeitgleich in Echtzeit zu annotieren. Zudem startet die Annotation jeweils im neutralen Zentrum der Annotationsfläche, dies entspricht im allgemeinen nicht dem tatsächlichen initialen emotionalen Zustand des Probanden. Typischerweise ist zu beobachten, dass Annoteure erst nach einiger Vorlaufzeit den Zustand des Probanden einschätzen können und beginnen die Trajektorie zu formen. Auch ist mit einem generellen zeitlichen Versatz der Trajektorie gegenüber dem tatsächlichen emotionalen Zustand zu rechnen. Dieser setzt sich aus der Zeit welche der Annoteur benötigt eine Änderung beim Probanden wahrnehmen zu können und der Zeit um

den Punkt korrekt im zwei dimensionalen Raum zu positionieren zusammen. Beide Phänomene sind bei der Annotatioon der kontinuierlichen AVEC Challenge zu beobachten[127].

GTrace

Das General Trace Program (GTrace) ist die Weiterentwicklung und somit Nachfolger von Feeltrace [55]. Es wurde im wesentlichen zur Annotation der HUMAINE Datanbank verwendet. Es erlaubt Annoteuren ebenfalls einfache Videos abzuspielen und während des Abspielens in zwei Dimensionen zu annotieren. Es beinhaltet über 50 vorbereitete Skalenvorlagen, erlaubt aber auch die Definition eigener Skalen. Während der Annotation sind typischerweise nur zwei Skalen zeitgleich in Verwendung.

Dieses Werkzeug löst somit die VA Dimensionsbeschränkungen von Feeltrace auf und erlaubt generelle kontinuierliche Annotation. Die wesentliche und elementare Problematik des neutralen Startpunktes, des Zeitversatzes aufgrund Wahrnehmungs- und Reaktionszeit des Annoteurs, und der Beschränkung auf eine Video Modalität bleiben jedoch unverändert bestehen.

Annotation der AVEC Datensätze

Feeltrace und GTrace fanden bei der Annotation der AVEC Datensätze Anwendung [223, 263, 189]. Bei der Erzeugung der kategoriellen Labels der ersten Challenge wurde das kontinuierliche Label mittels Schwellwert in diskrete Labels umgewandelt. Die Wahl des Schwellwertes ist dabei ein kritisches Kriterium, da die zugrunde liegende kontinuierliche Annotation zwar scheinbar auf einer absoluten Skala vollzogen wurde. Tatsächlich neigen Annoteure allerdings dazu innerhalb der gegebenen Grenzen eine relative Annotation zum vor kurzem wahrgenommen Teil der Daten zu beschreiben. Somit ist ein fester Schwellenwert nicht plausibel. Zudem sind die entstehenden binären Labels insbesondere im Bereich des Schwellwertes teils stark zerklüftet. Dies ist ebenfalls wenig plausibel.

Bei den späteren Challenges wurde das zugrundeliegende kontinuierliche Label als Lehrersignal zur Verfügung gestellt. Dies beinhaltet allerdings die zuvor beschriebenen Unzulänglichkeiten der Feeltrace Annotation. Kenntnis über den Annotationsprozess und Berücksichtigung der Eigenschaften des Korrelationskoeffizenten, welcher als Gütemaß für die Challenge herangezogen wurde, führt dazu, dass die bestmögliche Vorhersage im Testdatensatz erzeugt werden konnte, ohne die Daten überhaupt zu betrachten [127]. Fraglich ist an dieser Stelle also, in wie weit die im Rahmen von Challenges gewonnenen Erkenntnisse überhaupt auf prototypische Systeme in realen Bedingungen anwendbar sind.

ikannotate

Das von Siegert et. al., am Lehrstuhl Kognitive System der Otto-von-Guericke-Universität Magdeburg, entwickelte Werkzeug ikannotate bietet im wesentlichen die Möglichkeit Audiomaterial zu transkribieren und emotional zu annotieren [16]. Die Transkription erfolgt nach GAT1/2, die Annotation ist mit den Modellen Basisemotionen nach Ekman, Geneva Emmotion Wheel nach Plutchik oder VAD gemäß SAM möglich. Abbildung 3.2 zeigt einige Screenshots des Werkzeuges, die Bilder sind der Download Seite von ikannotate entnommen[1].

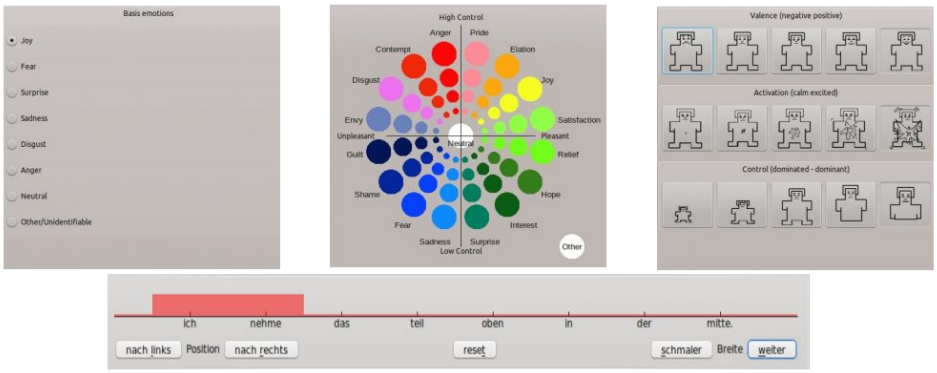

Abbildung 3.2.: ikannotate Screenshot. Oben: Annotation nach Eckman, Plutchik und SAM. Unten: Transkription und Emotionsintensität. Bildquelle: `http://www.iikt.ovgu.de/Lehrstühle+und+Fachgebiete/KS/Forschung/ikannotate.html` (Stand: 3. Juni 2019)

ikannotate besitzt seinen Schwerpunkt im Bereich der akustischen Transkription. In einer aktuellen Version wurde die Medienwiedergabe zwar auf Einkanal Videos erweitert, die Annotationsparadigmen jedoch nicht verändert, wodurch die wesentliche Anwendung auf die Transkription und Annotation von in Video enthaltem Sprachmaterials beschränkt bleibt. Im Bereich der akustischen Transkription stellt dieses Werkzeug ein probates Mittel dar. Durch die Verwendung der gängigsten Emotionsmodelle kann ein grundlegender Umfang von Annotationsschemata abgedeckt werden. Eine Annotation von MMI ausgerichteten Datensätzen mit freien und aktuelleren Emotionsmodellen ist jedoch zum gegenwärtigen Zeitpunkt nicht möglich.

ANNEMO

Der Versuch die zugrundeliegende technischen und infrastrukturellen Anforderungen zu reduzieren, um somit eine breitere Masse an potentiellen Annoteuren zu erreichen, liegt der Idee

[1] `http://www.iikt.ovgu.de/Lehrstühle+und+Fachgebiete/KS/Forschung/ikannotate.html` (Stand: 3. Juni 2019)

von ANNEMO zugrunde[2] [75]. Dieses Werkzeug ist als Web Applikation realisiert (Siehe Abbildung 3.3), wodurch eine größere Anzahl von Annoteuren zeitgleich auf einem Datenbestand operieren kann und dabei ihre eigenen Geräte (Laptops, Smartphones oder Tablets) verwenden können. Die Annotation ist zudem auf eine Dimension pro Annotationsdurchlauf begrenzt. Dies reduziert neben dem technischen Aufwand insbesondere die kognitiven Anforderungen an den Annoteur. Durch Einsatz simpler webbasierter Annotatiosnverfahren wird mit diesem Werkzeug eine Multirater Annotation praktisch realisierbar. Die Annotation ist allerdings auf einfache Videos begrenzt.

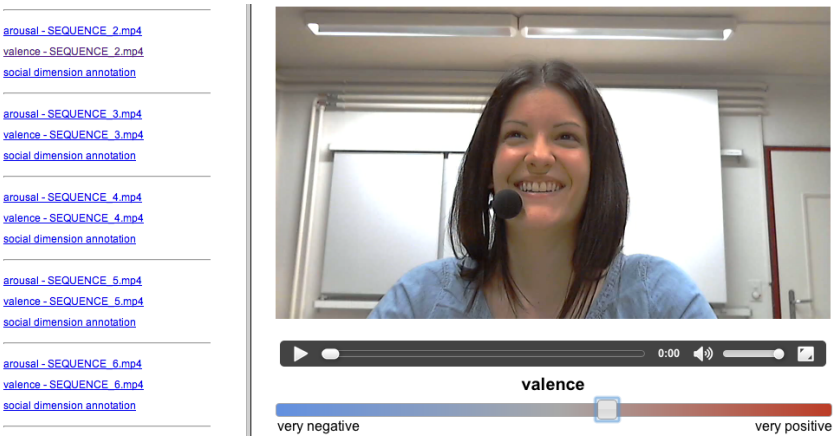

Abbildung 3.3.: ANNEMO Screenshot. Links: Videoauswahl, Rechts: Videowiedergabe mit darunter liegendem Annotationsregler. Bildquelle: `https://github.com/ilyabo/annemo` (Stand: 3. Juni 2019)

ANVIL

Bei ANVIL handelt es sich um ein verhältnismäßig umfangreiches Annotationswerkzeug, mit Schwerpunkt Video Annotation, entwickelt von Kipp et. al. [133][3]. Bis zur Erscheinung von ATLAS war es das einzige Werkzeug, welches frei definierbare Annotationsschmata in mehreren Ebenen erlaubte, eine strikte Bindung an klassische Emotionsmodelle entfällt somit. Die Annotation erfolgt vollständig graphisch in Form von diskreten Labeln, welche als Blöcke in der Zeitleiste dargestellt werden. Entgegen den vorherigen Werkzeugen bietet es die Möglichkeit in der Zeitleiste beliebig zu springen, Segmente erneut wiederzugeben oder die Wiedergabe zu pausieren. Somit überwindet ANVIL die Wahrnehmungsverzögerung des Annoteurs und die Problematik des neutralen Starts. Zudem kann der Annoteur an eventuell schwierigen Stellen die Annotationsgeschwindigkeit an seine Fähigkeiten und die Komplexität der Aufgabe anpassen. Ursprüngliche wurde ANVIL in den frühen 2000ern zur Annotation

[2] `https://github.com/ilyabo/annemo` (Stand: 3. Juni 2019)
[3] `http://www.anvil-software.org/` (Stand: 3. Juni 2019)

von Gesten eingesetzt, konnte aufgrund seiner Mächtigkeit aber auch in frühen MMI basierten Annotationen und Datensätzen eingesetzt werden. ANVIL besitzt zudem als erstes Werkzeug eingeschränkte Möglichkeiten Daten mit anderen Werkzeugen auszutauschen. ANVIL kann damit in seiner aktuellsten Version 5.0, veröffentlicht im Jahr 2011, als zu diesem Zeitpunkt mächtigstes Werkzeug angesehen werden. Abbildung 3.4, welche auf der ANVIL Projekt Website veröffentlicht ist, zeigt eine typische Annotationsansicht des Werkzeuges. Deutlich zu erkennen ist die Annotation in mehreren ggf. hierarchisch verbunden Ebenen.

Nachteilig ist die Begrenzung der Modalitäten auf Einkanal Audio- und Videoströme, welche zudem in nur wenigen, nicht zeitgemäßen Codes unterstützt werden. Ebenfalls sind keine interaktiven Schnittstellen nach außen vorgesehen, welche zum Einsatz aktiver Annotatinsverfahren zwingend notwendig sind. Die Annotation ist in der gegenwärtigen Version 5.0 auf kategorielle diskrete Labels begrenzt, auch ein Import von Merkmalen oder Klassifikationsresultaten ist nicht vorgesehen.

Abbildung 3.4.: ANVIL Screenshot. Oben (von Links nach Rechts): Hauptfenster mit Projektinformationen, Videoansicht, Labelinformationen. Unten: Labelspuren. Bildquelle: `http://www.anvil-software.org/` (Stand: 3. Juni 2019)

3.1.2. Defizite verfügbarer Annotationswerkzeuge

Zusammenfassend ist festzustellen, dass keines der bisher verfügbaren Werkzeuge, die Annotation umfangreicher Modalitäten, welche über klassische bimodale Audio- und Videodaten hinausgeht, unterstützt. Diese Beschränkungen sind mit Blick auf aktuellste Datensätze und der zu erwartenden steigenden Anzahl von messbaren Modalitäten durch zukünftige Techno-

logien nicht länger zeitgemäß. Zudem liegt der Fokus aller vorgestellten Werkzeuge rein auf der Annotation. Keines der Werkzeuge sieht einen Einsatz im gesamten Entwicklungszyklus von der Datenerhebung und früher explorativen Analyse über Annotation, Merkmalsgewinnung und Resultatbewertung hin zu prototypischen Systemen vor. Da in fast allen Stadien verschiedene Abstraktionsniveaus der Daten im Mittelpunkt stehen, wäre dies durchaus sehr wünschenswert. Lediglich die beiden Werkzeuge gTrace und ANVIL erlauben eine Annotation in weitgehend frei definierbaren Annotationsschemata, wodurch diese Werkzeuge immerhin eine der grundsätzlichen Voraussetzungn zum Einsatz auf MMI relevanten Daten erfüllen.

Steigender Umfang von Datenmaterial (Aufnahmedauer) und Modalitäten macht eine klassische, vollständig manuelle Annotation zudem kaum mehr realisierbar. Erste aktive Lernverfahren, welche hier Abhilfe schaffen können, befinden sich in der methodischen Erforschung. Keines der Werkzeuge verfügt über Schnittstellen und Strukturen, welche eine technische Einbindung externer Algorithmen erlaubt. Hierdurch ist eine praktische Erprobung oder tatsächlicher Einsatz erfolgversprechender, maschinell unterstützter aktiver Annotationsmethoden nicht möglich. Die wesentlichen Defizite bestehender Werkzeuge sind zusammengefasst:

- Beschränkte und unflexible Modalitäten Unterstützung
- Fokusierung auf klassische Emotionsmodelle, geringe Unterstützung freier Annotationsschemata
- Nichtverwendbarkeit im gesamten Entwicklungszyklus
- Vollständiger Verzicht auf externe Schnittstellen
- Fehlende Möglichkeiten zum Einsatz maschinell unterstützter Annotationsverfahren

Diese und weiter Defizite werden durch ATLAS behoben.

3.1.3. Aktive Lernverfahren für multimodale affektive Daten

Die Erzeugung eines informativen und repräsentativen Subsets der Datengesamtheit ist wesentlicher Bestandteil aktiver Lernverfahren [48, 29]. Allgemein setzen aktive Lernverfahren eine interaktive Kooperation mit einem menschlichen Experten voraus, welcher zumeist als Orakel bezeichnet wir. Aktive Lernalgorithmen wählen aus einer Menge (Pool) von ungelabelten Daten, anhand einer Selektionsstrategie, diejenigen Datenpunkte (Instanzen), oder Zeitabschnitte aus, welche als besonders informativ erscheinen. Diese werden dann, mit oder ohne Labelvorschlag, dem Experten zur Ansicht präsentiert. Der Experte vergibt mittels seiner Expertise ein Label (Im Falle von systemseitigen Vorschlägen durch bestätigen oder widerrufen). Annotierte Daten

werden anschließend im Pool der annotierten Daten abgelegt. Mittels der neu gewonnen Informationen werden die Selektionsparameter adaptiert und eventuell Klassifikatoren verändert, anhand derer erneut Instanzen gewählt werden, welche eine möglichst hohe Verbesserung bezogen auf das Gesamtziel versprechen. Dieser iterative Zyklus wird durch Abbruchkriterien wie Anzahl annotierter Daten, erreichter Qualitätsschranken oder geringe Modelländerung zwischen zwei Iterationen beendet. Dieser generelle Vorgang ist in Algorithmus 1 abstrakt als Pseudocode dargestellt.

Algorithmus 1 Poolbasiertes aktives Lernen

1: $\mathcal{U} = \{x_1, x_2, ...x_n\}$ ▷ Ungelabelte Menge
2: $\mathcal{L} = \{\}$ ▷ Gelabelte Menge
3: $Klassifikator\ k$
4: **while** $Abbruchkriterium \neq true$ **do**
5: $x \leftarrow Auswahlstrategie(\mathcal{U})$ ▷ Wähle ungelabelte Instanz
6: $y \leftarrow Orakel(x)$ ▷ Erfrage Label bei Annoteur
7: $\mathcal{L} \leftarrow \mathcal{L} \cup (x, y)$ ▷ Füge gelabelte Instanz zu gelabelter Menge
8: $\mathcal{U} \leftarrow \mathcal{U} \setminus x$ ▷ Entferne gelabelte Instanz aus ungelabelter Menge
9: **end while**
10: $k \leftarrow train(\mathcal{L})$ ▷ Trainiere Klassifikator
11: **return** k

Diese Verfahren werden in verschiedensten Anwendungen bereits erfolgreich eingesetzt [234, 215]. Im Bereich der Affekt- und Emotionserkennung wurde in der Literatur bisher allerdings noch kein erfolgreicher Einsatz beschrieben. Ein wesentlicher Grund dafür dürfte, neben der sehr jungen Disziplin der Affekterkennung, insbesondere angewendet auf gering expressive Daten, auch in der enormen Komplexität dieser Daten liegen. Bisher wurde davon ausgegangen, dass selbst die Verwendung aller verfügbaren Daten nicht ausreichend ist um optimale Klassifikatoren zu gewinnen.

Aktive Lernverfahren gehören zur Gruppe der teilüberwachten Lernverfahren, welche einen hybriden Ansatz aus überwachtem und unüberwachtem lernen verfolgen [43]. Multiview Lernverfahren erweitern und verbessern die Lernverfahren dadurch, dass sie einen Datensatz oder Datenstrom aus verschiedene Blickwinkeln betrachten [162]. Diese Blickwinkel sind durch verschiedene, aus einem Ursprungssignal erzeugte, Merkmale manifestiert [277]. Die zugrundeliegende Signalquelle und somit auch Grundwahrheit entspringt dabei einer einzigen Quelle. Listing 1 wird dabei um einen weiteren Datenpool A und B erweitert. Die Auswahl kann gleichmäßig oder asymmetrisch auf die Pools verteilt sein.

Im Falle multimodaler emotionaler Daten ist diese Signalquelle nicht durch einen einzelnen Sensor (oder mehrerer Sensoren), der einen zentralen physikalischen Prozess abtastet, gegeben. Vielmehr ist das zentrale Element die abstrakte Wahrnehmung des Menschen über sein eigenes

affektives und emotionales Empfinden. Dieser emotionale Zustand verursacht verschiedenste körperliche Reaktionen, welche sich in den unterschiedlichen Modalitäten unterschiedlich niederschlagen und somit messen lassen. Die Grundwahrheit ist dabei nur sehr vage gegeben, im Wesentlichen hat jede Modalität eine eigene Grundwahrheit, welche nicht zwingend über die gesamte Signaldauer mit der Grundwahrheit anderer Modalitäten übereinstimmen muss. Neben dem nichtauftreten und auftreten eines Labels, sind auch im Konflikt stehende Labels nicht ausgeschlossen. Dadurch unterscheidet sich das multimodale aktive Lernen von den klassischen Ansätzen des multiview aktiven Lernens, dessen Ansätze immer eine gemeinsame unumstößliche Grundwahrheit annehmen. Die Verfahren der Literatur sind somit nicht ohne Anpassung übertragbar. Zudem ist die Qualität emotionaler Grundwahrheiten gegenwärtig immer noch Gegenstand der Diskussion innerhalb der Forschergemeinschaft. Die meisten aktiven Lernverfahren nehmen allerdings an, dass die Qualität der Grundwahrheit nahezu perfekt ist. Arbeiten, welche aktives Lernen mit unscharfen Lehrersignalen kombinieren, sind bis heute weitgehend unbekannt. Settles et al. beschreibt in seiner Arbeit einige dieser elementaren Fragestellungen [234].

Aktive Lernszenarien

Aktive Lernverfahren unterscheiden sich im wesentlichen in der Beschaffenheit vorhandener Daten aus denen ausgewählt werden kann, und der Selektionsstrategie mit der relevante Instanzen bestimmt werden. Beide sind zentrale Bestandteile des aktiven Lernens [234, 235]. Abbildung 3.5 zeigt mögliche Selektionsstrategien.

Synthesebasierte Verfahren (Membership Query Synthesis) ist eine der ältesten Methoden [6]. Dabei wählt der Selektor Instanzen aus dem Eingaberaum durch Synthese prototypischer Instanzen, welche dem Orakel vorgelegt werden. Die gewählten Instanzen werden dabei weitgehend synthetisch generiert, anstelle von tatsächlich gemessene Eingangsgrößen, der Eingaberaum ist somit nicht beschränkt. Synthetische Daten sind in vielen Fällen für den Experten erheblich komplizierter zu annotieren, tragen durch Auswahl optimaler Punkte im Eingaberaum jedoch einen wesentlich höheren Informationsgehalt und sind somit als relevanter einzustufen. Im Falle affektiver Daten ist diese Methode nicht anwendbar, da schon real erfasste Daten schwierig zu annotieren sind. Kann die Annotation hingegen automatisiert werden (was einem seltenen Spezialfall entspricht) ist diese Methode enorm mächtig. King et al. verwenden einen Automaten zur Durchführung biologischer Testverfahren anhand ausgewählter synthetischer biologischer Proben [132].

Strombasierte Verfahren (Stream-based Selective Sampling) nutzten die Eigenschaft (zumeist zeitlicher) Ordnung von Eingabedaten aus [9]. Die Auswahl relevanter Instanzen ist

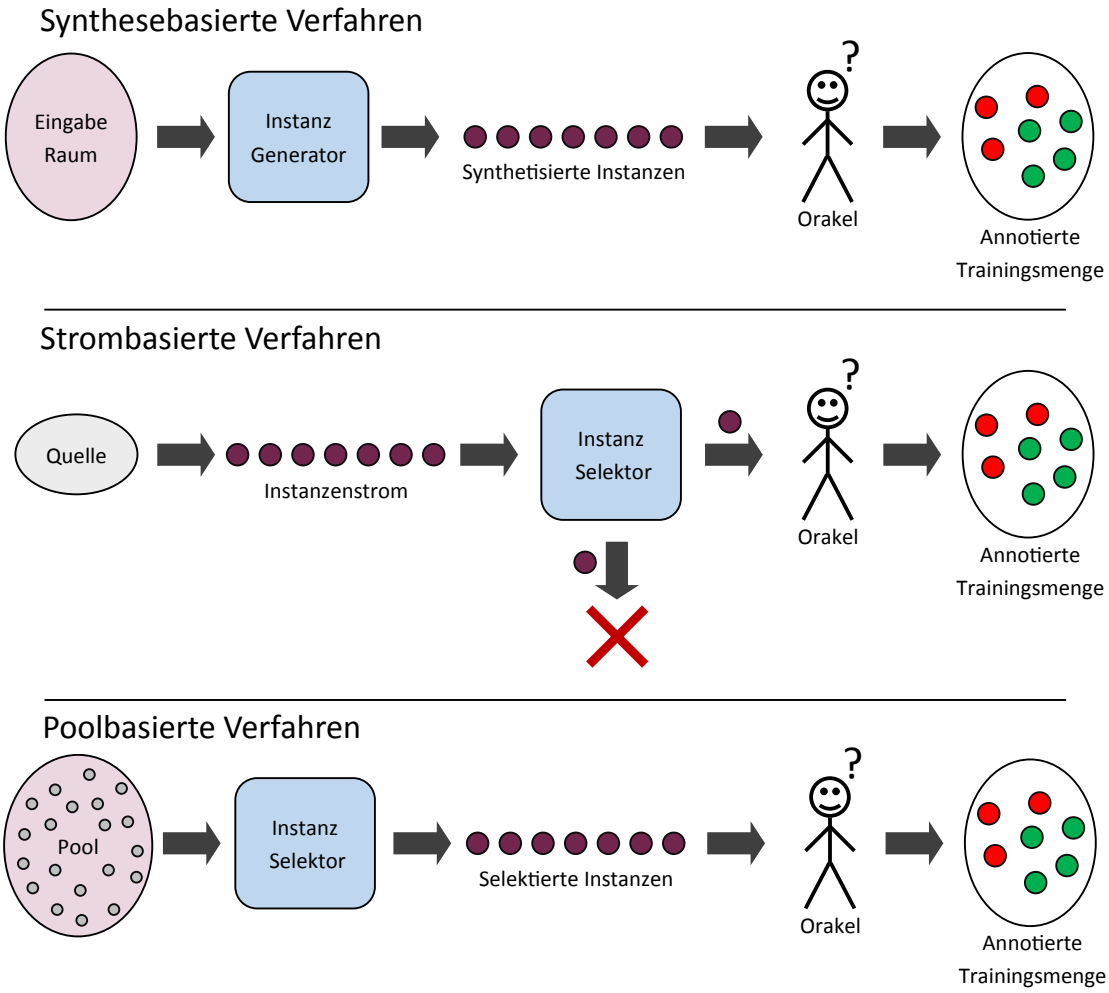

Abbildung 3.5.: Grundlegende aktive Lernverfahren. Abhängig von den Eigenschaften der Datengewinnung werden verschiedene Verfahren eingesetzt.

hierbei nur in der Reihenfolge ihres Auftretens möglich. Es wird angenommen, dass das Beobachten (Messen) einer Instanz keine Kosten im Sinne des Algorithmus verursacht, das Speichern einer Instanz jedoch Kosten verursacht. Jede Instanz wird dem Selektor präsentiert, dieser entscheidet ob eine kostenintensive Speicherung und Annotation sinnvoll erscheint, oder die Wahrscheinlichkeit entsprechend hoch ist in Kürze interessantere Instanzen beobachten zu können. Die Daten affektiver Szenarien sind zeitlich geordnet und könnten somit grundsätzlich mit dieser Methode untersucht werden. Da die Annotation und das Klassifikatortraining jedoch nicht in Echtzeit, sondern anhand aufgezeichneter Daten erfolgt, stehen dem Selektor zu jedem Zeitpunkt alle Daten einer Aufzeichnung zur Verfügung. Rücksprünge in die Vergangenheit sind also möglich, wodurch die Beschränkungen und Vorteile dieser Methode obsolet sind. Im Einsatz späterer, im realen Umfeld eingesetzter Companion-Systeme, kann diese Methode relevant werden, da hier Daten nur klassifiziert und anschließend gelöscht werden. Lassen sich

durch strombasierte Auswahl interessante Zeitpunkte erfassen, welche zudem annotierbar sind (z.B. durch Rückfragen beim Nutzer), können diese einen großen Beitrag zur Verbesserung und Individualisierung eines bestehenden Systems leisten.

Poolbasierte Verfahren (Pool-based Sampling) nutzen große Mengen von geordneten oder ungeordneten Instanzen aus, wie sie in den meisten Anwendungen erfassbar sind, auf welche wahlweise zugegriffen werden kann. Dies trifft auch auf affektive Datensätze zu, welche als großer aber beschränkter, zeitlich geordneter, Pool gesehen werden können aus denen der Selektor frei wählen kann [139]. Der Pool ändert sich dabei während der gesamten aktiven Lernphase nicht, Instanzen werden zumeist durch Greedy Verfahren bestimmt [63]. Die Selektionsverfahren können dabei vielfältig sein, sortieren den Datensatz jedoch immer anhand einer der folgenden Strategien. Zumeist wird in absteigender Reihenfolge des zu erwartenden Informationsgehaltes der im Pool enthaltenen Instanzen sortiert und die informativsten k Instanzen durch das Orakel annotiert.

Unsicherheitsselektion (Uncertainty Sampling) ist sicherlich die einfachste Selektionsstrategie [139]. Dabei werden alle vorhanden Instanzen mit einem initialen Klassifikator bewertet. Es werden diejenigen Instanzen vom Selektor an das Orakel gegeben, welche die größte Unsicherheit aufweisen, da anzunehmen ist, dass diese den größtmöglichen Beitrag zur Verbesserung des zur Selektion eingesetzten Klassifikators beitragen (Im Falle von Support Vektor Maschine (SVM) liegen diese Instanzen nahe an der Entscheidungsgrenze). Diese Instanzen sind auch für einen Experten eventuell schwierig zu klassifizieren, es wird davon ausgegangen das der Experte diese Aufgabe immer zuverlässig erfüllen kann. Das Verfahren kann für Mehrklassenproblem durch margin sampling [200] oder durch Einsatz von Entropiemaßen als Unsicherheitsmaß [241] erweitert werden.

Mehrheitsselektion (Query-by-Committee) erweitert das vorherige Verfahren durch Einsatz eines Klassifikator Ensembles zur Selektion [238]. Dabei werden Klassifikatoren mit unterschiedlichen Hypothesen auf den bereits klassifizierten Daten trainiert. Jeder dieser Klassifikatoren wählt dann die für seine Hypothese interessantesten Instanzen aus. Durch den Selektor wird diejenige Instanz gewählt welche die geringste Übereinstimmung aufweist. Das Verfahren unterscheidet sich dabei in die Varianten *wahrscheinlich nahezu richtige* aktive Lerner und *agnostisch* aktive Lerner. Erstere nehmen an, dass eine perfekte Hypothese existiert, während Zweitere eine perfekte Hypothese nur annäherungsweise approximieren.

Modelländerungsselektion (Expected Model Change) wählt diejenige Instanzen, welche die größte mögliche Modelländerung zu erwarten lässt [237]. Das Modell ist dabei zunächst beliebig abstrakt, es muss nicht zwingend dem Klassifikator des Selektors entsprechen. In vielen Fällen wird der Klassifikator jedoch als Modell verwendet. Die Idee des Verfahrens legt zu Grunde, dass eine möglichst große Änderung des Modells möglichst schnell zum optimalen Modell führt (Ähnlich dem Gradienten Abstiegsverfahren). Selektierte Instanzen müssen daher möglichst viel neue Information in das Modell einbringen.

Fehlerreduktionsselektion (Expected Error Reduction) versucht durch Auswahl geeigneter Instanzen den Fehler des Modells zu verringern [192]. Dabei wird die Instanz derart selektiert, dass das damit angepasste Modell einen möglichst geringen Fehler zum optimalen Modell aufweist. Dieses Selektionsverfahren ist somit ähnlich zu dem vorherig beschriebenen Verfahren. Da die Fehlerreduktion Ziel aller Klassifikatoren ist, ist die Fehlerreduktionsselektion sehr interessant, jedoch sehr rechenintensiv. Diese Rechenintensivität begründet sich in der Notwendigkeit einer Vorhersage zukünftiger erwarteter Fehler für jede mögliche denkbare Selektion. Zusätzlich muss in jeder Iteration das Modell vollständig neu trainiert werden, sowie jedes mögliche zukünftige Modell welches seinerseits wieder über den gesamten Pool iteriert. In seiner Reinform ist das Verfahren somit nicht anwendbar, es sei denn, es existieren leicht berechenbare Approximationen zukünftiger Modelle.

Varianzreduzierendeselektion (Variance Reduction) löst einige Probleme der Fehlerreduktionsselektion mittels indirekter Reduktion des Generalisierungsfehlers durch Minimierung der Ausgabevarianz [47]. Trotz Reduktion des Rechenaufwandes ist dieses Verfahren immer noch wesentlich rechenintensiver als einfache Verfahren und daher nur für kleine Datenmengen handhabbar.

Dichteselektion (Density-Weighted Methods) verwendet die den Daten zugrunde liegende Dichteverteilung im Eingaberaum als zusätzliches Selektionsmerkmal neben dem Maß der Unsicherheit [236]. So wird sichergestellt, dass die selektierte Menge repräsentativ für die Gesamtheit der Daten bleibt. Wurde zum Beispiel zur Selektion ein Klassifikator, basierend auf trennenden Hyperebenen, verwendet (z. B. SVM), gehen eventuell Informationen verloren (Es wird nur entlang der Hyperebene selektiert) welche es zum Beispiel abschließend nicht mehr erlauben Dichteschätzer, wie Gaussian Mixture Modelss (GMMs), zu trainieren. Auch wird verhindert, dass durch wiederholtes auswählen unsicherer Elemente nahe der Entscheidungsebene, keine strukturbeschreibende Instanz, welche prototypisch für eine Klasse ist, zum reduzierten Datensatz hinzugefügt wird. Dichteselektion gewichtet daher das Unsicherheitsranking zusätzlich mit der durchschnittlichen Ähnlichkeit der Instanz zu bereits gewählten Instanzen und interinstanzen Abständen zu anderen Instanzen der Eingabemenge.

Wie bereits erwähnt, haben gegenwärtige affektive Datensätze Eigenschaften welche nur ein, eventuell geordnetes, Poolbasiertes Samplingverfahren erlauben. Hinsichtlich der Selektionsstrategien haben alle in der Literatur bekannten Verfahren erheblich Nachteile. Unsicherheitssampling ist grundsätzlich möglich. Aufgrund der auch für Experten enorm schwierig zu klassifizierenden selektierten Instanzen, ist hier nicht mir Erfolgen zu rechnen. Es ist eher damit zu rechnen, dass durch Annotationsfehler viele falsche Labels in den Trainingsdatensatz geraten, bzw. dass selektierte Instanzen im wesentlichen aus verrauschten und klassenüberlappenden Bereichen des Datensatzes entnommen werden. Merheitsselektionen weisen ähnliche

Probleme auf. Modelländerungsselektion, Fehlerreduktionsselektion und Varianzreduktionsselektion sind aufgrund des schwer zu beschreibenden oder gänzlich fehlenden Modells, und der enormen zu erwartenden Rechenintensität auf Datensätzen der Größe typischer affektiver Datensätze ungeeignet. In den folgenden Abschnitten werden angepasste Selektionsverfahren beschrieben, welche Eigenschaften affektiver Datensätze ausnutzen, und so eine modifiziert Variante der Unsicherheitsselektion beschreiben. Es wird gezeigt, dass diese Verfahren die Fähigkeit besitzen im uni- und multimodalen aktiven Lernen in affektiven Daten eingesetzt zu werden.

3.2. Aufwand manueller Annotation

Bestehende Datensätze konnten, aufgrund ihres überschaubaren Umfangs, bisher manuell annotiert werden sofern ihre Labels nicht bereits durch Induktionsparadigmen gegeben sind. Die Annotation erfolgt dabei zumeist auf einem relativ hohen Abstraktionsniveau. Das bedeutet es wurden größere zusammenhängende Zeitabschnitten kategorische Labels zugeordnet. Die zeitliche Dichte der Labels war dabei durch die konzentrierte Form der Induktion als eng zu betrachten. Zudem wird auf eine manuelle Annotation kleiner Ereignisse, wie *Lächeln*, *Augen zwinkern*, *Stirn runzeln* etc. verzichtet, obwohl diese speziellen kurzen Momente für einige Anwendungen von großem Informationsgehalt und somit Bedeutung sind. Lediglich die FACS basierte Annotation bildet hier eine Ausnahme, ist allerdings enorm zeitaufwendig.

In modernen Datensätzen wie dem uulmMAC ist die Dichte emotionaler Ereignisse an das spärliche Auftreten realer Bedingungen angepasst. Die Annotationszeit verkürzt sich dadurch jedoch nicht, da weiterhin das gesamte Datenmaterial manuell untersucht werden muss. Die Datensätze wachsen in ihrem Umfang stetig an, wodurch der Annotationsaufwand etwa linear mit der Datensätzelänge steigt. Eine manuelle Annotation wird dadurch zunehmend kostenintensiver.

Eine empirische Untersuchung anhand der bimodalen manuellen Annotation von neun ausgewählten Probanden des uulmMAC zeigt, dass der zeitliche Aufwand für eine manuelle Annotation im Bereich des ca. 10 fachen der aufgenommen Dauer pro Modalität und Experten liegt. Wünschenswert ist die Annotation von mehreren Experten in allen Modalitäten. Daraus ergibt sich ein Aufwand zur Annotation von einer Stunde trimodalem Datenmaterials durch drei Experten von 45 bis 135 Stunden. Im Falle umfangreicher Datensätze wie dem uulmMAC würde dies einem erwartenden Annotationsaufwand von ca. 6.750 Stunden entsprechen. Angenommen dies würde von studentischen Hilfskräften ohne Abschluss ausgeführt (Stundensatz 9,58 €, Stand Oktober 2017), entspräche dies einem Kostenaufwand von ca. 65.000 €. Ist

Probanden ID	Daten Länge (Minuten)	davon Sprache (Minuten)	Ann. Aufw. Sprache (Minuten)	Ann. Aufw. Video (Minuten)	Faktor Audio	Faktor Sprache	Faktor Video
09	36	2	85	340	2,4	43	9
11	31	5	150	315	4,8	30	10
12	32	3	105	405	3,3	35	13
13	35	2	70	330	2,0	35	9
17	32	3	95	275	3,0	32	9
19	32	2	85	285	2,7	43	6
28	33	3	90	245	2,7	30	8
29	34	2	80	325	2,4	40	7
31	35	2	95	365	2,7	48	8
Schnitt:	**33,6**	**2,7**	**95,6**	**320,6**	**2,9**	**37,3**	**9,0**

Tabelle 3.1.: Aufwand der manuellen Annotation neun ausgewählter Probanden des uulmMAC. Durchschnittlich wird das neunfache der Aufnahmedauer für die Annotation einer Modalität benötigt.

wie z.B. im Falle von FACS eine gesonderte Qualifikation des Annoteurs notwendig, ist mit deutlich höheren Stundensätzen zu rechnen.

Der uulmMAC wurde daher nur in reduzierter Form vollständig manuell annotiert. Dabei wurden neun repräsentative Probanden von den Versuchsleitern ausgewählt und in den Modalitäten *frontale Kamera* und *Headset Akustik* annotiert. Die Annotation erfolgte dabei kategoriell in den Kategorien *emotionale Reaktion* und *neutrales Verhalten*. Die Segmentierung wurde im Audiokanal automatisch anhand einer Signalenergieschwelle durchgeführt, im Videokanal vollständig manuell. Die Annotation wurde von studentischen Hilfskräften übernommen, welche mit der Materie vertraut sind, jedoch keine gesonderte Ausbildung durchlaufen haben. Tabelle 3.1 zeigt den Zeitaufwand der Annotation und den daraus resultierenden Aufwandsfaktor.

3.3. Methoden aktiver Lernverfahren für multimodale affektive Daten

Entgegen der klassischen manuellen Annotation ist die aktive Annotation durch aktive Lernverfahren unterstützt. Die Grundidee besteht dabei darin, durch geeignete Verfahren Instanzen des Datensatzes so auszuwählen, dass mit möglichst wenig Daten ein guter Klassifikator erzeugt werden kann. Diese wenigen gewählten Instanzen müssen dabei durch den Experten annotiert werden. Abschnitt 3.1.3 beschreibt die gängigsten gegenwärtigen Verfahren des aktiven Lernens. Diese lassen sich nur mit geringem Erfolg direkt auf affektive Daten anwenden. Der wesentliche Grund liegt dabei in der hohen Komplexität dieser Daten, was selbst bei klassischem Lernen auf der Datengesamtheit nur zu geringen Erkennungsraten führt (siehe

Kapitel 4). Im Folgenden wird ein neu entwickeltes Verfahren vorgestellt, welches sich die speziellen Eigenschaften affektiver Daten zunutze macht und sehr gute Ergebnisse bei erheblich verringertem Annotationsaufwand liefert.

Affektive Datensätze, welche in realistischen MMI Situationen aufgenommen wurden, beinhalten nur wenige emotionale Momente. Tabelle 3.2 zeigt das Verhältnis von Instanzen, welche emotional behaftet sind, zu Instanzen von emotional neutralem Verhalten anhand, eines als Vorstudie zum uulmMAC eingesetzten Datensatzes. Tabelle 3.3 zeigt dasselbe Verhältnis, welches sich aus den manuell annotierten Probanden des uulmMAC ergibt. In allen Datensätzen und Modalitäten ist zu erkennen, dass in der Regel deutlich weniger als 10% der Daten emotionales Verhalten erkennen lassen. Für alle Lernprozesse ist es dabei unerheblich ob der Proband in den als neutral gekennzeichneten Instanzen Emotionen verspürt oder nicht. Da diese Emotionen nicht ersichtlich sind, können diese auch nicht von Sensoren erfasst werden und somit nicht zur Klassifikation verwendet werden. Elementar ist also die Auffindung emotional behafteter Sequenzen in einem überwiegend neutralen Datensatz.

Unter der Annahme, dass die Daten überwiegend einer neutralen Klasse zugeordnet sind, ist davon auszugehen, dass die neutrale Klasse den Schwerpunkt der Daten definiert und emotionale Instanzen einen erkennbar größeren Abstand zu diesem Schwerpunkt aufweisen als neutrale Instanzen. Die Richtung dieses Abstandes ist eventuell durch die zugrundeliegende Emotion gegeben. Die Richtungseigenschaft wird nicht verwendet, Ziel ist zunächst nur die Segmentierung emotionaler Events, welche zum Training von Klassifikatoren herangezogen werden können.

In den nächsten Abschnitten wird zunächst eine Vorstudie, welche auf unimodalen Daten mit inhaltlicher Ähnlichkeit zum uulmMAC aufweisen, beschrieben. Dieser Proof of Concept wurde bereits während der Konzeptionsphase des uulmMAC durchgeführt. Anschließend wird eine Erweiterung des Proof of Concept Ansatzes vorgestellt, welche unimodal auf dem uulmMAC durchgeführt wurde. Diese unimodalen Ansätze werden schließlich zu einem multimodalen Algorithmus erweitert. Abschließend seien einige Worte zum Ausblick und weiteren Möglichkeiten der entwickelten Verfahren gegeben.

3.3.1. Unimodale aktive Lernverfahren affektiver Daten

Unimodales aktives Lernen wird in dieser Arbeit nur anhand der Modalitäten Audio und Video untersucht, ohne methodisch auf diese begrenzt zu sein. Dies ist darin begründet, dass selbst Experten für Biophysiologie anhand grafischer Darstellungen biophysiologischer Daten, ohne Kontextwissen, keine belastbare Aussage über den Zusammenhang mit Emotionen und

Probanden ID	012	015	017	023	026	030
Emotionen	108	135	66	208	429	665
Neutral	3387	3365	4149	3406	3409	4489
Summe	3495	3500	4215	3614	3838	5154

Tabelle 3.2.: Datenverteilung der manuell annotierten, visuellen Merkmalsinstanzen der Vorstudie

Affekten machen können. Modalitäten mit dieser Eigenschaft können nur in unüberwachten Verfahren oder in Kombination mit anderen Modalitäten herangezogen werden.

Unimodale Vorstudie und Proof of Concept

Das im Folgenden vorgestellte Verfahren wurde in [253] publiziert. Es nutzt die spärlich auftretenden emotionale Expressionen in Videodaten des Datensatzes von Schüssel et. al. [218]. Abbildung 3.6 zeigt eine Ansicht dieses Datensatzes in ATLAS. Zu erkennen ist eine typische kurze emotionale Reaktion des Probanden. Die roten Labels entsprechen emotionalen Momenten, gelb den neutralen Phasen. Tabelle 3.2 zeigt die Anzahl emotionaler und nicht emotionaler Instanzen. Unter Ausnutzung der emotionalen Spärlichkeit wird angenommen, dass das Training einer One class support vector machine (OCSVM) [212] auf allen Daten im wesentlichen die neutrale Klasse modelliert und Instanzen außerhalb der Entscheidungshülle mit hoher Wahrscheinlichkeit eine Emotion darstellen. Diese werden dem Annoteur zur Überprüfung präsentiert und vom ungelabelten Datensatz entfernt. Anschließend wird die Hülle der OCSVM adaptiert und neue Instanzen ausgewählt. Dieser Prozess wird solange iteriert bis kein Daten mehr vorhanden sind[4]. In jeder Iteration wird ein Klassifikator auf allen bereits gelabelten Daten trainiert und mit dem Baseline Klassifikator, welcher auf den gesamten Daten (manuell gelabelt) trainiert wurde, verglichen. Die Resultate zeigen, dass nur ein geringer Bruchteil der Daten notwendig ist um ein vergleichbares Resultat zu erzielen. Aufgrund des enormen Annotationsaufwandes, welcher zum Vergleich notwendig ist, wurde der Proof of Concept auf sechs Probanden beschränkt.

Die OCSVM wurde von Schölkopf et. al. als Erweiterung der klassischen binären SVM vorgeschlagen [212]. Sie verwendet Instanzen einer einzelnen Klasse um eine Entscheidungsfunktion zu definieren, welche in einer kleinen Region den Wert $+1$ annimmt in dem der Großteil der Instanzen liegt, und -1 außerhalb dieser Hülle. Instanzen außerhalb der Hülle haben eine geringe Wahrscheinlichkeit zur gesuchten Klasse zu gehören. Die Instanzen werden in den Hilbert Raum $\mathbb{H}$ abgebildet und anschließend durch eine Hyperebene, mit minimalem Abstand,

[4] In der praktischen Anwendung muss ein geeignetes Abbruchkriterium verwendet werden

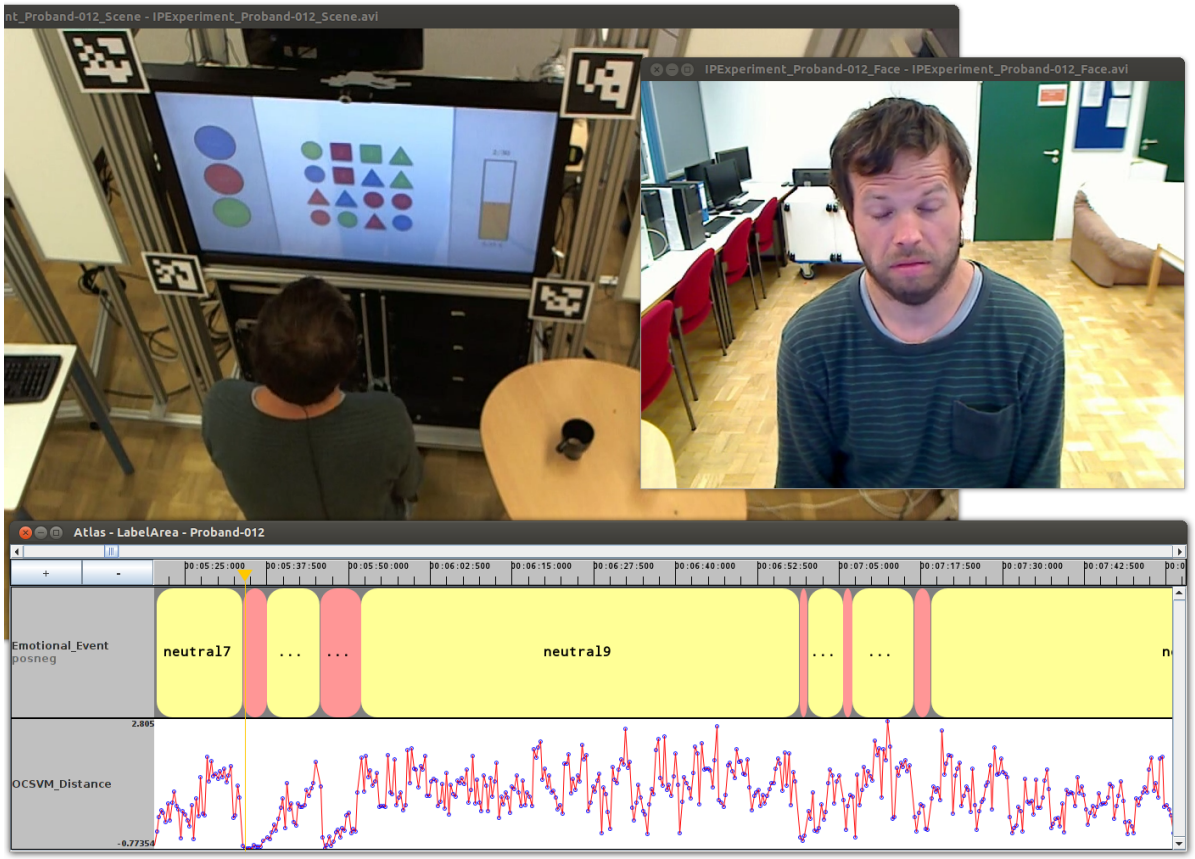

Abbildung 3.6.: Prof of Concept des aktiven Lernverfahrens anhand einer Vorstudie des uulmMAC. Oben: Videoansicht mit Aufgabenstellung ähnlich der des uulmMAC. Unten: Annoation Emotionaler Ereignisse in rot und neutraler Bereiche in gelb (erste Zeile). Wahrscheinlichkeit das eine Instanz zum Neutralmodell gehört (Niedrige Werte bedeuten hohe Emotionswahrscheinlichkeit).

von der Zielklasse getrennt. Die Klassifikation einer Instanz bezüglich der Zielklasse erfolgt durch die Seite zur Hyperebene auf der sich ein Punkt befindet.

Sei $\{x_i\}_{i=1}^{l}$ mit $x_i \in \mathbb{R}^d \ \forall i \in \{1,\ldots,l\}$ und $l \in \mathbb{N}^*$ ein Datensatz der Zielklasse. Weiter sei $\phi : \mathbb{R}^d \longrightarrow \mathbb{H}$ eine Funktion die eine Instanz in den Hilbert Raum $\mathbb{H}$ abbildet. Um den Datensatz vom Ursprung zu trennen muss das folgende quadratische Programmierproblem gelöst werden:

$$\begin{aligned} &\textit{minimize} \quad L(\omega,\xi,\rho) = \frac{1}{2}\|\omega\|^2 + \frac{1}{\nu l}\sum_{i=1}^{l}\xi_i - \rho, \\ &\textit{s.t.} \quad (\omega^T\phi(x_i)) \geq \rho - \xi_i, \xi_i \geq 0, \forall i \in \{1,\ldots,l\} \end{aligned} \tag{3.1}$$

Der Parameter $\nu \in (0,1]$ beschreibt eine obere Grenze des Anteils von Außenseitern zu einer Mindestanzahl von Supportvektoren. Das Optimierungsproblem wird durch den Lagrange Multiplikator α_i gelöst. Die Entscheidungsfunktion ist definiert durch:

$$f(x) = \textit{sign}(\sum_{i\in S}\alpha_i K(x_i,x) - \rho), \tag{3.2}$$

mit der Kernfunktion $K(x_i,x) = \phi(x_i)^T\phi(x)$ und der Supportvektor Indexmenge S.

Für die hier vorgestellte Anwendung wurde die Gauss Kernfunktion $K(x,x') = \textit{exp}(-\gamma\|x - x'\|)^2$ verwendet[5], der Parameter γ steuert die Form und den Radius der Entscheidungshülle. Daher beeinflussen sowohl ν als auch γ die Performanz der Entscheidungsfunktion. Da keine systematische Methode zur optimalen Bestimmung der Parameter existiert, muss dies halbautomatisch erfolgen. Da die OCSVM in ihrer Grundform keine Konfidenz berechnet, wird stattdessen stellvertretend die Distanz zur Hyperebene zur Instanz als Konfidenzmaß verwendet. Entsprechend sind große Distanzen ein Maß für hohe Sicherheit, kleine Distanzen nahe der Hyperebene entsprechen geringer Konfidenz.

Die Instanz Selektionsmethode ist eine Modifikation der Unsicherheitsselektion. Es werden diejenigen Instanzen zur manuellen Annotation ausgewählt, welche die größte Konfidenz haben, nicht zum Modell zu passen, also am weitesten außerhalb der umspannenden Hyperkugel liegen. Da diese am schlechtesten zur neutralen Klasse passen, handelt es sich dabei sehr

[5]Es ist jeder Kernfunktion welcher die Kerneleigenschaft erfüllt möglich. Sei X der Eingaberaum. Eine Abbildung $K\colon X \times X \to \mathbb{R}$ heißt Kernel, wenn es einen Skalarproduktraum $(F, \langle\cdot,\cdot\rangle)$ und eine Abbildung $\phi\colon X \to F$ in diesen Raum gibt mit: $K(x,y) = \langle\phi(x),\phi(y)\rangle \ \ \forall x,y \in X$

wahrscheinlich um spärlich auftretende emotionale Ereignisse. Dieses Vorgehen resultiert am schnellsten im Auffinden möglichst vieler emotionalen Ereignisse bei geringst möglichem manuellem Annotationsaufwand. Bei dem iterativen Verfahren werden pro Iteration mehrere Instanzen selektiert um den Berechnungsaufwand für das Adaptieren des Modells zu reduzieren. Die Anzahl der Iterationen ist auf einen festen Wert $N \in \mathbb{N}$ eingestellt.

Bei jeder Iteration wird ein OCSVM Klassifikator auf allen umgelabelten Instanzen trainiert, die Parameter ν und γ wurden zuvor anhand empirischer Erfahrungen, basierend auf einer Gittersuche, festgelegt. Mit diesem Klassifikator werden anschließend alle Instanzen reklassifiziert und die $k \in \mathbb{N}$ weitest außen liegenden Instanzen manuell annotiert und aus dem umgelabelten Pool entfernt. Zu Evaluationszwecken wird in jeder Iteration ein Klassifikator auf allen bereits gelabelten Instanzen trainiert und mit einem Klassifikator, welcher auf dem vollständig manuell annotierten Datensatz trainiert wurde, verglichen. Algorithmus 2 beschreibt diesen Vorgang in Pseudocode.

Algorithmus 2 One Class SVM aktives Lernverfahren

Voraussetzungen:
$T = \{x_i | x_i \in \mathbb{R}^d, i = 1, \ldots, l\}$ ▷ Trainingsmenge
$N \in \mathbb{N}, N \geq 1$ ▷ Anzahl Iterationen
$1 \leq k \leq l$ ▷ Anzahl gewählter Instanzen pro Iteration
ν, γ ▷ Parameter der OCSVM
Initialisierung:
$D_0 = T$ ▷ Ungelabelte Menge
$S_0 = \emptyset$ ▷ Gelabelte Menge
Iteration:
for $j = 1$ to N **do**
    Berechne OCSVM mit D_{j-1}: OCSVM$_{j-1}$
    Berechne Distanzen d_i für alle $x_i \in D_{j-1}$ durch OCSVM$_{j-1}$ auf D_{j-1}
    Annotiere die k größten Distanzen von OCSVM$_{j-1}$: k-FN$_{j-1}$
    $D_j \leftarrow D_{j-1} \backslash$ k-FN$_{j-1}$ ▷ Entferne gelabelte Instanzen von D_{j-1}
    $S_j \leftarrow S_{j-1} \cup$ k-FN$_{j-1}$ ▷ Füge gelabelte Instanzen zu gelabelter Menge hinzu
end for
Abschluss:
(D_N, S_N)
Trainiere binäre SVM auf gelabelter Menge S_N
Wende binäre SVM auf verbliebener Menge D_N an

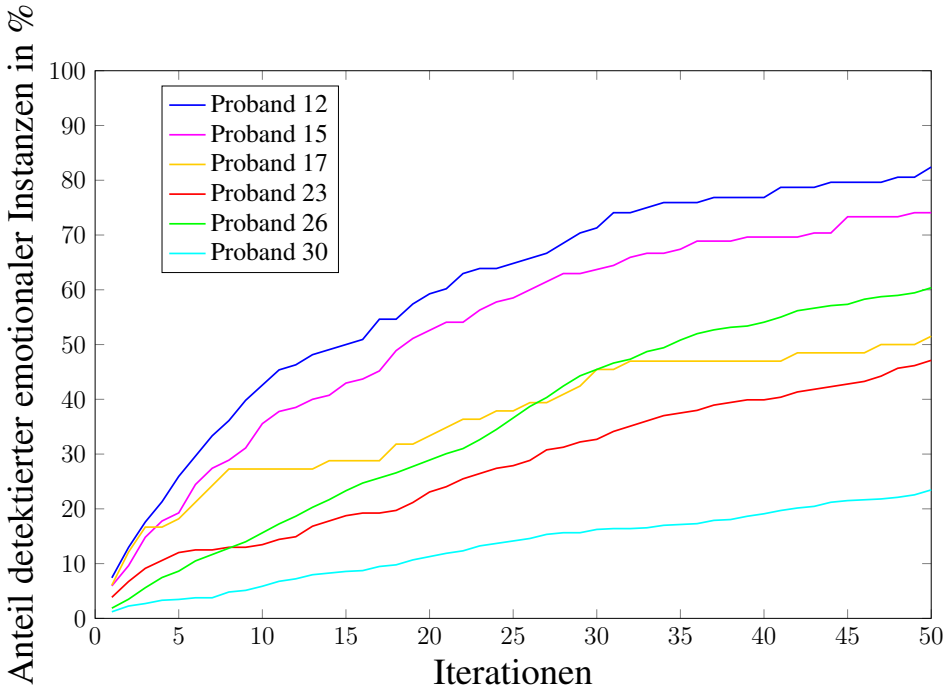

Abbildung 3.7.: Anzahl der Instanzen, welche als Emotion annotiert wurden nach jeder Iteration des aktiven Lernverfahrens. Pro Iteration wurden 10 Instanzen annotiert.

Da das meist verwendete Accuracy-Maß für stark unbalancierte Datensätze nur schlecht geeignet ist, wird in diesem und allen folgenden Ansätzen der Geometrischer Mittelwert (G-Mean) als Maß für die Güte der Klassifikatoren verwendet.

$$gmean = \sqrt{acc^{+} \times acc^{-}} \tag{3.3}$$

mit acc^{+} und acc^{-} den Erkennungsraten (Accuracy) der majoren und minoren Klasse.

Dieser stellt sicher, dass sowohl das Auffinden der spärlichen emotionalen Instanzen, als auch die Separierung der wesentlich größeren neutralen Klasse, gleichmäßig gewichtet ist.

Dieser Ansatz wurde auf dem Videokanal des als Vorstudie eingesetzten Datensatzes von Schüssel et. al. evaluiert. Die Klassendisbalance liegt in diesem Datensatz bei ca. 30 (Neutral) zu 1 (Events). Als Merkmal wurde das Local Binary Pattern - Tree Ortogonal Planes (LBP-TOP) beschrieben in Abschnitt 4.3.2, angewendet auf die Gesichtsregion, verwendet. Die Anzahl der aktiven Lerniterationen ist auf $N = 50$ mit jeweils $k = 10$ selektierten Instanzen festgelegt. Die Hyperparameter betragen $\nu = 0.25$ und $\gamma = 2$. Die Experimente wurden individualspezifisch für jeden Probenden durchgeführt. Abbildung 3.7 zeigt die Entwicklung der pro Iteration ausgewählten und als emotionale markierten Instanzen. Es ist zu erkennen, dass die Anzahl

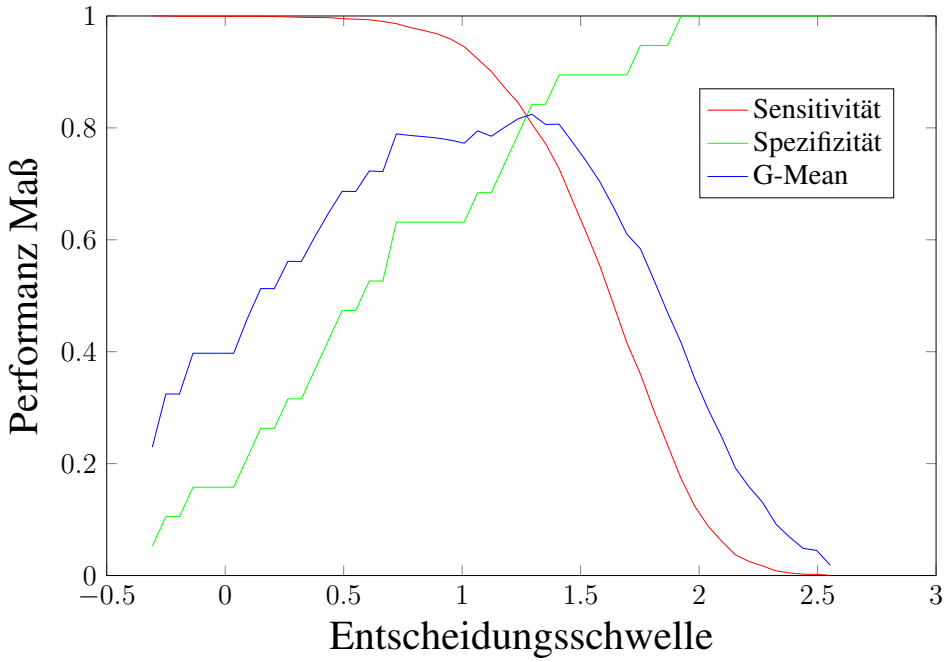

Abbildung 3.8.: Klassifikationsperformanz in Abhängigkeit der Schwelle. Probandenspezifisch (hier gezeigt für Proband 12). Local Binary Pattern (LBP) Merkmale, binäre SVM trainiert auf der reduzierten Trainingsmenge.

der gefunden emotionalen Instanzen zunächst schnell wächst und deutlich über der Apriori Wahrscheinlichkeit liegt. Nach einiger Zeit sinkt die Rate der gefunden emotionalen Instanzen ab, bleibt jedoch höher als bei einer zufälligen Auswahl. Dabei ist zu beachten, dass in jeder Iteration viele emotionale Instanzen aus dem umgelabelten Pool entfernt werden. Die Apriori Wahrscheinlichkeit für das Auffinden einer solchen Instanz sinkt somit in jeder Iteration, dennoch werden eine große Anzahl dieser Instanzen aufgefunden. Da die Parameter ν und γ in diesem Proof of Concept Ansatzes nicht variiert werden, ändert sich das neutrale Modell in jeder Iteration nur gering. Dadurch rücken die verbleibenden noch zu findenden Instanzen immer näher an die Hyperkugel des neutralen Models und beginnen sich zunehmend mit neutralen Instanzen zu vermischen.

Weiter zeigt Abbildung 3.7 das für Proband 12 insgesamt 89 emotionale Instanzen nach 50 Iterationen gefunden wurden, dies entspricht 82% aller vorhanden emotionalen Instanzen bei nur 14% des gesamten Datensatzes. Ähnliches Verhalten ist auch für Proband 15 erkennbar, hier wurden 100 von 135 emotionale Instanzen gefunden, dies entspricht 74% bei ebenfalls nur einer Annotation von 14% der Daten. Klassifikatoren, welche auf diesen 14% der Daten trainiert wurden, erreichen dabei Klassiifkationsraten, welche mit denen von Klassifikatoren trainiert auf den gesamten Daten, vergleichbar sind. Der Proof of Concept ist somit erfolgreich, eine weitergehende Untersuchung unter Optimierung des Verfahrens wird im folgenden Abschnitt

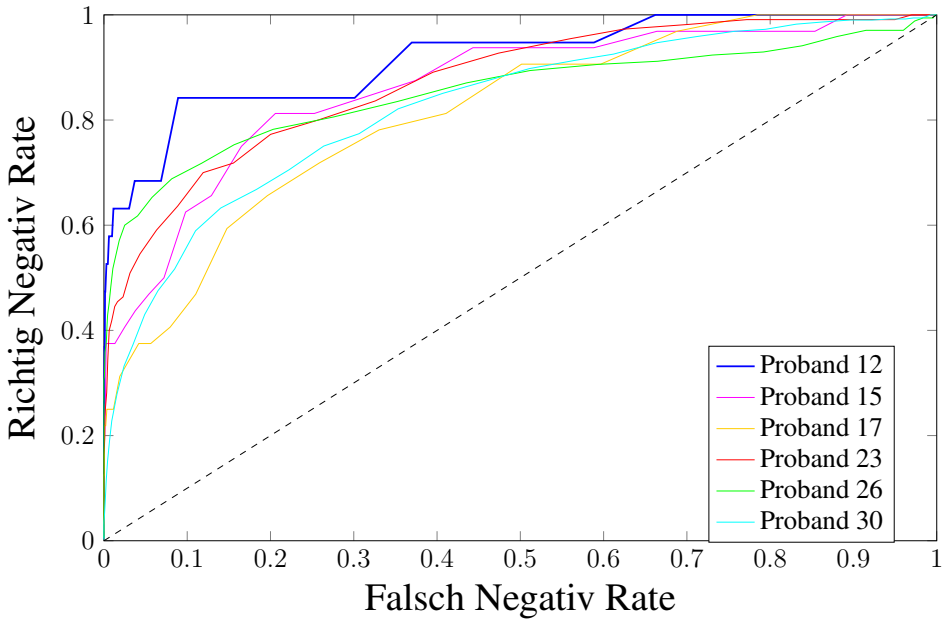

Abbildung 3.9.: Klassifikationsperformanz in Abhängigkeit der Schwelle als ROC Kurve für alle Probanden der Vorstudie. LBP Merkmale, binäre SVM trainiert auf der reduzierten Trainingsmenge.

unter Verwendung des uulmMAC beschrieben. Eine weitere kritische Größe ist die Wahl

der Konfidenzschwelle ab der eine Instanz als emotional klassifiziert wird. Auch hier wird als Gütemaß der G-Mean verwendet. Abbildung 3.8 zeigt die erreichte Klassifikatorgüte abhängig von einem gewählten Schwellwert, den eine klassifizierte Instanz erreichen muss um als emotionales Event eingestuft zu werden. Die Wahl dieses Schwellwertes beeinflusst die Klassifikationsperformanz erheblich, und ist daher eine kritische Größe. Es lassen sich in dieser Vorstudie maximale G-Mean Werte von ca. 0,8 erreichen. Abbildung 3.9 zeigt die Performanz der erreichten Modelle, individualspezifisch für jeden Probanden, dargestellt als ROC Verlauf, nach Durchlauf von 50 Iterationen.

Alle Modelle übertreffen einen zufälligen Klassifikator deutlich. Es ist davon auszugehen, dass bereits die meisten Instanzen, welche als Supportvektoren in Frage kommen, gefunden sind und ein binärer Klassifikator die Klassen bereits gut modellieren kann. Die Hinzunahme weiterer Instanzen liefert keinen relevanten Beitrag an Supportvektoren, die Hyperebene wird also nur noch geringfügig verändert. Es ist bereits früh eine gute Generalisierung erreicht. Diese Tatsache wird im Folgenden bei der Annotation des uulmMAC verwendet um das Problem einer nicht realisierbaren vollständigen manuellen Annotation zu umgehen. Aktive Lernverfahren

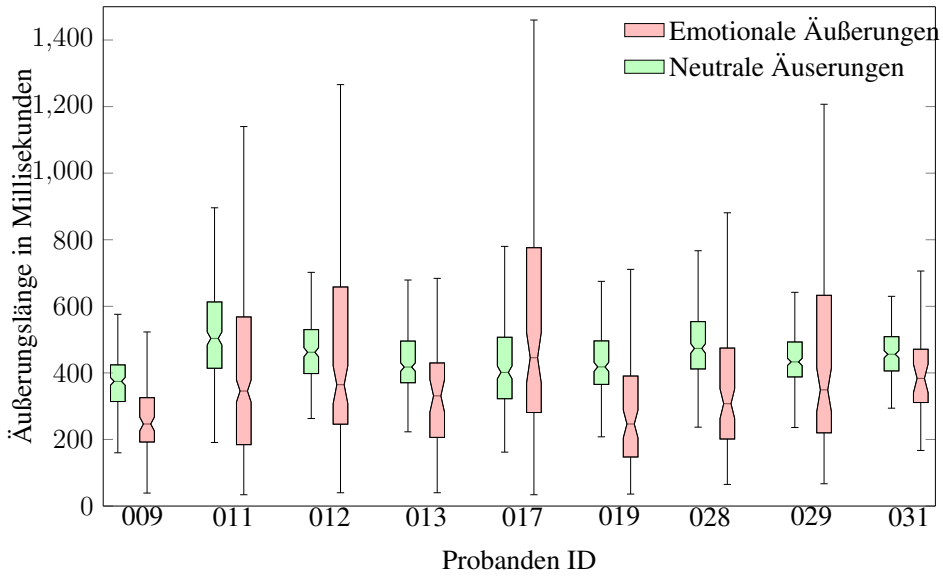

Abbildung 3.10.: Äußerungslängen bei manueller Annotation. Probandenabhänige Äußerungslänge für normale und emotionsbehaftete Äußerungen. Zu erkennen ist eine hohe Varianz zwischen den verschiedenen Probanden.

lassen sich so zum ersten mal, unter Ausnutzung der Spärlichkeitseigenschaft emotionaler Momente, auf realistische affektive Datensätze anwenden.

Audiobasiertes aktives Lernen

Nach Erhebung des uulmMAC wird das Proof of Concept Verfahren modifiziert und umfangreich auf einem affektiven Datensatz mit real vorkommenden emotionalen Elementen evaluiert. Das unimodale Verfahren für akustische Emotionen wurde in [255] publiziert. Für die Evaluation wurden neun Probanden ausgewählt und vollständig manuell annotiert (siehe Tabelle 3.1), auch für den uulmMAC zeigt sich das spärliche Auftreten von emotionalen Instanzen ähnlich der Vorstudie. Für den akustischen Kanal wurde eine automatische Segmentierung anhand der Sprachaktivierung (Signalenergie) durchgeführt [3], wodurch das zu untersuchende und annotierende Datenmaterial bereits reduziert wurde.

Basierend auf der manuellen Annotation wurde zunächst in Abbildung 3.10 die Länge sprachlicher Segmente abhängig von ihrer emotionalen Färbung für jeden Probanden untersucht. Zu erkennen ist, dass jeder Proband Unterschiede aufweist, abhängig von seinem emotionalen Zustand. Im Falle emotionaler sprachlicher Segmente sind diese etwas kürzer, weisen jedoch eine deutlich größere Varianz auf. Im wesentlichen zeigt die Untersuchung die großen individuellen Unterscheide auf, sowie große Varianzen. Die Länge einer Äußerung ist somit nicht gut geeignet um relevante Instanzen zu selektieren.

Probanden ID	009	011	012	013	017	019	028	029	031
Emotionen	350	2206	890	195	1036	297	439	632	196
Neutral	1033	2162	1574	1464	1330	1363	1851	1503	1555
Summe	1383	4368	2464	1659	2366	1660	2290	2135	1751

Tabelle 3.3.: Datenverteilung der akustischen Merkmalsinstanzen der manuell annotierten uulmMAC Probanden

Die mittlere Länge eines Segmentes beträgt ca. 100 bis 330 Millisekunden, das Analysefenster wird im Folgenden daher auf 215 Millisekunden festgelegt. Der Vorschub beträgt 65 Millisekunden. Tabelle 3.3 zeigt die daraus resultierende Verteilung der Instanzen auf die emotionalen Klassen. Da auch hier eine Disbalance zu beobachten ist, wird erneut der G-Mean als Gütemaß verwendet. Für jede Instanz, bestehend aus einem 215 Millisekunden Fenster, wird ein Merkmalsvektor basierend auf einem Subfenster von 25 Millisekunden mit 10 Millisekunden Vorschub berechnet. Dieser beinhaltet Linear Predictive Coding (LPC), Relative Spectral Perceptual Linaear Prediction (RastaPLP), Mel Frequency Cepstral Coefficent (MFCC) und F0, resultierend in 65 Merkmalsdimensionen bei 20 Subfenstern. (20×65 Merkmals Matrix) Die Extraktion wurde mit *openSMILE*[6] durchgeführt [77]. Zur Reduktion der Matrix auf einen Merkmalsvektor pro Analysefenster wurden 14 statistische Funktionen auf die Matrix angewendet (Mittelwert, Median, Standard Abweichung, Maximum, Minimum, Varianz, Schiefe, Wölbung, erstes und zweites Quartile, Inter Quartile Abstand, 1%-Perzentile und 99%-Perzentile, Abstand von 1% zu 99% Perzentile). Daraus resultiert ein 910 dimensionaler Merkmalsvektor je Analysefenster, welcher in die folgenden Algorithmen eingeht.

Ein wichtiger Schritt des aktiven Lernens ist die Bewertung des generierten Datensatzes, um die Relevanz der extrahierte Merkmalsvektoren für die zu Grunde liegende Lernaufgabe sicherzustellen. Dazu wurden zunächst folgende drei Klassifikatoren ausgewählt, mit denen eine personenspezifische fünffache Block-Kreuzvalidierung [21] durchgeführt wird: Ein Random Forest mit fester Anzahl von 300 Bäumen, eine SVM mit radialem Basisfunktions Kernel, ein Multi Layer Perzeptron (MLP) mit einer versteckten Schicht mit 100 Neuronen und radialer Basis Transfer Funktion. Die Block-Kreuzvalidierung teilt den Datensatz in fünf zeitlich aufeinander folgende Stücke anstelle in fünf zufällige Partitionen. Dies ist notwendig, da die eingesetzten Merkmalsvektoren eine zeitliche Überlappung aufweisen. Eine (teilweise) Abhängigkeit der Test- und Trainingsdaten wird somit vermieden. Vor dem Klassifikationsprozess wird zunächst eine Hauptkomponentenanalyse (engl. Principal Component Analysis) (PCA) durchgeführt um die Dimensionalität zu reduzieren. Gewählt werden die ersten zehn Komponenten. Auf dem Trainingsset wird die Synthetic Minority Over-sampling Technique (SMOTE) [44] angewendet

[6] `http://audeering.com/technology/opensmile/` (Stand: 3. Juni 2019)

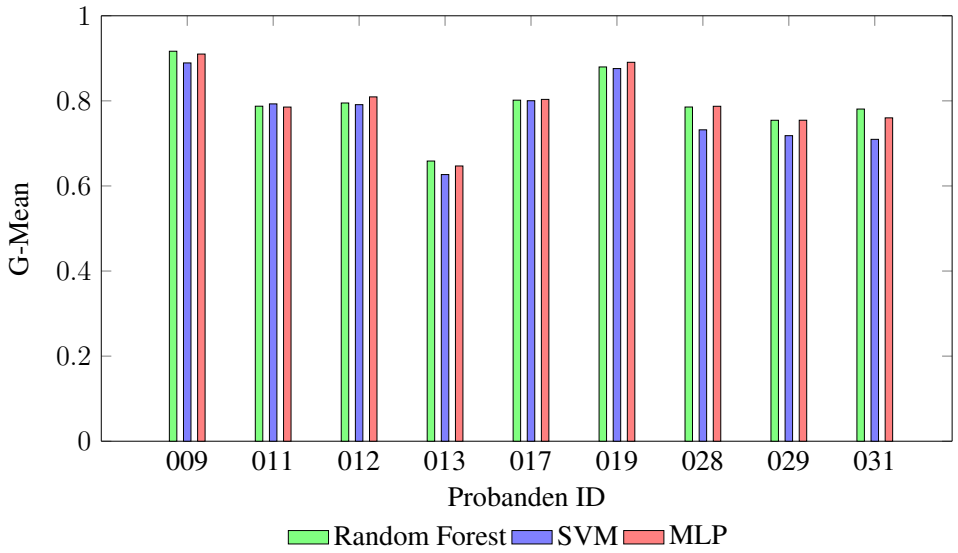

Abbildung 3.11.: Resultate des überwacht trainierten Klassifikators. Personenabhängige 5-fach Kreuzvalidierung. Die Wahl des Klassifikators hat nur unwesentlichen Einfluss auf die erreichbare Erkennungsrate.

um die Klassen zu balancieren. Die Klassifikationsergebnisse sind in Abbildung 3.11 dargestellt. Zu erkennen ist, dass der Unterschied der Klassifikatoren nicht signifikant ist. Das MLP weist eine minimal bessere Erkennungsrate auf als der Radom Forest und die SVM. Die Wahl des Klassifikators ist, bei geeigneter Parameterwahl, daher nicht elementar für die Anwendbarkeit auf realistische affektive Datensätze wie dem uulmMAC. Diese Klassifikationresultate gehen als Baseline Wert in den, im folgenden vorgestellten aktiven Lernalgorithmus ein. Ziel ist dabei diese Baseline Resultate zu erreichen und dabei eine möglichst kleine Untermenge von annotierten Daten zu verwenden, um so die Annotationskosten zu reduzieren.

Der vorgeschlagene Algorithmus basiert, wie die Proof of Concept Untersuchung, auf dem Pool basiertem Verfahren kombiniert mit Unsicherheitsselektion um die manuell zu annotierenden Instanzen zu bestimmen. Die vorgeschlagene Methode baut auf den zuvor publizierten Arbeiten [253] und [252] auf. Das Verfahren besteht ebenfalls darin, zuerst eine Teilmenge von Instanzen aus dem ungelabelten Datensatz so zu wählen und zu sortieren, dass daraus die Instanzen mit höchster Wahrscheinlichkeit für emotionalen Inhalt gewählt werden können. Die Selektion erfolgt dabei mittels zweier Selektoren deren Vorschläge anschließend mittels Rangaggregation zu einer Auswahl kombiniert werden. Eine Auswahl erfolgt mittels Modellbildung auf den ungelabelten Daten, unter der Spärlichkeitsannahme emotionaler Daten, ähnlich dem Proof of Concept Ansatzes. Zusätzlich wird auf den bereits annotierten Daten der vorherigen Iterationen eine SVM trainiert. Mittels dieser SVM werden dann die ungelabelten Daten klassifiziert und ebenfalls, anhand der Distanz zur Hyperebene, die Instanzen selektiert, welche mit

größter Wahrscheinlichkeit emotionalen Inhalt tragen. Nach erfolgter Rangaggregation werden die k Instanzen mit bestem Rang zur manuellen Annotation an den Annoteur gegeben.

Der vorgeschlagene Ansatz verwendet nicht länger die OCSVM des Vorversuchs, sondern erzeugt ein Ensemble von Supportvector data description (SVDD) [251] und wählt die informativsten Instanzen anhand einer Mehrheitsentscheidung (Voting) des Ensembles. Entgegen des von Thiam et. al. zunächst vorgeschlagen Ansatzes ist das Ensemble nicht zufällig erzeugt [252].

