Verbesserung der Störsicherheit bei der Mimikanalyse in mono- und binokularen Farbbildsequenzen durch Auswertung geometrischer und dynamischer Merkmale

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

von Dipl.-Ing. Robert Niese

geb. am 04.02.1977 in Halberstadt

genehmigt durch die Fakultät für Elektrotechnik und Informationstechnik

der Otto-von-Guericke-Universität Magdeburg

Gutachter:

J.-Prof. Dr.-Ing. habil. Ayoub Al-Hamadi

Prof. Dr.-Ing. habil. Bernd Michaelis

Prof. Dr. rer. nat. Heiko Neumann

Promotionskolloquium am 24.11.2010

Vorwort

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Elektronik, Signalverarbeitung und Kommunikationstechnik der Otto-von-Guericke-Universität Magdeburg.

Besonderer Dank gilt den Herren J.-Prof. Dr. A. Al-Hamadi und Prof. Dr. B. Michaelis, die mir stets wertvolle Hinweise und Anregungen mit auf den Weg gaben. Über die Zeit meiner Promotion unterstützten sie mich stetig bei meiner Arbeit und fanden immer Zeit für notwendige Diskussionen. Ebenso möchte ich Herrn Prof. Dr. Heiko Neumann für die Kooperation innerhalb des Sonderforschungsbereichs Transregio 62 danken, in dessen Rahmen diese Arbeit fertiggestellt wurde.

Dank sagen möchte ich ebenfalls meinen Kollegen, die mir nützliche Hinweise und Anregungen bei der Durchführung dieser Arbeit gaben.

Ganz besonderer Dank gilt meiner Frau Nancy, die mich bei der Realisierung dieser Arbeit stets motiviert und unterstützt hat.

Schriftliche Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Insbesondere habe ich nicht die Hilfe einer kommerziellen Promotionsberatung in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, 31.05.2010

Robert Niese

Zusammenfassung

Die vorliegende Arbeit befasst sich mit der Entwicklung und Analyse eines neuen Ansatzes zur automatischen kamerabasierten Mimikanalyse auf der Grundlage von mono- und binokularen Farbbildsequenzen. Mit dieser Arbeit wird versucht eine Reihe von Methoden einzuführen, die eine hohe Erkennungsleistung bei expressiver Mimik (Basisemotionen) ermöglichen und gleichzeitig verschiedene bekannte Probleme adressieren. Hierzu wird eine Systemstruktur vorgeschlagen, bei der eine Auswertung geometrischer und dynamischer Merkmale genutzt wird. Geometrische Merkmale beschreiben räumliche Parameter wie beispielsweise Abstände und Winkel, welche auf der Grundlage photogrammetrischer Berechnungen ermittelt werden. Anders als bei der differentiellen Erfassung dynamischer Merkmale können geometrische Merkmale nicht nur bei einer Veränderung der Mimik, sondern jederzeit berechnet werden.
Dynamische Merkmale werden zur schnellen Erfassung mimikbedingter Änderungen des Bildinhaltes benutzt und unter Berücksichtigung sogenannter physiologisch motivierter Regionen mit Hilfe des Optischen Flusses berechnet.
Dieser Arbeit liegt die Hypothese zugrunde, dass durch eine integrierte Auswertung geometrischer und dynamischer Merkmale eine verbesserte Erkennungsleistung und somit verbesserte Störsicherheit erzielt werden kann.

Abstract

This thesis presents a new camera based approach for automatic facial expression analysis, based on mono- and binocular color image sequences. In this work a series of new methods is introduced, that on the one hand achieve high recognition rates for expressive facial behavior and on the other hand address a couple of common problems in this area of research. For this purpose, a system architecture is proposed for the evaluation of geometric and dynamic features. Physiologically motivated image regions are employed for detection of dynamic features by using an optical flow method. Opposed, geometric features describe geometric parameters which correspond to 3D based Euclidean distances and angles. Particularly, the hypothesis of this thesis is that through integrated evaluation of geometric and dynamic features improved recognition rates can be achieved.

Inhaltsverzeichnis

Abkürzungen

Abkürzung	Bedeutung	Verwendung
AF	Appearance Feature	Texturbasierte Merkmalsart
ANN	Artificial Neural Network	Klassifikator
ASM	Active Shape Model	Formmodell
AU	Action Unit	Einheit für Muskelbewegung
BMU	Best Matching Unit	Neuron einer SOM
BSP	Binary Space Partitioning	Datenstruktur
BU_i	Binghamton University Database	Bildmaterial zur Mimikanalyse
BV	Bildverarbeitung	Auswertung von Bildmaterial
EM	Expectation-Maximization	Cluster-Algorithmus
FACS	Facial Action Coding System	System zur Mimikbeschreibung
FAP	Facial Animation Parameters	MPEG-4 Animationsparameter
FDP	Facial Definition Parameters	MPEG-4 Modellparameter
FR	Flussregion	Modellregion zur Bewegungsanalyse
HMM	Hidden Markov Model	Stochastisches Modell
ICA	Independent Component Analysis	Statistisches Verfahren, Klassifikator
ICP	Iterative Closest Point	Algorithmus zur Registrierung
KKF	Kreuzkorrelationsfunktion	Korrespondenzbestimmung
k-NN	k-Nearest Neighbor	Klassifikator
LDA	Lineare Diskriminanzanalyse	Statistisches Verfahren, Klassifikator
MAD	Mittlere absolute Differenz	Korrespondenzbestimmung
MD_i	Magdeburger Datenbank	Bildmaterial zur Mimikanalyse
MLP	Multilayer Perceptron	Klassifikator
MMI	Mensch-Maschine-Interaktion	Anwendung
OF	Optischer Fluss	Bewegungsanalyseverfahren
PCA	Hauptkomponentenanalyse	Statistisches Verfahren
PDM	Point Distribution Model	Modell
PIE	Pose, Illumination, Expression	Parameter bei der Gesichtsanalyse

ROI	Region of Interest	Bildbereich
SOM	Self Organizing Map	Klassifikator
SVM	Support Vector Machine	Klassifikator
UL_i	Ulmer Datenbank	Bildmaterial zur Mimikanalyse
VV	Verschiebungsvektor	Bewegungsmessung
VVF	Verschiebungsvektorfeld	Bewegungsmessung

Formelzeichen

Symbol	Bedeutung
$\mathbf{A} \in \mathbb{R}^3$	Menge aller Modell-Ankerpunkte
$\mathbf{a}_j \in \mathbb{R}^3$	Einzelner Ankerpunkt
$\mathbf{a}_i, \mathbf{b}_i \in \mathbb{R}^3$	3D Modellvertex und zugehöriger Normalenvektor
$\partial \mathbf{a}_i / \partial \mathbf{t}$	Ableitung der Modellkoordinaten nach dem Zustandsvektor $\mathbf{t}$
$\mathbf{C}_i \in \mathbb{N}$	Klasse, Zuweisung durch Klassifikator
$d_i, \alpha_j \in \mathbb{R}$	Geometrische Rohmerkmale, Abstände und Winkel
$\mathbf{D}_f, \mathbf{D}_r \in \mathbb{R}^{m \times n}$	Tiefenkarten des Modells in der frontalen bzw. aktuellen Pose
$\mathbf{E}_{ri}, \mathbf{E}_{tj} \in \mathbb{R}^{4 \times 4}$	Elementare Rotations-und Translationsmatrizen
$\mathbf{E} \in \mathbb{R}^3$	Hilfsebene zur Nullposetransformation
$e(\mathbf{t}) \in \mathbb{R}$	Fehlerfunktion bezüglich eines Modell-Zustandsvektors $\mathbf{t}$
$\mathbf{f}^t_{dyn} \in \mathbb{R}^m$	Merkmalsvektor, dynamisch
$\mu^{\mathrm{C}_i}_{dyn} \in \mathbb{R}^m$	Mittelwert der dynamischen Merkmale bezüglich der Klasse C_i
$\mathbf{f}^t \in \mathbb{R}^n$	Merkmalsvektor geometrischer Rohmerkmale zum Zeitpunkt t
$\mathbf{f}_{neutral} \in \mathbb{R}^n$	Merkmalsvektor bei neutraler Mimik
$\mathbf{f}^t_{ratio} \in \mathbb{R}^n$	Merkmalsvektor, Verhältniswert
$\mathbf{f}^t_{geo} \in \mathbb{R}^n$	Merkmalsvektor, geometrisch

$\mu_{geo}^{\mathrm{C}_i} \in \mathbb{R}^n$	Mittelwert geometrischer Merkmale bezüglich der Klasse C_i
$f_{\mathrm{CLASS}}(\mathbf{f}^t)$	Beliebige Klassifikation eines geo. bzw. dyn. Merkmalsvektors
$\mathbf{I}_t \in \mathbb{N}^{m \times n}$	Bild zum Zeitpunkt t
$\mathbf{I}_{fp} \in \mathbb{R}^2$	Menge der Merkmalspunkte in Bildkoordinaten
$\mathbf{i}_j \in \mathbb{R}^2$	2D Subpixel-Bildkoordinate
$\mathbf{K}$	Modell einer realen Kamera
$\mathbf{K}_{GL}$	Modell einer synthetischen Kamera in OpenGL
$k(.),\ \mathbb{R}^3 \rightarrow \mathbb{R}^2$	Transformation eines 3D Punktes in Subpixelkoordinate
$k^{-1}(.),\ \mathbb{R}^2 \rightarrow \mathbb{R}^3$	Transformation Bildpunkt zu 3D Weltpunkt
$\mathbf{p}_i \in \mathbb{R}^3$	3D Punkt in kartesischen Koordinaten
$\mathbf{P}_{fp} \in \mathbb{R}^3$	Menge der Merkmalspunkte in Weltkoordinaten
$\mathbf{p}_{k,i} \in \mathbb{R}^3$	Bewegungsabtastpunkt, Modellpunkt zur Bewegungsanalyse
$r \in \mathbb{R}$	Korrelationskoeffizient
$\mathbf{S}$	Gesichtsmodell als Dreiecksnetz
$t \in \mathbb{N}$	Diskreter Zeitpunkt, Bildnummer
$\mathbf{t} \in \mathbb{R}^6$	Zustandsvektor eines Modells
$\partial/\partial\mathbf{t}$	Partielle Ableitung des Zustandsvektors $\mathbf{t}$
$\mathbf{T} \in \mathbb{R}^{4\times4}$	Transformationsmatrix in homogenen Koordinaten
$\mathbf{v}_{j,i}^t \in \mathbb{R}^2$	Verschiebungsvektor j zum Zeitpunkt t in Flussregion i
$\tilde{\mathbf{v}}_{j,i}^t \in \mathbb{R}^2$	Akkumulierter Verschiebungsvektor
$\bar{\mathbf{v}}_i^t \in \mathbb{R}^2$	Verschiebungsvektor, arithmetischer Mittelwert
$v_{sum}(t) \in \mathbb{R}$	Aktivierungsfunktion
$\mathbf{W} \in \mathbb{R}^3$	Messpunktwolke aus Stereoberechnung
$z(.),\ \ \mathbb{R}^2 \rightarrow \mathbb{R}^2$	Nullposetransformation eines Verschiebungsvektors

Kapitel 1

Einleitung

Mit Hilfe bildverarbeitender Systeme wird heute in vielfältigen Anwendungsbereichen in Industrie und Forschung versucht, die menschliche Auge-Gehirn-Interaktion durch technische Lösungen nachzubilden. So können etwa Menschen von monotonen Überwachungsaufgaben entlastet, bei komplexen Handlungen unterstützt oder bestimmte Anforderungen an Verarbeitungszeit und Genauigkeit überhaupt erst realisiert werden. Mit der Entwicklung immer leistungsfähigerer und zugleich kostengünstiger Rechentechnik setzen somit immer mehr Unternehmen Bildverarbeitungssysteme zur Qualitätskontrolle ein.
Auch im Bereich der Mensch-Maschine Interaktion (MMI) finden bildverarbeitende Systeme vermehrt Eingang, welche die Interaktion um neue Dimensionen erweitern sollen. Neben der Entwicklung neuer Eingabemodalitäten liegt hierbei ein besonderer Fokus auf der Erweiterung der situativen Erkennungsfähigkeit technischer Systeme. Es ist somit nicht mehr nur das „Wer“ und „Was“ von Interesse, sondern ebenso das „Wie“. Zu diesem Zweck wird heute sehr aktiv an neuen Techniken gearbeitet, mit denen automatisch aus Bildern und Sprache, d.h. aus Mimik, Gestik und Sprachbetonung (Prosodie) auf die aktuelle Verfassung des Nutzers geschlossen werden kann. Für jeden gesunden Menschen ist die Aufgabe einfach, den Gefühlszustand zu erkennen, der dem Anderen "ins Gesicht geschrieben steht", ohne sich eingehender mit Körpersprache beschäftigt zu haben. Das gilt besonders für die häufig von Psychologen angeführten sechs grundlegenden Gefühlsäußerungen Glück, Trauer, Wut, Ekel, Angst und Verwunderung. Das Gesicht ist somit für den Menschen das am leichtesten zu interpretierende und ausdrucksstärkste Kommunikationsmittel, wenngleich die Mimik nicht immer ein verlässliches Mittel zur Interpretation sein muss.
In den vergangenen 15 Jahren wurde eine Vielzahl technischer Lösungsansätze für diese Problematik entworfen. Auch wenn hierbei beachtliche Fortschritte erzielt wurden, besteht für die Mimikerkennung im Sinne einer robusten Merkmalserfas-

sung, Verarbeitung und Klassifikation weiterhin Forschungsbedarf. Dies gilt insbesondere für die Realisierung konkreter Schnittstellen für die Mensch-Maschine-Interaktion mit entsprechenden Anforderungen.

1.1 Zielstellung und Fragestellungen

Im Fokus dieser Arbeit steht die automatisierte bildbasierte Mimikanalyse, welche perspektivisch für eine Anwendung in der erweiterten Mensch-Maschine-Interaktion einsetzbar ist. Insbesondere soll hierzu die Erkennung prototypischer Basisemotionen erzielt und untersucht werden. Dabei erlaubt die Analyse mittels Bildverarbeitungsmethoden Vereinfachungen zum reinen Zweck der MMI. Somit erfolgt eine Abgrenzung zur biologisch motivierten Emotionsforschung bei der tiefgreifendere Parameter erfasst und analysiert werden, etwa fMRT/EEG[1], etc.
Grundsätzlich sind technische Systeme zur bildbasierten Analyse von Gesichtern, etwa für Biometrie oder MMI Anwendungen mit einer Reihe von Problemen konfrontiert, die aus einer variablen Kopfpose, einer unbekannten Beleuchtungssituation und dem aktuellen Gesichtsausdruck resultieren, was in der Literatur als sogenanntes PIE (Pose, Illumination, Expression) Problem bezeichnet wird. Hinzu kommt häufig ein für eine automatisierte Auswertung schwieriger Hintergrund, z.B. durch Überfüllung mit Personen oder Gegenständen, die dem zu observierenden Gesicht ähnliche Charakteristika bei der Merkmalsgewinnung aufweisen. Eine weitere Herausforderung ergibt sich aus partiellen Verdeckungen des Gesichts, was die Detektion relevanter Parameter und damit die Erkennungsleistung beeinträchtigen kann. Ebenso kommt es in Anwendungen mit nicht kontrollierbaren Bedingungen außerhalb der Labore schnell zu grundsätzlichen Störungen, z.B. durch falsche Kameraeinstellungen, Überblendung oder Signalstörungen.
Eine Lösung aller genannten Probleme ist in naher Zukunft sicher nicht zu erwarten. Mit dieser Arbeit wird aber dennoch versucht, eine Reihe grundsätzlicher Methoden einzuführen, die eine hohe Erkennungsleistung bei expressiver Mimik ermöglichen und dabei verschiedene Probleme adressieren. Hierzu wird im Sinne einer qualitativ hochwertigen Erkennung eine Systemstruktur vorgeschlagen, bei der eine Auswertung geometrischer und dynamischer Merkmale benutzt wird.
Während geometrische Merkmale aus einer einzelnen Aufnahme ermittelt werden können und räumliche Parameter beschreiben, werden dynamische Merkmale

[1] Die funktionelle Magnetresonanztomographie (fMRT) bzw. Elektroenzephalographie (EEG) sind bildgebende Verfahren zur Messung der Hirnaktivität [Hei05].

durch ein differentielles Verfahren berechnet und zur schnellen Erfassung von mimikbedingten Änderungen des Bildinhaltes genutzt.
Geometrische Merkmale repräsentieren Abstandsmaße und Orientierungen in Weltkoordinaten. Bei deren Auswertung wird immer Bezug auf den Normalzustand, d.h. zur neutralen Mimik genommen, die im vorgeschlagenen Verfahren als bekannt angenommen wird. Zur Bestimmung dieses Merkmalstyps werden photogrammetrische Techniken wie z.B. Kamerakalibrierung sowie ein starres Gesichtsmodell eingesetzt, das die Oberfläche des Gesichts der Versuchsperson im Neutralzustand approximiert.
Dynamische Merkmale werden hingegen unter Berücksichtigung sogenannter physiologisch motivierter Regionen durch ein Verfahren nach dem Optischen Fluss bestimmt und ferner unter Nutzung photogrammetrischer Techniken weiterverarbeitet. Diese Art von Merkmal repräsentiert flächenhafte Verschiebungen auf der Gesichtsoberfläche und ermöglicht so eine schnelle Detektion von Bildänderungen aufgrund von Variationen der Mimik.
Insbesondere liegt dieser Arbeit die Hypothese zugrunde, dass durch eine integrierte Auswertung geometrischer und dynamischer Merkmale eine verbesserte Erkennungsleistung erzielt werden kann.
Aus der vorgeschlagenen Systemstruktur leiten sich eine Reihe grundlegender Fragestellungen ab, die durch die Arbeit untersucht und geklärt werden sollen:

- Ermöglichen geometrische und dynamische Merkmale eine Erkennung und Unterscheidung der sechs gängigen Klassen emotional expressiver Mimik? Wo liegen hierbei die Stärken und Schwächen bzw. Grenzen?
- Worin liegen Vor- und Nachteile der geometrischen und dynamischen Merkmale?
- Führt die Integration geometrischer und dynamischer Merkmale zur Verbesserung bei der Mimikerkennung im Sinne der Klassifikationsergebnisse?
- In welchen Situationen ist eine solche Integration sinnvoll und wann nicht?
- Führt die Nutzung starrer 3D Modelle zu hinreichend korrekten Merkmalen für die Mimikerkennung?

Mit Hilfe der vorliegenden Arbeit sollen Antworten auf diese Fragen gegeben und empirisch unterlegt werden (s. Abschnitt 6.7).

1.2 Aufbau der Arbeit

Die vorliegende Arbeit ist in sieben Teile gegliedert.

In *Kapitel 2* wird der Stand der Technik auf dem Arbeitsgebiet der bildbasierten Mimikanalyse dargelegt. Hierbei wird zunächst auf Aspekte der Mimikerkennung eingegangen, bevor die beiden grundlegenden Ansätze in der Literatur vorgestellt werden. Im Weiteren werden anhand einer typischen Verarbeitungskette einer Mimikanalyse aktuelle Methoden zur Gesichtsdetektion sowie zur Extraktion und Repräsentation von Merkmalen dargestellt. Dabei wird auf geometrische, texturbasierte und dynamische Merkmale eingegangen. Abschließend werden aktuelle Klassifikationsansätze zur Erkennung von Mimik vorgestellt.

In *Kapitel 3* werden die Grundlagen, welche zum besseren Verständnis und zur Beschreibung der vorgeschlagenen Systemstruktur erforderlich sind kurz dargelegt. Dabei wird auf photogrammetrische Techniken, insbesondere das verwendete Kameramodell zur Durchführung von Welt zu Bild Transformationen und umgekehrt sowie das Prinzip der Stereophotogrammetrie zu Messzwecken kurz beschrieben. Weiterhin wird auf intensitätsbasierte Verfahren zur Bewegungsanalyse und Korrespondenzbestimmung eingegangen, was zur Erfassung der dynamischen Merkmale verwendet wird. Desweiteren werden verschiedene Klassifikatorarchitekturen erläutert, die zur Durchführung der Experimente verwendet wurden.

In *Kapitel 4* werden dynamische und geometrische Merkmale motiviert und beschrieben. Weiterhin wird die Erstellung des geometrischen Gesichtsmodells auf der Grundlage stereophotogrammetrischer Messungen erläutert und es wird die Herangehensweise zur Merkmalspunktextraktion dargestellt.

In *Kapitel 5* wird die vorgeschlagene Systemstruktur zur Mimikanalyse vorgestellt. Insbesondere werden hierzu zwei Ansätze beschrieben, die im Rahmen dieser Arbeit untersucht wurden. Zum einen wird die Normierung des Gesichts vorgestellt, zum anderen die Normierung der extrahierten Merkmale. Dabei wird die Erfassung dynamischer sowie geometrischer Merkmale betrachtet. Ein Weg zur Vereinigung der beiden Merkmalsarten wird vorgeschlagen, durch den eine Verbesserung des Klassifikationsergebnisses erzielt werden kann.

In *Kapitel 6* werden Ergebnisse aus der Evaluation mit Hilfe dreier Datenbanken, welche emotionstypische Mimik beinhalten, dargestellt. Das darin enthaltene Datenmaterial wurde zum Teil selbst aufgezeichnet und zum Teil durch externe Gruppen bereitgestellt. Bei der Evaluation wurden die Merkmalsräume der geometrischen und dynamischen Merkmale auf Klassentrennung untersucht. Außerdem werden Klassifikationsergebnisse für die betrachteten Merkmalsarten sowie deren Fusion vorgestellt. Weiterhin werden die Auswirkungen der Kopfpose auf die Verarbeitung dargestellt und eine Gegenüberstellung mit vergleichbaren Verfahren aus der Literatur gegeben. Die Ergebnisse werden mit Blick auf die eingangs gestellten Fragen diskutiert.

Im Anschluss folgt in *Kapitel 7* eine Zusammenfassung der Arbeit mit Ausblick.

Im Anhang werden weitere Klassifikationsergebnisse gegeben, dynamische und geometrische Merkmale noch einmal kompakt dargestellt sowie Algorithmen und Transformationsmatrizen beschrieben.

Kapitel 2

Stand der Technik

Eine automatische bildbasierte Analyse von Gesichtern ermöglicht die Realisierung neuer Schnittstellen zwischen Mensch und technischem System und stellt ein aktuelles Forschungsgebiet in der Computer Vision dar. Die Einsatzmöglichkeiten sind äußerst vielfältig, das Spektrum reicht dabei von biometrischen Anwendungen zur Personenerkennung [Del07, Par08], über Mimikerkennung mit Anwendung in affektsensitiven Mensch-Maschine-Schnittstellen [Pan09, Zen09], zur Erhöhung der Sicherheit im Automotivebereich, etwa durch Kontrolle des Lidschlags, was der Erkennung von Müdigkeit oder Sekundenschlaf des Fahrers dient [Zha06], bis hin zu zukünftigen medizinischen Applikationen, in denen der Zustand von Patienten überwacht wird. Dabei könnten perspektivisch z.B. Schmerzzustände oder Vigilanz (Wachheit) bei nicht interaktionsfähigen Patienten automatisch bestimmt werden [Bra06, Nie09]. Weitere aktuelle Anwendungen liegen im Bereich Komfort und Unterhaltungsindustrie, z.B. zur automatischen Fokussierung von Gesichtern in modernen Digitalkameras.
Aktuelle Modelle detektieren dabei nicht nur Gesichter, sondern erkennen auch das Lächeln der Fotografierten und übernehmen die Bildaufnahme, ohne dass hierzu der Auslöser betätigt werden muss (z.B. “Smile Shutter”, Sony Cybershot® T70 Kamera Serie). Weitere Einsatzgebiete liegen in der Qualitätssicherung, z.B. zur Erfassung des Fahrkomforts [Oer08], zur Bestimmung des Immersionsgrades in Virtual Reality Anwendungen und Computerspielen [Boe05] sowie zur Steuerung von Avataren [Dig10].
Im professionellen Film wird dies bereits seit geraumer Zeit verwendet. Dazu wird durch markerbasiertes sogenanntes “Motion Capturing“ die Mimik eines Schauspielers auf eine computergenerierte Figur übertragen [Wes03]. Eine interessante zukünftige Applikation stellt die automatische Erkennung von Täuschungsversuchen bzw. Lügendetektion dar, was bisher nur durch eine detaillierte menschliche Observierung möglich war. Hierzu werden spezifische Gesichtsregionen in einer

erzeugten Stresssituation mit Hilfe von Wärmebildkameras analysiert und durch Mustererkennungsmethoden ausgewertet [Tsi06].

Funktionell betrachtet stellt das menschliche Antlitz eine multi-modale Kommunikationsschnittstelle zur Ein- und Ausgabe dar. Es liefert Informationen zum Gehirn und reflektiert getroffene Bewertungen und Entscheidungen. Einer der Ausgabekanäle ist die Mimik. Durch sie wird kommuniziert und es lassen sich Schlussfolgerungen über die aktuelle psychische Verfassung, wie den Gefühlszustand ziehen [Kel00]. Grundsätzlich werden nach Pantic et al. vier Arten von Signalen durch das Gesicht vermittelt [Pan07].

1. Permanente Signale werden für gewöhnlich zu Identifikationszwecken verwendet, da sie dauerhaft sind und die Grundlage für das Erscheinungsbild des Gesichts darstellen. Schädel, Knochen, weiches Gewebe und die bestehenden Proportionen insgesamt fallen in diese Kategorie.