Basierend auf der Arbeit von Chang et al. [42] werden insgesamt $m \times n$ SVDD-Modelle generiert, indem m Werte für den Parameter C, gleichmäßig verteilt im Intervall $\left[\frac{1}{N}, 1\right]$ gewählt werden, wobei N die Anzahl der Samples in der ungelableten Menge ist, und n Werte für den Parameter γ des RBF-Kernels, gleichmäßig verteilt innerhalb des Intervalls $\left[\frac{1}{f}, 1\right]$, wobei f die Dimensionalität der Merkmalsvektoren ist. Auf diese Weise wird das Raster der möglichen Werte für beide Parameter abgedeckt und die Vielfalt im Ensemble ist gewährleistet. Darüber hinaus wird ein Schwellwert verwendet, um das generierte Ensemble auf Grundlage der Reklassifizierungsergebnisse jedes generierten Modells zu reduzieren. Modelle mit Ausreißer-Klassifikationsraten, die höher als der spezifizierte Schwellenwert sind, werden verworfen. Anschließend werden die Instanzen der ungelabelten Menge in absteigender Reihenfolge der Abstimmungsanzahl des Ensembles geordnet. Durch die Ensemble Methode wird der Nachteil der Proof of Concept Methode durch zuvor manuell festgelegte Parameter gelöst.

Ein alternativer Ansatz besteht darin ein GMM auf den ungelabelten Daten zu erzeugen. Unter Verwendung der Mahalanobis Distanz wird der Abstand einer Instanz zu den Gauß Komponenten bestimmt und die Wahrscheinlichkeit für einen emotionalen Ausreißer errechnet. Ein Schwellenwert wird auch in diesem Fall zur Erkennung von Ausreißern verwendet, wobei Instanzen mit einem Mahalanobis-Abstand größer der angegeben Schwelle als Ausreißer gelten. Alle Instanzen der Teilmenge der detektierten Ausreißer werden in aufsteigender Reihenfolge der Wahrscheinlichkeitsdichtefunktion des GMM sortiert.

Zusätzlich zu den beiden unüberwachten Selektionsansätzen wird ein binäres SVM Ensemble auf Basis der bereits annotierten Instanzen erzeugt[7]. Das Vorgehen ist dabei ähnlich der SVDD Methode. Der Parameter für C ist ebenfalls gleichmäßig verteilt im Intervall $\left[\frac{1}{N}, 1\right]$ und γ gleichmäßig verteilt innerhalb des Intervalls $\left[\frac{1}{f}, 1\right]$. Die reklassifizierten Instanzen werden ebenfalls entsprechend ihrer Distanz zur Hyperebene in Richtung der emotionalen Klasse absteigend sortiert.

Die Selektion der zu annotierenden Instanzen wird mittels Borda's geometrischer Mittelwert Rang Aggregation Methode [141] auf den unüberwachten trainierten (SVDD oder GMM) und

[7] In der initialen Iteration existiert dieser Klassifikator nicht und trägt daher nicht zur Selektion bei.

überwacht trainierten (SVM) Vorschlägen durchgeführt. Abschließend werden die k Instanzen mit dem höchstem Rang für die Annotation ausgewählt.

Der vorgeschlagen Ansatz wird mittels fünffach Block Kreuzvalidierung bewertet. Pro Fold werden dabei vier Teile zum Training verwendet und ein Teil zum Test. Für jede Iteration des aktiven Lernverfahrens wird dabei ein SVM Klassifikator auf den, durch das Verfahren ausgewählten und vom Experten annotierten, Daten trainiert und mit dem SVM Baseline Klassifikator verglichen, welcher auf dem vollständig manuell annotierten Datensatz trainiert wurde. Die Parameter beider SVM Klassifikatoren sind identisch gewählt.

Für den unüberwachten Teil des aktiven Lernenansatzes wurde eine feste Anzahl von 25 SVDD Modellen gewählt ($m = 5, n = 5$), die Anzahl der GMM Komponenten beträgt 5. Weiter ist die Anzahl der maximal wählbaren Instanzen pro Iteration auf 5 begrenzt.

Die Abbildungen 3.12 und 3.13 zeigen die erfolgreichen Resultate beider alternativer Ansätze. Dargestellt ist die erreichte Klassifikatorgüte gemessen mittels G-Mean in Abhängigkeit des verwendeten Anteils aller zur Verfügung stehenden Instanzen des Datensatzes. Gestrichelt eingetragen ist die Performanz des klassisch überwacht trainierten Baseline Klassifikators. Der schattierte Bereich zeigt die Standardabweichung der 5 Kreuzvalidierungsdurchläufe der Evaluation des aktiven Lernverfahrens.

Die Verläufe der Probanden 013 und 019 zeigen dabei ein auffälliges Verhalten. Für beide Probanden bleibt die Standardabweichung in beiden Ansätzen durchgehend hoch. Auffallend ist bei diesen Probanden auch, dass der Baseline Klassifikator ebenfalls die geringste Performanz aufweist, zudem ist auch die Varianz des Baseline Klassifikators deutlich höher als die der anderen Baseline Klassifikatoren (nicht eingezeichnet). Beide Probanden zeichnen sich durch eine generell schwierige Annotation aufgrund ihres untypische geringen emotionalen Verhaltens aus. Es ist anzunehmen, dass die überschaubaren Resultate dieser beiden Probanden in der geringen Anzahl und auffallend geringen Expressivität begründet sind, nicht im aktiven Lernalgorithmus. Beide Probanden haben abgesehen von der hohen Varianz zwischen den Durchläufen einen grundsätzlich erwarteten Verlauf, bezogen auf das aktive Lernverfahren. Alle anderen Probanden zeigen durchgehend hervorragende Resultate. Die Verläufe beginnen jeweils zunächst auf schwachem Niveau, nähern sich jedoch schnell dem Baseline Klassifikator an. In allen Fällen wurde gezeigt, dass Baseline Resultate bereits mit wenigen Daten erreicht werden können. Das SVDD basierte Verfahren zeichnet sich dadurch aus, dass es initial bereits auf besserem Niveau startet und die Baseline schneller erreicht. Daher wird im folgenden auf weitere Untersuchungen des GMM basierten Ansatzes verzichtet.

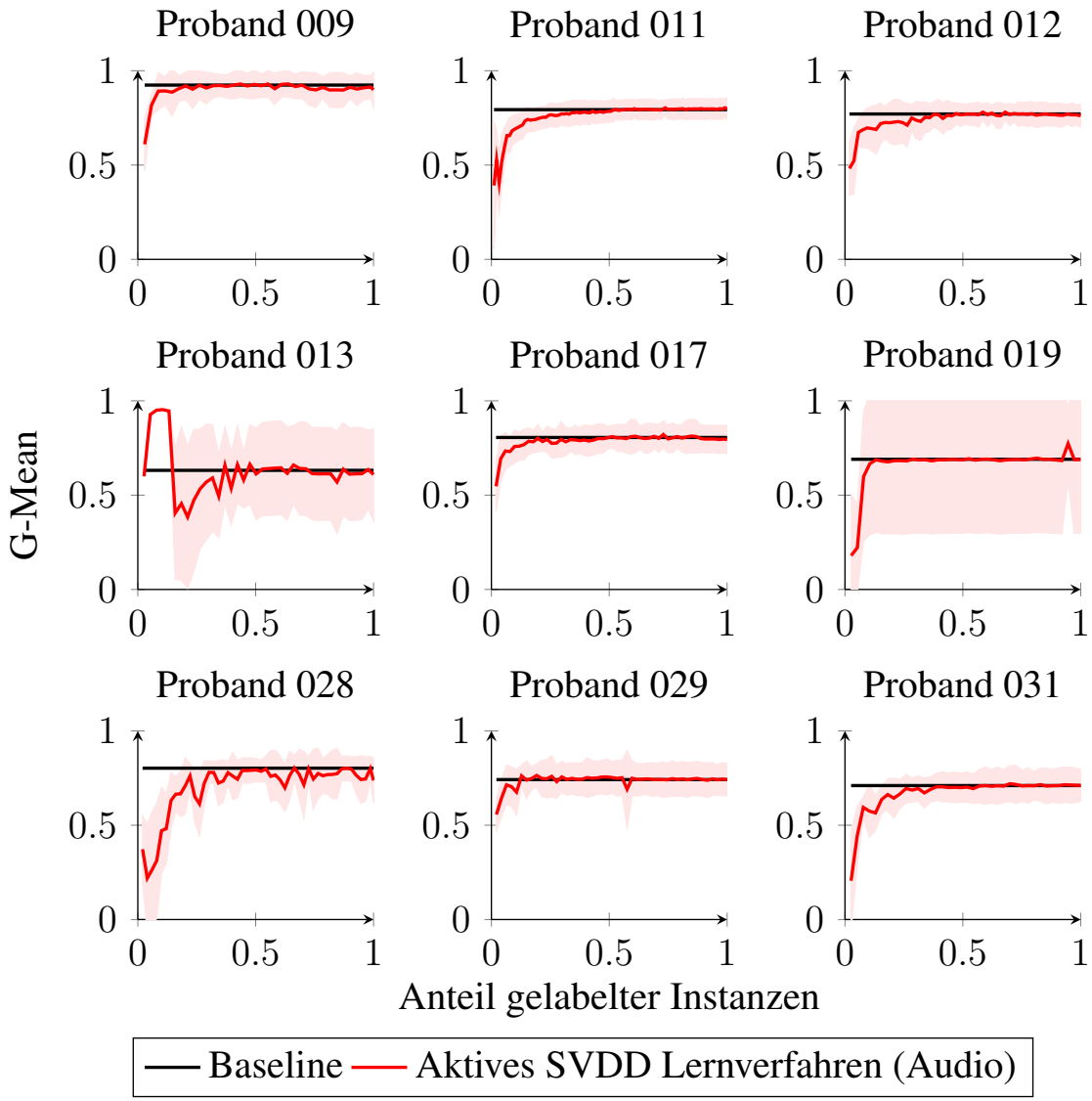

Abbildung 3.12.: Audio Resultate des aktiven Lernverfahrens bei Verwendung der SVDD Instanzselektionsmethode. Personenabhängige 5-fach Kreuzvalidierung. Die schwarze Linie entspricht dem Baselinewert eines auf den gesamten Daten trainierten SVM Klassifikators. Die rote Linie entspricht dem nach jeder Iteration des aktiven Lernverfahrens trainierten SVM Klassifikators, trainiert auf den zum Iterationszeitpunkt annotierten Daten. Der schattierte Korridor beschreibt die Standardabweichung der Kreuzvalidierung.

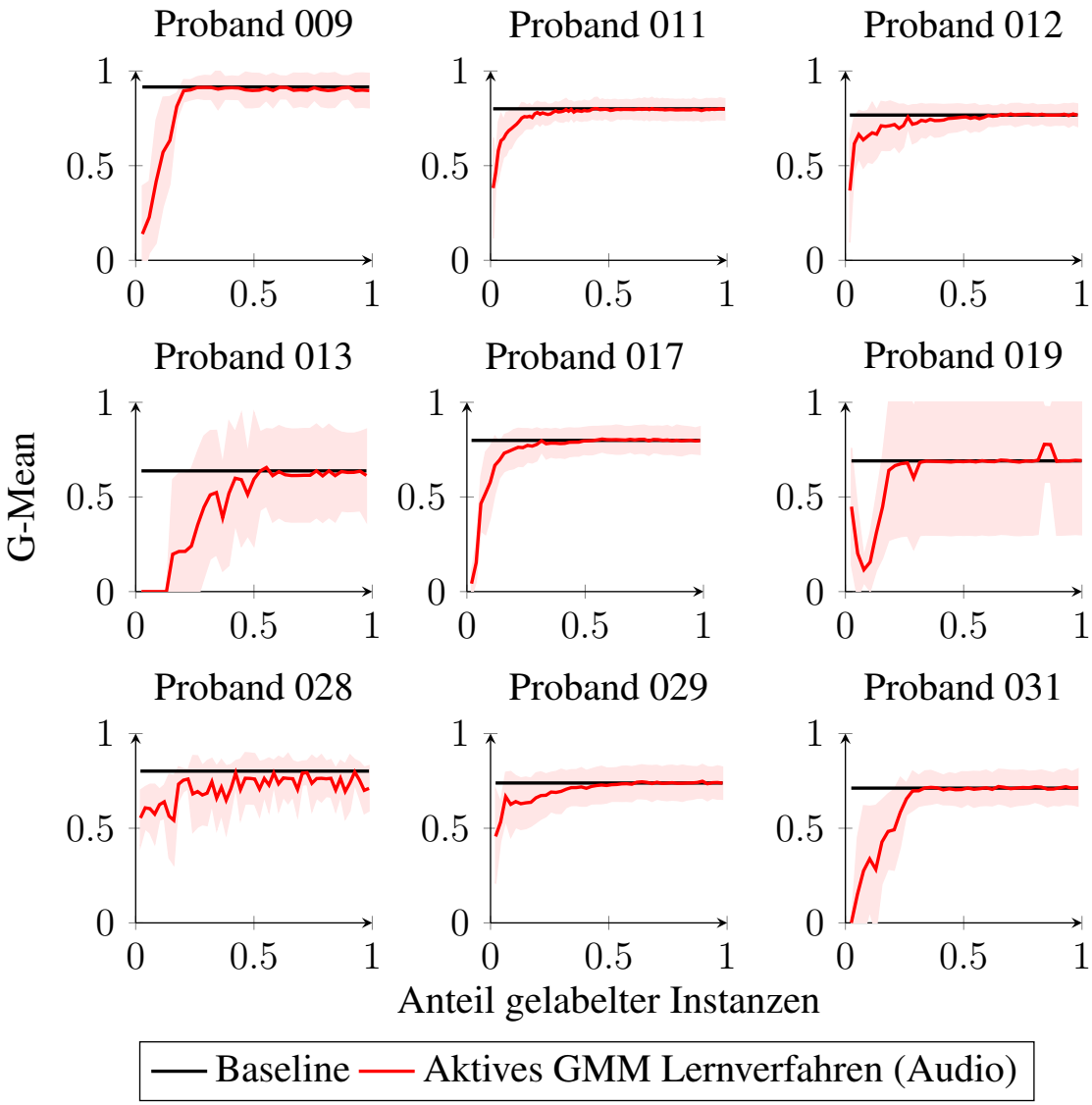

Abbildung 3.13.: Audio Resultate des aktiven Lernverfahrens bei Verwendung der GMM Instanzselektionsmethode. Personenabhängige 5-fach Kreuzvalidierung. Die schwarze Linie entspricht dem Baselinewert eines auf den gesamten Daten trainierten SVM Klassifikators. Die rote Linie entspricht dem nach jeder Iteration des aktiven Lernverfahrens trainierten SVM Klassifikators, trainiert auf den zum Iterationszeitpunkt annotierten Daten. Der schattierte Korridor beschreibt die Standardabweichung der Kreuzvalidierung.

Videobasiertes aktives Lernen

Analog zum vorherigen Abschnitt wurde eine Studie durchgeführt, welches die Performanz des vorgeschlagenen aktiven Lernverfahrens auf Basis unimodaler Videodaten zeigt. Es wurde derselbe SVM und SVDD basierte Selektor verwendet um die interessantesten Instanzen zu selektieren, welche manuell annotiert werden sollen. Die Evaluierungsprozedur ist ebenfalls identisch gehalten. Abbildung 3.14 zeigt ein ähnlich gutes Verhalten wie im Falle der audiobasierten Annotation. Auch hier ist ein schnelles Ansteigen der Qualität des erlernten Klassifikators zu erkennen, welches anschließend in eine Sättigung auf dem Level des Referenz-Klassifikators, welcher auf den gesamten Daten trainiert wurde, übergeht. Tabelle 3.4 zeigt die Anzahl an akustischen Beobachtungen, welche notwendig sind, um einen zugehörigen G-Mean zu erreichen. Erstaunlich ist, dass bereits mit 0.2% der gesamten Daten ein Klassifikator trainiert werden kann, welcher 90% der maximal möglichen Qualität erreicht. Voraussetzung ist dabei allerdings die korrekte Auswahl der zum Training selektierten Instanzen, welche die maximale Information tragen und repräsentativ für den gesamten Datensatz stehen. Die individuellen Unterschiede lassen sich aus Abbildung 3.14 entnehmen. Die schwarze Linie entspricht dem Baselineklassifikator, die blaue Linie stellt den G-Mean in Abhängigkeit der verwendeten Datenmenge dar. Der schattierte Bereich beschreibt die Varianz welche sich aus den einzelne Folds der Testdurchläufe ergibt. Es ist zu erkennen, dass die Varianz zu Beginn größer ist, da hier die Auswahl einzelner Datenpunkte erheblichen Einfluss auf das Resultat hat. Versehentlich ungünstig gewählte Instanzen sind in dem noch sehr geringen Trainingsdatensatz von Bedeutung und werden noch nicht durch Generalisierungseffekte korrigiert.

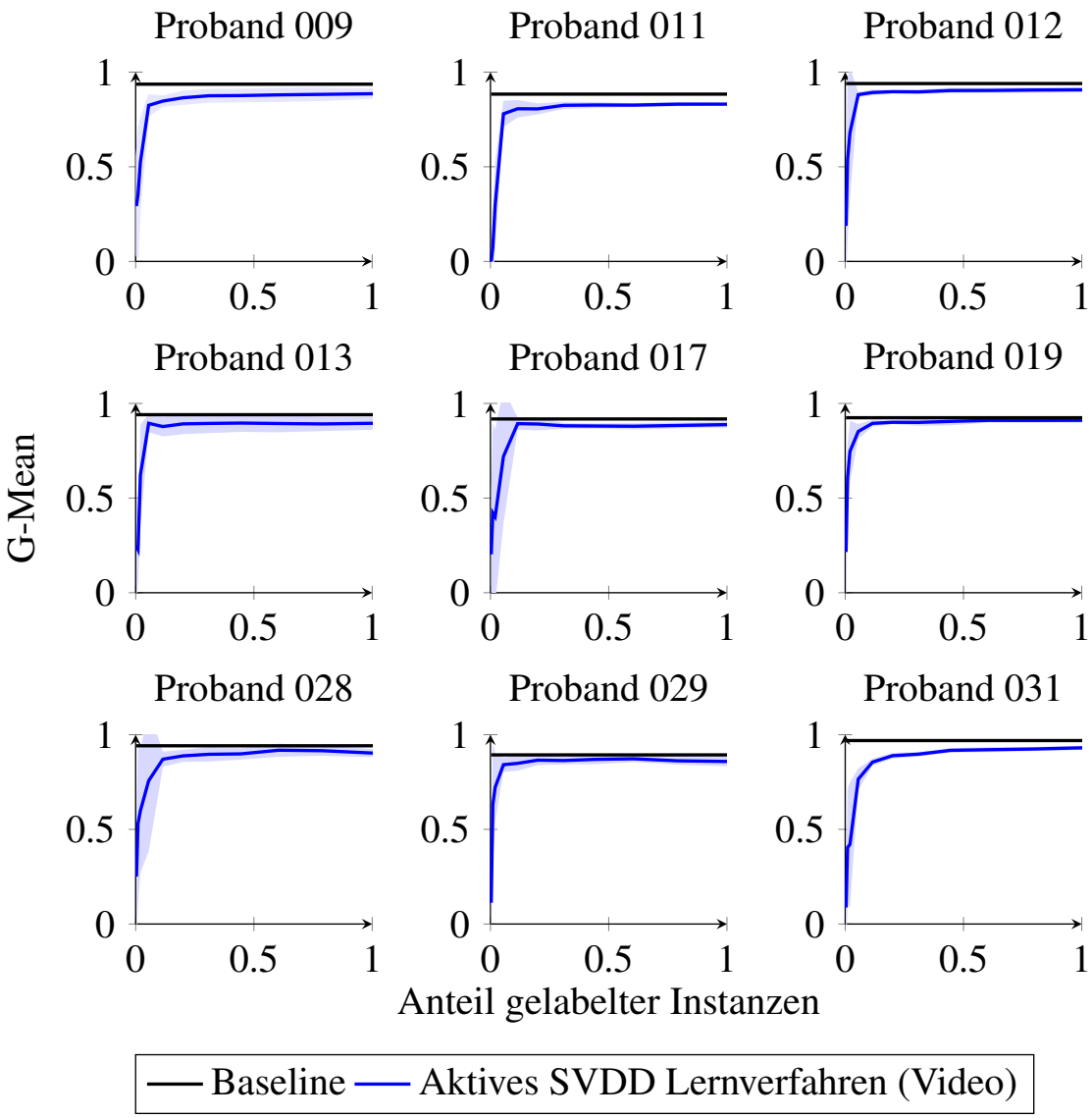

Abbildung 3.14.: Video Resultate des aktiven Lernverfahrens bei Verwendung der SVDD Instanzselektionsmethode. Personenabhängige 5-fach Kreuzvalidierung. Die schwarze Linie entspricht dem Baselinewert eines auf den gesamten Daten trainierten SVM Klassifikators. Die blaue Linie entspricht dem nach jeder Iteration des aktiven Lernverfahrens trainierten SVM Klassifikators, trainiert auf den zum Iterationszeitpunkt annotierten Daten. Der schattierte Korridor beschreibt die Standardabweichung der Kreuzvalidierung.

3.3.2. Multimodale aktive Lernverfahren affektiver Daten

Da affektive Datensätze realer Szenarien multimodal erhoben werden, müssen diese auch multimodal annotiert werden. Bisher wurde in der Literatur noch kein Vorschlag beschrieben, welcher aktive Lernverfahren auf multimodale Daten anwendet. Eine Betrachtung der multimodalen Eigenschaften fördert schnell kritische Fragestellungen zu tage. Fraglich ist zum Beispiel, ob eine Grundwahrheit für mehrere Modalitäten verwendet werden kann, oder ob es notwendig ist eine separate Grundwahrheit für jede Modalität zu verwenden. Auch ist nicht geklärt, inwieweit multiview Lernverfahren auf Multimodalität übertragbar sind. Neben der Frage der Grundwahrheit ist zudem zunächst nicht intuitiv klar wie die Selektionsstrategie auf mehrere Modalitäten erweitert werden soll. Sollen zum Beispiel die Datenpools mehrerer Modalitäten vereinigt und daraus Instanzen selektiert werden, oder ist eine getrennte Betrachtung zweier Modalitäten und ihrer Datenpools sinnvoll. Dabei muss dann mit zwei Selektionsrankings gearbeitet werden, welche anschließend zusammengefasst werden müssen. Auch ist nicht klar, inwieweit eine für menschliche Annoteure schwer beherrschbare Datenquelle, wie zum Beispiel Biophysiologie, einen Beitrag zur Verbesserung der Selektionsstrategie beitragen kann. Die folgenden Abschnitte beschreiben einen ersten Ansatz zum multimodalen aktiven Lernen in affektiven Datensätzen. Die vorgeschlagen Ansätze erweitern dabei die SVM basierten Ansätze der zuvor vorgestellten unimodalen Methode.

Erzeugung der Grundwahrheit

Die Grundwahrheit ist elementarer Bestandteil aller überwachten Lernverfahren. Im unimodalen Fall ist intuitiv klar wie diese zu erzeugen ist. Ist in einer Modalität eine emotionale Reaktion (z.B. *Lächeln* oder *erhobene Stimme*) zu erkennen, wird in der zugehörigen Modalität ein Label gesetzt. Im Falle multimodaler Daten kann die Grundwahrheit einzelner Modalitäten differieren. So ist im Falle eines Lächelns im Videokanal ein Label gesetzt, nicht jedoch in allen anderen Kanälen. Unklar ist in solchen Fällen nun wie die gemeinsame Grundwahrheit aller Modalitäten aussehen sollte.

Unsere menschliche Wahrnehmung erkennt Emotionen bereits wenn diese in nur einer Modalität gezeigt werden und erkennt so den emotionalen Zustand eines Gegenübers. Technisch gesehen entspricht dies einer OR Verknüpfung der Modalitäten. Im Folgenden wird die Grundwahrheit daher zunächst aus einer OR Verknüpfung aller unimodal manuell annotierten Grundwahrheiten gebildet.

Weiter haben einzelne Modalitäten unterschiedliche zeitliche Auflösungen in den darin gezeigten emotionalen Reaktionen, dadurch haben auch die darin annotierten Labels unterschiedliche

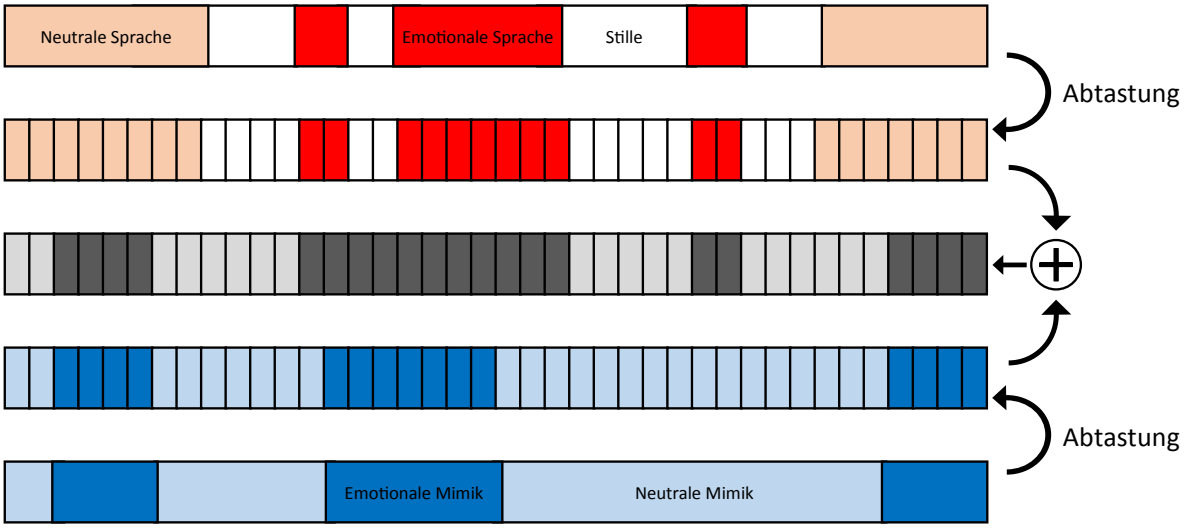

Abbildung 3.15.: Erzeugung der multimodalen Grundwahrheit des aktiven Lernverfahrens. Die Labels aller Modalitäten werden zunächst mit 1000Hz abgetastet und anschließend mittels OR Operator zusammengefasst. Dunkle Felder repräsentieren emotionale Sequenzen, helle Felder neutrale Sequenzen, weiße Felder entsprechen Stille. Audio ist rot, Video blau und Fusion schwarz dargestellt.

Auflösungen. Zur Vereinheitlichung der Zeitskala im multimodalen Fall werden alle unimodalen Labels mit einer Abtastrate von 1000Hz gesampelt, die daraus resultierende Auflösung ist mit einer Millisekunde deutlich kürzer als die darin zu erwartenden emotionalen Reaktionen. Die Veränderung der Segmentgrenzen im Millisekunden Raster ist daher vernachlässigbar. Der Vorgang der multimodalen Grundwahrheitsbildung ist in Abbildung 3.15 dargestellt.

Multimodale Selektionsstrategie

In der Emotionserkennung haben sich hierarchische Fusionsarchitekturen mit einer späten Fusionsstrategie bewährt (siehe hierzu Kapitel 4). Aufgrund dessen ist auch im Falle von multimodalem aktiven Lernens zu erwarten, dass es zielführender ist zunächst für jede Modalität einen Klassifikator zur Selektion relevanter Instanzen hinzusetzten. So werden zudem auch alle Probleme, welche aus unterschiedlichen Abtastraten, Fenstergrößen und Merkmalsvektorlängen resultieren, elegant umgangen. Demnach stehen dem aktiven Lernverfahren im Falle bimodaler Datensätze zwei getrennte Pools von Instanzen zur Verfügung. Diese Pools sind dabei unabhängig.

Eine Selektionsstrategie besteht dabei darin die beiden Pools separat zu betrachten und aus beiden Pools pro Iteration jeweils, mittels der für unimodale Fälle beschriebenen SVDD+SVM Strategie, eine gleiche und feste Anzahl von Instanzen zu selektieren. Diese werden dann ge-

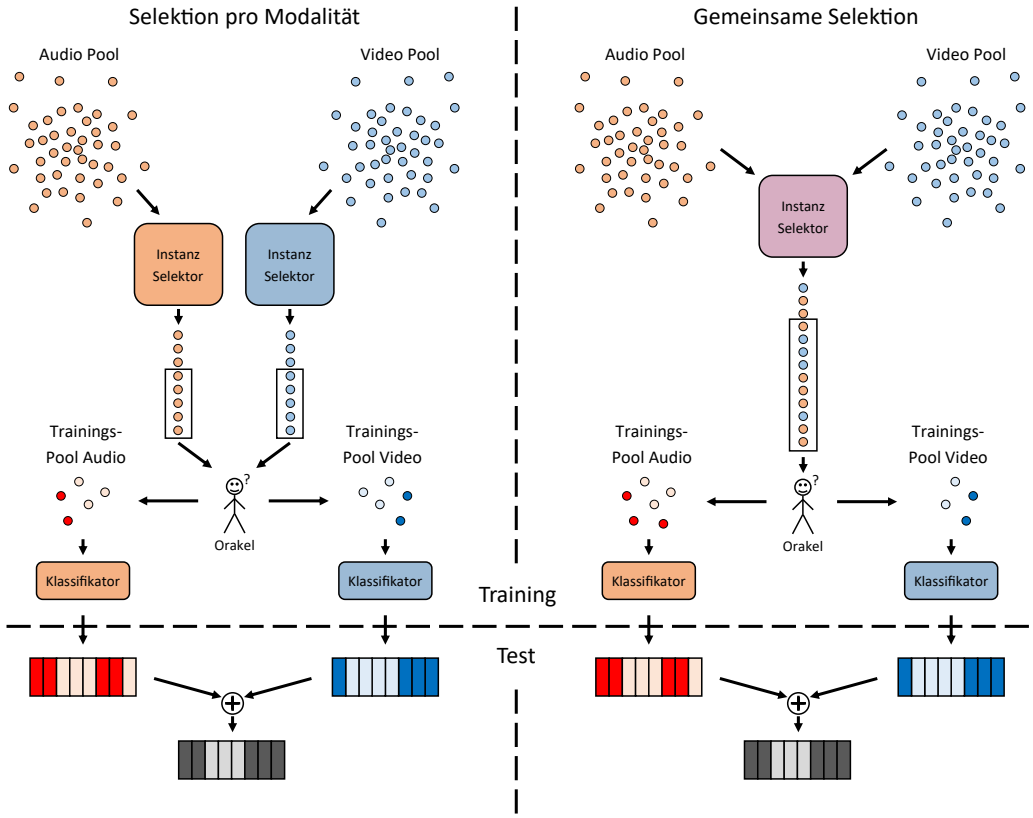

Abbildung 3.16.: Gegenüberstellung der multimodalen Selektionsstrategien des aktiven Lernverfahrens. Bei der Selektion pro Modalität ist der aktive Lernprozess vollständig getrennt. Der Selektor ist auf die Modalität optimiert, die Anzahl der selektierten Instanzen je Modalität ist leicht kontrollierbar. Erst in der Testphase erfolgt die Fusion. Bei der gemeinsamen Selektion verwendet der aktive Lernprozess einen gemeinsamen Selektor. Die Anzahl der selektierten Instanzen je Modalität hängt von den Eigenschaften der Daten und des Selektors ab. Es besteht die Gefahr das eine Modalität unterrepräsentiert wird. In beiden Fällen kennt das Orakel die Herkunft der Instanzen nicht, die Bewertung erfolgt immer multimodal.

meinsam an den Annoteur gegeben, damit dieser diese annotieren kann. Der Annoteur bekommt dabei beide Modalitäten präsentiert und entscheidet aufgrund seiner Einschätzung beider Modalitäten über das Label, er hat dabei keine Kenntnis aus welche Modalität der Vorschlag stammt. Das vom Annoteur vergebene Label wird dann der entsprechenden Quellmodalität zugeordnet.

So entstehen im Laufe des iterativen aktiven Lernens zwei gleich große separate Mengen von Trainingsdaten. Auch die abschließend überwacht trainierten Klassifikatoren bestehen aus zwei unabhängigen SVMs für jede Modalität. Die Vereinigung der beiden Klassifikatoren zu einer Entscheidung erfolgt dann mittels Fusion. Für die Untersuchung der multimodalen aktiven Lernverfahren wurde ein gewichtetes Voting als Fusion eingesetzt.

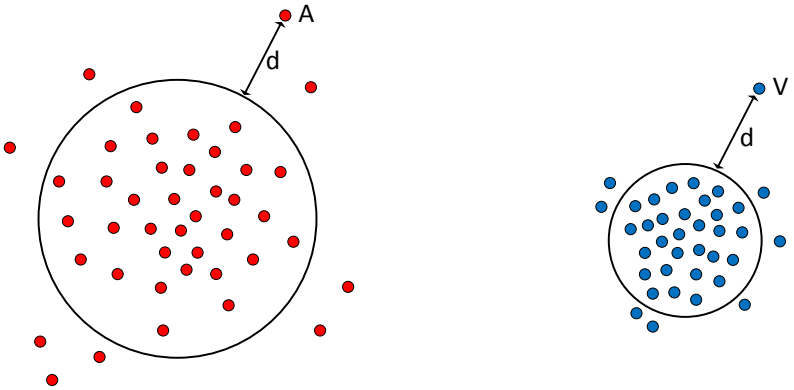

Abbildung 3.17.: Vergleichbarkeitsproblem der Distanzen zweier Modalitäten (Modelle). Für beide Modalitäten wurde unabhängig voneinander ein Modell einer Hyperebene trainiert (schwarzer Kreis). Die Punkte A und V haben identische Distanzwerte zu ihrer jeweiligen Hyperebene und würden bei der Selektion gleich gewichtet, obwohl Punkt V offensichtlich ein stärkerer Ausreißer ist als Punkt A. Sind A und V annotiert folgen anschließend überwiegend rote Punkte da diese größere Distanzwerte zur Hyperebene haben.

Eine weitere Selektionsstrategie versucht aus den beiden Pools Instanzen derart zu wählen, dass diese in ein gemeinsames Ranking überführt werden können. Anhand dessen wird dann eine feste oder variable Anzahl von Instanzen zur manuellen Annotation selektiert. Das Verhältnis der beiden beteiligten Modalitäten ist dabei nicht festgelegt. Unter Umständen ist sogar die Selektion aus nur einem Pool möglich. Abbildung 3.16 stellte beide Selektionsstrategien gegenüber. Da eine frühe Fusion in einen Klassifikator zur Selektion nicht möglich ist, bleiben die Selektoren weiterhin getrennt, lediglich die Auswahl anhand des aggregierten Rankings verbindet die beiden Modalitäten.

Für den Fall eines gemeinsamen Rankings resultieren allerdings einige Probleme. Die beiden eingehenden Modalitäten haben gänzlich unterschiedliche Eigenschaften, verwenden vollkommen andere Merkmalsvektoren, unterliegen gänzlich anderen Verteilungen und Dichten. Die Distanzen, welche bisher als Selektionsmaß in die Rankings eingingen, sind im multimodalen Falle nicht vergleichbar. Abbildung 3.17 verdeutlicht dieses Problem. Die Instanzen A und V haben einen ähnlichen Distanzwert, V ist allerdings eindeutig ein stärker Ausreißer. Einige Vorversuche die beiden Pools so zu normieren, dass eine Vergleichbarkeit möglich ist, sind bei der Anwendung auf realen Daten gescheitert. In der Regel führt dies zu einer Dominanz einer Modalität bei der die zweite Modalität nie berücksichtigt (ausgehungert) wird, wodurch der multimodale Ansatz auf den unimodalen Ansatz reduziert wird.

Ein weiterer Versuch das Distanz Skalierungsproblem zu lösen bestand darin, die Distanzen im Verhältnis zu den stärksten Ausreißer eines Pools zu skalieren. Dies führt allerdings dazu,

die tatsächliche Distanz als Maß zu missachten, zudem ist das Verfahren äußerst sensibel bezüglich eines einzelnen zufällig entstanden (durch Rauschen oder Messfehler) weit entfernten Ausreißers. Auch Versuche das Verfahren auf die Mittlere- oder Mediandistanz der k äußersten Instanzen zu erweitern hat keinen Erfolg gezeigt. Zudem unterscheiden sich die beiden Selektoren in ihren Kernels deutlich.

Aufgrund dieser Unwägbarkeiten wurden die folgenden multimodalen Experimente mittels der zuvor beschrieben Strategie der zwei unabhänigen Selektionen fester Größe durchgeführt. Der sich daraus ergebende bimodale aktive Lernalgorithmus ist in Algorithmus 3 beschrieben.

Algorithmus 3 Fusion im aktiven Lernen

$\mathcal{U} = \{\mathcal{U}_1, \mathcal{U}_2, ..., \mathcal{U}_n\}$ ▷ Ungelabelte Mengen verschiedener Modalitäten
$\mathcal{L} = \{\}$ ▷ Gelabelte Menge
$\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_n\}$ ▷ Testmengen der Modalitäten
$S = \{s_1, s_2, ..., s_n\}$ ▷ SVM für jede Modalität
m ▷ Anzahl gewählter Instanzen pro Iteration
while $Abbruchkriterium \neq true$ **do**
    for all $\mathcal{U}_k \in \mathcal{U}$ **do**
        $x \leftarrow selektion(\mathcal{U}_k)$ ▷ Wähle Instanzen der ungelabelten Mengen
        $y \leftarrow orakel(x)$ ▷ Fordere Label beim Orakel an
        $\mathcal{L}_k \leftarrow \mathcal{L}_k \cup (x, y)$ ▷ Füge gelabelte Instanz zu gelabelter Menge
    end for
end while
for all $\mathcal{L}_k \in \mathcal{L}$ **do**
    $s_k \leftarrow train(\mathcal{L}_k)$ ▷ Trainiere SVM auf gelabelter Menge
    $r_k \leftarrow test(s_k, \mathcal{T}_k)$ ▷ Wende SVM auf Testmenge an
end for
$r \leftarrow merge(\{r_1, r_2, ..., r_n\})$ ▷ Fusion
return r

Vergleich der uni- und multimodalen Resultate

Für die Erzeugung der Resultate wird die OR vereinigte Grundwahrheit als Referenz herangezogen. Der nach jeder Iteration erzeugte überwachte Klassifikator besteht, wie in allen vorangegangen Experimenten, aus einer SVM. Zur Selektion der meistrelevantesten Instanzen pro Iterationsschritt wird das SVDD Verfahren des zuvor beschrieben unimodalen aktiven Lernens für jede Modalität verwendet. Die Experimente sind wie zuvor in einer fünffach Blockkreuzvalidierung durchgeführt. Da die Größe der beiden Pools stark unterschiedlich ist, pro Iteration jedoch die gleich Anzahl an relevanten Instanzen selektiert wird, ist in der Ergebnisdarstellung nicht mehr der Anteil der gesamten Daten, sondern die absolute Zahl der verwendeten Instanzen auf der Abszissenachse abgetragen. Da nunmehr offensichtlich ist,

dass die aktiven Lernverfahren bereits früh in Sättigung gehen und die Qualität vergleichbarer Baseline Klassifikatoren erreichen, ist nur der Bereich der ersten 3000 selektierten Instanzen relevant. Abbildung 3.18 stellt die unimodalen und das fusionierte Ergebnis eines nach jeder Iteration trainierten SVM Klassifikators dar.

Die geringere Performanz des Audiosignals ist darin begründet, dass im Datensatz ein geringer Sprachanteil vorhanden ist. Der akustische Klassifikator hat somit eine geringere Grundmenge aus der relevante Instanzen selektiert werden können, es sind schlicht weniger emotionale Instanzen von guter Qualität vorhanden. Daher ist der Videoklassifikator früher performant. Dennoch ist deutlich zu erkennen, dass der Audioklassifikator einen wesentlichen Beitrag zur Fusion leistet. In allen Fällen liegt die Qualität des multimodalen Klassifikators deutlich über den beteiligten Einzelklassifikatoren.

Tabelle 3.4 zeigt nochmals den Erfolg der uni- und multimodalen Klassifikatoren, wenn deren Trainingsmenge mittels einer aktiven Lernstrategie ausgewählt wurde. Für den multimodalen Fall ist zu erkennen, dass bereits 10% der zur Verfügung stehenden Daten ausreichend sind, um einen Klassifikator zu gewinnen, der bis auf 1% an die Güte einer auf allen Daten trainierten Architektur heranreicht (Zur Erinnerung, im unimodalen Fall lag dieser Wert noch deutlich höher, bei ca. 14% aller Daten.). Bedenkt man weiter die generell überschaubaren Resultate der Emotionserkennung, verglichen mit anderen Anwendungen des maschinellen Lernens, ist sogar ein 5% schlechterer Klassifikator zumeist nicht signifikant schlechter als sein vergleichbares Pendant, trainiert auf allen Daten. Ein solcher nicht signifikant schlechterer multimodaler Klassifikator kann jedoch schon mit 3,2% der Daten und somit mit drei hundertstel des Annotationsaufwandes gewonnen werden. Bei einem Annotationsaufwand von 10% bezogen auf die klassische manuelle Annotation ist lediglich eine 1% schlechtere Klassifikationsrate zu erwarten.

Anteil Baseline	**Anteil Instanzen Audio**	**Anteil Instanzen Video**	**Anteil Instanzen Fusion**
90 %	6.3 %	0.2 %	2.6 %
95 %	15.1 %	1.2 %	3.2 %
99 %	53.7 %	41.5 %	10.6 %
100 %	82.6 %	80.9 %	20.0 %

Tabelle 3.4.: Mittlerer Anteil der Merkmalsinstanzen welche benötigt werden um einen bestimmten Prozentsatz der Güte des Baseline Klassifikators zu erreichen. Bereits kleine Datenmengen reichen aus um nahe an die Baseline Güte heran zu reichen.

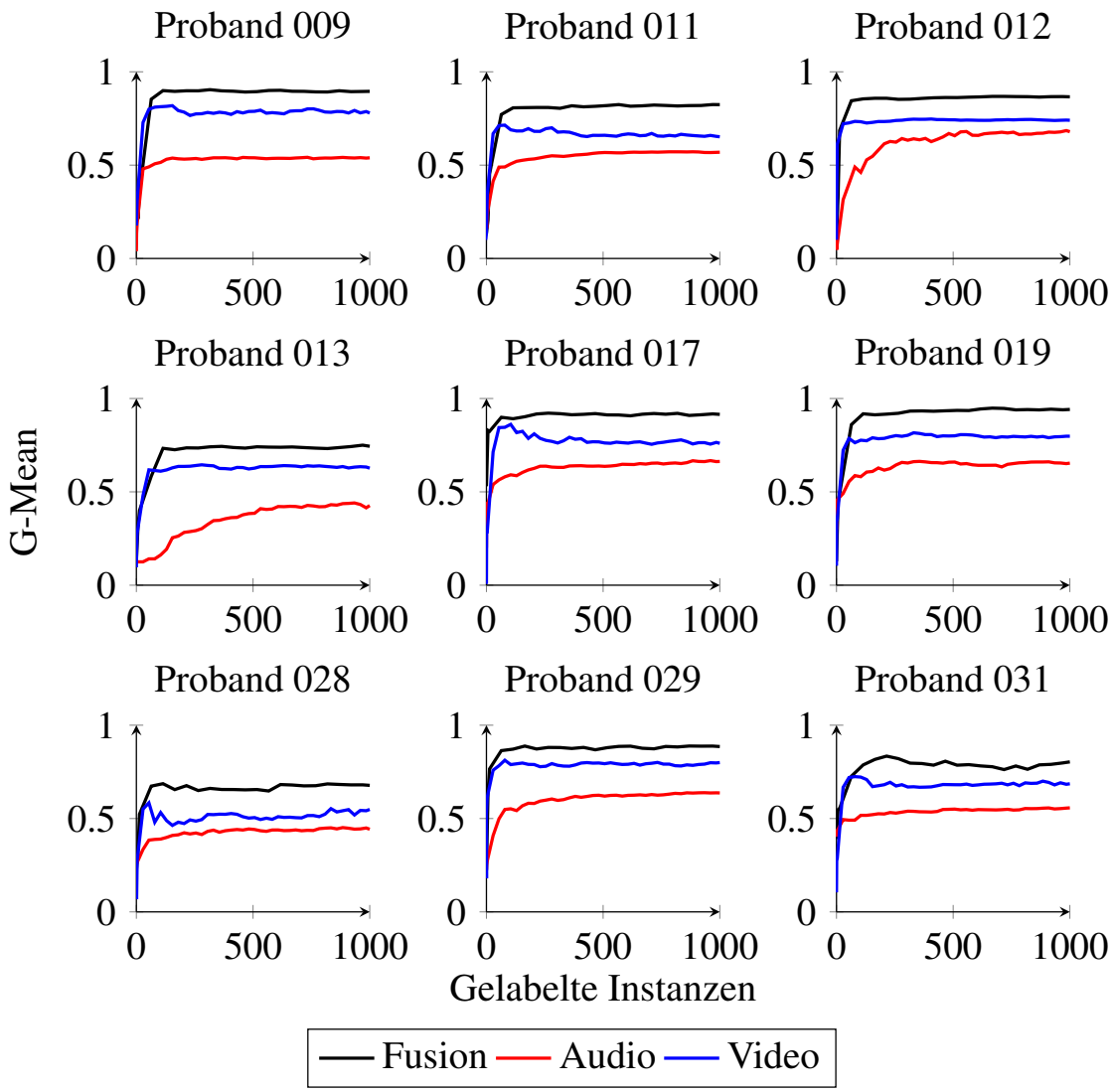

Abbildung 3.18.: Fusionsresultate des aktiven Lernverfahrens bei Verwendung der SVDD Instanzselektionsmethode. Personenabhängige 5-fach Kreuzvalidierung. Die schwarze Linie entspricht der Fusion, die rote und blaue Linie den Audio und Video Klassifikatoren bezogen auf die vereinte Grundwahrheit.

3.3.3. Abbruchkriterium

In den vorherigen Abschnitten wurde gezeigt, dass aktive Lernverfahren ein mächtiges Instrument sind um den Annotationsaufwand zu reduzieren. In allen gezeigten Experimenten wurde das Verfahren dabei entweder bis zum Ende, das heißt bis zur Annotation aller verfügbaren Daten, oder für eine zuvor festgelegte Anzahl an Iterationen durchgeführt. Der Vergleich mit den Baselineresultaten hat gezeigt, dass die Qualität des Baselineklassifikators zumeist früh erreicht wird. Ein Abbruchkriterium, welches sich intuitiv aufdrängt, wäre daher das Erreichen eines Schwellwertes abhängig von der Performanz des Baselineklassifikators. Dies ist in praktischen Anwendungen allerdings nicht möglich, denn das trainieren eines Baselineklassifikators setzt die manuelle Annotation aller Daten voraus, was erst nach Durchlauf aller aktiven Lerniterationen möglich ist. Fraglich ist daher anhand welchen Kriteriums erkannt werden könnte, dass eine dem Baseline Klassifikator entsprechende Qualität erreicht ist, ohne den Baseline Klassifikator zu kennen.

Eine scheinbar triviale Möglichkeit besteht darin, den Qualitätsunterschied des iterativen Klassifikators zur vorherigen Iteration, oder den l vorherigen Iterationen, zu bestimmen. Das Abbruchkriterium wäre dann erreicht, wenn die Differenz δ_l kleiner eines zuvor festgelegten Wertes ϵ ist. Dieser Ansatz scheitert allerdings auch daran, dass der iterative Klassifikator dabei alle bisher ungelabelten Instanzen klassifiziert. Um allerdings die Güte dieser Klassifikation bestimmen zu können ist es wiederum notwendig die gesamte Grundwahrheit aller Daten zu kennen. Alternativ könnte der Klassifikator auch nur die bereits annotierte Untermenge bewerten, dies entspricht allerdings seiner eigenen Trainingsmenge und somit einer Reklassifikation. Die dabei zu erwartenden Resultate sind vollkommen unrealistisch und sicherlich zu gut. Ein einfacher Versuch hat wie erwartet gezeigt, dass die Reklassifikation bereits in der ersten Iteration zum fast perfekten Klassifikator führt und das Verfahren terminiert.

Ein erfolgreicher Ansatz besteht darin, die Eigenschaften der eingesetzten Supportvektor Verfahren zu nutzen. Dabei wird in jeder Iteration die Anzahl, der aus den Trainingsdaten gewählten Instanzen, als Suportvektoren bestimmt. Es wird angenommen, dass sich das Modell nach erreichen der Sättigungsphase, welche immer nahe am Baseline Resultat liegt, nur noch geringfügig ändert. Für den Fall von Supportvektor Verfahren bedeutet dies, dass trotz neu hinzugekommener Instanzen in die Trainingsmenge, keine oder nur wenige Suportvektoren mehr in das Modell der nächsten Iteration aufgenommen werden. Weiter ändert sich auch die Gewichtung der Supportvektoren nur noch geringfügig. Die Differenz der Supportvektorzahl wird also als Maß für die Modelländerung verwendet und definiert ein Abbruchkriterium.

Abbildung 3.19 zeigt die Anzahl der gewählten Supportvektoren über der Anzahl der durchlaufenen Iterationen exemplarisch für die Videomodalität. Andere Modalitäten verhalten sich

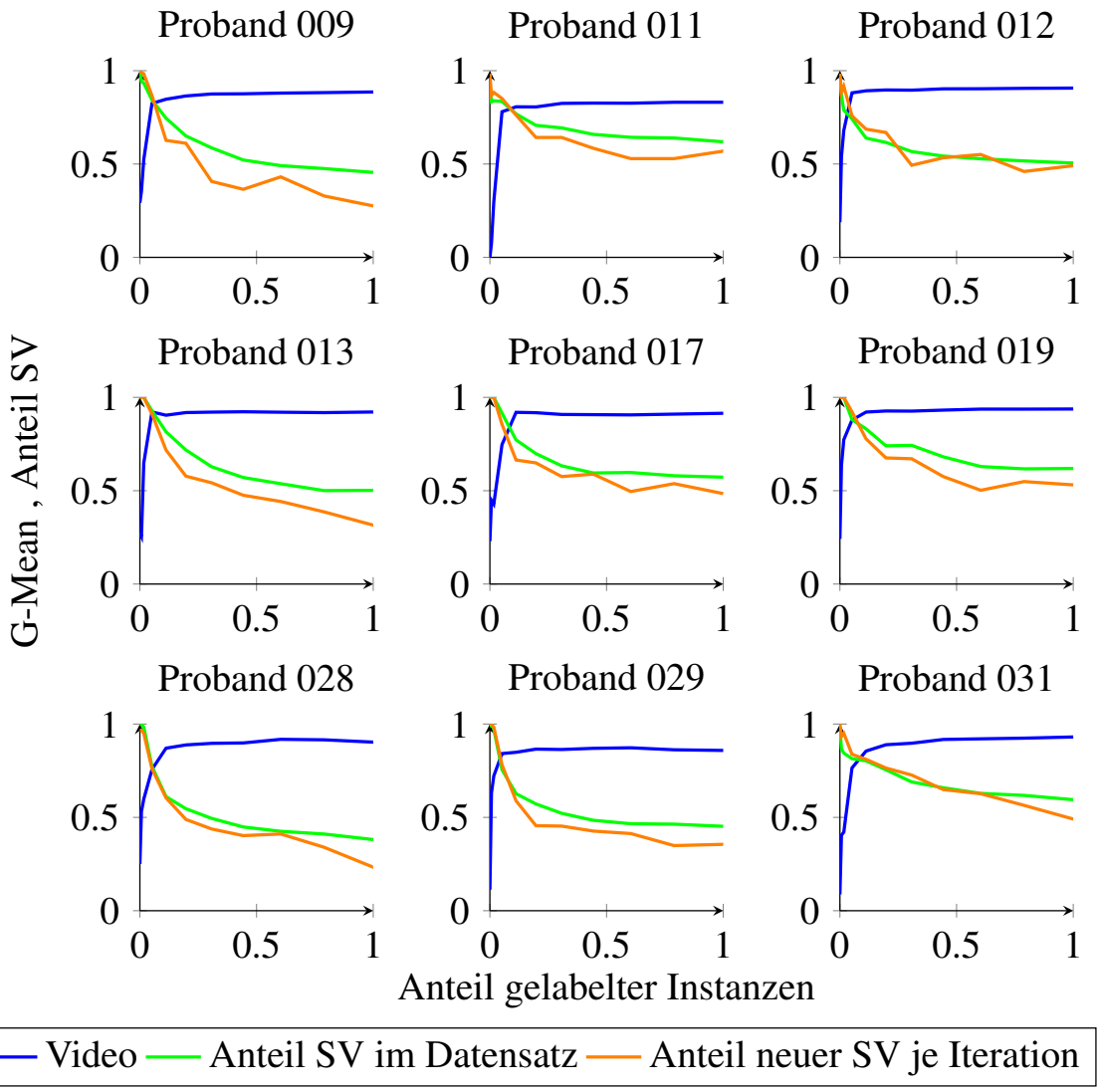

Abbildung 3.19.: Anzahl Support Vektoren je Iteration als Abbruchkriterium, am Beispiel des Videokanals. Anteil der Annotierten Instanzen je Iteration welche als Support Vektoren verwendet werden (orange) als dynamisches Abbruchkriterium des aktiven Lernverfahrens. Sinkt die Anzahl der Support Vektoren erreicht der Klassifikator gute Klassifikationsraten (blau). In grün dargestellt der Anteil der als Support Vektoren verwendeten Instanzen im gesamten annotierten Datensatz.

vergleichbar. Zu erkennen ist, dass in den frühen Iterationen des aktiven Lernens zunächst fast alle Instanzen der annotierten Untermenge als Supportvektoren verwendet werden. Der Anteil der Supportvektoren aus den zur Verfügung stehenden Trainingsdaten fällt dabei zunächst etwa exponentiell ab, und geht dabei bei etwa 50% aller zur Verfügung stehenden Instanzen in Sättigung. Der hohe Anteil von Supportvektoren auch am Ende aller Iterationen ist der Komplexität affektiver Modelle geschuldet und wird auch in klassischen überwachten Methoden beobachtet. Der ebenfalls aufgetragene Anteil neuer Supportvektoren aus den neu hinzugewonnen annotierten Instanzen folgt, wie zu erwarten, im wesentlichen dem Verlauf des Anteils aller Supportvektoren.

Betrachtet man nun den Verlauf der Supportvektor Anteile und den Verlauf der Klassifikatorgüte zeigt sich, dass die Klassifikatorgüte einem an der x-Achse gespiegelten Verlauf der Anzahl der Supportvektoren entspricht. Beide haben in etwa bei der selben Anzahl Iterationen ihren Knickpunkt und erreichen die Sättigungsphase. Da die Anzahl der Supportvektoren und Anzahl neu gewählter Supportvektoren auch ohne Kenntnis über die Klassifikatorgüte in jeder Iteration ohne zusätzlichen Rechenaufwand bestimmt werden kann, ist dies ein optimales Abbruchkriterium. Unterschreitet die Anzahl neu hinzugenommener Supportvektoren aus den neu gewählten Instanzen eine zuvor festgelegte Schranke, kann angenommen werden, dass das Modell weitgehend stabil ist und keine Verbesserung mehr eintritt. Der Selektionsalgorithmus des aktiven Lernens kann terminiert werden.

Ein solches Modelländerungsverfahren ist für die meisten Klassifikationsverfahren berechenbar [234], wodurch auch beim Einsatz nicht supportvektorbasierter Verfahren im aktiven Lernen ein vergleichbares Abbruchkriterium bestimmt werden kann.

3.3.4. Ausblick aktiver Lernverfahren

Die hier vorgeschlagen uni- und multimodalen aktiven Lernverfahren in affektiven Datensätzen sind erste Ansätze solcher Algorithmen in diesem Kontext. Sie beweisen die Mächtigkeit ihrer Möglichkeiten bei der Reduktion des Annotationsaufwandes affektiver Daten ohne dabei Qualitätsverluste bei der Bereitstellung affektiver Klassifikatoren hinnehmen zu müssen. Eine weitere Verbesserung der unimodalen Verfahren ist von Thiam und Meudt in [254] veröffentlicht. Darin ist die Rangaggregationsmethode zwischen unüberwachtem SVDD Selektor und überwachtem SVM Selektor dahingehend optimiert, dass eine Gewichtung erfolgt, die mit steigender Iterationszahl und somit Qualität des überwachten Klassifikators diesem gegen Ende des Verfahrens ein erhöhtes Gewicht gibt. Zusätzlich wird in dieser Veröffentlichung auch das multimodale Verfahren dahingehend erweitert, dass eine Gewichtung der Modalitäten ermöglicht wird.

Bisher wird bei keinem der Verfahren die Abhängigkeit zeitlich benachbarter Instanzen berücksichtigt. Es ist davon auszugehen, dass auch die zeitlichen Nachbarn einer als relevant selektierten Instanz, emotionales Verhalten beinhalten. Es ist daher zu erwarten, dass eine Anpassung der Selektionsgewichte aufgrund benachbarter Instanzen eine weitere Verbesserung beim schnellen Auffinden emotionaler Instanzen bringt. Ein Ansatz und eine Evaluation hierzu ist ebenfalls in [254] beschrieben. Durch die gewichtete Berücksichtigung von zeitlich und räumlich (Modalitäten) benachbarter Instanzen konnte so ein temporal-spatiales aktives Lernverfahren vorgeschlagen werden, welches zu einer weiteren Verbesserung des Auswahlprozesses führt. Mit diesem Verfahren ist es möglich enorme Datenmengen mit geringem Aufwand zu annotieren und somit für den Einsatz in der Companion-Technologie nutzbar zu machen.

Zudem kann die Hinzunahme weitere Modalitäten, welche von menschlichen Experten leicht annotierbar sind, mit den vorgestellten Methoden eine weitere Verbesserung erzielen. Vorstellbar sind hierbei zum Beispiel Körperpose und Bewegungsinformationen wie sie von einem Kinect Sensor erzeugt werden. Auch die Hinzunahme komplexer Modalitäten wie physiologische Kanäle (Elektroenzephalografie (EEG), EKG oder SCL) oder funktionelle Magnetresonanztomographie (engl. functional Magnetic Resonance Imaging) (fMRI) Daten, welche durch Experten nicht mehr ohne weiteres hinsichtlich emotionalem Inhalt annotiert werden können, bieten eine Möglichkeit die Verfahren weiter zu verbessern. Es ist problemlos möglich die unüberwachten Anteile der vorgeschlagenen Methoden (Neutrale Hintergrundmodelle: OCSVM, SVDD und GMM) auf komplexe Modalitäten zu erweitern. Dort gefundene relevante Instanzen können zwar nicht durch den Experten annotiert werden, aufgrund ihres zeitlichen Zusammenhanges kann die gefunden Information allerdings in andere Modalitäten propagiert werden. Dort können Selektionsgewichte angepasst werden, um so bessere Selektionsrankings zu erstellen. Andersrum kann ein Label, welches beispielsweise in der Audio und Video Modalität sicher vergeben wurde, auf die nicht menschlich annotierbare Datenquelle übertragen werden, sofern sich in dem zugehörigen unüberwachten Modell ebenfalls Hinweise auf einen emotionalen Inhalt (Ausreiser) finden lassen. Somit werden komplexe Spuren implizit annotiert und es können in Folge auch dort überwachte Klassifikatoren trainiert werden.

Eine abschließende Idee zur zukünftigen Annotation affektiver Daten sei hier nur kurz angerissen, da diese nicht methodisch, sondern technisch, begründet ist. Im Zuge der omnipräsenten smarten Technologie und weltweiten Vernetzung schlägt Morris et. al. vor Crowdsourcing und Crowdfunding Methoden auch auf die Annotation von affektiven Daten in Form von Crowdannotation anzuwenden [160]. Das am Anfang dieses Kapitels beschriebene Werkzeug ANNEMO verfolgt einen ähnlichen Ansatz in abgeschwächter Form. Diese sicherlich spannenden Herangehensweisen ließen sich hervorragend mit den Methoden der vorgestellten aktiven Lernverfahren kombinieren in denen die Crowd als Ensamble-Orakel fungiert.

3.4. EmotionWheel

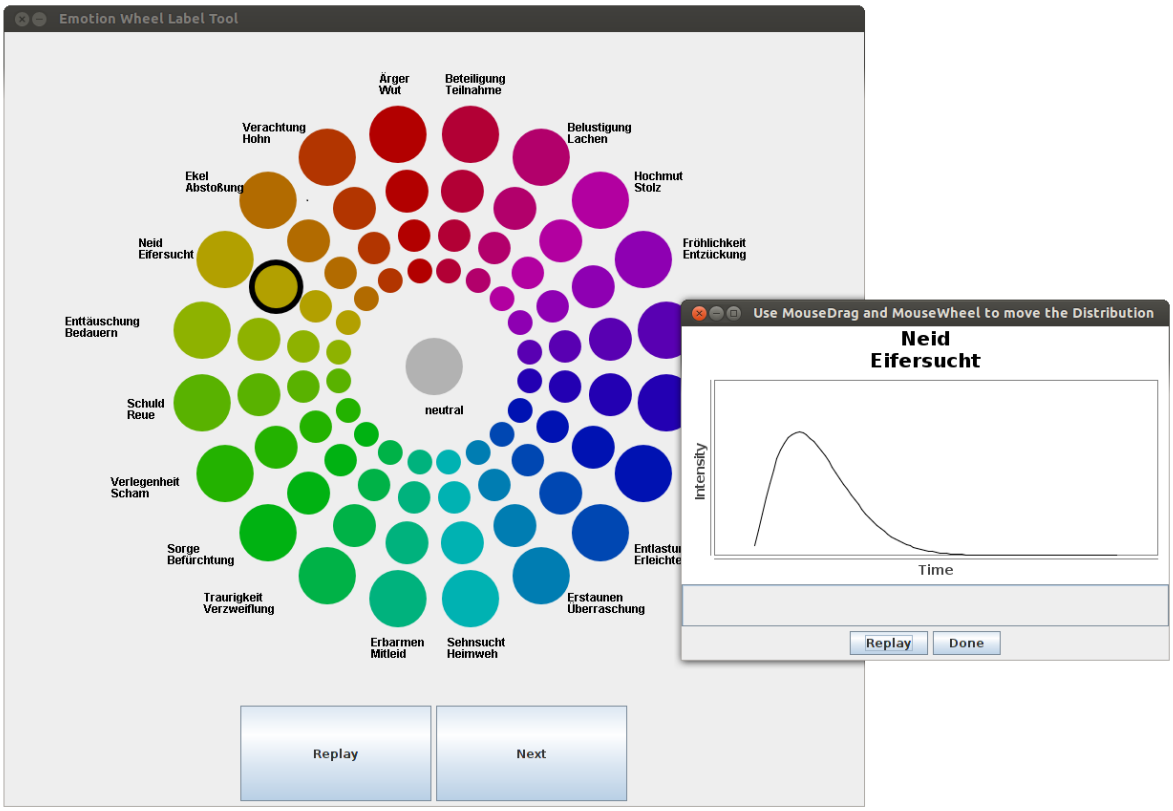

Abbildung 3.20.: EmoWheel Screenshot. Links: Emotionsanordnung nach Plutchik in 4 Intensitäten. Rechts: Modellierung der Intensitätsverteilung (Hier: hohe Intensität nur am Satzanfang).

Im Rahmen dieser Arbeit wurde zunächst in Zusammenarbeit mit Scherer und Bigalke das spezialisierte Annotationstool *EmotionWheel* entwickelt[8] [211]. Mit diesem Tool können kurze Audio oder Video Sequenzen entsprechend dem emotionalen Modell nach Plutchik annotiert werden. Die zu annotierenden Sequenzen werden dem Annoteur durch akustische und visuelle Echtzeitwiedergabe präsentiert, und ggf. in einer Schleife wiederholt. Der Annoteur wählt dann die zutreffende Emotion und Intensität aus. Zusätzlich ist es möglich eine Intensitätsverteilung durch Modellierung einer Betafunktion über die Dauer der Sequenz zu bestimmen. Die Betaverteilung ist auf $(0, 1)$ definiert durch die Wahrscheinlichkeitsdichte

$$f(x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1} \tag{3.4}$$

[8] `http://www.uni-ulm.de/fileadmin/website_uni_ulm/iui.inst.130/Mitarbeiter/smeudt/EmoWheel.tar.gz` (Stand: 3. Juni 2019)

mit den reellen Parametern p und q $(p, q > 0)$ und der namensgebenden Betafunktion

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 u^{p-1}(1-u)^{q-1}\mathrm{d}u \tag{3.5}$$

Dabei bezeichnet Γ die Gammafunktion.

Abbildung 3.20 zeigt die Oberfläche des Annotationstools mit dem Modell nach Plutchik und der Erweiterung durch die Betaverteilung. Die Form der Betaverteilung wird dabei grafisch, durch ziehen mit der Maus, modelliert. Kenntnisse der Betaverteilung und ihren Parametern sind nicht erforderlich.

3.5. ATLAS

Aufbauend auf die Anlyse bestehender Datensätze und Annotationstools und der erkannten Schwächen wurde im Rahmen dieser Arbeit das umfangreiche Annotationstool ATLAS als eine der zentralen Komponenten spezifiziert und implementiert[9] [151]. Ziel dieses Werkzeugentwurfs ist es, alle Modalitäten aller verfügbarer Benchmark Datensätze und des erhobenen uulmMAC, sowie potentieller zukünftiger Datensätze, in einer einheitlichen Umgebung darzustellen und synchron wiedergeben zu können. Weiter sollen alle im Arbeitsablauf des Klassifikatorentwurfs entstehenden Zwischenschritte, wie aufgearbeiteter Sensordaten, Annotationen, Merkmalsmatritzen, Zwischenresultate und Klassifikationsergebnisse darstellbar sein, um so neben den Rohdaten jederzeit einen Überblick über den Gesamtansatz visualisieren zu können [150]. Zusätzlich zur Visualisierungsanforderung solle es ermöglicht werden Datenbestände auf unterschiedlichsten Abstraktionsebenen zu annotieren, ohne dabei durch den Softwareentwurf in ein spezielles Annotationsschma oder gar emotionales Modell gedrängt zu werden. ATLAS ermöglicht so zum Beispiel die Annotation von Emotionen nach Eckman [70], Plutchik [179] oder VAD [197], genauso wie die Annotation nach eigenen definierten Schemata. Aber auch die Annotation nicht emotionaler Elemente wie Experimantalsequenzen, Trailgrenzen, Nutzerverhalten, Bewegungen oder allgemeiner Ereignisse mit und ohne zeitliche Ausdehnung. Aufgrund der Komplexität auftretender Annotationsaufgaben bietet ATLAS darüber hinaus die Möglichkeit mittels aktiver Lernverfahren oder anderer annotationsunterstüzender Methoden dem Annoteur Hilfestellungen bei seiner Arbeit zur Seite zu stellen [156]. Abschließend werden diverse Assistenzfunktionen und umfangreiche Schnittstellen zu gängigen Datenformaten und Werkzeugen bereitgestellt.

[9] `http://www.uni-ulm.de/in/neuroinformatik/mitarbeiter/s-meudt/` (Stand: 3. Juni 2019)

In den folgenden Abschnitten wird zunächst der spurenbasierte Grundaufbau von ATLAS erläutert, welcher die synchrone Darstellung und Wiedergabe aller sensorischen und extrahierten Daten sowie Resultate realisiert. Anschließend werden die zur Verfügung gestellten Hilfsfunktionen, Schnittstellen und aktive Lernfunktionen, welche ein effektives analysieren und annotieren von Daten und Resultaten ermöglichen, erläutert. Abschließend wird ein MAR^{2}S basierter Ansatz zum Umgang mit besonders umfangreichen Korpora vorgestellt und einige technische Details zur Realisierung erläutert. Aufgrund des enormen Funktionsumfangs werden die Funktionen im Rahmen dieses Dokumentes nur auf abstraktem Niveau in ihrer grundsätzlichen Funktionsweise und Idee dargestellt. Für eine detaillierte Beschreibung aller Menüpunkte und Parametern sei an dieser Stelle auf das Softwarehandbuch von ATLAS verwiesen, welches in seiner aktuellsten Form als Teil der in ATLAS integrierten Onlinehilfe verfügbar ist, oder als Basisversion im pdf-Format zum Download bereit steht[10]. Ebenfalls stehen zum Download einige Demonstrationsdatensätze[11], ergänzende Werkzeuge und Codebeispiele[12] zur Verfügung. Auf eine umfangreiche Beschreibung aller ATLAS Datenformate und Definitionen wird aus Gründen des Umfangs an dieser Stelle verzichtet. Aufgrund der freien Verfügbarkeit von ATLAS wird das Werkzeug zum gewärtigen Zeitpunkt weit über die Grenzen dieser Arbeit hinaus eingesetzt[13]. Seine flexible Gestaltung erlaubt dabei auch den Einsatz in MMI Kontext fernen Anwendungen.

Sofern nicht anders angegeben wird in den folgenden abgebildeten ATLAS Screenshots immer der uulmMAC als Beispieldatensatz verwendet, die gezeigten Funktionen sind jedoch auf jeden beliebigen Datensatz anwendbar. Abbildung 3.21 zeigt zunächst eine Übersichtsdarstellung einer typischen ATLAS Annotationsansicht für Experten dargestellt auf zwei Monitoren. Im linken Bildbereich sind alle dem Datensatz zugehörigen Daten- und Audiospuren zu sehen, im mittleren Bereich sind Videospuren dargestellt, im rechten Teilbereich sind einige Fenster geöffnet, welche Zusatzinformationen und Hilfsfunktion bereitstellen. Die Größe und Anordnung ist in ATLAS grundsätzlich flexibel gehalten um eine Anpassung an den dargestellten Datensatz und die Präferenzen des Annnoteurs zu ermöglichen. Darstellungseinstellungen werden gespeichert und beim öffnen von ATLAS berücksichtigt. Experten verwenden zumeist eine recht detaillierte Darstellung mit vielen Informationen und Hilfestellungen, dies erfordert jedoch einen geübten Umgang, nicht nur mit ATLAS, sondern mit der gesamten Materie

[10] `http://www.uni-ulm.de/fileadmin/website_uni_ulm/iui.inst.130/Mitarbeiter/smeudt/AtlasHelp.pdf` (Stand: 3. Juni 2019)

[11] `http://www.uni-ulm.de/fileadmin/website_uni_ulm/iui.inst.130/Mitarbeiter/smeudt/demo_clock.tar.gz` (Stand: 3. Juni 2019)

[12] `http://www.uni-ulm.de/fileadmin/website_uni_ulm/iui.inst.130/Mitarbeiter/smeudt/AtlasTools.tar.gz` (Stand: 3. Juni 2019)

[13] Bekannte Verwendung von ATLAS u.A. durch SFB-TRR 62 Teilprojekte A1, A2, B3, B4, C4, C5, Z3, T3, Cluster 8; BMBF Projekt SENSE-Emotion; Arbeitsgruppen Prof. Dr. H. C. Traue (Ulm), Dr. A. O. Andrade (Uberlandia), Dr. A. Brechmann (LIN Magdeburg), Jun.-Prof. Dr. C. Schrader (Ulm). Durch die freie Verfügbarkeit ist darüber hinaus mit weiteren, nicht bekannten, Verwendungen zu rechnen.

Abbildung 3.21.: Mögliches Aussehen der ATLAS Nutzerschnittstelle bei Verwendung von zwei Monitoren. Links: Liniendarstellung der Daten- und Labelspuren. Mitte: Videoansichten. Rechts: Eigenschafts- und Zusatzfunktionsfenster. Die Anordnung der Fenster ist in ATLAS vollständig flexibel gehalten, das ermöglicht große Freiheiten in der Verwendung. Es existiert somit nicht *die Eine* ATLAS Ansicht.

multimodaler Datensätze und deren Bearbeitung. Annoteure ohne umfangreiche Expertise bevorzugen zumeist eine einfachere Darstellung mit geringerem Funktionsumfang wie z.B. in Abbildung 3.6 dargestellt. Sie bearbeiten dabei jedoch auch nur eine, meist atomare, Aufgabe. Beide Ansichten operieren dabei auf demselben Datenbestand und können im Falle verteilter Annotation (siehe Abschnitt 3.5.5) sogar simultan in zwei ATLAS Instanzen existieren.

3.5.1. Aufbau und Funktionsweise von ATLAS

Jeder Datenquelle wird in ATLAS durch eine sogenannte Spur (engl. Track) repräsentiert und wird typischerweise aus einer Datei gespeist. Die Superklasse der Spur stellt dabei die Grundfunktionalitäten bereit, welche alle Spuren gemeinsam haben, wie z.B. Name, Anfangszeit oder Länge. Eine solche Spur ist ein abstraktes Element ohne konkrete Wiedergabeinstanz und wird erst durch davon abgeleitete Spuren instanziiert. Dabei wird zunächst noch zwischen Medien- und Linienspuren unterschieden, bevor in den Blättern des Vererbungsbaumes konkrete Spuren folgen, welche instanziiert werden können und eine grafische oder akustische Darstellungskomponente enthalten. Abbildung 3.22 zeigt den hierachischen Aufbau der ATLAS zugrunde liegenden spurbasierten Struktur.

Videospuren

Videospuren werden in ATLAS durch Darstellung in einem Videofenster präsentiert. Die Wiedergabe erfolgt dabei unter Verwendung des Video Lan Client (VLC)-Players[14] und seiner Java

[14] `https://www.videolan.org/vlc/index.html` (Stand: 3. Juni 2019)

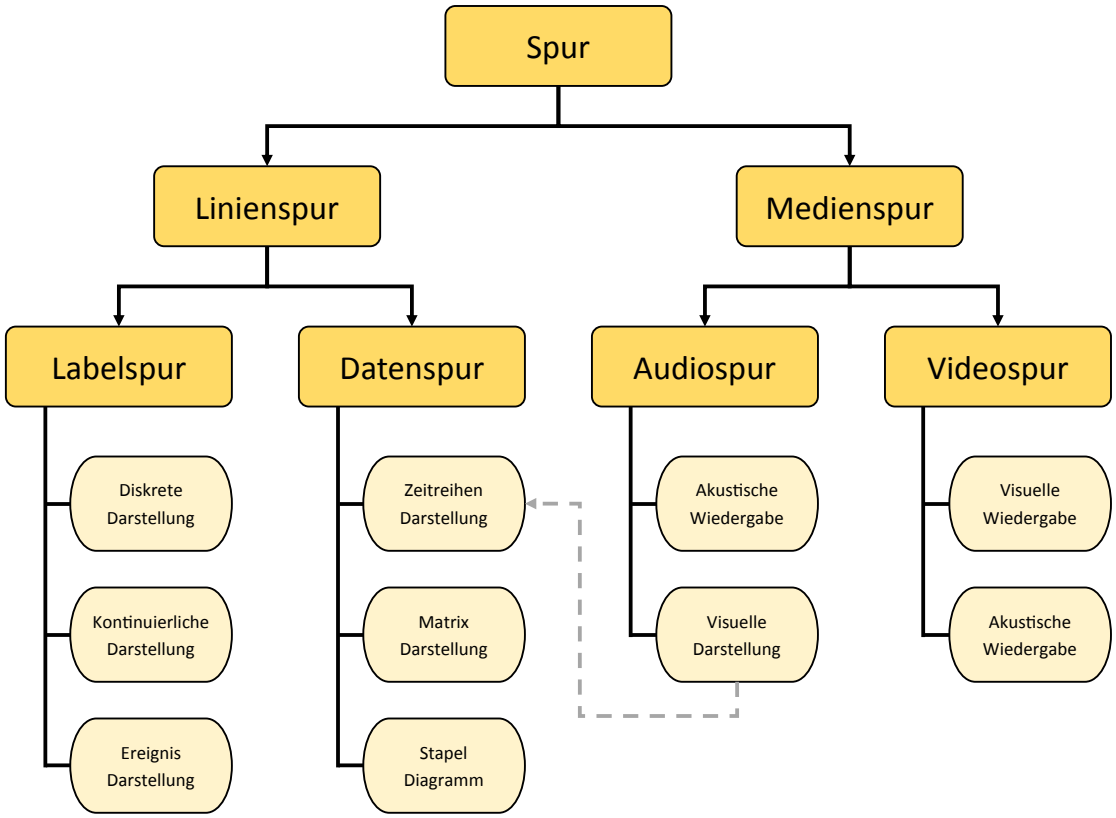

Abbildung 3.22.: Spurbasierter Aufbau von Atlas. Die abstrakte Klasse Spur wird zunächst in abstrakte Linen- und Medienspuren unterschieden. Diese werden durch Label-, Daten-, Audio- und Videospuren instanziiert. Diese haben verschiedene Formen der akustischen und visuellen Repräsentation.

Anbindung vlcj[15]. Videospuren können neben den offensichtlichen visuellen Daten zusätzliche Informationen wie akustische Kanäle, transkribierte Untertitel oder Metainformationen (z.B. aus dem Moving Picture Experts Group (MPEG) Standard) beinhalten. Diese Zusatzkanäle werden ebenfalls wiedergegeben, haben jedoch keine weitere Repräsentation in Form einer Linienspur. Pro Videospur wird ein Ausgabefenster erzeugt, es ist dabei möglich einzelne Videospuren auf unsichtbar zu schalten.