2. Langsame Signale repräsentieren allmählich im Laufe der Zeit entstehende Änderungen im Erscheinungsbild des Gesichts. Hierzu gehören dauerhaft bestehende Falten und Veränderungen der Hauttextur, was zur Altersbestimmung genutzt werden kann, jedoch negative Auswirkungen auf die Qualität der Merkmalsbestimmung und Erkennung haben kann.

3. Künstliche Signale entstehen aus temporären Veränderungen durch Zusätze wie Brillen oder Kosmetik und können zusätzliche Informationen, z.B. zur Erkennung des Geschlechts bereitstellen. Dabei können diese die Detektion von Gesichtsmerkmalen sowohl erschweren als auch vereinfachen.

4. Schnelle Signale sind spontane und temporäre Änderungen, welche aus neuromuskulärer Aktivität resultieren und zu visuell detektierbaren Veränderungen im Erscheinungsbild des Gesichts führen und damit die Grundlage für die Mimik bilden. Die durch schnelle Signale erzeugten Merkmale werden auch als transient bezeichnet.

In der vorliegenden Arbeit werden die schnellen Signale zur Analyse plötzlich auftretender mimikbedingter Veränderungen im Gesicht genutzt. Diese Signale werden dabei abhängig von der Art ihrer Erfassung als dynamisch oder geometrisch in Kapitel 4 eingeführt und auf der Grundlage einer neuartigen Systemstruktur ausgewertet, welche in Kapitel 5 vorgestellt wird.
In diesem Kapitel wird eine Übersicht zu aktuellen Ansätzen, Techniken und wichtigen Aspekten im Forschungsgebiet der Mimikanalyse gegeben.

2.1 Aspekte der Mimikerkennung

Wie in jedem Forschungsgebiet, gibt es auch bei der Mimikanalyse eine Reihe bisher ungelöster Probleme bzw. Einschränkungen. Während die Erkennung von Mimik für einen gesunden Menschen scheinbar ohne jede Anstrengung vollführt werden kann, stellt die Entwicklung eines automatischen Systems, welches diese Aufgabe in variablen Szenarien zuverlässig realisiert, eine schwierige Angelegenheit dar. Aus diesem Grund wurde in den letzten 15 Jahren ein beträchtlicher Beitrag der verschiedenen beteiligten Forschungsfelder geleistet. In den folgenden Unterpunkten werden einige wichtige Aspekte der automatischen bildbasierten Mimikanalyse diskutiert.

2.1.1 Facial Action Coding System

Veränderungen der Mimik werden durch Muskelkontraktionen verursacht. Es gibt dabei eine Anzahl von 43 Muskeln, die grundlegend für die Entstehung der Gesichtsausdrücke verantwortlich sind. Ekman&Friesen schlugen das Facial Action Coding System (FACS) vor, mit dem es möglich ist, jeden nur denkbaren Gesichtsausdruck genau zu beschreiben [Ekm02]. Da Mimik meist aus einer Kombination von verschiedenen Muskelaktivierungen resultiert, wird als Maßeinheit für das FACS nicht die Aktivität einzelner Muskeln verwendet, sondern sogenannte Action Units (AUs), von denen 64 definiert wurden. Insbesondere beschreiben diese Einheiten Kontraktionen und Relaxationen von Gesichtsmuskeln bzw. Muskelgruppen. Abbildung 2-1 stellt die Muskulatur der Stirnregion und eine Reihe assoziierter Action Units dar, die bei der durch Mimik verursachten Bewegung der Augenbrauen, Stirn und der Augenlieder zur detaillierten Beschreibung verwendet werden. Da das FACS ein Kodierungssystem für Gesichtsbewegungen darstellt, sind einige der Action Units nur schwer in einem statischen Bild darstellbar. Zur Veranschaulichung wird häufig lediglich das Ergebnis der mimischen Bewegung abgebildet. Eine Übersicht ist z.B. hier zu finden [CMU10].
Speziell ausgebildete FACS Coder sind in der Lage in zeitintensiver Arbeit Gesichtsausdrücke in einzelne Action Units manuell zu „zerlegen“.
Die automatische Zuweisung einer Mimikklasse durch einen Entscheider kann auch durch Nutzung vereinfachter Größen erfolgen und erfordert nicht notwendigerweise die Bestimmung der Action Units. Diese stellen jedoch eine allgemein anerkannte Grundlage zur Beschreibung der Mimik dar. Automatische Systeme

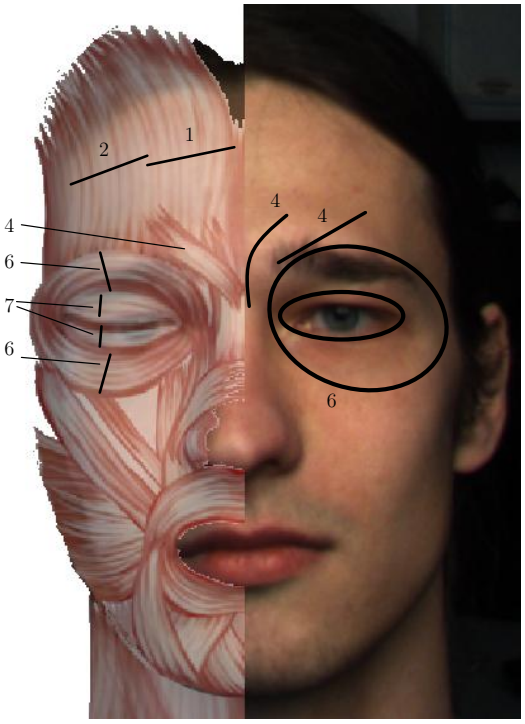


Abbildung 2-1: Anatomie und Untermenge an Action Units in der Stirnregion nach dem FAC System [Ekm02]. In der linken Hälfte wird die Gesichtsmuskulatur abgebildet, Quelle [Flo10]. Rechts wird die Auswirkung der Muskelaktivität skizziert.

zur Erfassung von Action Units aus Bildern sind Gegenstand aktueller Forschung [Rot09, Val06a, Coh01].

2.1.2 Gestellte vs. spontane Mimik

Gespielte Gesichtsausdrücke sind bewusst erzeugt, während spontane Mimik originär durch die beobachtete Person in einer bestimmten Situation hervorgebracht wird. Psychologische Studien legen nahe, dass diese beiden Arten von Mimik völlig verschiedener Natur sind. Neurowissenschaftlich wurde gezeigt, dass hier verschiedene Hirnaktivierungen stattfinden, welche in unterschiedlicher Muskelaktivierung und entsprechender Dynamik resultieren [Ekm05]. In einer Studie von Valstar et al. wurde gezeigt, wie spontane und absichtliche Augenbrauenaktivierung zu verschiedenen Resultaten bezüglich Apex und onset/offset Geschwindigkeit führt [Val06b].
In der Konsequenz bedeutet dies, dass für Anwendungen in der Mensch-Maschine-Schnittstelle grundsätzlich auch Untersuchungen mit authentischer Mimik durchgeführt werden sollten, um die Funktionsfähigkeit für den Einsatz sicherzustellen.

2.1.3 Mimik-Darstellung und Dynamik

Die Darstellung der aktuellen Mimik resultiert aus dem Zusammenspiel beteiligter Action Units. Dabei beschreibt die Dynamik den zeitlichen Verlauf der Aktivierungsphasen. Dieses Timing aus Aktivierung und Fortdauer kann zur Analyse der mimischen Aktivität wertvolle Information bereitstellen. Grundsätzlich lässt sich ein präsentierter Gesichtsausdruck in drei Phasen untergliedern, d.h. onset, apex und offset, oder anders ausgedrückt, Aktivierung, Halten und Entspannung. Die Dynamik der präsentierten Mimik wird in den kognitiven Wissenschaften als zentrales Merkmal zur Interpretation angesehen [Amb05]. So lässt sich z.B. ein Lächeln mit kleiner Amplitude und langsamer onset und offset Phase als höflich kategorisieren [Ekm03]. Dies zeigt, wie sehr der zeitliche Kontext für eine korrekte Interpretation erforderlich ist. Cohn et al. haben in einer Studie gezeigt, das spontanes Lächeln mehrfache Apex Phasen aufweisen kann und sich dabei von gespieltem Lächeln abhebt [Coh04].

2.1.4 Quantifizierung und Kontextinformation

Durch eine Quantifizierung der Mimik wird bestimmt, wie ausgeprägt der Gesichtsausdruck dargestellt wird, d.h. leicht, stark, evtl. graduell. Anders ausgedrückt wird durch die Quantifizierung die Stärke der Abweichung vom neutralen Gesicht im entspannten Zustand gemessen. Somit lässt sich z.B. sagen, wie weit die Augenbrauen angehoben wurden, oder wie weit der Mund seine Breite bzw. Höhe verändert hat. Insbesondere wenn kein Zusatzwissen wie etwa der Neutralzustand bekannt ist, stellt die Quantifizierung ein schwieriges Problem dar und ist selbst für den Menschen nicht immer sicher zu bestimmen.
Während für das FACS eine 5-Punkteskala zur Beschreibung der Variation der Action Unit Intensität vorgeschlagen wurde [Ekm02], sind automatische Systeme hierfür noch zu entwickeln. Ansätze zur Quantifizierung wurden bereits publiziert. Lien et al. beschreiben, wie die Intensität grundsätzlich aus der Änderung der Mimik abgeleitet werden kann [Lie98], während Zhang et al. eine implizierte Kodierung der Intensität vorschlagen [Zha05]. Bartlett et al. führten einen Vergleich zwischen manuell ermittelten Intensitätswerten und ihrem System zur automatischen Codierung der Intensitätsvariation durch. Wie zu erwarten zeigte sich dabei eine signifikante Korrelation zwischen dem Abstand der trennenden Hyperebene des Klassifikators und den manuell bestimmten Intensitätswerten [Bar06].

Der vorgeschlagene Ansatz zur Mimikanalyse stellt eine gute Grundlage dar, um in weitergehenden Untersuchungen eine Quantifizierung zu realisieren.
Mimikerkennung ohne Berücksichtigung situativer Kontextinformation kann zu Fehlinterpretationen führen, beispielsweise können zugekniffene Augen das Ergebnis blendenden Lichts sein oder auf eine ablehnende, verärgerte Haltung hindeuten. Daher ist es für eine situative Erkennung der Mimik entscheidend, Kontextinformation zu berücksichtigen. Das Problem des "context-sensing" ist jedoch für den allgemeinen Fall nur schwer zu lösen, falls überhaupt. Bis auf sehr wenige Arbeiten, die eine nutzerprofilorientierte Interpretation der Mimik verwenden, z.B. [Fas04] oder [Pan04] wo ein nutzerzentrierter Ansatz zum Anlernen des mimischen Verhaltens vorgeschlagen wird, ist die Mehrzahl der Erkennungssysteme nicht kontextsensitiv.
Eine Berücksichtigung und Auswertung des Kontexts im Sinne intelligenter Systeme ist ebenfalls nicht das Anliegen der vorliegenden Arbeit.

2.1.5 Mimik-Datenbanken

Zur Entwicklung von Algorithmen, zur Merkmalsextraktion, zum Trainieren von Klassifikatoren und zur Validierung neuer Systeme wird generell eine große Menge an Daten benötigt. Grundsätzlich existiert keine allumfassende Datenbank, die als Grundlage für die vielfältigen bisher veröffentlichten Ansätze zur maschinellen Mimikanalyse dienen könnte. Entsprechend des untersuchten Forschungsgegenstandes decken eine Vielzahl separater Datenbanken jeweils nur bestimmte Teilaspekte ab, z.B. Farbe oder Schwarzweis, Einzelbilder oder Bildfolgen, 2D oder 3D, gestellte oder spontane Mimik, Zustände definierter Pose, Beleuchtung, Mimikklassen, Alter, Geschlecht, Hautfarbe, verschiedene Ethnien, Annotation, FACS Kodierung, etc.
Das erste umfangreiche Archiv zur Untersuchung der Auswirkung der Parameter "Pose, Illumination, Expression" auf die Gesichtserkennung stellt die PIE Database dar, eine Datenbank der Carnegie Mellon University, Pittsburgh, welche im Jahr 2000 veröffentlicht wurde und mehr als 40,000 Bilder von 68 Personen umfasst, von denen jede in 13 verschiedenen Posen, 43 Beleuchtungssituationen und vier Gesichtsausdrücken abgelichtet wurde [Sim02]. Dabei wurde auch das Sprechen der Versuchspersonen als Mimik berücksichtigt. Dieses Datenarchiv wurde mittlerweile zur Multi-PIE Database erweitert, welche für mehr als 300 Personen eine dreiviertel Million Bilder mit einer Vielzahl von Parametern enthält [Gro08].