Audiospuren

Audiospuren werden in ATLAS akustisch, ebenfalls unter Verwendung des VLC-Players, wiedergegeben und haben zusätzlich eine Darstellung in Form einer Linenspur, in welcher die Abtastwerte für Mono- und Stereokanäle dargestellt sind. Grundsätzlich ist auch die akustische Wiedergabe von Mehrkanal Inhalten wie Dolby Digital 5.1 möglich, jedoch werden dabei

[15] `https://github.com/caprica/vlcj` (Stand: 3. Juni 2019)

lediglich die enthaltenen Steroinformationen grafisch dargestellt. Sind in einem ATLAS Projekt mehrere akustische Spuren enthalten, werden diese bei der Wiedergabe durch die Systemtreiber zu einer Stereoausgabe zusammengemischt. Es ist möglich innerhalb von ATLAS einzelne Spuren auf lautlos zu schalten.

Datenspuren

Datenspuren beinhalten ein oder mehrdimensionale Zeitreihen mit variabler oder konstanter Abtastrate. Jeder Punkt der Zeitreihe wird dabei durch seinen Zeitpunkt, Dimension und Abtastwert repräsentiert. Es stehen dabei zwei grundsätzliche Darstellungsoptionen zur Verfügung. Abbildung 3.23 zeigt eine Auswahl der gebräuchlichsten Darstellungsvarianten in Form ein- oder mehrzeiliger Linienzüge (zum Beispiel EEG), Abtastpunktdarstellung (zum Beispiel Kontextwerte oder Zustandsänderungen), Matrixdarstellung (zum Beispiel Merkmalsmatrix, Spektrogramm) oder Stapelbalkendiagramm (zum Beispiel mehrklassen Klassifikationsresultate)

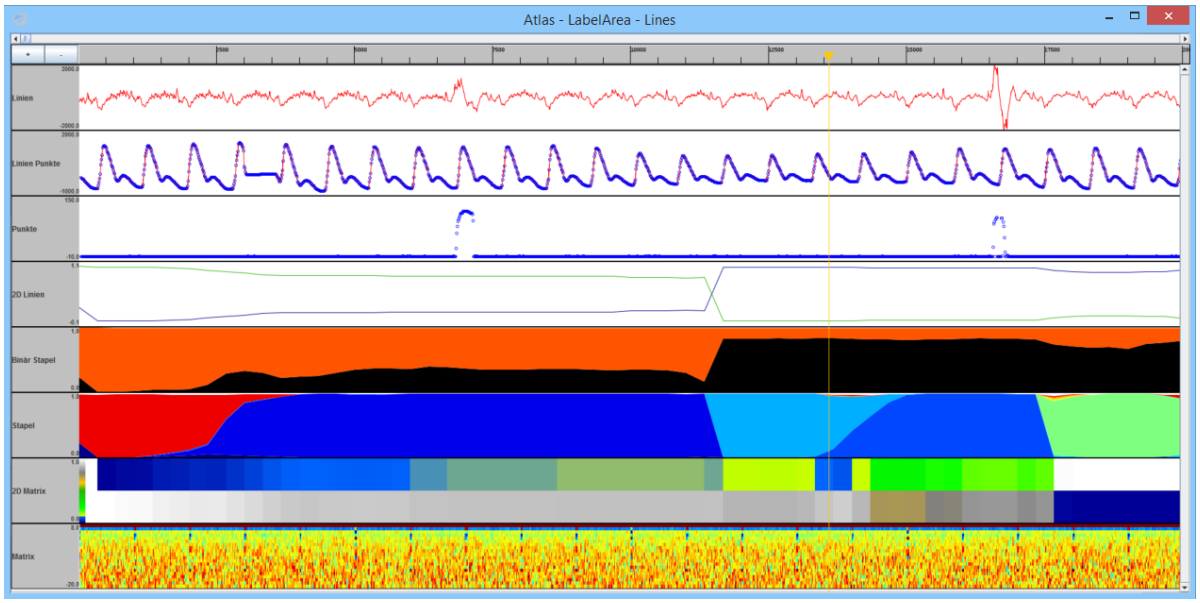

Abbildung 3.23.: Darstellung der ATLAS Linienspuren. Von oben nach unten: Zeitreihe als Linienzug, Zeitreihe als Linienzug mit Abtastwerten, Zeitreihe als Abtastwerte, zweidimensionale Zeitreihe als Linienzug, Klassifikationsresultat als binäre Stapeldarstellung, Klassifikationsresultat als mehrdimensionale Stapeldarstellung, duale Intensitäten als zweidimensionale Matrixdarstellung und Merkmalsvektoren in Matrixdarstellung.

Labelspuren

Labelspuren stellen das zentrale Element der Annotation dar. Jede Labelspur ist dabei genau einer Labelklasse zugeordnet. Labelklassen beschreiben dabei eine Menge von Entitäten welche als Zustandselement für ein Label ausgewählt werden können. Am Beispiel der Basis Emotionen nach Eckmann ist dies die Klasse *BasisEmotionen* mit den Entitäten *Freude, Wut, Ekel, Furcht, Verachtung, Trauer* und *Überraschung*. Gängige Klassen und Modelle wie Basisemotionen, VAD oder FACS stehen als Vorlagen zur Verfügung. Die Klassen können jedoch auch vom Anwender frei definiert werden, somit entfällt eine Beschränkung auf Emotionen oder feste Annotationsschemata wie in anderen Annotationswerkzeugen. Auch Klassen wie *Nutzerverhalten* mit Entitäten wie *Steht, Läuft, Zeigt, etc.* sind modellierbar. Weiter kann diese Struktur auch Informationen der Metaebene eines Experimentes durch Klassen wie *ExperimentalPhase* mit Entitäten *Intro, 1, 2, 3, 4,* und *SAM* abbilden. Durch diese adaptive Klassenstruktur kann jede denkbare Annotationsaufgabe abgebildet werden.

Ein Label beschreibt ein atomares Element, welchem eine der Klassenentitäten zugeordnet ist. Darüber hinaus kann ein Label über die optionalen Felder *Text, Wert, Kommentar* und *Status* verfügen. Die Bedeutung dieser Felder kann vom Versuchsleiter frei definiert werden. So kann der *Wert* wahlweise für eine Intensität, Konfidenz oder ähnliches stehen. Labels können in den drei Grundformen *Ereignis Label, Diskretes Label* und *Kontinuierliches Label* vorliegen. Ein *Ereignis Label* enthält neben seiner Entität und optionalen Parametern nur einen Zeitpunkt als weitere Information. Es dient zum Markieren von Ereignissen ohne zeitliche Ausdehnung und wird durch eine Fahne in der Labelspur dargestellt. Diskrete und kontinuierliche Labels haben hingegen eine zeitliche Ausdehnung, definiert durch Start- und Endzeitpunkt des Labels. Diese werden als rechteckige Box in der Labelspur dargestellt. Die Farbe des Labels entspricht dabei der Farbe, welche der gewählten Entität zugeordnet ist. Kontinuierliche Labels bieten als weitere Option die Möglichkeit einen kontinuierlichen Verlauf darzustellen. Dieser ist durch Stützstellen definiert, welche wahlweise linear, als Treppenfunktion, durch B-Splines[16] [59] oder durch geglättete Approximation verbunden sind. Abbildung 3.24 zeigt eine Auswahl der beschrieben Möglichkeiten der Darstellung und Modellierung von Labelinformationen.

Labels können durch grafische Interaktion mit der Maus frei in einer Labelspur modelliert werden. Die Anzahl der Labels unterliegt dabei keiner Einschränkung. Zeitlich überlappende Labels sind innerhalb einer Spur nicht möglich, können aber jederzeit durch Hinzunahme einer Hilfsspur modelliert werden. In den meisten Fällen ist der Bedarf nach überlappenden Labels ein Hinweis auf unzulänglich definierte Annotationsschemata, es sollte erst nach reiflicher Überlegung auf die Hilfsspurmodellierung zurückgegriffen werden.

[16] `https://en.wikipedia.org/wiki/B-spline` (Stand: 3. Juni 2019)

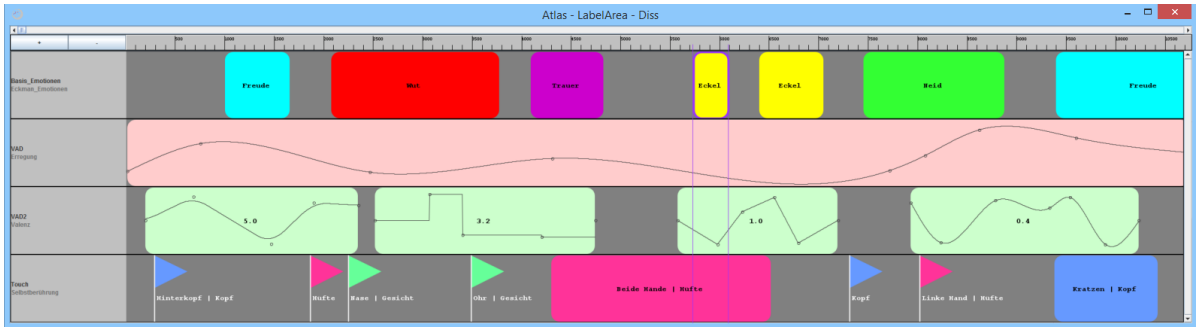

Abbildung 3.24.: Darstellung der ATLAS Labelspuren. Die erste Zeile zeigt eine rein kategorielle Labelspur, darin ausgewählt das aktuell gewählte und bearbeitet Label. Die zweite Zeile visualisiert eine rein kontinuierliche Labelspur. Die dritte Zeile zeigt eine Kombination beider Varianten mit dargestelltem Label, optionalem Labelwert und kontinuierlicher Modellierung. Kontinuierliche Modellierungen von links nach rechts: Approximation, Stufen, lineare Interpolation und B-Spline Interpolation. Die letzte Spur zeigt die Option neben Labels mit zeitlicher Ausdehnung auch ereignisbasierte Flaggenlabels zu verwenden. Alle gezeigten Optionen lassen sich frei kombinieren.

3.5.2. Aktive und passive Hilfsfunktionen

ATLAS bietet eine umfangreiche Palette von Hilfsfunktionen. Dabei wird zwischen aktiven und passiven Hilfsfunktionen unterschieden. Aktive Hilfsfunktionen können den Daten- und Labelbestand durch schreibende Zugriffe aktiv verändern, während passive Funktionen lediglich die Darstellungseigenschaften der Daten und Labels beeinflussen. Im folgenden werden lediglich die wichtigsten Funktionen beschrieben, für eine vollständige Übersicht sei erneut auf die ATLAS Online Dokumentation verwiesen.

LabelProperties Fenster

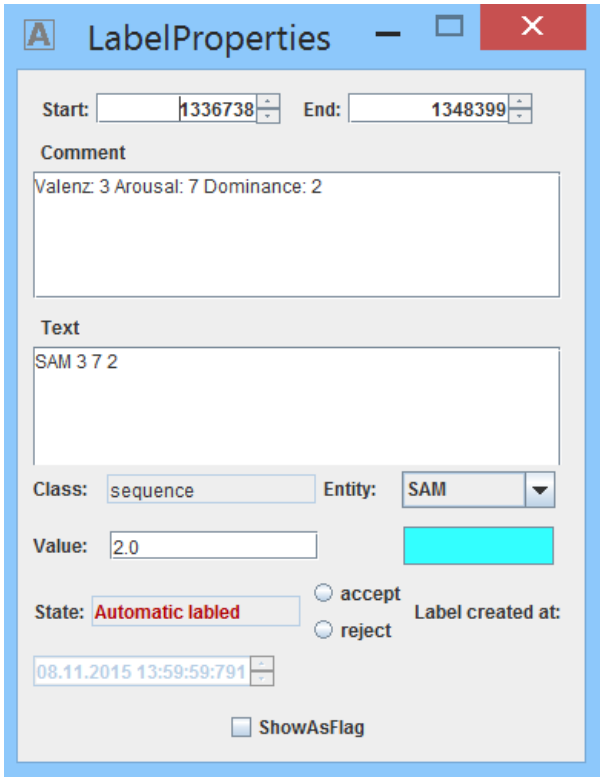

Abbildung 3.25.: Screenshot des ATLAS LabelProperties Fensters

Dieses Fenster (siehe Abbildung 3.25) zeigt die aktuellen Eigenschaften eines Labels an und kann diese verändern. Die wesentlichen Eigenschaften eines Labels können zwar auch durch Drag&Drop verändert werden, im LabelsProperties Fenster können die Werte jedoch, falls notwendig, exakter verändert werden. Insbesondere bei millisekundengenauer Positionierung eines Labels ist dies sehr hilfreich. Die Erzeugungsmethode eines Labels wird dargestellt, dabei wird zwischen *manuell* und *automatic* unterschieden. Automatisch beschreibt dabei die Herkunft eines Labels entweder aus automatisch verarbeitenden Logdateien eines Experimentes oder die Entstehung durch Vorschläge aktiver Lernverfahren. Im Falle automatischer Labels kann weiter zwischen *akzeptiert* (da korrekt) und *abgelehnt* (falscher Vorschlag) unterschieden werden, dies wird den weiteren aktiven Lernprozess beeinflussen.

ScalarToLabels Fenster

In vielen Fällen werden kontinuierliche Eingangsdaten mittels Schwellwertentscheidung in diskrete Labels umgewandelt. Dies geschieht beispielsweise beim Transfer kontinuierlicher Klassifikatorausgaben hin zu Klassenlabels, aber auch bei kontinuierlich abgetasteten Sensor-

intensitäten, welche als binäre Eingangsgröße benötigt werden. Dieser Vorgang ist in ATLAS modelliert, nach Auswahl eines Datenkanals und Schwellwertes, können Labels mit entsprechend gewählten Voreinstellung erzeugt werden. Im Falle mehrdimensionaler Eingangswerte können so auch Klassenentitäten automatisch zugewiesen werden.

Clear and Shift Labels Fenster

Die Funktion eine Labelspur zu leeren kommt im wesentlichen in frühen Phasen des vertrautmachens mit der Umgebung vor. Relevanter ist hingegen die Möglichkeit alle Labels einer oder mehrerer Labelspuren um einen konstanten Offset zu verschieben. Dies erfolgt häufig wenn nach der Annotationsphase eines Experimentes festgestellt wird, dass zumeist aufgrund technischer Fehler, der Nullpunkt der Datenspur an einem anderen Zeitpunkt liegt als zunächst angenommen[17].

Search for Labels Fenster

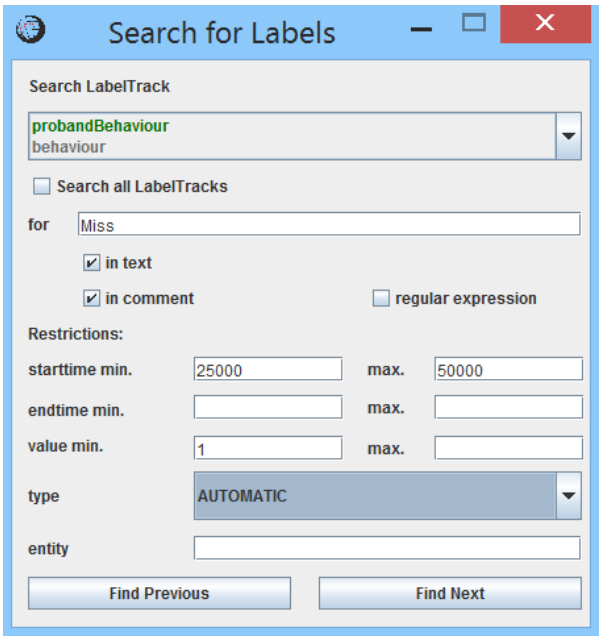

Abbildung 3.26.: Screenshot des ATLAS Search for Labels Fensters

In umfangreichen Projekten mit zahlreichen Labelklassen, Entitäten und Labelspuren besteht schnell die Gefahr den Überblick zu verlieren. ATLAS bietet hier, neben der hier nicht beschrieben tabellarischen Darstellungsmöglichkeit von Labels, die Möglichkeit mittels filtern

[17] Vorgekommen zum Beispiel bei den Datensätzen LMC und Teilen des EmoRec II

nach Labels zu suchen. Die Suche ist dabei so flexibel gehalten, dass nach einer beliebigen Kombination von Suchkriterien gefiltert werden kann. Abbildung 3.26 zeigt die umfangreichen Möglichkeiten der Suchfunktion.

LableLegend Fenster

Dieses Fenster zeigt eine farbige Legende aller verwendeten Klassen und ihrer Entitäten in ihrer Baumstruktur an. Entgegen dem Klasseneditor kann die Struktur hier nicht geändert werden.

ClassEditor Fenster

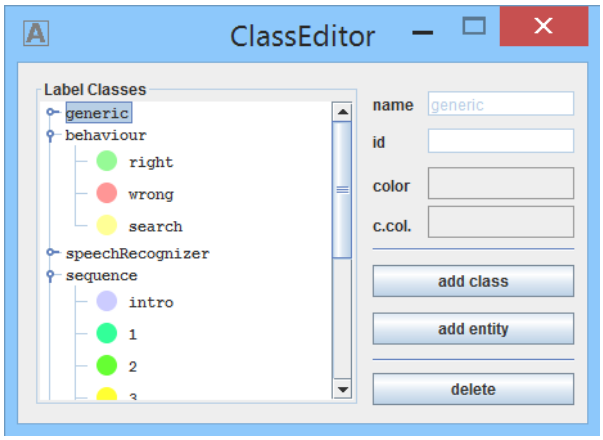

Abbildung 3.27.: Screenshot des ATLAS ClassEditor Fensters

Der Klasseneditor (Abbildung 3.27) dient zur Erstellung und Veränderung von Klassen und ihren Entitäten. Er stellt das zentrale Element bei der Definition von Annotationsschemata dar. Im Falle von Änderungen an der Klassenstruktur nach erfolgter Annotation kann es zu Konflikten kommen, sofern in der Annotation bereits verwendete Klassen und Entitäten geändert oder gelöscht werden. ATLAS führt daher vor jeder Änderung eine Konsistenzprüfung durch.

ProjectCollection Fenster

In vielen Fällen kann es sinnvoll sein schnell zwischen zahlreichen ATLAS Projekten zu wechseln. Dies kann zum einen bei der Bearbeitung von Experimenten mit zahlreichen Probanden durch den Versuchsleiter sinnvoll sein, zum anderen für Präsentationen bei denen flexibel zwischen verschiedenen Themengebieten und somit Demonstratoren umgeschaltet werden soll.

Für diese Fälle kann eine Liste von Projekten mit zugehörigen grafischen Anordnungen aller ATLAS Fenster in einer Kollektion definiert werden. Durch einfaches klicken auf einen Eintrag dieser Liste wird das aktuelle Projekt ohne Rückfrage geschlossen, das neu gewählte Projekt geladen und die Fensteranordnung entsprechend eingestellt.

TrackEditor Fenster

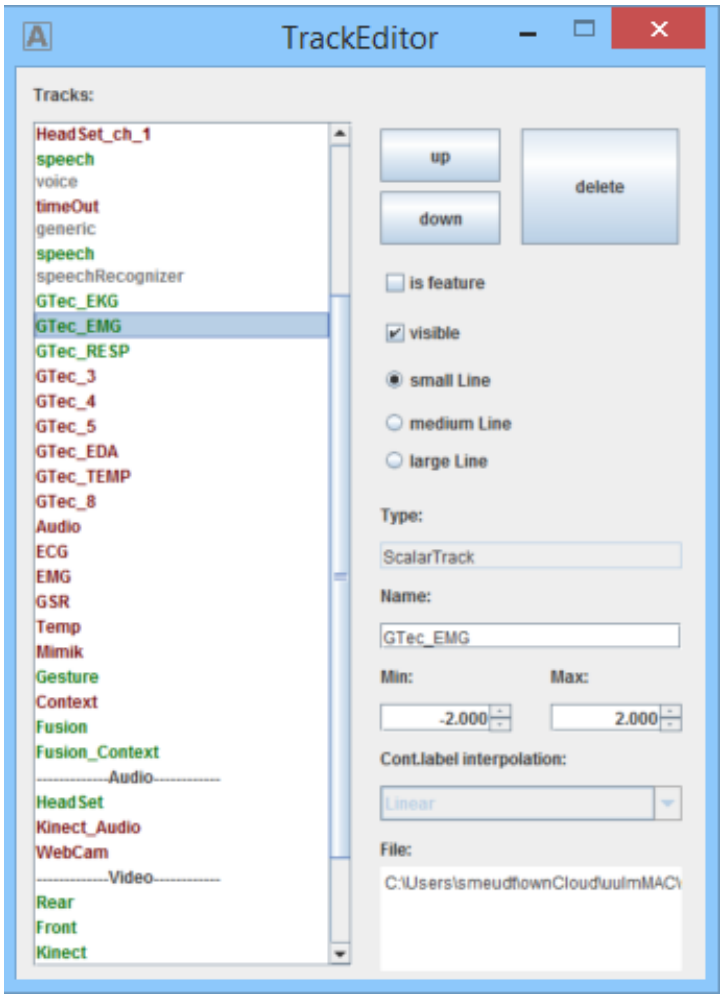

Abbildung 3.28.: Screenshot des ATLAS TrackEditor Fensters

Dieser Editor (Abbildung 3.28) zeigt alle im Projekt vorhanden Daten-, Label-, Audio- und Videospuren an. Hier können die Spuren in ihrer Anzeigereihenfolge, Größe und Sichtbarkeit geändert werden. Weiter wird im Falle von Datenspuren die Skalierung definiert. Die Option *ist Merkmal* (engl. is Feature) gibt eine Datenspur als Merkmalsdaten für die Verwendung aktiver Lernverfahren frei.

Video und Playback Kontrolle

Neben der Annotation ist die synchrone Wiedergabe multimodaler Daten ein zentrales Element von ATLAS. Um dies zu ermöglichen bestehen verschiedene Möglichkeiten der Wiedergabe. Zentral ist das ATLAS Kontrollfenster (Abbildung 3.29) mit den Möglichkeiten des Abspielens in normaler Geschwindigkeit, des bildweisen Springens und der wiederholten Wiedergabe in einer Endlosschleife. Anfang und Endpunkt der wiederholten Wiedergabe werden dabei wahlweise durch ein ausgewähltes Label oder einen frei definierbaren Ein- und Aussprungspunkt definiert. Darüber hinaus kann die Wiedergabegeschwindigkeit aller Videos und Daten bis zum Faktor 10 beschleunigt oder verlangsamt werden. Die akustische Wiedergabe ändert dabei entsprechend die Tonhöhe (Frequenz). Sind in einem Projekt mehrere Videos vorhanden können diese in ihrer Größe skaliert werden und automatisch auf dem Bildschirm angeordnet werden.

Abbildung 3.29.: Screenshot des ATLAS Haupt und Playback Fensters

3.5.3. Aktive und passive Schnittstellen

Neben den internen Hilfsfunktionen bietet ATLAS auch die Möglichkeit mit externen Programmen mittels Datenschnittstellen zu interagieren. Dabei wird, wie auch bei den internen Funktionen, zwischen aktiven und passiven Schnittstellen unterschieden. Passive Schnittstellen kennzeichnen sich dadurch, dass in ATLAS Dateien auf einen Datenträger geschrieben werden, welche dann von beliebigen anderen Programmen geladen und bei Bedarf verändert werden können. Anschließend erfolgt ein Reimport in ATLAS. Die beiden Programminstanzen kommunizieren also nicht direkt miteinander. Anders stellt sich dies bei aktiven Schnittstellen dar, hier kommuniziert ATLAS direkt mit einer im Hintergrund laufenden Instanz eines anderen Programmes mittels Datenpipelines direkt auf der Central Processing Unit (CPU). Änderungen an den Daten und Labels durch andere Programme werden somit automatisch in ATLAS sichtbar.

Exportfunktionen

ATLAS bietet die Möglichkeit Labelspuren neben dem intern verwendeten Extensible Markup Language (XML) Format, definiert in der ATLAS Onlinehilfe auch im gängigen Komma separierte Werte (engl. comma sepperated value) (CSV) Format zu exportieren. Die Felder jedes Labels werden dabei in eine Zeile geschrieben. Im Falle kontinuierlicher Labels werden die Stützstellen und die gewählte Darstellungsform gespeichert. Darüber hinaus besteht die Option kontinuierliche Labels unter freier Wahl einer Abtastrate als Folge von Abtastwerten zu exportieren. Die Abhängigkeit die genaue Definition der z.B. Spline-Funktionsparameter zu kennen entfällt dabei. Neben den Exportfunktionen besteht auch die Möglichkeit alle, dem Projekt zugehörigen Dateien (zumeist XML), außerhalb von ATLAS zu editieren, ohne einen expliziten Export durchführen zu müssen. Wurden bei der Änderungen die Formatdefinitionen berücksichtigt, werden Änderungen beim nächsten Laden des Projektes automatisch übernommen.

Importfunktionen

Um eine möglichst universelle Einsatzmöglichkeit und Kompatibilität sicherzustellen, können in ATLAS zahlreiche Datenformate importiert werden. Dazu gehören derzeit der Import von:

- Audiodaten im Resource Interchange File Format Wave Extention (WAVE) und MPEG-2 Audio Layer III (MP3) Format.
- Videos in jedem durch VLC unterstützen Format[18]
- Klassendateien im XML Format.
- ATLAS Labelspuren im ATLAS XML Format, AVEC Format, Volker FLK Transkriptionen, sowie zahlreicher SFB-TRR 62 intern definierter Formate der verschiedenen Arbeitsgruppen.
- Datenspuren im XML, CSV, MATLAB (.mat), Java Memory Map (.raw) und SFB-TRR 62 interne Formate.

Durch Definition und Implementation geeigneter Import Filter, welche ein Abbildung der Information vom Eingangsformat in ATLAS interne Formate definieren, kann diese Liste beliebig erweitert werden.

[18] Eine Liste aller unterstützen Codecs findet sich unter `https://wiki.videolan.org/VLC_Features_Formats` (Stand: 3. Juni 2019)

Execute Command Fenster

Hier besteht die Möglichkeit externe Systemaufrufe auf Betriebssystemebene zu definieren und auszulösen. Typischerweise werden von hier aus externe Skripte aufgerufen, welche umfangreiche Funktionen beinhalten können. Da der Prozess von ATLAS aus gestartet wird, sein Resultat jedoch nicht automatisch in ATLAS übernommen wird (für die Übername ist ein Import notwendig), kann diese Funktion als aktiv-passiv hybride Funktion angesehen werden.

Execute Matlab Funktion

Diese Option, wie in Abbildung 3.30 dargestellt, stellt eines der mächtigsten Mittel dar um von ATLAS aus interaktiv mit Matlab zu kommunizieren. Von hier aus können alle Inhalte eines ATLAS Projektes mittels Daten Pipeline an Matlab übertragen werden. In der aufgerufen Matlab Instanz kann dann zunächst optional ein einfaches Kommando ausgeführt werden. (z.B. hinzufügen von Pfaden) um anschließend eine beliebige Matlab Funktion aufzurufen, welche auf den übergebenen Daten agiert. Diese Matlab Instanz kann dann blockierend oder nicht blockierend Algorithmen ausführen. Blockierend bedeutet dabei, dass keine Änderungen innerhalb von ATLAS möglich sind während der Matlab Prozess nicht beendet ist. Da diese Prozesse teilweise umfangreiche Algorithmen aufrufen, kann alternativ im nicht blockierenden Fall in ATLAS weiter gearbeitet werden. Nach Beendigung der Matlab Prozesse werden deren Resultate automatisch in ATLAS übernommen und angezeigt. So können z.B. beliebige aktive Lernalgorithmen oder Klassifikationsalgorithmen in Matllab implementiert und erprobt werden und deren Funktionalität mit geringem Aufwand zu ATLAS hinzugefügt werden. Eine Neu- oder Reimplementation für ATLAS in Java ist nicht notwendig. Auch müssen die aufgerufenen Matlab Funktionen nicht zwingend auf dem selben Personal Computer (PC) rechnen wie die ausgeführte Matlab Instanz. So ist es möglich auch die gesamte Rechenkapazität eines Computeservers oder sogar Clusters zu nutzen um enorme Rechenperformanz für einen echtzeit aktiven Lernprozess zur Verfügung zu stellen.

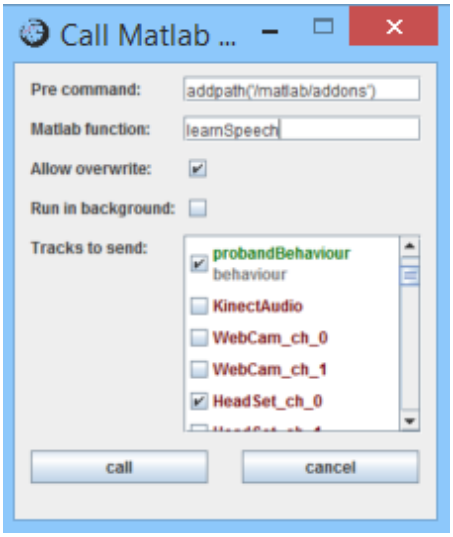

Abbildung 3.30.: Screenshot der ATLAS - Matlab Schnittstelle

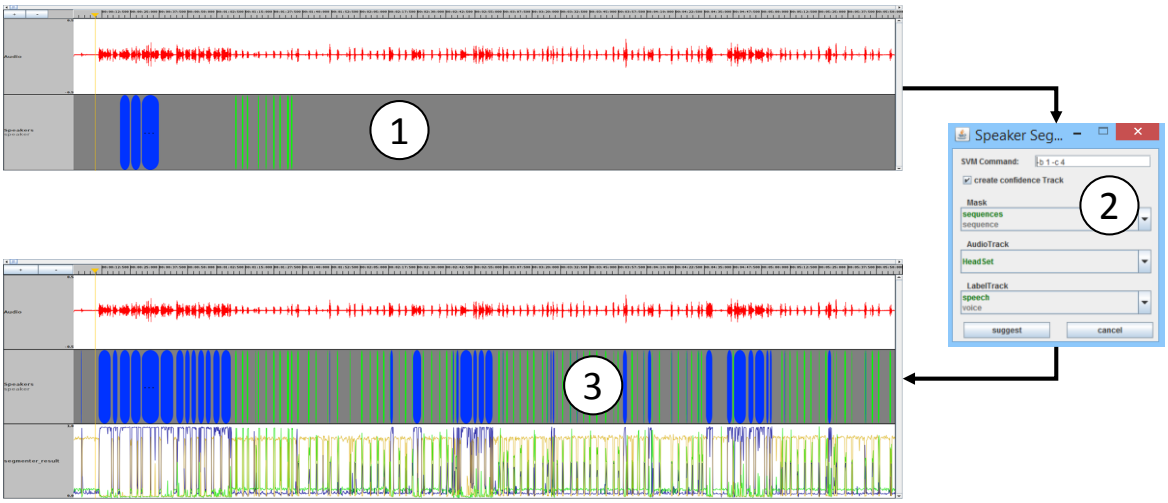

Abbildung 3.31.: Screenshot der automatischen ATLAS Sprecher Segmentierung. Im ersten Schritt werden kurze Abschnitte aller Sprecher manuell annotiert. Im zweiten Schritt werden die Einstellungen der automatischen Segmentierung gewählt. Im dritten Schritt zeigt ATLAS das Resultat der automatischen Segmentierung an. Ebenfalls dargestellt ist die optionale Datenspur der Segmentierungskonfidenzen.

3.5.4. Aktive Annotationsschnittstellen und -Funktionen

ATLAS verfügt über eingebaute Funktionen, welche aktives lernen in verschiedenen Ausprägungen unterstützt. In den folgen drei Abschnitten werden diese zunächst anhand eines konkret umgesetzten Spezialfalls der halbautomatischen Sprechersegmentierung gezeigt. Anschließend wird eine generelle ATLAS interne Implementierung eines SVM Klassifikators beschrieben, gefolgt von einer weiteren Generalisierung, welche durch externe Skripte und Funktionen realisiert werden kann.

SpeakerSegmentation Fenster

Eine relative einfache Aufgabe stellt das automatische segmentieren von verschiedenen Sprechern in einer akustischen Spur dar. Hierzu annotiert der Nutzer zunächst einige Sprachsegmente aller an einem Gespräch beteiligten Personen. Über den Menüpunkt *Speaker Segmentation* wird dann der Dialog zur automatischen Segmentierung der verbleibenden Daten aufgerufen. Hier wählt der Nutzer zunächst die Parameter der SVM sofern diese von den Default-Parametern abweichen sollen. Weiter wird der zu verwendende Audiokanal und die Labelspur gewählt, sofern mehrere im Projekt enthalten sind. Über eine Maske kann die Analyse und Segmentierung auf Teilbereiche der Zeitachse beschränkt werden. Anschließend wird der Vorschlagsprozess gestartet. ATLAS berechnet dabei zunächst die Signalenergie der Audiospur. Bereiche deren

Energie sich im unteren 3% Perzentiel befinden werden intern als zusätzliche Stilleklasse automatisch annotiert. Auf der gesamten Audiospur wir nun eine MFCC Merkmalsmatrix berechnet (Siehe Abschnitt 4.3.1). Auf Basis der MFCC Merkmale und der um Stille erweiterten Segmente wird nun ein Klassifikator trainiert, mit welchem die nicht annotierten Bereiche klassifiziert werden, und dem Nutzer als Vorschlag angeboten werden. Der Annoteur überprüft diese Vorschläge, ist die Qualität bereits ausreichend ist der Prozess abgeschlossen, andernfalls werden die vorgeschlagenen Labels vom Annoteur angenommen oder abgelehnt. Mit den angenommen Labels wird das Training und die Klassifikation iterativ wiederholt, während zurückgewiesene oder nicht betrachtete Labels vor dem generieren neuer Vorschläge entfernt werden. In den meisten Fällen ist bereits nach ein bis zwei Iterationen eine hervorragende Segmentierung möglich. Abbildung 3.31 zeigt Menue und Resultat einer beispielhaften Segmentierung zweier Sprecher. Dabei wird optional auch die Konfidenz der SVM mit ausgegeben, diese optionale Ausgabe kann auch unterdrückt werden.

SuggestLabels Fenster

Diese Funktion verallgemeinert den zuvor beschriebenen Ansatz. Dem Annoteur wird hier die Möglichkeit gegeben eine beliebige, als Merkmal markierte Datenspur, als Merkmalsvektoren für die Klassifikation zu wählen. Auch hier besteht die Möglichkeit den Zeitraum durch Maskierung zeitlich einzuschränken. Entgegen der Sprechersegementierung existiert hier keine automatische Erkennung einer Rest- oder Garbageklasse (z.B. Stille). Ist im konkreten Falle eine solche Klasse relevant, muss diese vom Annoteur angelegt und in der Labelspur annotiert werden. Auch in diesem Fall wird eine SVM als Klassifikator eingesetzt deren Parameter frei wählbar sind. Da im Falle allgemeiner und wesentlich komplexerer Daten nicht mit einem solch guten Ergebnis wie im Falle der Sprecher Segmentierung gerechnet werden kann, kann hier die Vorschlagsmenge durch den *amount* Parameter begrenzt werden. Geringe Werte bedeuten dabei, dass nur wenige, hoch konfidente Vorschläge präsentiert werden. (Ein amount Parameter von 100% wirkt wie eine Deaktivierung der Konfidenzschwelle). Über den *granularity* Parameter wird eine Glättung der Ausgabe bewirkt. Geringe Granularität bewirkt kurze aber unter Umständen zerklüftete Vorschläge, während ein hoher Wert unter Umständen nicht mehr in der Lage ist Segmentgrenzen erfolgreich zu trennen. Die Wahl der Parameter hängt stark vom jeweiligen Klassifikationsproblem ab und muss individuell angepasst werden. Abbildung 3.32 zeigt das Einstellungsfenster sowie ein Beispiel von generierten Vorschlägen der Klassen positive und negative Valenz. Auch hier können Vorschläge angenommen, zurückgewiesen oder nicht beachtet werden. Wird der Prozess iterativ gestartet werden angenommene Labels im nächsten Umlauf berücksichtigt um die Qualität der Vorschläge zu verbessern.

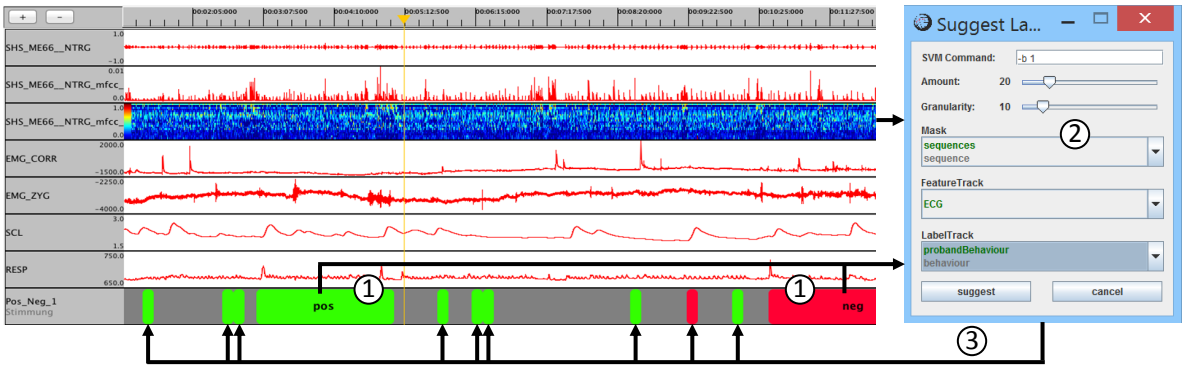

Abbildung 3.32.: Automatische Labelvorschläge unterbreitet durch ATLAS. Hier gezeigt am Beispiel zweier Klassen (*positve Valenz* und *negative Valenz*). Zunächst werden einige Bereiche manuell annotiert (1). Im SuggestLabels Fenster (2) werden anschließend alle Parameter gewählt (Mindestens: eine Merkmalsmatrix und Labelspur). ATLAS erzeugt daraufhin Vorschläge mit maximaler Konfidenz zu allen Klassen (3).

ActiveLearning Fenster

In der letzten Ausbaustufe der ATLASeigenen Klassifikationsmethoden ist ein aktiver Lernzyklus am Beispiel einer SVM vollständig implementiert. Auch hier werden zunächst Merkmalsspur, Labelspur und eine optionale Maske als Quelle gewählt. Nach Wahl der SVM Parameter beginnt das Training des Klassifikators im Hintergrund. Der iterative Ablauf der Annotation ist als solcher nicht mehr direkt erkennbar. Sobald überschüssige CPU Ressourcen zur Verfügung stehen werden Vorschläge auf Basis bisheriger Annotationen, angenommener und abgelehnter Vorschläge generiert und interaktiv eingeblendet. Alte Vorschläge werden dabei unter Umständen entfernt oder abgeändert, solange sie noch nicht vom Annoteur als richtig akzeptiert wurden. Manuell erzeugte Labels werden generell nicht verändert. Liegen zu Beginn der Session vortrainierte Klassifikatoren vor (z.B. aus einer vorherigen Session oder eines anderen Probanden) können diese geladen werden. Der aktuelle Stand eines Klassifikators kann umgekehrt zur späteren Verwendung gespeichert werden. Neben dem veränderten Parameterauswahlfenster ist die Ausgabe diese Verfahrens nicht vom vorherigen Vorschlagsverfahren zu unterscheiden. Lediglich die Kontrolle über den Start eines weiteren aktiven Lernzykluses ist auf Basis der Systemressourcen automatisiert, neue Labels erscheinen daher scheinbar auf magische Art und Weise aus dem nichts. Dabei kann sich das aktive Lernszenario im Falle ausreichender Systemressourcen in mehreren Iterationen zeitgleich befinden. Dies bedeutet, sind aktuell ausreichend Ressourcen vorhanden und eine genügend große Änderung an den Labels vorgenommen worden, startet ein aktiver Lernzyklus auf Basis der aktuellen Merkmale und Labels, ungeachtet ob ein weiterer Zyklus im Training ist. Lediglich bei der Klassifikation wird überprüft ob neu generierte Vorschläge eventuell in Konflikt mit anderen Schleifen oder

Nutzereingaben stehen. Tritt eine Konflikt auf, haben zunächst Nutzereingaben Vorrang, in Falle von Konflikten automatischer Labels wird das Label übernommen dessen Klassifikator auf dem aktuellsten Stand der Labels trainiert ist.

Externes Aktives Lernen

Reichen die auf SVM beschränkten internen Möglichkeiten von ATLAS nicht mehr aus, oder soll auf externe Computeserver zurückgegriffen werden, wird die zuvor beschriebene Execute Matlab Function Option verwendet. Die aufgerufene Matlab Funktion bekommt die Merkmale und Labels, welche zum aktiven Lernen verwendet werden sollen, übergeben, führt die Vorschlagsberechnung durch und übergibt diese zurück an ATLAS. ATLAS wird so um die Funktionsmächtigkeit von Matlab erweitert. Auch dieser Vorschlagsprozess kann im Hintergrund erfolgen. Abbildung 3.33 zeigt einen Screenshot des OCSVM basierten Ansatzes [253, 254], welcher im Rahmen einer Begehung des SFB-TRR 62 als externe Matlab Funktion realisiert und demonstriert wurde. Das gezeigte Diagramm ist nicht Bestandteil der üblichen ATLAS Ansicht, es wurde lediglich zur Visualisierung im Rahmen der Demonstration angezeigt. Realisiert wurde die automatische Erkennung von emotionalen Reaktionen im akustischen und visuellen Kanal des uulmMAC.

3.5.5. Verteilte Annotation

Im Falle umfangreicher Datensätze wird in manchen Fällen die Ressourcengrenze eines einzelnen handelsüblichen PCs erreicht. Tritt dies im Falle aktiven Lernens auf, besteht über die integrierte Verbindung zu Matlab die Möglichkeit die Trainings-Algorithmen nach Matlab auszulagern, von Matlab aus kann dann über den *Distributed Computing Server*[19] und die *Parallel Computing Toolbox*[20] via Netzwerk auf Computeserver oder Cluster Infrastrukturen, wie das Baden-Württembergische Universitäts Rechen Cluster (bwUniCluster)[21] der *bwHPC-C5* Initiative[22] zurückgegriffen werden. Durch die hierbei enorme erreichbare Resourcenmenge ist das Resourcenproblem des Echtzeit aktiven Lernens gelöst.

Ungelöst ist dabei zunächst allerdings das Problem möglicher zu geringer lokaler Ressourcen für das Abspielen multimodaler Datensätze mit großer Anzahl oder datenintensiven Modalitäten. Für diesen Fall ist ATLAS in das MAR^{2}S integriert und verfügt über die Möglichkeit

[19] `https://de.mathworks.com/products/distriben.html` (Stand: 3. Juni 2019)
[20] `https://de.mathworks.com/products/parallel-computing.html` (Stand: 3. Juni 2019)
[21] `https://www.uni-ulm.de/einrichtungen/kiz/service-katalog/high-performance-computing/` (Stand: 3. Juni 2019)
[22] `http://www.bwhpc-c5.de/` (Stand: 3. Juni 2019)

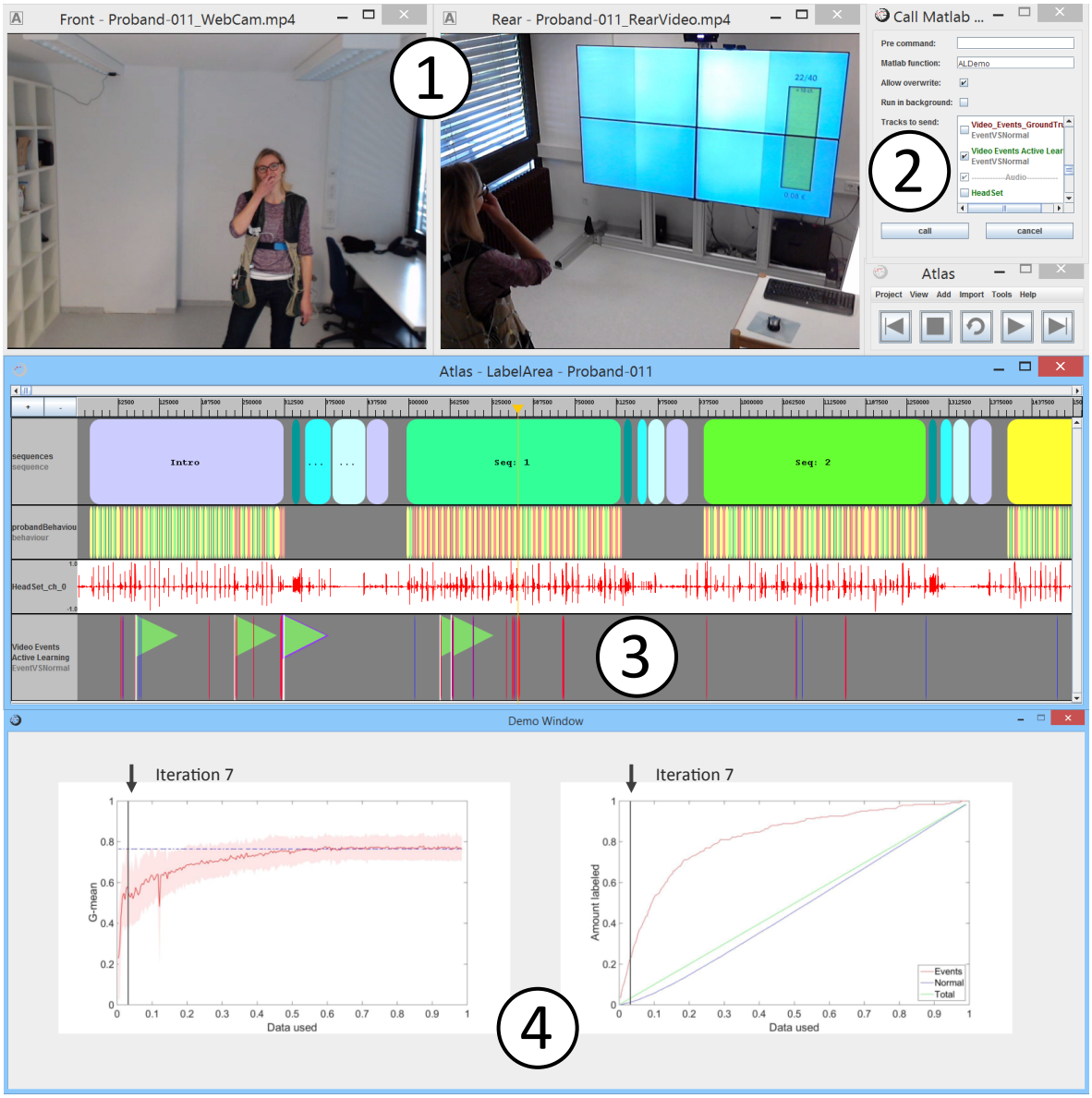

Abbildung 3.33.: Screenshot einer Demonstration des aktiven Lernverfahrens in ATLAS. Beispielhaft dargestellt ist die semiautomatische Annotation des Videokanals (1) mittels der ATLAS Matlab Schnittstelle (2). Die Berechnungen des aktiven Lernalgorithmus wurden bei der Demonstration auf einem remote Matlab Computeserver in Echtzeit durchgeführt. In der Labelspur (3) markieren grüne Flaggenlabels die selektierten Zeitpunkte, die als maximal relevant identifiziert wurden. Diese werden durch den Annoteur klassifiziert. Annotierte Instanzen vorangegangener Iterationen sind durch blaue (*keine Emotion*) und rote (*Emotion*) Labels markiert. Das Demofenster (4) zeigt eine Analyse des aktiven Lernprozesses. Dieser ist nur zu Demonstrationszwecken eingeblendet, diese Ansicht erfordert ein vollständiges Durchlaufen des Verfahrens in der Vorbereitung der Demonstration.

mehrere ATLAS Instanzen via MAR^{2}S zu synchronisieren. Dabei operieren alle synchronen Instanzen auf einem identischen Datenbestand, wobei für Daten welche nur wiedergeben werden (typischerweise alle Spuren außer Labelspuren) zur Performanzsteigerung und Entlastung der Netzwerk Infrastruktur auf dem jeweiligen PC als lokale Cache-Kopie vorhanden sein können. Diese Architektur erlaubt es einen Datensatz auf beliebig vielen PCs wiederzugeben und simultan zu bearbeiten. Die Anzeige Einstellungen von ATLAS können dabei auf jedem PC individuell angepasst werden (z.B. Exporten und Novizen Ansicht).

Abbildung 3.34 zeigt eine mögliche verteilte Annotationskonfiguration dreier ATLAS Instanzen bei der alle Optionen des verteilten Annotierens verwendet werden. Systemmanager und Datenserver könnten dabei auch auf einem der lokalen PCs ausgeführt werden, sofern die Systemressourcen dies zulassen.

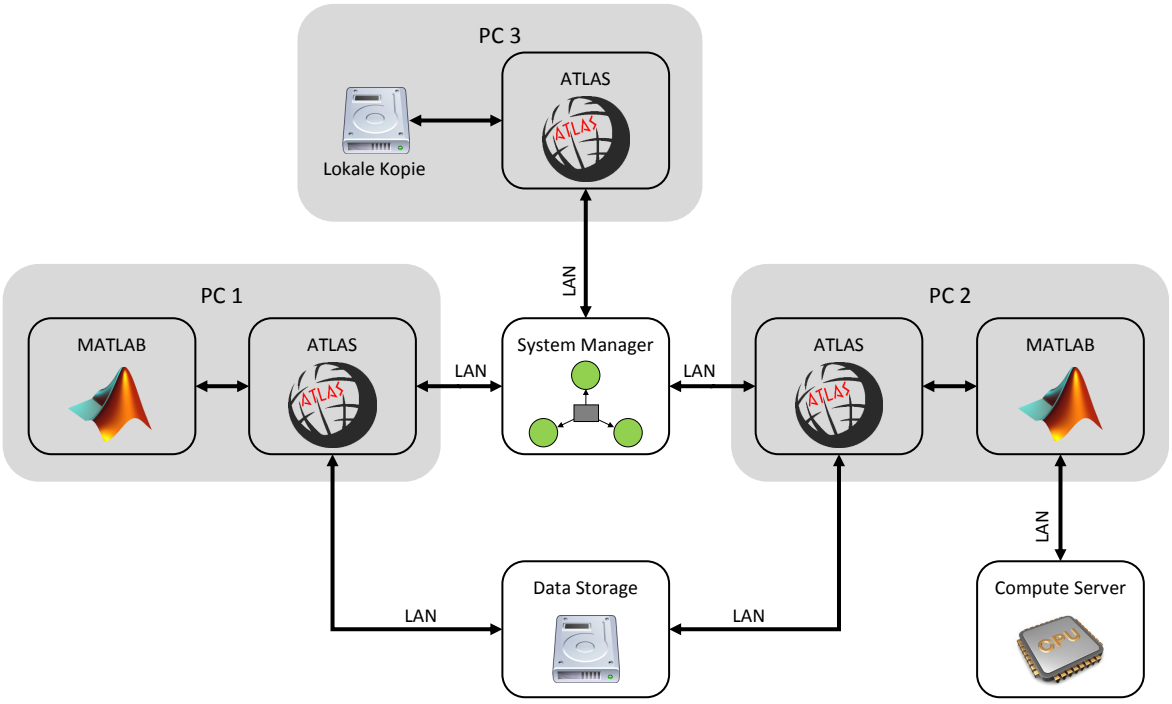

Abbildung 3.34.: Mögliche verteilte Annotationssituation mit drei ATLAS Instanzen, synchronisiert via Middleware. Die ATLAS Instanz auf PC3 greift via lokaler Kopie auf die Daten zu, Instanz 1 und 2 via Netzwerk auf einen Datenstorage. Die Instanzen 1 und 2 verwenden zusätzlich je eine Matlabinstanz zur interaktiven Berechnung komplexer Algorithmen (z.B. aktive Lernverfahren). Die Matlabinstanz 2 greift zusätzlich via Netzwerk auf einen Computeserver zu. Alle gezeigten Module können in beliebiger Anzahl und Struktur kombiniert werden.

3.5.6. Technische Realisierung

Mit der Implementierung von ATLAS wurde bereits in einer frühen Phase dieser Arbeit begonnen. Der Implementierung ging zunächst eine Bedarfs- und Machbarkeitsanalyse voraus. Zunächst wurden die Basis Funktionen *Wiedergabe multimodaler Daten* und *einfache*, zunächst *kategorielle Annotation*. Mit Weiterentwicklung der Datensätze wurden der Funktionsumfang von ATLAS stetig erweitert. Ebenfalls wurden die Möglichkeiten der Annotation kontinuierlich verbessert. Bedarfsorientiert wurden neue Funktionen hinzugefügt und ATLAS an die dezentrale MAR^{2}S Architektur angepasst. In der aktuell letzten Ausbaustufe ist die aktive Lernfunktionalität vollumfänglich hinzugefügt worden. In einer zukünftigen Version ist eine live Echtzeitanbindung an datenerzeugende Module der MAR^{2}S Architektur vorgesehen.

ATLAS liegt gegenwärtig in der stabilen Version 1.6.8 vor. Es wurde bis Version 0.8 in Java jdk 1.7 implementiert. Seid dem Versionssprung zu 1.0 erfolgt die Implementierung 64Bit kompatibel in Java jdk 1.8. bei der Implementierung von ATLAS wurde neben den Java 1.8 Systembibliotheken auf folgende Bibliotheken zurückgegriffen:

- Medienwiedergabe: vlcj 2.1.0 und VLC 2.0
- Interaktive Matlab-Kommunikation: MatlabControl v2010
- SEMAINE Anbindung: JMS 1.1
- XML Parsing: jDOM 2.0.6
- Onlinehilfe: hsViewer 1.4
- Nutzereingabekontrolle: UIV 0.1
- MFCC Extraktion: CoMIRVA 0.36

Bei der Auswahl aller Bibliotheken wurde darauf geachtet diese neben inhaltlichen Kriterien so auszuwählen, dass die Plattform Unabhängigkeit von Java nicht verletzt wird. ATLAS ist somit für die Betriebssystem MS Windows[23] Linux[24], und macOS[25] verfügbar.

Der interne Klassenaufbau von ATLAS ist ähnlich wie das MAR^{2}S möglichst modular gehalten. Klassen kommunizieren dabei neben klassischen Vererbungsstrukturen und Objektgraphen mittels interner Nachrichten Transfers (EventBus) miteinander, welche in einer eigens entwickelten *Message Manager* Komponente realisiert sind. So wird ein ressourcensparendes Koppeln inter-

[23] Erfolgreich getestet unter Windows Vista, 7, 8 und 10
[24] Erfolgreich getestet unter ubuntu 12.04 bis 17.04, CentOS 7.4 und openSUSE 42.3
[25] Erfolgreich getestet unter macOS 10.5 Leopard bis 10.12 Sierra

ner Abläufe realisiert. Der Codeumfang von ATLAS beläuft sich dabei auf ca. 15.000 Zeilen, verteilt auf 119 Klassen, welche die graphischen und technischen Komponenten realisieren. Bei der gesamten Entwicklung wurde dabei auf eine effiziente Implementierung geachtet um dem Problem der zu erwartenden großen Datenbestände Rechnung zu tragen. Insbesondere im Falle der Datenspuren wurde das *Memory Mapped File und IO* Konzept des *java.nio* Paketes[26] genutzt und angepasst. Für verschieden Zoomstufen wird dabei zwischen unterschiedlichen Parametern der Nutzung umgeschaltet. So wird jederzeit eine Übersichtsansicht im RAM gehalten, welche schnell verfügbar ist. Beim Zoomen in Daten werden kurze Elemente der Daten von der Festplatte nachgeladen, wodurch ein ruckelfreies Zoomen realisiert ist. Auf eine tiefgreifende Fokussierung der Beschreibung der Implementierung wird an dieser Stelle verzichtet. Bei Interesse kann dazu jederzeit mit dem Autor Kontakt aufgenommen werden. Auch eine Weitergabe des Source Codes ist nicht ausgeschlossen.

Alle ATLAS eigenen Datenformate sind im XML Format gehalten, eine Ausnahme bilden hier lediglich die *raw* Dateien des der Mapped File Puffers welche aus Performanzgründen im overheadfreien Byteformat gehalten sind. Grundlage jeden ATLAS Projektes bildet zunächst das Projekt XML welches alle zum Projekt gehörenden Spuren und Labelklassen referenziert. Labelklassen und Labelspuren sind ebenfalls im XML Format gehalten. Aufgrund der guten menschlichen Lesbarkeit von XML kann so der Aufbau eines ATLAS Projektes einfach nachvollzogen werden. Eine vollständige Definition aller verwendeten XML Einträge findet sich in der Onlinehilfe und der Dokumentation von ATLAS.

[26] `https://docs.oracle.com/javase/7/docs/api/java/nio/MappedByteBuffer.html` (Stand: 3. Juni 2019)

KAPITEL 4

MASCHINELLE EMOTIONSERKENNUNG

Dieses Kapitel beschreibt Verfahren der maschinellen Emotionserkennung und Informationsfusion zur multimodalen Affekt- und Emotionsschätzung anhand von sensorischen Daten. In den vorstehenden Kapiteln wurde beschrieben, wie die nun eingesetzten Trainingsdaten gewonnen und annotiert werden können. Bei den verwendeten und vorgestellten Klassifikationsverfahren handelt es sich um Verfahren der Literatur, welche derart modifiziert und erweitert wurden, damit sie sich zur Lösung der gegebenen Fragestellungen der Companion-Technologie eignen. Alle vorgeschlagenen Methoden und Architekturen wurden dabei immer mit dem Ziel der praktischen Einsetzbarkeit in prototypischen Systemen gewählt und auf ihre Tauglichkeit hin bewertet.

Nachdem zunächst in Abschnitt 4.1 der Stand der Forschung der maschinellen Emotionserkennung beleuchtet wird, werden im darauffolgenden Abschnitt 4.2 zunächst zwei Verfahren vorgestellt um die Dimensionalität und Instanzenmenge der Trainingsdaten mit dem Ziel der Informationsverdichtung zu reduzieren. Anschließend werden in Abschnitt 4.3 für alle in der Arbeit verwendeten Modalitäten zunächst unimodale Merkmale, Klassifikatoren und Experimente vorgestellt, welche im darauffolgenden Abschnitt 4.4 als Eingangsdaten zur multimodalen Fusion herangezogen werden. Abschnitt 4.4 zu multimodalen Erkennung beschreibt anschließend mögliche Architekturen zur ganzheitlichen Schätzung welche sich bewährt haben. Ebenfalls wird am Ende des Abschnittes zur Multimodalität die Möglichkeit der Verbesserung der Erkennungsleistung durch Einbezug von situativem Kontextwissen beleuchtet.

4.1. Stand der Forschung - Maschinelle Emotionserkennung

Die maschinelle Emotions- und Affekterkennung ist von den vier Säulen dieser Arbeit die am intensivsten beforschte Thematik. Die maschinelle Emotions- und Affekterkennung ist

ein elementares Teilgebiet des *affektiv Computing*. Fragestellungen des affektive Computing sind nicht auf die Bereiche der Companion-Technologie begrenzt. Zu weiteren Anwendungen gehören auch medizinische Anwendungen wie das automatische Erkennen von Schmerz [119, 269, 264] oder die computergestützte Erkennung von Depressionen [58] oder Krankheitsbildern wie der *Post Traumatic Stress Dissorder* (PTSD) [270]. Letztere wird dabei meist durch das US Militär gefördert. In ihrem Kern unterscheiden sich die dabei eingesetzten Methoden nur geringfügig, häufig wird die Anwendung durch Verbundsmitgliedschaften oder bewilligte Fördermittel der Forschergruppen beeinflusst. Die eingesetzten Erkenner sind dabei häufig sehr ähnlich, oder gar identisch, lediglich beim Entwurf spezieller modellgetriebener Merkmale bestehen teilweise geringe Unterschiede. Unterscheide bestehen im wesentlichen bei den eingesetzten Trainingsdatensätzen, bzw. der darin annotierten Grundwahrheiten. So sind zum Beispiel im Falle der AVEC Challenges seid einigen Jahren zwei Subchallenges etabliert, eine zur Schätzung von depressiven Verhaltensmustern und eine zur Erkennung des VAD Zustandes des Nutzers. Bei der methodischen Beschreibung des Standes der Forschung wird daher nicht nach der zugrundeliegenden Anwendungsfragestellung der verschiedenen in der Literatur bekannten Methoden unterschieden.

Die Arbeiten unterscheiden sich in der Anzahl der verwendeten Modalitäten. Zum einen existieren diverse Forschungen zu unimodalen Methoden, deren Ziel ist es dabei die Klassifikation mittels einer einzigen Modalität zu optimieren. Dabei liegt der Fokus in der Regel auf der Entwicklung verbesserter Merkmale [279], der Optimierung von signalverarbeitenden Prozessen [5] (Filterung, Kompression, etc.), Optimierung der physikalischen Eigenschaften von Sensoren (Größe, Abtastgenauigkeit, Abtasraten, etc.) [137] und natürlich dem Entwurf idealer Klassifikationsmethoden angepasst auf die jeweilige Modalität [228, 91, 69].

Multimodale Ansätze hingegen zielen darauf ab aus einer gegeben Anzahl von schwachen unimodalen Erkennungen eine optimale Gesamtschätzung zu erreichen. Dabei sind Fragestellungen der temporalen und spatialen Fusion zu beantworten [69, 208]. Dazu gehört der Umgang mit unimodalen Eingaben, welche auf verschiedenen Zeitskalen erfasst wurden [123, 91], Umgang mit vorübergehend fehlenden und falschen Sensoreingaben [69, 210] und der Umgang mit Unsicherheit [279, 275, 240]. Gründe für fehlende Daten können technische Probleme wie Sensorfehlfunktionen sein oder situative Gründe haben, wenn zum Beispiel der Nutzer nicht in die Kamera schaut, nicht spricht oder keine sensorische Oberfläche berührt. Ein anderes Problem liegt im emotionalen Verhalten selbst. Nicht jedes emotionale Verhalten kann in jeder Modalität erkannt werden. Zum Beispiel verursacht ein kurzes Lächeln nicht unbedingt eine Veränderung der Hautleitfähigkeit. In der Literatur haben sich daher hierarchische Ansätze durchgesetzt. Dabei wird wie in Abbildung 4.1 dargestellt in drei Varianten der hierarchischen Fusion unterschieden.

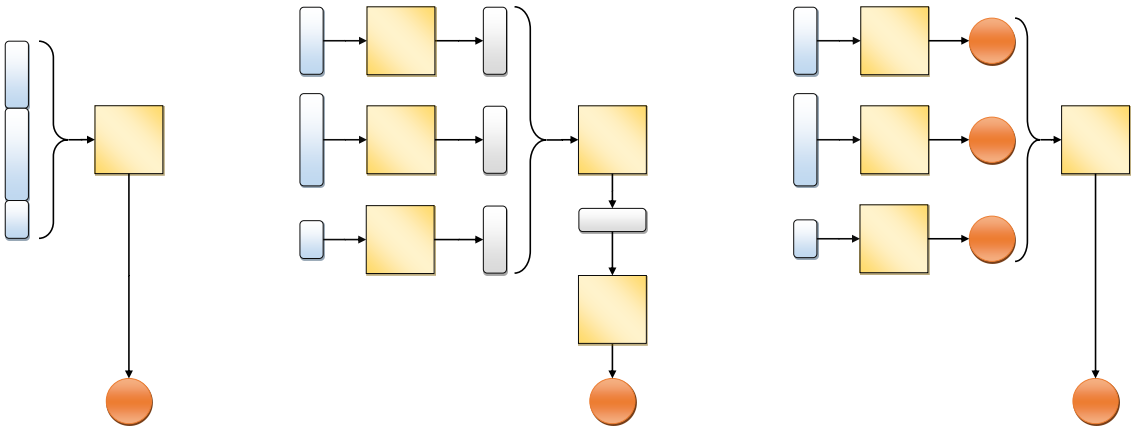

Abbildung 4.1.: Hierarchische Fusionsarchitekturen. Von links nach rechts: Frühe Fusion, hybride Fusion, späte Fusion. Merkmalsvektoren der Eingabeschicht (blau), Klassifikatoren (gelb), Klassenentscheidung (orange), Zwischenrepräsentation (grau).

Bei der *frühen Fusion* werden die Merkmalsvektoren der beteiligten Modalitäten bereits in der untersten Ebene durch Konkatenation zusammengefasst. Dies setzt allerdings voraus, dass diese zu einen gemeinsamen Zeitpunkt extrahiert wurden und die Analysefenster vergleichbare Längen haben. Der frühe Fusionsansatz kann nicht ohne weiteres mit fehlenden Signalen einzelner Modalitäten umgehen, da bei der Vektorkonkatenation vorausgesetzt wird, dass alle Vektoren vorhanden sind. Behelfsweises auffüllen mit Platzhaltern wie Nullvektoren oder Wiederholungen des vorherigen Vektors ist nicht zielführend. Diese Architektur hat sich daher nur in einigen speziellen Anwendungsfällen durchgesetzt.

Der Ansatz der *späten Fusion* geht davon aus, dass jede Modalität zunächst separat betrachtet wird. Dabei kann für jede Modalität ihr optimales Merkmal, mit den jeweils optimalen Parametern eingesetzt werden, ohne dabei Rücksicht auf die spätere Fusion nehmen zu müssen. Die unimodale Klassifikation kann dabei selbst ebenfalls hierarchische Architekturen aufweisen [124]. Jede unimodale Klassifikation liefert an ihrer Ausgansschicht eine Klassenentscheidung aus der selben Klassenmengen wie die der späteren fusionierten Ausgabe. Die Klassenentscheidung kann dabei entweder binär, kategoriell oder kontinuierlich sein und eventuelle zusätzliche Unsicherheitsangaben beinhalten. Da zum Zeitpunkt der Fusion bereits Klassenentscheidungen vorliegen, wird dieser Ansatz auch als *Entscheidungsebenenansatz* bezeichnet. Vorteilhaft ist dabei die lange Zeit unabhängige Betrachtung der Modalitäten, dies erlaubt eine gute praktische Einsetzbarkeit dieses Ansatzes [201, 131] und eine große methodische Freiheit. Auch kann mit fehlende Signalen einzelner Modalitäten gut umgegangen werden. Späte Fusion reicht von einfachen gewichteten Summierungsvarianten bis hin zu komplexen Modellen wie dem Markov Fusionsnetzwerk (MFN), das ein probabilistisches Modell für multimodale und temporale Fusion beschreibt [90, 89]. Nachteilig ist hingen dass

zum Zeitpunkt der Fusion bereits eine starke Informationskompression stattgefunden hat. Die Möglichkeiten der Fusion reduzieren sich dadurch auf das zusammenfassen der Klassenvektoren. Auch eventuelle Querbezüge zwischen Modalitäten auf Merkmalsebene können durch das späte aufeinandertreffen der Modalitäten kaum berücksichtigt werden. In der Literatur ist dies die meistverwendetste Architektur.

Die *Zwischenschicht Fusion* beschreibt einen hybriden Ansatz, der die eingehenden Modalitäten zunächst unabhängig betrachtet, die unimodalen Klassifikatoren haben also große methodische Freiheit. Die Ausgabe der unimodalen Klassifikatoren beinhalten entgegen der späten Fusion noch keine Klassenlabels, sondern entsprechen einer Zwischenrepräsentation. Diese könnten für einen Videokanal beispielsweise einer FACS Kodierung entsprechen, im akustischen Signal einem Voicequality Merkmal entsprechen und im biophysiologischen Kanal einer Repräsentation im VAD Modell. Die darauf aufbauende Fusionschicht bildet diese Zwischenrepräsentation auf die finale Klassenentscheidung ab (z.B. kognitive Belastung). Tiefe neuronale Netze können ebenfalls als solche hybride Fussionsarchitektur gesehen werden [92]. Dabei repräsentieren die untersten Schichten die unimodalen Elemente der Architektur. Diese verschmelzen zunehmend miteinander, in Form von Zwischenrepresentationen, bis in der Ausgabeschicht die Klassenentscheidung getroffen wird. Die Zwischenschichten sind dabei nicht so gut interpretierbar wie bei einer manuell aufgebauten Architektur, haben allerdings ungeahnte Möglichkeiten [45, 114]. Tiefe neuronale Netze sind im Bereich des affektive Computing aufgrund ihrer Komplexität allerdings bisher wenig verwendet, werden allerdings in Zukunft eine immer größere Rolle spielen.

Der Umgang mit unsicheren Daten und Entscheidungen ist von entscheidender Bedeutung bei dem Ziel, Klassifikatoren für realistische Interaktionssysteme zu entwickeln. Die Grundwahrheit von Trainingsdaten ist aufgrund der Komplexität der MMI unsicher, so dass die Klassifikatoren auf ihrer Eingabeseite mit Unschärfe oder Unsicherheit umgehen müssen. Auch die erreichbaren Klassifikationsraten sind nicht vergleichbar mit anderen Anwendungen, so dass ein System einem Klassifikator-Output nicht bedingungslos vertrauen kann. Daher ist es notwendig, die Unsicherheit von der Eingabe auf die Ausgabe des Klassifikators abzubilden und dem System Informationen über die Zuverlässigkeit der gegebenen emotionalen Schätzung zu geben.

Ein Ansatz zum Umgang mit Unsicherheit wird von Dempster und Shafer, bekannt als *Dempster-Shafer-Theorie*, beschrieben [239]. Die Theorie erlaubt die Kombination von Unsicherheit bzw. Evidenz aus mehreren Quellen und erzeugt einen Vertrauenswert (*Believe*), beschrieben durch die Believefunktion. Die Dempster-Shafer-Theorie verallgemeinert die Bayessche Theorie der subjektiven Wahrscheinlichkeit. Eine Evidenz kann als eine Erweiterung einer Wahrscheinlichkeit betrachtet werden, wobei ein zweidimensionales Maß verwendet

wird. Dies setzt sich zusammen aus dem Grad des Dafürhaltens bzw. dem Grad des Vertrauens darin, dass die Aussage einer Quelle zutrifft (engl. degree of belief), und der Plausibilität des Ereignisses, d.h. aus einem Wahrscheinlichkeitsbereich mit einer unteren Grenze Believe und einer oberen Grenze Plausibilität. Die Evidenztheorie ist nicht in Situationen einsetzbar, in denen der Grad des Vertrauens in eine Quelle nicht beziffert werden kann. In den Anwendungen der Mustererkennung, bei denen sich mittels der Evidenztheorie Aussagen verschiedener unzuverlässiger Algorithmen kombinieren lassen, ist die Theorie einsetzbar. Dennoch spielt sie in der gegenwärtigen Forschungslandschaft eine eher untergeordnete Rolle. In den meisten Fällen werden traditionelle Bayssche Ansätze eingesetzt um einen Grad der Erkennungssicherheit zu beschreiben.

Unimodale Ansätze begannen ursprünglich mit den bereits früh verfügbaren audio-visuellen Datensätzen. Sprachsignale sind für die Erkennung von Emotionen interessant, da sie einfach zu verarbeiten sind und ihre Analyse verspricht vielversprechende Ansätze für die Forschung [80, 203, 206, 209]. In den letzten Jahren haben sich verschiedene akustische Merkmale als nützlich für die Emotionserkennung erwiesen. Nicholson et. al. verwenden Merkmale basierend auf Tonhöhe und LPC in Kombination mit künstlichen neuronalen Netzen [165]. In frühen Arbeiten wurden Mehrklassifikatorsysteme mit drei Merkmalstypen (LPC, Modulations Spektrum (ModSpec), RastaPLP) trainiert, um sieben verschiedene expressive Emotionen mit einer Genauigkeit von mehr als 70% zu erkennen, durchgeführt auf der EmoDB [36, 208, 170, 210, 230]. Eines der ersten eingesetzten Merkmale sind die MFCC welche für die Spracherkennung entwickelt wurden [138, 154]. In der jüngeren Vergangenheit haben Sprachqualitätsmerkmale (engl. Voicequality) aufgrund ihrer Fähigkeit, verschiedene Sprachstile zu repräsentieren, erhöhte Aufmerksamkeit erhalten und wurden für die Emotionserkennung erprobt [148, 147, 205], konnten dort jedoch auch keine überlegenen Resultate erzielen.

Anstatt sich nur auf eines oder wenige der erwähnten Merkmale zu verlassen, werden in Emotionserkennungswerkzeugen wie openEAR [76] mehrere tausend Merkmale berechnet um so hohe Klassifizierungsraten zu erzielen. Eine große Anzahl von Merkmalen bringt jedoch den Nachteil mit sich, dass die darin enthalten zielführenden Informationen wenig dicht gepackt sind und Klassifikatoren darauf häufig mit Überanpassung reagieren [163, 22]. Daher werden Merkmal-Selektions-Algorithmen verwendet, um die informativen Teile der großen Merkmalsmenge zu kondensieren und die Klassifikation so signifikant zu verbessern [96]. Da eine vollständige Untersuchung des Merkmals-Unterraums praktisch unmöglich ist [22], werden verschiedene Heuristiken verwendet, um die Merkmalsmengen zu beschneiden oder zu erweitern. Ein beliebter heuristischer Algorithmus ist die vorwärts-rückwärts Merkmalsselektion [136].

Die Erkennung von Emotionen anhand von Bildmaterial begann zunächst durch die Analy-

se statischer Bilder, dabei werden texturbasierte oder geometriebasierte Merkmale verwendet [283, 281]. Diese Arbeiten verwendeten ursprünglich nur einzelne Bilder aus Videosequenzen. Erste videobasierte Arbeiten verwendeten monochrome Datensätze wie der Cohn-Kanade [121]. Diese Arbeiten gehen dabei auf das letzte Jahrzehnt des vorigen Jahrhunderts zurück. Mittlerweile ist die videobasierte Erkennung weit vorangeschritten. Typische eingesetzte Merkmale berücksichtigen nun auch die zeitliche Komponente der Bildfolge, auch wenn in neueren Arbeiten immer noch Verbesserungen der Merkmale erzielt werden ist der Merkmalsextraktionsprozess mittlerweile derart stabil, dass es mehrere frei verfügbare Werkzeuge wie CERT [144] oder openSMILE [78] gibt, die den Merkmalsgewinnungsprozess automatisch übernehmen. Diese extrahieren dabei alle gängigen Merkmale und werden stetig durch neue Merkmale erweitert.

Kipp und Martin [134] untersuchen vier grundlegende Handgestenmerkmale und ihre jeweilige Beziehung zu Emotionen. Ihr Hauptziel war es, Ganzkörperavatare zu erzeugen, die mittels Handgesten Emotionen unterscheiden und ausdrücken können. Sie haben ein Lexikon eingeführt, um Handgesten anhand von *Händigkeit*, *Handgeometrie*, *Handflächenorientierung* und *Bewegungsrichtung* zu beschreiben. Die Daten wurden aus dem Film *Death of a Salesman* (1966 DS-1 und 1985 DS-2) gewonnen. Um die Korrelation zwischen Emotionskategorie und Gesten zu bestimmen, berechneten die Autoren paarweise χ^2-Werte. Die Ergebnisse deuteten auf Korrelationen zwischen Emotionskategorie und Händigkeit hin (χ^2 = 40.14; p $<$ .001 im Film DS-1, χ^2 = 35.37; p $<$ 0,001 in Film DS-2). In Film DS-2 fand sich auch eine Korrelation zwischen Emotionskategorie und Handflächenorientierung (χ^2 = 42.50; p $<$ .05).

Kapur et al. [122] haben im Jahr 2005 eine Studie zu gestenbasiertem affektive Computing durchgeführt. Ihr Ziel war es, einen Klassifikator zu entwerfen, um zwischen den vier Emotionen *Trauer, Freude, Ärger* und *Angst* zu unterscheiden. Dazu haben die Autoren fünf Schauspieler mit Markierungen eines Bewegungserfassungssystems ausgerüstet und forderten sie auf, diese Emotionen zu schauspielern. Anschließend wurden 10 Teilnehmer gebeten, jede Emotion nur durch beobachten der sich bewegenden Punkte (Positionen der Marker) zu identifizieren. Im nächsten Schritt berechneten die Autoren Mittelwerte aus Geschwindigkeit und Beschleunigung der Positionen. Diese Daten wurden verwendet, um mehrere Klassifikatoren zu trainieren. Die Teilnehmer erreichten eine durchschnittliche Erkennungsrate von 93%, während die Klassifikatoren zwischen 66.2% und 91.8% erreichten.

Physiologische Signale wie EKG, EMG und SCL werden in der Affekterkennung erst seit etwa einem Jahrzehnt eingesetzt und sind so relativ jung im Vergleich zur audiovisuellen Affekterkennung. Die Vorteile physiologischer Signale besteht darin, dass die Regulierung der Werte wie Herzfrequenz oder Aktivität der Schweißdrüsen nicht willentlich beeinflusst werden kann, im Gegensatz zum Sprache, Mimik und Gestik. Dies führt zu einem fast unverzerrten Bild über

den emotionalen Zustand einer Person [280, 171]. Für biophysiologische Signale existieren Studien über die Auswirkungen kognitiver Belastung und die Erkennung dieser [86, 97]. Die Kombination von Sprache und biophysiologischen Signalen wurde von Kim et. al. untersucht und eine Verbesserung der Erkennungsrate im Vergleich zur unimodalen Klassifikation erreicht [130]. Die meisten Studien basieren auf geschauspielerten oder starken expressivem emotionalen Verhalten unter Einsatz des Emotionsmodells von Eckman [71]. Diese Methoden können nicht ohne weiteres auf typische schwache Expressionen der MMI übertragen werden.

4.2. Informationsverdichtung

Nicht alle Elemente eines Datensatzes tragen gleich viel zur Beschreibung der jeweiligen Klassen bei, vielmehr sind meist nur wenige Instanzen oder Merkmale hoch relevant. Anhand der relevantesten Elemente eines Datensatzes könne Strukturen in den Daten meist bereits sehr gut abgebildet werden. Neben den weniger relevanten Merkmalen und Instanzen (z. B. Instanzen mit hohem Abstand zur Hyperebene einer SVM) kommen zusätzlich noch diejenigen hinzu, die überwiegend Rauschen beinhalten und somit keine Information tragen. Neben dem Auffinden der wichtigsten Informationen haben informationsverdichtende Methoden also noch zum Ziel die verrauschten Daten zu eliminieren. Hier vorgestellt wird dies exemplarisch anhand von Instanz Selektion und Merkmals Selektion.

4.2.1. Selektion von Instanzen

Das in diesem Abschnitt vorgestellte Verfahren zur rückweisungsbasierten Instanz Selektion und seine Resultate wurden in [154] publiziert.

Bei der Instanz Selektion wird eine kleine Teilmenge S des ursprünglichen Trainingssatzes T gesucht, so dass ein Klassifikator, trainiert auf S, eine ähnliche oder sogar bessere Klassifikationsleistung als ein auf den vollständigen Datensatz T trainierter Klassifikator zeigt. Das Verfahren verwendet eine auf der Reklassifikationskonfidenz basierte Auswahlstrategie. Als Datensatz werden die akustischen Daten der Sequenzen 4 und 6 des EmoRec Datensatzes, bezeichnet als *positv* und *negativ*, herangezogen. Als Instanzen werden daraus extrahierte MFCC Merkmalsvektoren betrachtet. Aufgrund der Tatsache, dass die Emotionserkennung eine sehr individuelle Problematik darstellt, werden die Selektionsexperimente individualspezifisch durchführt. Die Architektur verwendet eine binäre Radiale Basis Funktion (RBF)-SVM. Eine wahrscheinlichkeitsbehaftete Klassifikationsausgabe wird durch Verwendung des von Platt et. al. vorgeschlagenen Verfahrens erreicht [178].

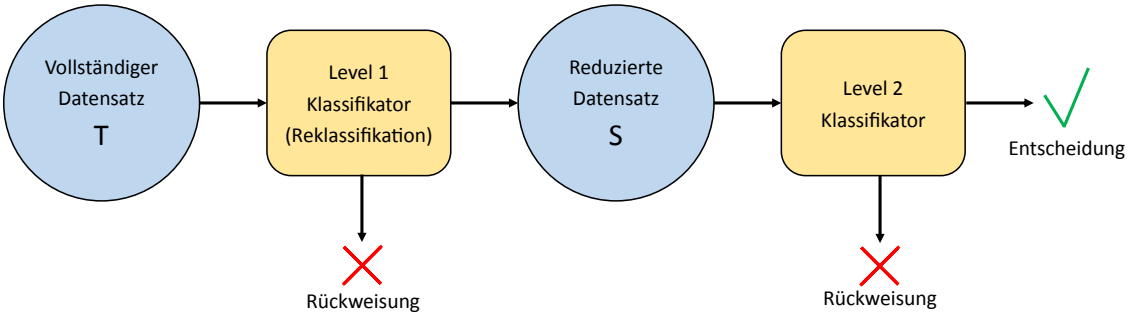

Abbildung 4.2.: Architektur der Instanzselektion durch Reklassifikation. Der vollständige Datensatz wird zunächst reklassifiziert. Die konfidentesten Instanzen bilden den reduzierten Datensatz. Auf diesem wird der zweite Klassifikator trainiert, dieser hat die Möglichkeit niedrig konfidente Klassifikationen zurückzuweisen.

Zunächst wird der gesamte Datensatz mittels Kreuzvalidierung reklassifiziert. Dazu wird der ursprüngliche Datensatz T in N Teile geteilt. Auf $N-1$ Teilen wird eine probabilistische SVM trainiert, mittels derer der verbleibende Teil reklassifiziert wird. Dies wird für alle N Teile wiederholt. Jede Instanz des Datensatzes wird so ein Konfidenzwert der Reklassifikation zugeordnet. Anschließend werden alle schwach Konfidenten, oder falsch reklassifizierten Instanzen aus dem Datensatz entfernt. Dabei wird angenommen, dass schwache Reklassifikationen ihren Ursprung darin haben, dass es sich um schlechte Prototypen oder Rauschen handelt. Zusätzlich wird im Falle unbalancierter Klassen eine Balancierung der Klassen erzwungen. Der reduzierte Datensatz S wird anschließend verwendet um den finalen Klassifikator zu trainieren. Abbildung 4.2 zeigt den Ablauf der Schritte auf.

Für die statistische Evaluation wird der Klassifikator bewertet, welcher auf den reklassifizierten und reduzierten Daten trainiert wurde. Dazu wird der Baselineklassifikator, welcher auf den gesamten Daten (ohne Reduzierung) trainiert wurde, mit Reduzierungen von 80%, 90% und 98% verglichen. Die Bewertung wurde mittels 5-fach Kreuzvalidierung durchgeführt. Als Klassifikator wird eine RBF-SVM mit den Paramtern $C = 2$ und $\gamma = 4 \cdot 10^{-3}$ eingesetzt.

Der Baselineklassifikator erreicht dabei eine Erkennungsrate von 65,9%. Die reduzierten Klassifikatoren erreichen 68,1% (bei 80% Reduktion), 67,6% (bei 90% Reduktion) und 63,3% (bei 98% Reduktion). Es ist also zu beobachten, dass eine Reduktion der Trainingsdaten eine geringfügige Verbesserung erreichen kann. Insbesondere ist jedoch festzustellen, dass die vorgeschlagene Reduktion auf die wesentlichen Instanzen zu keinem Qualitätsverlust führt, den Trainingsdatensatz allerdings erheblich verkleinert. Die Ergebnisse dieses Vergleichs sind in Abbildung 4.3 in der jeweils ersten (linken) Spalte der 50% Konfidenz Level dargestellt.

Wird dem finalen Klassifikator die Möglichkeit eingeräumt im Falle schwacher Konfidenz keine Resultate auszugeben, kann die Klassifikationsleistung deutlich gesteigert werden. Die

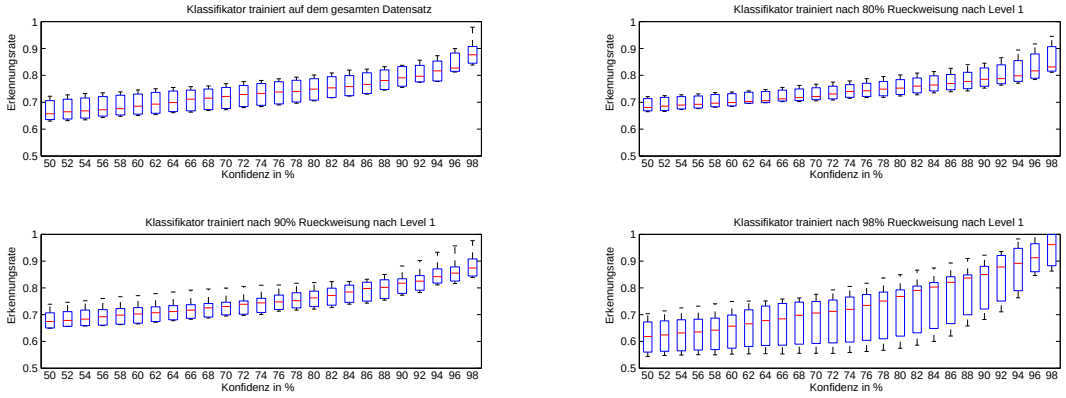

Abbildung 4.3.: Erkennungsrate abhängig vom Grad der Rückweisung in Level 2. Level 2 Klassifikator trainiert auf 4 verschieden Reduzierungsstärken in Level 1.

Boxplots in Abbildung 4.3 zeigen den Verlauf der Accuracy mit steigender Konfidenzschranke in der Ausgabeschicht. Neben dem auf allen Daten trainierten Klassifikator lassen sich auch die reduziert trainierten Klassifikatoren auf Erkennungsraten über 80% steigern, sofern ihnen erlaubt wird einige unsichere Entscheidungen nicht zu treffen. Diese fehlenden Entscheidungen können in den meisten Fusionsarchitekturen kompensiert werden. Diese fehlenden Entscheidungen führen dabei zu wesentlich weniger Problemen, als eventuell falsch bewertete unimodale Eingabedaten. Maximale Klassifikationsraten lassen sich erzielen, wenn der Trainigsdatensatz um 98% reduziert wurde und dem finalen Klassifkator eine hohe Rückweisungsrate erlaubt wurde. Hier konnten Resultate von über 90% Accuracy erreicht werden, allerdings ist das Verfahren recht instabil, was sich in einer wesentlich höheren Varianz äußert, verglichen mit den moderateren Reduktionen.

4.2.2. Selektion von Merkmalen

Eine Vielzahl von merkmalsextrahierenden Algorithmen liefern Instanzen mit einer großen Anzahl von Merkmalen (hochdimensionale Merkmale), einige Beispiele hierzu finden sich in Abschnitt 4.3. Diese tragen dabei zunächst meist wenig Information bezogen auf ihre Dimensionszahl. Um diese zu reduzieren werden häufig dimensionsreduzierende Verfahren wie PCA [274] und Linear Discriminant Analysis (LDA) [11] eingesetzt. Einen weiteren Ansatz bietet die Vorwärts-Rückwärts Merkmalsselektion [183], welche im folgenden um ein Langzeitabbruchkriterium erweitert dargestellt wird. Der Vorschlag des Verfahrens wurde in [158] und [285] veröffentlicht. Die Evaluation erfolgte unter Verwendung der EmoDB aus der hochdimensionale akustische Merkmalsvektoren extrahiert wurden.

	Spektrale Merkmale		Sprachqualitätsmerkmale						Prosodie
Name	MFCC	LFPC	NAQ	QO	GH	SGP	PSP	QOQ	f_0
Dimension	21	18	1	1	8	27	1	1	1

Tabelle 4.1.: Dimensionen der akustischen Merkmale

Diese umfassen spektrale, prosodische und Sprachqualität Merkmale:

- Spektral: MFCC [138], LPC [167]
- Prososdisch: Grundfrequenz f_0
- Sprachqualität: NAQ, QO, GH, SGP, PSP, QOQ [148]

Tabelle 4.1 zeigt alle extrahierten Merkmale und ihre Dimensionen, welche sich zunächst zu einem 79 dimensionalen Merkmalsvektor summieren. Aus der Abfolge dieser Vektoren und ihrer Ableitungen werden nun die sechs statistischen Maße Mittelwert, Varianz, Minimum, Bereich, Schiefe und Krümmung berechnet. Daraus resultieren Merkmalsvektoren der Dimension $79 \times 3 \times 6 = 1422$, welche als Instanzen in den Merkmalsselektionsalgorithmus eingehen.

Aufgrund der Rauschempfindlichkeit von Klassifikatoren sollten nur die nützlichsten Merkmale für die Klassifizierung verwendet werden. Das Entfernen der nichtinformativen Merkmale erhöht die Genauigkeit eines Klassifikators und verringert die Wahrscheinlichkeit von Überanpassung. Ein weiterer vorteilhafter Effekt der Verringerung der Dimensionalität ist ein geringerer Rechenaufwand beim Klassifikatortraining.

Im ersten Schritt wurde der Vorwärts-Selektions-Algorithmus (Alg. 4) angewendet, um die vielversprechendsten Merkmale zu finden. Beginnend mit einer leeren Merkmalsmenge wird in jeder Iteration versucht, die Genauigkeit zu steigern, indem das beste Merkmal zum aktuellen hinzugefügt wird. Der entscheidende Teil des Algorithmus ist das Abbruchkriterium. In vielen Arbeiten ist eine feste Dimension d des resultierenden reduzierten Merkmalsraums das gewünschte Kriterium. In anderen Arbeiten wird der Algorithmus terminiert, sobald die die Genauigkeit abnimmt, d.h. $acc(Y_{i+1}) < acc(Y_i)$. Dies führt jedoch schnell zu einer vorzeitigen Beendigung des Algorithmus bei kurzzeitigem Absinken der Resultate. In dieser Arbeit wird alternativ ein Langzeitabbruchkriterium benutzt. Dabei werden die Genauigkeiten der letzten i Iterationen berücksichtigt.

Angenommen, $v(i)$ ist der Genauigkeitsvektor der Erkennungsraten,

$$v(k) = (acc(Y_{m-k}), acc(Y_{m-k+1}), ..., acc(Y_{m-1}), acc(Y_m)) \tag{4.1}$$

dann ist das Abbruchkriterium definiert durch:

$$Terminate\, if:\ argmax_{i=1..k}(v_i(k)) = 1 \tag{4.2}$$

Dies bedeutet, dass der Algorithmus beendet wird, wenn die Genauigkeit der letzten $k-1$ Iterationen die Genauigkeit nicht verbessert hat. Sobald das Abbruchkriterium erfüllt ist, wird die Merkmalsmenge auf $m-k$ Dimensionen zurückgesetzt und die letzten $k-1$ Dimensionen die keine weitere Verbesserung erbrachten entfernt.

Die resultierende Merkmalsmenge wird dann als anfängliche Merkmalsmenge für einen Rückwärtseliminierungsalgorithmus verwendet. Hierbei werden die am wenigsten erfolgversprechenden Merkmale in jeder Iteration entfernt. Der Algorithmus wird beendet, wenn alle Merkmale eliminiert wurden. Die finale Merkmalsmenge ist diejenige, die die beste Genauigkeit während der rückwärts gerichteten Eliminierung erreicht hat.

Algorithmus 4 Vorwärts Selektion

1: $Y_0 = \{\emptyset\}$ ▷ Initialisiere Merkmalsmenge
2: $m = 0$
3: **while** $argmax_{i=1..k}(v_i(k)) \neq 1$ **do**
4: $\quad x^+ = argmax_{x \notin Y_m}[Accuracy(Y_m + x)]$ ▷ Suche nächstes optimales Merkmal
5: $\quad Y_{m+1} = Y_m \cup x^+$ ▷ Update Merkmalsmenge
6: $\quad m = m + 1$
7: **end while**
8: **return** Y_{m-k}

Die Bewertung der vorgeschlagenen Methoden erfolgte anhand der 535 Äußerungen der EmoDB in Form einer *leave one subject out* (LOSO) Kreuzvalidierung.

Aufgrund der großen Anzahl von Dimensionen des ursprünglichen Merkmalvektors und des geringen informativen Charakters der meisten Merkmale, ist ein Klassifikator trainiert auf all diesen Merkmalen nur unwesentlich performanter als eine a priori Schätzung und ein Vergleich mit dieser als Baseline wenig hilfreich. Die Genauigkeit wird drastisch verbessert, wenn nur die informativsten Merkmale ausgewählt werden. Abbildung 4.4 zeigt den Verlauf der Erkennungsgenauigkeit in Abhängigkeit der ausgewählten Merkmalsmenge. Zunächst werden die vielversprechendsten Merkmale hinzugefügt, was zu einer schnellen Steigerung der Erkennungsrate führt. Der linke Teil der Vorwärts Selektion zeigt zu dem die Wichtigkeit den Auswahlalgorithmus nicht vorzeitig zu beenden, da vorübergehende kleiner Einbrüche der Erkennungsrate auftreten. Das langfristige Abbruchkriterium verhindert einen solchen vorzeitigen Abbruch. Das Abbruchkriterium mit $k = 100$ ist erfüllt, nachdem 334 der anfänglichen 1422-Dimension ausgewählt wurden, welche eine Erkennungsrate von 82,21%

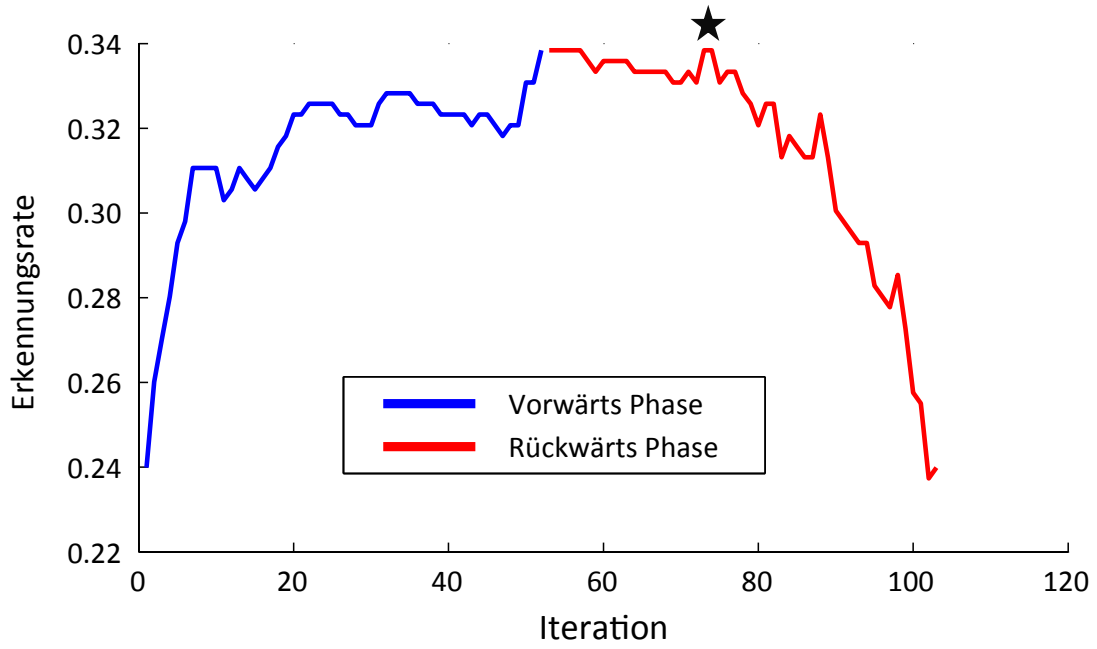

Abbildung 4.4.: Resultat der Vorwärts-Rückwärts-Merkmalsselektion mit Langzeitabbruchkriterium. Die mit Stern gekennzeichnete Stelle entspricht dem Resultat des Algorithmus. Diese erreicht dieselbe Erkennungsrate wie zu Beginn der Rückwärtsphase, benötigt jedoch weniger Merkmale.

erreichen. Anschließend wird ausgehend von dieser Merkmalsteilmenge die Merkmalsmenge durch Rückwärtssuche reduziert. Dabei konnte im speziellen Fall keine Verbesserung mehr erreicht werden. Die Merkmalsmenge konnte bei gleichbleibender Erkennungsrate (siehe Abbildung 4.4 rechter Teil) jedoch nochmals um 226 Merkmale auf 108 reduziert werden, was nur 8% aller ursprünglichen Merkmale entspricht.

Tabelle 4.2 zeigt die Verwechslungsmatrix des mit den besten 106 Merkmalen trainierten Klassifikators. Die Emotionen *Ärger*, *Langeweile* und *Trauer* wurden mit Erkennungsraten über 90% erkannt. Mit 67,9% ist *Glück* die schwierigste erkennbare Emotion, welche oft mit *Wut* verwechselt wird. Dieses Problem ist bereits von anderen Studien bekannt [169], welche zu dem Schluss kommen, dass *Glück* und *Wut* Ähnlichkeiten in der Erregungsdimension des VAD Modells aufweisen.

	Ärger	Langeweile	Ekel	Angst	Freude	Trauer	Neutral
Ärger	**0.961**	0	0	0.0177	0.021	0	0
Langeweile	0	**0.946**	0	0.011	0	0	0.043
Ekel	0.014	0.147	**0.783**	0	0	0	0.056
Angst	0.070	0.017	0.013	**0.802**	0.033	0.024	0.041
Freude	0.221	0	0.009	0.091	**0.679**	0	0
Trauer	0	0	0	0.014	0	**0.986**	0
Neutral	0	0.146	0	0.009	0.035	0	**0.81**

Tabelle 4.2.: Verwechslungsmatrix des Klassifikators trainiert auf den durch vorwärts rückwärts Merkmalsselektion gewählten Merkmalen.

4.3. Unimodale Emotionserkennung

4.3.1. Erkennung anhand von Prosodie

Für die Erkennung von Emotionen steht eine breite Palette von Merkmalen zur Verfügung. Diese werden im Folgenden vorgestellt, und dabei wie in [155] beschrieben, um das Merkmal der Enhanced Auto Correlation (EAC) erweitert. Die meisten Merkmale basieren dabei auf verschiedenen Formen der Frequenzanalysen des akustischen Signals. Dabei wird das Eingangssignal in kleine Zeitabschnitte, sogenannte Fenster, zerlegt welche dann über das Signale geschoben und analysiert werden. Zur Vermeidung von Randeffekten bei der Analyse wird zunächst typischerweise eine Hamming-Funktion $w(i) = 0.54 + 0.46 \cdot cos(\frac{2\pi i}{M})$, $i = -\frac{M}{2}, ..., \frac{M}{2} - 1$ mit der Fenstergröße M und Abtastwertindex i auf das Fenster angewendet. Abbildung 4.5 zeigt eine Übersichtsdarstellung der verwendeten Merkmale.

Akustische Merkmale

Mel Frequency Cepstral Coefficent (MFCC) [265]: Die MFCC-Darstellung eines Signals wird durch die Art und Weise motiviert, wie Menschen eine Audio-Modulation im Ohr wahrnehmen. Ein Filterbank mit Filtern linearen Abstandes in den unteren Frequenzen, und logarithmisch skalierten Filtern in den oberen Frequenzen, genannt die MEL-Filterbank, wird zur Merkmalsextraktion verwendet. MFCC werden häufig zur Analyse sehr kurzer Fenster verwendet, da sie die Fähigkeit besitzen Informationen in kompakter Weise darzustellen. Die wichtigen Information für sprachverarbeitende Anwendungen wie Phoneminformationen bleiben in der kompakten Darstellung erhalten [145].

Auf jedem Fenster wird eine schnelle Fourier-Transformation (engl. Fast Furier Transformation)

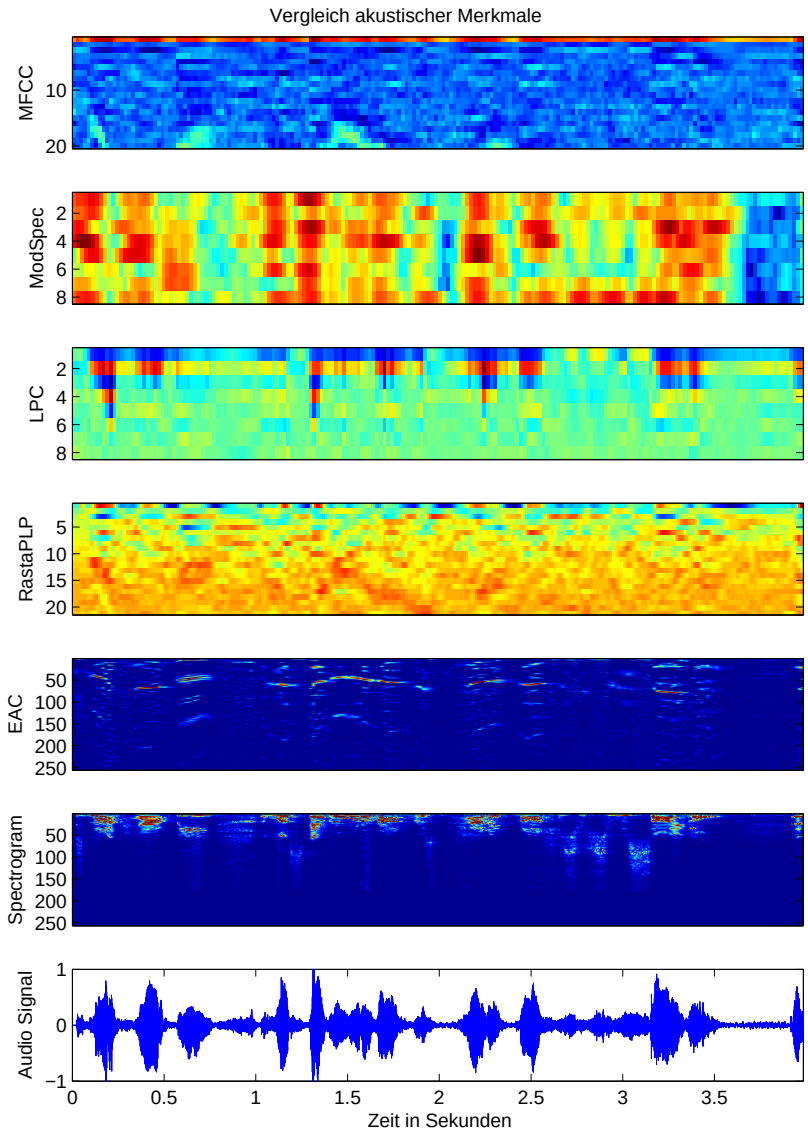

Abbildung 4.5.: Vergleichsdarstellung der akustischen Merkmale anhand einer kurzen Audiospur. Zu erkennen sind die unterschiedlichen Dimensionszahlen und Ausprägungen der Merkmale.

(FFT) berechnet, auf deren Resultat die MEL-Filterbank von Dreiecksfiltern angewendet wird. Die Diskrete Cosinus Transformation (DCT) angewendet auf jedes Filterresultat erzeugt die dekorrelierte Cepstral-Koeffizienten jedes Analysefensters. Eine typische Filteranzahl (Merkmalsdimension) liegt im Bereich von 8 bis 24 Filtern. Typische Analysefenster haben eine Breite von ca. 40 Millisekunden und werden meist um die halbe Fensterbreite vorgeschoben.

Modulations Spektrum (ModSpec) [103]: Um den Einfluss von Umgebungsgeräuschen zu reduzieren, hat Hermansky die Verwendung des Modulationsspektrums vorgeschlagen, welches die zeitliche Dynamik des Sprachsegments erhält. Um eine Modulationen im Bereich von 2 Hz

bis 16 Hz zuverlässig modellieren zu können, sind Fenstergrößen von 200 ms bis zu mehreren Sekunden notwendig. Die Wahl der Fensterlänge hängt dabei häufig vom Ausgangsmaterial ab. Beinhaltet dieses nur einzelne Kommandowörter begrenzen diese das maximale Analysefenster.

Wie bei MFCC wird zuerst eine FFT auf überlappenden Unterfenstern berechnet und die MEL-Filterbank angewendet. Anstelle der DCT Anwendung auf jedem Filterresultat werden diese stattdessen zu einer Folge von Werten aggregiert. Auf diesem aggregierten Signal wird nun erneut eine FFT durchgeführt, welche zum endgültigen Merkmalswert führt. Starke Änderungsraten des Vokaltrakts können so beschrieben werden, diese enthalten dabei wichtige sprachliche Informationen.

Linear Predictive Coding (LPC) [117]: Anstatt das Signal mithilfe von FFT-basierten Ansätzen zu analysieren, wird hier das Sprachmuster $s(t)$ durch Linearkombination der letzten p Abtastwerte approximiert.

$$s(t) \approx a_1 s(t-1) + ... + a_p s(t-p), \tag{4.3}$$

Dabei werden die Koeffizienten $a_1, ..., a_p$ in einem kurzen Signalsegment als konstant angenommen. Daraus ergibt sich ein p dimensionaler Vektor, der einer Kurve entspricht, welche die Spitzen des kurzzeitigen logarithmischen Signalspektrums approximiert. Wie in den vorherigen Merkmalen werden die Informationen dabei komprimiert, jedoch ohne Transformation von Zeit zu Frequenz Darstellung. Typische Werte für p liegen in der Größenordnung von 12, extrahiert auf kurzen Fenstern mit ca. 40 Millisekunden Signalbreite.

Relative Spectral Perceptual Linaear Prediction (RastaPLP) [104]: Die biologische Wahrnehmung von kritischen Frequenzbändern und deren gleicher Lautheitswahrnehmung bilden die Grundlage dieser Merkmale. Der Schalldruck (in dB), welcher erforderlich ist, um einen Ton einer beliebigen Frequenz so laut wahrzunehmen, wie ein Referenzton von 1 kHz wird dabei angenähert durch:

$$E(w) = 1.151 \cdot \sqrt{\frac{(w^2 + 144 \cdot 10^4) \cdot w^2}{(w^2 + 16 \cdot 10^4) \cdot (w^2 + 961 \cdot 10^4)}} \tag{4.4}$$

Als Folge dieser Gleichung benötigen Frequenzen unter 1 kHz höhere Schalldruckpegel als der Referenzton, während Frequenzen zwischen 2 und 5 kHz geringere Schalldruckpegel benötigen um gleich laut wahrgenommen zu werden. Wie bei den MFCC wird eine Filterung durchgeführt, im Gegensatz zu MFCC wird allerdings die lineare *Bark-Skala* verwendet, deren Filter zudem nicht dreieckig sind.

Da diese Resultate zunächst sehr empfindlich gegenüber spektralen Änderungen sind, welche zum Beispiel durch Übertragungskanäle oder durch verschiedene Mikrofone und Telefone oder durch Signalkompressionsalgorithmen verursacht werden, schlägt Hermansky nach der Bandfilterung eine Transformation des Spektrums in die logarithmische Domäne vor. Auf dieser wird dann eine relative spektrale Filterverarbeitung durch exponentielle Retransformation angewendet. Schließlich werden daraus LPC-Koeffizienten berechnet. Typische Filderdimensionen liegen bei 21 Filtern, angewendet auf Fensterbreiten zwischen 20 und 40 Millisekunden.

Enhanced Auto Correlation (EAC) [256, 155]: Die EAC-Funktion wurde ursprünglich von Tolonen und Karjalainen als Multipitch Analyse eingeführt und findet in verschiedenen Bereichen der Audiosignalverarbeitung Verwendung. In dieser Arbeit wird eine modifizierte Version des Extraktionsprozesses verwendet, um die Hauptharmonien des Stimmbandes, also der Stimmquelle des Sprachsignals, zu extrahieren. Unterschiedliche emotionale Zustände können eine Vielzahl von Muskelspannungen im Vokaltrakt verursachen, die den erzeugten Klang beeinflussen. Das Extrahieren der Mehrfachtonhöhen basierend auf der Periodizität des Signals ist daher eine zuverlässige Quelle mit vielen emotionalen Informationen, unabhängig von den gesprochenen Inhalten. Zusätzlich liefert die EAC Analyse ein sehr hochdimensionales Merkmal der Audioverarbeitung

Wie bei den meisten anderen Merkmalsextraktionsverfahren wird auch hier das Signal zuerst in überlappende Hamming-Fenster zerlegt. Für jedes Fenster mit dem Eingangssignal S wird die Autokorrelationsfunktion

$$acf(S) = \text{Re}\left(\text{iFFT}\left(\left|\text{FFT}\left(\text{Hamming}(S)\right)\right|^{k}\right)\right) \tag{4.5}$$

berechnet um so periodische Anteile des Signals zu extrahieren. Der Parameter k ermöglicht die Kontrolle der Periodizitätserkennung mit einer nichtlinearen Verarbeitung in der Frequenzdomäne. Spitzen in der Autokorrelationskurve sind ein Indikator für die Tonhöhe der Grundfrequenzen. Häufig sind viele redundante Informationen und Rauschen Teil dieser Kurve. Um die Zuverlässigkeit der Erkennung mittels Autokorrelation zu erhöhen, wird deren Ergebnis bei Null abgeschnitten (Alle negativen Werte werden auf Null gesetzt). Um multiple Spitzen der Grundfrequenz, die durch die Autokorrelationsfunktion verursacht werden, zu entfernen, wird ein zeitverdoppeltes Signal der Autokorrelation vom Autokorrelationssignal subtrahiert. Anschließend werden erneut alle negativen Werte auf Null gesetzt, dies entspricht dann dem endgültigen Merkmalsvektor.

Entgegen andere Merkmale hängt die Dimensionalität der EAC Signale von der Größe des verwendeten Analysefensters ab. Diese sind in ihrer Wahl dadurch eingeschränkt. In dieser Arbeit

Merkmal	Instanzweises Voting Späte Fusion	Fenster Durchschnitt Frühe Fusion	Fenster Varianz Frühe Fusion
MFCC	24.7 (2.44)	25.6 (1.56)	22.8 (2.06)
ModSpec	28.4 (1.71)	28.0 (3.86)	25.1 (1.95)
LPC	21.9 (0.80)	20.6 (2.18)	20.6 (3.23)
RastaPLP	22.8 (1.61)	19.9 (1.36)	29.1 (1.60)
EAC	**29.2** (1.13)	**30.7** (3.82)	**30.7** (2.80)

Tabelle 4.3.: Erkennungsraten der kategoriellen Emotionserkennung der Emotions in the Wild (EmotiW) Challenge für verschiedene akustische Merkmale in Prozent. Evaluiert auf dem *ValidationSet* der Challenge. In Klammer ist die Standardabweichung der 10-fach Kreuzvalidierung angegeben.

werden Fenstergrößen von 1024 Abtastwerten (entspricht 40 Millisekunden) verwendet, daraus resultieren Merkmalsvektoren der Dimension 512, der Parameter k wurde nach empirischen Erfahrungen auf $1/3$ festgelegt.

Langzeitmerkmale: Alle zuvor beschriebenen Merkmale sind von kurzen Zeitfenstern eines Signals abgeleitet, was zu einer großen Anzahl (200-300) von Merkmalsvektoren pro Satz oder Äußerung führt. Eine solche Anzahl von Instanzen ist, abhängig von den nachfolgenden Fusionsarchitekturen, nicht immer sinnvoll. Um die Anzahl der Instanzen zu reduzieren, können zwei verschiedene frühe Fussionstechniken eingesetzt werden um die Anzahl der Instanzen auf eine Instanz pro Äußerung zu reduzieren.

Eine Möglichkeit besteht darin den Durchschnittsvektor $x_\mu = \frac{1}{T}\sum_{t=1}^{T} x_t$ aller $t = 1, \ldots, T$ Merkmalsvektoren pro Sequenz zu bilden. Dabei geht allerdings die Vielfalt innerhalb einer Äußerung verloren. Um diesen Nachteil zu verringern kann als alternative Methode auch die Varianz $\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \mu)^2$ einer Vektorsequenz eingesetzt werden. Natürlich ist auch eine Konkatenation beider Varianten als Eingangsvektor der Klassifikation möglich.

Vergleich der Merkmalsperformanz

Um die Leistung der Merkmale zu vergleichen wurden die akustischen Daten der EmotiW Challenge verwendet. Diese bestehen aus 962 Filmsequenzen aus denen alle beschriebenen Merkmale extrahiert wurden. Dabei wurden die fensterweisen Merkmale sowie die beschrieben Durchschnitts- und Varianzvektoren erzeugt. Eine SVM mit RBF Kernel wird als Klassifikator eingesetzt. Jede SVM wurde mittels Gittersuche auf ihre spezifischen Parameter C und γ hin optimiert, da die verschiedenen Merkmale stark unterschiedliche Parameter erfordern. Die Evaluation erfolgte mittels 10-fach Kreuzvalidierung. Tabelle 4.3 gibt einen Überblick über die

Merkmal	Instanzweises Voting Späte Fusion	Fenster Durchschnitt Frühe Fusion	Fenster Varianz Frühe Fusion
MFCC	35.6	28.7	28.7
ModSpec	36.1	32.9	26.5
LPC	30.5	28.7	28.7
RastaPLP	35.6	20.4	28.7
EAC	**36.6**	**37.6**	**35.6**
Fusion	**40.1**		
Baseline	26.2		

Tabelle 4.4.: Erkennungsraten der kategoriellen Emotionserkennung der EmotiW Challenge für verschiedene akustische Merkmale in Prozent. Evaluiert auf dem geheimen *TestSet* der Challenge. In Klammer ist die Standardabweichung der 10-Fach Kreuzvalidierung angegeben.

	Ärger	Ekel	Angst	Freude	Neutral	Trauer	Überra.	Erk.rate
Ärger	34	0	1	11	10	2	0	58.06%
Ekel	4	0	0	4	17	1	0	0.0%
Angst	14	0	13	4	11	4	0	28.3%
Freude	9	1	1	19	43	8	0	23.5%
Neutral	3	0	4	23	73	12	2	62.4%
Trauer	10	0	2	7	20	14	0	26.4%
Überraschung	8	0	1	7	8	2	0	0.00%

Tabelle 4.5.: Verwechslungsmatrix des EAC Merkmals auf dem geheimen *Testset* der EmotiW Challenge. Erkennungsrate in Prozent, Verwechslung in absoluten Werten.

erreichten Ergebnisse der einzelnen Merkmalsvarianten. Im Falle von einzelfensterbasierten sowie Durchschnitts- und Varianzmerkmalen übertrifft das EAC-Merkmal alle anderen Merkmale, jedoch nicht signifikant. Das in der Literatur am häufigsten verwendete MFCC-Merkmal führt nur zu moderaten Resultaten auf dem EmotiW Datensatz. Dies deckt sich auch mit Erkenntnissen der Challengeteilnahme aus den Vorjahren [158]. Auch fensterweise Analyse sowie Durchschnitts- und Varianzmerkmale bilden keinen nennenswerten Unterschied.

Ergebnisse der EmotiW Challenge Teilnahme

Basierend auf den vorherigen Ergebnissen wurde beschlossen, die ModSpec und EAC Merkmale in allen drei Varianten (Voll, Durchschnitt und Varianz) für die Teilnahme an der EmotiW Challenge zu verwenden. Die sechs SVMs wurden mittels der Trainings und Validierungsdaten der Challenge trainiert und optimiert. Tabelle 4.4 gibt einen Überblick über die Ergebnisse, welche auf dem bis heute geheimen Testset der Challenge erreicht wurden. Zusätzlich wur-

de eine Fussionsarchitektur eingereicht, welche die probabilistischen Ausgaben des sechs Einzelklassifikatoren umfassenden Ensembles mittels einer weiteren SVM zusammengefasst. Die Fussionsarchitektur erreichte eine Performanz von 40.1% auf 7 Klassen und ist damit signifikant besser als jedes einzelne Merkmal. Der Vergleich der Merkmale untereinander zeigt dasselbe Verhalten wie die vorherige Merkmalsanalyse.

In Tabelle 4.5 ist die Verwechslungsmatrix des besten Challengeresultates (Akustische Fusion) dargestellt. Diese zeigt ein starkes Ungleichgewicht der Klassifikation entlang der verschiedenen Klassen. *Neutral* und *Ärger* konnten sehr gut erkannt werden, *Ekel* und *Überraschung* konnten hingegen überhaupt nicht erkannt werden. Diese Eigenschaft wurde bereits bei der Teilnahme im Vorjahr beobachtet [158]. Dieses Ungleichgewicht ist nicht überraschend. Die schlecht erkennbaren Klassen weisen im Trainingsdatensatz einen signifikant geringeren Umfang als alle anderen Klassen auf. Dies ist zwar zunächst etwas unbefriedigend, die für die MMI relevanteren Klassen *Neutral* und *Ärger* lieferten allerdings gute Ergebnisse.

Akustische Analyse des uulmMAC

Zur Erkennung kognitiver Last aus gesprochener Sprache am Beispiel des uulmMAC wurde der *Random Forest* Ansatz zur Klassifikation eingesetzt. Da nur das Sprachmaterial (ohne Stilleanteil) einer Sequenz zur akustischen Klassifikation herangezogen werden kann, reduziert sich die Durchschnittliche Länge der Trainingsdaten je Proband auf 22 bis 362 Sekunden für Überforderung und 18 bis 324 Sekunden für die Unterforderung. Die großen Unterschiede kommen aufgrund der stark unterschiedlichen Kommunikationsfreude der Probanden zustande. Pro Sequenz und Proband treten durchschnittliche 44 sprachliche Äußerungen (Sätze oder Kommandoworte) auf. Daraus wurden die bereits zuvor verwendeten Merkmale extrahiert. Die Resultate der akustischen Erkennung wurden in [101] veröffentlicht.

Bei der Evaluation wurde zwischen einer individuellen und einer interindividuellen LOSO Analyse unterschieden. Beide Varianten wurden sowohl auf Einzelfensterebene als auch auf Basis ganzer Äußerungen untersucht. Tabelle 4.6 zeigt die erreichten Resultate des Zwei-Klassen-Problems. Zu erkennen ist, dass, wie zu erwarten, die Klassifikation ganzer Äußerungen besser funktioniert, als die Klassifikation einzelner Analysefenster. Dies ist wenig verwunderlich, da bei der Bewertung ganzer Äußerungen eine implizite temporale Fusion einzelner Entscheidungen enthalten ist. Weiter ist das für die Emotionserkennung typische Verhalten des individuellen Problems erkennbar, so erzielen auf ein Individuum optimierte Klassifikatoren wesentlich bessere Resultate, hier um ca 10% besser.

Merkmal	Individuell		Interindividuell (LOSO)	
	Fenster	Äußerung	Fenster	Äußerung
LPC	66.52 %	77.75 %	61.22 %	68.93 %
MFCC	**76.27 %**	**83.80 %**	**68.54 %**	**75.53 %**
Rasta-PLP	74.49 %	82.75 %	68.70 %	74.40 %
ModSpec	70.20 %	78.82 %	63.90 %	71.29 %
EAC	73,91 %	79,23 %	63,72 %	72,36 %
Audio-Fusion	**81.67 %**	**86.58 %**	**72.72 %**	**77.50 %**

Tabelle 4.6.: Evaluation aller akustischen Merkmale für die individuelle und interindividuelle (LOSO) Erkennung von Überforderung im uulmMAC. Durchgeführt jeweils auf Einzelinstanzen (Fenster) und einer ganzen Äußerung (Voting Fusion).

4.3.2. Erkennung anhand von Gesichtsexpressionen

Die Merkmalsextraktionsmethoden können in zwei Gruppen aufgeteilt werden: Modellgetriebene Methoden und datengetriebene Methoden. Modellgetriebene Methoden nutzen Modellwissen der Anatomie des menschlichen Gesichtes und deren Bedeutung bei der Expression von Emotionen. Datengetriebene Methoden betrachten hingegen die reinen Bildeigenschaften eines Gesichtsausschnittes. Da diese Arbeit keinen Beitrag zur Verbesserung der gesichtsbasierten Emotionserkennung liefert, sondern Methoden aus der Literatur verwendet, werden diese Methoden im folgenden nur kurz beschrieben. Auch auf eine tiefgehende Evaluation wurde verzichtet, da diese ebenfalls in umfangreicher Literatur zu finden ist [121, 140, 79, 166, 142, 283]. Die hier beschriebenen Methoden finden Anwendung als Eingangsdaten späterer Fusionsarchitekturen.

Gesichtsmerkmale

Facial Action Coding System (FACS) [72]: Die zugrunde liegenden psychologischen Gedanken wurden bereits in Abschnitt 2.1 beschrieben. Ziel der maschinellen Verfahren ist es die beschriebenen AU, welche bisher manuell codiert wurden, automatisch zu erkennen. Ein frei verfügbares Werkzeug, welches diese automatische Codierung leisten kann, ist in der Literatur durch die *Computer Expression Recognition Toolbox* (CERT) [144] beschrieben.

Graphbasierte Ausdruckserkennung [107, 113]: Diese Ansätze verwenden eine graphbasierte Gitterdarstellung des Gesichtes, dessen Knoten sich an erkennbaren markanten Punkten im Gesicht befindet. Solche Punkte sind definiert durch Augenmitte, Augenränder, Augenbrauen, Nasenspitze, Nasenflügel, Ober- und Unterlippe, etc. Das Gitter liegt als elastischer Graph über dem Gesicht. Diese Graphen stellen ein Merkmal dar, welches direkt in graphbasierten

Klassifikationsmethoden eingesetzt werden können. Darüber hinaus können aus dem Graphen weitere Merkmale wie Verhältnis der Mundbreite zu Mundhöhe (Öffnung des Mundes), Augenbrauenwinkel (Stirnrunzeln), Abstand Nasenspitze zu Mund (Naserümpfen) und weitere Merkmale abgeleitet werden.

Orientierungshistogramme [81]: Dieses Histogramm enthält alle Kantenausrichtungen der im Bild befindlichen Kanten. Die Kantenerkennung erfolgt durch Faltung mit zwei *Sobel-Filtern* [246] in horizontaler (x) und vertikaler (y) Richtung. Das Ergebnis sind zwei Bilder G_x und G_y, die an jedem Pixel eine Approximation der Ableitung des Urbildes enthalten. Daraus werden die Richtungen aller Gradienten mit $\theta(x, y) = atan2(G_y, G_x)$ berechnet. Diese Richtungen werden dann in einem Histogramm aufgenommen, welches als Merkmal zur Klassifikation verwendet wird.

Optischer Fluss [249]: Dieser beschreibt die Bewegung der Pixel von einem Bild zum nächsten und nutzt so die zeitliche Information in Videodaten aus. Der optische Fluss beschreibt für jedes Pixel einen Vektor $u = (u_1, u_2)$ der die Richtung der Bewegung des Pixels enthält. Diese Vektoren können mit $s_u = atan2(u_1, u_2)$ in einen skalaren Wert transformiert werden, welche in einem Histogramm erfasst werden. Wie auch die Orientierungshistogramme, bilden die Flusshistogramme den Ausgangspunkt der Klassifizierung.

Local Binary Pattern - Tree Ortogonal Planes (LBP-TOP) [168, 282]: LBP Merkmale nutzen die Textur in einem Bild, bzw. Bildfolgen. Dafür wird die $3x3$-Nachbarschaft jedes Pixels betrachtet. Daraus wird die Texturinformation der Region bestimmt. Diese Werte sind in einem Histogramm erfasst und können direkt als Eingabe für einen Klassifikator verwendet werden. Eine Optimierung unter Berücksichtigung der zeitlichen Abfolge von Bildern wird durch das LBP-TOP Merkmal erreicht. Ein kurzes Videosegment wird dabei als Voxel-Quader mit drei orthogonalen Ebenen betrachtet: XY, XT und YT. Jede Richtung der Ebenen führt zu einem eigenen LBP Histogramm. Die drei Histogramme werden abschließend zu einem Vektor konkateniert, welcher neben der Texturinformation auch eine Texturänderungsinformation über die Zeit beinhaltet

Ergebnisse der gesichtsexpressionsbasierten Erkennung

Wie bereits erwähnt, wurden die Merkmale in dieser Arbeit nur in ihren Standardversion eingesetzt. Belastbare Bewertungen aller Merkmale finden sich wie bereits im Stand der Forschung beschrieben in großem Umfang in der Literatur [121, 140, 79, 166, 142, 283].

In dieser Arbeit wurden Gesichtsausdrücke mit einem multivariaten Gauß Ansatz klassifiziert. Zur stabilen Klassifizierung, wurde ein Baggingverfahren durchgeführt, dabei wurden 99% der

Testframes aufgrund der geringen Konfidenz des Klassifikators durch diesen zurückgewiesen (Frames sind in Videoanwendungen typischerweise im Übermaß vorhanden). Die verbleibenden klassifizierten Merkmale erreichen auf dem Zwei-Klassen-Problem des uulmMAC für interindividuelle Experimente lediglich eine frameweise Genauigkeit von 56,5%, im Falle individueller Klassifikatoren erreichten diese durchschnittlich immerhin 67,3%.

4.3.3. Erkennung anhand von Körperposen

Tiefensensoren wie die Microsoft Kinect v2 erlauben eine ganzheitliche Betrachtung der Körperpose des Nutzers. Die Körperpose wird dabei durch einen Graphen bestehend aus Gelenken und Knochen beschrieben. Abbildung 2.13 zeigt die Darstellung des erfassten Graphen in der Szene. Erfasst wird die Gelenkposition im Raum, und somit implizit auch die Länge der Knochen, und der Winkel der Gelenkstellung (Winkel zwischen benachbarten Knochen). In dieser Arbeit wurden Merkmale entwickelt, welche aus diesen Körperposegraphen abgeleitet sind und eine gute individualspezifische Beschreibung der kognitiven Belastung erlauben. Die Merkmale und Resultate wurden in [106] und [105] veröffentlicht.

Körperpose Merkmale

Die Literatur beschreibt keine typischen Körperposen welche kognitive Belastung in der mṃi ausdrücken. Durch explorative Betrachtungen der erfassten Videodaten können folgende gut automatisch erkennbare statische Merkmale als häufig auftretend identifiziert werden:

- **Arme verschränken**: Die rechte Hand hält den linken Ellenbogens und vice versa.
- **Hände hinter dem Rücken:** Beide Hände befinden sich (verdeckt) hinter dem Rücken (meist auf Höhe des Gesäß).
- **Hände in die Hüfte gestemmt:** Die rechte und/oder linke Hand befindet sich nahe der jeweiligen Hüfte.
- **Beine überkreuzt:** Der linke Fuß befindet sich rechts vom rechten Fuß.
- **Schrittstellung:** Ein Fuß befindet sich deutlich näher am Sensor als der andere.

Die Vorkommen sind dabei fast gleichverteilt im gesamten Datensatz und weisen somit nur eine geringe interindividuelle Unterscheidungskraft auf. In Kombination mit ihrer zeitlichen Ausprägung (Dauer) wird die Unterscheidungskraft jedoch stark verbessert. Problematisch ist dabei allerdings die Wahl einer geeigneten Schwelle, die den exakten Beginn und Ende einer

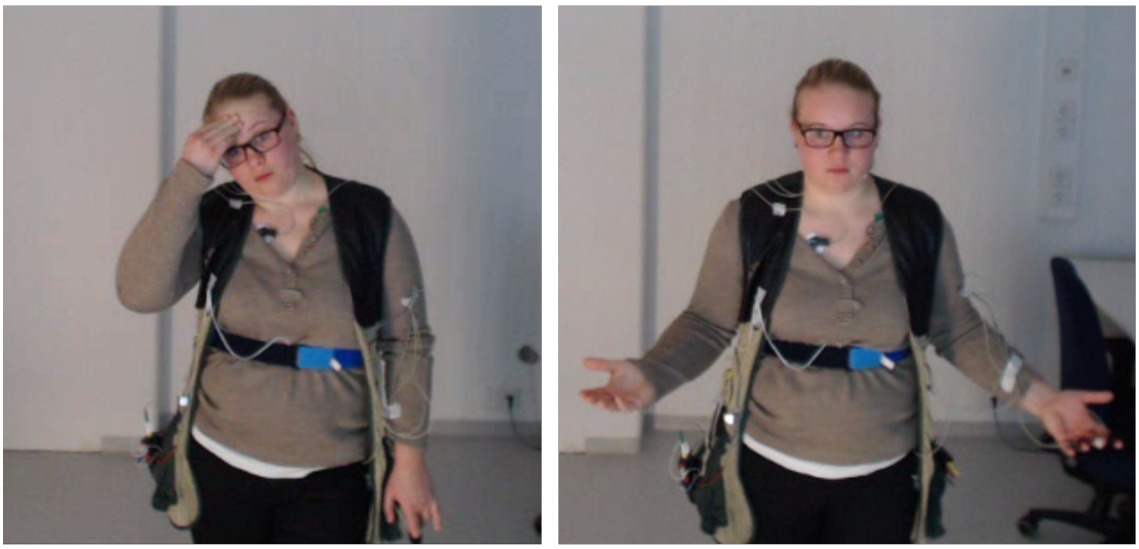

Abbildung 4.6.: Beispieldarstellung von typischen Körperposen der uulmMAC Probanden. Links: Selbstberührung des Kopfes durch kratzen am Kopf. Dies Verhalten tritt im Datensatz weitgehend gleichverteilt auf, es existieren jedoch starke individuelle Unterschiede. Einige Probanden neigen unter Überforderung vermehrt zu Selbstberührungen, andere eher bei Unterforderung. Rechts: Die Probandin streckt ihre Hände vom Körper weg. Dieses Verhalten ist vermehrt bei Überforderung in Folge falscher Antworten zu beobachten.

oben genannten Bedingung beschreibt. Kann eine Schwelle für eine bestimmte Person gefunden werden, gilt die Schwelle nicht notwendigerweise für andere Personen. Zur Vermeidung einer individuellen Schwelle wird die mittlere Entfernung der jeweiligen Gelenke direkt als kontinuierliches Eingangssignal für die Klassifikation verwendet. Die Wahl einer geeigneten schwelle wird so in das Lernproblem verschoben und dort gelöst. Abbildung 4.8 gibt ein Beispiel ausgewählter Abstände.

Durch weiteres exploratives Betrachten des Videomaterials und das Beobachten des Verhaltens der Teilnehmer konnten zwei lineare Bewegungen als dynamische Merkmale identifiziert werden: *Kurzzeitiges wegbewegen der Hände vom Torso* und *kurzzeitiges berühren oder kratzen am Kopf oder Gesicht*. Diese beiden Merkmale sind in Abbildung 4.6 dargestellt und treten ebenfalls weitgehend gleichverteilt im Datensatz auf. Eine auffällig häufige Kombination mit nicht neutralen Gesichtsausdrücken konnte dabei beobachtet werden und lässt auf eine gute Fusionierbarkeit mit Videodaten hoffen. Diese Gesten werden erfasst, indem ebenfalls die mittleren Gelenkabstände berechnet werden. Da die Merkmale jedoch nur kurz auftreten wird die zeitliche Ausdehnung weitgehend ignoriert und nur das reine Auftreten als Merkmal verwendet.

Abschließend wurden noch die mittleren Geschwindigkeiten und Beschleunigungen aller Gelenke berechnet. Diese sind ein Indikator für die Unruhe oder auch Zappeligkeit eines Probanden.

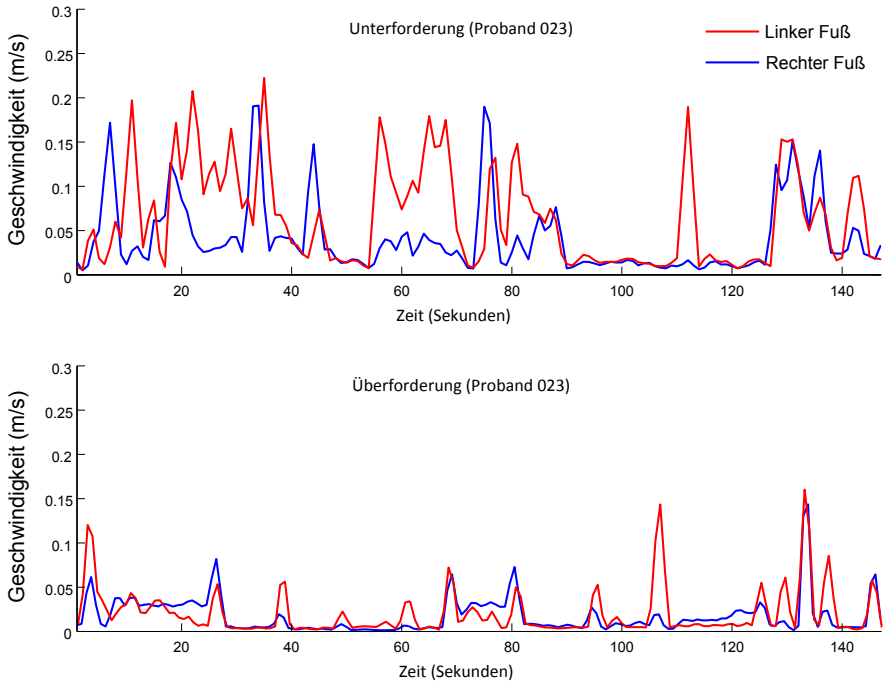

Abbildung 4.7.: Gelenkgeschwindigkeit als Körperpose Merkmal. Oben: Geschwindigkeiten der Fußgelenke der Unterforderungsphase für Proband 023. Unten: Geschwindigkeiten der Fußgelenke der Überforderungsphase für Proband 023. In überforderten Phasen steht der Proband konzentriert still, während er im Falle von Unterforderung hin und her läuft. Diese Verhalten ist stark individualspezifisch.

Für jedes Gelenk wurden die Werte in einem 4 Sekunden Fenster ermittelt. Abbildung 4.7 zeigt dies beispielhaft für die Füße eines Probanden. Auffällig ist dabei die deutlich größere Unruhe in der Unterforderung. Die Bewegungsfreude ist individuell stark verschieden. Die Gelenkgeschwindigkeit wird durch die erste Ableitung der Gelenkpositionen (x, y, z) über die Zeit $\bar{v} = \frac{\Delta \mathbf{s}}{\Delta t}$ beschrieben. Die Gelenkbeschleunigung berechnet sich näherungsweise durch durch $\bar{a} = \frac{\Delta \mathbf{v}}{\Delta t}$.

Die meisten Teilnehmer zeigten eine ziemlich aktive Kopfbewegung. Der Kinect Sensor misst die Rotationswinkel des Kopfes in Gier-, Pitch- und Rollrichtung. Der Gierwinkel kann verwendet werden, um festzustellen, ob ein Teilnehmer in die Kamera schaut oder nicht. Der Neigungswinkel gibt die Bewegung zum Boden oder zur Decke an, der Rollwinkel entspricht der Kopfneigung. Die Winkel werden vom Sensor bereits in aufbereiteter Form geliefert und können direkt als Merkmal eingesetzt werden.

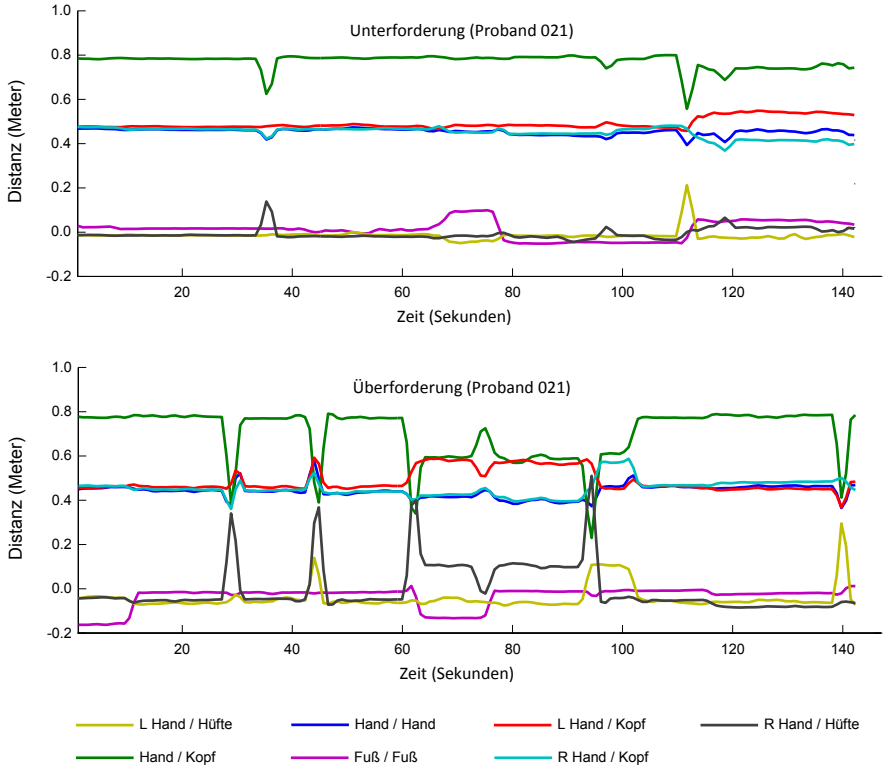

Abbildung 4.8.: Gelenkdistanzen als Körperpose Merkmal. Oben: Distanzen ausgewählter Gelenke der Unterforderungsphase für Proband 023. Unten: Distanzen ausgewählter Gelenke der Überforderungsphase für Proband 023. Gelenkdistanzen zeigen Selbstberührungen und abgestreckte Gliedmaßen gut an.

Ergebnisse der Poseerkennung

Bereits bei der Definition der Merkmale konnten große individuelle Unterschiede identifiziert werden. Eine interindividuelle LOSO Klassifikation erschien daher wenig vielversprechend. Versuche diese dennoch zu realisieren zeigte, dass die Resultate nicht signifikant besser als raten sind. Alle weiteren Untersuchungen fokussieren sich daher auf individualspezifische Experimente.

Um zunächst eine Vorstellung davon zu bekommen, wie gut sich verschiedene Klassifikationsmethoden eignen die erzeugten Merkmale zu klassifizieren, wurden mehrere Tests mit verschiedenen Klassifikatoren durchgeführt. Abbildung 4.9 zeigt die Ergebnisse eines 3-NN-Ensemble, einer LDA und einem hybriden Ansatz aus beiden. Dieser wurde mit einer Zufallsunterraummethode, welche eine Verallgemeinerung des HF-Algorithmus ist, erzeugt [33]. Dieser arbeitet mit einer zufälligen Teilmenge des Merkmalsraum. In diesem Fall wurde dieser, wie von Breiman im Originalpapier vorgeschlagen, auf $m = \lfloor log_2(M) \rfloor$ gesetzt. Zusätzlich wurde ein Random Forest trainiert. Dieser ist leicht zu trainieren und kann große Datenmengen gut beherrschen. Außerdem benötigt er nur wenige Parameter.

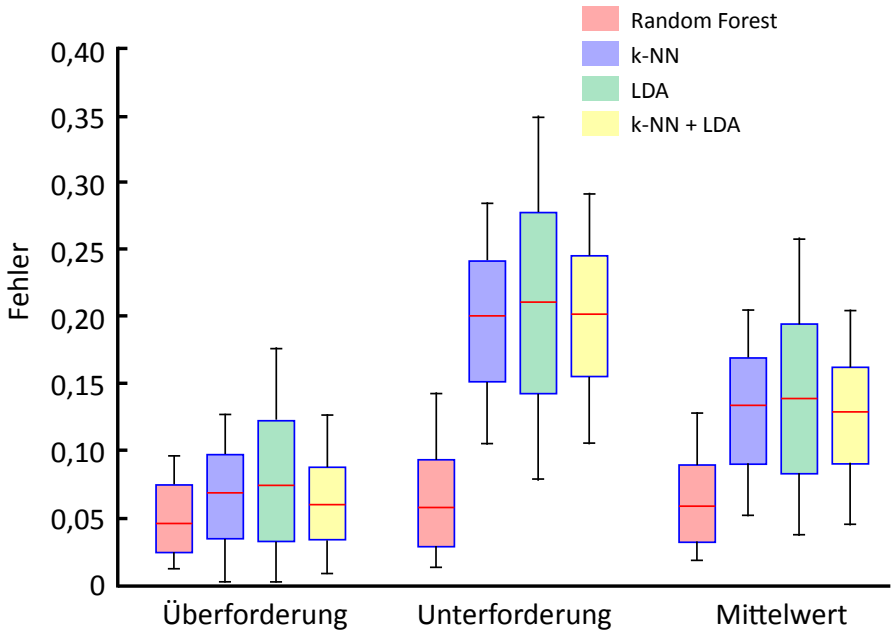

Abbildung 4.9.: Vergleich verschiedener Klassifikatoren zur Erkennung von Über- und Unterforderung anhand der Körperpose

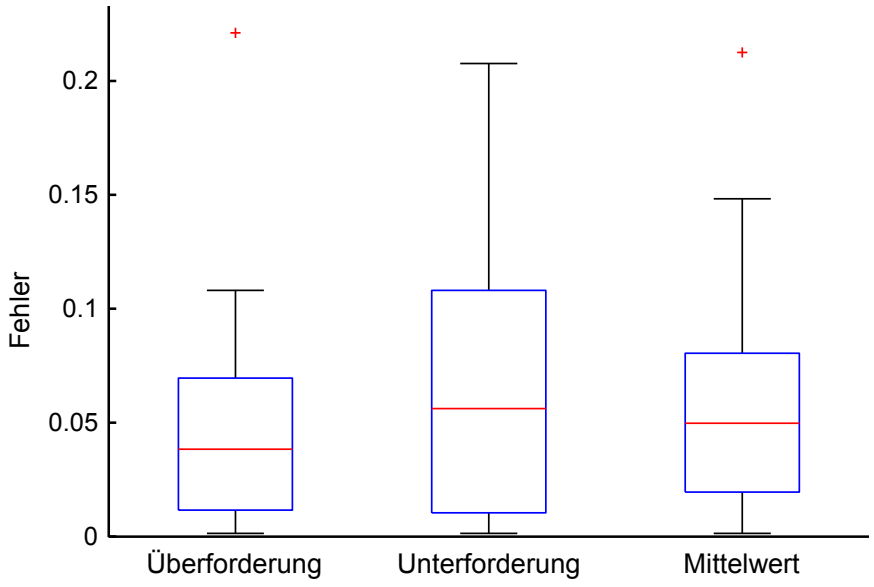

Abbildung 4.10.: Klassifikationsresultat der Körperpose basierten Erkennung kognitiver Last, individualspezifische Klassifikation für alle Probanden des uulmMAC. Für jeden Probanden wurde eine 10-fach Kreuzvalidierung unter Verwendung eines Random-Forest Klassifikators durchgeführt.

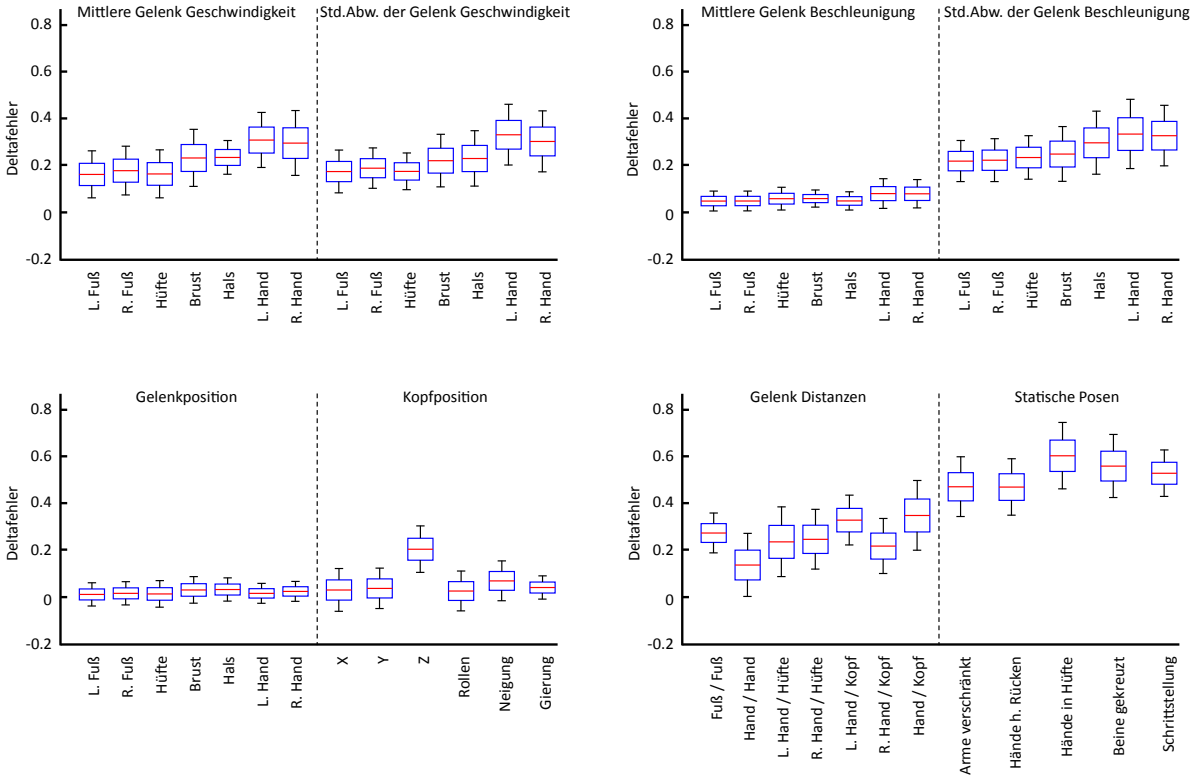

Abbildung 4.11.: Mittlere Merkmalsbedeutung der Körperpose Merkmale. Der Deltafehler zeigt die Änderung der Klassifikationsrate an, welche auftritt wenn Merkmale permutiert werden. Hohe Werte bedeuten somit einen hohen Einfluss eines Merkmals auf die Separierbarkeit im gegebenen Problem und markieren so wichtige Merkmale.

Ein Random Forest [32] mit 200 Bäumen bildet aufgrund seiner guten Performanz und leichten Handhabbarkeit unter Verwendung einer 10-fach Kreuzvalidierung die Grundlage der Evaluation. Eventuell überlappende Merkmale im Trainings und Testset wurden entfernt. Die individuellen Ergebnisse sind in der Abbildung 4.10 dargestellt. Aufgrund des meist expressiveren Verhaltens bei kognitiver Überforderung führte diese Klasse zu einem geringeren Klassifizierungsfehler. Insgesamt wurde eine Fehlerrate von etwa 5% erreicht. (Der interindividuelle Fehler liegt bei 47%).

Die folgende Untersuchung wird nun zeigen, welche der vorgeschlagen Merkmale die größte Bedeutung haben. Die Bedeutung der Merkmale wurde untersucht, indem die Werte eines Merkmals nach dem Zufallsprinzip permutiert und anschließend klassifiziert wurden [33]. Die resultierenden Klassifkationsfehler geben an, um wieviel sich der mittlere Klassifizierungsfehler, bezeichnet als *Delta-Fehler*, durch die Permutation geändert hat. Hohe Deltafehlerwerte repräsentieren also wichtige Merkmale. Die mittleren Deltafehler für alle Probanden und alle Merkmale sind in Abbildung 4.11 dargestellt. Die großen Varianzen zeigen auch hier wieder den

enormen Grad der Individualisierung. Nur sehr wenige Merkmale zeigen über die individuelle Grenze hinweg eine große allgemeine Bedeutung. Auffallend ist, dass Merkmale der mittleren Beschleunigung der Gelenke wenig bedeutsam sind (Abbildung 4.11 rechts oben, Merkmale 1-7), während die Standardabweichung der Beschleunigungen relevant ist (Abbildung 4.11 rechts oben, Merkmale 8-14). Dies bestätigt den subjektiven Eindruck, dass Probanden in der ein oder anderen Situation deutlich anderes individuelles Bewegungsverhalten zeigen. Zudem fällt auf das der Z-Wert der Kopfposition von Bedeutung ist (Abbildung 4.11 links unten). Dies bestätigt den explorativ gewonnen Eindruck einer leicht nach vorne gebeugten Haltung hin zum Bildschirm unter kognitiver Belastung. Die durch explorative Beobachtungen entworfenen statischen Merkmale haben die größte Merkmalsrelevanz

4.3.4. Erkennung anhand von Biophysiologie

Anders als bei allen zuvor genannten Modalitäten ist es fast unmöglich, das menschliche biophysiologische Verhalten willentlich zu beeinflussen. Zudem ist sind diese Signalquellen dauerhaft und unterbrechungsfrei vorhanden. Dies macht diese Modalität sehr interessant für die Erkennung von Emotionen. Allerdings ist es ungleich schwerer diese Signal zuverlässig zu messen. Es werden wesentlich spezialisiertere Geräte benötigt. Die meisten Sensoren müssen zum gegenwärtigen Zeitpunkt zudem noch direkt auf dem Körper angebracht sein. Im folgenden werden die gängigsten Signalquellen und ihre Merkmale vorgestellt. Diese wurden auch in den unimodalen Experimenten zur Erkennung kognitiver Last des uulmMAC verwendet, die Ergebnisse sind in [101] veröffentlicht

Biophysiologische Merkmale

Hierbei werden entgegen der modellgetriebenen Merkmale der Akustik und der Gesichtsexpresionen eine Vielzahl von statistischen Merkmalen der Zeit- und Frequenzdomäne der Signale erfasst. Dies resultiert in einen 58 Merkmale umfassenden Vektor je Signalquelle. Vor Extraktion der statistischen Merkmale werden die Signale durch Butterworth-Filterung vorverarbeitet, dabei wird eine untere Grenzfrequenzen von 10 Hz und eine obere Grenzfrequenz von 125 Hz, eingesetzt bei einer Filterordnung von vier. Das EKG Signal bildet hier eine Ausnahme, da hier modellgetriebene Merkmale verfügbar sind, werden für diese Signalquelle zusätzliche Merkmale erfasst. Da biophysiologische Merkmale einen eher langzeitigen Charakter haben, werden hier Analysefenster mit 5 Sekunden Länge verwendet. Die Dauer eines Herzschlages beträgt ca. 1 Sekunde, die Dauer eines Atemzuges beträgt sogar mehrere Sekunden. Um dennoch eine

ausreichende Merkmalsanzahl zu bekommen, werden die Fenster nicht wie sonst üblich um halbe Breite über das Signal geschoben, sondern nur um 0,1 Sekunden verschoben.

Elektroenzephalografie (EEG) [20]: Das EEG ist eine Methode, um die Gehirnaktivität zu messen, indem elektrische Potentialdifferenzen (Spannung) entlang der Kopfhaut abgetastet werden. Heutzutage werden die meisten Messungen mit dem *10-20-System* durchgeführt, bei dem 21 Elektroden fest positioniert sind, je nach Größe des darunter liegenden Bereichs.

Unterschiedliche Wachheitsstufen können unterschiedlichen Frequenzspektren zugeordnet werden. Zum Beispiel steht das *Delta-Frequenzband* von 0,1 bis 4 Hz im Zusammenhang mit dem *nicht REM Schlaf*, während die *Beta-Wellen* (16-31 Hz) auf treten, wenn der Mensch aktiv denkt. Typische Analysen in der MMI sind jedoch die Suchen nach sogenannten ereignisbezogenen Potentialen, die Reaktionen des Menschen auf Sinneswahrnehmungen abbilden. Aufgrund der aufwendigen und unzuverlässigen Messmethode können diese Signale bisher in realistischen MMI Szenarien nicht erfasst werden. Diese Arbeit verwendet diesen biophysiologischen Kanal daher nicht.

acfemg [14]: Mittels der EMG Methode werden elektrischer Aktivitäten von Muskeln erfasst. Für die MMI sind die zwei Muskeln *Musculus Zygomaticus* und *Musculus Corrugator Supercilii* im Gesicht und der *Musculus Trapezius* im Nacken von besonderem Interesse. Die beiden Gesichtsmuskeln sind verantwortlich für die Auf- und Abwärtsbewegung des Mundwinkels (Lächeln) und der Augenbraue (Strinrunzeln), der Nackenmuskel ist typischerweise mit Stressempfinden korreliert (verspannter Nacken). Das EMG erfasst die Aktivitäten der Muskeln durch Ableitung von elektrischen Potentialen auf der Hautoberfläche.

Hautleitwert (engl. Skin Conductance Level) (SCL) [57]: Der deutsche Physiologe Emil Heinrich Du Bois-Reymond war der erste Wissenschaftler, der die elektrische Leitfähigkeit der menschlichen Haut im Jahr 1849 untersucht hat. Die Hautleitfähigkeit ist in hohem Maße mit den Aktivitäten der Schweißdrüsen verbunden. Wird das sympathische Nervensystem einer Person angeregt, erhöht sich der Schweiß und somit der Hautleitwert. Die elektrodermale Aktivität kann kaum beeinflusst werden und ist zudem relativ leicht messbar. Über den Hautleitwert können plötzliche Zustandsänderungen gut erfasst werden.

Atmung: Zur Untersuchung der Atmung kann ein dehnbares Sensorband um den Thorax gelegt werden. Dieses misst die relative Expansion / Kontraktion der Brust. Die Atemfrequenz liefert unter anderem Informationen über den Stresslevel des Menschen.

Elektrokardiographie (EKG) [284, 250]: Für das EKG-Signal besteht die Vorverarbeitung aus linearem Detrending und die Normierung der Signale mit den mittleren *R-Peaks*. Dies ist notwendig, da die Sensoren bei jedem Teilnehmer nicht exakt an der gleichen Position sitzen,

Merkmal	Individuell	Interindividuell (LOSO)
EKG	**91.52 %**	**66.53 %**
EMG	74.53 %	51.03 %
SCL	64.68 %	51.51 %
Atmung	64.68 %	51.50 %
Temperatur	60.89 %	54.23 %
Fusion	**91.91 %**	65.61 %

Tabelle 4.7.: Evaluation der biophysiologischen Merkmale zur Erkennung von Über- und Unterforderung im uulmMAC für individuelle und interindividuelle Klassifikation.

wodurch individuelle jedoch aussagelose Spitzenwerte für jeden Teilnehmer auftreten. Die aus dem EKG-Signal extrahierten Merkmale basieren auf 25 Merkmalen der Waveletanalyse und statistischen Merkmalen gewonnen aus dem *PQRST-Komplex*, anstelle der reinen Abtastwerte.

Ergebnisse der biophysiologischen Erkennung

Um die kognitive Belastung auf Basis der Biophysiologie zu erkennen, werden Random Forest Klassifikatoren für jeden Kanal eingesetzt, welche jeweils auf den 58 statistischen Merkmalen bzw. auf den 25 Merkmalen der EKG Wavletanalyse trainiert wurden. Auch hier wurde wieder zwischen individuell optimierten Klassifikatoren und der interindividuellen LOSO Methode, bei der der Klassifikator zur Trainingszeit keine Daten des zu klassifizierenden Individuums verwendet hat, unterschieden. Neben der Untersuchung der einzelnen Signalquellen wurde auch eine Fusion aller biophysiologischen Merkmale durchgeführt. Die Fusion erfolgte mittels Moore-Penrose Pseudoinverse. Die Evaluation erfolgte, wie bei allen anderen Modalitäten, unter Verwendung der *40-Singles* Probanden Gruppe des uulmMAC.