Die Cohn-Kanade Database wurde nach dem Facial Action Coding System annotiert, so dass für jede der dargestellten Mimiken die Aktivierung in Action Units zur Verfügung steht [Kan00]. Dabei wurden Sequenzen von 100 Probanden im Alter von 18 bis 30 Jahren mit verschiedenem Hautfarbtyp (African-American, Asian, Latino) mit einer Videokamera aufgezeichnet.
Yin et al. von der Binghamton University, New York, haben mit der BU-3DFE Database die erste umfangreiche 3D Datenbank zur Mimikanalyse mit Fokus auf sechs Basisemotionen vorgestellt. Diese Datenbank beinhaltet hochaufgelöste statische 3D Modelle inklusive Farbtextur von 100 Probanden verschiedenen Alters, Geschlechts und Hautfarbtyps [Yin06]. Mit der BU-4DFE Database wurde die Datenbank auf Bildfolgen mit insgesamt mehr als 60,000 einzelnen 3D Scans erweitert, so dass hier neben der rein statischen Analyse auch die Dynamik ausgewertet werden kann [Yin08].
Im Rahmen der vorliegenden Arbeit entstand ebenfalls ein Archiv an Daten, welches Bildfolgen von zwanzig männlichen Probanden mit jeweils 4 verschiedenen Basisemotionen sowie die Klasse neutral enthält.
Mittlerweile sind eine ganze Reihe weiterer Datenbanken veröffentlicht wurden, eine Zusammenstellung ist hier zu finden [Fac10].

2.2 Grundlegende Ansätze zur Mimikerkennung

In der Vergangenheit wurde eine Reihe verschiedener Ansätze zur Mimikerkennung vorgestellt, die sich prinzipiell in zwei Kategorien unterteilen lassen. Die erste Gruppe von Verfahren wertet die durch Mimik hervorgerufene Deformation aus, während die zweite auf der Analyse der zugrundeliegenden Dynamik beruht. Im Folgenden wird ein kurzer Abriss der beiden Ansätze und der verschiedenen involvierten Techniken gegeben. Da diese mittlerweile sehr vielzählig sind wird dabei kein Anspruch auf Vollständigkeit erhoben.

2.2.1 Deformationsbasierte Verfahren

Die mimikbedingte Deformation der Gesichtsmerkmale manifestiert sich in einer Änderung von Form plus Textur und führt zu einer Menge an Gradienten im Bildsignal, welche im Orts- oder Frequenzraum analysiert werden können. Im Frequenzraum wird diese Änderung durch Hochpass- oder Gabor Wavelet basierte Filter ermittelt [Jae05]. Der Gaboransatz ermöglicht dabei eine gewisse Toleranz bei einer Variation der Beleuchtungsverhältnisse und entsprechenden Signal-

schwankungen im Bild. Die verschiedenen bekannten Verfahren zur Messung der Deformation lassen sich prinzipiell wie folgt untergliedern.

- Holistische Verfahren

Hierbei wird das untersuchte Gesicht ganzheitlich, d.h. holistisch betrachtet und in einem Bildblock analysiert. Dabei ist es zur Fehlervermeidung entscheidend, dass lediglich das Gesicht ausgewertet wird, welches hierzu vom Hintergrund klar abgegrenzt sein muss. Für holistische Verfahren werden insbesondere Filtertechniken wie Gabor Wavelets eingesetzt. Bei diesem Ansatz wird die Hauptlast der Arbeit dem Klassifikator übertragen [LiS05].
Ma et al. führten einen systematischen Vergleich maschineller Lernverfahren zur holistischen Mimikanalyse durch [MaW05]. Dabei fanden sie, dass die linearen Klassifikatoren in der überwiegenden Zahl der Fälle die höchste Performanz erreichen. Fellenz et al. verglichen bei der Gaborfilterung drei verschiedene Herangehensweisen zur Erkennung von vier Klassen emotional expressiver Mimik in statischen Gesichtsaufnahmen [Fel99].

- Lokale bildbasierte Verfahren

Bei den lokalen bildbasierten Verfahren wird die Deformation regional begrenzt ausgewertet, indem an markanten Stellen Bildfenster definiert werden. Zur Festlegung dieser Positionen werden nicht-transiente Gesichtsmerkmale, z.B. die Augen herangezogen. Die Deformation lokaler transienter Merkmale wie Fältchen, erfolgt zumeist durch Auswertung von Bildgradienten. Shan et al. haben hierzu ein Verfahren vorgestellt, welches lokale Binärmuster, sogenannte “Local Binary Patterns“ zur Mimikerkennung auswertet [Sha09].

- Modellbasierte Verfahren

Bei den modellbasierten Verfahren werden Gesichter mit Hilfe geometrischer Modelle, wie z.B. Active Shape Models (ASM), die insbesondere Kanteninformation und texturbasierte Point Distribution Models (PDM) nutzen, ausgewertet [AlH06b]. Bei den sogenannten “Appearance“ Modellen, welche auf der Modellierung und dem Abgleich des Erscheinungsbildes des Gesichts basieren, werden ASMs entlang nicht-transienter Merkmale orientiert [LiS05, Mor10]. Häufig kommen dabei Gabor Wavelets zur Rauschunterdrückung zum Einsatz.

2.2.2 Bewegungsbasierte Verfahren

Neben dem deformationsbasierten Ansatz stellt die Erfassung der Dynamik, d.h. der durch Mimik verursachten Veränderung den zweiten wesentlichen Ansatz zur

Erkennung dar. Die Detektion basiert dabei auf verschiedenen Techniken der Bewegungsanalyse, beispielsweise auf dem Tracking von Merkmalspunkten, Auswertung des Optischen Flusses, Differenzbildtechnik, etc. Im Folgenden werden einige wesentliche Herangehensweisen kurz erläutert.

- Optischer Fluss (Dense Optical Flow)

Durch die Technik des optischen Flusses, welche zu den differentiellen Verfahren gehört (Abschnitt 3.3.1), kann Objekt- oder Kamerabewegung erfasst werden. Diese resultiert in einem Verschiebungsvektorfeld (VVF), durch das die Richtung und Stärke der gemessenen Bewegung dargestellt wird. Der optische Fluss kann sowohl holistisch als auch lokal zur Auswertung der Verschiebung von Gesichtsmerkmalen eingesetzt werden [Pan09]. Lien et al. nutzten z.B. den optischen Fluss in Zusammenhang mit einer Wavelet-basierten holistischen Methode [Lie98]. Otsuka et al. führten eine Mimikanalyse auf der Grundlage eines lokalen Ansatzes zur Auswertung des optischen Flusses mit Hidden Markov Modellen durch [Ots98]. Die Berechnung eines dichten VVFs über das gesamte Gesicht ist jedoch rechenaufwendig und schwierig auszuwerten, weshalb häufig eine lokale Berechnung in interessanten Regionen, etwa auf der Grundlage eines Muskelmodells durchgeführt wird [Nie07b]. Die Kontraktion und Entspannung der Muskulatur gibt dabei die Bewegungsrichtung in den verschiedenen Gesichtsregionen vor.

- Merkmalspunkt Tracking

Ein anderer Weg zur Erfassung der Bewegung im Gesicht besteht im Verfolgen ausgewählter Punkte entlang markanter Gesichtsmerkmale, z.B. Augenbrauen, Mund, etc. [Nie09, Lie98]. Bei diesem Ansatz werden die Merkmalspunkte im ersten Bild der Sequenz initialisiert und anschließend verfolgt mit entsprechenden Re-Initialisierungsschritten. Dieses Tracking ist jedoch anfällig für Variationen der Aufnahmebedingungen, wie z.B. Bildrauschen, Beleuchtungsänderungen, Verdeckungen, etc. Aus diesem Grund werden für das Tracking meist Beobachtungsfenster mit einer geeigneten Nachbarschaft definiert und Techniken wie z.B. Blockmatching eingesetzt.

- Differenzbildtechnik

Die mimische Bewegung lässt sich weiterhin durch Differenzbildtechniken erfassen. Bartlett et al. haben eine holistische differenzbildbasierte Bewegungsdetektion eingesetzt [Bar98]. Dabei werden die einzelnen Bilder von einem Referenzbild subtrahiert, welches das Gesicht mit neutraler Mimik beinhaltet. Um rotierte Gesichter mit dem Referenzbild verarbeiten zu können, ist ein Gesichtsnormierungs-

schritt erforderlich, ebenso ist eine Adaption der Beleuchtungssituation notwendig. Weiterhin ist dieser Ansatz empfindlich gegenüber Bildrauschen.

2.3 Verarbeitungskette der kamerabasierten Mimikanalyse

Die grundsätzliche Gemeinsamkeit aller Verfahren zur automatischen bildbasierten Mimikanalyse ist, dass nach einer Merkmalsextraktion aus Bildern oder Bildsequenzen ein Erkennungsmodul folgt, in dem die Merkmale interpretiert werden, entweder im Sinne einer zugewiesenen Mimikkategorie oder einer Beschreibung entsprechend des Facial Action Coding Systems. Unterschiede bestehen im verwendeten Ansatz zur Merkmalsextraktion und der nachfolgenden Klassifikation. Dementsprechend stellt Abbildung 2-2 das prinzipielle Schema zur kamerabasierten Mimikanalyse dar. Die wesentlichen Komponenten werden in den nachfolgenden Abschnitten anhand aktueller Ansätze aus der Literatur erläutert.

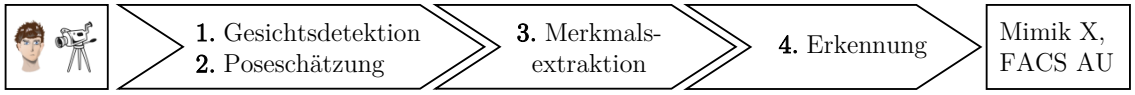


Abbildung 2-2: Grundlegende Schritte der kamerabasierten Mimikanalyse.

Entsprechend der dargestellten Verarbeitungskette erfolgt initial die Aufnahme der Versuchsperon in Einzelbildern oder Bildfolgen mit Hilfe von zumeist monokularen Kamerasystemen, welche im sichtbaren Spektrum des Lichts arbeiten und dabei Farbe oder Grauwerte aufzeichnen. Es gibt ebenfalls Arbeiten, in denen Kamerasysteme verwendet werden, die das Spektrum des nahen Infrarot (NIR) Bereichs nutzen und somit eine bezüglich Beleuchtungsänderungen robuste Bildauswertung erzielen [Tai08, Kha06]. Aufgrund des zurzeit noch hohen Anschaffungspreises ist eine Mimikerkennung auf der Grundlage von Infrarotkameras weniger verbreitet. Ähnliches gilt für die Benutzung von Mehrkamerasystemen. Obwohl deren Verwendung grundsätzlich mehr Zuverlässigkeit bei der Gesichtsdetektion und Merkmalsextraktion ermöglicht, werden diese für die Mensch-Maschine-Interaktion mittels Mimik, aufgrund des Mehraufwandes, z.B. notwendige Kalibrierung, höheres Datenaufkommen und hoher Anschaffungspreis, bisher nur in Spezialanwendungen eingesetzt.

2.3.1 Gesichtsdetektion

Unabhängig von der konkreten Aufgabe, stellt die Gesichtsdetektion den ersten Schritt eines jeden vollautomatischen Systems zur Analyse, der in Gesichtern enthaltenen Information, z.B. zur Feststellung der Identität, Geschlecht, Alter, Mimik, Pose, etc., dar. Dabei ist es die prinzipielle Aufgabe der Gesichtsdetektion herauszufinden, ob in einem gegebenen Bild Gesichter vorkommen und falls ja, deren Position festzustellen. Diese Aufgabe ist nicht trivial, da alle möglichen auftretenden Variationen im Erscheinungsbild von Gesichtern berücksichtigt werden müssen. Insbesondere betrifft dies die als PIE (Pose, Illumination, Expression) Problem bekannten Variationen [LiS05], d.h. mögliche Änderungen in der Orientierung, was zu veränderter Größe des Gesichts, Rotationen innerhalb und außerhalb der Ebene mit entsprechender perspektivischer Verkürzung bzw. Selbstverdeckung führen kann. Variierende Beleuchtung führt weiterhin zu beträchtlichen Schwankungen im Erscheinungsbild des Gesichts und zugehöriger Merkmale, ebenso wie expressive Mimik. Darüber hinaus haben Gesichter verschiedener Ethnien eine unglaubliche Vielfalt in Farbe und Form. Zusätzlich kann eine Detektion, durch Teilverdeckungen oder Besonderheiten wie Brille, Bart, etc. erschwert werden.
In der Vergangenheit wurde eine Vielzahl von Verfahren vorgestellt, mit denen versucht wird, diese Probleme oder einzelne Teilprobleme, durch spezielle Algorithmen auf der Grundlage von Grauwertbildern [Yan09, Hua07, Vio04], durch Verwendung von Farbinformation [HsuAJ02] bzw. durch Kombination von beiden [Kim08] oder durch Zuhilfenahme von 3D-Information zu lösen [Mia06]. Während die frühen Ansätze hauptsächlich auf frontale Aufnahmen beschränkt waren, sind mittlerweile Verfahren verfügbar, die die Aufgabe der Gesichtsdetektion auch bei moderater Kopfrotation um alle drei Achsen recht zuverlässig und in Echtzeit, sowohl mit Standard PC-Hardware als auch portablen Geräten, wie handelsüblichen Digital- und Videokameras, lösen.