Tabelle 4.7 zeigt die durchschnittlichen erzielten Klassifikationsraten. Auffallend ist hier, wie in bisher allen Modalitäten, eine signifikant bessere Performanz der individualisierten Klassifikatoren. Zudem fällt auf, dass das EKG Signal deutlich bessere Resultate erzielt. Dies dürfte in dem modellgetriebenen Merkmalsansatz der Waveletanalyse begründet sein, welcher vermutlich mehr Informationen aus dem Signal gewinnen kann als eine rein generische statistische Signalbeschreibung. Aufgrund der hohen Performanz des EKG konnte eine Fusion dieses Signals mit allen anderen Signalen keinen Zugewinn der Erkennungsleistung mehr erzielen.

Die EMG Verläufe zeigen bei einigen Probanden Auffälligkeiten. Es kann grundsätzlich zwischen den 3 Varianten der Trapetiusmuskelaktivität unterschieden werden.

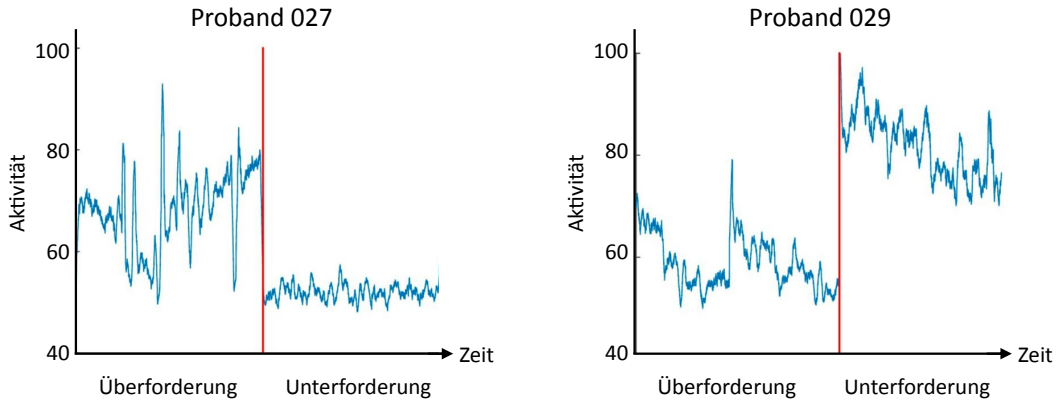

Abbildung 4.12.: Verlauf der Aktivität des Trapetiusmuskels abhängig von der kognitiven Belastung für Proband 027 und 029 im Vergleich.

- Größere Aktivität in der Überforderunssituation (Abbildung 4.12 links)
- Größere Aktivität in der Unterforderungssituation (Abbildung 4.12 rechts)
- Kein signifikanter Unterschied in beiden Phasen

Im Falle individualisierter Klassifikatoren kann dieser Aktivitätsunterschied sehr gut zur Klassifikation genutzt werden. Dies zeigt sich auch in den Resultaten. Im Falle eines globalen Klassifikators für alle Probanden liegt die Erkennungsrate in etwa auf Chancelevel, wärend der individualisierte Klassifikator mit 75% Erkennungsrate nach dem EKG Klassifikator am zweitbesten performt. Ein ähnliches Verhalten hat sich bereits zuvor in der Körperposenmodalität gezeigt. Dort konnten starke individuelle Unterschiede bei der situationsabhängigen Bewegungsfertigkeit der Probanden nachgewiesen werden.

Zusammenfassend ist festzustellen, dass die Biophysiologie gut geeignet ist kognitive Belastungszustände zu schätzen, die Klassifikatoren jedoch unbedingt individualspezifisch trainiert werden sollten.

4.4. Multimodale Emotionserkennung

Für diese Arbeit wurden zahlreiche multimodale Architekten für verschiedenste Fragestellungen, Demonstratoren und Untersuchungen eingesetzt. Übersichten zu den zahlreich eingesetzten Methoden finden sich in den Veröffentlichungen zu dieser Arbeit unter anderem in [226, 201, 101, 131, 157, 243, 153, 123, 202]. Im folgenden werden drei ausgewählte Architekturen vorgestellt. Bei der ersten Architektur handelt es sich um frühe Arbeiten audio-

visueller Fusion auf der ersten affektiven Benchmark-Datenbank AVEC [223], beschrieben in Abschnitt 2.2. Die zweite vorgestellte bimodale Architektur kombiniert physiologische Daten mit akustischen Daten und verzichtet dabei auf die in vielen Arbeiten vorausgesetzten visuellen Komponenten. Außerdem beantwortet die Untersuchung dieser Fusion die Frage, ob Sprache die Analyse von Atmungssignalen und andere biophysiologischen Signalen nachteilig beeinflusst. Die dritte vorgestellte Architektur kombiniert alle Modalitäten des uulmMAC zu einer gemeinsamen Schätzung kognitiver Belastung. Dabei wird zusätzlich der Einfluss der Individualität des Menschen auf die Klassifikatoren beleuchtet. Abschließend wird eine Möglichkeit vorgeschlagen den situativen Kontext als weitere Informationsquelle nutzbar zu machen.

4.4.1. Hierarchische audio-visuelle Emotionserkennung

Diese Arbeit entstammt der Teilnahme am ersten internationalen Vergleichskampf zur Affektschätzung AVEC. Der damals neu veröffentlichte AVEC Datensatz bildet dabei die Grundlage. Er beinhaltet die Kommunikation eines Menschen mit einem Avatar, bei welcher der Mensch von erlebten Geschichten erzählt. Die Annotation erfolgte binär in den Kategorien *Erregung*, *Erwartung*, *Ausdruckskraft* und *Valenz*. Methode und Ergebnis der Wettkampfteilnahme sind in [123] veröffentlicht.

Für jede der Labeldimension und für jedes akustische Merkmal wurde eine Menge von Hidden Markov Modellen (HMM) trainiert [32, 184]. Die versteckten Zustände und die Anzahl der Komponenten des HMM wurden mittels einer Parametersuche optimiert, was zur Wahl von drei versteckten Zuständen und zwei Komponenten im GMM mit vollen Kovarianzmatrizen geführt hat. Die Bewertung des Optimierungsprozesses zeigte zudem, dass einige Merkmale keinen Beitrag zur Klassifikation leisten können. Der akustische Merkmalssatz wurde daher auf Energie, Grundfrequenz und MFCC beschränkt.

Für jedes Label werden die Log-Likelihoods jedes HMM, die je Merkmal trainiert wurden, summiert. Um robustere Modelle zu erhalten wurden zusätzlich fünfmal soviele Modelle pro Klasse verwendet wie notwendig, und deren Ausgaben summiert. Des Weiteren wurde angenommen, dass sich die Labels im Laufe der Zeit nur langsam verändern. Die Klassifizierung wurde daher auf der Basis von Turns durchgeführt, und die darin enthaltenen Einzelentscheidungen über die Zeit summiert. Die Ergebnisse dieses Ansatzes sind in Tabelle 4.8 (Audio) aufgeführt.

Zur Video Klassifikation wurde eine ν-SVM eingesetzt [244]. Dabei bilden konkatenierte Merkmalsvektoren aus 300 Form- und 300 Bewegungsmerkmale die Grundlage, um die ν-acsvm mit einem linearen Kernel zu trainieren und probabilistische Ausgaben nach Platt

zu erzeugen [178]. Aufgrund von damaligen Speicherbeschränkungen wurden nur 10.000 zufällig gezogene Stichproben zum Training verwendet. Wieder wurde eine Parametersuche angewendet, um möglichst geeignete Parameter zu erhalten, was zur Einstellung von $\nu = 0{,}3$ für Erregung und Ausdruckskraft und $\nu = 0{,}7$ für Erwartung und Valenz führte. Tabelle 4.8 (Video) zeigt die Ergebnisse.

Basierend auf den Ergebnissen der unimodalen Resultate wurde die Fusion unter Verwendung eines MLP durchgeführt, um die endgültige Schätzung zu erhalten. Die Ergebnisse sind in Tabelle 4.8 (Fusion) aufgeführt. Dabei bildet ein achtdimensionaler Eingangsvektor (Vier Klassen je Modalität) die Eingangsschicht des MLP. Die MLP Ausgabeschicht liefert die finale Klassifikationsentscheidung. Die zunächst scheinbar sehr schlechten Resultate der Schätzung sind, objektiv betrachtet, als hervorragend zu bewerten, und führten zu einem zweiten Platz im weltweiten Vergleich. Zudem erlaubte die Challenge nur interindividuelle Architekturen. Nach heutigem Stand der Forschung ist klar, dass nichtindividualisierte Klassifikatoren kaum in der Lage sind bessere Ergebnisse zu liefern.

	Erregung	**Erwartung**	**Ausdruckskraft**	**Valenz**
Audio	66,9%	**62,9%**	**63,2%**	65,7%
Video	58,2%	53,5%	53,7%	53,2%
Fusion	**69,3%**	61,7%	61,3%	**68,8%**

Tabelle 4.8.: Erkennunsraten der ersten AVEC Challenge auf Basis binärer Labels in 4 Dimensionen (*Erregung*, *Erwartung*, *Ausdruckskraft* und *Valenz*).

4.4.2. Bimodale Erkennung von kognitiver Last

Dieser, in [101] veröffentlichte Ansatz, kombiniert die akustische und biophysiologische Modalität mit dem Ziel die kognitive Belastung der Probanden des uulmMAC zu schätzen. Neben der Untersuchung möglicher Fusionsarchitekturen wird zudem der Frage nachgegangen, inwieweit die Qualität der biophysiologischen Signale durch Sprache beeinflusst wird, und ob es sinnvoller ist, während Nutzer sprechen, nicht auf biophysiologische Daten zurückzugreifen.

Um bimodale Klassifizierungsergebnisse zu erzielen, wurden zwei unterschiedliche Fusionsansätze (Moore-Penrose Pseudoinverse und Konfidenzsummation) verwendet, um die zuvor vorgestellten unimodalen Ergebnisse zu fusionieren. Zur Fusion werden dabei die probabilistischen Ausgaben von vier unimodalen Klassifikatoren eingesetzt. Je Modalität ein Klassifikator für Überforderung und einer für Unterforderung. Für die bimodale Auswertung wurden die Daten aus beiden Modalitäten entnommen, sofern ein Proband spricht. Gibt es keine Eingabe aus Sprache basiert das Klassifikationsergebnis nur auf der physiologischen Klassifikation.

Moore-Penrose Pseudoinverse Ansatz

Die Pseudoinverse wird verwendet, um Gewichtungen für Eingangsklassifikatoren zu erstellen. Dabei werden zwei separate Gewichtsvektoren betrachtet. Einer für die Erkennung von Überforderung und einer für die Erkennung von Unterforderung. Entgegen aller bisherigen Ansätze wird also nicht anhand einer Schwelle zwischen den beiden Klassen unterschieden, sondern es werden zwei separate Schätzungen berechnet. Die Klassenentscheidung erfolgt durch den größeren Wert der beiden Einzelklassifikatoren. Die Gewichte werden berechnet durch:

1. Training des unimodalen Ensembles.
2. Bilden eines Fusionstrainingssatz aus N Merkmalsinstanzen, welche nicht für die unimodalen Trainings verwendet wurden.
3. Klassifikation (unimodal) alle Instanzen des Fusinstrainingssatzes
4. $X = N \times 2$ je Klasse (2, da bimodal)
5. Berechne die Pseudoinverse $X^+ = (X^t X)^{-1} X^t$ bzw. $X^+ = \lim_{\alpha \to 0}(X^t + \alpha I)^{-1} X^t$
6. Berechne die Gewichte $w = X^+ Y$

Die gleichen Schritte werden unabhängig für beide Klassen durchgeführt. Durch diese Schritte werden die Gewichte für die bimodale Fusionsschicht gesetzt. Die berechneten Gewichte hängen dabei stark von den Daten ab, die präsentiert werden, um die Fusionsschicht zu trainieren. Für die individuelle Klassifikation existieren durchschnittlich 284 Instanzen, für interindividuelle Klassifikation etwa 16450 Instanzen zum Training der Fusionsschicht. Dies entspricht etwa 11% der Daten, die zum Trainieren der Ensembles verwendet werden.

Tabelle 4.9 zeigt die Klassifikationsresultate. Für die individuelle Klassifikation wurden 94.07 % Erkennungsrate erreicht. Das Gewicht für die Eingangskanäle der individuellen Klassifikation hat durch Berechnung der Pseudoinversen eine mittlere optimale Gewichtung von 0,23 für Sprache und 0,77 für Biophysiologie ergeben, bei $\sigma^2 = 0.38$. Die große individuelle Differenz der einzelnen Modalitätsgewichte ist in der großen Varianz der physiologischen Klassifikation zwischen den verschiedenen Probanden begründet.

Für die interindividuelle LOSO Klassifikation beträgt das Gewicht für Sprache 0,76 und für physiologische 0,24. Die Varianz der Gewichte für die LOSO Klassifikation ist viel kleiner, da sich nur etwa 1,7% der Daten innerhalb des Trainingssatzes für die Klassifizierung des nächsten Probanden ändern. Die durchschnittliche Erkennungsrate für den interindividuellen Ansatz

	Individuell	Interindividuell (LOSO)
Überforderung	95.59 %	75.80 %
Unterforderung	92.55 %	76.82 %
Fusion	94.07 %	76.31 %

Tabelle 4.9.: Akustische biophysiologische bimodale Erkennunsraten von Überforderung, Unterforderung und Fusion durch Moore-Penrose Pseudoinverse für individuelle und interindividuelle (LOSO) Klassifikation.

beträgt allerdings auch nur 76,31%, was erneut die Problematik eines personenunabhänigen Ansatzes bestätigt.

Konfidenz Summen Ansatz

Die zweite verwendete Fusionsmethode ist ein konfidenzbasierter Summationsansatz. Diese Methode hat keine Lernphase und kann direkt verwendet werden, um die unimodalen Klassifikationsergebnisse zu fusionieren. Die unimodalen Ausgaben als Fusionseingaben bestehen wie zuvor aus je einem Wert für Überforderung und Unterforderung je Modalität. Also insgesamt vier kontinuierlichen Werten. Der kontinuierliche Wert wird dabei als Konfidenz interpretiert. Die Gewichtung der Modalitäten ist entgegen des vorherigen Ansatzes gleich für beide Modalitäten. Allerdings wird jede einzelne Stimme im Summationsprozess mit der Ausgabekonfidenz der vorangegangen Schicht gewichtet.

Tabelle 4.10 zeigt, dass die individuelle Klassifizierung eine durchschnittliche Erkennungsrate von 96.41% erreicht. Der Unterschied zwischen der Erkennbarkeit von Unterforderung und Überforderung ist mit 3,04% moderat. Verglichen mit den unimodalen Ergebnissen konnte die Erkennungsleistung durch Fusion um 1% bis zu 29% je nach Individum verbessert werden. Der Verbesserungseffekt ist besonders ausgeprägt für Probanden, die niedrige unimodale Klassifikationsraten im biophysiologischen Kanal aufweisen. Ein weiterer Effekt ist die stabilere Erkennung, welche sich in der Verringerung der Varianz von unimodal Audio ($\sigma^2 = 2.9$) und unimodal biophysiologisch ($\sigma^2 = 5.6$) auf bimodal Fusion ($\sigma^2 = 1.6$) verbessert hat. Die interindividuelle LOSO Klassifikation zeigt erneut deutlich schwächere Werte. Die beiden vorgestellten Fusionsansätze unterscheiden sich in ihren Resultaten nur geringfügig. Die Pseudoinverse konnte das interindividuelle Problem etwas besser lösen, während der Summationsansatz das individuelle Problem besser klassifizieren konnte. Signifikant sind die Unterschiede allerdings nicht, der Aufwand eine Pseudoinverse zu berechnen ist demnach fraglich.

	Individuell	Interindividuell (LOSO)
Überforderung	97.93 %	82.56 %
Unterforderung	94.89 %	61,15 %
Fusion	96.41 %	72.55 %

Tabelle 4.10.: Akustische biophysiologische bimodale Erkennunsraten von Überforderung, Unterforderung und Fusion durch Summe der Konfidenzen für individuelle und interindividuelle (LOSO) Klassifikation.

Einfluss von Sprache auf die biophysiologischen Signale

Die Veränderung der physiologischen Signale durch Sprache beruht auf zwei Faktoren. Der erste ist der direkte Einfluss durch die Vibration der Stimmbänder oder die Veränderung des Luftvolumens in der Lunge und der damit einhergehende Einfluss auf die Biophysiologie. So weist die Atmung beim Sprechen eine höhere gemessene Atemfrequenz auf, dies wiederum beeinflusst unter Umständen den Herzschlag, das Herz-Kreislauf System wiederum eventuell den Hautleitwert. Diese Änderungen sind sicher minimal, da aber auch schwach expressive emotionale Empfindungen nur geringe Signaländerungen erzeugen, ist eine Überlagerung, welche den Klassifikationsprozess stört, nicht auszuschließen. Ein zweiter Einflussfaktor ist die, durch Sprache erzeugte Bewegung selbst, dies könnte im Falle sensibler Sensorik eventuell zu leichter Artefaktbildung und somit Signalqualitätseinschränkungen führen.

Um die genannten Einflüsse datengetrieben bewerten zu können, wurde der Datensatz in zwei Teile geteilt. Ein Teil enthält nur Instanzen deren zugrundeliegendes Analysefenster Sprache und Biophysiologie beinhaltet und einen Vergleichsdatensatz, dessen physiologische Instanzen aus Zeitbereichen stammen, in denen der Proband nicht gesprochen hat. Die Analysefenster sind auf fünf Sekunden festgelegt. Fenster die auf die Segmentgrenzen von Sprachsegmenten fallen wurden nicht berücksichtigt. Die biopysiologischen Daten werden im folgenden wieder einzeln anhand ihrer Signalquelle bewertet um unterschiedliche Einflüsse auf die unterschiedlichen Signale erkennen zu können. Die Klassifikation erfolgt identisch zu den unimodalen vorgestellten biophysiologischen Ansätzen.

Tabelle 4.11 zeigt die Resultate für sprachüberlagerte und sprachfreie biophysiologische Merkmalsinstanzen für individuell spezifische und interindividuelle Klassifikatoren. Sowohl für den individuellen als auch interindividuellen Ansatz konnten in allen Kanälen unter Entfernung sprachüberlagerter Instanzen, Verbesserungen erzielt werden.

Um zu untersuchen, ob es einen statistisch signifikanten Unterschied zwischen der Klassifikation aller biophysiologischen Signale, abhängig von sprachlicher Überlagerung, gibt, wird ein T-Test verwendet. Die Ergebnisse dieses Signifikanz Tests sind in der Tabelle 4.12 ange-

Merkmal	Individuell		Interindividuell(LOSO)	
	mit Sprache	ohne Sprache	mit Sprache	ohne Sprache
EKG	91.08 %	92.72 %	65.70 %	67.85 %
EMG	74.54 %	75.19 %	49.55 %	53.62 %
SCL	64.18 %	66.08 %	49.95 %	54.15 %
Atmung	57.78 %	60.27 %	49.72 %	54.73 %
Temperatur	60.47 %	62.27 %	53.47 %	55.65 %
Fusion	91.73 %	93.00 %	65.82 %	68.42 %

Tabelle 4.11.: Evaluation der biophysiologischen Erkennungsraten abhängig von sprachlicher Überlagerung für individuelle und interindividuelle (LOSO) Experimente. In allen Fällen beeinflusst das Sprechen des Probanden die Erkennungsraten der biophysiologischen Merkmale negativ.

zeigt. Die Ergebnisse bestätigen die initiale Annahme, dass es einen Einfluss von sprachlicher Überlagerung auf die Qualität der biophysiologischen Signale gibt. Zum Beispiel werden bei $\alpha = 0,05$ alle Nullhypothesen der einzelnen Kanäle, außer dem individualspezifischen EMG Kanal, zurückgewiesen. Wie erwartet ist die Bedeutung überlagerter Sprache im Atemsigal besonders groß, und auch die individuellen EKG Signale zeigen deutliche Signifikanzen auf.

Resultierend aus dieser Erkenntnis ist für multimodale Anwendungen zu empfehlen, das Audiosignal zu nutzen um die Gewichtung biophysiologischer Klassifikationsergebnisse herabzusetzen, während Spracheingaben getätigt werden.

T-Test	Individuell mit/ohne Sprache	Interindividuell (LOSO) mit/ohne Sprache
EKG	0.004	$2.0\ 10^{-3}$
EMG	0.465	$4.4\ 10^{-5}$
SCL	0.041	$2.2\ 10^{-4}$
Atmung	0.007	$2.3\ 10^{-10}$
Temperatur	0.009	$4.0\ 10^{-4}$
Fusion	0.003	$4.9\ 10^{-9}$

Tabelle 4.12.: Signifikanz des sprachlichen Einflusses auf biophysiologische Merkmale mittels T-Test.

4.4.3. Multimodale Erkennung von kognitiver Last

Nachdem nun viele uni und multimodale Experimente durchgeführt wurden, wird abschließend ein Ansatz vorgestellt der alle Modalitäten des uulmMAC berücksichtigt, um daraus

eine gemeinsame Schätzung der kognitiven Belastung der Probanden zu ermöglichen. Die dazu eingesetzte Architektur sieht eine späte Fusion durch Einsatz eines MFN vor, welches die kontinuierlichen Schätzungen aller vorgestellten unimodalen Architekturen als Eingangssignal bekommt. Diese Architektur kann problemlos mit den unterschiedlichen Zeitskalen der unimodalen Analysefenstern und eventueller fehlender Sensoreingaben umgehen. Nach Vorstellung der interindividuellen LOSO Resultate wird anschließend eine individuelle Optimierung vorgestellt, welche zu verbesserten Ergebnissen führt, sofern der zu erkennende Proband der Architektur bekannt ist. Die Resultate dieses Ansatzes sind unter anderem in [131] veröffentlicht und werden dort zudem mit einem weiteren, hier nicht gezeigten, Kalmann Filter basierten Fusionsansatz verglichen.

Markov Fusionsnetzwerk

Das von Glodek et al. vorstellte MFN ist in der Lage multiple kontinuierliche Signale in einem gegeben Zeitfenster zu fusionieren [89]. Das hier verwendete MFN ist dabei in diskreten Zeitscheiben organisiert (feste Abtastrate), methodisch aber auch in der Lage für jede neue Eingabe eine neue Ausgabe zu erzeugen (variable Abtastrate). Zu unterschiedlichen Zeitpunkten auftretende unimodale Eingaben werden so zu einer technisch leichter handhabbaren konstanten Abtastung synchronisiert und liefern zu jedem Zeitschritt t eine Gesamtschätzung y_t, auch wenn keine Eingabe für den Zeitschritt t vorliegt. Dies wird durch eine Markov-Kette realisiert, welche die Schätzungen von y_t erzeugt und dabei sicherstellt das diese ähnlich zu y_{t-1} und y_{t+1} ist. y_t wird also von den Eingangswerten die dem Zeitschritt t zugeordnet sind beeinflusst, sowie von seinen zeitlichen Nachbarn.

Das MFN wird charakterisiert durch die drei Potentiale *Datenpotential*, *Glättungspotential* und *Verteilungspotential*. Das Datenpotential stellt sicher, dass die Gesamtschätzung nahe an den Eingangswerten der unimodalen Quellen liegt. Das Glättungspotential stellt die Glättung der Ausgaben y_{t-1}, y_t, y_{t+1} sicher. Das Verteilungspotential sorgt für die Summation aller Ausgabewahrscheinlichkeiten zu 1.

Alle drei Potentiale werden zu einer Wahrscheinlichkeitsdichtefunktion $p(\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_M)$ zusammengefasst. Dabei beinhaltet $\mathbf{Y}$ alle Schätzungen des MFN und $\mathbf{X}_1, \ldots, \mathbf{X}_M$ sind die unimodalen Schätzungen der M Modalitäten. Um die wahrscheinlichste Schätzung $\mathbf{Y}$ zu gewinnen, muss derren Erwartungswert bestimmt werden. Dies geschieht mittels Gradientenabstiegsverfahren. Alle Gradienten $\frac{\partial \mathbf{Y}}{\partial y_{i,t}}$ werden berechneten, dabei beschreibt i die Klasse und t die Zeitschritte. Die Schätzungen $y_{i,t}$ wandern dabei hin zum Erwartungswert der Wahrscheinlichkeitsdichtefunktion. Dies wird für eine festgelegte Iterationszahl wiederholt. Die Initialisierung von $\mathbf{Y}$ erfolgt durch die apriori Wahrscheinlichkeit aller Klassen oder den letz-

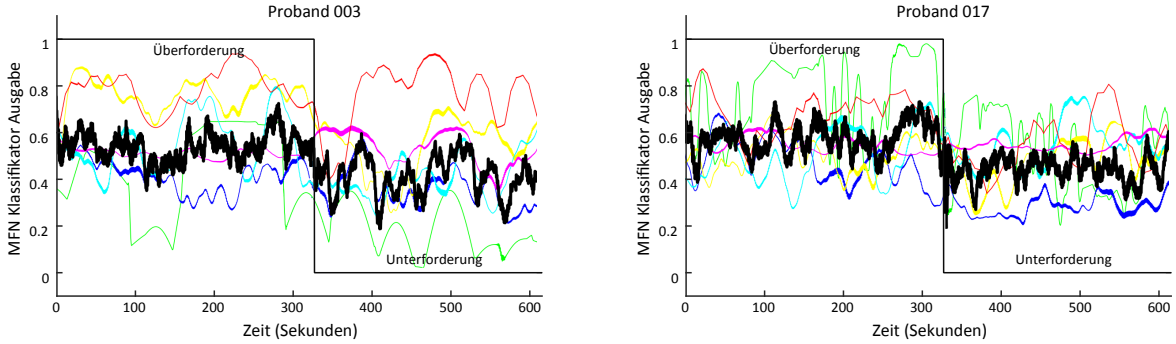

Abbildung 4.13.: Multimodale Performanz des MFN im Livebetrieb, exemplarisch für Proband 003 und 017. Die grünen, blauen, gelben, cyan, magenta und roten Kurven beschreiben die unimodalen Erkennungsresultate der Körperpose, EMG, SCL, EKG, Video und Audio Klassifikation als Eingangsgrößen des MFN. Die schwarze Kurve beschreibt die Ausgabe des MFN.

ten Stand der Klassifikation. Detaillierte Information zur Berechnung der MFN Architektur beschreibt Glodeck in [89].

Für die *offline* Fusion werden alle Daten aller Zeitschritte gemeinsam betrachtet, d.h. auch die unimodalen Eingaben, welche vom aktuell bewerteten Zeitschritt aus in der Zukunft liegen. Wird eine *online* Fusion durchgeführt werden nur die vergangenen Zeitschritte betrachtet, da zukünftige Messwerte noch nicht vorliegen. Die online Fusion wird dabei meist als gleitendes Fenster über der Vergangenheit realisiert, um die Anzahl der berücksichtigten Zeitschritte zu begrenzen.

Ergebnisse der vollständigen multimodalen Erkennung des uulmMAC

Abbildung 4.13 zeigt die Performanz eines MFN für die Probanden 3 und 17. Auf der x-Achse sind die diskreten Abtastpunkte bei einer Fusionsfrequenz von 2Hz abgetragen. Der y-Wert beschreibt den geschätzten Grad der kognitiven Überforderung. Blau gestrichelt dargestellt ist die Grundwahrheit, gewonnen aus der Induktionshypothese der Experimentalphase, bestätigt durch SAM. Die dünnen farbigen Linien repräsentieren die unimodalen Klassifikationen aller uulmMAC Modalitäten. Die dicke schwarze Linie repräsentiert die Gesamtschätzung des MFN.

Für Proband 3 ist zu erkennen, dass die akustische Erkennung (rot) sehr schlechte Ergebnisse geliefert hat, dies wurde durch das MFN kompensiert. Typisch für alle Modalitäten ist auch der teilweise sprunghafte Verlauf der Schätzungen mit durchaus großen Amplituden. Das MFN zeigt ebenfalls einen leicht verrauschten Verlauf, jedoch sind die Amplituden der Schätzung deutlich geringer, was wesentlich näher an einen tatsächlichen kognitiven Belastungsverlauf

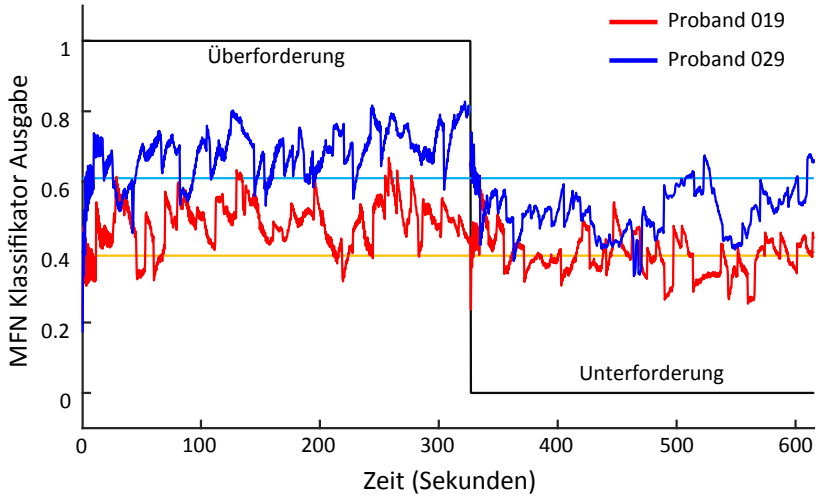

Abbildung 4.14.: MFN Ausgabe der kognitiven Belastungsschätzung für die Probanden 019 (rot) und 029 (blau). Die Grundwahrheit ist in schwarz eingezeichnet. Die optimale Entscheidungsschwelle für Proband 019 ist in orange angegeben, die optimale Schwelle für Proband 029 in hellblau. Zu erkennen ist der individuelle Unterschied der optimalen Schwelle. Proband 019 würde bei Verwendung der festen Schwelle von 0,5 fälschlicherweise überwiegend als unterfordert eingeschätzt, während Proband 029 überwiegend als überfordert eingestuft würde.

kommt. Die Schätzung entspricht bei weitem nicht dem binären Grundwahrheitssignal, dies war allerdings auch nicht zu erwarten, erkennbar ist allerdings ein Trend der Intensität hin zur Grundwahrheit. Im Vergleich beider dargestellter Probanden ist auch erkennbar, dass die Performanz der unimodalen Schätzungen stark variiert. Es kann keine Modalität ausgemacht, werden die in allen Fällen eine signifikant bessere Performanz als alle anderen Modalitäten liefert. Da nicht vorhersagbar ist welche Modalität im gegebenen Fall (Proband und Situation) gut funktioniert, besteht hier auch wenig Hoffnung zur Optimierung hin zu einer dominanten unimodalen Schätzung. Die Fusion ist in fast allen Fällen in der Lage eine bessere gemeinsame Schätzung zu liefern, dennoch liegt die personenunabhängige erreichte Erkennungsrate der kognitiven Last unter realistischen Bedingungen nur bei 64.42%, bei Verwendung der festen Schwelle 0,5 als Entscheidungsgrenze zwischen Überforderung und Unterforderung.

Bedeutung der Individualität für die Klassifikation

Die personenunabhängige Schätzung verwendet 0,5 als Schwellwert für die Klassifikation. Diese Schwelle ist in den wenigsten Fällen optimal, da es individuelle Unterschiede gibt in der Expressivität mit der Probanden ihre kognitive Belastung zeigen. Im folgenden wird diese Schwelle individuell angepasst. Alle zugrundeliegenden unimodalen Klassifikatoren bleiben

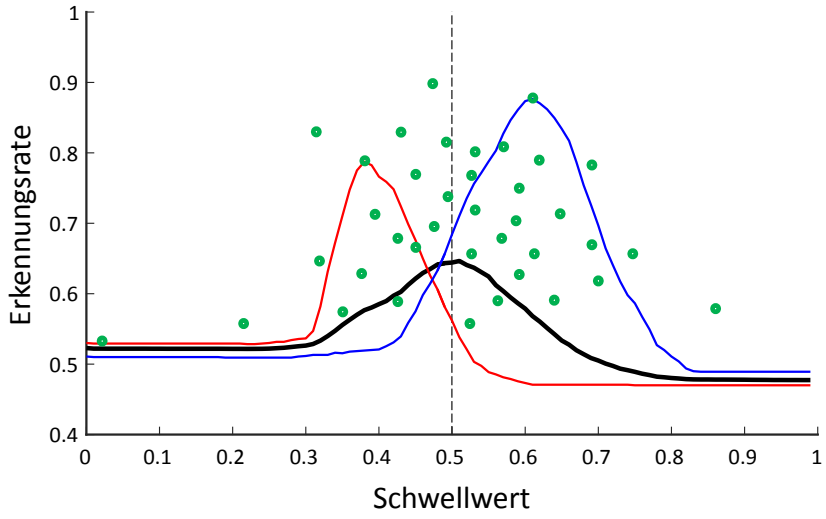

Abbildung 4.15.: Multimodale Erkennungsrate der kognitiven Belastung (Über- und Unterforderung) bei Verwendung individueller Schwellwerte bei der MFN Fusion. Die farbigen Kurven zeigen den Verlauf der Erkennungsrate abhängig vom Schwellwert der zur Bildung der binären Entscheidung verwendet wird für die Probanden 29 (blau) und 19 (rot). Die schwarze Kurve zeigt die durchschnittliche Erkennungsrate bei Verwendung einer gemeinsamen Schwelle für alle Probanden. Die grünen Markierungen zeigen die maximale individuelle Erkennungsrate und zugehörige Schwelle für alle Probanden des uulmMAC. Es ist deutlich zu erkennen, dass die individuelle Schwelle deutlich bessere Erkennungsraten erzeugt.

dabei unberührt, d.h. personenunabhängig trainiert. Die Parameter des MFN bleiben über die Probandengrenzen hinweg identisch. Die einzige Individualisierung liegt in der Anpassung des Schwellwertes selbst. So kann eine aufwendige Individualisierung durch Training von allen unimodalen Klassifikatoren und dem MFN selbst umgangen werden. (Würde diese durchgeführt, würden nahezu perfekte Klassifikationsraten erreicht werden, da die unimodalen individuellen Erkennungsarten bereits wie gezeigt sehr gut sind.)

Abbildung 4.14 zeigt zwei Verläufe des Fusinsresultates zweier unterschiedlicher Probanden. Für beide Probanden sind Unterschiede der mittleren Überforderungsschätzungen erkennbar, abhängig von der kognitiven Belastung, die in den untersuchten Belastungssequenzen induziert wurde. Proband 29 weist dabei einen generell höheren Schätzwert auf als Proband 19, zusätzlich hat Proband 29 eine etwas größere Differenz zwischen den beiden Sequenzen der maximalen Über- und Unterforderung. Eine vollständig individuell trainierte Architektur würde neben anderen individuellen Faktoren auch diesen Grundwert der Expressivität lernen. Wird nun allerdings die Entscheidungsschwelle an den Probanden angepasst, (rote und grüne Leine in Abbildung 4.14) kann auch unter Verwendung des interindividuellen Klassifikators eine deutliche Verbesserung erzielt werden.

Modalität	Erkennungsrate
Körperose	58,39%
EMG	56,27%
SCL	52,55%
Temperatur	54,14%
EKG	50,04%
Sprache	62,26%
Mimik	53,76%
Fusion: Gewichtetes Voting	55,40%
Fusion: MFN (feste Schwelle)	64,42%
Fusion: MFN (individuelle Schwelle)	71,95%

Tabelle 4.13.: Uni- und multimodale Erkennungsraten für die Detektion von Über- und Unterforderung im uulmMAC (interindividuell)

Die schwarze Kurve in Abbildung 4.15 zeigt die mittlere Erkennungsrate aller Probanden abhängig von der Schwelle. Wie durch das Training optimiert, liegt ihr Maximum bei einer Schwelle von 0,5, bei einem Wert von 65%. Betrachtet man nun allerdings die individuelle Erkennungsrate für jeden einzelnen Probanden ist festzustellen, dass bei Wahl optimaler individueller Schwellen wesentlich bessere Ergebnisse zu erzielen sind. Die grüne und orange Kurve in Abbildung 4.15 zeigt dies exemplarisch für die Probanden 19 und 29. Ihre Maxima liegen bei 80% bzw. 88% bei Verwendung individueller Schwellen von 0,37 und 0,62. Die orangen Kreise stellen die optimale Schwelle und die entsprechend erreichte Genauigkeit für die Probanden des 40Singles Subsets des uulmMAC dar.

Tabelle 4.13 zeigt eine Übersicht der unimodalen Erkennungsraten der für dieses Experiment eingesetzten unimodalen Klassifikatoren. Weiter zeigt Tabelle 4.13 stellt die durch Fusion errichten Erkennungsraten dar. Die durchschnittliche Erkennungsrate bei Verwendung interindividueller Klassifikatoren mit individueller Schwellen liegt bei 71,95% und somit um 7,53% höher als bei der Verwendung einer statischen Schwelle und 6,69% höher als die beste Einzelmodalität.

4.4.4. Einbezug von situativem Kontextwissen

Die Integration von Kontextinformationen ist eine Herausforderung für Algorithmen des maschinellen Lernens [88]. Sei eine Kontextinformation gegeben, welche in Form eine Aprioriwahrscheinlichkeit einer zu erwartenden emotionalen Reaktion entspricht, kann diese an unterschiedlichen Stellen der hierarchischen multimodalen Emotionserkennung eingesetzt werden. Eine Möglichkeit wird von Glodeck durch Einsatz von Conditioned Hidden Markov

Modell (CHMM) beschrieben. Das CHMM ist eine Erweiterung des klassischen HMM, bei dem die versteckten Zustände durch eine zusätzliche Folge von Bedingungen erweitert sind. Diese Bedingungen können durch eine externe Klassifikatorentscheidung bereitgestellt werden, die als zusätzlicher Kontext dient [90, 88, 87].

Neben dieser sehr spezialisierten Form kann eine Kontextinformation auch verwendet werden, um den Glättungsparameter eines klassischen MFN vorübergehend anzupassen. Wird von einem System eine eventuell kritische Information an den Nutzer gegeben (in Form von audio-visueller Ausgabe), ist basierend auf dem Kontextwissen über den Inhalt der Nachricht eventuell eine emotionale Reaktion zu erwarten. In diesem Fall kann der Glättungsparameter vorübergehend reduziert werden, um so eine auftretende sprunghafte Änderung der Klassifikatorausgabe zu ermöglichen. Werden im Gegenzug keine emotionalen Reaktionen erwartet, kann ein erhöhter Glättungswert eventuell helfen vorübergehende sensorische Fehleingaben zu kompensieren. Diese Methode wurde im Demonstrator DS3, beschrieben im Abschnitt 5.3.2, erfolgreich eingesetzt.

Sowohl CHMM als auch MFN sind Spezialfälle multimodaler Architekturen. Im folgenden wird ein Ansatz vorgestellt, welcher es erlaubt eine Kontextinformation in jeder allgemeinen Form multimodaler Architekturen einzusetzen. Zunächst muss allerdings geklärt werden, woher eine solche Kontextinformation stammen kann. Eine stets vorhandene Informationsquelle bietet die MMI Anwendung selbst, zusammen mit der Interaktionshistorie des Nutzers. Im Falle der uulmMAC Selektionsanwendung kann dieser Kontext zu jedem Zeitpunkt t definiert werden durch:

$$Konetxt(t) = \alpha Inter(t) + (1 - \alpha) Komplex(t) \tag{4.6}$$

Dabei beschreibt $Inter(t)$ die Kontextkomponenten der Interaktionshistorie, und $Komplex(t)$ die Komplexität der aktuell gestellten Aufgabe. Der Parameter α erlaubt eine Gewichtung der beiden Elemente. Der Kontext der Iteraktionshistorie wird beschrieben durch:

$$Inter(t) = a \sum_{t-T \leq i < t} \frac{t_i}{t_{max}} + (1 - a) \sum_{t-T \leq i < t} \frac{miss_i}{|T|} \tag{4.7}$$

Wobei T die Anzahl der betrachten Elemente der Historie, t_i die benötigte Bearbeitungszeit im Schritt i, t_{max} die maximal erlaubte Bearbeitungszeit und $miss_i$ die Anzahl der falsch erkannten Symbole beschreibt. Mittels des Parameters a werden die beiden bestimmenden Terme gewichtet. Die Historie hängt also nur von der Leitungsfähigkeit des Nutzers ab. Die Komplexität $Komplex(t)$ beschreibt hingegen die systemseitigen Eigenschaften ohne den Nutzer zu betrachten. Sie ist definiert durch:

$$Komplex(t) = b \frac{N_t}{N_{max}} + c \frac{Farb_t}{Farb_{max}} + (1 - b - c) \frac{Form_t}{Form_{max}} \tag{4.8}$$

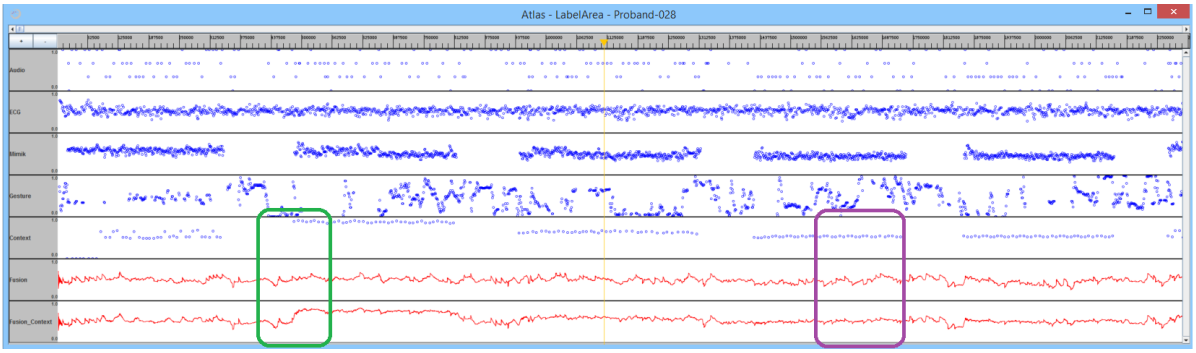

Abbildung 4.16.: ATLAS Screenshot der Auswirkung von situativem Kontext auf die Schätzung von kognitiver Last. Dargestellt sind vier unimodale Eingaben einer MFN Fusion (Audio-, EKG-, Mimik- und gestenbasierter Schätzung). Zeile fünf zeigt den Kontext aus Interaktionshistorie und Anwendungskomplexität. Spur sechs (zweite von Unten) zeigt die Fusion ohne Kontext, die untere Spur zeigt die Fusion mit zusätzlicher Kontexteingabe. Grün markiert ist eine Stelle bei der der Kontext einen Sprung der kognitiven Belastung begünstigt, während die lila markierte Stelle ein Beispiel für die glättende Wirkung von Kontext zeigt.

Dabei beschreiben N_{max}, $Farb_{max}$ und $Form_{max}$ die maximale Anzahl an Elementen, Farben und Formen die in der Anwendung möglich sind, während N_t, $Farb_t$ und $Form_t$ die tatsächlich zum Zeitpunkt t verwendete Anzahl von Elementen Farbe und Formen beschreibt. Die Parameter b und c dienen der Gewichtung. In Abbildung 4.16 ist eine, auf diese Weise erzeugte, Kontextspur für den uulmMAC exemplarisch dargestellt.

Diese Kontextinformation kann nun entweder als zusätzliche sensorische Spur am Eingang der Fusionsarchitektur angelegt werden. Alternativ ist auch eine Veränderung der Klassifikatorschwellen der Ausgangsschicht mittels Kontext möglich. Da beide Varianten sich lediglich in ihrer technischen Umsetzung unterscheiden, methodisch jedoch äquivalent sind, wurde im Rahmen hier entschieden, die Kontextinformation als zusätzlichen Eingang anzulegen. Abbildung 4.16 zeigt alle unimodalen Klassifikatorausgaben der der Fusion vorgeschalteten Schichten, sowie die Kontextspur. Weiter ist ein fusioniertes Klassifikationsresultat mit und ohne Kontext dargestellt. Zu erkennen ist, dass der Kontext schwankende, rauschende und teilweise leicht widersprüchliche Eingaben der unimodalen Schicht glättet (Abbildung 4.16 links). Wird durch unimodale Eingaben hingegen eine Änderung der kognitiven Last bereits schwach angezeigt, und diese durch den Kontext plausibilisiert, begünstigt der Kontext eine sprunghafte Antwort der Ausgabe (Abbildung 4.16 rechts).

KAPITEL 5

TECHNISCHE REALISIERUNG VON COMPANION-SYSTEMEN

In den letzten Jahren ist die Interaktion mit Computersystemen, wie bereits von von Mark Weiser im Jahr 1991 vorausgesagt [271], allgegenwärtig geworden. Menschen sind heute im täglichen Leben von Computern umgeben, sei es in der Form von Smartphones, verschiedenen Komponenten in modernen Autos, Smart-TVs oder Haushaltsgeräten wie Vakuumroboter und interaktiven Kühlschränken. MMI als wissenschaftliche Disziplin beschränkt sich daher nicht nur auf Systeme die von wenigen Experten verwendet werden, sondern umfasst auch Systeme die Einfluss auf fast jeden Menschen in modernen Zivilisationen haben. Die Qualität der MMI ist mitverantwortlich für die Erstellung von nützlichen, universellen und akzeptierten Technologien [242]. Die MMI erfolgt dabei nicht nur auf expliziter Kommandoebene, sondern auch auf der Erkennung impliziter Signale und emotionaler Nutzerzuständen.

Darin wird der Fokus auf die eng mit der MMI verbundenen Aspekte der Individualität gelegt. Ein System kann als nutzerindividuell bezeichnet werden, wenn es die Fähigkeiten, Präferenzen, Anforderungen, sowie die aktuellen Bedürfnisse des Nutzers berücksichtigt. Darüber hinaus kann ein empathisches System mittels Kenntnis der Situation eines Nutzers und dessen emotionalen Zustandes sein Verhalten anpassen [15]. Solche Systeme werden als Companion-Systeme bezeichnet [272] und bestehen aus verschiedenen Komponenten, die sich kontinuierlich an den Nutzungskontext anpassen können [88, 109]. Das Zusammenspiel einer solchen Vielzahl von beteiligten Systemkomponenten, die dabei unter Umständen auch auf verschiedensten Geräten verteilt kooperieren müssen, bedarf dabei einiger Besonderheiten bei der technischen Realisierung.

Dieses Kapitel beschreibt zunächst den Stand der Forschung zur technischen Realisierung von Companion-Systemen aus wissenschaftlicher und industrieller Sicht. Darauf folgend wird die MAR^{2}S Architektur vorgestellt, welche auf den Aufnahmekomponenten des uulmMAC aufbaut und eine Möglichkeit bietet, modulare Architekturen für Companion-Systeme effizient

und dezentral zu realisieren. Abschließend werden einige prototypisch realisierte Companion-Systeme vorgestellt und auf technische Details der Umsetzung eingegangen.

5.1. Stand der Forschung - Technische Realisierung von Companion-Systemen

Um sich an den Nutzer anpassen zu können und mit ihm zu interagieren, muss das System über sensorische Eingänge wie Kameras, Mikrofone oder Verstärker für biophysiologische Daten verfügen. So werden explizite und implizite Eingaben des Nutzers abgetastet. Dabei ist die Fähigkeit, den emotionalen Zustand des Kommunikationspartners zu erkennen für ein technisches System ebenso notwendig, wie in der Mensch-Mensch-Interaktion [52, 176].

In den letzten Jahren wurde umfangreich am komplexen Thema der maschinellen Emotionserkennung geforscht. Wissenschaftliche Methoden sind mittlerweile in der Lage zwischen Emotionen oder verwandten Begriffen wie Stimmung und Gefühl zu unterscheiden (Methoden hierzu sind in Kapitel 4 umfangreich beschrieben). Emotionen wurden von Fridja et. al. als objekt- oder handlungsbezogen identifiziert [84]. emotionen können in der Companion-Technologie als Reaktion auf eine bestimmte Situation gesehen werden, die entweder vom System oder dem umgebenden Kontext ausgelöst wurde [31].

Offen ist nun die Frage wie Systeme technisch realisiert werden sollen, die auf der einen Seite eine Anwendung realisieren, und auf der anderen Seite über umfangreiche sensorische Module verfügen. Eine weitere Problemstellung ist darin gegeben, dass moderne technisch Systeme nicht mehr durch eine zentrale technische Komponente realisiert sind. Vielmehr bestehen moderne Systeme aus dem Zusammenspiel verschiedenster elektronischer und Softwarekomponenten. Demzufolge ist es offensichtlich, dass eine zukünftige effiziente Realisierung, nicht wie bisher als zentrales Softwareelement welches auf einer Hardwareplattform läuft, möglich sein wird. Vielmehr muss über dezentrale modulare Strukturen nachgedacht werden.

Wissenschaftliche Forschung ist jedoch zumeist nicht primär an der konkreten Systemrealisierung bis hin zur Produktebene interessiert. Zumeist werden nicht einmal echte Prototypen realisiert, methodische Forschung endet zumeist bereits nach der Evaluierung eines Algorithmus. Unter echten Prototypen wird in diesem Kontext ein Vorläufer eines späteren Produktes verstanden, der die meisten Funktionen und Eigenschaften des Produktes bereits weitgehend realisiert. Ein Prototyp wird bereits mit einer konkreten Produktvision als Grundlage realisiert. Eine Vorstufe hierzu bietet die Herstellung eines „Pretotyps“. Darunter wird ein System verstanden, welches lediglich die Machbarkeit einer Methode zeigt. Die realisierten Komponenten

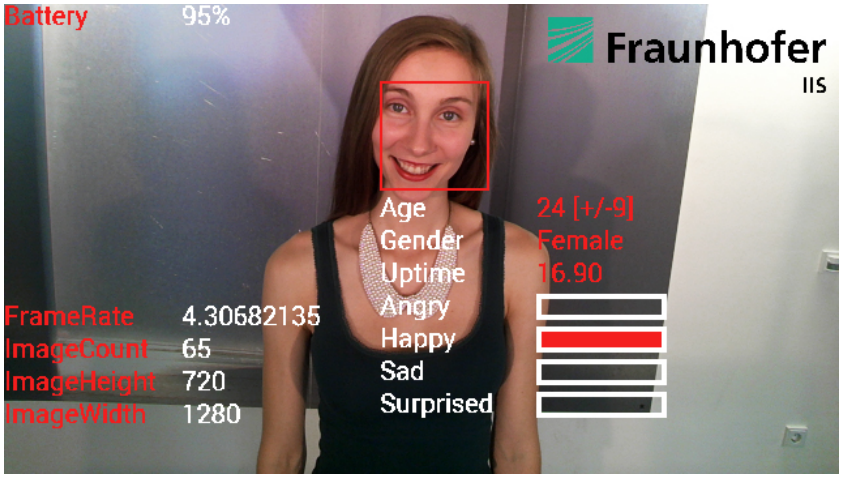

Abbildung 5.1.: Screenshot des Frauenhofer IIS Frameworks zur visuellen Emotionserkennung. Bildquelle: `https://www.iis.fraunhofer.de/de/ff/sse/ils/tech/shore-facedetection.html` (Stand: 3. Juni 2019)

sind dabei auf ein Minimum reduziert welches notwendig ist um die Kernfunktion zu zeigen. Eigenschaften und Beschränkungen eventueller Zielplattformen werden nicht berücksichtigt. Wissenschaftliche Systeme entsprechen somit eher einem Pretotyp. Diese werden zumeist derart realisiert, dass die Funktionalität auf Basis zuvor entwickelten Quellcodes der methodischen Untersuchung realisiert wird. Architekturfragen werden dabei in der Regel vollständig ignoriert. Ziel ist nur die Herstellung eines irgendwie zum laufen gebrachten Demonstrators, egal wie ineffizient und anfällig eine solche Lösung ist. Dabei werden zumeist nur einzelne Komponenten demonstriert, was die Frage nach einer übergeordneten Architektur überflüssig macht [248, 100, 12].

Ein Beispiel hierfür ist das weit entwickelte SHORE System des Fraunhofer IIS auf Basis der Google Glass[1] [193], gezeigt in Abbildung 5.1. Dies bietet eine Software Framework zur Erkennung von Emotionen in Videos, welches echtzeitfähig ist und so zum Beispiel Autisten im Alltag unterstützen kann. Dies erfolgt durch Assistenz bei der Erkennung von Emotionen eines Gesprächspartners. Das System zeigt einen sehr schönen wissenschaftlichen Demonstrator, der es bis zur Prototypenreife geschafft hat. Dies bildet eine seltene Ausnahme im wissenschaftlichen Kontext. Die eingesetzte Architektur besteht allerdings aus einem zentral implementierten Softwaremodul.

Eine Ausnahme zu den als Zentralmodul realisierten Ansätzen bildet lediglich das Social Signal Interpretation (SSI) Framework [268][2] aus Abbildung 5.2. Dieses realisiert mehrere Komponenten, die dezentral via Netzwerk kommunizieren können. Wie der Name bereits suggeriert,

[1] `https://www.iis.fraunhofer.de/de/ff/sse/ils/tech/shore-facedetection.html` (Stand: 3. Juni 2019)
[2] `https://hcm-lab.de/projects/ssi/` (Stand: 3. Juni 2019)

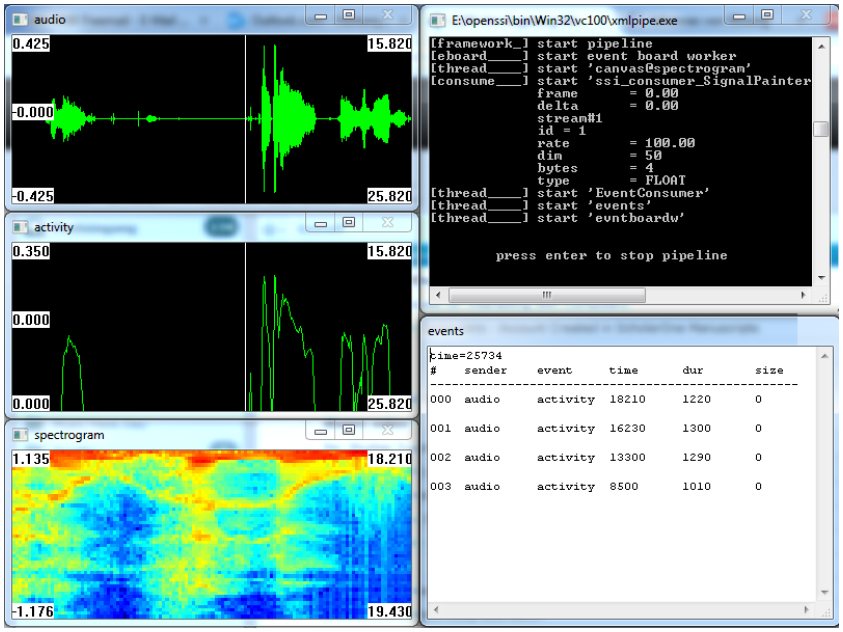

Abbildung 5.2.: Screenshot des SSI Frameworks zur Emotionserkennung am Beispiel einer akustischen Analyse.

ist die Zielsetzung dieses Frameworks nicht darin gegeben ein Companion-System zu realisieren. Vielmehr wird mittels SSI eine modulare Pipelinearchitektur zur Erfassung, Aufnahme, Verarbeitung und Klassifikation von affektiven Datenströmen realisiert. Komponenten, die durch eine Pipeline, multimodal erkannte emotionale Zustände in einem Companion-System weiterverarbeiten, sind nicht vorgesehen.

Industrielle Herangehensweisen unterscheiden sich von wissenschaftlichen Methoden dahingehend, dass zumeist bereits in sehr frühen Entwicklungsphasen Produkte oder Produktvisionen im Fokus stehen. Dies beeinflusst die technische Entwicklung erheblich dahingehend, dass nicht jede, zur Realisierung eingesetzte, Methode bis ins letzte Detail erforscht wird. Vielmehr wird eine Methode dann als einsetzbar angesehen, wenn diese entsprechende Software- und empirische Tests bestanden hat. Im Fokus industrieller Entwicklung steht viel mehr die eingesetzte Software Architektur mit Fokus auf Modularisierbarkeit, Wiederverwendbarkeit und Einsetzbarkeit in Software Entwicklungsprozessen [28, 182]. Leider werden diese technischen Informationen von Firmen typischerweise geheim gehalten, da hier eine wesentliche Komponente des Firmen Know-hows zu suchen ist. Die folgend exemplarisch vorgestellten kommerziellen Produkte erfüllen einige, jedoch bei weitem nicht alle, Companion-Eigenschaften.

Kleinere Systeme wie die smarte Pillendose von *AdhereTech*[3] sind in der Lage die Pilleneinnahme seiner Nutzer zu überwachen. Dabei können zwar einige adaptive Anpassungen an den

[3] `https://www.adheretech.com/` (Stand: 3. Juni 2019)

Nutzer erfolgen, die Funktionalität bleibt aber auf diese eine Aufgabe beschränkt. Auch die Realisierung ist wenig modular. Dieses Produkt erscheint somit durchaus smart, realisiert aber keine Komponente der Companion-Technologie.

Ein weiteres einfaches Produkt ist die Gedächtnishilfe von *Cognitive Prosthetics*[4]. Diese unterstützt Patienten welche unter Gedächtnisproblemen leiden (z.b. durch Gehirnverletzungen oder Alzheimer). Das Produkt ist als App realisiert und bietet individualisierte Erinnerungshilfen im Alltag an. Es kann durch langsames Training zur Verbesserung der Gedächtnisleistung beitragen. Auch hier wirkt die Anwendung smart, ist aber auf einen Spezialfall in ihrer Anwendbarkeit begrenzt. Eine modulare Technologie ist innerhalb der Realisierung der App nicht zu vermuten.

Wesentlich höher entwickelt ist das *Google Now* bzw. *Google Assistent* System[5] und das *Amazon Echo* System *Alexa*[6]. Beide realisieren einen persönlichen Assistenten, der nach Aussage der Firmen „ die richtige Information zur richtigen Zeit bereitstellen“soll. Der Nutzen der Anwendung besteht in der Antizipation und Aufbereitung für den Nutzer relevanter Funktionen (Navigation, Termine, Nachrichten, Wetter, Smarthome[7], ...). Die technische Realisierung wurde mittels eines elektronischen Dienstes realisiert, der auf einem einzelnen Device (Alexa) oder einem Smartphone (Google) realisiert ist. Beide Systeme erfüllen jedoch nicht den vollständigen Funktionsumfang eines Companion-Systemes. Zwar kann das System aufgrund einer modularen Architektur durch Kombination mit Modulen anderer Hersteller erweitert werden. Insbesondere die Adaptivität ist dabei aber in vielen Bereichen nur durch explizite Konfiguration durch den Nutzer realisiert. Eine Erkennung und Berücksichtigung emotionaler Zustände vorgesehen im Systementwurf nicht vorgesehen. Bei genauerer Betrachtung ist erkennbar, dass auch hier die Companion-Eigenschaften nur in geringem Umfang realisiert wurden. Auch die Modularität ist nur in geringem Umfang nach außen geöffnet. Systemerweiterungen sind nur in Zusammenarbeit mit dem Hersteller unter großem Aufwand möglich. Es ist allerdings zu vermuten das die interne Realisierung auf einer vollständig modularen Architektur aufbaut. Für beide Systeme ist allerdings mit einer rasanten technischen Weiterentwicklung, hin zu umfangreichen Companion-Systemen zu rechnen.

Nach bestem Wissen gibt es derzeit kein System, das die folgenden Beiträge in Kombination voll erfüllt:

- Ein System das Emotionen mit multimodalen Sensoren erfasst.

[4] `http://www.beyondliteracy.com/cognitive-prosthetics/` (Stand: 3. Juni 2019)
[5] `https://www.google.com/intl/de/landing/now/` (Stand: 3. Juni 2019)
[6] `https://www.amazon.de/Amazon-Echo-2nd-Generation-Anthrazit-Stoff-/dp/B06ZXQV6P8` (Stand: 3. Juni 2019)
[7] `https://de.wikipedia.org/wiki/Smart_Home` (Stand: 3. Juni 2019)

- Ein System das die Möglichkeit bietet, die emotionsauslösende Ursache zu erfassen, mit der Absicht, Veränderung des emotionalen Nutzerzustandes herbeizuführen.
- Ein System das adaptiv und modular aufgebaut ist.
- Ein System das adäquat im ursächlichen Modul seiner Architektur mit einer Anpassung reagiert.
- Ein System das seinen umgebenden Kontext erfasst und berücksichtigt.

Ein Companion-System ist somit bis heute nicht existent, wenngleich einige industrielle Produkte dies auf den ersten Blick durchaus vermuten lassen.

5.2. Modulare dezentrale Architektur, MAR^{2}S

Companion-Systeme sind hoch komplexe technische Systeme. Bei ihrer Realisierung sind eine Vielzahl von Fachdisziplinen beteiligt [27]. Dabei interagieren verschiedenste methodische Verfahren miteinander. Eine solch komplexe Struktur kann nicht in Form einer einzigen enorm umfangreichen Komponente realisiert werden, vielmehr ist es sinnvoll eine modulare Architektur zu verwenden. Dies ermöglicht es nahezu beliebig große und inhomogene Gruppen von Entwicklern, Ingenieuren und Fachleuten zur Systemrealisierung einzusetzen. Komponenten können so zunächst in kleinen Gruppen realisiert werden und anschließend durch entsprechende Integration zu immer größeren und komplexeren Gesamtsystemen zusammengefasst werden.

Die hier vorgeschlagene MAR^{2}S Architektur erfüllt alle Eigenschaften, wie sie für einen modernen Systementwurf notwendig sind. Bereits in Abschnitt 2.3.3 wurde ein erster Einblick in den modularen Aufbau der MAR^{2}S Architektur gegeben. Darin wurden exemplarische Implementierungen am Spezialfall von Sensordatenaufzeichnungsmodulen vorgestellt. Diese beschreiben jedoch nur einen Teil der Möglichkeiten der MAR^{2}S Architektur. Im folgenden werden weitere Eigenschaften beschrieben wie sie für einen modernen Systementwurf notwendig sind.

5.2.1. Modularer Aufbau und Kapselung der Funktionalitäten

MAR^{2}S sieht vor, jede elementare Funktionalität in einem separaten Modul zu realisieren. Der interne Aufbau und die technische Realisierung des Moduls sind dabei nach außen für andere Module nicht sichtbar. Der gesamte Aufbau erinnert dabei an moderne objektorientierte Pro-

grammierung [187, 56]. Module übernehmen dabei die Rolle der Klassen, Modulschnittstellen die Rolle öffentlicher Methoden.

Jedes Modul definiert eine Menge von Schnittstellen die eine Kommunikation nach außen realisieren. Dies Schnittstellen beschreiben eingangsseitig die Anforderungen, welche ein Modul an das Gesamtsystem stellt um seine Funktionalität bereitstellen zu können. Diese Anforderungen bestehen zum einen in obligatorischen Anforderungen, zum anderen in optionalen Anforderungen. Ohne Erfüllung der obligatorischen Anforderungen kann ein Modul seine Aufgabe nicht erfüllen, optionale Anforderungen können beitragen dass ein Modul effizienter läuft oder genauere Ergebnisse liefern kann. Ausgangseitig definiert eine Schnittstelle welche Ausgaben ein Modul erzeugen kann und in welcher Form diese bereitgestellt werden. Auch hier können obligatorische und optionale Schnittstellen unterschieden werden.

Systeme werden wie in Abbildung 5.6, 2.10 und 5.10 anhand einer konkreten Implementierung gezeigt, durch einen Vielzahl von Modulen realisiert. Diese realisieren dabei unterschiedlichste Elemente des Gesamtsystems. Die folgende Liste enthält eine Menge von Grundmodulen:

- Sensorische Aufnahmemodule (Zeichnen Sensordaten auf, Siehe Abschnitt 2.3.3)
- Sensorische Erfassungsmodule (Kommunizieren mit Sensoren und stellen Datenströme bereit)
- Merkmalextrahierende Module (Erzeugen Merkmalsvektoren aus Sensordatenströmen)
- Klassifikator Module (Klassifizieren Merkmale)
- Fusionsmodule (Fassen explizite und/oder implizite Eingaben zusammen)
- Anwendungsmodule (Realisieren Funktionalitäten konkreter Anwendungen)
- Dialogmodule (Bewerten und optimieren den Dialogverlauf)
- Wissensmodule (Realisieren die Funktionalität der Wissensbasis, evtl. inkl. Wissensakquise)
- Fissionsmodule (Bewerten alle denkbaren Systemausgabegeräte)
- Ausgabemodule (Realisieren konkrete Systemausgaben auf Geräten)

Dies Liste erhebt keinen Anspruch auf Vollständigkeit. In konkreten Umsetzungen können durchaus weitere Module notwendig werden. Insbesondere die Anwendung kann abhängig von ihrer Komplexität eine enorme Anzahl weiter Module erfordern.

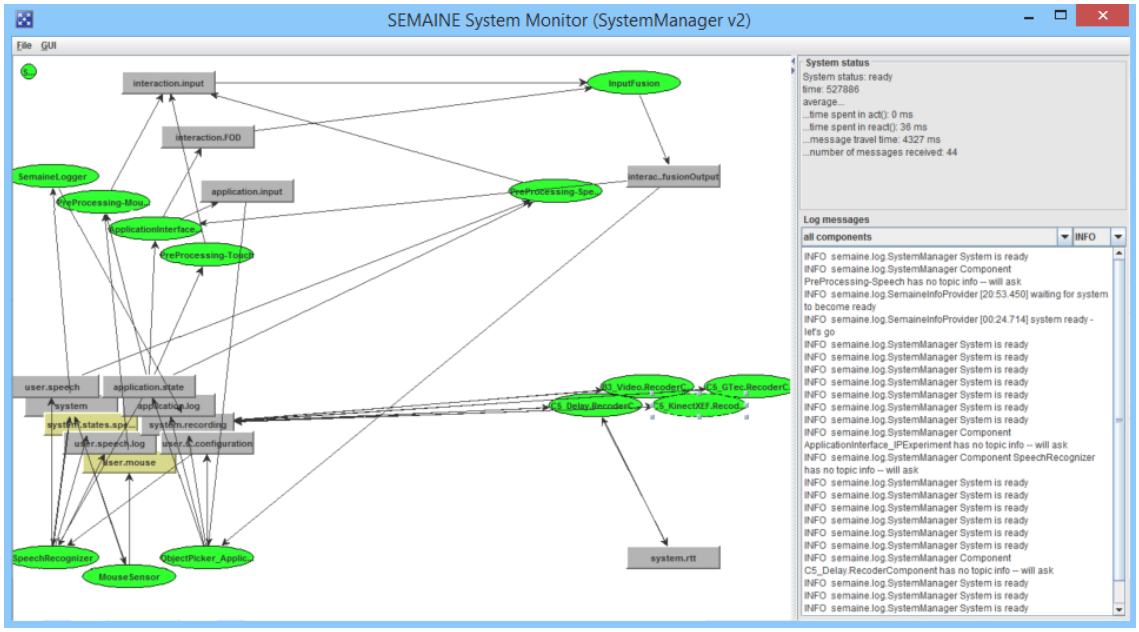

Abbildung 5.3.: Darstellung aller für die Realisierung des uulmMAC Experimentes instanziierter Module (grün) und Nachrichtenpfade (grau) in der Kontrollansicht des Middleware Systemmanagers.

5.2.2. Netzwerkbasierte Modulinteraktion

Die Kommunikation der MAR^{2}S Architektur Module erfolgt über Netzwerkschnittstellen. Die an den Schnittstellen übertragenen Informationen werden in zuvor gemeinsam definierten XML Schemata verpackt und via XML-String versendet. XML hat hier den Vorteil, dass auf einfach Weise auch komplexe Strukturen (Bäume, Objekte und Objektgraphen) verschickt werden können. Zudem hat eine XML Kommunikation den Vorteil, dass unterschiedliche Interpretationen an beiden Enden der Kommunikationsschnittstelle vermieden werden[8]. Die XML Kommunikation ist dabei jedoch nicht zwingend, in begründeten Fällen kann jede beliebige andere binäre Struktur definiert werden. Dies kann bei hohen Anforderungen an den Datendurchsatz sinnvoll sein, da XML Strings zugunsten ihrer Interpretierbarkeit einen nicht unerheblichen Datenoverhead mit sich bringen.

Um die Vielzahl von gleichzeitigen Netzwerk Kommunikationen unterscheiden und koordinieren zu können, wird das durch die SEMAINE Architektur[9] verwendete System der Nachrichten *Topics* zur Verwendung vorgeschlagen [217]. Darin kommunizieren alle Module ausschließlich über Topics. Jedes am System beteiligte Modul definiert dabei Topics auf denen es empfangen möchte und Topics auf denen es senden wird. Die SEMAINE Middleware übernimmt

[8] Beispielsweise ist die Interpretation binärer Zahlendarstellung in diversen Umgebungen und Programmiersprachen unterschiedlich. (Big vs. Little Endian, Signed vs. Unsigned Darstellung, etc.)

[9] `https://github.com/SEMAINE` (Stand: 3. Juni 2019)

dabei die Koordination und den Transport der Nachrichten. Es existiert dabei keine Beschränkung der Anzahl von Sendern und Empfängern auf einem Topic oder der Anzahl der Topics selber. Bildlich gesprochen kann ein Topic als Diskussionsgruppe in einem Chatraum oder ähnlich einer Mailing-Liste gesehen werden, wobei Module mehreren Chaträumen (Topics) oder Mailing-Listen angehören können. Die Middleware fungiert dabei als Chatraum Betreiber und Administrator. In dieser Arbeit wurden nahezu alle vorgestellten Komponenten auf Basis der SEMAINE Middleware realisiert. Das grundsätzliche MAR^{2}S Konzept ist jedoch nicht auf diese konkrete Middleware beschränkt.

Durch die Verwendung einer Middleware ist eine direkte Kopplung der Module mittels IP Adressen und Ports nicht notwendig. Module können zur Laufzeit in den Komponentenverbund aufgenommen werden und diesen verlassen. Dem Modul muss dazu keine der anderen Modulinstanzen explizit bekannt sein. Dies ist insbesondere bei der Verwendung dezentraler Hardware hilfreich. In einem Szenario, an dem mehrere Geräte beteiligt sind (Smartphones, Tablets, PCs, Wearables[10], etc.), entstehen so keine zusätzlichen Probleme durch einen dynamisch adaptiven Geräteverbund. Es muss lediglich sichergestellt sein, dass zu jeder Zeit die zur Grundfunktionalität notwendigen Module verfügbar sind. Kann dies nicht sichergestellt werden, müssen Module über Ausnahmebehandlungen verfügen um Systemabstürze durch erreichen ungültiger Systemzustände zu verhindern.

Abbildung 5.3 zeigt eine exemplarische Ansicht der uulmMAC Module inklusive aller Anwendungskomponenten die zur Realisierung des interaktiven Spiels verwendet wurden, sowie den Ausgabe Modulen zur Nutzerinteraktion.

5.2.3. Organisation der Module

Die zugrundeliegende Netzwerktechnologie ist durch die MAR^{2}S Architektur nicht vorgegeben oder beschränkt. Auch ist es nicht Gegenstand dieser Arbeit diese untergeordnete technische Schicht der Netzwerkfunktionalitäten zu untersuchen. Da die MAR^{2}S Architektur allerdings direkt auf Netzwerkinfrastrukturen aufbaut, sei dennoch an dieser Stelle ein kurzer Ausflug gestattet, der die wesentlichen Grundlagen und ihrer Unterscheide beschreibt.

In Netzwerken kann grundsätzlich zwischen serverbasierter und Peer to Peer (P2P) Kommunikation unterschieden werden [46, 216, 4] (Siehe Abbildung 5.4). Bei einem serverbasierten Ansatz ist jedem Modul (Client) die Netzwerkadresse eines zentralen Servers bekannt der jederzeit erreichbar ist. Möchte ein Modul eine Kommunikation eröffnen meldet es sich am Server an, da es weiß wie es diesen erreichen kann. Der Server verwaltet alle seine registrier-

[10] `https://de.wikipedia.org/wiki/Wearable_Computing` (Stand: 3. Juni 2019)

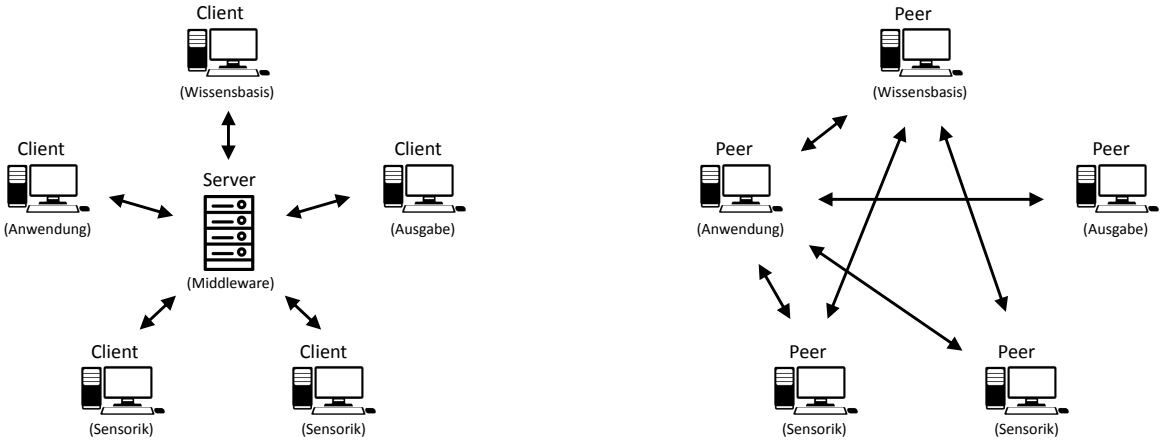

Abbildung 5.4.: Vergleich der zugrundeliegenden Netzwerkstrukturen. Links: Client-Server Modell. Rechts: Peer To Peer Modell. Das Peer To Peer Modell benötigt keinen zentralen Server, ist in der Handhabung allerdings aufgrund der fehlenden zentralen Steuerinstanz kritischer. Die in der Arbeit vorgestellten Prototypen verwenden den Client-Server Ansatz, sind jedoch nicht darauf beschränkt. In Klammern sind mögliche Module, welche auf den Clients laufen, angegeben.

ten Module (Klienten) und kommuniziert exklusiv mit diesen. Möchte ein Modul mit einem anderen kommunizieren übermittelt es alle seine Nachrichten an den Server, dieser leitet die Nachricht an alle Adressaten weiter. Ein Modul kennt dabei die Identität der Module (Instanzen) mit denen es indirekt kommuniziert nicht zwangsläufig. Vorteilhaft an dieser Architektur ist die relativ einfache Realisierung, Nachteile bestehen allerdings im Server als Flaschenhals. Fällt der Server aus oder reichen seine Ressourcen nicht aus um den Nachrichtenverkehr zu bearbeiten, fällt das gesamte System aus, obwohl alle Module intakt sind. Die SEMAINE verwendet dieses serverbasierte Konzept. Der *Systemmanager* realisiert den zentralen Nachrichtenserver. Allen Modulen wird bei der Initatlisierung die Serveradresse, der Serverport und die relevanten Topics mittels Konfigurationsdatei mitgegeben. Da die, in dieser Arbeit exemplarisch umgesetzten Instanzen der abstrakten MAR^{2}S Architektur auf Basis der SEMAINE Middleware kommunizieren, wurde zwangsläufig auch das Client Server Prinzip übernommen.

Einen gänzlich anderen Ansatz verfolgt eine P2P Netzwerk Struktur [216]. In einem P2P-Netz sind alle Komponenten gleichberechtigt und können sowohl Funktionalität in Anspruch nehmen, als auch zur Verfügung stellen. Durch die fehlende zentralsteuernde Instanz (Server) ist die praktische Realisierung ungleich komplexer und fehleranfälliger. So weis ein Modul bei seiner Initialisierung zunächst nicht, an welche Stelle es sich wenden soll, um Informationen über andere Module und Topics zu bekommen. Mit der Suchfunktion (*lookup*) können Module (*Peers*) im Netzwerk diejenigen Module identifizieren, die für eine bestimmte Objektkennung zuständig sind. Kann mittels Suchfunktion mindestens ein weiteres Modul gefunden werden,

werden von diesem Modul seine Informationen über den aktuellen Netzwerkverbund und die Identität andere Modulinstanzen geteilt (Diese Informationen können dabei falsch oder veraltet sein). Mittels der Suchfunktion können zudem weitere Peers im Netzwerk gesucht werden, die gewisse Kriterien erfüllen (z. B. bestimmte Topics bereitstellen). Grundsätzlich scheinen solche P2P Netzwerke für dezentral organisierte Companion-Systeme sehr geeignet zu sein. Bisher wurden sie allerdings aufgrund ihrer schwierigen Beherrschbarkeit noch nicht eingesetzt. Problematisch ist in einem solchen Netzwerk neben der Tatsache, dass Module sich direkt sehen müssen, im wesentlichen die Herausforderung, dass alle Module eine einheitlich Sicht auf den gesamten Modulverbund haben sollten, was nicht zu jeder Zeit sichergestellt werden kann. Unter bestimmten Voraussetzungen ist es sogar denkbar, dass P2P Netzwerke in zwei Teilnetze zerfallen und nicht mehr untereinander kommunizieren können [216].

Eine weitere Technologie besteht im Einsatz von sogenannten selbstorganisierenden Netzwerken (SON) [2, 149, 112, 257]. Diese sind bisher aber praktisch kaum erprobt, was es schwierig macht diese im Bezug auf ihrer Einsetzbarkeit in Companion-Architekturen zu bewerten. Die grundsätzliche Idee steht in keinem Konflikt mit den MAR^{2}S Ansätzen und liese sich kombinieren.

5.2.4. Systemrealisierung unter Verwendung von Softwaretechniken

Alle größeren Systementwicklungen verwenden Bausteine der Softwaretechnik (engl. System Engineering) um eine Realisierung überhaupt erst zu ermöglichen. Softwaretechnik beschäftigt sich mit der Herstellung oder Entwicklung von Software, der Organisation und Modellierung der zugehörigen Datenstrukturen und dem Betrieb von Softwaresystemen. Eine Definition von Balzert beschreibt das Gebiet als „*Zielorientierte Bereitstellung und systematische Verwendung von Prinzipien, Methoden und Werkzeugen für die arbeitsteilige, ingenieurmäßige Entwicklung und Anwendung von umfangreichen Softwaresystemen.*“ [13] Die Definition von Softwaretechnik lässt sich problemlos auch auf die Erstellung von Softwaresystemen mit Hardwarekomponenten erweitern.

Softwaretechnik umfasst dabei eine Vielzahl von Teilgebieten, welche die Softwareentwicklung begleiten [182, 13]. Dazu zählen: Projektmanagement, Qualitätsmanagement, Risikomanagement, Anforderungsmanagement, Systemdesign, Implementierung, Softwaretest und Inbetriebnahme. Die MAR^{2}S Architektur ermöglicht dabei insbesondere den Einsatz in den Gebieten Systemdesign, Implementierung und Softwaretest.

In der Systemdesignphase wird eine typische *top down*[11] Zerlegung unterstützt. So können zunächst Modulgruppen weitgehend abstrakt definiert werden, welche dann, in Form von (teilweise abstrakten) Modelldefinitionen verfeinert werden können, und abschließend in konkreten externen und schließlich internen Modellbeschreibungen münden. Während der Implementierungsphase wird dann ein *bottom up* Ansatz unterstützt. Dabei werden zunächst einzelne Funktionen implementiert, diese werde zu Modulen zusammengefasst, Module wiederum bilden Modulgruppen, welche anschließend wiederum zu einem Gesamtsystem integriert werden können.

Auch der Softwaretest [13] wird in Form von Whitebox-, Blackbox- und Modul- und Integrationstest voll unterstützt. Die modulare Eigenschaft erlaubt es, dass sich verschiedenste Module in unterschiedlichen Entwicklungsphasen befinden können. Soll zum Beispiel ein Modul getestet werden während seine Schnittstellen Partnermodule noch nicht, oder nur teilweise realisiert sind, ist es möglich Dummymodule einzubinden, welche die Schnittstellendefinition erfüllen ohne dabei selbst voll umfänglich zu funktionieren. Im Falle von Integrations- und Systemtests können zusätzlich weitere Module eingebunden werden, welche zum Beispiel Systemzustände protokollieren[12] oder Datendurchsätze messen.

Diese Eigenschaften erlauben den Einsatz von MAR^{2}S in allen gängigen Vorgehensmodellen [13] der Softwareentwicklung wie dem Wasserfallmodell [161], dem V-Modell [34, 35] (gezeigt in Abbildung 5.5), dem Extremeprogramming [161] und Prototyping [181] (eingesetzt in der Demonstrator Entwicklung des SFB-TRR 62) oder dem SCRUM [225]. Insbesondere die Einsatzfähigkeit im V-Modell und V-Modell XT ist dabei hervorzuheben, da dieses Vorgehensmodell sehr weit verbreitet und in hohem Maße standardisiert ist. Es stellt einen defacto Industriestandard aller größeren Entwicklungshäuser dar, und ist zudem IT-Entwicklungsstandard der öffentlichen Hand in Deutschland und Modell der deutschen Bundesverwaltung

5.3. Prototypische Realisierung von Companion-Systemen

Die folgenden drei Abschnitte beschreiben zwei realisierte Architekturen und Anwendungen von Companion-Systemen und ein vollständig spezifiziertes Anwendungsszenario eines hypothetischen Companion-Systemes. Die erste Anwendung beschreibt eine Vorstufe des MAR^{2}S mit rein sensorischen Modulen. Die zweite Anwendung beschreibt einen Demonstrator, welcher im Rahmen des SFB-TRR 62 realisiert wurde. Dabei wurde eine vollständige Companion-

[11] `https://de.wikipedia.org/wiki/Top-down_und_Bottom-up` (Stand: 3. Juni 2019)
[12] Siehe SEMAINE-logger Modul von Honold und Schüssel

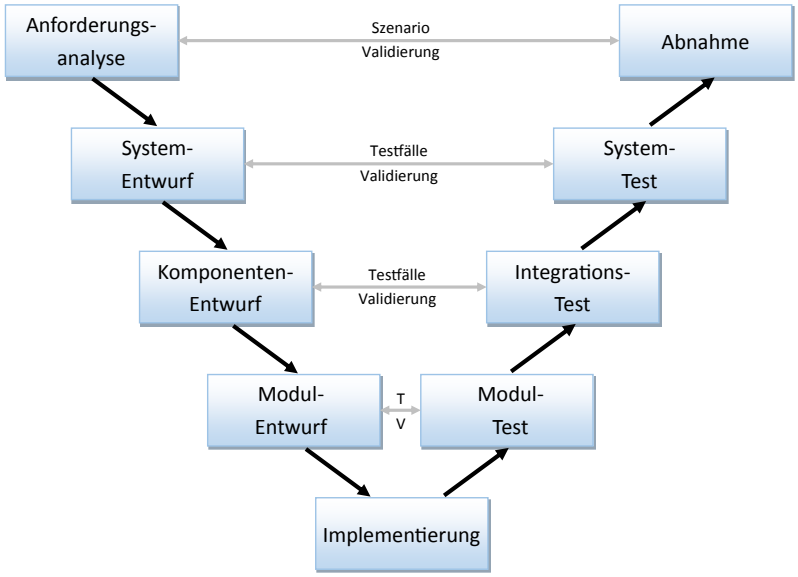

Abbildung 5.5.: Das V-Modell des Systementwurfs beschreibt die top down Zerlegung eines Problems, bis hin zur Implementierung, gefolgt von einem bottom up Ansatz der Integration. Jede Ebene definiert dabei während des Entwurfs (linke Seite) Testfälle, welche in der Test- und Integrationsphase (rechte Seite) validiert werden.

Anwendung, inklusive Nutzerinteraktion, unter Einsatz eines dezentralen modularen Systems wie der MAR^{2}S Architektur, entwickelt. Die letzte Anwendung beschreibt ein mögliches Szenario, welches unter Einsatz der MAR^{2}S realisierbar wäre, und dabei alle Aspekte der Companion-Technologie berücksichtigt.

5.3.1. Verteiltes Nutzertracking und dezentrale Nutzeridentifikation

Eine erste Erprobung des Einsatzes einer verteilten Architektur fand als Vorstudie zur Entwicklung der MAR^{2}S statt. Dabei wurde als exemplarische Anwendung das Tracking und die Identifikation mehrerer Nutzer mittels Multi-Kamera-System gewählt. Dabei ist jede eingesetzte Kamera an jeweils einem PC angeschlossen[13], die Auswertung der Bilder erfolgt dabei zunächst lokal für jede Kamera. Mittels Netzwerkkommunikation werden Zwischenergebnisse in Form von erkannten Personen und ihren Bildpositionen zu einem zentralen Modul gesendet. Dort findet die Fusion der Einzelbilddaten zu einer Gesamtschätzung der Identität und der aktuellen Nutzerposition statt. Diese kann anschließend wiederum mittels einer weiteren Netzwerkkommunikation an weitere dezentrale Komponenten der Companion-Technologie

[13] Die Einsatzmöglichkeit ist nicht auf PCs als Hardwarebasis begrenzt, auch smarte Geräte wie Mobiltelefone, Tablets und Laptops mit eingebauten oder externen Kameras sind möglich

gesendet werden. Die Resultate und Methoden des im folgenden vorgestellten Identifikations- und Trackingmoduls wurden in [152] veröffentlicht.

Modulare Architektur

Der grundsätzliche Ablauf der Nutzeridentifikation und Nutzerpositionsschätzung ist in Abbildung 5.6 dargestellt. Zunächst werden die Bilder aller Kameras separat betrachtet, in der konkreten Umsetzung werden vier Webcams C_1 bis C_4 eingesetzt. Auf jedem Einzelbild wird eine Adaboost-Kaskade, beschrieben von Viola und Jones, zur Gesichtsregionserkennung eingesetzt [266]. Da diese Kaskade zwar alle Gesichter zuverlässig findet, jedoch eine hohe falsch positive Rate aufweist, werden alle gefunden Regionen mit einem Farbmodell verglichen um sicherzustellen, dass ein ausreichender Anteil hautfarbene Fläche enthalten ist.

Die Identifikation der Nutzer erfolgt zunächst auf allen Einzelbildern unter Einsatz der Eigenvektor (EV) Methode [278] und Orientierungshistogramm Methode [61]. Dabei werden alle, als Merkmal extrahierten EV und Orientierungshistogramme, mittels k Nächste Nachbarn (kNN) und SVM klassifiziert. Die Klassifikationsergebnisse und Koordinaten der Gewichtsmittelpunkte werden anschließend mittels Inter-Modul-Kommunikation an die temporalen Fussionsmodule übergeben. Dort werden die letzten t Klassifikationen mittels gewichtetem Voting bewertet, die Koordinaten werden mittels gleitendem Durchschnitt geglättet. Die Fusionsmodule können dabei bereits auf einem anderen PC ausgeführt werden[14]. Das Modul der spatialen Fusion erhält alle 40 Millisekunden eine Schätzung von jedem temporalen Modul, sofern darin Schätzungen berechnet werden konnten. Das spatiale Fusionsmodul kombiniert alle temporal fusionierten Gesichtsinformationen der Sensoreingänge. Dabei werden die Gesichtsregionen mittels Epipolargeometrie in den dreidimensionalen Raum abgebildet und dort zugeordnet. Die Positionsschätzung erfolgt durch Mittelung aller Schnittpunkte der Epipolarlinien. Die Identitätsschätzung durch gewichtetes Voting aller räumlich einander zugeordneten Identitätsschätzungen. Das spatiale Fusionsmodul kommuniziert dabei eingangsseitig mit allen temporalen Modulen, ausgangsseitig mit dem Companion-Technologie Schnittstellenmodul jeweils mittels Netzwerkkommunikation. Im letzten Modul fließt zusätzlich noch der akustische Kanal ein. Dieser bewertet im ersten Modul anhand eines 40 Millisekunden Fensters ob ein Nutzer spricht und sofern Sprache erkannt wurde, wird die Identität des sprechenden Nutzers geschätzt. Diese wird anschließend ebenfalls temporal fusioniert und via Netzwerk an das Ausgansschnittstellen Modul übermittelt. Dieses Ausgabemodul kombiniert die Positionsschätzungen und Identitätsschätzung (Audio und Video) aller erkannten Nutzer zu einer finalen Nachricht, welche wiederum via Netzwerk an die Anwendungsmodule einer

[14] In der prototypischen Umsetzung wurden die Fusionsmodule auf demselben PC ausgeführt wie die sensorischen Eingangsmodule da ausreichend Ressourcen zur Verfügung standen.

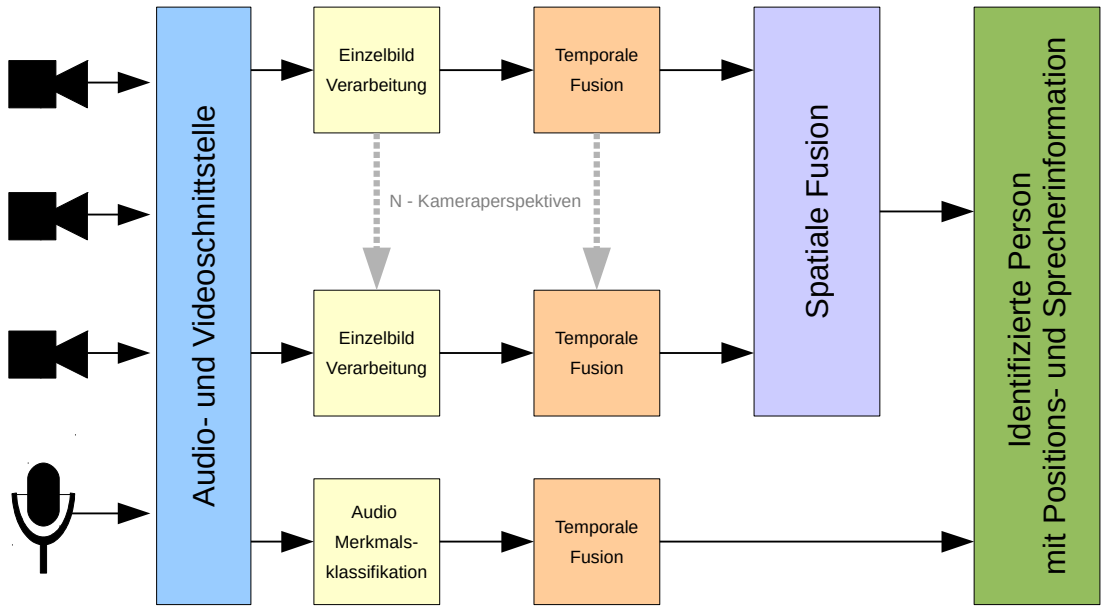

Abbildung 5.6.: Zusammenhang der zur Identifikation und Positionsschätzung realisierten Module.

Companion-Architektur übermittelt werden können. Die Positionsschätzung wird dabei um einen Dispertionswert ergänzt, welcher die Güte der Positionsschätzung beschreibt (Umkreis aller am Mittelwert beteiligten Positionen).

Abbildung 5.7 zeigt die Oberfläche des Systems. Links sind alle vier Eingansbilder und die daraus extrahierten zweidimensionalen Informationen gezeigt. Rechts oben eine Projektion der Draufsicht der Systemumgebung. Rechst unten Details zum Systemstatus.

Technisches Setup und Datenaquise

Der aufgenommene Referenzdatensatz zur Evaluierung der Systemperformance besteht aus 6 Personen. Dabei wurden pro Person 45 Bilder mit den Systemkameras aufgenommen, davon 15 in frontaler Ansicht und jeweils 15 Bilder im rechten und linken Profil. Das Kammerasetup besteht aus vier Logitech Pro 9000 Webcams, mit einer Auflösung von 1280 x 720 bei maximal 30 Bildern pro Sekunde in 3x8Bit Farbtiefe. Das Koordinatensystem von Kamera 1 definiert den Ursprung des Weltkoordinatensystems.

Zur Evaluation wurde zunächst je eine Minute Videomaterial pro Proband an jeweils drei fest definierten Koordinaten aufgenommen. Um auch dynamische Bewertungen zu ermöglichen, sind anschließen pro Proband drei festgelegte Trajektorien abgelaufen worden. In Summe wurden so sechs Minuten Videomaterial pro Proband, also insgesamt 36 Minuten aufgenommen. Der Hintergrund der Testaufnahmen besteht in normaler Büroumgebung, es wurde nicht

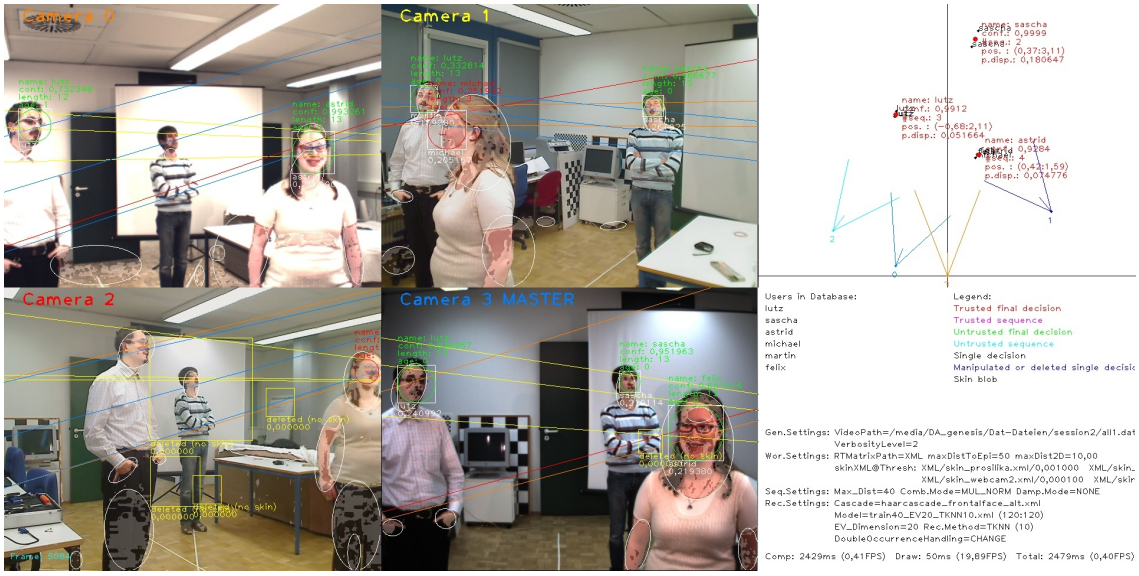

Abbildung 5.7.: Screenshot des Identifikations- und Trackingsystems. Links sind die vier Kameraperspektiven und die daraus extrahierten Informationen dargestellt. Diese sind erkannte Gesichter, Epipolarlinien und hautfarbene Bereiche. Rechts oben ist eine 2D Kartenprojektion der dreidimensionalen Szene gezeigt. Systeminformationen sind rechts unten abgebildet.

auf homogene Hintergründe zurückgegriffen, um die reale Einsatzfähigkeit zeigen zu können. Die Daten wurden unter Einsatz von ATLAS annotiert. Zur Bewertung der Laufzeitgeschwindigkeit und Echtzeitfähigkeit wurde ein Desktop PC (Intel Core i7 960 at 3.2GHz, 12GB Ram, Nvidia GeForce GTX 260, Ubuntu 10.04) verwendet. Das Gesamtsystem erreichte eine durchschnittliche Framerate von 10.3fps, was ausreichend ist um alle Nutzer unter realen Interaktionsbedingungen zu verfolgen und zu identifizieren.

Resultate der Gesichtserkennung

Die Bewertung der Gesichtssegmentierung erfolgt anhand 560 zufällig gewählter Bilder des gesamten Datensatzes. Darauf wurde der Segmentierungsalgorithmus angewendet und dessen Resultat manuell bewertet (*korrekte Segmentierung*, *Fehlalarm* und *übersehenes Gesicht*).

Die Gesichtserkennung erreicht eine erwartet gute Erkennungsrate von 97,4% , dabei jedoch nur eine Relevanz von 59,2%. Diese niedrige Relvanz resultiert aus 245 Fehlalarmen der Gesichtserkennung. Nach Erweiterung durch einen Hautfarbenklassifikator sinkt die Erkennungsrate nur geringfügig auf 95,2% ab. Die Relevanz erhöht sich allerdings signifikant auf 90,8%.

Für jedes segmentierte Gesicht wurden Eigenwert (EW) der Dimension 20 als Merkmal extrahiert. Weiter wurden zum Vergleich Orientierungshistogramme mit 3x3 Teilbildern und einer

Merkmal	Klassifikator	Fusion	Accuracy
EV dim=20	SVM	Einzelbild	81.3 %
EV dim=20	kNN	Einzelbild	78.2 %
OH 16-3	SVM	Einzelbild	73.2 %
OH 16-3	kNN	Einzelbild	87.0 %
EV dim=20	SVM	Temporal (50 Bilder)	93.1 %
EV dim=20	kNN	Temporal (50 Bilder)	89.8 %
OH 16-3	SVM	Temporal (50 Bilder)	86.2 %
OH 16-3	kNN	Temporal (50 Bilder)	96.3 %
EV dim=20	SVM	Spatio-temporal	98.3 %
EV dim=20	kNN	Spatio-temporal	97.4 %
OH 16-3	SVM	Spatio-temporal	99.9 %
OH 16-3	kNN	Spatio-temporal	99.5 %

Tabelle 5.1.: Resultate der videobasierten Nutzeridentifikation auf Merkmalsinstanzebene, nach temporaler Fusion und nach spatiotemporaler Fusion.

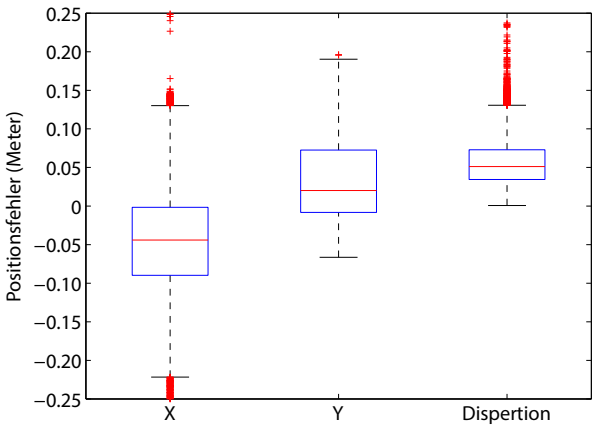

Abbildung 5.8.: Boxplot des Fehlers der geschätzten Position, Angaben in Zentimetern. Die linken beiden Boxen beschreiben den Fehler in x und y Richtung. Die Dispertion bewertet die Güte der Positionsschätzung durch das System selbst. Ausreißer mit hohen Dispertionswerten entsprechen den Ausreißern mit schlechter Positionsschätzung. Das System kann eine schlechte Schätzung somit korrekt erkennen.

Histogrammauflösung von 16 als 144 dimensionaler Merkmalsvektor extrahiert. Als Klassifikatoren wurden kNN mit $k = 10$ und eine SVM mit radialer Kernfunktion verglichen. Tabelle 5.1 zeigt die erreichten Resultate. Klar zu erkennen ist der Zugewinn der Erkennungsleistung durch Einsatz von temporaler und spatialer Fusion.

Resultate der Positionsschätzung

Zur Bewertung der Güte der Positionsschätzung wurden die vom System geschätzten Positionen mit zuvor manuell gemessenen Koordinaten verglichen. Abbildung 5.8 zeigt die Abweichungen in x und y Richtung. Der Positionsfehler kann als gering bezeichnet werden, insbesondere ist er wesentlich kleiner als der Durchmesser eines menschlichen Körpers. Auch die systemeigene Bewertung der Positionsschätzung anhand des Dispertionswertes zeigt bei guten Schätzungen kleine Werte. Insbesondere wurden nahezu alle Fehler der Positionsschätzung auch durch große Dispertionswerte vom System angezeigt (In der Abbildung nicht dargestellt).

5.3.2. Das Demonstrationsszenario 3 (DS3) des SFB-TRR 62

Dieser Abschnitt beschreibt den Demonstrator 3 des SFB-TRR 62 welcher das Zusammenspiel mehrerer sensorischer Komponenten realisiert. Im Rahmen der Demonstration wurden alle sensorischen Werte in Echtzeit erfasst und ausgewertet. Als Anwendung wurde ein Fahrkartenautomat gewählt, welcher durch Einsatz von Companion-Technologie um Elemente der Companion-Eigenschaften erweitert wurde. Die Anwendung ist dabei im wesentlichen abstrakt gehalten und kann beliebig ausgewechselt werden[15]. Nicht realisiert wurden Komponenten einer Wissensaquise und Wissensmodellierung. Auch die Nutzerinteraktion ist auf ein Ausgabesystem begrenzt[16]. Die Implementierung erfolgte unter Einsatz ähnlicher Konzepte wie die MAR^{2}S, dezentral und ohne Einsatz von Triggersignalen. Alle beteiligten Module kommunizieren über Netzwerk mittels SEMAINE Middleware und den darin modellierten Nachrichtenkanälen (Topics). Die folgenden Abschnitte beschreiben die Funktionsweise des realisierten Systems, beschrieben in [157] und [243].

Motivation und involvierte Komponenten

Companion-Systeme sollen unter verschiedenen Umgebungsparametern und technischen Gegebenheiten arbeiten und stehen somit vor der Herausforderung, Informationen aus verschiedenen sensorischen Quellen unter verschiedenen Bedingungen zu analysieren und zu integrieren. Der technische Aufbau definiert die für die Signalverarbeitung verfügbaren Sensortypen. Die schwer kontrollierbaren Umgebungsbedingungen schränken die Zuverlässigkeit dieser Sensoren ein, wodurch verschiedene Arten der Informationsverarbeitung und -interpretation erforderlich werden. Das aus den Sensoren abgeleitete Wissen über den Zustand des Nutzers und den Nutzungskontext ermöglicht es dem System schließlich, nicht nur seine Funktionalität, sondern auch die Art der Interaktion mit dem Nutzer in Bezug auf verfügbare Ein- und Ausgabemodalitäten zu adaptieren. Unter Verwendung des Szenarios *Fahrkartenkauf* wird diese Adaptivität anhand multimodaler, dynamischer Interaktion zwischen Nutzer und technischem System gezeigt.

Um den Ticketautomaten zu realisieren, wurden mehrere PCs mit Video- und Audio-Aufnahme-Geräten, Laserscannern, einem Touchscreen und einem Tiefensensor ausgestattet. Diese werden

[15] Im Laufe der Implementierung wurde zunächst ein Getränkeautomat realisiert. Aufgrund der Abstraktion von einer konkreten Anwendung, strikter Trennung von Anwendungsoberfläche und Funktionalität und der modularen Gesamtarchitektur konnte innerhalb weniger Tage auf einen Fahrkartenautomat als exemplarische Anwendung gewechselt werden.

[16] Wissensbasis, Wissensaquise und multimodale Nutzerinteraktion durch Fission (ausgabeseitige Multimodalität) wurden im Rahmen eines weiteren Demonstrators ohne sensorische Komponenten durch andere Arbeitsgruppen realisiert [19]

Abbildung 5.9.: Ein Nutzer während der Interaktion mit dem prototypischen Fahrkartenautomaten. Die Anwendung wird auf dem rechten Bildschirm dargestellt, während die linken Bildschirme interne Zustände und sensorische Werte des System zu Entwicklungszwecken darstellen.

von verschiedenen Softwarekomponenten verwendet, die explizite und implizite Nutzersignale verarbeiten. Explizite Signale beschreiben Befehle, die vom Nutzer in der Absicht ausgeführt werden, mit dem System zu interagieren, wie z.B. Interaktionsgesten, Sprache und Berührungseingabe. Implizite Signale dagegen sind nicht direkt an das System gerichtet, enthalten aber dennoch eine Fülle von Informationen, die für das System relevant sind. Diese Signale umfassen den situativen Kontext des Nutzers, seine Gestik, Sprache, Körperhaltung, Mimik und Prosodie. Der Informationsfluss zwischen den beteiligten Modulen ist in Abbildung 5.10 dargestellt.

Während explizite Nutzersignale direkt in die Eingangsfusions eingehen und so Signale an die Ticketanwendung senden, werden implizite Signale zuerst kombiniert und in einer dedizierten Datenfusionskomponente weiter abstrahiert. Die technische Architektur und die Kommunikations-Middleware des zugrundeliegenden Systems sind durch eine ähnliche Architektur wie die MAR^{2}S realisiert. Unterschiede kommen aufgrund der unterschiedlichen und

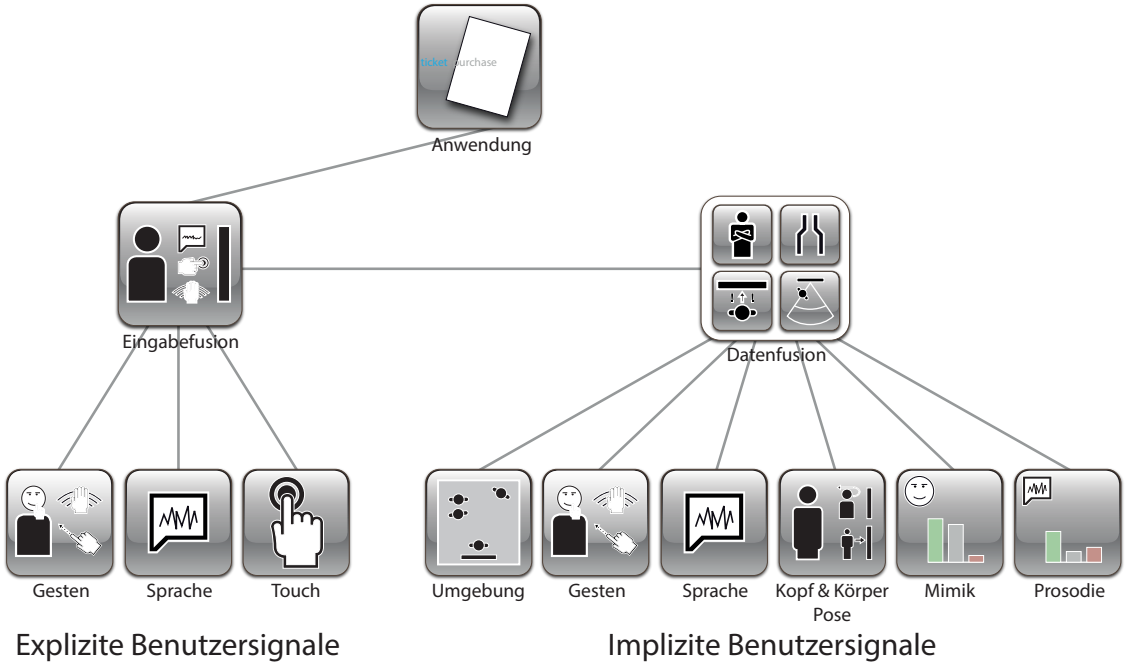

Abbildung 5.10.: Sensorischer Bottom-Up-Informationsfluss im prototypischen Fahrkartenautomaten. Explizite Nutzersignale (Gesten-, Sprach- und Berührungseingaben) werden direkt in der Eingabefusion verarbeitet und an die Anwendung gesendet. Implizite Signale stammen von verschiedenen Sensoren. Laserscanner beobachten die Umgebung des Nutzers. Videokameras erfassen Gesten, Mimik, sowie die Kopf- und Körperhaltung. Mikrophone analysieren die Sprache und nonverbale Signale. In der Datenfusionskomponente impliziter Signale werden diese zu symbolischen Informationen über den Zustand des Nutzers kombiniert (z.B. *unerwartetes Systemverhalten*, *Aufmerksamkeit* oder *Hastigkeit*) und an die Eingabefusion als implizites Signal weitergeleitet.

teilweise parallelen Entwicklung zustande. Während die hier verwendetet Architektur nur die schnelle Systemrealisierung des Demonstrators im Fokus hatte, fokussierte sich die MAR^{2}S zunächst auf ihre Aufnahme Komponenten. Erst zu späteren Zeitpunkten wurde aufgrund der großen strukturellen Gemeinsamkeiten ein einheitliches System entworfen.

Die relativ einfachen Planungs- und Dialogverwaltungsaufgaben werden innerhalb der Anwendungskomponente realisiert. Das Gleiche gilt für die Speicherung von Nutzerinformationen und Präferenzen, die ohne eine dedizierte Wissensbasiskomponente ausgeführt wird. In einem tatsächlichen Companion-Systeme würde die Anwendung in einer weitere Vielzahl von Modulen realisiert werden[17]. Die Anwendung bietet hier lediglich schrittweise Dialoge an, welche die wichtigsten Informationen für den Kauf eines Zugtickets zusammenstellen.

[17] Siehe hierzu den Schwester Demonstrator DS1 mit vollständiger Anwendung ohne sensorische Komponenten [19]

Im folgenden Abschnitt wird ein vollständiger Durchlauf eines normalen Ticketkaufs beschrieben. Verarbeitung, Informationsfusion, Nutzeranpassung und ggf. Interaktion. Der anschließende Abschnitt hebt dann die Möglichkeiten eines verkürzten Ticketkaufs im Falle eines erfahrenen eiligen Nutzers hervor.

Beispielhafter Durchlauf eines normalen Ablaufs

Der beschriebene Durchlauf beginnt mit einem Nutzer, der sich dem System nähert. Der Verkauf wird durch den Sonderfall einer Unterbrechung gestört, ansonsten wird ein normaler Ticketkauf beschrieben, der aus Angabe des Ziels, der Reisezeit, der Anzahl der Tickets und der spezifischen Zugverbindung besteht. Das Verlassen des Geräts markiert schließlich das Ende des normalen Kaufs.

Annäherung an den Automaten: Die Aktivierung des Ticketsystems wird basierend auf dem Laser-Umgebungswahrnehmungssystem und dem tiefenkamerabasierten Kopf- und Körperhaltungs Modul erkannt. Mithilfe von Multi-Objekt-Tracking-Algorithmen [188] werden die Zustände aller Personen in der Nähe des Systems mit zwei Laser-Entfernungsmessern geschätzt. Alle Personen, die sich auf das Ticketsystem zubewegen, werden auf der Grundlage ihrer Trajektorien als potenzielle Nutzer angesehen. Um eine Aktivierung des Systems durch Passanten zu verhindern, kombiniert das Datenfusionsmodul den Zustand des Trackings mit den Ergebnissen der Kopf- und Körperposeschätzung [98]. Nutzer müssen sich auf das System zubewegen und mit Körperhaltung und Kopfpose dem System zugewandt sein. Somit lösen nur tatsächliche potenzielle Nutzer den Beginn eines Ticketkaufs aus. Nachdem ein neuer Nutzer erkannt wurde, startet die Anwendung bereichsabhängig eine Überblendung vom Standby-Bildschirm zum Begrüßungsbildschirm des Kaufprozesses. Mit einem kalibrierten Koordinatensystem, wird die globale Position des aktiven Nutzers an andere Module übertragen. So werden von anderen Sensoren keine Sensordaten von nichtaktiven Nutzern verarbeitet.

Wahl des Reiseziels: Der erste Schritt im Ticketkauf Prozess ist die Auswahl des Reiseziels. Ist der Nutzer dem System bekannt, sind individuelle Kontextinformationen über den Nutzer verfügbar. In diesem Dialogschritt werden automatisch die häufigsten Reiseziele des Nutzers auf einer Karte vorgeschlagen. Der Nutzer kann nun ein Ziel auswählen, indem er die Toucheingabe auf der Karte verwendet oder ein Ziel über eine direkte Spracheingabe spezifiziert.

Wahl der Abfahrtszeit: Der zweite Schritt ist die Auswahl von Datum und Uhrzeit der Reise. Ist der Nutzer dem System bekannt, ist sein persönlicher Terminplan als individuelle Kontextinformation verfügbar und wird angezeigt. Der Nutzer kann so ein Zeitfenster für die Reise wählen.

Die Reisezeitwahl kann entweder durch Spracheingabe oder durch Gesten erfolgen. Es ist dabei meist einfacher ein bestimmtes Datum verbal zu benennen (Zum Beispiel „Ich möchte am Mittwoch um 8 Uhr fahren“), hingegen ist das Blättern in einem Kalender via Gesten eher umständlich. Es stehen im Falle von Gesten nur bestimmte Gesten wie *nach links wischen* zur Verfügung um zur nächsten Woche zu gelangen. Das Zeigen auf kleine Elemente wie einzelne Kalendereinträge ist schwer zu erkennen und somit fehleranfällig.

Die Zeigerichtung wird in virtuellen Bildschirmkoordinaten an die Anwendung übertragen. Ihr Wert wird auf Grundlage der globalen Position des Nutzers und seiner Hand am ausgestreckten Arm vom Gestenmodul bestimmt und an andere Module übertragen. Befindet sich die Hand des Nutzers nahe an seinem Körper, wird die Zeigerichtung durch lokale Handbewegungen eingestellt. Zusätzlich wird eine grafische Rückmeldung gegeben, um die erkannte Zeigegeste zu bestätigen.

Beide Eingabemethoden werden unabhängig erkannt und in der Eingabefusion verarbeitet (siehe Abbildung 5.10) und das ausgewählte Zeitfenster wird angezeigt. Nach einer Nutzerbestätigung wird dies für den weiteren Vorgang verwendet. Durch die Verwendung der Eingabefusion können beide Modalitäten auch kombiniert werden. Der Nutzer kann einen bestimmten Kalendereintrag durch Zeigen auswählen, und durch Verwendung eines zusätzlichen Sprachbefehls, wie „wöchentlich“ergänzen

Wahl der Ticketanzahl: Das Ticketsystem ermöglicht zusätzlich eine Anpassung der Anwendung an den aktuellen Kontext, dies wird anhand einer automatischen Auswahl Ticketanzahl veranschaulicht. Durch Verwendung eines Mehrfachobjekt-Verfolgungsalgorithmus [188] kann das System erkennen, dass sich der aktuelle Nutzer mit mehreren anderen Personen genähert hat. Die Bestimmung der Gruppegröße basiert auf räumlichen Regionen sowie auf gemeinsamen Trajektorien der Vergangenheit. Kennt die Anwendung die Gruppengröße schlägt es automatisch vor, entweder ein einzelnes Ticket oder Tickets für alle Personen in der Gruppe zu kaufen.

Unterbrechung des Kaufvorgangs: Eine Herausforderung mit der sprachgesteuerte technische Systeme umgehen müssen, ist die Unterscheidung zwischen beabsichtigten Nutzerkommandos und Nutzer-Nutzer Kommunikation. Die zweite Art von Äußerungen wird üblicherweise als *Off-Talk* bezeichnet. Solange sich der Inhalt dieses Off-Talks von Systembefehlen unterscheidet, kann die Differenzierung rein aus dem Sprachinhalt unterschieden werden. Aber in Situationen, bei denen der Off-Talk ähnliche Phrasen wie die der erwarteten Nutzerkommandos enthält, wird dieser Ansatz erschwert. Dies kann zum Beispiel entstehen, wenn der Nutzer klärende Dialoge mit seinen Gruppenmitgliedern führt oder via Mobiltelefon über eine Reise spricht. In diesem Fall kann die Entscheidung, ob die Äußerungen des Nutzers das System

steuern sollen oder nicht, multimodal bestimmt werden. In unserem Szenario untersuchen wir zwei Ereignisse, bei denen wir diese Art von Off-Talk anhand multimodaler Sensoren erkennen: 1) Nutzer wendet sich zeitweise vom System ab, 2) Nutzer telefoniert.

Für den ersten Fall muss das System nur die Pose eines Nutzers erkennen. Blickt der Nutzer nicht auf das System, wird die Eingabe mittels Sprache vom Spracherkenner ignoriert, bis der Blick des Nutzers wieder auf das System gerichtet ist.

Für den zweiten Fall muss das System sowohl die erkannte Sprache als auch die Geste interpretieren. Um einen Anruf zu tätigen oder anzunehmen wird das Mobiltelefon ans Ohr gehalten. Weiter werden zu Beginn typische Phrasen zur Begrüßung erkannt. Die Selbstberührung des Ohrs des Nutzers kann von unserem Gestenerkenner erkannt werden, während der Spracherkenner die begrüßenden Äußerungen registriert. Beide Ereignisse werden in den einzelnen Modulen unabhängig voneinander bewertet. Treten beide innerhalb kurzen Zeitspanne auf, werden sie innerhalb der Datenfusion erkannt und als implizites Nutzersignal verwendet. Das System deaktiviert den Spracherkenner und den Sprachsynthesizer, solange das Telefonat andauert.

Wahl der Verbindung: Nachdem alle benötigten Informationen vom System erfasst wurden, kann es nun nach Zugverbindungen suchen, Präferenzen sind dabei aus der Wissensbasis bekannt. Gibt es keine geeignete Verbindung die der Präferenz voll entspricht muss das System Verbindungen wählen, welche nur teilweise übereinstimmen. In dieser Demonstration zeigt das System nur die Verbindungen an, bei denen der Nutzer eine Reservierung vornehmen kann, leider muss er dabei sehr oft umsteigen.

Die Reaktion des Nutzers, die mittels Gesichtsausdruck und Stimme erfasst wird, zeigt an, dass er oder sie nicht mit der Vorauswahl zufrieden ist. Als implizite Eingabe für das Datenfusionsmodul kann das trainierte System den emotionalen Zustand des Nutzers erkennen. Die negative implizite Reaktion wird somit erkannt. Der Gesichtsausdruck wird anhand von Merkmalen klassifiziert, die aus geometrischen Abständen der Gesichtsregionen bewertet werden (z. B. Mundbreite zu Mundhöhe oder Augenbrauen Abstand) [166].

Zusätzlich werden während der gesamten Interaktion MFCC Merkmale extrahiert und durch eine probabilistische SVM bewertet, so werden positive und negative Emotionen identifiziert (vgl. Kapitel 4). Beide Modalitätsbewertungen werden an das Datenfusionsmodul gesendet, welches daraus den emotionalen Zustand des Nutzers mit einem MFN bewertet.

Nach Erkennen der negativen Reaktion des Nutzers werden diese Informationen vom impliziten Eingabemodul, über das zentrale Eingabemodul, an die Anwendung gesendet. Diese löst eine Erklärung der Vorschläge aus und fragt beim Nutzer nach, ob die Vorauswahl angepasst werden

soll. Wünscht der Nutzer dies, wird die Liste der Zugverbindungen erweitert. Der Nutzer wählt eine Verbindung die am ehesten seiner Präferenz entspricht und fährt mit dem Bezahlvorgang fort. Die Nutzerentscheidung der sekundären Präferenz im Falle der Unerfüllbarkeit der primären Präferenzen wird zur späteren Verwendung und Systemoptimierung im Nutzermodell abgelegt.

Ende des Kaufvorgangs: Nach dem Kaufprozess bleibt das System aktiv, bis sich der Nutzer vom System abwendet. Ähnlich der Methodik die verwendet wird um sich dem Gerät zu nähern, wird das Ende der Interaktion unter Verwendung des Zustands und der Position des Nutzers ausgelöst. Das System kehrt nur in den Stand-by-Modus zurück, wenn die Entfernung des Nutzers einen bestimmten Schwellenwert überschreitet und dieser nicht mehr dem System zugewandt ist.

Beispielhafter Durchlauf eines eiligen Expertenablaufs

Das Ticketsystem unterstützt zwei grundlegende Interaktionsmodi. Neben dem zuvor beschriebenen normalen Modus gibt es einen schnellen Kauf für erfahrene Nutzer, die es eilig haben.

Im Expertenmodus benötigt die Interaktion weniger Zeit und verwendet weniger Modalitäten: Nur Berührungs- und Sprachinteraktionen sind möglich. Das heißt, die Eingabe über Gesten ist deaktiviert, da angenommen werden kann, dass die Gesten eines Nutzers, der in Eile ist, nicht präzise genug sind, um vom System korrekt erkannt zu werden.

Die Auswahl des Interaktionsmodus erfolgt automatisch anhand einer Analyse der Annäherungsgeschwindigkeit des Nutzers an das Ticketsystem. Falls sich das System für den Expertenmodus entscheidet, prüft es zunächst mittels Rückfrage, ob der Nutzer den vorgeschlagenen Expertenmodus wünscht.

Um kürzere Interaktionszeiten zu erreichen, muss der Nutzer beim Kauf weniger Auswahlen treffen. Im Schnellmodus wird der Nutzer nicht nach der Auswahl von Optionen gefragt, da das System sie automatisch unter Berücksichtigung der verfügbaren Informationen einstellt. Zum Beispiel legt es die Anzahl der Tickets auf die Größe der Gruppe von Personen fest, die mit dem System interagieren. Die Anzahl wird basierend auf den Daten der Laser-Entfernungssensoren abgeleitet. Informationen können aber auch aus den Präferenzen des Nutzermodells stammen. Da das System weis, dass der Nutzer in Eile ist, schlägt es vor, einen Zug zu nehmen, der als nächstes fährt. Zur Abfahrt benutzt es den aktuellen Bahnhof. Das Ziel kann durch den Nutzer unter Verwendung einer Spracheingabe ausgewählt werden. Treten Mehrdeutigkeiten auf, behebt das System das Problem, indem zusätzliche Nutzerabfragen angezeigt werden. Die

letzte erforderliche Nutzereingabe ist die Bestätigung der Zugverbindung. Danach ist der Kauf erfolgreich abgeschlossen.

Zusammenfassung

Im beschriebenen Szenario wurden einige Schlüsselkomponenten der Companion-Technologie exemplarisch in einem Prototypen eines Ticketsystems realisiert. Neben der Möglichkeit, mehrere Eingabemodalitäten zu verwenden, passt das vorgeschlagene System sein Verhalten an die aktuelle Situation an und interpretiert implizite Eingabedaten. Die Anpassung des Systems an die aktuelle Situation wird durch die Unterbrechungsbehandlung, die Auswahl der Ticketzahl und die automatische Umschaltung zwischen Standard- und Expertenmodus demonstriert. Beispiele für implizite Eingabedaten sind die Interpretation der Mimik und Stimme.

Obwohl der präsentierte Prototyp die Interaktionen zwischen Mensch und Maschine, im Vergleich zu derzeit verfügbaren Systemen, deutlich verbessert, ist die Integration einer Wissensbasis und eines umfangreichen Nutzermodells zur vollständigen Erfüllung der Companion-Eigenschaften unerlässlich. Es wird erwartet, dass die Anpassungsfähigkeit des Systems an die Präferenzen des Nutzers dadurch weiter verbessert wird. Die Wissensbasis könnte dabei zum Beispiel als Webdienst auf Basis einer Cloudanwendung realisiert sein, auf das die Anwendung jederzeit zugreifen kann, während das Nutzermodell auf dem Smartphone des Nutzers gespeichert ist und nur bei Bedarf an die Anwendung übergeben wird.

5.3.3. Adaptive Nutzer-Interaktion

Die folgenden Überlegungen des Zusammenspiels der emotionserkennenden Komponenten mit den internen Anwendungskomponenten und den Nutzerschnittstellen beschreiben ein mögliches Szenario eines Companion-Systemes. Dabei wurden die Ergebnisse dieser Arbeit und die Arbeiten von F. Honold zur multimodalen Fission[18] und F. Schüssel zur Nutzermodellierung und individuellen Interaktionshistorie zu einem hypothetischen Gesamtsystem kombiniert. Die Resultate sind in einer gemeinsamen Arbeit veröffentlicht [153].

Wir betrachten Companion-Systeme als intelligente Systeme, die in der Lage sind, Emotionen zu erfassen und auf diese durch Anpassung auf verschiedenen Ebenen zu reagieren (siehe Abbildung 5.11). Dies folgt dem in [142] beschriebenen Prinzip und der daraus resultierenden Herausforderung.

[18] Ausgabeseitige Nutzerinteraktion in mehreren Geräten und Modalitäten

Diese Möglichkeiten der Anpassung auf verschiedenen Ebenen erzeugen ein Entscheidungsproblem, das emotionsauslösende Objekt in der Anwendung zu finden und das richtige Maß der Anpassung zu wählen. Mag zum Beispiel ein Nutzer die aktuelle Nutzeroberfläche nicht (z. B. die Verwendung eines Smartphone-Bildschirms in Sonnenschein) ist es nicht notwendig andere Ebenen als die Ausgabe anzupassen. Der Systemvorschlag eine andere (leichtere) Aufgabe durchzuführen, oder die Anpassung der aktuell verwendeten Dialogstruktur wird nicht effektiv sein. Eine alternative Interaktion via Sprache anzubieten, würde in einer solchen Situation mehr zum Wohlbefinden des Nutzers beitragen.

Im restlichen Abschnitt liegt der Focus auf Companion-Systemen mit der Fähigkeit, Emotionen des Nutzers wahrzunehmen, jedoch ohne die Fähigkeit Emotionen in Form eines Avatars auszudrücken. Dies bedeutet dabei nicht, dass Emotionen nicht durch Systemausgaben ausgelöst werden können. Dies geschieht allerdings eher subtil (z. B. durch die Verwendung einer adaptiven Dialog Struktur mit Elementen für Entschuldigungen (vgl. [1]) und Ermutigungen oder durch den Einsatz von Hintergrundmusik).

Im Folgenden wird zunächst eine mögliche Systemarchitektur nach Honold und Schüssel beschrieben, welche es erlaubt emotional adaptive Systeme zu realisieren. Anschließend werden die Möglichkeiten solcher Systeme aufgezeigt.

System Architektur

Die verschiedenen Schichten der Systemarchitektur (siehe Abbildung 5.11) können auf das Architektur-Modell von Gram und Cockton [93] abgebildet werden. Die vorgestellten Komponenten ermöglichen einen Systemaufbau in Ebenen [109]. Die höchste Ebene aus Abbildung 5.11 enthält den *Funktionskern* einer Anwendung (z. B. ein Mediaplayer mit all seinen Funktionen). Diese Ebene enthält auch eine Komponente für die *KI-Planung*. Benötigt der Nutzer Unterstützung bei der Lösung eines komplexen Problems, ist der Planer verantwortlich eine Folge von Aufgaben abzuleiten, die zu einer Lösung des Problems führen.

Die Ebene des *Dialog- und Kernadapters* dient als interaktionsmodalitätenunabhängige Schnittstelle zur Funktionalität des Kerns auf der einen Seite, und ist in der Lage, auf der anderen Seite den Inhalt jedes notwendigen Aufgabenschritts zu kommunizieren. Der *Phraser* muss eine nutzerindividuelle Dialogstruktur für das System zur Verfügung stellen. Der *Interpreter* muss Nutzereingaben interpretieren und entscheiden, welche Funktionen des Kerns aufgerufen werden müssen oder welche Planaufgaben erfüllt wurden.

Die nächste Ebene stellt die *logische Interaktions-Schicht* dar. Die *Fissions-Komponente* verfeinert die abstrakte Ausgabe des Dialogmanagements und bewertet die modalitätsspezifischen

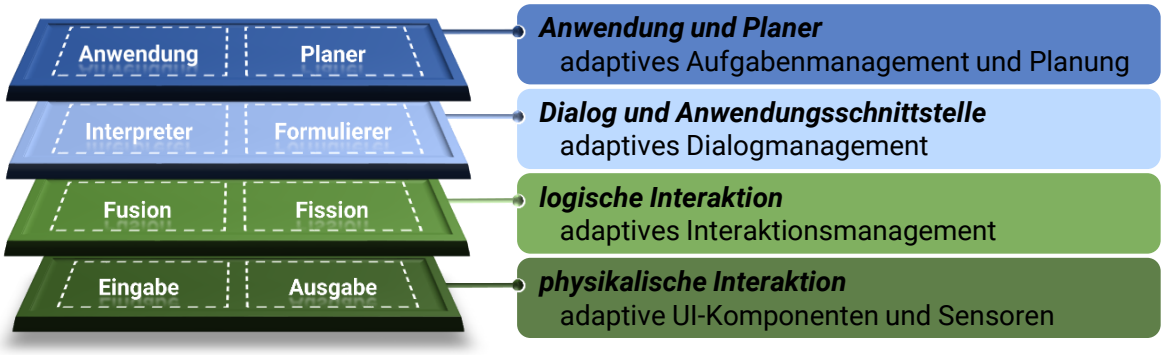

Abbildung 5.11.: Ebenenmodell der adaptiven Komponenten. Affektiv-adaptive Entscheidungen können auf unterschiedlichen Ebenen getroffen werden. (Bildquelle: Dissertation Honold [108], basierend auf Meudt et. al. [153])

Ausgabemöglichkeiten. Dabei entscheidet diese über die Verwendung von später verwendeten Gerätekomponenten und Sensoren [109]. Die zweite Komponente dieser Ebene ist die Komponente der *Eingabefusion*. Diese Komponente ist in der Lage, multimodale Eingaben und semantische Bedeutung abzuleiten, die als Eingabe für den Dialog-Manager dienen.

Die unterste Ebene für die *physische Interaktion* umfasst alle verfügbaren Gerätekomponenten der Eingabe und Ausgabe. Ausgabekomponenten werden durch die Fissionskomponente ermittelt und erzeugen die Ausgabe des Dialogmanagers in einer multimodale Weise. In gleicher Weise werden Eingangssensoren mit spezifischen Grammatiken oder Konfigurationsdaten bewertet. Dabei werden mögliche implizite Nutzereingaben (z. B. emotionaler Status oder Standort des Nutzers) sowie explizite Nutzereingaben (z. B. Spracheeingabe oder Gesteneingabe des Nutzers) erfasst.

Anpassungsmöglichkeiten auf jeder dieser verschiedenen Ebenen werfen die Frage auf: Wann soll welche Komponente, auf welcher Ebene, aufgrund von erkannten Emotionen adaptiert werden? Es ist offensichtlich, dass Anpassungen auf allen Ebene, wann immer es möglich ist, nicht zu dem gewünschten Ergebnis in Bezug auf emotionales Verhalten führen. Als Beispiel: Mag ein Nutzer das aktuelle Layout der System-Nutzeroberfläche bei der Arbeit an einer bestimmten $Aufgabe_1$ nicht, ist es besser die Nutzeroberfläche für diese Aufgabe anzupassen, als dem Nutzer eine andere $Aufgabe_2$ vorzuschlagen. Daher zielt dieser Ansatz darauf ab, eine Identifizierung der Ursache einer plötzlichen Emotionsänderung durchzuführen, und nur die als ursächlich angenommen Systemkomponente zu adaptieren.

Der Erfolg hängt dabei entscheidend von der Fähigkeit ab, den höchstwahrscheinlichen Grund einer wahrgenommenen, unerwünschten, emotionalen Wirkung zu bestimmen. Daher sollte jeder Schritt im Systemverhalten protokolliert werden. Abbildung 5.12 zeigt eine beispielhafte

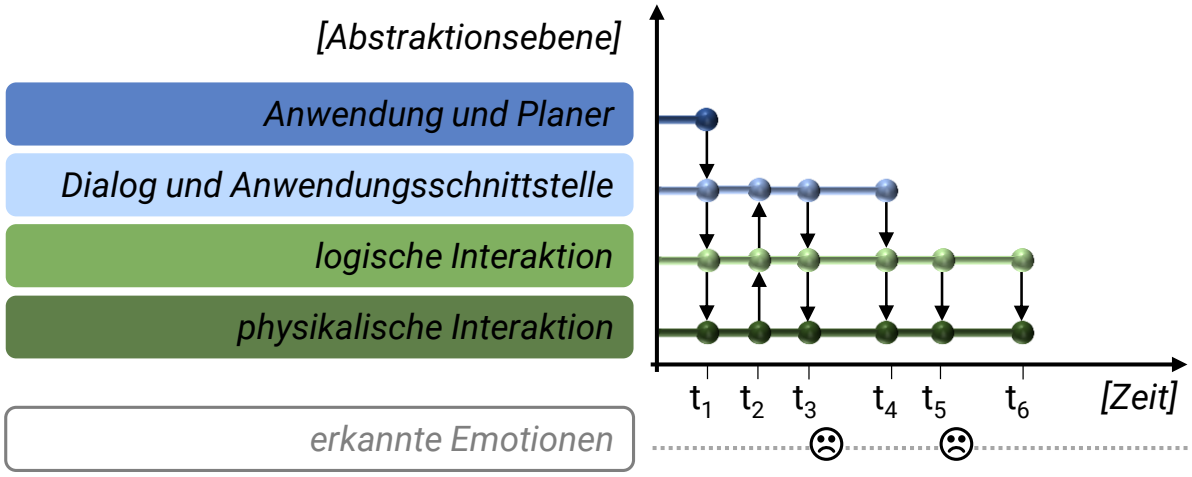

Abbildung 5.12.: Zeitliche Abfolge ebenenspezifischer Aktionen mit erkannten Emotionen. Jede Kugel repräsentiert eine Aktion welche, durch eine Beschreibung, als Aktionsereignis in die Interaktionshistorie aufgenommen wird. Aktionen erfolgen dabei meist als Kaskade, entweder durch Konkretisierung einer Ausgabe (Zeitpunkt t1) oder in Folge der Abstraktion einer Eingabe (Zeitpunkt t2). Im Beispiel wurden nach t3 und t5 negative Emotionen erkannt. (Bildquelle: Dissertation Honold [108], basierend auf Meudt et. al. [153])

Sequenz von Systemaus- und Eingaben mit ebenenspezifischen Ereignissen. Diese komponentenspezifischen Aktionsereignisse werden gespeichert und bilden eine hierarchische, zeitbasierte Ereignishistorie, welche als *Interaktionshistorie* bezeichnet wird (siehe Abbildung 5.13). Jedes Element in der Historie repräsentiert ein 5-Tupel bestehend aus ⟨ *Zeitstempel, Komponente, Trigger, Aktion, erwartete emotionale Reaktion* ⟩, bezeichnet als *Aktionsgegenstand*. Wobei *Komponente* eine der Systemkomponenten, die zu einem bekannten *Zeitstempel* eine bestimmte *Aktion* ausgeführt hat, aktiviert durch einen bestimmten *trigger*. Da das System sich seiner Aktionen und seiner eigenen Ausgabe bewusst ist [110], kann es in machen Situationen sogar *erwartete emotionale Reaktionen* vorhersagen. Weis das System zum Beispiel, dass es die Privatsphäre nicht respektieren kann, weil ein Problem in einer bestimmten Ausgabemodalität besteht, ist mit Unmut durch den Nutzer zu rechnen. Nicht jedes Aktionselement besitzt eine solche Information. Wenn aber ein Aktionselement solche Informationen bereitstellt, kann es als nützlicher Hinweis verwendet werden, um das auslösende Objekt zu identifizieren, wenn die Reaktion nicht der Erwartung entspricht (siehe Abbildung 5.13).

Immer wenn eine unerwartete Veränderung der Emotion eines Nutzers erkannt wird, überprüft das System seine Interaktionshistorie und identifiziert mögliche Aktionen, die unmittelbar vor Beginn der Veränderung der Emotionen stattgefunden haben. Dabei kann das System zwei verschiedene Fälle unterscheiden.

Eine mögliche Ursache: Im besten Fall zeigt die Überprüfung der Interaktionshistorie nur eine mögliche Komponente, welche als emotionales Objekt identifiziert werden kann. Dreht der Nutzer zum Beispiel sein Smartphone, wechselt das Layout möglicherweise von Querformat zu Hochformat, dies kann zu einer beschnittenen Ansicht der Nutzeroberfläche führen. Eine solche Aktion könnte in der Interaktionshistorie als Aktionselement 5-Tupel ⟨*1430923042637, phone.display, enterLandscapeMode, adaptationOfLayout, negativeReaktion*⟩ gespeichert worden sein (vgl. Aktion t_6 in Abbildung 5.12). Die Aktion der referenzierten Komponente des Aktionselementes kann als emotionales Objekt identifiziert werden.

Multiple mögliche Ursachen: In einer beispielhaften Situation wird die Ausgabe wegen eines sich bewegenden Nutzers von einem Bildschirm zu einem anderen übertragen. In einem solchen Fall erfolgte die Entscheidung über die Ausgabe durch die Fissionskomponente des Interaktionsmanagements (vgl. Aktion t_5 in Abbildung 5.12). In dem Beispiel verursacht die erste Aktion eine Folgeaktion auf der untersten Ebene mit ihren beteiligten Ausgabekomponenten. Der alte Bildschirm muss in den Standby Modus wechseln und der neu adressierte Bildschirm muss die designierte Ausgabe rendern. Basierend auf diesem Ansatz kann in einem solchen Fall eine Veränderung der Emotionen auf mehrere mögliche Gründe abgebildet werden. Im gegebenen Beispiel bietet die Interaktionshistorie zwei mögliche Aktionselemente, die mit t_5 in Verbindung stehen. Die nächsten Abschnitte beschreiben Strategien, wie mehrere Möglichkeiten reduziert werden können, um die emotionsauslösende Komponente identifizieren zu können.

Erwartungswertbasierte Identifikation: Wie oben erwähnt, können Aktionselemente optionale Unterelemente enthalten, die Informationen über eine erwartete emotionale Reaktion bereitstellen. Solche Informationen können verwendet werden um mögliche emotionale Objekte zu disambiguieren und die Anzahl der mögliche Gründe zu reduzieren. Elemente, für welche die erwartete Reaktion gleich der erkannten Emotion sind, können mit hoher Wahrscheinlichkeit als Ursache für die emotionale Änderung angenommen werden.

Reduktion, basierend auf Kausalität und historischer Evidenz: Da die Aktionsebenen voneinander abhängen, ist es wahrscheinlich, dass Aktionselemente in Kausalzusammenhang stehen sofern sie zeitnah erfolgen. In solchen Fällen können kausale Abhängigkeiten abgeleitet werden, indem *trigger* und *Aktion* Marker von jedem beteiligten Element in der Interaktionshistorie bewertet werden. Für jede erzeugte kausale Kette (oder Hierarchie) wird versucht untergeordnete Aktionselemente zu entfernen, basierend auf der Evidenz der Historie. In diesem Zusammenhang können historische Informationen früherer Handlungen, bei denen keine negativen Auswirkungen festgestellt wurden, analysiert werden. Wenn zum Beispiel eine Medienanwendung eine Playlist abspielt und eine neues Lied wählt, kann dies auf allen unteren Ebenen Aktionen auslösen (vgl. t_2 in Abbildung 5.12). Wenn aber der Nutzer mit dem früheren

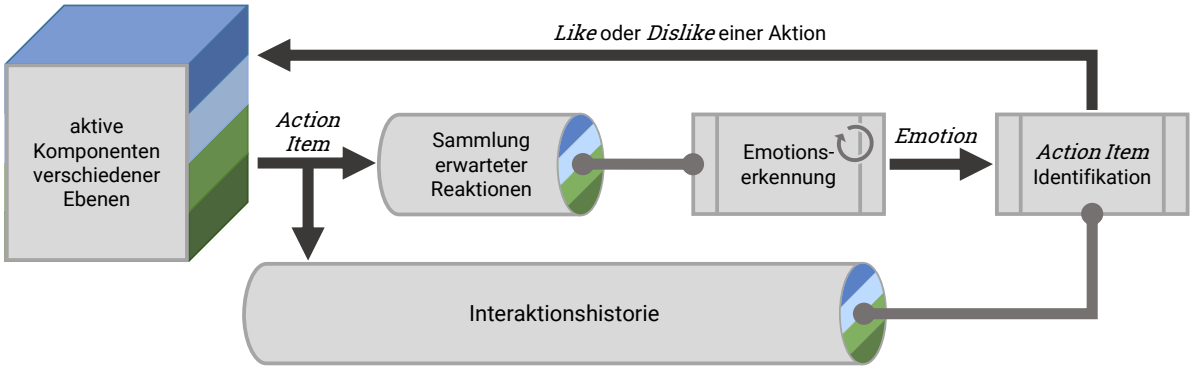

Abbildung 5.13.: Prozess zur Realisierung eines affektiv-adaptiven Systemverhaltens. Systemaktionen werden in der Interaktionshistorie ebenenweise verwaltet. Das Auftreten von Emotionen ermöglicht eine zeitbasierte Identifkation einzelner Aktionen innerhalb der Historie. Die relevanten Komponenten erhalten eine entsprechende Information. (Bildquelle: Dissertation Honold [108], basierend auf Meudt et. al. [153])

Systemverhalten zufrieden war, und der Kontext sich nicht geändert hat, kann die auslösende Komponente auf niedrigeren Ebenen vermutet werden. So kann vermutet werden, dass der gewählte Song nicht gut gewählt war.

Rückfrage beim Nutzer: Sind die oben genannten Strategien nicht in der Lage, ein Aktionselement eindeutig zu bestimmen, kann ein Companion-Systeme selbsttätig einen neuen Dialog erzeugen. Damit kann das Companion-System zur Laufzeit beim Nutzer, mit der Bitte um Klärung, rückfragen. Der Inhalt einer solchen Rückfrage wird dabei durch die Auswahl verbleibender, möglicher, ursächlicher Komponenten erzeugt. Jeder der Auswahlpunkte bezieht sich dabei auf ein mögliches emotionales Objekt.

Beispiele von affektbasiertem Systemverhalten

Die Interaktion eines Nutzers mit einem Companion-System wird über die logischen und physischen Interaktionsschichten realisiert (vgl. Abbildung 5.11). Um also eine sinnvolle Anpassung innerhalb dieser Schichten zu ermöglichen, muss die Ursache einer Veränderung im emotionalen Zustand des Nutzers ebenfalls innerhalb dieser Schichten liegen. Ziel der emotionalen Adaptivität der Interaktionsschichten ist es, emotional negative Einflüsse zu verhindern, um so Dialogunterbrechungen zu reduzieren und die Nutzerzufriedenheit zu steigern. Jenachdem, ob die Eingabe oder Ausgabe des Systems ursächlich ist, können die Systemanpassungen stark voneinander abweichen. Um einige Beispiele für emotionale Anpassung zu geben, werden diese an einem vereinheitlichenden Szenario des Autofahrens vorgestellt. Auf der einen Seite erlaubt

ein solches Szenario vielfältige Interaktionen mit technischen Systemen, andererseits kann die Umgebung leicht mit diversen Sensoren zur Emotionserkennung ausgestattet werden. In den folgenden Abschnitten werden konkrete Beispiele von Interaktionssituationen mit emotionaler Adaptivität im Szenario aufgezeigt.

Physikalische Interaktionsebene: Automatische Lautstärken Kontrolle. Die meisten modernen Fahrzeuge verfügen über ein eingebautes geschwindigkeitsabhängiges Lautstärken-Steuerungssystem. (Bekannt als: SDVC, SCV, GAL or GALA). Mit dieser Funktion wird die Lautstärke des Audiosystems an die aktuelle Geschwindigkeit des Fahrzeugs angepasst. Auf diese Weise müssen die Passagiere die Lautstärke nicht manuell einstellen, wenn das Fahrgeräusch mit der Geschwindigkeit zunimmt. In Fällen, in denen die Geschwindigkeit abrupt sinkt kann die Lautstärke merklich sinken (vgl. Aktion zur automatischen Lautstärkeanpassung bei t_6. Einige Fahrer schaffen es nicht die Einstellungen für diese Anpassung fein abzustimmen. Daher ist es möglich, dass die automatische Lautstärkeregelung in manchen Situationen nicht mit der Erwartung des Fahrers übereinstimmt. Ein negativer emotionaler Ausdruck kann durch eine Kamera mit Blick auf das Gesicht des Fahrers erkannt werden. Hochfrequente Merkmale wie LBP können dies erfassen. Wegen möglicher Veränderungen in der Umgebungslautstärke steht der Audiokanal eventuell nicht als Eingabemodalität zur Verfügung. Da der Fahrer in ständigem Kontakt mit dem Lenkrad steht, ist es möglich zusätzliche Informationen über den SCL zu erhalten, der sensibel für schnelle Veränderungen in Richtung negativer Empfindungen ist. Entscheidungslevel-Fusion ist in der Lage, die Gesichts- und SCL-Informationen zu einer endgültige Schätzung des Auftretens von negativen emotionalen Ereignissen zu vereinigen.

Ein zeitbasiertes Nachschlagen in der Interaktionshistorie zeigt die automatische Lautstärke Anpassung als Ursache der negativen Emotion (vgl. Prozess in Abbildung 5.13). Diese Feststellung wird an das verknüpfte Aktionselement, die aktive Komponente der automatischen Lautstärkeregelung, übergeben. Auf Grundlage dieses Feedbacks kann die Lautstärke neu eingestellt werden.

Multiple Ebenen: Ansage von Stauwarnungen. Angenommen wird eine Situation im Auto, in der das Navigationssystem die aktuellen Verkehrswarnungen ankündigen will (vgl. Aktion bei t_2 auf der *Funktionskern und Planungs* Schicht, Abbildung 5.12). Die Aktion wird als Teil der Interaktionshistorie im Speicher gehalten. Das Navigationssystem weis durch seine Implementierung, dass Warnungen vor bevorstehenden Staus möglicherweise negative Emotionen verursachen. Deshalb gibt dieses Aktionselement optionale Information über eine erwartete negative Emotion an seine Unterschichten weiter.

Der durch das Navigationssystem aufgerufe Dialogmanager wendet seine Strategie für einen solchen Fall an, er bereitet die Ausgabe einer Liste aller Verkehrsmeldungen vor. Danach wird

die Fissionskomponente mit einer Ausgabemodalitäten unabhängigen Liste der Verkehrsmeldungen aufgerufen. Als nächstes wird auf der physischen Ausgabeebene jede Warnung über verschiedene Ausgabekonzepte wie Text To Speech (TTS) oder eine Grafische Benutzeroberfläche (engl. Graphical User Interface) (GUI) ausgegeben. In diesem vereinfachten Beispiel muss die Fission zwischen den folgenden drei Ausgabeoptionen entscheiden: nur GUI, nur TTS oder GUI + TTS. Basierend auf dem Wissen über den aktuelle Kontext, bewertet das System die Optionen und entscheidet, dass die Verwendung von GUI + TTS optimal ist. Die TTS Ausgabe ersetzt die aktuelle Musikwiedergabe des Radios. In der GUI werden Metainformationen über Sänger und Lied gegen die Verkehrswarnungen ersetzt.

Direkt nach der akustischen Ansage, welche die Aufmerksamkeit des Fahrers geweckt hat, erkennt das System eine negative Emotion und datiert ihren Ursprung bei t_2. Durch die Reaktion des Fahrers, welche wieder durch die Kamera und biophysiologische Daten erfasst wurde, werden negative emotionale Veränderungen erkannt. Zusätzlich werden eventuelle verbale Äußerungen erfasst. Mittels MFN lassen sich alle sensorischen Werte zu einer Entscheidung fusionieren.

In diesem Beispiel führt eine Kette von kausalen Abhängigkeiten zur Identifizierung von vier mögliche Aktionselementen, angeordnet auf jeder Ebene. Da der Fahrer sich in der Vergangenheit nicht über die Verwendung von Display und Lautsprecher beschwert hat, kann eine emotionale Reaktion aufgrund der physischen Ausgabeebene ausgeschlossen werden. Auch die anderen Ebenen können so anhand kausaler Zusammenhänge ausgeschlossen werden. Somit bleibt nur der Inhalt der Ausgabe als auslösendes Element übrig. Da dieser außerhalb des Einflussbereichs des Systems liegt wird das Systemverhalten nicht adaptiert.

Interaktionseingabeebene: Fehlerhafte Spracherkennung. Eine typische Aufgabe in einem Autofahrszenario ist die Spezifikation eines Ziels, für das moderne Autos Spracheingabe anbieten. Um sicherzustellen, dass die Spracherkennung korrekt ist, muss jede Spracheingabe dabei ausdrücklich bestätigt werden. Mit einem emotionsbewussten System könnten diese mühsamen Dialogschritte leicht vermieden werden, da keine emotionale Reaktion implizit die vorherige Eingabe bestätigt. Explizite Anfragen sind nur notwendig, wenn eine negative Emotion erkannt wird, wie im folgenden beschrieben.

Angenommen eine Situation, in der der Nutzer „Newark“ sagt, um ein Ziel einzugeben. Der Spracherkenner missversteht „New York“ als wahrscheinlichste Eingabe und leitet sie an die oberen Schichten im Komponenten-Stack weiter (vgl. Aktionen t_1 in Abbildung 5.12). Sobald das erkannte (aber falsche) Ziel im Navigationsdisplay angezeigt wird (vgl. T_2) zeigt der Nutzer eine negative Emotion in Form von Verwirrung und / oder Ärger.

Überraschungen verursachen meist Gesichtsreaktionen der Augenbrauen, die mit FACS erkannt

werden können. Die verbale Reaktion eines Fahrers wie „oh“ ist möglich, was mittels MFCC oder LPC Merkmalen erkennbar ist. Im Falle einer Interaktion mit dem System während der Fahrt berührt der Fahrer zusätzlich das Lenkrad und stellt biophysiologische Informationen zur Verfügung. Aufgrund der unklaren zu erwartenden Modalität in der eine Reaktion messbar ist, ist ein MFN eine geeignete Fusionsarchitektur. In diesem Fall ist eine Überlagerung der Emotionen Ärger und Überraschung denkbar. Dies kann entweder durch zwei separate Klassifikatoren realisiert werden, oder über ein speziell auf diese Kombination trainiertes Netz erfolgen.

Da vorherige Spracheingaben wie erwartet funktioniert haben, kann der Grund für die Emotion der physischen Interaktionsschicht, d.h. dem Spracherkenner zugeordnet werden. Das System kennt nun das Missverständnis und zeigt die möglichen Alternativen (in der Reihenfolge ihrer Erkennungskonfidenzen) oder eine Option zum Wiederholen der Eingabe an. Da sich das richtige Ziel in der Liste der Alternativen befindet, kann der Nutzer es auswählen und die Navigation starten.

Interaktionseingabeebene: Sicherheitsgurt-Erinnerung, falsch positiv. Angenommen der Fahrer legt einen schweren Gegenstand (z.B. eine schwere Reisetasche) auf einen der Rücksitze, könnte dies vom Insassendetektionssystem als eine Person, die dort sitzt, fehlinterpretiert werden. So würde die Sicherheitsgurt-Erinnerung einen Event auslösen, welches zu einem nervigen Warnsignal führt (t_1 und t_2 in Abbildung 5.12). Dies kann zunächst zu einer emotionalen Reaktion der Überraschung später sogar zu Verärgerung führen. Diese können auf ähnliche Weise wie die zuvor beschrieben Reaktionen im Navigationsbeispiel erkannt werden.

Da die Interaktionshistorie nur eine Aktionen enthält, die keine Variationen in der Handhabung einer solchen Warnung erlaubt, ist die einzige verbleibende Möglichkeit die Eingabe des Sitzbelegungssystems als falsch positiv zuzuordnen. Daher könnte das System den Nutzer fragen ob die Warnmeldung vorübergehend deaktiviert werden soll, sofern der Sitz nicht von einem Menschen besetzt ist.

Interaktionseingabeebene: Mediensteuerung, falsch negativ. Die Rücksitz-Unterhaltung ermöglicht eine bequeme multimodale Interaktion mittels Sprachbefehlen und Zeigegesten um den Wiedergabezustand von Medien zu steuern. Erkennt das System den Blick auf dem Medienbildschirm, können Sprachbefehle verwendet werden, um den aktuellen Film oder die aktuelle Audiospur abzuspielen oder anzuhalten. So können mühsame Berührungsinteraktionen vermieden werden, was besonders hilfreich für Kinder ist, die Schwierigkeiten haben den Bildschirm zu erreichen, wenn der Sicherheitsgurt befestigt ist. Nehmen wir eine Situation an, in der das Kind Musik hört und die Landschaft am Seitenfenster vorbeizieht. Jetzt möchte es die Musik stoppen, um die Mitreisenden nach dem Namen der Region zu fragen, die sie

durchqueren. Dazu blickt es auf den Medienbildschirm und gibt den Befehl „Stop“ aus. In einer solchen Situation kann es leicht passieren, dass der Sprachbefehl zu leise ist, um eine Spracheingabe auszulösen. Ein sogenanntes falsches negatives Ereignis. Weil nichts passiert, reagiert das Kind mit einem Gefühl der Ungeduld, das durch eine einfache Kamera die im Rahmen des Medienbildschirms platziert wurde, erkannt werden kann. Zusätzliche akustische Informationen, die für Spracherkennung kaum zu erfassen sind, können dennoch eventuell weniger konfidente emotionale Zustandsinformationen tragen. Darüber hinaus sind Reaktionen wie „hmm“ als Reaktion der Ungeduld denkbar. Aufgrund der fehlenden intelligenten Oberflächen im Bereich der Rücksitze, die biophysiologische Daten erfassen können, muss sich die Erkennung auf den visuellen und unsicheren akustischen Kanal stützen. Ungeduld benötigt eventuell mehrere Sekunden um aufzutreten, daher sind Merkmale mit Fenstergrößen bis zu einer Sekunde verwendbar, was alle Video- und Audiomerkmale einschließt. Die Dimensionalität von Videomerkmalen sollte durch die Verwendung von PCA oder LDA-Methoden reduziert werden. Aufgrund der ähnlichen Fensterlängen von Audio- und Videomerkmalen kann in diesem Sonderfall auch eine frühe Fusion in Betracht gezogen werden um Ungeduld zu erfassen.

In dieser besonderen Situation der komplementären Eingabe gemäß den CARE-Eigenschaften [51] kann die Sprachmodalität (Eingabegerätekomponente) eindeutig für eine falsche negative Situation verantwortlich gemacht werden. Auf diese Weise wird eine Evidenz für ein falsch negatives Ereignis der Eingabeschicht erzeugt, zusätzlich zum Eintrag t_7 in der Interaktionshistorie.

Als unmittelbare Reaktion entschuldigt sich das System für den Fehler und fordert freundlichst dazu auf, die nicht erkannte Eingabe zu wiederholen. Um solche falsche negativen Ereignisse in Zukunft zu vermeiden, kann das System die Empfindlichkeit der betroffenen Mikrofone erhöhen. Diese Anpassung birgt jedoch das Risiko anderer Fehler, wie falsch positiven Fehleingaben. Wie in [221] ausgeführt kann dies allerdings mittels typischen zeitlichen Relationen der beteiligten Modalitäten erkannt und somit verhindert werden.

KAPITEL 6

DISKUSSION

Nach erfolgreicher Betrachtung aller grundlegenden Voraussetzungen zur Bereitstellung von Erkennerkomponenten der Companion-Technologie sollen zunächst die Erkenntnisse aus den vier Teilbereichen der Arbeit (Kapitel 2-5) diskutiert werden, gefolgt von einer ganzheitlichen und zusammenhängenden Betrachtung der Arbeit auf Basis der gesammelten Erfahrungen. Im Fokus steht dabei immer die praktische Einsetzbarkeit der vorgeschlagenen Methoden und der erzielten Ergebnisse.