In den meisten Verfahren zur grauwertbasierten Gesichtsdetektion werden charakteristische Eigenschaften, wie lokale Merkmale oder holistische Intensitätsmuster aus einem Satz von Trainingsbildern in einer Anlernphase extrahiert, in denen Gesichter in einer definierten Pose vorgegebenen sind [LiS05]. Um eine verbesserte Resistenz gegen Beleuchtungsschwankungen zu erreichen, werden häufig Bildverarbeitungstechniken wie Histogrammspreizung und Varianznormierung verwendet [Vio04]. Auf der Grundlage der extrahierten Merkmale wird im Suchschritt das komplette Bild an jeder Position und in verschiedenen Größenskalierungen auf

das Vorkommen von Gesichtern untersucht. Das Skalierungsproblem wird dabei für gewöhnlich durch eine wiederholte Detektion in einer Pyramide von Bildern gelöst, in denen das Originalbild um einen gegebenen Faktor verkleinert wird. Wie in [Kim08] gezeigt, kann durch Hinzunahme von Farbinformation und der damit verbundenen Verkleinerung des Suchraumes das Detektionsergebnis deutlich verbessert und die Rate an Falsch-Positiven gesenkt werden.
In der Literatur werden zahlreiche Ansätze zur Gesichtsdetektion vorgeschlagen, die zum gegenwärtigen Stand performanteste Detektion beruht dabei auf der Auswertung sogenannter "Haar-like features", welche durch Viola und Jones vorgestellt wurden [Vio04, Hua07]. Weitere Ansätze basieren auf Wavelets [Sch04, Pap98] und sogenannten "Parts-based" Methoden [Hei07], die auf der separaten Detektion einzelner Gesichtsmerkmale beruhen. Häufig werden die verschiedenen Techniken kombiniert. Mittlerweile ältere Methoden sind pixelbasiert [Sun98, Yan02] oder nutzen lokale Kantenmerkmale [Fle01].
Die hohe Performanz des von Viola und Jones vorgestellten "Haar-like-feature" basierten Detektors wird durch Verwendung einer effizienten Datenverarbeitung erreicht. Hierzu werden sogenannte Integralbilder verwendet, welche an jeder Koordinate (x, y) die Summe $sat(x, y)$ (2.1) der Intensitätswerte aller links oberhalb befindlichen Pixel enthalten. Integralbilder wurden zuvor als "summed area table" (sat) zum Texturmapping in der Computergrafik eingeführt [Cro84].

$$sat\,(x, y) = \sum_{x' \leq x, y' \leq y} \mathbf{I}\,(x', y') \tag{2.1}$$

mit $sat()$ als Funktion zur Berechnung des Integralbildes einer Bildquelle $\mathbf{I}$.

Mit Hilfe eines einmal berechneten Integralbildes, lässt sich nach (2.2), durch Verwendung von nur vier Randpunkten eines beliebigen Bildbereichs, dessen integriertes Grauwertsignal in konstanter Zeit bestimmen. Bezugnehmend auf die Bezeichnungen in Abbildung 2-3 berechnet sich diese Summe wie folgt:

$$\sum_{\mathbf{a}(x) \leq x' \leq \mathbf{c}(x),\ \mathbf{a}(y) \leq y' \leq \mathbf{c}(y)} \mathbf{I}\,(x', y') = sat\,(\mathbf{a}) + sat\,(\mathbf{c}) - sat\,(\mathbf{b}) - sat\,(\mathbf{d}) \tag{2.2}$$

Die Integralbilder stellen die Grundlage zur schnellen Berechnung der "Haar-like features" dar, welche zur Repräsentation von Gesichtern oder anderer Grauwert- bzw. Intensitätsmuster in Pixelrastern benutzt werden können. In Abbildung 2-4 werden beispielhaft die durch Viola und Jones vorgestellten Merkmalsarten darge-

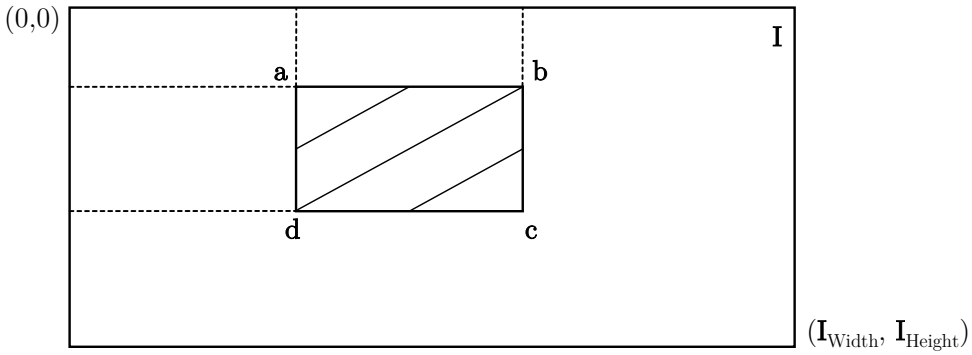


Abbildung 2-3: Schnelle Berechnung der Summe aller Grauwerte innerhalb einer rechteckigen Region (**abcd**) in Bild **I**.

stellt, die zur Erfassung horizontaler, vertikaler und diagonaler Pixelstrukturen benutzt werden. In nachfolgenden Arbeiten wurde der Satz rein horizontaler und vertikaler Merkmale zur Erhöhung der Erkennungsgenauigkeit um rotierte Blöcke erweitert [Xia09].
Haar-like features stellen somit Merkmale dar, die lokale Kanteninformation in unterschiedlicher Orientierung und Skalierung repräsentieren. Mit einem gegebenen Beispielbild der Größe 24x24 Pixel beträgt der vollständige Satz parametrierter Merkmale über 160.000 [Vio04]. Zur Gesichtsdetektion wird jedoch nur eine kleine, vom AdaBoost Algorithmus selektierte Menge von Merkmalen zum Anlernen aus positiven und negativen Beispieldaten verwendet.
Im Gegensatz zu den meisten älteren Verfahren, die nur einen einzigen starken Klassifikator verwenden, der zur Erkennung alle Merkmale einbezieht, wie z.B.

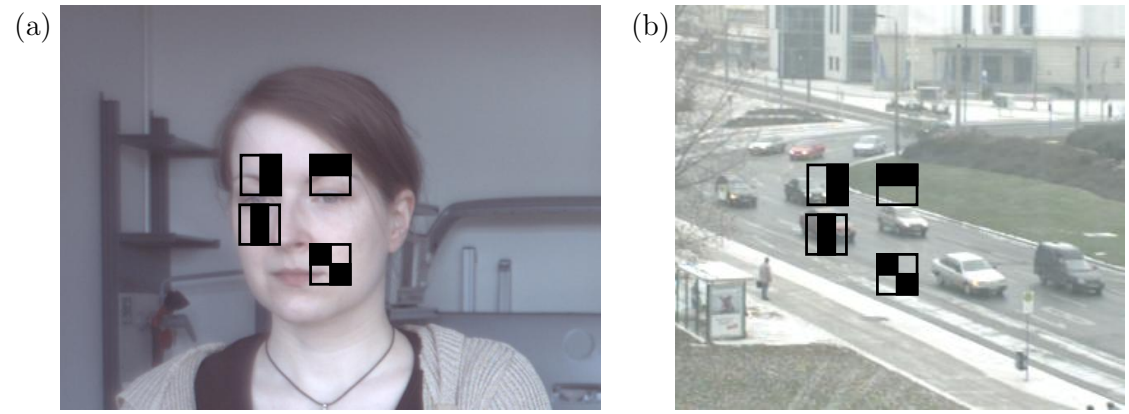


Abbildung 2-4: “Haar-like features“ nach Viola und Jones in einem Bild mit Gesicht (a) und ohne (b). Die Merkmale werden mit unterschiedlicher Skalierung und Position durch Subtraktion der dunklen und hellen Regionen auf der Grundlage des Integralbildes berechnet.

künstliche Neuronale Netze oder Support Vector Machines, wird in der Viola&Jones Methode eine Anzahl schwacher Klassifikatoren kombiniert, von denen jeder eine Schwellwertfilterung eines "Haar-like" Merkmals benutzt. Die schwachen Klassifikatoren werden dabei durch einen Ada-Boost Algorithmus ausgewählt und zu einem Mehrheitsentscheidungskriterium kombiniert.
Im Allgemeinen werden mehrere so antrainierte Klassifikatoren in einer Kaskade kombiniert. Dazu wird in jeder Stufe der Kaskade eine kleine Gruppe schwacher Klassifikatoren mit verschieden strenger werdenden Vorgaben antrainiert. Auf diese Weise lassen sich bereits zu einem frühen Zeitpunkt die negativen Fälle, d.h. Bildregionen ohne Gesicht effiziert ausschließen, so dass die aufwendigeren und somit stärkeren Klassifikatoren, die auf strengen Vorgaben basieren, nur noch gesichtsähnliche Merkmalskombinationen auswerten müssen. In der ursprünglich vorgestellten Implementierung bestand der finale Detektor aus einer Kaskade von 38 Stufen von Klassifikatoren, die insgesamt eine Menge von 6060 Merkmalen umfassen. Während die eigentliche Detektion sehr performant ist, ist das Anlernen eines solchen Systems rechenaufwendig und speicherintensiv. Möglichkeiten zu einer drastischen Geschwindigkeitssteigerung im Trainingsprozess werden z.B. in [Wu08] gegeben, wo ein Greedy Algorithmus zur Vorselektion der Merkmale im Trainingsprozess angewendet wird.
Aufgrund des großen Potentials des Viola&Jones Verfahrens wurden in der Vergangenheit eine Vielzahl an Anwendungen und Erweiterungen vorgeschlagen, z.B. zur Detektion von gedrehten Gesichtern, ebenso zur Erfassung von Menschen, Fußgängern, Autos und anderer Objekte [Hua07, Yam08].
Neben der Detektion von Gesichtern hat insbesondere die Detektion markanter Gesichtsmerkmalspunkte auf der Grundlage des Viola&Jones Verfahrens für die vorliegende Arbeit eine besondere Bedeutung (Abschnitt 4.4.1).

2.3.2 Bestimmung der Kopfpose

Neben der Detektion von Gesichtern als initialer Schritt, kommt der Bestimmung der Kopfpose, d.h. der Orientierung mit jeweils drei Freiheitsgraden zur Translation und Rotation in Bezug zu einem übergeordneten Koordinatensystem eine entscheidende Bedeutung für die Merkmalsextraktion zu. Eine Auswertung der 2D Bildinformation ohne Berücksichtigung der aktuellen Orientierung des Kopfes ist fehlerträchtig. In biometrischen Applikationen wie Gesichtserkennung lässt sich eine definierte frontale Kopfhaltung erzwingen, bei einer freien Mensch-Maschine Interaktion nicht. Des Weiteren stellt die Kopfpose selbst eine nützliche

Informationsquelle dar, wie z.B. zur Erkennung von Ablehnung oder Zustimmung des Nutzers durch Nicken oder Kopfschütteln sowie zur Feststellung der Blickrichtung, durch welche sich prinzipiell auf die durch den Nutzer fokussierten Objekte schließen lässt.

In [Mur09, Pat09] wird ein Überblick zu einer Vielzahl aktueller und gegenüberstellend älterer Ansätze zur Kopfposeschätzung gegeben. Grundsätzlich unterschieden werden muss zwischen Verfahren, die Daten aus monokularen bzw. stereoskopischen Kamerasystemen auswerten [Mal05]. Die Bestimmung der Pose aus monokularen Aufnahmen basiert üblicherweise auf einer modellgestützten Analyse von Parametern bzw. Merkmalen, die aus der Gesichtsregion extrahiert werden, nachdem diese durch einen Gesichtsdetektor erfasst wurde. Hierzu werden häufig Modelle verwendet, die die Gesichtsgeometrie berücksichtigen, etwa die relative Position markanter Punkte wie Augen und Nasenspitze.

Weidenbacher et al. beschreiben Methoden, die neurobiologische Wirkmechanismen zur Schätzung der Kopfpose und der Blickrichtung nutzen. Dabei verwendeten sie zum einen biologisch motivierte Filtertechniken [Wei06] sowie Lernverfahren wie z.B. STDP (Spike-Time Dependent Plasticity) zum Anlernen prototypischer Repräsentationen von Kopfposen [Wei08].

Andere Ansätze simulieren das Erscheinungsbild der Textur. Bei diesen sogenannten “Appearance“ Modellen werden die Variationen von Helligkeit und Farbe in der Gesichtsregion modelliert [LiS05, Mor10]. Anders als bei den geometriebasierten Modellen wird bei den Appearance-Modellen das Gesicht als Ganzes ausgewertet, weshalb diese auch als global oder holistisch bezeichnet werden.