6.1. Multimodale Datenbasen

Festzustellen ist, dass der elementare Einfluss der Trainingsdaten auf einen praktischen Einsatz bisher zumeist unterschätzt wurde. Eine Erkenntnis dieser Arbeit ist, dass prototypische Realisierungen von Companion-Technologie-Komponenten nur dann erfolgreich waren, wenn die eingesetzten Erkennerkomponenten mit Daten trainiert wurden, welche möglichst ähnlich zum späteren Einsatzszenario waren[1]. In der Literatur wurden bisher zumeist Datensätze betrachtet, welche zwar emotionalen Inhalt haben der teilweise auch von guter Qualität ist, jedoch ist das Szenario zur Emotionsinduktion, sofern überhaupt vorhanden, nicht im Kontext der Companion-Technologie gelegen. Aufgrund unzureichender Emotionsmodelle (Klassisch statt MMI-Relevant) und Induktionsparadigmen (Passiv statt situativ induziert) bleiben diese Datensätze allerdings in ihrer Verwendbarkeit begrenzt auf methodische Erforschungen von Merkmalen und Klassifikatoren.

Wie bereits mehrfach erwähnt, spielen die bereitgestellten Trainingsdaten eine entscheidende Rolle für die erfolgreiche Realisierung der MMI-Schnittstelle eines Companion-Systems. Worin aber bestanden die Herausforderungen bei der Datenerhebung? Problematisch war die

[1] Die ersten Demonstratoren mussten mit exakt der Person trainiert werden, welche später in der Demonstration als Nutzer fungierte. Auch die räumliche Umgebung musste im Training exakt der Umgebung der Demonstration entsprechen.

Bereitstellung von emotionalen Modellen und somit entsprechenden Labels im MMI-Kontext. Auf Basis fundierter psychologischer Betrachtungen existieren bis heute keine umfassenden und akzeptablen theoriebasierten Emotionsmodelle für die MMI. Dies zeigt sich zum einen anhand der Literatur, in der sich keine vergleichbar akzeptierten und standardisierten Modelle wie die Basisemotionen oder das VAD-Modell für den Bereich der MMI finden lassen. Auch konnte in vielen Jahren der Zusammenarbeit mit anderen Fachdisziplinen, speziell der Psychologie, im Rahmen des SFB-TRR 62 kein einheitliches und breit akzeptiertes Modell der MMI Emotionen publiziert werden. Die Ingenieursdisziplinen und Informatik haben daher einen wesentlich pragmatischeren datengetriebenen Ansatz gewählt. Dabei wurden zahlreiche Vorstudien auf Basis kleinerer WoZ Experimente durchgeführt. Diese Daten wurden zum einen manuell untersucht und relevante Ereignisse identifiziert, zum anderen wurden diese Daten mittels unüberwachter und semiüberwachter Datamining-Verfahren (z. B. Clustering und aktives Lernen wie in Abschnitt 3.3) analysiert. Konnten durch Einsatz solcher Verfahren interessante Instanzen identifiziert werden, wurden diese weiter verfolgt und auf eine Nutzbarkeit und Mehrwert in Anwendungen bewertet. So wurden auch die Begriffe der *Überforderung* und *Unterforderung* als relevante Elemente der MMI identifiziert, worauf dann der uulmMAC konzipiert wurde. Das daraus trainierte und realisierte System wurde erfolgreich eingesetzt, eine breite Akzeptanz dieser datengetriebenen Kategorien durch Experten der Psychologie ist bis heute allerdings nicht gegeben (Einen Widerspruch allerdings auch nicht). Da die Technologie schnell voranschreitet und immer mehr Produkte entstehen, welche Elemente der Companion-Technologie einsetzen, die psychologische Erforschung damit allerdings bisher nicht schritthalten kann, ist der pragmatische datengetriebene Ansatz derzeit am vielversprechendsten. Durch Einsatz modularer Architekturen wie der MAR^{2}S können Pretotypen und Prototypen oder WoZ Studien mit geringem Aufwand aufgebaut werden, diese lassen sich dann in (Vor-)Studien einsetzen.

Die Möglichkeit erste Komponenten der Companion-Technologie realisieren zu können bietet zudem die Chance geeignetere Trainingsdaten aus der Anwendungsdomäne zu gewinnen. Ideale Trainingsdaten entstammen genau der Einsatzdomäne, in der ein Klassifikator später funktionieren muss. Es ist also wünschenswert, Daten in realen Situationen vollständig außerhalb des Labors zu erheben. Dies wird im allgemeinen als *in the wild* bezeichnet. Frühe Companion-Systeme, die eventuell noch nicht alle Eigenschaften eines Companion-Systems erfüllen, diesen aber nahe kommen, stellen die Anwendungsgrundlage dar. Die sensorische Erfassung erfolgt dabei durch allgegenwärtige Sensoren. Dazu zählen auf den Nutzer gerichtete Kameras und Mikrophone wie sie in unseren Smartphones und Laptops bereits eingebaut sind, Tiefenkameras und Infrarotsensoren unserer Spielekonsolen im heimischen Wohnzimmer, diverse Fitness- und Gesundheitsarmbänder wie zum Beispiel das *E4* Band der Firma *Empatika*, welche als sogenannte *Waerables* immer weitere Verbreitung finden, Fahrerüberwachungs-

systeme im Automobilsektor oder zahlreiche weitere uns täglich umgebende elektronische Geräte.

6.2. Annotation multimodaler Daten

Gewonnene Daten können mit Werkzeugen wie ATLAS analysiert und annotiert werden. ATLAS ist nach bestem Wissen des Autors bisher das einzige Annotationswerkzeug, welches die Anforderungen der Multimodalität vollumfänglich unterstützt, ohne dabei die Annotation und Analysemöglichkeiten in Form von vorgegebenen Schemata zu beschränken. Dadurch ist ATLAS ideal geeignet moderne MMI-Datensätze zu analysieren und zu annotieren. Es ist zudem zukunftssicher entwickelt, sodass Anpassungen an zu erwartende Datensätze schnell realisierbar sind, sofern überhaupt notwendig. Zukünftige Datensätze werden sich dabei dahingehend entwickeln, dass mehr Modalitäten erfasst werden und diese zudem in größerem zeitlichen Umfang bereitstehen werden. Neben sensorischen Komponenten, welche Parameter des Menschen direkt erfassen (wie Mimik oder Biophysiologie), ist zu erwarten das zukünftig auch sensorische Messungen smarter Interaktionsobjekte (wie analoge Touch-Oberflächen oder spezialisierte Eingabegeräte) Teil der Datensätze werden und zur besseren Erfassung des emotionalen Zustandes des Menschen beitragen werden. Die Ausdehnung des zeitlichen Umfangs der Daten wird durch immer kleinere und omnipräsente Sensoren erreicht werden, wodurch langzeit datenerfassende Szenarien in weitgehend natürlicher Umgebung nicht länger in unerreichbarer Ferne liegen werden.

Aktive Lernverfahren reduzieren dabei den notwendigen Aufwand. Ohne die hier vorgestellten Ergebnisse der Methoden aktiver Lernverfahren für multimodale affektive Daten (Abschnitt 3.3) sind aktuelle Datensätze wie der uulmMAC nur schwer beherrschbar. Zukünftige Datensätze werden aufgrund ihres Umfangs ohne solche Verfahren überhaupt nicht mehr beherrschbar sein. Dies hat auch die Community des *affective Computing* erkannt. Seit Vorstellung und internationaler Diskussion der ersten Ergebnisse dieser Arbeit finden sich in der Literatur zunehmend weitere Arbeiten, die Ideen des aktiven Lernens in verschiedenster Form aufgreifen. Problematisch und weitgehend ungelöst ist nach wie vor der Umgang mit großen Anzahlen von Modalitäten. Insbesondere setzt der hier vorgestellte Ansatz voraus, dass ein Orakel (hier der Mensch) immer in der Lage ist, eine durch den Selektor bestimmte kurze Zeitspanne zu annotieren. Für die verwendeten Modalitäten Audio und Video, die wir Menschen täglich wahrnehmen, ist dies wie vorgestellt möglich. Unklar ist der Umgang mit im Alltag weniger geläufigen Parametern wie den biophysiologischen Kanälen, oder gar den zukünftig zu erwartenden Daten smarter Oberflächen. Hier ist es selbst für Experten nicht oder nur schwer möglich, eine eindeutige Abbildung von sensorischen Werten auf emotionale Kategorien zu

bestimmen. Allerdings zeigen die Erkennungsresultate, dass diese Daten zu wertvoll sind um ignoriert zu werden. Es bleibt also die Aufgabe eine Lösung zu finden, welche es ermöglicht u. A. biophysiologische Daten mit den vorgestellten Verfahren zu annotieren. Denkbar sind dabei Mischformen aus einer direkten Annotation dieser Modalität (z. B. erkennen von Sprüngen im SCL oder Veränderungen im EKG) gepaart mit Annotationen intuitiven Audio- und Videokanälen und situativem Kontextwissen. So können vermutlich Labels mit guter Qualität durch Methoden indirekter Annotation gewonnen werden.

6.3. Maschinelle Emotionserkennung

Betrachtet man die zahlreichen Publikationen, welche in den letzten Jahren im Themengebiet des *affective Computing* vorgestellt wurden, lässt sich erkennen, dass zunächst schnell, durch Optimierung bestehender Methoden und neu entwickelte Methoden, eine Verbesserung der Resultate erreicht werden konnte. So ist es mittlerweile möglich, emotionale Zustände nicht nur auf überexpressiv geschauspielerten Daten zu erkennen, sondern auch anhand von mittelmäßig expressiven induzierten Daten lassen sich Erkennungserfolge erzielen. Allerdings ist auch zu erkennen, dass die Erkennungsraten trotz immer ausgefeilterer Methoden immer noch auf einem eher schlechten Niveau liegen, verglichen mit vielen anderen Anwendungen der Mustererkennung. So werden heute noch eine Vielzahl von Arbeiten präsentiert, deren Ergebnisse zwar über Zufallsniveau liegen, jedoch weit von üblichen Erkennungsraten von deutlich über neunzig Prozent entfernt sind. Die besten erreichbaren Ergebnisse werden im Falle von vollständig auf den jeweiligen Datensatz spezialisierten und auf den Probanden induviualsierten multimodalen Erkennern erreicht. Selbst diese Ergebnisse sind dabei immer noch nicht zufriedenstellend. Zudem lassen sich diese Klassifikatoren bisher in keiner Weise auf andere Individuen oder Situationen übertragen ohne dass die Erkennungsleistung dabei dramatisch einbricht.

Die Erfahrungen aus dieser Arbeit lassen daran zweifeln, dass zukünftige Verbesserungen der Merkmals- und Klassifikationsmethoden alleine ausreichend sein werden, um die notwendige Steigerung der bisherigen Erkennungsniveaus zu ermöglichen. Die Resultate welche in Kapitel 4 vorgestellt wurden, stützen diese Behauptung, da in diesem Kapitel verschiedenste Merkmale und Klassifikatoren in uni- und multimodaler Form verwendet wurden, die Resultate allerdings alle auf ähnlichem Niveau liegen. Der Klassifikator (inkl. Merkmale) ist nicht die alles entscheidende Komponente. Auch weitere, hier nicht beschriebene Experimente, Untersuchungen und Studien deuten darauf hin, dass die Klassifikationsmethode allein nicht ausschlaggebend ist. Die Klassifikationsmethode sollte eher auf Basis einer konkreten Situation gewählt werden. Spielt der zeitliche Charakter eines Signals eine wesentliche Rolle sind Methoden wie MFN einer SVM vorzuziehen. Verwenden zwei Merkmale stark unterschiedliche

Analysefenster, ist eine frühe Fusion schwer realisierbar. Resultierend aus den Erfahrungen der Arbeit wird empfohlen, die Klassifikatorarchitektur auf Basis solcher, eher äußerer (teilweise rein technischer) Einflüsse, zu wählen, anstatt eine Methode zu erzwingen, welche laut Literatur auf einem ähnlichen aber dennoch verschiedenen Datensatz geringfügig bessere Resultate gezeigt hat. Der Methodenkatalog, aus dem gewählt werden kann, ist umfangreich.

Welche Möglichkeiten bestehen nun zur Verbesserung der Resultate? Zunächst ist dabei zu prüfen, ob eine Verbesserung zwingend notwendig ist, oder ob es sinnvoller ist, Companion-Technologien derart zu entwickeln, dass sie mit diesen Resultaten funktionieren können. Dass dies möglich ist wurde durch die vorgestellten Prototypen und Demonstratoren gezeigt. In Abschnitt 4.4.4 wurde gezeigt, dass die Hinzunahme von Kontextwissen eine Verbesserung der Resultate ermöglicht. Ein solcher Kontext besteht aus mindestens zwei Teilbereichen. Einem Anwendungskontext, wie er in dieser Arbeit exemplarisch verwendet wurde, und einem situativen Kontext durch Interpretation der Umwelt. Ein Anwendungskontext kann bei jeder Realisierung eines Companion-Systems abgeleitet werden. Er sollte daher bei der Klassifikation bzw. Interpretation der Klassifikationsresultate immer berücksichtigt werden. Erzeugt wird der Kontext durch die Anwendungskomponente und wird in Form von symbolischer Information in der Klassifikation berücksichtigt. Inwieweit ein situativer Umweltkontext erfasst werden kann ist sicherlich von Fall zu Fall unterschiedlich. Kann ein situativer Umweltkontext erfasst werden, sollte dieser allerdings unbedingt berücksichtigt werden. Er könnte zum Beispiel aus trivialen Informationen wie der Wetterlage und dem persönlichen Kalender des Nutzers abgeleitet werden (Der Nutzer wollte unbedingt seinem Hobby im Freien nachgehen, aber es regnet in Strömen). Auch der Einbezug komplexerer Zusammenhänge, aus denen zum Beispiel eine zu erwartende grundlegende Stimmung abgeleitet werden kann, ist denkbar (Der Nutzer wurde die ganze Nacht von seinen erkrankten Kindern wach gehalten, zusätzlich steht ein wichtiger beruflicher Termin an.). Es ist zu vermuten, dass sich derart starke Kontexte konstruieren lassen, die in manchen Situationen eine sensorische Zustandserfassung nahezu überflüssig machen. Eine praktische Überprüfung steht bisher jedoch aus.

6.4. Prototypische Companion-Systeme

Die Grundlagen für die erfolgreiche Realisierung eines Companion-Systems werden bereits in der frühen Phase der Konzeption einer Trainingsdatenerhebung gelegt, wenngleich zu diesem sehr frühen Zeitpunkt eine prototypische Realisierung eher visionären Charakter hat und nur sehr unscharf definiert ist. Aufgrund dessen sind zum Zeitpunkt der Trainingsdatenerhebung Details späterer Anforderungen nur schwer zu spezifizieren und es fehlen wesentliche Erfahrungen. Um diese Defizite zu reduzieren ist es zu empfehlen, so früh wie möglich mit kleinen

Vorstudien und einfachen Prototypen Erfahrungen und Information über zukünftig auftretende Anforderungen zu sammeln. Damit dabei das eigentliche Ziel nicht aus den Augen verloren wird, sollten Software- und System-Engineering Methoden eingesetzt werden. Die Konzeption von Trainingsdatensätzen und prototypischen Systemen erfolgt dabei parallel in gegenseitiger Wechselwirkung. Erkenntnisse der verschiedenen Entwicklungs- und Analysestufen des Systementwurfs können dabei eingesetzt werden, um Studien und Prototypen zu verbessern und so neue und verbesserte Trainingsdaten zu erheben. Eine solche iterative Herangehensweise resultiert schließlich in einem guten System, welches auf die Bedürfnisse seiner Nutzer angepasst ist.

Problematisch hat sich bei der Realisierung der hier vorgestellten prototypischen Companion-System die hohe thematische Komplexität herausgestellt, welche sich unter anderem aus der Zusammenarbeit der unterschiedlichen Fachdisziplinen ergibt. Durch stetige Weiterentwicklung zunächst sehr rudimentärer, wenige Funktionen umfassender Systeme bzw. Komponenten wird es ermöglicht, dass bereits zu sehr frühen Phasen eine einheitliche Sicht aller beteiligten Partner auf das System ermöglicht wird. Durch konsequente Weiterentwicklung dieser anfänglichen Systeme entlang sorgfältig definierter Anforderungen wird die Systementwicklung Schritt für Schritt voran getrieben, bis letztlich ein finaler Prototyp entsteht, welcher alle Anforderungen realisiert. Dabei sind Vorgensmethoden der agilen Entwicklung wie z.B. SCRUM ein geeignetes Mittel um den Entwicklungsprozess zu steuern. Dies erfordert ein hohes Maß an Modularität im System-Entwicklungsprozess, was durch die vorgestellte MAR^{2}S-Technik ermöglicht wird.

6.5. Ganzheitliche und zusammenhängende Betrachtung der Teilbereiche

Um ein Companion-System zu realisieren ist die Zusammenarbeit der Fachdisziplinen aus dem Bereich der Psychologie und den Bereichen der Informatik und Ingenieurswissenschaften notwendig. Wie die Erstellung dieser Arbeit gezeigt hat liegt hierin eine besondere Herausforderung. Neben dem unterschiedlichen Vokabular und daraus resultierenden kommunikativen Herausforderungen, unterschieden sich diese Disziplinen teilweise erheblich in ihrer Sichtweise auf die Vision eines Companion-Systems und in ihrer Herangehensweise an das Ziel eines funktionsfähigen Companion-Systems. Während das Ingenieurswesen und die Informatik in ihrer Herangehensweise die typischen technischen Methoden der datengetriebenen Analyse kleiner Feldstudien und Vorgehensmodelle des *top-down*, *bottom-up* bevorzugt, setzt die Psychologie auf die statistische Auswertung von umfangreichen Laborstudien mit möglichst geringen Freiheitsgraden innerhalb der Studien, da die Umsetzung solcher psychologischen Studien

meist enorm zeitintensiv ist und ihr Resultat oft in der fundierten Beantwortung einer einzigen kleinteiligen Fragestellung liegt. Die Feldstudien der Informatik finden meist in weniger kontrollierter Umgebung, teilweise außerhalb des Labors, statt. Dies bedingt mehr Freiheitsgrade, und die Analyse der Studie erfolgt datengetrieben und explorativ. Diese unterschiedlichen Sicht- und Vorgehensweisen der Disziplinen, welche erst seit wenigen Jahren durch die gemeinsame Erforschung von MMI Schnittstellen zusammenarbeiten, sollten stets berücksichtigt werden.

Bei der Entwicklung der emotionserkennenden Komponente eines Companion-Systems dürfen die vorgestellten vier Teilbereiche nicht isoliert betrachtet werden. Vielmehr müssen die Wechselwirkungen der Bereiche berücksichtigt werden. Es besteht eine enge Kopplung der Datenerhebung und der Annotation. So setzt z. B. die vorgestellte Annotation mittels aktiver Lernverfahren voraus, dass auftretende Emotionen eher spärlich sind. Nur so kann sichergestellt werden, dass das automatisch generierte Hintergrundmodell (Neutralmodell) zuverlässig erzeugt werden kann. Diese Bedingung ist bei natürlichen MMI Datensätzen gegeben, nicht aber bei früheren expressiveren oder geschauspielerten Korpora. Aus diesen Gründen muss das Datensatzdesign und die Annotationsmethodik aufeinander abgestimmt sein.

Die Qualität der annotierten Daten und ihrer Labels hat unmittelbaren Einfluss auf das Training von Klassifikatoren und deren maximal erreichbarer Performanz, sowohl im individuellen als auch im interindividuellen Fall. Beim Design der Klassifkationsarchitektur muss also darauf geachtet werden, dass diese mit eventuellen Fehlern im Datenerhebungs- und Annotationsprozess zurecht kommt. Dies bedingt meist den Einsatz von Klassifikatoren, welche mit konfidenzbehafteten Eingaben (oder fuzzy Labels) aus niedrigeren Verarbeitungsebenen umgehen müssen. Die Ausgabe einer hierarchischen Klassifikation ist im Bereich der Emotionserkennung mit einer nicht unerheblichen Fehlerrate behaftet. Ein Companion-System muss daher immer so gestaltet werden, dass es in der Lage ist mit einer erhöhten Fehlerrate angemessen umgehen zu können. Dies kann mittels verschiedener Herangehensweisen erreicht werden. Z. B. sollte immer die Möglichkeit eingeplant werden, in kritischen Situationen beim Nutzer rückzufragen. Alternativ kann auch eine Absicherung durch Einbezug von situativem Kontextwissen des Systems erfolgen. Dabei können so gewonnene Informationen in einem real eingesetzten System durchaus dazu verwendet werden, den ursprünglichen Datensatz zu erweitern und so mittels erneutem Durchlauf der Verarbeitungskette den Klassifikator zu adaptieren, um zukünftig bessere individualisierte Ergebnisse zu erhalten. Derartige Systeme unterscheiden sich dabei von bisherigen technischen Systemen dadurch, dass sie sich im Laufe ihrer Einsatzzeit weiter entwickeln und verbessern können, im Gegensatz zu bisherigen Systemen mit fester Funktionalität.

Diese Rückkopplung zeigt sich auch bei der iterativen Vorgehensweise beim Systementwurf. Hierbei werden zu Beginn einfache Systeme zunächst rudimentär realisiert. Mittels dieser

Systeme werden anschließend Daten erhoben, annotiert und klassifiziert um einen initialen Klassifikator für die nächste Ausbaustufe des Sytsems zu erhalten. Mittels dieses Klassifikators wird dann der Systementwurf verfeinert und weitere Komponenten realisiert, welche dann wiederum die Grundlage für verbesserte Datenaufnahmeszenarien bilden. Es besteht also neben der direkten Kopplung der Teilbereiche auch noch eine Rückkopplung vom prototypischen System zur Datenerhebung. Systeme in den ersten Phasen eine Datenerhebung werden dabei meist als WoZ-Systeme realisierte, wobei die WoZ gesteuerten Komponenten mit jeder Entwicklungsstufe reduziert und durch tatsächliche funktionale Komponenten ersetzt werden. Durch den initialen Einsatz von WoZ Komponenten kann auch das *Henne-Ei Problem*[2] oder *Bootstrapping Problem*[3] gelöst werden, welches zunächst ein lauffähiges System voraussetzt um Daten erfassen zu können, welches wiederum auf erfassten Daten trainiert werden muss.

Ein Companion-System hat zum Ziel eine größtmögliche Akzeptanz bei einer größtmöglichen Anzahl von Nutzern zu erreichen. Diese kann nach bisherigem Kenntnisstand noch nicht garantiert werden. So ist es im Verlaufe dieser Arbeit immer wieder im Rahmen von Demonstrationen, Vorversuchen, Studien oder Datenerfassungen vorgekommen, dass ein weitgehend akzeptabel funktionierendes System bei einigen Personen eine überraschend schlechte Performanz zeigte. Die Gründe hierfür lagen dabei häufig darin, dass die erfassten sensorischen Werte nicht, oder nur schlecht, zu den erlernten Eingaben passten. Gründe hierfür können z. B. in physiologischen Eigenschaften der Personen liegen, z. B. extrem lederartiger Hautoberfläche durch jahrelange körperliche Arbeit im Freien, welche eine Abtastung von biophysiologischen Parametern nahezu unmöglich macht. Daraus folgt ein Totalausfall einer nicht unerheblichen Modalität, mit entsprechender Auswirkung auf das Gesamtresultat. Auch ist es vorgekommen, dass starke Gesichtsbehaarung (üppiger Vollbart) die visuelle Abtastung derart negativ beeinflusst hat, dass diese stark fehlerhafte Ergebnisse lieferte, was durch Fusion nicht in allen Fällen ausgeglichen werden konnte. Neben diesen physischen Eigenschaften zeigte sich zudem bei einigen Personen ein abweichendes emotionales Ausdrucksverhalten. So werden Companion-Systeme auch zukünftig bei der Erfassung stark introvertierter Personen erheblichen Herausforderungen gegenüber stehen. Aber auch stark extrovertiertes Verhalten stellt überraschender Weise ein Problem dar. So ist im uulmMAC mit Probandin 11 eine solch extrovertierte Person enthalten. Diese zeigt ein derart intensives emotionales Verhalten, dass die emotionale Spärlichkeitsannahme im Trainingsdatensatz teilweise verletzt wurde, dies wiederum beeinflusst die Annotation und somit die Klassifikation negativ. Zudem ist für ein Companion-System zunächst nicht klar wie es auf eine derart große Anzahl von emotionalen Ereignissen adäquat reagieren soll. Es bleibt festzustellen, dass die realisierten prototypischen Systeme erfolgreich auf einer umfang-

[2] `https://de.wikipedia.org/wiki/Henne-Ei-Problem` (Stand: 3. Juni 2019)
[3] `https://de.wikipedia.org/wiki/Bootstrapping_(Programmierung)` (Stand: 3. Juni 2019)

reichen Anzahl von Personen funktionieren, es allerdings auch einige wenige Totalausfälle zu verzeichnen gibt. Unklar ist dabei wie in einem solchen Fall vorgegangen werden muss, wenn prototypische Companion-Systeme zu serienreifen Companion-Systemen für den Massenmarkt weiterentwickelt werden.

6.6. Fazit

In dieser Arbeit wurde untersucht, welche Eigenschaften Trainingsdatensätze erfüllen müssen und wie diese effizient annotiert werden können, um daraus multimodale Klassifikatorarchitekturen zur Schätzung emotionaler Zustände zu entwerfen, welche in prototypischen Realisierungen von Companion-Systemen eingesetzt werden können. Der Schwerpunkt lag dabei auf Arbeiten zur Emotionserkennung im MMI Kontext als Bestandteil des Forschungsfeldes der Companion-Technologie. Eine Zusammenfassung der wesentlichen wissenschaftlichen Beiträge der vier Bereiche dieser Arbeit wurde bereits in Abschnitt 1.3 gegeben. Der Arbeit zugrundeliegende Publikationen sind dort den vorgestellten Teilgebieten zugeordnet. Im Anhang *Abstracts eigener Publikationen* sind diese Publikationen zudem mit ihren Abstracts aufgelistet. Diese Arbeit geht zudem an vielen Stellen über bereits veröffentlichte Inhalte hinaus.

Es wurde gezeigt, dass klassische Emotionsmodelle nicht direkt geeignet sind um in der MMI eingesetzt zu werden. Auch ist ein Transfer von Klassifikatoren, welche auf expressiven Daten klassischer emotionaler Korpora trainiert wurden, hin zum Einsatz auf realistischen Daten nicht möglich. Mit dem uulmMAC wurde daher ein Datensatz erhoben, der klassische, aus der Psychologie bekannte Emotionsmodelle um Kategorien der MMI Emotionen erweitert, und realistische schwach expressive Daten bereitstellt. Die Erweiterung der Emotionsmodelle ist dabei exemplarisch anhand von kognitiver Last durchgeführt worden. Für die zukünftige Weiterentwicklung von affektiven MMI Schnittstellen, welche Bestandteil von Companion-Systemen sind, sind sicherlich weitere Arbeiten notwendig welche diese Emotionsmodelle auf einer psychologisch fundierten Grundlage erweitern und verfeinern.

Die Annotation der gewonnen Daten erfolgte mittels entwickelter Verfahren zum aktiven Lernen in multimodalen Korpora. Vor dieser Arbeit war der Einsatz solcher Verfahren im Kontext der Emotionserkennung weitgehend unbekannt. Die hier vorgeschlagenen Verfahren konnten die Mächtigkeit des aktiven Lernens in großen Datenmengen hervorheben. Dabei ist die präsentierte methodische Entwicklung erst ein Anfang. Aus den vorgestellten Methoden und Resultaten lassen sich zahlreiche weitere Fragestellungen ableiten. Z.B.: Können komplexe Modalitäten den Annotationsprozess weiter optimieren? Wie kann der zeitliche Zusammenhang zwischen Modalitäten besser genutzt werden? Wie kann der situative Kontext, welcher nachweislich

die Klassifikation verbessern kann, ebenfalls bei der aktiven Annotation eingesetzt werden? Können Methoden des Transfer-Lernens helfen die initial schwachen Klassifikatoren des aktiven Lernens, durch Wissenstransfers aus anderen Situationen oder von anderen Probanden, zu verbessern?

Die vorgestellte MAR^{2}S Architektur erlaubt den modularen Aufbau sowohl von datenerzeugenden Szenarien, als auch die schnelle Erstellung prototypischer Companion-Systeme, ohne dass dabei ein Paradigmen- oder Entwicklungsumgebungswechsel notwendig wird. Dies ermöglicht eine schnelle Umsetzung weiterer wissenschaftlicher Experimente.

6.7. Ausblick

Daten können zukünftig immer einfacher mit genau den Sensoren erfasst werden, welche in späteren tatsächlichen Systemen verwendet werden, und in Umgebungen welche exakt dem späteren Anwendungsfeld entsprechen. Dies ist die Folge der Entwicklung in Richtung von *in the wild* Szenarien. Auch die genannten Wearable-Sensoren oder spezielle Sensoren aus dem automobilen Umfeld können so erfasst werden. Zudem liefern zahlreiche Eingabegeräte wie Touchsensoren oder Tasten immer häufiger druck- und geschwindigkeitsabhängige analoge Werte (Im Gegensatz zu klassischen binären Tasten). Auch diese Tastendruckdynamik kann emotionale Information tragen, z. B. durch unsicheres und zaghaftes drücken gegenüber selbstsicherem und kräftigem drücken einer Taste. Bisher sind noch keine solch umfangreichen Datensätze in ausreichender Form verfügbar. Sind diese zukünftig verfügbar, stellen sie eine hervorragende Ausgangslage zum Klassifikatortraining dar. Kritisch ist dabei nur die Annotationsfrage. Ein Teil der Daten wird sicherlich auch zukünftig manuell zu annotieren sein. Durch geringfügige Anpassungen der präsentierten aktiven Lernverfahren wird es möglich sein, den Sensordatenstrom kontinuierlich zu erfassen und dabei relevante Instanzen zu identifizieren. Kann das System diese nicht eindeutig zuordnen sind speziell in der Einlernphase eines Systems Rückfragen beim Nutzer denkbar, welcher dann als Orakel im aktiven Lernprozess fungiert. Initiale Klassifikatoren solcher *in the wild* aktiver Lernverfahren werden dabei durch Training auf Datensätzen wie dem uulmMAC bereitgestellt und ermöglichen einen Start der iterativen aktiven Lernverfahren auf hohem Erkennungsniveau, wodurch die Anzahl von *in the wild* Iterationen auf ein für den Nutzer akzeptables Maß reduziert wird.

Ein zukünftiges prototypisches Szenario könnte, ausgehend von den Ergebnissen dieser Arbeit, ein tutorielles System sein, welches sich mittels Erkennung der kognitiven Belastung eines lernenden Probanden in der Geschwindigkeit seiner Wissensvermittlung adaptiert. Tutorielle Systeme sollen den Nutzer in einer Lernumgebung beim Erreichen eines Lernziels unterstützen

und dabei auf seine individuellen Fähigkeiten in pädagogisch sinnvoller Weise eingehen. Hierfür sind umfassende Nutzermodellierungen erforderlich, welche die Interaktionshistorie und den augenblicklichen emotionalen Zustand des Nutzers berücksichtigen. Neben klassischen sensorischen Parametern (Audio, Video) sind dabei neurophysiologische Parameter (fMRI, EEG) und psychobiologische Merkmale (EKG, Hautleitwert, Atmung) von besonderem Interesse, um auftretende Änderungen emotionaler und kognitiver Nutzerzustände mittels multimodaler Datenanalyse abzuleiten. Daraus resultiert eine interdisziplinäre Analyse und Interpretation der Zusammenhänge zwischen den teils schwachen Signalen einzelner Modalitäten und die Entwicklung und Optimierung von Klassifikatoren zur multimodalen Emotionserkennung und affektiven aktiven Annotation, welche weitere zukünftige Forschungsfragen aufwerfen.

ANHANG A

IMPLEMENTIERTE KOMPONENTEN

Diese Arbeit umfasst neben theoretischen und methodischen Untersuchungen auch eine Vielzahl von implementierten Softwarekomponenten. Dabei handelt es sich unter anderem um das Annotationswerkzeug ATLAS, eine Vielzahl von MAR^{2}S Komponenten (MAR^{2}S-Module), sowie diverse Demonstratoren und Demonstratorkomponenten. Die technische Realisierung dieser Komponenten hat aufgrund ihrer zumeist hohen Komplexität, resultierend aus den speziellen Anforderungen, einen bedeuten Zeitaufwand dargestellt. In diesem Dokument wurde überwiegend auf die methodischen Grundlagen und erzielten experimentellen Resultate eingegangen. Die technische Dokumentation der implementierten Komponenten findet sich jeweils in der (Online)-Hilfe, Sourcecodedokumentation und den HowTo-Dokumenten der jeweiligen Komponente. Eine Vielzahl der realisierten Komponenten haben Verwendung in diversen weiteren Arbeitsgruppen gefunden. Aufgrund der freien Verfügbarkeit einiger Komponenten kann nicht in allen Fällen eine vollständiger Liste anderer Nutzer angegeben werden. Die folgende Auflistung fasst die Implementierungen dieser Arbeit mit ihren wesentlichen technischen Eigenschaften zusammen.

Annotationswerkzeuge

Bezeichnung:	**ATLAS**
Kurzbeschreibung:	Multimodales universelles Annotationswerkzeug ohne Beschränkung der Annotationsschemata (z.B. Emotionsmodelle) mit diversen Schnittstellen zur Anbindung aktiver Lernverfahren.
Beschrieben in:	Abschnitt 3.5
Programmiersprache:	Java 1.7 und 1.8
Codeumfang:	15.000 Zeilen
Klassenzahl:	119
Online Verfügbar:	`http://www.uni-ulm.de/in/neuroinformatik/mitarbeiter/s-meudt/` (Stand: 3. Juni 2019)
Weitere Verwendung:	SFB-TRR 62 Teilprojekte A1, A2, B3, B4, C4, Z3 SFB-TRR 62 Transferprojekt T3 SFB-TRR 62 Demonstrator Cluster 8 BMBF Projekt SENSE-Emotion Arbeitsgruppe Prof. Dr. H. C. Traue (Ulm) Arbeitsgruppe Dr. A. O. Andrade (UF Uberlandia) Arbeitsgruppe Dr. A. Brechmann (LIN Magdeburg) Arbeitsgruppe Jun.-Prof. Dr. C. Schrader (Ulm)

Bezeichnung:	**EmotionWheel**
Kurzbeschreibung:	Annotationswerkzeug zur kontinuierlichen und diskreten Annotation von Audio und Videodaten nach Plutchik
Beschrieben in:	Abschnitt 3.4
Programmiersprache:	Java 1.6
Codeumfang:	1.500 Zeilen
Klassenzahl:	21
Online Verfügbar:	`http://www.uni-ulm.de/in/neuroinformatik/mitarbeiter/s-meudt/` (Stand: 3. Juni 2019)
Weitere Verwendung:	Assistant Prof. Dr. S. Scherer (USC Kalifornien)

MAR²S Komponenten

Bezeichnung:	**GTecRecorder**
Kurzbeschreibung:	MAR²S Modul zur synchronen Aufzeichnung von biophysiologischen Daten mittels gTec mobiLab+ Bluetooth Signalverstärker
Beschrieben in:	Abschnitt 2.3.3
Programmiersprache:	C#
Codeumfang:	1.300 Zeilen
Online Verfügbar:	Auf Nachfrage beim Autor
Weitere Verwendung:	SFB-TRR 62 Teilprojekt Z3 (Ergebung GF Datensatz) BMBF ProjektSENSE-Emotion (Angepasste Version)

Bezeichnung:	**EnobioRecorder**
Kurzbeschreibung:	MAR²S Modul zur synchronen Aufzeichnung von biophysiologischen Daten mittels Neuroelectrics Enobio Bluetooth Signalverstärker
Programmiersprache:	C#
Codeumfang:	1.250 Zeilen
Online Verfügbar:	Auf Nachfrage beim Autor
Weitere Verwendung:	EU Projekt BRISC Arbeitsgruppe Prof. Dr. Dr. D. A. Braun (Ulm)

Bezeichnung:	**KinectRecorder**
Kurzbeschreibung:	MAR²S Modul zur synchronen Aufzeichnung von Audio-, Video und 3D Tiefendaten mittels Microsoft Kinect v2 Sensor
Beschrieben in:	Abschnitt 2.3.3
Programmiersprache:	C#
Codeumfang:	1.400 Zeilen
Online Verfügbar:	Auf Nachfrage beim Autor
Weitere Verwendung:	SFB-TRR 62 Teilprojekt Z3 (Ergebung GF Datensatz)

Bezeichnung:	**DelayRecorder**
Kurzbeschreibung:	MAR^{2}S Modul zur Erfassung der Netzwerk Latenz
Beschrieben in:	Abschnitt 2.3.3
Programmiersprache:	C#
Codeumfang:	900 Zeilen
Online Verfügbar:	Auf Nachfrage beim Autor
Weitere Verwendung:	Teilprojekt Z3 (Ergebung GF Datensatz) EU Projekt BRISC Arbeitsgruppe Prof. Dr. Dr. D. A. Braun (Ulm)

Bezeichnung:	**AudioVideoRecorderApp**
Kurzbeschreibung:	MAR^{2}S Modul zur synchronen Aufzeichnung von Audio und Videodaten mittels Webcam, realisiert als App für smarte Geräte mit eingebauten Kameras
Programmiersprache:	C#
Codeumfang:	1.600 Zeilen
Online Verfügbar:	Nein (Experimentelle Version)

Bezeichnung:	**SEMAINEtranslator**
Kurzbeschreibung:	MAR^{2}S Modul zur Anbindung alter, nicht SEMAINE tauglicher Komponenten an die MAR^{2}S Architektur
Programmiersprache:	C#
Codeumfang:	700 Zeilen
Online Verfügbar:	Nein (Sonderentwicklung für D. Braun et. al.)
Weitere Verwendung:	EU Projekt BRISC Arbeitsgruppe Prof. Dr. Dr. D. A. Braun (Ulm)

Demonstratoren und Demonstratormodule

Bezeichnung:	**EmoVoiceEstimator**
Kurzbeschreibung:	Modul zur Schätzung emotionaler Zustände anhand von sprachlichen Äußerungen in Echtzeit. SEMAINE angebunden, Vorläufer Architektur der MAR^{2}S
Beschrieben in:	Abschnitt 5.3.2
Programmiersprache:	Java 1.5
Codeumfang:	1.800 Zeilen
Online Verfügbar:	Nein (Sonderentwicklung für Demonstrator DS3)

Bezeichnung:	**FusionModule**
Kurzbeschreibung:	Modul zur multimodalen Fusion emotionaler Zustandsschätzungen in Echtzeit. SEMAINE angebunden, Vorläufer Architektur der MAR^{2}S
Programmiersprache:	Java 1.8
Codeumfang:	2.200 Zeilen
Online Verfügbar:	Nein (Sonderentwicklung für Demonstrator Cluster 8)

Bezeichnung:	**AL-Demo**
Kurzbeschreibung:	Komponente zur Visualisierung von Resultaten und Methoden aktiver Lernverfahren in ATLAS
Programmiersprache:	Matlab R15a Java 1.8 C\C++
Codeumfang:	1.800 Zeilen
Online Verfügbar:	Nein (Sonderentwicklung für Demonstrator C5)

Bezeichnung:	**UserIdentifier**
Kurzbeschreibung:	Komponente zur Audio/Video basierten verteilten Nutzeridentifikation und Positionsschätzung mehrerer Nutzer
Beschrieben in:	Abschnitt 5.3.1
Programmiersprache:	C\C++
Codeumfang:	4.000 Zeilen
Online Verfügbar:	Auf Nachfrage beim Autor

ANHANG B

ABSTRACTS EIGENER PUBLIKATIONEN

HELD, Dennis ; **MEUDT, Sascha** ; SCHWENKER, Friedhelm: Bimodal Recognition of Cognitive Load Based on Speech and Physiological Changes. In: *Proceedings of Multimodal pattern recognition of social signals in human computer interaction (MPRSS 2016)*, 2016, S. 1–8

Abstract: An essential component of the interaction between humans is the reaction through their emotional intelligence to emotional states of the counterpart and respond appropriately. This kind of action results in a successful interpersonal communication. The first step to achieve this goal within HCI is the identification of these emotional states.

This paper deals with the development of procedures and an automated classification system for recognition of mental overload and mental underload utilizing speech an physiological signals. Mental load states are induced through easy and tedious tasks for mental underload and complex and hard tasks for mental overload. It will be shown, how to select suitable features, build uni modal classifiers which then are combined to a bimodal mental load estimation by the use of early and late fusion. Additionally the impact of speech artifacts on physiological data is investigated.

HIHN, Heinke ; **MEUDT, Sascha** ; SCHWENKER, Friedhelm: Inferring Mental Overload based on Postural Behavior and Gestures. In: *Proceedings of Emotion Recognition and Modeling for Companin Technologie (ERM4CT 2016)*, 2016, S. 1–8

Abstract: Humans rely on body gestures and posture when communicating. This topic has been covered in great detail by researchers from various fields. Within this work, approaches to transfer findings in psychology and behavioral studies regarding the relation between gestures and emotions to machine learning methods, will be investigated. Knowledge about the users emotional state is important to achieve human like, natural HCI in modern technical

systems. The main focus lies on discriminating between mental overload and mental underload, when completing a given task, which for instance can be useful in an e-tutorial system. Mental underload is a new term used to describe the state a person is in when completing a dull or boring task. A suggestion how the affective states of overload and underload can be expressed using the established notation in the Valence, Arousal and Dominance (V,A,D) space will be given. In a further step it will be shown how to select suitable features, such as gestures, movement and postural behavior patterns. Based on the features selected, a classifier is designed and trained capable of deciding whether a person is experiencing mental overload or underload.

HIHN, Heinke ; **MEUDT, Sascha** ; SCHWENKER, Friedhelm: On Gesture and Postural Behavior as a Modality in Ensamble Methods. In: *Proceedings of Artificial Neural Networks in Pattern Recognition (ANNPR'16)*, 2016, S. 312–323

Abstract: Knowledge about the users emotional state is important to achieve human like, natural HCI in modern technical systems. Humans rely on body gestures and posture when communicating. We investigate the relation between gestures and human emotion, specifically when completing tasks. The main focus of this work lies on discriminating between mental overload and mental underload, which can e.g. be useful in an e-tutorial system. Mental underload is a new term used to describe the state a person is in when completing a dull or boring task. It will be shown how to select suited features, such as gestures, movement and postural behavior. Furthermore those features will be investigated regarding their discriminative power. After features are selected, a multiple classifier system will be designed, trained and evaluated.

KÄCHELE, Markus ; GLODEK, Michael ; ZHARKOV, Dimitrij ; **MEUDT, Sascha** ; SCHWENKER, Friedhelm: Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. In: MARSICO, Maria D. (Hrsg.) ; TABBONE, Antoine (Hrsg.) ; FRED, Ana (Hrsg.): *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM'14)*, SciTePress, 2014, S. 671–678

Abstract: Reliable prediction of affective states in real world scenarios is very challenging and a significant amount of ongoing research is targeted towards improvement of existing systems. Major problems include the unreliability of labels, variations of the same affective states amongst different persons and in different modalities as well as the presence of sensor

noise in the signals. This work presents a framework for adaptive fusion of input modalities incorporating variable degrees of certainty on different levels. Using a strategy that starts with ensembles of weak learners, gradually, level by level, the discriminative power of the system is improved by adaptively weighting favorable decisions, while concurrently dismissing unfavorable ones. For the final decision fusion the proposed system leverages a trained Kalman filter. Besides its ability to deal with missing and uncertain values, in its nature, the Kalman filter is a time series predictor and thus a suitable choice to match input signals to a reference time series in the form of ground truth labels. In the case of affect recognition, the proposed system exhibits superior performance in comparison to competing systems on the analysed dataset.

KÄCHELE, Markus ; SCHELS, Martin ; **MEUDT, Sascha** ; KESSLER, Viktor ; GLODEK, Michael ; THIAM, Patrick ; TSCHECHNE, Stephan ; PALM, Günther ; SCHWENKER, Friedhelm: On Annotation and Evaluation of Multi-Modal Corpora in Affective Human-Computer Interaction. In: MARSICO, Maria D. (Hrsg.) ; TABBONE, Antoine (Hrsg.) ; FRED, Ana (Hrsg.): *Proceedings of Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction (MA3HMI'14)*, SciTePress, 2014, S. 35–44

Abstract: In this paper, we discuss the topic of affective human-computer interaction from a data driven viewpoint. This comprises the collection of respective databases with emotional contents, feasible annotation procedures and software tools that are able to conduct a suitable labeling process. A further issue that is discussed in this paper is the evaluation of the results that are computed using statistical classifiers. Based on this we propose to use fuzzy memberships in order to model affective user state and endorse respective fuzzy performance measures.

KÄCHELE, Markus ; SCHELS, Martin ; **MEUDT, Sascha** ; PALM, Günther ; SCHWENKER, Friedhelm: Revisiting the EmotiW challenge: how wild is it really? In: *Journal on Multimodal User Interfaces (JMUI)* (2016), S. 1–12

Abstract: The focus of this work is emotion recognition in the wild based on a multitude of different audio, visual and meta features. For this, a method is proposed to optimize multi-modal fusion architectures based on evolutionary computing. Extensive uni- and multi-modal experiments show the discriminative power of each computed feature set and fusion

architecture. Furthermore, we summarize the EmotiW 2013/2014 challenges and review the conclusions that have been drawn and compare our results with the state-of-the-art on this dataset.

KINDSVATER, Daniel ; **MEUDT, Sascha** ; SCHWENKER, Friedhelm: Fusion Architectures for Multimodal Cognitive Load Recognition. In: *Proceedings of Multimodal pattern recognition of social signals in human computer interaction (MPRSS 2016)*, 2016, S. 1–8

Abstract: Knowledge about the users emotional state is important to achieve human like, natural Human Computer Interaction (HCI) in modern technical systems. Humans rely on implicit signals like body gestures and posture, vocal changes (e.g. pitch) and mimic expressions when communicating. We investigate the relation between them and human emotion, specifically when completing easy or difficult tasks. Additionally we include physiological data which also differ in changes of cognitive load. We focus on discriminating between mental overload and mental underload, which can e.g. be useful in an e-tutorial system. Mental underload is a new term used to describe the state a person is in when completing a dull or boring task. It will be shown how to select suited features, build uni modal classifiers which then are combined to a multimodal mental load estimation by the use of Markov Fusion Networks (MFN) and Kalman Filter Fusion (KFF).

MEUDT, Sascha ; BIGALKE, Lutz ; SCHWENKER, Friedhelm: ATLAS - Annotation tool using partially supervised learning and Multi-view Co-learning in Human-Computer-Interaction Scenarios. In: *Proceedings of the 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA'12)*, IEEE, 2012, S. 1309–1312

Abstract: In this paper we present ATLAS, a new graphical tool for annotation of multi-modal data streams. Although Atlas has been developed for data bases collected in human computer interaction (HCI) scenarios, it is applicable for multimodal time series in general settings. In our HCI scenario, besides multi-channel audio and video inputs, various bio-physiological data has been recorded, e.g. complex multi-variate signals such as ECG, EEG, EMG as well as simple uni-variate skin conductivity, respiration, blood volume pulse, etc. All these different types of data can be processed through ATLAS. In addition to processing raw data, intermediate data processing results, such as extracted features, and even (probabilistic or crisp) outputs of pre-trained classifier modules can be displayed. Furthermore, annotation and transcription

tools have been implemented. ATLAS's basic structure is briefly described. Besides these basic annotation features, active learning (active data selection) approaches have been included into the overall system. Support Vector Machines (SVM) utilizing probabilistic outputs are the current algorithms to select confident data. Confident classification results made by the SVM classifier support the human expert to investigate unlabeled parts of the data.

MEUDT, Sascha ; BIGALKE, Lutz ; SCHWENKER, Friedhelm: ATLAS – An Annotation Tool for HCI Data Utilizing Machine Learning Methods. In: *Proceedings of the 1st International Conference on Affective and Pleasurable Design (APD'12) [jointly with the 4th International Conference on Applied Human Factors and Ergonomics (AHFE'12)]*, CRC Press, 2012 (Advances in Human Factors and Ergonomics Series), S. 5347–5352

Abstract: ATLAS is a graphical tool for the annotation of multi-modal data streams collected in human computer interaction (HCI) scenarios. The input modalities to ATLAS are not limited to audio and video, any type of input data collected in HCI settings such as ECG, EEG, EMG etc can be processed by ATLAS. In addition to displaying raw input data the intermediate data processing outputs such as the results of many feature extraction procedures or the classification results of pre-trained classifier modules can be displayed as well. Annotation and transcription of HCI scenes by human annotators is also integrated and supported in the ATLAS system. In this paper ATLAS's basic structure is described, furthermore it is explained how different types of data and label information can be presented. Besides these basic annotation features an active learning (active data selection) approach has been included into the overall system. Support Vector Machines (SVM) utilizing probabilistic outputs are the current algorithms to select confident data. Confident classification results made by the SVM classifier support the human annotor to investigate unlabeled parts of a dataset. Finally we give an overview of the upcoming development steps of the tool, and the features of the next versions.

MEUDT, Sascha ; GLODEK, Michael ; SCHELS, Martin ; SCHWENKER, Friedhelm: Multi-View Video Based Tracking and Audio-Visual Identification of Persons in a Human-Computer-Interaction Scenario. In: *Proceedings of IEEE International Conference on Cybernetics (CybConf'13)*, Springer, 2013, S. 1–8

Abstract: User identification and tracking are definitely the basic tasks in any human computer interaction (HCI) scenario. For these tasks we propose a multi-view approach utili-

zing multi-camera systems and audio processing systems. Face detectors and face recognizers are based on orientation histogram and eigenface techniques, and Mel Frequency Cepstral Coefficients (MFCC) are applied for speaker identification. In order to achieve a robust user identification and localization spatio-temporal classifier fusion methods have been integrated into the overall classifier system, support vector machines (SVM) and k nearest neighbor (kNN) models are used as base classifiers. A general office environment with up to six persons was the test bed for data collection and numerical evaluation.

MEUDT, Sascha ; SCHMIDT-WACK, Miriam ; HONOLD, Frank ; SCHÜSSEL, Felix ; WEBER, Michael ; SCHWENKER, Friedhelm ; PALM, Günther: Going Further in Affective Computing: How Emotion Recognition Can Improve Adaptive User Interaction. In: ESPOSITO, Anna (Hrsg.) ; JAIN, Lakhmi C. (Hrsg.): *Toward Robotic Socially Believable Behaving Systems - Volume I: Modeling Emotions*. Springer, 2016, S. 73–103

Abstract: This article joins the fields of emotion recognition and human computer interaction. While much work has been done on recognizing emotions, they are hardly used to improve a user's interaction with a system. Although the fields of affective computing and especially serious games already make use of detected emotions, they tend to provide application and user specific adaptions only on the task level. We present an approach of utilizing recognized emotions to improve the interaction itself, independent of the underlying application at hand. Examining the state of the art in emotion recognition research and based on the architecture of Companion-System, a generic approach for determining the main cause of an emotion within the history of interactions is presented, allowing a specific reaction and adaption. Using such an approach could lead to systems that use emotions to improve not only the outcome of a task but the interaction itself in order to be truly individual and empathic.

MEUDT, Sascha ; SCHWENKER, Friedhelm: On Instance Selection in Audio Based Emotion Recognition. In: *Proceedings of the 5th IAPR TC3 Workshop on Artificial Neural Networks for Pattern Recognition (ANNPR'12)*, Springer, 2012, 186-192

Abstract: Affective computing aim to provide simpler and more natural interfaces for human-computer interaction applications, e.g. recognizing automatically the emotional status of the user based on facial expressions or speech is important to model user as complete as possible in order to develop human-computer interfaces that are able to respond to the

user's action or behavior in an appropriate manner. In this paper we focus on audio-based emotion recognition. Data sets employed for the statistical evaluation have been collected through Wizard-of-Oz experiments. The emotional labels have been are defined through the experimental set up therefore given on a relatively coarse temporal scale (a few minutes) which This global labeling concept might lead to miss-labeled data at smaller time scales, for instance for window sizes uses in audio analysis (less than a second). Manual labeling at these time scales is very difficult not to say impossible, and therefore our approach is to use the globally defined labels in combination with instance/sample selection methods. In such an instance selection approach the task is to select the most relevant and discriminative data of the training set by using a pre-trained classifier. Mel-Frequency Cepstral Coefficients (MFCC) features are used to extract relevant features, and probabilistic support vector machines (SVM) are applied as base classifiers in our numerical evaluation. Confidence values to the samples of the training set are assigned through the outputs of the probabilistic SVM.

MEUDT, Sascha ; SCHWENKER, Friedhelm: Enhanced Autocorrelation in Real World Emotion Recognition. In: *Proceedings of the 16th ACM International Conference on Multimodal Interaction (ICMI'14)*, ACM, 2014, S. 502–507

Abstract: Multimodal emotion recognition in real world environments is still a challenging task of affective computing research. Recognizing the affective or physiological state of an individual is difficult for humans as well as for computer systems, and thus finding suitable discriminative features is the most promising approach in multimodal emotion recognition. In the literature numerous features have been developed or adapted from related signal processing tasks. But still, classifying emotional states in real world scenarios is difficult and the performance of automatic classifiers is rather limited. This is mainly due to the fact that emotional states can not be distinguished by a well defined set of discriminating features. In this work we present an enhanced autocorrelation feature as a multi pitch detection feature and compare its performance to feature well known, and state-of-the-art in signal and speech processing. Results of the evaluation show that the enhanced autocorrelation outperform other state-of-the-art features in case of the challenge data set. The complexity of this benchmark data set lies in between real world data sets showing naturalistic emotional utterances, and the widely applied and well-understood acted emotional data sets.

MEUDT, Sascha ; SCHWENKER, Friedhelm: ATLAS - Machine learning based annnotation of multimodal data recorded in human-computer interaction scenarios. In: *Proceedings of the 1st International Symposium on Companion-Technology (ISCT 2015)*, ACM, 2015, S. 181–186

Abstract: We present ATLAS a graphical tool for the annotation of multimodal data streams in human computer interaction (HCI) scenarios. In typical HCI scenarios, besides multi-channel audio and video inputs, various bio-physiological data streams are recorded. ATLAS is able to process all these different types of data. In addition to the data itself, extracted features, and even the outputs of pre-trained classifiers can be displayed. Furthermore, annotation and transcription are realized in ATLAS. Furthermore, partially supervised learning techniques, such as active data selection, are integrated together with probabilistic support vector machines (prob-SVM).

MEUDT, Sascha ; SIEGERT, Ingo ; SCHÜSSEL, Felix ; SCHMIDT, Miriam ; REUTER, Stephan ; LAYHER, Georg ; HÖRNLE, Thilo ; HANDRICH, Sebastian ; AL-HAMADI, Ayoub ; DIETMAYER, Klaus ; NEUMANN, Heiko ; PALM, Günther ; SCHWENKER, Friedhelm ; WENDEMUTH, Andreas: Multi-Modal Information Processing in Companion Systems—A Ticket Purchase System. In: BIUNDO, Susanne (Hrsg.) ; WENDEMUTH, Andreas (Hrsg.): *Companion Technology*. 1. Springer, 2016

Abstract: The demonstration scenario of the SFB/TRR-62 shows multimodal, dynamic interactions between a human being and a technical system that are adaptive to the situation and the emotional state. It uses the example of purchasing a train ticket to demonstrate how a companion system is able to adapt its dialog with the user according to the situational context and the emotions of the user. One special feature of this scenario are the simultaneous analyses and evaluations of explicit and implicit input data. The scenario demonstrates further how background knowledge about the user can be included; for example often visited destinations, the user's timetable or the number of travelers.

MEUDT, Sascha ; ZHARKOV, Dimitri ; KÄCHELE, Markus ; SCHWENKER, Friedhelm: Multi Classifier Systems and Forward Backward Feature Selection Algorithms to Classify Emotional Coloured Speech. In: *Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI'13)*, ACM, 2013 (ICMI 2013). – ISBN 978–1–4503–2129–7, S. 551–556

Abstract: Systems for the recognition of psychological characteristics such as the emotional state in real world scenarios have to deal with several difficulties. Amongst those are unconstrained environments and uncertainties in one or several input channels. However a more crucial aspect is the content of the data itself. Psychological states are highly person-dependent and often even humans are not able to determine the correct state a person is in. A successful recognition system thus has to deal with data, that is not very discriminative and often simply misleading. In order to succeed, a critical view on features and decisions is essential to select only the most valuable ones. This work presents a comparison of a common multi classifier system approach based on state of the art features and a modified forward backward feature selection algorithm with a long term stopping criteria. The second approach takes also features of the voice quality family into account. Both approaches are based on the audio modality only. The dataset used in the challenge is an in between dataset of real world datasets which are still very hard to handle and over acted datasets which were famous in the past and today are well understood.

SCHELS, Martin ; GLODEK, Michael ; **MEUDT, Sascha** ; SCHERER, Stefan ; SCHMIDT, Miriam ; LAYHER, Georg ; TSCHECHNE, Stephan ; BROSCH, Tobias ; HRABAL, David ; WALTER, Steffen ; TRAUE, Harald ; PALM, Günther ; NEUMANN, Heiko ; SCHWENKER, Friedhelm: Multi-Modal Classifier-Fusion for the Recognition of Emotions. In: ROJC, Matej (Hrsg.) ; CAMPBELL, Nick (Hrsg.): *Coverbal Synchrony in Human-Machine Interaction*. 1. CRC Press, 2013, S. 73–98

Abstract: Research activities in the field of human-computer interaction increasingly addressed the aspect of integrating features that characterize different types of emotional intelligence. Human emotions are expressed through different modalities such as speech, facial expressions, hand or body gestures, and therefore the classification of human emotions should be considered as a multi-modal pattern recognition problem. In recent time, a multitude of approaches have been proposed to enhance the training and recognition of multiple classifier systems (MCSs) utilizing multiple modalities to classify human emotional states. The work summarizes the progress of investigating such systems and presents aspects of the problem namely fusion architectures and training of statistical classifiers based on only marginal informative features. Furthermore, it describes how the effects of missing values, e.g. due to missing sensor data or classifiers

Exploiting reject options in order to circumvent false classifications, can be mitigated. Another aspect is the usage of partially supervised learning, either to support annotation or to improve classifiers. Parts of these aspects are then exemplified using two recent examples of emotion recognition, showing a successful realization of MCSs in emotion recognition.

SCHELS, Martin ; GLODEK, Michael ; **MEUDT, Sascha** ; SCHMIDT, Miriam ; HRABAL, David ; BÖCK, Ronald ; WALTER, Steffen ; SCHWENKER, Friedhelm: Multi-Modal Classifier-Fusion for the Classification of Emotional States in WOZ Scenarios. In: *Proceedings of the 1st International Conference on Affective and Pleasurable Design (APD'12) [jointly with the 4th International Conference on Applied Human Factors and Ergonomics (AHFE'12)]*, CRC Press, 2012 (Advances in Human Factors and Ergonomics Series), S. 5337–5346

Abstract: Learning from multiple sources is an important field of research in many applications. Amongst of the benefits of such an approach is that different sources can correct each other or that a failure of a channel can be easier compensated. The emotion of a subject can give helpful cues for a computer in a human machine dialog. The problem of emotion recognition is inherently multimodal. The most intuitive way of inferring a user state is to use facial expression and spoken utterances. However, bio-physiological readings can be helpful in this context. In this study, a novel information fusion architecture for the classification of human emotions in a computer interaction is proposed. We use information from the three modalities above mentioned. It turned out that the combination of different sources can be helpful for the classification. Also, a reject option for the classifiers is evaluated and yields promising results.

SCHERER, Stefan ; SIEGERT, Ingo ; BIGALKE, Lutz ; **MEUDT, Sascha**: Developing an Expressive Speech Labeling Tool Incorporating the Temporal Characteristics of Emotion. In: CALZOLARI, N. (Hrsg.) ; CHOUKRI, K. (Hrsg.) ; MAEGAARD, B. (Hrsg.) ; MARIANI, J. (Hrsg.) ; ODJIK, J. (Hrsg.) ; PIPERIDIS, S. (Hrsg.) ; ROSNER, M. (Hrsg.) ; TAPIAS, D. (Hrsg.): *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, ELRA, 2010, S. 1172–1175

Abstract: A lot of research effort has been spent on the development of emotion theories and modeling, however, their suitability and applicability to expressions in human computer interaction has not exhaustively been evaluated. Furthermore, investigations con-

cerning the ability of the annotators to map certain expressions onto the developed emotion models is lacking proof. The proposed annotation tool, which incorporates the standard Geneva Emotional Wheel developed by Klaus Scherer and a novel temporal characteristic description feature, is aiming towards enabling the annotator to label expressions recorded in human computer interaction scenarios on an utterance level. Further, it is respecting key features of realistic and natural emotional expressions, such as their sequentiality, temporal characteristics, their mixed occurrences, and their expressivity or clarity of perception. Additionally, first steps towards evaluating the proposed tool, by analyzing utterance annotations taken from two expressive speech corpora, are undertaken and some future goals including the open source accessibility of the tool are given.

SCHWENKER, Friedhelm ; BÖCK, Ronald ; SCHELS, Martin ; **MEUDT, Sascha** ; SIEGERT, Ingo ; GLODEK, Michael ; SCHMIDT-WACK, Miriam ; THIAM, Patrick ; WENDEMUTH, Andreas ; KRELL, Gerald: Multimodal Affect Recognition in the Context of Human-Computer Interaction for Companion Systems. In: BIUNDO, Susanne (Hrsg.) ; WENDEMUTH, Andreas (Hrsg.): *Companion Technology*. 1. Springer, 2016

Abstract: In general, humans interact with each other using multiple modalities. The main channels are speech, facial expressions, and gesture. But also bio-physiological data such as biopotentials can convey valuable information which can be used to interpret the communication in a dedicated way. Companion system can use these modalities to perform an efficient humans-computer interaction (HCI). To do so, the multiple sources need to be analyzed and combined in technical systems. However, so far only few studies have been published dealing with the fusion of three or even more such modalities. This chapter addresses the necessary processing steps in the development of a multimodal system applying fusion approaches,

The ATLAS and Ikannotate are presented which are designed for the pre-analyzing of multimodal data streams and the labeling of relevant parts. ATLAS allows to display raw data, extracted features and even outputs of pre-trained classifier modules. Further, the tools integrates annotation, transcription and an active learning module. The Ikannotate can be directly used for transcription and guided step-wise emotional annotation of multimodal data. The tool includes the three mainly used annotation paradigms, namely the basic emotions, the Geneva emotion wheel and the self-assessment manikins (SAM). Furthermore, annotators using Ikannotate can assign an uncertainty to samples.

Classifier architectures need to realize a fusion system in which the multiple modalities are combined. A large number of machine learning approaches were evaluated such as data, feature, score and decision level fusion schemes, but also temporal fusion architectures and partially super-vised learning.

The proposed methods are evaluated on either multimodal benchmark corpora or on the datasets of the Transregional Collaborative Research Centre SFB/TRR 62, i.e. Last Minute Corpus and the EmoRec Dataset. Furthermore, we present results which were achieved in international challenges.

SIEGERT, Ingo ; REUTHER, Stefan ; SCHÜSSEL, Felix ; LAYHER, Georg ; HÖRNLE, Thilo ; **MEUDT, Sascha** ; WENDEMUTH, Andreas: Multimodal Information Processing: The Ticket Purchase - a Demonstration Scenario of the SFB/TRR-62. In: *Elektronische Sprachsignalverarbeitung*, 2016, S. 111–118

Abstract: The demonstration scenario of the SFB/TRR-62 shows multimodal, dynamic interactions between a human being and a technical system that are adaptive to the situation and the emotional state. It uses the example of purchasing a train ticket to demonstrate how a companion system is able to adapt its dialog with the user according to the situational context and the emotions of the user. One special feature of this scenario are the simultaneous analyses and evaluations of explicit and implicit input data. The scenario demonstrates further how background knowledge about the user can be included; for example often visited destinations, the user's timetable or the number of travelers.

THIAM, Patrick ; **MEUDT, Sascha** ; KÄCHELE, Markus ; PALM, Günther ; SCHWENKER, Friedhelm: Detection of Emotional Events Utilizing Support Vector Methods in an Active Learning HCI Scenario. In: *Proceedings of the 2014 Workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems (ERM4HCI'14)*, 2014, S. 31–36

Abstract: In recent years the fields of affective computing and emotion recognition have experienced a steady increase in attention and especially the creation and analysis of multi-modal corpora has been the focus of intense research. Plausible annotation of this data, however is an enormous problem. In detail emotion annotation is very time consuming, cumbersome and sensitive with respect to the annotator. Furthermore emotional reactions

are often very sparse in HCI scenarios resulting in a large annotation overhead to gather the interesting moments of a recording, which in turn are highly relevant for powerful features, classifiers and fusion architectures. Active learning techniques provide methods to improve the annotation processes since the annotator is asked to only label the relevant instances of a given dataset. In this work an unsupervised one-class Support Vector Machine is used to build a background model of non-emotional sequences on a novel HCI dataset. The human annotator is iteratively asked to label instances that are not well explained by the background model, which in turn renders them candidates for being interesting events such as emotional reactions that diverge from the norm. The outcome of the active learning procedure is a reduced dataset of only 14% the size of the original dataset that contains most of the significant information, in this case more than 75% of the emotional events.

THIAM, Patrick ; **MEUDT, Sascha** ; PALM, Günther ; SCHWENKER, Friedhelm: A Temporal Dependency Based Multi-modal Active Learning Approach for Audiovisual Event Detection. In: *Neural Processing Letters* (2017). `http://dx.doi.org/10.1007/s11063-017-9719-y`. – DOI 10.1007/s11063–017–9719–y. – ISSN 1573–773X

Abstract: In this work, two novel active learning approaches for the annotation and detection of audiovisual events are proposed. The assumption behind the proposed approaches is that events are susceptible to substantively deviate from the distribution of normal observations and therefore should be lying in regions of low density. Thus, it is believed that an event detection model can be trained more efficiently by focusing on samples that appear to be inconsistent with the majority of the dataset. The first approach is an uni-modal method which consists in using rank aggregation to select informative samples which have previously been ranked using different unsupervised outlier detection techniques in combination with an uncertainty sampling technique. The information used for the sample selection stems from an unique modality (e.g. video channel). Since most active learning approaches focus on one target channel to perform the selection of informative samples and thus do not take advantage of potentially useful and complementary information among correlated modalities, we propose an extension of the previous uni-modal approach to multi-modality. From a target pool of instances belonging to a specific modality, the uni-modal approach is used to select and manually label a set of informative instances. Additionally, a second set of automatically labelled instances of the target pool is generated, based on a transfer of information stemming from an auxiliary modality which is temporally dependent to the target one. Both sets of labelled instances (automatically and manually labelled instances) are used for the semi-supervised training of a classification model to be used in the next active learning iteration. Both methods have been

assessed on a set of participants selected from the UUlmMAC dataset and have proven to be effective in substantially reducing the cost of manual annotation required for the training of a facial event detection model. The assessment is done based on two different methods: Support Vector Data Description and expected similarity estimation. Furthermore, given an appropriate sampling approach, the multi-modal approach outperforms its uni-modal counterpart in most of the cases.

THIAM, Patrick ; **MEUDT, Sascha** ; SCHWENKER, Friedhelm ; PALM, Günther: Active Learning for Speech Event Detection in HCI. In: *Proceedings of Artificial Neural Networks in Pattern Recognition (ANNPR'16)*, 2016, S. 285–297

Abstract: In this work, a pool-based active learning approach combining outlier detection methods with uncertainty sampling is proposed for speech event detection. Events in this case are regarded as atypical utterances (e.g. laughter, heavy breathing) occurring sporadically during a Human Computer Interaction (HCI) scenario. The proposed approach consists in using rank aggregation to select informative speech segments which have previously been ranked using different outlier detection techniques combined with an uncertainty sampling technique. The uncertainty sampling method is based on the distance to the boundary of a Support Vector Machine with Radial Basis Function kernel trained on the available annotated samples. Extensive experimental results prove the effectiveness of the proposed approach.

VALENA, Maria ; GRUSS, Sascha ; LAYHER, Georg ; THIAM, Patrick ; ZHANG, Yan ; SCHORK, Daniel ; KESSLER, Viktor ; **MEUDT, Sascha** ; NEUMANN, Heiko ; KIM, Jonghwa ; SCHWENKER, Friedhelm ANDRE, Elisabeth ; TRAUE, Harald C. ; WALTER, Steffen: The SenseEmotion Database: A Multimodal Database for the Development and Systematic Validation of an Automatic Painand Emotion-Recognition System. In: *Proceedings of Artificial Neural Networks in Pattern Recognition (ANNPR'16)*, 2016, S. 127–139

Abstract: In our modern industrial society the group of the older (generation 65+) is constantly growing. Many subjects of this group are severely affected by their health and are suffering from disability and pain. The problem with chronic illness and pain is that it lowers the patient's quality of life, and therefore accurate pain assessment is needed to facilitate effective pain management and treatment. In the future, automatic pain monitoring may enable health care professionals to assess and manage pain in a more and more objective way. To this

end, the goal of our SenseEmotion project is to develop automatic painand emotion-recognition systems for successful assessment and effective personalized management of pain, particularly for the generation 65+. In this paper the recently created SenseEmotion Database for pain- vs. emotion-recognition is presented. Data of 45 healthy subjects is collected to this database. For each subject approximately 30min of multimodal sensory data has been recorded. For a comprehensive understanding of pain and affect three rather different modalities of data are included in this study: biopotentials, camera images of the facial region, and, for the first time, audio signals. Heat stimulation is applied to elicit pain, and affective image stimuli accompanied by sound stimuli are used for the elicitation of emotional states.

ZHARKOV, Dimitrij ; KÄCHELE, Markus ; **MEUDT, Sascha** ; SCHWENKER, Friedhelm: A long-term stopping criterion for the forward-backward feature selection algorithm for the classification of emotions in speech. In: *Joint Proceedings of the 2013th T2CT and CCGL Workshops at IVA 2013*, 2013, S. 25–33

Abstract: Emotion recognition from speech is an important field of research in human-machine-interfaces, and has various applications, for instance for call centers. In the proposed classifer system different spectral, prosodic and voice quality features are extracted from the speech signals. These features are then used to represent long-term information of the speech, leading to utterance-wise suprasegmental features. The most promising of these features are selected using a forward-selection/backwardelimination algorithm using a long-term termination criterion for the selection. The overall system has been evaluated by using recordings from the public Berlin emotion database. Utilizing the resulted features a recognition rate of 88,97% has been achieved which surpasses the performance of humans on this database.

ABKÜRZUNGSVERZEICHNIS

ABC	Airplane Behaviour Corpus
API	Application Programming Interface
ATLAS	Annotation Tool with a Learning Assistive System
AU	Action Unit
AVEC	Audi Visual Emotion recognition Challenge
BT	Bluetooth
bwUniCluster	Baden-Württembergische Universitäts Rechen Cluster
CD	Compact Disk
CHMM	Conditioned Hidden Markov Modell
CPU	Central Processing Unit
CSV	Komma separierte Werte (engl. comma sepperated value)
DCT	Diskrete Cosinus Transformation
DVD	Digital Video Disk
DVI	Digital Video Interface
EAC	Enhanced Auto Correlation
EEG	Elektroenzephalografie
EKG	Elektrokardiographie
EMG	Elektromyographie
EmoDB	Berlin Database of Emotional Speech
EmoRec	Emotion Recognition Corpus

EmotiW	Emotions in the Wild
EV	Eigenvektor
EW	Eigenwert
FACS	Facial Action Coding System
FFT	schnelle Fourier-Transformation (engl. Fast Furier Transformation)
fMRI	funktionelle Magnetresonanztomographie (engl. functional Magnetic Resonance Imaging)
GAT	Gesprächsanalytisches Transkriptionssystem
GMM	Gaussian Mixture Models
G-Mean	Geometrischer Mittelwert
GUI	Grafische Benutzeroberfläche (engl. Graphical User Interface)
HD	High Definition
HIAT	Halbinterpretative Arbeitstranskriptionen
HMM	Hidden Markov Modell
IAPS	International Affective Picture System
GF	Ganganalyse und Fitness
IR	Infrarot
kNN	k Nächste Nachbarn
LAN	Local Area Network
LBP-TOP	Local Binary Pattern - Tree Ortogonal Planes
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LMC	Last Minute Corpus
LOSO	Ein Proband ausgelassen (engl. leave one subject out)

LPC	Linear Predictive Coding
MAR^{2}S	Modular Affect Recording and Recognition System
MFCC	Mel Frequency Cepstral Coefficent
MFN	Markov Fusionsnetzwerk
MLP	Multi Layer Perzeptron
MMI	Mensch Maschine Interaktion
ModSpec	Modulations Spektrum
MP3	MPEG-2 Audio Layer III
MP4	MPEG-4 Part 14
MPEG	Moving Picture Experts Group
OCSVM	One class support vector machine
P2P	Peer to Peer
PCA	Hauptkomponentenanalyse (engl. Principal Component Analysis)
PC	Personal Computer
PHOG	Pyramid Histogramm of Gradients
RastaPLP	Relative Spectral Perceptual Linaear Prediction
RBF	Radiale Basis Funktion
SAM	Self Assessment Manikin
SCL	Hautleitwert (engl. Skin Conductance Level)
SD	Standart Definition
SFB-TRR 62	Sonderforschungsbereich Transregio 62
SMOTE	Synthetic Minority Over-sampling Technique
SNTP	Simple Network Time Protocol
SSI	Social Signal Interpretation

SVDD	Supportvector data description
SVM	Support Vektor Maschine
TCP / IP	Transmission Control Protocol / Internet Protocol
TTS	Text To Speech
UHF	Ultra Hoch Frequenz
UI	Benutzer Schnittstelle (engl. User Interface)
USB	Universal Serial Bus
USF	umgangssprachliche Fragen
uulmMAC	Ulm University Multimodal Affective Corpus
VAD	Valenz-Arousal-Dominance
VaM	Vera am Mittag
VLC	Video Lan Client
WAVE	Resource Interchange File Format Wave Extention
WMV	Windows Media Video
WoZ	Wizzard of Oz
XML	Extensible Markup Language

ABBILDUNGSVERZEICHNIS

1.1. Companion-Technologie, Nutzer und Umwelt 6

2.1. Mimischer Ausdruck der Basisemotionen nach Ekman 17
2.2. Emotionsmodel nach Plutchik 20
2.3. Annotationsdarstellung des Emotionsmodells nach Plutchik 21
2.4. VAD Model 22
2.5. Standardisierte Piktogramme des SAM 23
2.6. Screenshot der uulmMAC Nutzerschnittstelle 30
2.7. Screenshot der uulmMAC Selbstannotation 31
2.8. Sequenzieller Ablauf des uulmMAC 33
2.9. SAM Einschätzungen der uulmMAC Experimentalsequenzen 34
2.10. Technische Aufnahme-Architektur des uulmMAC 41
2.11. Bauliche Anordnung des uulmMAC Equipments. 42
2.12. Screenshot des MAR^{2}S Kinect Aufzeichnungsmoduls (KinectRecorder) . . . 43
2.13. Ansichten des Kinect Sensors des uulmMAC 44
2.14. Screenshot des MAR^{2}S biophysiologischen Aufzeichnungsmoduls (gRecorder) 45
2.15. Anbringung der biophysiologischen Sensoren im uulmMAC 46
2.16. Screenshot des MAR^{2}S Audio-Video Aufzeichnungsmoduls. 47
2.17. Screenshot des uulmMAC in ATLAS 51
2.18. Screenshot der uulmMAC Labels in ATLAS 52

3.1. Feeltrace Screenshot 60
3.2. Ikannotate Screenshot 62
3.3. ANNEMO Screenshot 63
3.4. ANVIL Screenshot 64
3.5. Aktive Lernverfahren 68
3.6. Prof of Concept des aktiven Lernverfahrens 76
3.7. Anzahl der als Emotion annotierten Instanzen 79
3.8. Klassifikationsperformanz in Abhängigkeit der Schwelle (Proband 12) 80
3.9. Klassifikationsperformanz in Abhängigkeit der Schwelle als ROC Kurven . . 81
3.10. Äußerungslängen bei Manueller Annotation 82
3.11. Resultate des überwacht Trainierten Klassifikators 84

3.12. Audio Resultate des aktiven Lernverfahrens bei Verwendung der SVDD Instanzselektionsmethode 87
3.13. Audio Resultate des aktiven Lernverfahrens bei Verwendung der GMM Instanzselektionsmethode 88
3.14. Video Resultate des aktiven Lernverfahrens bei Verwendung der SVDD Instanzselektionsmethode 90
3.15. Erzeugung der Multimodalen Grundwahrheit 92
3.16. Gegenüberstellung der multimodalen Selektionsstrategien des aktiven Lernverfahrens 93
3.17. Vergleichbarkeitsproblem der Distanzen zweier Modalitäten (Modelle) 94
3.18. Fusionsresultate des aktiven Lernverfahrens bei Verwendung der SVDD Instanzselektionsmethode 97
3.19. Anzahl Support Vektoren je Iteration als Abbruchkriterium des aktiven Lernverfahrens 99
3.20. EmoWheel Labeltool Screenshot 102
3.21. Darstellung der ATLAS Nutzerschnittstellenelemente. 105
3.22. Spur basierter Aufbau von Atlas 106
3.23. Darstellung der ATLAS Linienspuren 107
3.24. Darstellung der ATLAS Labelspuren 109
3.25. Screenshot des ATLAS LabelProperties Fensters 110
3.26. Screenshot des ATLAS Search for Labels Fensters 111
3.27. Screenshot des ATLAS ClassEditor Fensters 112
3.28. Screenshot des ATLAS TrackEditor Fensters 113
3.29. Screenshot des ATLAS Haupt und Playback Fensters 114
3.30. Screenshot der ATLAS - Matlab Schnittstelle 116
3.31. Screenshot der automatischen ATLAS Sprecher Segmentierung 117
3.32. Label Vorschläge unterbreitet durch ATLAS 119
3.33. Screenshot einer Demonstration des aktiven Lernverfahrens in ATLAS 121
3.34. Verteilte Annotation in ATLAS 122

4.1. Fusionsarchitekturen 127
4.2. Architektur der Instanzselektion durch Reklassifikation 132
4.3. Erkennungsrate abhängig vom Grad der Rückweisung. 133
4.4. Resultat der Vorwärts-Rückwärts-Merkmalsselektion 136
4.5. Vergleichsdarstellung der Akustischen Merkmale 138
4.6. Beispieldarstellung von typischen Körperposen der uulmMAC Probanden . . 147
4.7. Gelenkgeschwindigkeit als Körperpose Merkmal 148
4.8. Gelenkdistanzen als Körperpose Merkmal 149

4.9. Vergleich verschiedener Klassifikatoren zur Erkennung von Über- und Unterforderung anhand der Körperpose 150
4.10. Klassifikationsresultat der Körperpose basierten Erkennung kognitiver Last . 150
4.11. Mittlere Merkmalsbedeutung der Körperpose Merkmale 151
4.12. Aktivität des Trapetius Muskels abhängig von der kognitiven Belastung . . . 155
4.13. Multimodale Performanz des MFN im Livebetrieb 163
4.14. MFN Ausgabe der kognitiven Belastungsschätzung. 164
4.15. Multimodale Erkennungsrate der kognitiven Belastung bei Verwendung individueller Schwellen 165
4.16. ATLAS Screenshot der Auswirkung von situativem Kontext auf die Schätzung von kognitiver Last 168

5.1. Screenshot des Frauenhofer IIS Frameworks zur visuellen Emotionserkennung. 171
5.2. Screenshot des SSI Frameworks zur Emotionserkennung 172
5.3. Darstellung aller Module des uulmMAC Experimentes im Systemmanager . . 176
5.4. Vergleich der Netzwerkstrukturen 178
5.5. V-Modell des Systementwurfs 181
5.6. Zusammenhang der zur Identifikation und Positionsschätzung realisierten Module. 183
5.7. Screenshot des Identifikations- und Trackingsystems 184
5.8. Boxplot des Fehlers der geschätzten Position der Identifikations- und Trackinganwendung 185
5.9. Ein Nutzer bei der interaktion mit einem prototypischen Fahrkartenautomaten 188
5.10. Sensorischer Bottom-Up-Informationsfluss im prototypischen Fahrkartenautomaten 189
5.11. Ebenenmodell der adaptiven Komponenten 196
5.12. Zeitliche Abfolge ebenenspezifscher Aktionen mit erkannten Emotionen . . . 197
5.13. Prozess zur Realisierung eines affektiv-adaptiven Systemverhaltens 199

TABELLENVERZEICHNIS

2.1. Beschreibung der AU gemäß FACS 18
2.2. Basisemotionen nach Ekman und den daran beteiligten AUs 19
2.3. Übersicht bestehender Datensätze 28
2.4. Eigenschaften der uulmMAC Experimentalsequenzen 33
2.5. Korrelation zwischen SAM und USF für die Überforderungssequenz 36
2.6. Korrelation zwischen SAM und USF für die Unterforderungssequenz 36
2.7. Durchschnittliche Längen der Labels in verschieden Modalitäten 38
2.8. uulmMAC Synchronisationsfehler 39
2.9. Datendurchsatz und Speichervolumen der verwendeten Sensoren und Aufnahmeformate 40

3.1. Manueller Annotationsaufwand des uulmMAC 73
3.2. Verteilung der visuellen Mermalsinstanzen der Vorstudie 75
3.3. Verteilung der Akustischen Mermalsinstanzen des uulmMAC 83
3.4. Mittlerer Anteil der Merkmalsinstanzen welche benötigt werden um die Baseline anzunähern 96

4.1. Dimensionen der akustischen Merkmale 134
4.2. Verwechslungsmatrix des Klassifikators trainiert auf den durch vorwärts rückwärts Merkmalsselektion gewählten Merkmalen. 137
4.3. Erkennungsraten der Emotionserkennung der EmotiW Challenge evaluiert auf dem *Validationset*. 141
4.4. Erkennungsraten der Emotionserkennung der EmotiW Challenge evaluiert auf dem *Testset*. 142
4.5. Verwechslungsmatrix des EAC Merkmals der EmotiW Challenge 142
4.6. Evaluation aller akustischen Merkmale für die Erkennung von Überforderung im uulmMAC 144
4.7. Evaluation der biophysiologischen Merkmale zur Erkennung von Über- und Unterforderung 154
4.8. Erkennunsraten der ersten AVEC Challenge 157
4.9. Bimodale Erkennunsraten von Überforderung, Unterforderung und Fuison durch Moore-Penrose Pseudoinverse 159

4.10. Bimodale Erkennunsraten von Überforderung, Unterforderung und Fuison durch Konfidenz Summe 160
4.11. Evaluation der biophysiologischen Erkennungsraten abhängig von Sprachlicher Überlagerung 161
4.12. Signifikanz des Sprachlichen Einflusses auf biophysiologische Merkmale. . . 161
4.13. Uni- und multimodale Erkennungsraten für die Detektion von Über- und Unterforderung im uulmMAC (interindividuell) 166

5.1. Resultate der Video basierten Nutzeridentifikation 185

BIBLIOGRAPHIE

[1] AKGUN, Mahir ; CAGILTAY, Kursat ; ZEYREK, Deniz: The effect of apologetic error messages and mood states on computer users' self-appraisal of performance. In: *Journal of Pragmatics* 42 (2010), Nr. 9, S. 2430 – 2448. http://dx.doi.org/10.1016/j.pragma.2009.12.011. – DOI 10.1016/j.pragma.2009.12.011. – ISSN 0378–2166

[2] AKYILDIZ, Ian F. ; SU, Weilian ; SANKARASUBRAMANIAM, Yogesh ; CAYIRCI, Erdal: A survey on sensor networks. In: *IEEE Communications magazine* 40 (2002), Nr. 8, S. 102–114

[3] ALAM, Md J. ; KENNY, Patrick ; OUELLET, Pierre ; STAFYLAKIS, Themos ; DUMOUCHEL, Pierre: Supervised/Unsupervised Voice Activity Detector for Text-dependent Speaker Recognition on RSR2015 Corpus. In: *Odyssey Speaker and Language Recognition Workshop*, 2014, S. 123–130

[4] ALUR, R ; ARZEN, K-E ; BAILLIEUL, John ; HENZINGER, TA: *Handbook of networked and embedded control systems*. Springer Science & Business Media, 2007

[5] ALZOUBI, Omar ; CALVO, Rafael A. ; STEVENS, Ronald H.: Classification of EEG for Affect Recognition: An Adaptive Approach. In: *Australasian Conference on Artificial Intelligence* Bd. 5866 Springer, 2009, S. 52–61

[6] ANGLUIN, Dana: Queries and concept learning. In: *Machine learning* 2 (1988), Nr. 4, S. 319–342

[7] ARONSON, Elliot ; AKERT, Robin M. ; WILSON, Timothy D.: *Sozialpsychologie*. Pearson Deutschland GmbH, 2010

[8] ATKINSON, Anthony P. ; DITTRICH, Winand H. ; GEMMELL, Andrew J. ; YOUNG, Andrew W.: Emotion perception from dynamic and static body expressions in point-light and full-light displays. In: *Perception* 33 (2004), Nr. 6, S. 717–746

[9] ATLAS, Les E. ; COHN, David A. ; LADNER, Richard E.: Training connectionist networks with queries and selective sampling. In: *Advances in neural information*

processing systems, 1990, S. 566–573

[10] AVIEZER, Hillel ; TROPE, Yaacov ; TODOROV, Alexander: Body cues, not facial expressions, discriminate between intense positive and negative emotions. In: *Science* 338 (2012), Nr. 6111, S. 1225–1229

[11] BALAKRISHNAMA, Suresh ; GANAPATHIRAJU, Aravind: Linear discriminant analysis-a brief tutorial. In: *Institute for Signal and information Processing* 18 (1998)

[12] BALOMENOS, Themis ; RAOUZAIOU, Amaryllis ; IOANNOU, Spiros ; DROSOPOULOS, Athanasios ; KARPOUZIS, Kostas ; KOLLIAS, Stefanos: Emotion analysis in man-machine interaction systems. In: *International Workshop on Machine Learning for Multimodal Interaction* Springer, 2004, S. 318–328

[13] BALZERT, Helmut: *Software-Technik*. Spektrum Verlag Heidelberg, 2000

[14] BASMAJIAN, John V. ; DE LUCA, Carlo J.: *Muscles alive: their functions revealed by electromyography*. Williams & Wilkins, 1985

[15] BASTIDE, R. ; PALANQUE, P.: A visual and formal glue between application and interaction. In: *Journal of Visual Languages and Computing* 10 (1999), Nr. 5, S. 481–507. `http://dx.doi.org/10.1006/jvlc.1999.0127`. – DOI 10.1006/jvlc.1999.0127

[16] BÖCK, Ronald ; SIEGERT, Ingo ; HAASE, Matthias ; LANGE, Julia ; WENDEMUTH, Andreas: ikannotate–a tool for labelling, transcription, and annotation of emotionally coloured speech. In: *Affective Computing and Intelligent Interaction* (2011), S. 25–34

[17] BEALE, Russell ; PETER, Christian: The role of affect and emotion in HCI. In: *Affect and emotion in human-computer interaction*. Springer, 2008, S. 1–11

[18] BENESTY, Jacob ; CHEN, Jingdong ; HUANG, Yiteng ; DMOCHOWSKI, Jacek: On microphone-array beamforming from a MIMO acoustic signal processing perspective. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15 (2007), Nr. 3, S. 1053–1065

[19] BERCHER, Pascal ; RICHTER, Felix ; HÖRNLE, Thilo ; GEIER, Thomas ; HÖLLER, Daniel ; BEHNKE, Gregor ; NIELSEN, Florian ; HONOLD, Frank ; SCHÜSSEL, Felix ; REUTER, Stephan u. a.: Advanced User Assistance for Setting Up a Home Theater. In: *Companion Technology*. Springer, 2017, S. 485–491

[20] BERGER, Hans: Über das Elektrenkephalogramm des Menschen. In: *Archiv für Psychiatrie und Nervenkrankheiten* 87 (1929), Nr. 1, S. 527–570. `http://dx.doi.org/`

10.1007/BF01797193. – DOI 10.1007/BF01797193. – ISSN 0003–9373

[21] BERGMEIR, Christoph ; BENITEZ, Jos.: On the use of cross-validation for time series predictor evaluation. In: *Information Sciences* (2012), S. 192–213

[22] BISHOP, Christopher M.: *Neural networks for pattern recognition*. Oxford university press, 1995

[23] BIUNDO, Susanne ; WENDEMUTH, Andreas: *Finazierungsantrag: Sonderforschungsbereich / Transregio 62 - „Eine Companion-Technologie für kognitive technische Systeme“(2009-2012)*. Universität Ulm, 2008

[24] BIUNDO, Susanne ; WENDEMUTH, Andreas: *Finazierungsantrag: Sonderforschungsbereich / Transregio 62 - „Eine Companion-Technologie für kognitive technische Systeme“(2013-2016)*. Universität Ulm, 2012

[25] BIUNDO, Susanne ; WENDEMUTH, Andreas: Companion-technology for cognitive technical systems. In: *KI-Künstliche Intelligenz* 30 (2016), Nr. 1, S. 71–75

[26] BIUNDO, Susanne ; WENDEMUTH, Andreas: *Finazierungsantrag: Sonderforschungsbereich / Transregio 62 - „Eine Companion-Technologie für kognitive technische Systeme“(2017-2020)*. Universität Ulm, 2016

[27] BIUNDO, Susanne ; WENDEMUTH, Andreas: *Companion Technology: A Paradigm Shift in Human-Technology Interaction*. Springer, 2017

[28] BLANCHARD, Benjamin S. ; FABRYCKY, Wolter J. ; FABRYCKY, Walter J.: *Systems engineering and analysis*. Bd. 4. Prentice Hall Englewood Cliffs, NJ, 1990

[29] BLUM, Avrim L. ; LANGLEY, Pat: Selection of relevant features and examples in machine learning. In: *Artificial intelligence* 97 (1997), Nr. 1, S. 245–271

[30] BRADLEY, Margaret M. ; LANG, Peter J.: Measuring emotion: the self-assessment manikin and the semantic differential. In: *Journal of behavior therapy and experimental psychiatry* 25 (1994), Nr. 1, S. 49–59

[31] BRAVE, Scott ; NASS, Clifford: Emotion in Human-computer Interaction. In: JACKO, Julie A. (Hrsg.) ; SEARS, Andrew (Hrsg.): *The Human-computer Interaction Handbook*. Hillsdale, NJ, USA : L. Erlbaum Associates Inc., 2003. – ISBN 0–8058–3838–4, S. 81–96

[32] BREIMAN, Leo: Bagging Predictors. In: *Machine Learning* 24 (1996), Nr. 2, S. 123–140

[33] BREIMAN, Leo: Random Forests. In: *Machine learning* 45 (2001), Nr. 1, S. 5–32

[34] BRÖHL, Adolf-Peter: *Das V-Modell: Der Standard für die Softwareentwicklung mit Praxisleitfaden*. Oldenbourg, 1993

[35] BROY, Manfred ; RAUSCH, Andreas: Das neue v-modell® xt. In: *Informatik-Spektrum* 28 (2005), Nr. 3, S. 220–229

[36] BURKHARDT, Felix ; PAESCHKE, Astrid ; ROLFES, Miriam ; SENDLMEIER, Walter F. ; WEISS, Benjamin: A database of german emotional speech. In: *Interspeech* Bd. 5, 2005, S. 1517–1520

[37] BUSSO, Carlos ; DENG, Zhigang ; YILDIRIM, Serdar ; BULUT, Murtaza ; LEE, Chul M. ; KAZEMZADEH, Abe ; LEE, Sungbok ; NEUMANN, Ulrich ; NARAYANAN, Shrikanth: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: *Proceedings of the 6th international conference on Multimodal interfaces* ACM, 2004, S. 205–211

[38] CARBONELL, Jaime G. ; MICHALSKI, Ryszard S. ; MITCHELL, Tom M.: An overview of machine learning. In: *Machine learning*. Springer, 1983, S. 3–23

[39] CARIDAKIS, George ; CASTELLANO, Ginevra ; KESSOUS, Loic ; RAOUZAIOU, Amaryllis ; MALATESTA, Lori ; ASTERIADIS, Stelios ; KARPOUZIS, Kostas: Multimodal emotion recognition from expressive faces, body gestures and speech. In: *Artificial intelligence and innovations 2007: From theory to applications* (2007), S. 375–388

[40] CASTELLANO, Ginevra ; VILLALBA, Santiago D. ; CAMURRI, Antonio: Recognising human emotions from body movement and gesture dynamics. In: *International Conference on Affective Computing and Intelligent Interaction* Springer, 2007, S. 71–82

[41] CATTELL, James M.: *Mental tests and measurements*. JSTOR, 1890

[42] CHANG, Wei-Cheng ; LEE, Ching-Pei ; LIN, Chih-Jen: A Revisit to Support Vector Data Description (svdd). In: *Technical Reports*, 2013

[43] CHAPELLE, Olivier ; SCHOLKOPF, Bernhard ; ZIEN, Alexander: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. In: *IEEE Transactions on Neural Networks* 20 (2009), Nr. 3, S. 542–542

[44] CHAWLA, Nitesh V. ; BOWYER, Kevin W. ; HALL, Lawrence O. ; KEGELMEYER, W. P.: SMOTE: Synthetic Minority Over-sampling Technique. In: *Journal of Artificial Intelligence Research* (2002), S. 321–357

[45] CIREŞAN, Dan ; MEIER, Ueli ; MASCI, Jonathan ; SCHMIDHUBER, Jürgen: Multi-column deep neural network for traffic sign classification. In: *Neural Networks* 32 (2012), S. 333–338

[46] CIVANLAR, Mehmet R. ; HASKELL, Barin G.: *Client-server architecture using internet and public switched networks*. 1999

[47] COHN, David ; ATLAS, Les ; LADNER, Richard: Improving generalization with active learning. In: *Machine learning* 15 (1994), Nr. 2, S. 201–221

[48] COHN, David A. ; GHAHRAMANI, Zoubin ; JORDAN, Michael I.: Active learning with statistical models. In: *Journal of artificial intelligence research* (1996)

[49] COHN, Jeffrey F. ; ZLOCHOWER, Adena J. ; LIEN, James ; KANADE, Takeo: Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. In: *Psychophysiology* 36 (1999), Nr. 1, S. 35–43

[50] COOPER, Andrew ; PETRIDES, KV: A psychometric analysis of the Trait Emotional Intelligence Questionnaire–Short Form (TEIQue–SF) using item response theory. In: *Journal of Personality Assessment* 92 (2010), Nr. 5, S. 449–457

[51] COUTAZ, Joelle ; NIGAY, Laurence ; SALBER, D. ; BLANDFORD, A. ; MAY, J. ; YOUNG, R. M.: Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The CARE Properties. In: *Proceedings of INTERACT95*, 1995, S. 115–120

[52] COWIE, R. ; DOUGLAS-COWIE, E. ; TSAPATSOULIS, N. ; VOTSIS, G. ; KOLLIAS, S. ; FELLENZ, W. ; TAYLOR, J.G.: Emotion recognition in human-computer interaction. In: *Signal Processing Magazine, IEEE* 18 (2001), Nr. 1, S. 32–80. http://dx.doi.org/10.1109/79.911197. – DOI 10.1109/79.911197. – ISSN 1053–5888

[53] COWIE, Roddy ; DOUGLAS-COWIE, Ellen ; SAVVIDOU*, Susie ; MCMAHON, Edelle ; SAWEY, Martin ; SCHRÖDER, Marc: 'FEELTRACE': An instrument for recording perceived emotion in real time. In: *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000

[54] COWIE, Roddy ; DOUGLAS-COWIE, Ellen ; TSAPATSOULIS, Nicolas ; VOTSIS, George ; KOLLIAS, Stefanos ; FELLENZ, Winfried ; TAYLOR, John G.: Emotion recognition in human-computer interaction. In: *IEEE Signal processing magazine* 18 (2001), Nr. 1, S. 32–80

[55] COWIE, Roddy ; SAWEY, Martin ; DOHERTY, Cian ; JAIMOVICH, Javier ; FYANS, Cavan ; STAPLETON, Paul: Gtrace: General trace program compatible with emotionml.

In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* IEEE, 2013, S. 709–710

[56] COX, Brad J.: *Object-oriented programming: an evolutionary approach*. Addison-Wesley Pub. Co., 1986

[57] CRITCHLEY, Hugo D. ; ELLIOTT, Rebecca ; MATHIAS, Christopher J. ; DOLAN, Raymond J.: Neural activity relating to generation and representation of galvanic skin conductance responses: a functional magnetic resonance imaging study. In: *Journal of Neuroscience* 20 (2000), Nr. 8, S. 3033–3040

[58] CUMMINS, Nicholas ; SCHERER, Stefan ; KRAJEWSKI, Jarek ; SCHNIEDER, Sebastian ; EPPS, Julien ; QUATIERI, Thomas F.: A review of depression and suicide risk assessment using speech analysis. In: *Speech Communication* 71 (2015), S. 10–49

[59] CURRY, Haskell B. ; SCHOENBERG, Isaac J.: On spline distributions and their limits-the polya distribution functions. In: *Bulletin of the American Mathematical Society* Bd. 53, 1947, S. 1114–1114

[60] DE GELDER, Beatrice: Towards the neurobiology of emotional body language. In: *Nature Reviews Neuroscience* 7 (2006), Nr. 3, S. 242–249

[61] DÉNIZ, Oscar ; BUENO, Gloria ; SALIDO, Jesús ; TORRE, Fernando De l.: Face recognition using histograms of oriented gradients. In: *Pattern Recognition Letters* 32 (2011), Nr. 12, S. 1598–1603

[62] DEVILLERS, Laurence ; VIDRASCU, Laurence ; LAMEL, Lori: Challenges in real-life emotion annotation and machine learning based detection. In: *Neural Networks* 18 (2005), Nr. 4, S. 407–422

[63] DEVORE, Ronald A. ; TEMLYAKOV, Vladimir N.: Some remarks on greedy algorithms. In: *Advances in computational Mathematics* 5 (1996), Nr. 1, S. 173–187

[64] DHALL, Abhinav ; GOECKE, Roland ; GHOSH, Shreya ; JOSHI, Jyoti ; HOEY, Jesse ; GEDEON, Tom: From individual to group-level emotion recognition: EmotiW 5.0. In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction* ACM, 2017, S. 524–528

[65] DHALL, Abhinav ; GOECKE, Roland ; JOSHI, Jyoti ; HOEY, Jesse ; GEDEON, Tom: Emotiw 2016: Video and group-level emotion recognition challenges. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction* ACM, 2016, S. 427–432

[66] DHALL, Abhinav ; GOECKE, Roland ; JOSHI, Jyoti ; SIKKA, Karan ; GEDEON, Tom: Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In: *Proceedings of the 16th International Conference on Multimodal Interaction* ACM, 2014, S. 461–466

[67] DHALL, Abhinav ; GOECKE, Roland ; JOSHI, Jyoti ; WAGNER, Michael ; GEDEON, Tom: Emotion recognition in the wild challenge 2013. In: *Proceedings of the 15th ACM on International conference on multimodal interaction* ACM, 2013, S. 509–516

[68] DHALL, Abhinav ; RAMANA MURTHY, OV ; GOECKE, Roland ; JOSHI, Jyoti ; GEDEON, Tom: Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* ACM, 2015, S. 423–426

[69] DIETRICH, Christian ; SCHWENKER, Friedhelm ; PALM, Günther: Classification of time series utilizing temporal and decision fusion. In: *Multiple Classifier Systems* (2001), S. 378–387

[70] EKMAN, Paul: *Are there basic emotions?* American Psychological Association, 1992

[71] EKMAN, Paul: An argument for basic emotions. In: *Cognition & emotion* 6 (1992), Nr. 3-4, S. 169–200

[72] EKMAN, Paul ; FRIESEN, Wallace: Facial action coding system: a technique for the measurement of facial movement. In: *Palo Alto: Consulting Psychologists* (1978)

[73] EKMAN, Paul ; FRIESEN, Wallace V.: Constants across cultures in the face and emotion. In: *Journal of personality and social psychology* 17 (1971), Nr. 2, S. 124

[74] EKMAN, Paul ; FRIESEN, Wallace V.: *Manual for the facial action coding system*. Consulting Psychologists Press, 1978

[75] ESKIMEZ, Sefik E. ; STURGE-APPLE, Melissa ; DUAN, Zhiyao ; HEINZELMAN, Wendi B.: WISE: Web-based Interactive Speech Emotion Classification. In: *SAAIP@ IJCAI*, 2016, S. 2–7

[76] EYBEN, F. ; WOLLMER, M. ; SCHULLER, B.: OpenEAR – Introducing the munich open-source emotion and affect recognition toolkit. In: *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009, S. 1–6

[77] EYBEN, Florian ; WENINGER, Felix ; GROSS, Florian ; SCHULLER, Björn: Recent

Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In: *ACM Multimedia (MM)*, 2013, S. 835–838

[78] EYBEN, Florian ; WÖLLMER, Martin ; SCHULLER, Björn: Opensmile: the munich versatile and fast open-source audio feature extractor. In: *Proceedings of the 18th ACM international conference on Multimedia* ACM, 2010, S. 1459–1462

[79] FASEL, Beat ; LUETTIN, Juergen: Automatic facial expression analysis: a survey. In: *Pattern recognition* 36 (2003), Nr. 1, S. 259–275

[80] FRAGOPANAGOS, N. ; TAYLOR, J.G.: Emotion recognition in human-computer interaction. In: *Neural Networks* 18 (2005), S. 389–405

[81] FREEMAN, William T. ; ROTH, Michal: Orientation histograms for hand gesture recognition. In: *International workshop on automatic face and gesture recognition* Bd. 12, 1995, S. 296–301

[82] FREUD, Sigmund: *Vorlesungen zur Einführung in die Psychoanalyse*. Offizin Haag-Drugulin AG, 1920

[83] FREY, Markus ; SCHELS, Martin ; GLODEK, Michael ; MEUDT, Sascha ; KÄCHELE, Markus ; SCHWENKER, Friedhelm: Investigation of Fuzzy Support Vector Machines. In: *Proceedings of the Statistical Computing 2013*, Springer, 2013 (Ulmer Informatik Berichte), S. 17–17

[84] FRIJDA, Nico H.: Varieties of affect: emotions and episodes, moods, and sentiments. In: EKMAN, Paul (Hrsg.) ; DAVIDSON, Richard J. (Hrsg.): *The Nature of Emotion, fundamental questions*. New York : Oxford University Press, 1994, S. 197–202

[85] GARNEFSKI, Nadia ; KRAAIJ, Vivian: The cognitive emotion regulation questionnaire. In: *European Journal of Psychological Assessment* 23 (2007), Nr. 3, S. 141–149

[86] GEVINS, Alan ; SMITH, Michael E. ; LEONG, Harrison ; MCEVOY, Linda ; WHITFIELD, Susan ; DU, Robert ; RUSH, Georgia: Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 40 (1998), Nr. 1, S. 79–91

[87] GLODEK, Michael ; GEIER, Thomas ; BIUNDO, Susanne ; PALM, Günter: A layered architecture for probabilistic complex pattern recognition to detect user preferences. In: *Journal of Biologically Inspired Cognitive Architectures* 9 (2014), S. 46–56

[88] GLODEK, Michael ; HONOLD, Frank ; GEIER, Thomas ; KRELL, Gerald ; NOTH-

DURFT, Florian ; REUTER, Stephan ; SCHÜSSEL, Felix ; HÖRNLE, Thilo ; DIETMAYER, Klaus ; MINKER, Wolfgang ; BIUNDO, Susanne ; WEBER, Michael ; PALM, Günther ; SCHWENKER, Friedhelm: Fusion Paradigms in Cognitive Technical Systems for Human-Computer Interaction. In: *Neurocomputing* 161 (2015), Nr. 0, S. 17–37. http://dx.doi.org/10.1016/j.neucom.2015.01.076. – DOI 10.1016/j.neucom.2015.01.076. – ISSN 0925–2312

[89] GLODEK, Michael ; SCHELS, Martin ; SCHWENKER, Friedhelm ; PALM, Günther: Combination of sequential class distributions from multiple channels using Markov fusion networks. In: *Journal on Multimodal User Interfaces* 8 (2014), Nr. 3, S. 257–272

[90] GLODEK, Michael ; SCHERER, Stefan ; SCHWENKER, Friedhelm ; PALM, G: Conditioned Hidden Markov Model Fusion for Multimodal Classification. In: *Interspeech*, 2011, S. 2269–2272

[91] GLODEK, Michael ; TSCHECHNE, Stephan ; LAYHER, Georg ; SCHELS, Martin ; BROSCH, Tobias ; SCHERER, Stefan ; KÄCHELE, Markus ; SCHMIDT, Miriam ; NEUMANN, Heiko ; PALM, Günther u. a.: Multiple classifier systems for the classification of audio-visual emotional states. In: *Affective Computing and Intelligent Interaction* (2011), S. 359–368

[92] GLOROT, Xavier ; BENGIO, Yoshua: Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, S. 249–256

[93] GRAM, Christian ; COCKTON, Gilbert: *Design principles for interactive software*. Chapman & Hall, Ltd., 1997. – ISBN 0–412–72470–7

[94] GRIMM, Michael ; KROSCHEL, Kristian ; NARAYANAN, Shrikanth: The Vera am Mittag German audio-visual emotional speech database. In: *Multimedia and Expo, 2008 IEEE International Conference on* IEEE, 2008, S. 865–868

[95] GUNES, Hatice ; PANTIC, Maja: Automatic, dimensional and continuous emotion recognition. In: *Creating Synthetic Emotions through Technological and Robotic Advancements*. IGI Global, 2010, S. 73–95

[96] GUYON, Isabelle ; ELISSEEFF, André: An introduction to variable and feature selection. In: *J. Mach. Learn. Res.* 3 (2003), S. 1157–1182. – ISSN 1532–4435

[97] HAAPALAINEN, Eija ; KIM, SeungJun ; FORLIZZI, Jodi F. ; DEY, Anind K.: Psychophysiological measures for assessing cognitive load. In: *Proceedings of the 12th ACM*

international conference on Ubiquitous computing ACM, 2010, S. 301–310

[98] HANDRICH, Sebastian ; AL-HAMADI, Ayoub: A robust method for human pose estimation based on geodesic distance features. In: *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on* IEEE, 2013, S. 906–911

[99] HASSOUN, Mohamad H.: *Fundamentals of artificial neural networks*. MIT press, 1995

[100] HEIMGÄRTNER, Rüdiger: Culturally-Aware HCI Systems. In: *Advances in Culturally-Aware Intelligent Systems and in Cross-Cultural Psychological Studies*. Springer, 2018, S. 11–37

[101] HELD, Dennis ; MEUDT, Sascha ; SCHWENKER, Friedhelm: Bimodal Recognition of Cognitive Load Based on Speech and Physiological Changes. In: *Proceedings of Multimodal pattern recognition of social signals in human computer interaction (MPRSS 2016)*, 2016, S. 1–8

[102] HERBART, Johann F.: *Allgemeine Pädagogik: aus dem Zweck der Erziehung abgeleitet*. JF Röwer, 1806

[103] HERMANSKY, Hynek: The modulation spectrum in the automatic recognition of speech. In: *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on* IEEE, 1997, S. 140–147

[104] HERMANSKY, Hynek ; MORGAN, Nelson ; BAYYA, Aruna ; KOHN, Phil: RASTA-PLP speech analysis technique. In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on* Bd. 1 IEEE, 1992, S. 121–124

[105] HIHN, Heinke ; MEUDT, Sascha ; SCHWENKER, Friedhelm: Inferring Mental Overload based on Postural Behavior and Gestures. In: *Proceedings of Emotion Recognition and Modeling for Companin Technologie (ERM4CT 2016)*, 2016, S. 1–8

[106] HIHN, Heinke ; MEUDT, Sascha ; SCHWENKER, Friedhelm: On Gesture and Postural Behavior as a Modality in Ensamble Methods. In: *Proceedings of Artificial Neural Networks in Pattern Recognition (ANNPR'16)*, 2016, S. 312–323

[107] HONG, Hai ; NEVEN, H. ; MALSBURG, C. von d.: Online facial expression recognition based on personalized galleries. In: *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 1998, S. 354–359

[108] HONOLD, Frank: *Interaktionsmanagement und Modalitätsarbitrierung für eine adaptive und multimodale Mensch-Computer Interaktion*, Ulm University, Doktorarbeit, 2017

[109] HONOLD, Frank ; BERCHER, Pascal ; RICHTER, Felix ; NOTHDURFT, Florian ; GEIER, Thomas ; BARTH, Roland ; HÖRNLE, Thilo ; SCHÜSSEL, Felix ; REUTER, Stephan ; RAU, Matthias ; BERTRAND, Gregor ; SEEGEBARTH, Bastian ; KURZOK, Peter ; SCHATTENBERG, Bernd ; MINKER, Wolfgang ; WEBER, Michael ; BIUNDO, Susanne: Companion-Technology: Towards User- and Situation-Adaptive Functionality of Technical Systems. In: *Intelligent Environments (IE), 2014 10th International Conference on*, IEEE, 2014, S. 378–381

[110] HONOLD, Frank ; SCHÜSSEL, Felix ; WEBER, Michael: Adaptive Probabilistic Fission for Multimodal Systems. In: *Proceedings of the 24th Australian Computer-Human Interaction Conference*, ACM, 2012 (OzCHI '12). – ISBN 978–1–4503–1438–1, S. 222–231

[111] HOROWITZ, Mardi J.: *Stress response syndromes: PTSD, grief, and adjustment disorders*. Jason Aronson, 1997

[112] HÖRNLE, Thilo ; TORNOW, Michael ; HONOLD, Frank ; SCHWEGLER, Reinhard ; HEINEMANN, Ralph ; BIUNDO, Susanne ; WENDEMUTH, Andreas: Companion-Systems: A Reference Architecture. In: *Companion Technology*. Springer, 2017, S. 449–469

[113] HUANG, Chung-Lin ; HUANG, Yu-Ming: Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification. In: *Journal of Visual Communication and Image Representation* 8 (1997), Nr. 3, S. 278 – 290. http://dx.doi.org/10.1006/jvci.1997.0359. – DOI 10.1006/jvci.1997.0359. – ISSN 1047–3203

[114] HUANG, Jui-Ting ; LI, Jinyu ; YU, Dong ; DENG, Li ; GONG, Yifan: Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* IEEE, 2013, S. 7304–7308

[115] HUDLICKA, Eva: To feel or not to feel: The role of affect in human–computer interaction. In: *International journal of human-computer studies* 59 (2003), Nr. 1, S. 1–32

[116] HUIJNEN, Claire ; BADII, Atta ; HEUVEL, Herjan van d. ; CALEB-SOLLY, Praminda ; THIEMERT, Daniel: "Maybe it becomes a buddy, but do not call it a robot"–Seamless Cooperation between Companion Robotics and Smart Homes. In: *Ambient Intelligence* (2011), S. 324–329

[117] ITAKURA, Fumitada: Line spectrum representation of linear predictor coefficients of

speech signals. In: *The Journal of the Acoustical Society of America* 57 (1975), Nr. S1, S. S35–S35

[118] IZARD, Carroll E.: *Human emotions*. Springer Science & Business Media, 2013

[119] KÄCHELE, Markus ; WERNER, Philipp ; AL-HAMADI, Ayoub ; PALM, Günther ; WALTER, Steffen ; SCHWENKER, Friedhelm: Bio-visual fusion for person-independent recognition of pain intensity. In: *International Workshop on Multiple Classifier Systems* Springer, 2015, S. 220–230

[120] KAMIJO, Ken-ichi ; TANIGAWA, Tetsuji: Stock price pattern recognition-a recurrent neural network approach. In: *Neural Networks, 1990., 1990 IJCNN International Joint Conference on* IEEE, 1990, S. 215–221

[121] KANADE, Takeo ; COHN, Jeffrey F. ; TIAN, Yingli: Comprehensive database for facial expression analysis. In: *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on* IEEE, 2000, S. 46–53

[122] KAPUR, Asha ; KAPUR, Ajay ; VIRJI-BABUL, Naznin ; TZANETAKIS, George ; DRIESSEN, Peter F.: Gesture-Based Affective Computing on Motion Capture Data. In: *Affective Computing and Intelligent Interaction*. Springer, 2005, S. 1–7

[123] KÄCHELE, Markus ; GLODEK, Michael ; ZHARKOV, Dimitrij ; MEUDT, Sascha ; SCHWENKER, Friedhelm: Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. In: MARSICO, Maria D. (Hrsg.) ; TABBONE, Antoine (Hrsg.) ; FRED, Ana (Hrsg.): *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM'14)*, SciTePress, 2014, S. 671–678

[124] KÄCHELE, Markus ; GLODEK, Michael ; ZHARKOV, Dimitrij ; MEUDT, Sascha ; SCHWENKER, Friedhelm: Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. In: *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRA)*, SciTePress, 2014, S. 671–678

[125] KÄCHELE, Markus ; SCHELS, Martin ; MEUDT, Sascha ; KESSLER, Viktor ; GLODEK, Michael ; THIAM, Patrick ; TSCHECHNE, Stephan ; PALM, Günther ; SCHWENKER, Friedhelm: On Annotation and Evaluation of Multi-Modal Corpora in Affective Human-Computer Interaction. In: MARSICO, Maria D. (Hrsg.) ; TABBONE, Antoine (Hrsg.) ; FRED, Ana (Hrsg.): *Proceedings of Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction (MA3HMI'14)*, SciTePress, 2014, S. 35–44

[126] KÄCHELE, Markus ; SCHELS, Martin ; MEUDT, Sascha ; PALM, Günther ; SCHWENKER, Friedhelm: Revisiting the EmotiW challenge: how wild is it really? In: *Journal on Multimodal User Interfaces (JMUI)* (2016), S. 1–12

[127] KÄCHELE, Markus ; SCHELS, Martin ; SCHWENKER, Friedhelm: Inferring depression and affect from application dependent meta knowledge. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* ACM, 2014, S. 41–48

[128] KÄCHELE, Markus ; ZHARKOV, Dimitrij ; MEUDT, Sascha ; SCHWENKER, Friedhelm: Prosodic, spectral and voice quality feature selection using a long-term stopping criterion for audio-based emotion recognition. In: *Proceedings of the International Conference on Pattern Recognition (ICPR'14)*, 2014, S. 803–808

[129] KESSLER, Viktor ; KÄCHELE, Markus ; MEUDT, Sascha ; SCHWENKER, Friedhelm ; PALM, Günther: Machine Learning Driven Heart Rate Detection with Camera Photoplethysmography in Time Domain. In: *Proceedings of Artificial Neural Networks in Pattern Recognition (ANNPR'16)*, 2016, S. 324–334

[130] KIM, Jonghwa: *Bimodal emotion recognition using speech and physiological changes*. Citeseer, 2007

[131] KINDSVATER, Daniel ; MEUDT, Sascha ; SCHWENKER, Friedhelm: Fusion Architectures for Multimodal Cognitive Load Recognition. In: *Proceedings of Multimodal pattern recognition of social signals in human computer interaction (MPRSS 2016)*, 2016, S. 1–8

[132] KING, Ross D. ; WHELAN, Kenneth E. ; JONES, Ffion M. ; REISER, Philip G. ; BRYANT, Christopher H. ; MUGGLETON, Stephen H. ; KELL, Douglas B. ; OLIVER, Stephen G.: Functional genomic hypothesis generation and experimentation by a robot scientist. In: *Nature* 427 (2004), Nr. 6971, S. 247–252

[133] KIPP, Michael: Anvil-a generic annotation tool for multimodal dialogue. In: *Seventh European Conference on Speech Communication and Technology*, 2001

[134] KIPP, Michael ; MARTIN, Jean-Claude: Gesture and Emotion: Can Basic Gestural Form Features Discriminate Emotions? In: *3rd International Conference on Affective Computing and Intelligent Interaction Workshops* IEEE, 2009, S. 1–8

[135] KNAPP, Mark L. ; HALL, Judith A. ; HORGAN, Terrence G.: *Nonverbal communication in human interaction*. Cengage Learning, 2013

[136] KOHAVI, Ron ; JOHN, George H.: Wrappers for feature subset selection. In: *Ar-*

tif. Intell. 97 (1997), Nr. 1-2, S. 273–324. http://dx.doi.org/10.1016/S0004-3702(97)00043-X. – DOI 10.1016/S0004–3702(97)00043–X. – ISSN 0004–3702

[137] LARA, Oscar D. ; LABRADOR, Miguel A.: A survey on human activity recognition using wearable sensors. In: *IEEE Communications Surveys and Tutorials* 15 (2013), Nr. 3, S. 1192–1209

[138] LEE, C. M. ; YILDIRIM, S. ; BULUT, M. ; KAZEMZADEH, A. ; BUSSO, C. ; DENG, Z. ; LEE, S. ; NARAYANAN, S. S.: Emotion recognition based on phoneme classes. In: *Proceedings of ICSLP 04*, 2004

[139] LEWIS, David D. ; GALE, William A.: A sequential algorithm for training text classifiers. In: *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* Springer-Verlag New York, Inc., 1994, S. 3–12

[140] LIEN, James J. ; KANADE, Takeo ; COHN, Jeffrey F. ; LI, Ching-Chung: Automated facial expression recognition based on FACS action units. In: *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on* IEEE, 1998, S. 390–395

[141] LIN, Shili: Rank Aggregation Methods. In: *Wiley Interdisciplinary Reviews: Computational Statistics* (2010), S. 555–570

[142] LISETTI, Christine L. ; SCHIANO, Diane J.: Automatic Facial Expression Interpretation: Where Human-Computer Interaction, Artificial Intelligence and Cognitive Science Intersect. In: *Pragmatics and Cognition (Special Issue on Facial Information Processing: A Multidisciplinary Perspective)* 8 (2000), Nr. 1, S. 185–235

[143] LISETTI, Christine L. ; NASOZ, Fatma: Using noninvasive wearable computers to recognize human emotions from physiological signals. In: *EURASIP Journal on Advances in Signal Processing* 2004 (2004), Nr. 11, S. 929414

[144] LITTLEWORT, Gwen ; WHITEHILL, Jacob ; WU, Tingfan ; FASEL, Ian ; FRANK, Mark ; MOVELLAN, Javier ; BARTLETT, Marian: The computer expression recognition toolbox (CERT). In: *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* IEEE, 2011, S. 298–305

[145] LOGAN, Beth u. a.: Mel Frequency Cepstral Coefficients for Music Modeling. In: *ISMIR*, 2000

[146] LUCEY, Patrick ; COHN, Jeffrey F. ; KANADE, Takeo ; SARAGIH, Jason ; AMBADAR,

Zara ; MATTHEWS, Iain: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on* IEEE, 2010, S. 94–101

[147] LUENGO, Iker ; NAVAS, Eva ; HERNÁEZ, Inmaculada: Feature analysis and evaluation for automatic emotion identification in speech. In: *Multimedia, IEEE Transactions on* 12 (2010), Nr. 6, S. 490–501

[148] LUGGER, Marko ; YANG, Bin: Classification of different speaking groups by means of voice quality parameters. In: *ITG-Fachbericht-Sprachkommunikation 2006* (2006)

[149] MARSHALL, Ian W. ; ROADKNIGHT, Christopher ; WOKOMA, Ibiso ; SACKS, Lionel: Self-organising Sensor networks. In: *1st UK-UbiNet Workshop*, 2003

[150] MEUDT, Sascha ; BIGALKE, Lutz ; SCHWENKER, Friedhelm: ATLAS - Annotation tool using partially supervised learning and Multi-view Co-learning in Human-Computer-Interaction Scenarios. In: *Proceedings of the 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA'12)*, IEEE, 2012, S. 1309–1312

[151] MEUDT, Sascha ; BIGALKE, Lutz ; SCHWENKER, Friedhelm: ATLAS – An Annotation Tool for HCI Data Utilizing Machine Learning Methods. In: *Proceedings of the 1st International Conference on Affective and Pleasurable Design (APD'12) [jointly with the 4th International Conference on Applied Human Factors and Ergonomics (AHFE'12)]*, CRC Press, 2012 (Advances in Human Factors and Ergonomics Series), S. 5347–5352

[152] MEUDT, Sascha ; GLODEK, Michael ; SCHELS, Martin ; SCHWENKER, Friedhelm: Multi-View Video Based Tracking and Audio-Visual Identification of Persons in a Human-Computer-Interaction Scenario. In: *Proceedings of IEEE International Conference on Cybernetics (CybConf'13)*, Springer, 2013, S. 1–8

[153] MEUDT, Sascha ; SCHMIDT-WACK, Miriam ; HONOLD, Frank ; SCHÜSSEL, Felix ; WEBER, Michael ; SCHWENKER, Friedhelm ; PALM, Günther: Going Further in Affective Computing: How Emotion Recognition Can Improve Adaptive User Interaction. In: ESPOSITO, Anna (Hrsg.) ; JAIN, Lakhmi C. (Hrsg.): *Toward Robotic Socially Believable Behaving Systems - Volume I: Modeling Emotions*. Springer, 2016, S. 73–103

[154] MEUDT, Sascha ; SCHWENKER, Friedhelm: On Instance Selection in Audio Based Emotion Recognition. In: *Proceedings of the 5th IAPR TC3 Workshop on Artificial Neural Networks for Pattern Recognition (ANNPR'12)*, Springer, 2012, 186-192

[155] MEUDT, Sascha ; SCHWENKER, Friedhelm: Enhanced Autocorrelation in Real World Emotion Recognition. In: *Proceedings of the 16th ACM International Conference on Multimodal Interaction (ICMI'14)*, ACM, 2014, S. 502–507

[156] MEUDT, Sascha ; SCHWENKER, Friedhelm: ATLAS - Machine learning based annotation of multimodal data recorded in human-computer interaction scenarios. In: *Proceedings of the 1st International Symposium on Companion-Technology (ISCT 2015)*, ACM, 2015, S. 181–186

[157] MEUDT, Sascha ; SIEGERT, Ingo ; SCHÜSSEL, Felix ; SCHMIDT, Miriam ; REUTER, Stephan ; LAYHER, Georg ; HÖRNLE, Thilo ; HANDRICH, Sebastian ; AL-HAMADI, Ayoub ; DIETMAYER, Klaus ; NEUMANN, Heiko ; PALM, Günther ; SCHWENKER, Friedhelm ; WENDEMUTH, Andreas: Multi-Modal Information Processing in Companion Systems—A Ticket Purchase System. In: BIUNDO, Susanne (Hrsg.) ; WENDEMUTH, Andreas (Hrsg.): *Companion Technology*. 1. Springer, 2016

[158] MEUDT, Sascha ; ZHARKOV, Dimitri ; KÄCHELE, Markus ; SCHWENKER, Friedhelm: Multi Classifier Systems and Forward Backward Feature Selection Algorithms to Classify Emotional Coloured Speech. In: *Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI'13)*, ACM, 2013 (ICMI 2013). – ISBN 978–1–4503–2129–7, S. 551–556

[159] MITRA, Sushmita ; ACHARYA, Tinku: Gesture recognition: A survey. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37 (2007), Nr. 3, S. 311–324

[160] In: MORRIS, R.R. ; MCDUFF, D.: *Crowdsourcing Techniques for Affective Computing*. Oxford University Press, 2014

[161] MUNASSAR, Nabil Mohammed A. ; GOVARDHAN, A: A comparison between five models of software engineering. In: *IJCSI International Journal of Computer Science Issues* 7 (2010), Nr. 5, S. 94–101

[162] MUSLEA, Ion ; MINTON, Steven ; KNOBLOCK, Craig A.: Active + semi-supervised learning= robust multi-view learning. In: *ICML* Bd. 2, 2002, S. 435–442

[163] NASRABADI, Nasser M.: Pattern recognition and machine learning. In: *Journal of electronic imaging* 16 (2007), Nr. 4, S. 049901

[164] NGUYEN, Thien-Nghia ; MICHAELIS, Bernd ; AL-HAMADI, Ayoub ; TORNOW, Michael ; MEINECKE, Marc-Michael: Stereo-camera-based urban environment perception using

occupancy grid and object tracking. In: *IEEE Transactions on Intelligent Transportation Systems* 13 (2012), Nr. 1, S. 154–165

[165] NICHOLSON, J. ; TAKAHASHI, K. ; NAKATSU, R.: Emotion recognition in speech using neural networks. In: *Neural Computing and Applications* 9 (2000), S. 290–296

[166] NIESE, Robert ; AL-HAMADI, Ayoub ; PANNING, Axel ; MICHAELIS, Bernd: Emotion recognition based on 2d-3d facial feature extraction from color image sequences. In: *Journal of Multimedia* 5 (2010), Nr. 5, S. 488–500

[167] NWE, Tin L. ; FOO, Say W. ; DE SILVA, Liyanage C.: Speech emotion recognition using hidden Markov models. In: *Speech communication* 41 (2003), Nr. 4, S. 603–623

[168] OJALA, Timo ; PIETIKÄINEN, Matti ; HARWOOD, David: A comparative study of texture measures with classification based on featured distributions. In: *Pattern Recognition* 29 (1996), Nr. 1, S. 51 – 59. http://dx.doi.org/10.1016/0031-3203(95)00067-4. – DOI 10.1016/0031–3203(95)00067–4. – ISSN 0031–3203

[169] PAESCHKE, Astrid: *Prosodische Analyse emotionaler Sprechweise*. Logos-Verlag, 2003

[170] PALM, Günther ; SCHWENKER, Friedhelm: Sensor-Fusion in Neural Networks. In: SHAHBAZIAN, Elisa (Hrsg.) ; ROGOVA, Galina (Hrsg.) ; DEWEERT, Michael J. (Hrsg.): *Harbour Protection Through Data Fusion Technologies*. Springer, 2009, S. 299–306

[171] PANTIC, Maja ; ROTHKRANTZ, Leon J.: Toward an affect-sensitive multimodal human-computer interaction. In: *Proceedings of the IEEE* 91 (2003), Nr. 9, S. 1370–1390

[172] PARASURAMAN, Raja ; SHERIDAN, Thomas B. ; WICKENS, Christopher D.: A model for types and levels of human interaction with automation. In: *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30 (2000), Nr. 3, S. 286–297

[173] PICARD, Rosalind W.: Affective Computing for HCI. In: *HCI (1)*, 1999, S. 829–833

[174] PICARD, Rosalind W.: Affective computing: challenges. In: *International Journal of Human-Computer Studies* 59 (2003), Nr. 1, S. 55–64

[175] PICARD, Rosalind W. ; KLEIN, Jonathan: Computers that recognise and respond to user emotion: theoretical and practical implications. In: *Interacting with computers* 14 (2002), Nr. 2, S. 141–169

[176] PICARD, Rosalind W. ; PICARD, Roalind: *Affective computing*. Bd. 252. MIT press Cambridge, 1997

[177] PIERRE-YVES, Oudeyer: The production and recognition of emotions in speech: features and algorithms. In: *International Journal of Human-Computer Studies* 59 (2003), Nr. 1, S. 157–183

[178] PLATT, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Advances in large margin classifiers* 10 (1999), Nr. 3, S. 61–74

[179] PLUTCHIK, Robert: A general psychoevolutionary theory of emotion. In: *Theories of emotion* 1 (1980), Nr. 3-31, S. 4

[180] PLUTCHIK, Robert: *The psychology and biology of emotion.* HarperCollins College Publishers, 1994

[181] POHL, Klaus: *Requirements engineering: fundamentals, principles, and techniques*. Springer Publishing Company, Incorporated, 2010

[182] PRESSMAN, Roger S.: *Software engineering: a practitioner's approach*. Palgrave Macmillan, 2005

[183] PUDIL, Pavel ; NOVOVIČOVÁ, Jana ; KITTLER, Josef: Floating search methods in feature selection. In: *Pattern recognition letters* 15 (1994), Nr. 11, S. 1119–1125

[184] RABINER, L. R.: A tutorial on hidden Markov models and selected applications in speech recognition. In: *Proceedings of the IEEE* 77 (1989), Nr. 2, S. 257–286

[185] RABINER, Lawrence R. ; JUANG, Biing-Hwang: *Fundamentals of speech recognition*. PTR Prentice Hall, 1993

[186] RAMMSTEDT, Beatrice ; JOHN, Oliver P.: Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. In: *Journal of research in Personality* 41 (2007), Nr. 1, S. 203–212

[187] RENTSCH, Tim: Object oriented programming. In: *ACM Sigplan Notices* 17 (1982), Nr. 9, S. 51–57

[188] REUTER, Stephan ; WILKING, Benjamin ; WIEST, Jurgen ; MUNZ, Michael ; DIETMAYER, Klaus: Real-time multi-object tracking using random finite sets. In: *IEEE Transactions on Aerospace and Electronic Systems* 49 (2013), Nr. 4, S. 2666–2678

[189] RINGEVAL, Fabien ; SCHULLER, Björn ; VALSTAR, Michel ; JAISWAL, Shashank ; MARCHI, Erik ; LALANNE, Denis ; COWIE, Roddy ; PANTIC, Maja: Avec 2015: The first affect recognition challenge bridging across audio, video, and physiological data.

In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* ACM, 2015, S. 3–8

[190] RIPLEY, Brian D.: *Pattern recognition and neural networks*. Cambridge university press, 2007

[191] RÖSNER, Dietmar F. ; FROMMER, Jörg ; FRIESEN, Rafael ; HAASE, Matthias ; LANGE, Julia ; OTTO, Mirko: LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions. In: *LREC*, 2012, S. 2559–2566

[192] ROY, Nicholas ; MCCALLUM, Andrew: Toward optimal active learning through monte carlo estimation of error reduction. In: *ICML, Williamstown* (2001), S. 441–448

[193] RUF, Tobias ; ERNST, Andreas ; KÜBLBECK, Christian: Face detection with the sophisticated high-speed object recognition engine (SHORE). In: *Microelectronic Systems*. Springer, 2011, S. 243–252

[194] RUKAVINA, Stefanie ; GRUSS, Sascha ; WALTER, Steffen ; HOFFMANN, Holger ; TRAUE, Harald C.: Open emorec II - A multimodal corpus of human-computer interaction. In: *International Journal of Computer, Electrical, Automation, Control and Information Engineering* 9 (2015), Nr. 5, S. 977–983

[195] RUSSELL, James A.: Core affect and the psychological construction of emotion. In: *Psychological review* 110 (2003), Nr. 1, S. 145

[196] RUSSELL, James A. ; BACHOROWSKI, Jo-Anne ; FERNÁNDEZ-DOLS, José-Miguel: Facial and Vocal Expressions of Emotion. In: *Annual Review of Psychology* 54 (2003), Nr. 1, S. 329–349

[197] RUSSELL, James A. ; MEHRABIAN, Albert: Evidence for a three-factor theory of emotions. In: *Journal of research in Personality* 11 (1977), Nr. 3, S. 273–294

[198] SAVVIDOU, Soultana: *Validation of the Feeltrace tool for recording impressions of expressed emotion*, Queen's University Belfast, Diss., 2011

[199] SCHALKOFF, Robert J.: *Artificial neural networks*. Bd. 1. McGraw-Hill New York, 1997

[200] SCHEFFER, Tobias ; DECOMAIN, Christian ; WROBEL, Stefan: Active hidden markov models for information extraction. In: *International Symposium on Intelligent Data Analysis* Springer, 2001, S. 309–318

[201] SCHELS, Martin ; GLODEK, Michael ; MEUDT, Sascha ; SCHERER, Stefan ; SCHMIDT, Miriam ; LAYHER, Georg ; TSCHECHNE, Stephan ; BROSCH, Tobias ; HRABAL, David ;

WALTER, Steffen ; TRAUE, Harald ; PALM, Günther ; NEUMANN, Heiko ; SCHWENKER, Friedhelm: Multi-Modal Classifier-Fusion for the Recognition of Emotions. In: ROJC, Matej (Hrsg.) ; CAMPBELL, Nick (Hrsg.): *Coverbal Synchrony in Human-Machine Interaction*. 1. CRC Press, 2013, S. 73–98

[202] SCHELS, Martin ; GLODEK, Michael ; MEUDT, Sascha ; SCHMIDT, Miriam ; HRABAL, David ; BÖCK, Ronald ; WALTER, Steffen ; SCHWENKER, Friedhelm: Multi-Modal Classifier-Fusion for the Classification of Emotional States in WOZ Scenarios. In: *Proceedings of the 1st International Conference on Affective and Pleasurable Design (APD'12) [jointly with the 4th International Conference on Applied Human Factors and Ergonomics (AHFE'12)]*, CRC Press, 2012 (Advances in Human Factors and Ergonomics Series), S. 5337–5346

[203] *Kapitel* 23. In: SCHERER, K. R. ; JOHNSTONE, T. ; KLASMEYER, G.: *Handbook of Affective Sciences - Vocal expression of emotion*. Oxford University Press, 2003 (Affective Science), S. 433–456

[204] SCHERER, Klaus R.: What are emotions? And how can they be measured? In: *Social science information* 44 (2005), Nr. 4, S. 695–729

[205] SCHERER, Stefan ; KANE, John ; GOBL, Christer ; SCHWENKER, Friedhelm: Investigating Fuzzy-Input Fuzzy-Output Support Vector Machines for Robust Voice Quality Classification. In: *Computer Speech and Language* 27 (2012), Nr. 1, S. 263–287

[206] SCHERER, Stefan ; OUBBATI, Mohamed ; SCHWENKER, Friedhelm ; PALM, Günther: Real-time emotion recognition from speech using echo state networks. In: *Artificial Neural Networks in Pattern Recognition*. Springer Berlin Heidelberg, 2008, S. 205–216

[207] SCHERER, Stefan ; SCHWENKER, Friedhelm ; CAMPBELL, Nick ; PALM, Günther: Multimodal laughter detection in natural discourses. In: *Human Centered Robot Systems* 6 (2009), S. 111–120

[208] SCHERER, Stefan ; SCHWENKER, Friedhelm ; PALM, Günther: Classifier fusion for emotion recognition from speech. In: *3rd IET International Conference on Intelligent Environments*, IET, 2007, S. 152–155

[209] SCHERER, Stefan ; SCHWENKER, Friedhelm ; PALM, Günther: Emotion recognition from speech using multi-classifier systems and rbf-ensembles. In: *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*. Springer Berlin Heidelberg, 2008, S. 49–70

[210] SCHERER, Stefan ; SCHWENKER, Friedhelm ; PALM, Günther: Classifier fusion for emotion recognition from speech. In: MINKER, Wolfgang (Hrsg.) ; WEBER, Michael (Hrsg.) ; HAGRAS, Hani (Hrsg.) ; CALLAGAN, V. (Hrsg.) ; KAMEAS, A. D. (Hrsg.): *Advanced intelligent environments*. Springer, 2009, S. 95–117

[211] SCHERER, Stefan ; SIEGERT, Ingo ; BIGALKE, Lutz ; MEUDT, Sascha: Developing an Expressive Speech Labeling Tool Incorporating the Temporal Characteristics of Emotion. In: CALZOLARI, N. (Hrsg.) ; CHOUKRI, K. (Hrsg.) ; MAEGAARD, B. (Hrsg.) ; MARIANI, J. (Hrsg.) ; ODJIK, J. (Hrsg.) ; PIPERIDIS, S. (Hrsg.) ; ROSNER, M. (Hrsg.) ; TAPIAS, D. (Hrsg.): *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, ELRA, 2010, S. 1172–1175

[212] SCHÖLKOPF, Bernhard ; WILLIAMSON, Robert ; SMOLA, Alex ; SHAWE-TAYLOR, John ; PLATT, John: Support vector Method for Novelty Detection. In: *Advances In Neural Information Processing Systems* 12 (1999), S. 582–588

[213] SCHMIDHUBER, Jürgen: Deep learning in neural networks: An overview. In: *Neural networks* 61 (2015), S. 85–117

[214] SCHMIDT, Thomas ; WÖRNER, Kai: Erstellen und analysieren von gesprächskorpora mit exmaralda. In: *Gesprächsforschung* 6 (2005), S. 171–195

[215] SCHOHN, Greg ; COHN, David: Less is more: Active learning with support vector machines. In: *ICML*, 2000, S. 839–846

[216] SCHOLLMEIER, Rüdiger: A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications. In: *Peer-to-Peer Computing, 2001. Proceedings. First International Conference on* IEEE, 2001, S. 101–102

[217] SCHRÖDER, Marc: The SEMAINE API: towards a standards-based framework for building emotion-oriented systems. In: *Advances in human-computer interaction* 2010 (2010), S. 2

[218] SCHÜSSEL, Felix ; HONOLD, Frank ; BUBALO, Nikola ; HUCKAUF, Anke ; TRAUE, Harald ; HAZER-RAU, Dilana: In-depth analysis of multimodal interaction: An explorative paradigm. In: *HCI International 2016*. Springer Berlin Heidelberg, July 2016 (Communications in Computer and Information Science)

[219] SCHÜSSEL, Felix ; HONOLD, Frank ; BUBALO, Nikola ; HUCKAUF, Anke ; WEBER, Michael: Getting rid of „ok google“: Individual multimodal input adaption in real world applications. In: *Proc. of the Int'l Symposium on Companion Technology (ISCT 2015)*,

2015

[220] SCHÜSSEL, Felix ; HONOLD, Frank ; SCHMIDT, Miriam ; BUBALO, Nikola ; HUCKAUF, Anke ; WEBER, Michael: Multimodal interaction history and its use in error detection and recovery. In: *Proceedings of the 16th International Conference on Multimodal Interaction* ACM, 2014, S. 164–171

[221] SCHÜSSEL, Felix ; HONOLD, Frank ; WEBER, Michael ; SCHMIDT, Miriam ; BUBALO, Nikola ; HUCKAUF, Anke: Multimodal Interaction History and its use in Error Detection and Recovery. In: *Proceedings of the 16th ACM International Conference on Multimodal Interaction*, ACM, 2014 (ICMI '14). – ISBN 978–1–4503–2885–2, S. 164–171

[222] SCHULLER, Björn ; ARSIC, Dejan ; RIGOLL, Gerhard ; WIMMER, Matthias ; RADIG, Bernd: Audiovisual behavior modeling by combined feature spaces. In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on* Bd. 2 IEEE, 2007, S. II–733

[223] SCHULLER, Björn ; VALSTAR, Michel ; EYBEN, Florian ; MCKEOWN, Gary ; COWIE, Roddy ; PANTIC, Maja: Avec 2011–the first international audio/visual emotion challenge. In: *Affective Computing and Intelligent Interaction* (2011), S. 415–424

[224] SCHULLER, Björn ; VALSTER, Michel ; EYBEN, Florian ; COWIE, Roddy ; PANTIC, Maja: AVEC 2012: the continuous audio/visual emotion challenge. In: *Proceedings of the 14th ACM international conference on Multimodal interaction* ACM, 2012, S. 449–456

[225] SCHWABER, Ken: Scrum development process. In: *Business object design and implementation*. Springer, 1997, S. 117–134

[226] SCHWENKER, Friedhelm ; BÖCK, Ronald ; SCHELS, Martin ; MEUDT, Sascha ; SIEGERT, Ingo ; GLODEK, Michael ; SCHMIDT-WACK, Miriam ; THIAM, Patrick ; WENDEMUTH, Andreas ; KRELL, Gerald: Multimodal Affect Recognition in the Context of Human-Computer Interaction for Companion Systems. In: BIUNDO, Susanne (Hrsg.) ; WENDEMUTH, Andreas (Hrsg.): *Companion Technology*. 1. Springer, 2016

[227] SCHWENKER, Friedhelm ; FREY, Markus ; GLODEK, Michael ; KÄCHELE, Markus ; MEUDT, Sascha ; SCHELS, Martin ; SCHMIDT, Miriam: A New Multi-class Fuzzy Support Vector Machine Algorithm. In: NEAMAT, El G. (Hrsg.) ; SCHWENKER, Friedhelm (Hrsg.) ; CHENG, Suen (Hrsg.): *Proceedings of the 6th IAPR TC3 Workshop on Artificial Neural Networks for Pattern Recognition (ANNPR'14)*, Springer, 2014 (LNCS 8774), S. 153–164

[228] SCHWENKER, Friedhelm ; KESTLER, Hans A. ; PALM, Günther: Three learning phases for radial-basis-function networks. In: *Neural networks* 14 (2001), Nr. 4, S. 439–458

[229] SCHWENKER, Friedhelm ; MEUDT, Sascha: On instance selection in multiple classifier systems. In: *Proceedings of the 36th Annual Conference of the German Classification Society (GfKl'12)* Bd. Data Analysis, Machine Learning and Knowledge Discovery, 2012, S. 50

[230] SCHWENKER, Friedhelm ; SCHERER, Stefan ; SCHMIDT, Miriam ; SCHELS, Martin ; GLODEK, Michael: Multiple Classifier Systems for the Recognition of Human Emotions. In: GAYAR, Neamat E. (Hrsg.) ; KITTLER, Josef (Hrsg.) ; ROLI, Fabio (Hrsg.): *Proceedings of the 9th International Workshop on Multiple Classifier Systems (MCS'10)*, Springer, 2010 (LNCS 5997), S. 315–324

[231] SEBE, Nicu ; LEW, Michael S. ; COHEN, Ira ; GARG, Ashutosh ; HUANG, Thomas S.: Emotion recognition using a cauchy naive bayes classifier. In: *Pattern Recognition, 2002. Proceedings. 16th International Conference on* Bd. 1 IEEE, 2002, S. 17–20

[232] SELTING, Margret ; AUER, Peter ; BARDEN, Birgit ; BERGMANN, Jörg ; COUPER-KUHLEN, Elizabeth ; GÜNTHNER, Susanne ; MEIER, Christoph ; QUASTHOFF, Uta ; SCHLOBINSKI, Peter ; UHMANN, Susanne u. a.: Gesprachsanalytisches transkriptionssystem (GAT). In: *Linguistische Berichte* (1998), S. 91–122

[233] SELTING, Margret ; AUER, Peter ; BARTH-WEINGARTEN, Dagmar ; BERGMANN, Jörg R ; BERGMANN, Pia ; BIRKNER, Karin ; COUPER-KUHLEN, Elizabeth ; DEPPERMANN, Arnulf ; GILLES, Peter ; GÜNTHNER, Susanne u. a.: Gesprächsanalytisches transkriptionssystem 2 (GAT 2). In: *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion* (2009)

[234] SETTLES, Burr: Active learning literature survey. In: *University of Wisconsin, Madison* 52 (2010), Nr. 55-66, S. 11

[235] SETTLES, Burr: Active learning. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6 (2012), Nr. 1, S. 1–114

[236] SETTLES, Burr ; CRAVEN, Mark: An analysis of active learning strategies for sequence labeling tasks. In: *Proceedings of the conference on empirical methods in natural language processing* Association for Computational Linguistics, 2008, S. 1070–1079

[237] SETTLES, Burr ; CRAVEN, Mark ; RAY, Soumya: Multiple-instance active learning. In: *Advances in neural information processing systems*, 2008, S. 1289–1296

[238] SEUNG, H S. ; OPPER, Manfred ; SOMPOLINSKY, Haim: Query by committee. In: *Proceedings of the fifth annual workshop on Computational learning theory* ACM, 1992, S. 287–294

[239] SHAFER, Glenn: *A mathematical theory of evidence*. Bd. 42. Princeton university press, 1976

[240] SHAFER, Glenn: Dempster-shafer theory. In: *Encyclopedia of artificial intelligence* (1992), S. 330–331

[241] SHANNON, Claude E.: A mathematical theory of communication. In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5 (2001), Nr. 1, S. 3–55

[242] SHNEIDERMAN, Ben: Creativity support tools: Accelerating discovery and innovation. In: *Communications of the ACM* 50 (2007), Nr. 12, S. 20–32

[243] SIEGERT, Ingo ; REUTHER, Stefan ; SCHÜSSEL, Felix ; LAYHER, Georg ; HÖRNLE, Thilo ; MEUDT, Sascha ; WENDEMUTH, Andreas: Multimodal Information Processing: The Ticket Purchase - a Demonstration Scenario of the SFB/TRR-62. In: *Elektronische Sprachsignalverarbeitung*, 2016, S. 111–118

[244] SMOLA, Alex J. ; SCHÖLKOPF, Bernhard: A tutorial on support vector regression. In: *Statistics and computing* 14 (2004), Nr. 3, S. 199–222

[245] SMYTH, Padhraic ; FAYYAD, Usama M. ; BURL, Michael C. ; PERONA, Pietro ; BALDI, Pierre: Inferring ground truth from subjective labelling of venus images. In: *Advances in neural information processing systems*, 1995, S. 1085–1092

[246] SOBEL, I. ; FELDMAN, G.: *A 3x3 Isotropic Gradient Operator for Image Processing*. 1968. – Never published but presented at a talk at the Stanford Artificial Project

[247] SOLEYMANI, Mohammad ; LICHTENAUER, Jeroen ; PUN, Thierry ; PANTIC, Maja: A multimodal database for affect recognition and implicit tagging. In: *IEEE Transactions on Affective Computing* 3 (2012), Nr. 1, S. 42–55

[248] STEPHANIDIS, Constantine ; SALVENDY, Gavriel ; AKOUMIANAKIS, Demosthenes ; ARNOLD, Albert ; BEVAN, Nigel ; DARDAILLER, Daniel ; EMILIANI, Pier L. ; IAKOVIDIS, Ilias ; JENKINS, Phil ; KARSHMER, Arthur u. a.: Toward an Information Society for All: HCI challenges and R&D recommendations. In: *International Journal of Human-Computer Interaction* 11 (1999), Nr. 1, S. 1–28

[249] SUN, Deqing ; ROTH, S. ; BLACK, M.J.: Secrets of optical flow estimation and their

principles. In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010. – ISSN 1063–6919, S. 2432–2439

[250] TANG, X ; SHU, L: Classification of Electrocardiogram Signals with RS and Quantum Neural Networks. In: *International Journal of Multimedia and Ubiquitous Engineering* 9 (2014), Nr. 2, S. 363–372

[251] TAX, David M. ; DUIN, Robert P.: Support Vector Data Description. In: *Machine Learning* 54 (2004), Nr. 1, S. 45–66

[252] THIAM, Patrick ; KÄCHELE, Markus ; SCHWENKER, Friedhelm ; PALM, Günther: Ensembles of Support Vector Data Description for Active Learning Based Annotation of Affective Corpora. In: *2015 IEEE Symposium Series on Computational Intelligence*, 2015, S. 1801–1807

[253] THIAM, Patrick ; MEUDT, Sascha ; KÄCHELE, Markus ; PALM, Günther ; SCHWENKER, Friedhelm: Detection of Emotional Events Utilizing Support Vector Methods in an Active Learning HCI Scenario. In: *Proceedings of the 2014 Workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems (ERM4HCI'14)*, 2014, S. 31–36

[254] THIAM, Patrick ; MEUDT, Sascha ; PALM, Günther ; SCHWENKER, Friedhelm: A Temporal Dependency Based Multi-modal Active Learning Approach for Audiovisual Event Detection. In: *Neural Processing Letters* (2017). http://dx.doi.org/10.1007/s11063-017-9719-y. – DOI 10.1007/s11063–017–9719–y. – ISSN 1573–773X

[255] THIAM, Patrick ; MEUDT, Sascha ; SCHWENKER, Friedhelm ; PALM, Günther: Active Learning for Speech Event Detection in HCI. In: *Proceedings of Artificial Neural Networks in Pattern Recognition (ANNPR'16)*, 2016, S. 285–297

[256] TOLONEN, Tero ; KARJALAINEN, Matti: A computationally efficient multipitch analysis model. In: *Speech and Audio Processing, IEEE Transactions on* 8 (2000), Nr. 6, S. 708–716

[257] TORNOW, Michael ; AL-HAMADI, Ayoub ; BORRMANN, Vinzenz: A multi-agent mobile robot system with environment perception and HMI capabilities. In: *Signal and Image Processing Applications (ICSIPA), 2013 IEEE International Conference on* IEEE, 2013, S. 252–257

[258] TORNOW, Michael ; KASZUBIAK, Jens ; KUHN, Robert W. ; MICHAELIS, Bernd ;

SCHINDLER, Thomas: Hardware approach for real time machine stereo vision. In: *WMSCI 2005: 9th World Multi-Conference on Systemics, Cybernetics and Informatics* Bd. 5, 2005, S. 111–116

[259] TOU, Julius T. ; GONZALEZ, Rafael C.: *Pattern recognition principles*. Addison-Wesley, 1974

[260] TRAUE, Harald C. ; OHL, Frank ; BRECHMANN, André ; SCHWENKER, Friedhelm ; KESSLER, Henrik ; LIMBRECHT, Kerstin ; HOFFMANN, Holger ; SCHERER, Stefan ; KOTZYBA, Michael ; SCHECK, Andreas u. a.: A framework for emotions and dispositions in man-companion interaction. In: *Coverbal Synchrony in Human-Machine Interaction. New Hampshire, USA: Science Publishers* (2013), S. 99–140

[261] VALSTAR, Michel ; GRATCH, Jonathan ; SCHULLER, Björn ; RINGEVAL, Fabien ; LALANNE, Dennis ; TORRES TORRES, Mercedes ; SCHERER, Stefan ; STRATOU, Giota ; COWIE, Roddy ; PANTIC, Maja: Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge* ACM, 2016, S. 3–10

[262] VALSTAR, Michel ; SCHULLER, Björn ; SMITH, Kirsty ; ALMAEV, Timur ; EYBEN, Florian ; KRAJEWSKI, Jarek ; COWIE, Roddy ; PANTIC, Maja: Avec 2014: 3d dimensional affect and depression recognition challenge. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* ACM, 2014, S. 3–10

[263] VALSTAR, Michel ; SCHULLER, Björn ; SMITH, Kirsty ; EYBEN, Florian ; JIANG, Bihan ; BILAKHIA, Sanjay ; SCHNIEDER, Sebastian ; COWIE, Roddy ; PANTIC, Maja: AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge* ACM, 2013, S. 3–10

[264] VELANA, Maria ; GRUSS, Sascha ; LAYHER, Georg ; THIAM, Patrick ; ZHANG, Yan ; SCHORK, Daniel ; KESSLER, Viktor ; MEUDT, Sascha ; NEUMANN, Heiko ; KIM, Jonghwa u. a.: The senseemotion database: A multimodal database for the development and systematic validation of an automatic pain-and emotion-recognition system. In: *IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction* Springer, 2016, S. 127–139

[265] VERGIN, Rivarol ; O'SHAUGHNESSY, Douglas ; FARHAT, Azarshid: Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. In: *Speech and Audio Processing, IEEE Transactions on* 7 (1999), Nr. 5, S. 525–532

[266] VIOLA, Paul ; JONES, Michael: Rapid object detection using a boosted cascade of simple features. In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* Bd. 1 IEEE, 2001, S. I–I

[267] VOGT, Thurid ; ANDRÉ, Elisabeth: Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on* IEEE, 2005, S. 474–477

[268] WAGNER, Johannes ; LINGENFELSER, Florian ; BAUR, Tobias ; DAMIAN, Ionut ; KISTLER, Felix ; ANDRÉ, Elisabeth: The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In: *Proceedings of the 21st ACM international conference on Multimedia* ACM, 2013, S. 831–834

[269] WALTER, Steffen ; GRUSS, Sascha ; EHLEITER, Hagen ; TAN, Junwen ; TRAUE, Harald C. ; WERNER, Philipp ; AL-HAMADI, Ayoub ; CRAWCOUR, Stephen ; ANDRADE, Adriano O. ; SILVA, Gustavo M.: The BioVid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In: *Cybernetics (CYBCONF), 2013 IEEE International Conference on* IEEE, 2013, S. 128–131

[270] WEATHERS, Frank W. ; LITZ, Brett T. ; HERMAN, Debra S. ; HUSKA, Jennifer A. ; KEANE, Terence M. u. a.: The PTSD Checklist (PCL): Reliability, validity, and diagnostic utility. In: *annual convention of the international society for traumatic stress studies, San Antonio, TX* Bd. 462 San Antonio, TX., 1993

[271] WEISER, Mark: The computer for the 21st century. In: *Scientific american* 265 (1991), Nr. 3, S. 94–104

[272] WENDEMUTH, Andreas ; BIUNDO, Susanne: A companion technology for cognitive technical systems. In: *Cognitive behavioural systems* (2012), S. 89–103

[273] WITTEN, Ian H. ; FRANK, Eibe ; HALL, Mark A. ; PAL, Christopher J.: *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016

[274] WOLD, Svante ; ESBENSEN, Kim ; GELADI, Paul: Principal component analysis. In: *Chemometrics and intelligent laboratory systems* 2 (1987), Nr. 1-3, S. 37–52

[275] WU, Huadong ; SIEGEL, Mel ; STIEFELHAGEN, Rainer ; YANG, Jie: Sensor fusion using Dempster-Shafer theory [for context-aware HCI]. In: *Instrumentation and Measurement Technology Conference, 2002. IMTC/2002. Proceedings of the 19th IEEE* Bd. 1 IEEE, 2002, S. 7–12

[276] WUNDT, Wilhelm M.: *Grundzüge de physiologischen Psychologie*. Bd. 1. W. Engelman,

1874

[277] XU, Chang ; TAO, Dacheng ; XU, Chao: A survey on multi-view learning. In: *arXiv preprint arXiv:1304.5634* (2013)

[278] YAMBOR, Wendy S. ; DRAPER, Bruce A. ; BEVERIDGE, J R.: Analyzing PCA-based face recognition algorithms: Eigenvector selection and distance measures. In: *Empirical evaluation methods in computer vision*. World Scientific, 2002, S. 39–60

[279] ZENG, Zhihong ; PANTIC, Maja ; ROISMAN, Glenn I. ; HUANG, Thomas S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. In: *IEEE transactions on pattern analysis and machine intelligence* 31 (2009), Nr. 1, S. 39–58

[280] ZHANG, Lin ; RUKAVINA, Stefanie ; GRUSS, Sascha ; TRAUE, Harald C. ; HAZER, Dilana: Classification analysis for the emotion recognition from psychobiological data. In: *International Symposium on Companion-Technology (ISCT)*, 2015

[281] ZHANG, Zhengyou ; LYONS, Michael ; SCHUSTER, Michael ; AKAMATSU, Shigeru: Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In: *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on* IEEE, 1998, S. 454–459

[282] ZHAO, Guoying ; PIETIKAEINEN, Matti: Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2007), S. 915–928

[283] ZHAO, Guoying ; PIETIKAINEN, Matti: Dynamic texture recognition using local binary patterns with an application to facial expressions. In: *IEEE transactions on pattern analysis and machine intelligence* 29 (2007), Nr. 6, S. 915–928

[284] ZHAO, Qibin ; ZHANG, Liqing: ECG feature extraction and classification using wavelet transform and support vector machines. In: *Neural Networks and Brain, 2005. ICNN&B'05. International Conference on* Bd. 2 IEEE, 2005, S. 1089–1092

[285] ZHARKOV, Dimitrij ; KÄCHELE, Markus ; MEUDT, Sascha ; SCHWENKER, Friedhelm: A long-term stopping criterion for the forward-backward feature selection algorithm for the classification of emotions in speech. In: *Joint Proceedings of the 2013th T2CT and CCGL Workshops at IVA 2013*, 2013, S. 25–33