In einer typischen appearance-basierten Methode wird das Bild des Gesichts so in eine andere Darstellung transformiert, dass eine Poseschätzung vereinfacht realisiert werden kann. Hierzu wird das Bild von einer pixelbasierten Darstellung in eine Pose basierte Repräsentation überführt [Mur09]. Eine solche Transformation wird typischerweise durch Auswertung großer Mengen an Trainingsdaten mit Bildern von Gesichtern in unterschiedlicher Pose angelernt. Dabei zielt die Transformation darauf ab, eine Repräsentation des Gesichts in Form einer Merkmalsmenge zu erreichen, in der andere als die posebedingten Variationen in den Vordergrund treten. Dies vereinfacht zum einen die nachfolgende Auswertung, z.B. zur Erkennung der Person, Mimik, etc., zum anderen kann durch Unterdrückung dieser nicht-posebedingten Variationen eben genau jene Pose berechnet werden.

In der Literatur werden die Unterschiede in den Merkmalsmengen, die von Gesichtern in derselben Pose berechnet werden als Intra-Class Variation bezeichnet, während die Unterschiede, die aus Gesichtern in verschiedener Pose resultieren,

als Inner-Class Variation bezeichnet werden. Eine leistungsfähige Transformation berechnet Merkmalsmengen mit kleiner Intra-Class und großer Inner-Class Variation [LiS05].
Nachdem für jedes Bild der Merkmalssatz berechnet wurde, wird ein Klassifikator zur Erkennung einer Reihe von bekannten Gesichtsposen angelernt. Bei einem einfachen Klassifikator wird dann die Pose in einem neuen Bild mit unbekanntem Gesicht, nach Berechnung des Merkmalssatzes, auf der Grundlage des nächsten Nachbarn, d.h. der nächsten bekannten Pose in den Trainingsdaten festgestellt [Pat09].

Geometriebasierte Methoden, zu denen auch die vorliegende Arbeit gehört, verwenden 3D Modelle, welche die Oberfläche des Gesichts mit einem gewissen Grad an Genauigkeit approximieren. In der Vergangenheit wurde hierzu eine Vielzahl von Ansätzen vorgestellt, in denen z.B. einfache Formen wie Halbzylinder [Xia02] oder Ebenen [Hor96] bis hin zu komplexeren 3D Geometrien [Vac04] verwendet werden, welche personenspezifisch oder generisch sein können.
Die zur Ermittlung der Pose benötigten Korrespondenzen zwischen markanten 3D Modell- und 2D Bildpunkten, wird normalerweise auf der Grundlage sogenannter Landmarken realisiert. Grundsätzlich beruht die Schätzung der Pose auf der Tatsache, dass sich bei bekannter Orientierung des 3D Modells und bekannten Parametern des Kameramodells, ebenfalls die 2D Projektionen der markanten Modellpunkte im Bild berechnen lassen (entsprechend Abschnitt 3.1). In der umgekehrten Reihenfolge heißt dies, dass bei bekannter Position der Landmarken im Bild auch die Pose in Weltkoordinaten berechnet werden kann (entsprechend Abschnitt 5.2.1).
Die Poseparameter, d.h. die Translationen und Rotationen sind zu Beginn der Berechnung meist nur näherungsweise bekannt. Daher stimmt die Projektion der markanten 3D Modellpunkte mit den detektierten Landmarken normalerweise nicht überein. Praktisch besteht hier ein Versatz zwischen „Soll“ und „Ist“ Wert. Aus diesem Grund lässt sich die Schätzung der Pose als ein Optimierungsproblem betrachten, bei dem unter Berücksichtigung der Freiheitsgrade des Modells, die Abweichungen zwischen den erwarteten 2D Positionen im Bild und den ermittelten Landmarken iterativ minimiert werden. Da die Bestimmung der Pose eine starre Transformation darstellt, dürfen nur solche Punkte im Gesicht verwendet werden, die nicht durch die Mimik verändert werden. Da die Berechnung weiterhin von einer nur kleinen Anzahl an korrespondierenden Punkten abhängt, erfordern geometriebasierte Ansätze zur Poseschätzung eine gewisse Mindestauflösung

des verwendeten Bildmaterials, um eine hinreichend genaue Detektion der Landmarken zu ermöglichen.

2.3.3 Merkmalsextraktion

Bei der bildbasierten Analyse von Gesichtern lassen sich die Ansätze zur Merkmalsextraktion nach Pantic et al. in mindestens drei Richtungen untergliedern [Pan07]. Zum einen stellt sich die Frage, ob die extrahierten Merkmale holistischer Natur sind, d.h. das Gesicht in seiner Textur als Ganzes betrachtet wird, oder analytisch-geometrisch einzelne Komponenten in Teilregionen beschrieben werden. Des Weiteren ist zu unterscheiden, ob der zeitliche Kontext berücksichtigt wird, d.h. die Auswertung auf statische Einzelaufnahmen beschränkt ist, oder die Dynamik der Mimik berücksichtigt wird. Als drittes Kriterium bietet sich eine Betrachtung der Dimensionalität extrahierter Merkmale an, womit zu unterscheiden ist, ob die erfassten Merkmale bildbasiert (2D) oder volumenbasiert (3D) sind. Während sich die Mehrzahl der in der Literatur vorgeschlagenen Verfahren zur Gesichtserkennung entsprechend der zuvor genannten Einteilung als holistisch, statisch und 2D basiert beschreiben lässt, werden zur Mimikerkennung vorwiegend analytisch-geometrische, regional definierte sowie dynamische Merkmale verwendet.

- Geometrische Merkmale

Mit geometrischen Merkmalen werden beispielsweise die Lage markanter Punkte und die Form von Mund, Augen, Augenbrauen, etc. aus dem Gesicht extrahiert. In der Vergangenheit wurde eine Vielzahl an Arbeiten veröffentlicht, die auf geometrischen Merkmalen basieren [Pan07, Tia05]. Große Unterschiede bestehen hier in der Komplexität der verwendeten Parameter, der damit verbundenen Realisierbarkeit in automatisierten Systemen und der erreichten Erkennungsrate bei der Mimikerkennung andererseits. Wang et al. werten zur Mimikerkennung dreidimsionale Formmerkmale auf der Grundlage von 64 manuell selektierten Markerpunkten aus, die durch Evaluierung einer 3D Datenbasis gewonnen wurden. Dabei führten sie einen Vergleich mit einer Reihe verschiedener Klassifikatoren durch [Wan06]. Chang et al. verwenden ein Shape Model, das mit 58 Kontrollpunkten parametriert wird [Cha06]. Andere Arbeiten zur Mimikerkennung kommen bei der Merkmalsextraktion mit deutlich kleineren Mengen an Merkmalspunkten aus, was den Vorteil bietet, dass diese recht zuverlässig vollautomatisch erfasst werden können [Pan08].

- Texturmerkmale

Verbreitet finden sogenannte "appearance features" (AF) Verwendung, die das aktuelle Erscheinungsbild einzelner Merkmale, wie z.B. mimikbedingte Fältchen in den betrachteten Gesichtsregionen erfassen. AF Daten eignen sich zum Anlernen von Bildfiltern mittels PCA/ICA[2], Gabor-Filtern oder Integralbildfiltern, wie z.B. "Haar-like features" und lokale Gradienten Histogramme und dienen weiterhin als Grundlage der Appearance Modelle [Del07]. Zur Überwindung von Skalierungseffekten, Bewegung, Beleuchtungsschwankungen und anderer Variationen, wird normalerweise eine Normierung des Gesichts in Bezug zu einem Referenzbild durchgeführt. Auch wenn häufig von einer Überlegenheit der AF Methoden über die analytischen geometriebasierten Verfahren berichtet wurde, z.B. mittels Gabor-Wavelets oder sogenannten Eigenfaces, ist ein pauschaler Vorteil eines Ansatzes momentan nicht sicher auszumachen [Pan07]. Es wurden weiterhin hybride Methoden vorgeschlagen, in denen geometrische und appearance Merkmale kombiniert und die Vorteile beider Techniken vereint wurden [Tia05].

- Dynamische Merkmale

Mimik wird durch Muskelkontraktionen verursacht und führt zur Bewegung der darüber liegenden Hautschichten. Dies wiederum bewirkt eine Verschiebung interessanter Gesichtsmerkmale mit einer entsprechenden Texturänderung in verschiedenen Regionen. Diese Veränderungen können mit Hilfe von Bewegungsanalyseverfahren, etwa durch Verfolgung markanter Punkte oder Konturmodelle, z.B. Active Shape Models [AlH06b], mit Hilfe intensitätsbasierter Matching-Verfahren oder durch die Analyse des Optischen Flusses (OF) erfasst werden. Eine solche Auswertung des Verschiebungsvektorfeldes (VVF) zur Erfassung mimikbedingter Bewegung bietet den Vorteil, dass keine exakte Beschreibung der Geometrie observierter Gesichtsmerkmale erforderlich ist [Nie07b]. Die Messung eines dichten VVFs ist somit im gesamten Gesicht möglich, selbst in Regionen mit evtl. schwach ausgeprägter Textur, wie Stirn oder Wangen. Der OF als das direkte Ergebnis mimikbedingter Bewegungsinformation kann fast unmittelbar zur Erkennung eingesetzt werden, was seit der ersten Veröffentlichung durch Yacoob et al. im Jahr 1994 [Yac94] vielfach in verschiedenen Erweiterungen und Abwandlungen publiziert wurde [Tia05]. In neueren Arbeiten zur automatischen Mimika-

[2] Die Hauptkomponentenanalyse (PCA) und Independent Component Analysis (ICA) sind statistische Verfahren zur Datenanalyse und werden u.a. zur Klassifikation von Texturmerkmalen eingesetzt.

nalyse wird versucht, grundsätzliche Probleme, die bei der Messung der Bewegungsinformation und dem Tracking von Merkmalspunkten auftreten, z.B. Rauscheinflüsse, Beleuchtungsänderungen oder Verdeckung durch Schätzverfahren wie Kalman- oder Partikelfilter zu umgehen bzw. zu reduzieren [Pan06].

2.3.4 Mimikerkennung

Im Erkennungsschritt wird aus den extrahierten Gesichtsmerkmalen eine Beschreibung der dargestellten Mimik generiert, wobei die überwiegende Zahl der aktuellen Verfahren dabei personenunabhängig arbeitet. Diese Beschreibung entspricht einer Interpretation, welche normalerweise in Form affektiver Zustände, d.h. Emotionen oder in Form aktivierter Muskeln gegeben wird, die dem dargestellten Gesichtsausdruck zugrunde liegen. Konzeptionell entspricht diese Unterteilung der Herangehensweise in der psychologischen Forschung, dem sogenannten "message" bzw. "sign judgment", d.h. der Bewertung des zugrundeliegenden Sachverhaltes einerseits, was der Zuordnung einer Emotionskategorie dienen kann und der Beschreibung der vorliedenden Mimik durch entsprechende Bewegungs- oder geometrische Formmerkmale andererseits [Coh08]. Während es beim "message judgment" einzig um die Interpretation und höhere Entscheidungsprozesse geht, ist das "sign judgment" völlig objektiv. Die am häufigsten verwendeten Kategorien für Gesichtsausdrücke bei interpretationsbasierten Verfahren sind die sechs Basisemotionen Freude, Überraschung, Wut, Ekel, Angst und Trauer. Diese wurden von Ekman und anderen Vertretern der diskreten Emotionstheorien als kulturübergreifend universell beschrieben, sowohl in der Präsentation als auch bei der Erkennung durch den Menschen [Kel00].
Die am häufigsten verwendeten Deskriptoren beim "sign judgment" sind die Action Units (AU) des Facial Action Coding Systems (s. Absatz 2.1.1).
Die meisten der bisher entwickelten automatischen Verfahren zur Mimikerkennung widmen sich der Analyse affektiver Zustände und versuchen eine Anzahl prototypischer emotionaler Gesichtsausdrücke zu erkennen [Pan07, Zen09]. Darüber hinaus sind auch Versuche unternommen wurden, andere Gesichtsausdrücke wie z.B. akute Schmerzen durch Bildverarbeitungs- und Mustererkennungsmethoden zu klassifizieren. Brahnam et al. haben dazu drei Klassifikatoren auf der Basis der Hauptkomponentenanalyse (PCA), Linearer Diskriminanzanalyse (LDA) und Support Vector Machines (SVM) vergleichend eingesetzt [Bra06].
Prinzipiell geschieht die Zuweisung einer Mimikklasse durch einen Klassifikator, entweder direkt auf der Grundlage benutzter Merkmale oder auf der Basis ermit-

telter Action Units. Eine Zwischenstufe zur Berechnung einzelner AUs wird häufig verwendet, stellt jedoch zur Kategorienzuweisung grundsätzlich kein Erfordernis dar. Eine mögliche Überführung der FACS Beschreibung in semantische Emotionskategorien wurde z.B. durch Fasel et al. beschrieben, bei der verschiedene Arten künstlicher neuronaler Netze vergleichend eingesetzt wurden [Fas04].
Die in der Literatur beschriebenen Ansätze zur Mimikerkennung basieren auf einer Vielzahl von Klassifikationsmethoden, vornehmlich maschineller Lernverfahren, welche überwacht auf einer geeigneten Datenbasis antrainiert werden. Besonders häufig werden dabei die Verfahren SVM (Support Vector Machines), künstliche Neuronale Netze, dabei insbesondere das Multilayer Perceptron, k-Nearest Neigbor sowie Bayes'sche Netze und regelbasierte Klassifikatoren verwendet [Tia05]. Zur Auswertung von Appearance Merkmalen kommen häufig PCA, LDA und ICA (Independent Component Analysis) zum Einsatz [Pan07]. Zur Analyse von Bildfolgen lassen sich HMMs (Hidden-Markov Modelle) nutzbringend einsetzen. Ein Beispiel stellt die Arbeit von Torre et al. dar, in der die zeitliche Segmentierung mimischer Gesten untersucht und das Ergebnis in das FACS überführt wurde [Tor07]. Dabei wurden die Merkmale mit Appearance Modellen verfolgt und die Formvariation mittels eines Zwei-Zustands-HMMs ausgewertet.
Eine häufig in der Literatur anzutreffende Aussage ist, dass sowohl dynamische als auch geometrische deformationsbasierte Mimikmerkmale wichtige Informationen zur Erkennung beinhalten [Pan07]. Dies motiviert den Versuch der vorliegenden Arbeit, eine kombinierte Auswertung dieser Daten durchzuführen.

Kapitel 3

Grundlagen

Zum besseren Verständnis der vorgeschlagenen Systemstruktur zur Mimikanalyse, welche konkret in Kapitel 5 vorgestellt wird, wird in diesem Kapitel ein kurzer Einblick in relevante Methoden aus tangierten Themengebieten gegeben. Insbesondere wird dazu im Hinblick auf die Bildaufnahme, Merkmalsextraktion und Mustererkennung auf Techniken der Photogrammmetrie, Stereobildverarbeitung, Verfahren der Bewegungs- und Korrespondenzanalyse und Farbbildverarbeitung eingegangen. Weiterhin werden die in dieser Arbeit eingesetzten Klassifikationsverfahren zur Erkennung der Mimik kurz vorgestellt.

3.1 Kameramodell

Kameramodelle ermöglichen eine mathematische Beschreibung der wesentlichen Eigenschaften eines Bildaufnahmesystems. Grundlagen solcher Modelle und die Technik der Photogrammmetrie im Allgemeinen finden sich z.B. in [Alb89, Luh03, McG04]. Das Lochkameramodell stellt dabei eine Vereinfachung dar, bei der die tatsächlichen physikalischen Eigenschaften des Lichts vernachlässigt werden, so dass für Lichtstrahlen eine gradlinige Bewegung in homogenen Medien angenommen wird. Im Beispiel in Abbildung 3-1 wird das Objekt von der linken Seite auf die Bildebene auf der rechten Seite projiziert. Die Bildebene ist dabei die Fläche in einer Box, zu der sich auf gegenüberliegender Seite ein Loch befindet, welches das Projektionszentrum darstellt. Die Lochkamera erzeugt damit ein auf dem Kopf stehendes Abbild des Objektes.
Aus dem Dreiecksverhältnis in Abbildung 3-1 wird ersichtlich, dass die Höhe des Objektabbildes durch h' (3.1) gegeben ist.

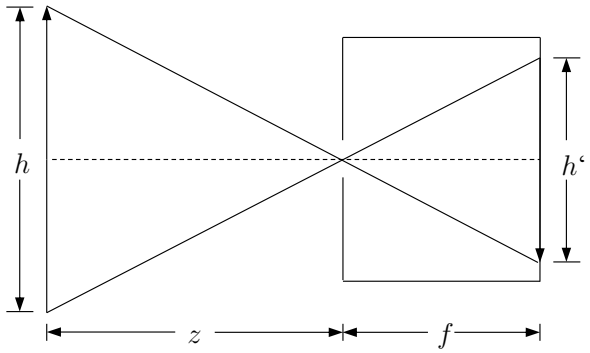


Abbildung 3-1: Lochkameramodell. Objekte auf der linken Seite werden durch ein Loch (Projektionszentrum) in eine Box abgebildet.

$$h' = hf/z, \tag{3.1}$$

wobei h die Höhe des Objektes repräsentiert, z die Distanz des Objektes zum Projektionszentrum und f die Brennweite als Abstand der Bildebene zum Projektionszentrum. Im Lochkameramodell entspricht die sogenannte Kamerakonstante c in etwa der Brennweite des Objektivs.

Um den physikalischen Gegebenheiten des Bildaufnahmesystems besser gerecht zu werden, müssen weitere Parameter, wie Linsenverzeichnung, Bildhauptpunkt und Seitenverhältnis der Pixel berücksichtigt werden, welche als sogenannte innere Kameraparameter $\mathbf{K}_i$ bezeichnet werden. Zusammen mit den äußeren Parametern der Kameraorientierung $\mathbf{K}_e$ werden diese in einem Kalibriervorgang, wie er aus der Vermessung bekannt ist, gewonnen [Luh03] (Abbildung 3-2). Im Weiteren werden diese Parameter als Kameramodell $\mathbf{K}$ zusammengefasst.

$$\mathbf{K} = \{\mathbf{K}_e, \mathbf{K}_i\} \tag{3.2}$$

Die äußeren Parameter $\mathbf{K}_e$ (3.3) beschreiben die Orientierung, d.h. Translation und Rotation im Weltkoordinatensystem, welche durch ein Kalibrierfeld festgelegt werden.

$$\mathbf{K}_e = \{t_{cx}, t_{cy}, t_{cz}, t_{c\omega}, t_{c\phi}, t_{c\kappa}\}, t_{cj} \in \mathbb{R} \tag{3.3}$$

Die inneren Parameter $\mathbf{K}_i$ beschreiben die geometrischen Kameraeigenschaften.

$$\mathbf{K}_i = \{c, s_y, \mathbf{h}, a_1, a_2\}, \tag{3.4}$$

dabei repräsentiert $c \in \mathbb{R}$ die Kamerakonstante, $s_y \in \mathbb{R}$ das Seitenverhältnis der Pixel, $\mathbf{h} \in \mathbb{R}^2$ den Bildhauptpunkt und $a_1, a_2 \in \mathbb{R}$ die Koeffizienten der radialsymmetrischen Linsenverzeichnung.

Auf der Grundlage des Kameramodells $\mathbf{K}$ (3.2) erfolgt die Berechnung der Transformation zwischen Bild-, Welt- und Kamerakoordinatensystem. Dabei lässt sich die Transformation eines 3D Punktes in Weltkoordinaten in den entsprechenden Bildpunkt in Pixelkoordinaten einfach durch die projektive Abbildung beschreiben, bei der zunächst eine Überführung in das Kamerakoordinatensystem erfolgt. Die gesamte Transformation eines Weltpunktes $\mathbf{p}$ in eine Subpixelkoordinate $\mathbf{i}_t$ wird im Folgenden als Funktion k (3.5) bezeichnet [Nie07a].

$$\mathbf{i}_t = k(\mathbf{p}, \mathbf{K}), \quad \mathbf{p} \in \mathbb{R}^3, \mathbf{i}_t \in \mathbb{R}^2, \text{ Kameramodell } \mathbf{K}. \tag{3.5}$$

Die inverse Funktion k^{-1} (3.6) transformiert einen Bildpunkt $\mathbf{i}_t$ in den 3D Weltpunkt $\mathbf{p}$. Da $\mathbf{i}_t$ mit nur zwei Koordinaten unterbesetzt ist, wird ein zusätzlicher Tiefenparameter d eingeführt, wobei dieser aus der aktuellen Tiefe der Szene bestimmt wird. Hierbei entspricht Parameter d dem Abstand zwischen der Bildebene an der Koordinate $\mathbf{i}_t$ und der Stelle, an der der entsprechende Kamerasehstrahl auf einen Oberflächenpunkt in der Szene trifft. In Kapitel 5 wird beschrieben, wie

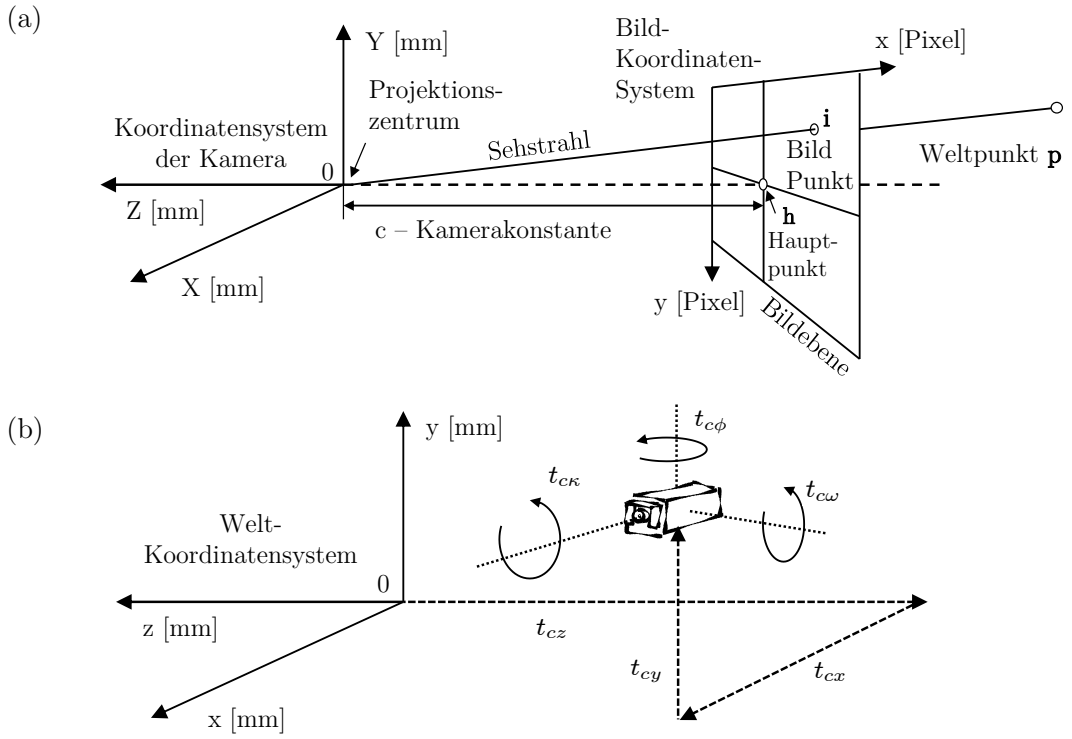


Abbildung 3-2: Kameramodell, (a) innere und (b) äußere Parameter.

eine solche Oberfläche durch ein geometrisches Modell des beobachteten Gesichts repräsentiert werden kann (s. Abschnitt 5.2.2).

$$\mathbf{p} = k^{-1}(\mathbf{i}_t, d, \mathbf{K}),\ \ \mathbf{p} \in \mathbb{R}^3,\ \mathbf{i}_t \in \mathbb{R}^2,\ d \in \mathbb{R}\,,\ \text{Kameramodell } \mathbf{K}. \tag{3.6}$$

3.2 Stereophotogrammetrische Messung

Mit Hilfe zweier gezielt ausgerichteter Kameras lässt sich die Position eines Objektes aufgrund geometrischer Beziehungen exakt bestimmen. Die Basis hierfür bilden sogenannte Stereoanalyseverfahren. Ein Ziel der Stereoanalyse ist die Rekonstruktion der Tiefeninformation von Objektoberflächen. Grundsätzlich unterschieden wird dabei zwischen aktiven Verfahren, welche mittels projizierter Lichtmuster aktiv die Szene beeinflussen und somit hochgenaue Messungen ermöglichen, und passiven Verfahren, die lediglich aus der Bildtextur Information über die Tiefeninformation gewinnen. Passive Verfahren weisen somit naturgemäß eine geringere Genauigkeit auf.
Prinzipiell werden Stereomessungen durch den Versatz eines Objektpunktes in zwei Aufnahmen, was auch als Disparität bezeichnet wird, realisiert [Mal00]. Abbildung 3-3 zeigt den schematischen Aufbau eines Stereokamerasystems. Im dargestellten Beispiel nehmen zwei Kameras mit gleichen inneren Parametern die Szene auf. In der vorliegenden Situation eines sogenannten Stereonormalfalls sind ihre optischen Achsen parallel zueinander ausgerichtet. Der Punkt $\mathbf{p}$ sei nun ein Punkt auf der Oberfläche des Körpers O. Dieser wird dabei auf die Bildebenen der beiden Kameras projiziert, so dass die Projektion durch die Punkte $\mathbf{i}_l(x_l, y_l)$ und $\mathbf{i}_r(x_r, y_r)$ beschrieben werden kann.

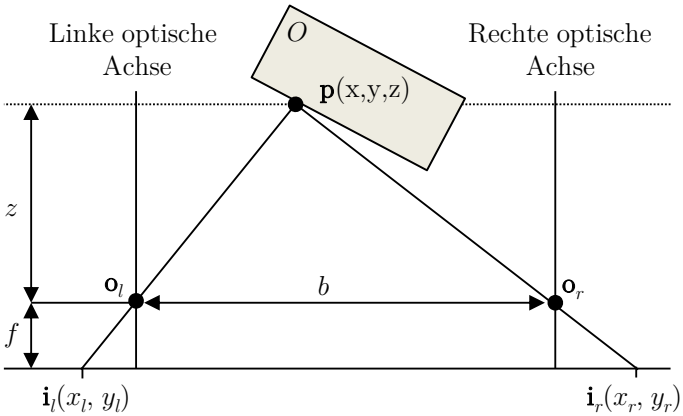


Abbildung 3-3: Schematischer Aufbau eines Stereokamerasystems mit Projektion eines Punktes $\mathbf{p}$ auf die Bildebenen der Kameras mit den Projektionszentren $\mathbf{o}_l$ und $\mathbf{o}_r$.

Dabei gelten die folgenden Beziehungen:

$$x_l = \frac{f \cdot x}{z} \qquad y_l = \frac{f \cdot y}{z}. \tag{3.7}$$

$$x_r = \frac{f \cdot (x-b)}{z} \qquad y_r = \frac{f \cdot y}{z}. \tag{3.8}$$

Um sicherzustellen, dass $y_l = y_r$, müssen beide Kameras auf gleicher Höhe montiert werden. Der Punkt $\mathbf{p}$ lässt sich nun aus der Disparität zweier Punkte $\mathbf{i}_l$ und $\mathbf{i}_r$, welche sich aus dem Abstand zwischen x_l und x_r, der Brennweite und dem Abstand der Kameras b ergeben, berechnen.

$$\frac{f \cdot x}{x_l} = \frac{f \cdot (x-b)}{x_r} = z \tag{3.9}$$

Die Koordinaten des Punktes lassen sich daraus umstellen zu:

$$x = \frac{b \cdot x_l}{x_l - x_r}, \quad y = \frac{b \cdot y_l}{x_l - x_r}, \quad z = \frac{b \cdot f}{x_l - x_r} \tag{3.10}$$

Bei der Stereoanalyse liegt das wesentliche Problem darin, den Abstand x_l-x_r zu bestimmen. Hierfür benötigt man Kenntnisse über die exakte Lage der Projektion des Punktes $\mathbf{p}$ in den beiden Kamerabildern. Durch Nutzung der Epipolargeometrie wird die Suche nach korrespondierenden Punkten zwar auf eine Zeile, die sogenannte Epipolarlinie beschränkt, was bei flächenhaften Pixelfeldern mit großflächig homogenen Texturen ohne hinreichende Grauwertunterschiede aber dennoch problematisch ist. Die durchzuführende Korrespondenzanalyse stellt daher keine triviale Aufgabe dar. Lösungsansätze finden sich in der intensitätsbasierten Korrespondenzanalyse und werden in [Kle96] detailliert behandelt. Häufig kommen zur Korrespondenzbestimmung sogenannte Matching-Verfahren wie z.B. die MAD Funktion zum Einsatz (s. Abschnitt 3.3.2).
Durch eine Projektion von Lichtmustern auf das zu vermessende Objekt, welches sich im Blickwinkel einer Kamera befindet, ist es möglich aufgrund der Veränderung der Struktur des Musters Rückschlüsse auf die Tiefeninformation des Objektes zu ziehen. Dieses Verfahren findet besonders dort große Verbreitung, wo eine dreidimensionale Vermessung automatisch und mit hoher Genauigkeit durchzuführen ist. Lichtmuster werden dabei mittels Laserlichtquelle oder Diaprojektor erzeugt. Die Berechnung der Tiefeninformation findet durch Anwendung der sogenannten Triangulation, bei bekannter Anordnung der Kameras statt. Die Verfahren zur 3D Objektvermessung setzen dabei je nach Zielstellung einfache geo-

metrische Lichtmuster wie Lichtpunkt- und Lichtstreifenprojektion, z.B. sogenannte Phasenshift-Verfahren [Alb98] und anderweitig kodierte, z.B. stochastische Muster ein.

3.3 Bewegungsanalyse und Korrespondenzbestimmung

Das grundsätzliche Anliegen der Bildfolgenanalyse ist die Bestimmung von Bewegungen in Bildszenen. Dabei wird versucht, die Änderungen innerhalb einer Bildsequenz als Bewegung der in ihr enthaltenen Objekte bzw. Strukturen zu interpretieren. Die Aufgabe der Verfahren zur Bewegungsanalyse besteht daher vornehmlich in der Bestimmung spezifischer Bewegungsgrößen. Eine typische Verarbeitungskette für eine Bewegungsanalyse zeigt Abbildung 3-4.

Abbildung 3-4: Verarbeitungskette bei der Bewegungsanalyse.

Entsprechend dieses Schemas erfolgt im ersten Schritt die Extraktion von Bildprimitiven wie etwa Punkten, Kanten, Flächen, usw. Durch die Zuordnung von Bildprimitiven in aufeinander folgenden Bildern entstehen Korrespondenzen, aus denen dann Parameter wie z.B. Bewegungsgrößen abgeleitet werden können. Die klassischen Verfahren zur Bestimmung von Korrespondenzen in zeitlichen wie auch räumlichen Bildpaaren, z.B. bei der Korrespondenzsuche in Stereobildpaaren, lassen sich wie folgt einordnen.

- Differentielle Verfahren nach dem Optischen Fluss [Jae05, Luc81, Hor81].
- Intensitätsorientierte Matching-Verfahren [Asc93, Mec99, AlH01],
- Merkmalsorientierte Matching-Verfahren [AlH06a],

Aufgrund ihrer Leistungsfähigkeit finden zur Erfassung mimischer Bewegung insbesondere die differentiellen Verfahren in der vorliegenden Arbeit Anwendung, während die intensitätsorientierten Matching-Verfahren ausschließlich zur Korrespondenzbestimmung bei der Stereoberechnung eingesetzt werden. Merkmalsorientierte Matching-Verfahren werden in dieser Arbeit nicht betrachtet, da diese auf das Vorhandensein bestimmter Bildprimitiven wie z.B. Kanten angewiesen sind [AlH06a], welche bei der Analyse von Gesichtern nicht immer vorhanden sind.

3.3.1 Differentielle Verfahren

Bei der Bildfolgenanalyse wird im Allgemeinen davon ausgegangen, dass die Änderungen der Intensität von Bild $\mathbf{I}_t$ zu Bild $\mathbf{I}_{t+1}$ durch relative oder absolute Bewegungen verursacht werden. Man darf jedoch nicht verallgemeinern, dass jede Grauwertänderung im Verlauf einer Bildsequenz ausschließlich aus einer Bewegung resultiert, da auch Beleuchtungsänderungen oder Reflexionen diese hervorrufen können. Es besteht daher das Problem, Grauwertänderungen zu unterscheiden, die entweder direkt oder indirekt aus einem Bewegungsmuster resultieren.
Die Verschiebung eines korrespondierenden Grauwertes im Bild $\mathbf{I}$ zwischen den Aufnahmezeitpunkten t_k und t_{k+1} resultiert in der Regel aus der Bewegung eines Oberflächenpunktes (Abbildung 3-5). Die ortsabhängige Verschiebung lässt sich dabei durch einen Verschiebungsvektor darstellen, der für den Bildpunkt im Ursprungsbild Richtung und Geschwindigkeit der Verschiebung beschreibt. Unter dieser Annahme ist das differentielle Verfahren eine Approximation des lokalen Verschiebungsvektorfeldes [Kle96]. Praktisch beschreibt das Verfahren den Übergang der Bildintensität $\mathbf{I}_t$ zu $\mathbf{I}_{t+1}$, d.h. den Verlauf der Änderungen der Pixel von Bild t zu t+1. Davon ausgehend soll die folgende Gleichung der sogenannten Bildwerttreue erfüllt sein:

$$\mathbf{I}_t\,(x, y, t) = \mathbf{I}_{t+1}\,(x + dx, y + dy, t + dt) \tag{3.11}$$

Unter der Annahme einer konstanten Beleuchtung und kleinen auftretenden Bewegungen ist die Grauwertfunktion $\mathbf{I}(x,\, y,\, t)$ linear. Somit lässt sich (3.11) durch eine Taylor-Reihenentwicklung entsprechend (3.12) für kleine $(dx,\, dy,\, dt)$ approximieren. Bei konstanten Grauwerten in aufeinander folgenden Bildern können die

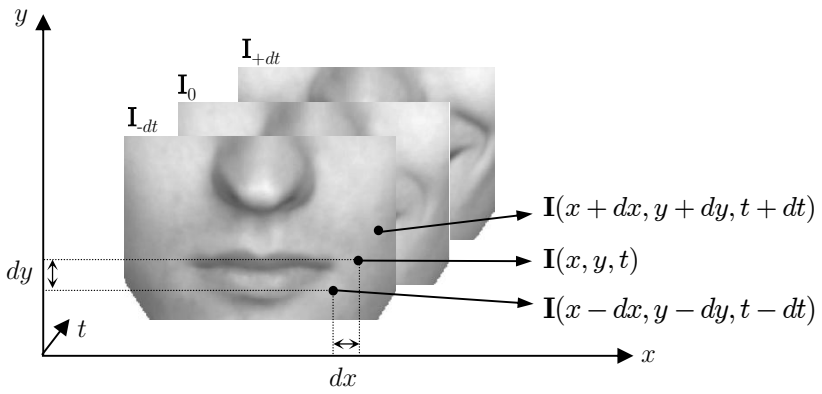


Abbildung 3-5: Verschiebung korrespondierender Grauwerte.

höheren nichtlinearen Terme γ vernachlässigt werden.

$$\mathbf{I}(x+dx, y+dy, t+dt) = \mathbf{I}(x,y,t) + \frac{\partial \mathbf{I}}{\partial x}dx + \frac{\partial \mathbf{I}}{\partial y}dy + \frac{\partial \mathbf{I}}{\partial t}dt + \gamma \tag{3.12}$$

$$u = \frac{dx}{dt} \Rightarrow dx = u \cdot dt \text{ und } v = \frac{dy}{dt} \Rightarrow dy = v \cdot dt \tag{3.13}$$

Durch Umformulierung beider Gleichungen ergibt sich die folgende, aus der Thermodynamik bekannte Kontinuitätsgleichung (3.14), welche man zur Ermittlung der Flussgeschwindigkeiten u und v verwenden kann.

$$\frac{\partial \mathbf{I}}{\partial x} \cdot u(x,y,t) + \frac{\partial \mathbf{I}}{\partial y} \cdot v(x,y,t) + \frac{\partial \mathbf{I}}{\partial t} = 0 \tag{3.14}$$

Bei der Projektion der Bewegung wird daher in Analogie zur Strömungslehre auch vom Optischen Fluss gesprochen [Hor81]. Da in der Gleichung nur eine Bedingung für die zwei gesuchten Parameter u und v formuliert wird, besteht jedoch keine eindeutige Lösung. In der Vergangenheit wurde eine Reihe von Methoden zur Berechnung des Optischen Flusses vorgestellt, die verschiedene Zusatzbedingungen annehmen. So wird im Verfahren nach Horn und Schunk eine Glattheitsbedingung eingeführt, welche davon ausgeht, dass bei bekanntem dt die benachbarten Bildpunkte ähnliche Bewegungen ausführen. Durch diese zusätzliche Annahme ist es möglich, die Berechnung des Verschiebungsvektors $\mathbf{h} = [u\ v]^T$ durchzuführen. Es ist jedoch zu beachten, dass besonders bei großen Bewegungen und starken Änderungen der lokalen Bildhelligkeit eine genaue Schätzung nicht möglich ist. Dies resultiert aus der Vernachlässigung der nichtlinearen Terme γ in (3.12). Ein weiterer Nachteil ist die Rauschempfindlichkeit des ermittelten Verschiebungsvektorfeldes, was darauf zurückzuführen ist, dass das Verfahren nach Horn-Schunk auf Pixelebene arbeitet [Kle96].
Im Gegensatz hierzu bietet der Ansatz nach Lucas-Kanade eine größere Rauschtoleranz und Robustheit. Die grundlegende Idee ist es, ein Fenster Ψ des Bildsignals zwischen dem aktuellen Bild t und nachfolgendem Bild t+1 zu finden. Dabei wird die Suche in der lokalen Nachbarschaft des Fensters Ψ_k durchgeführt (3.15). Folglich wird das LK Verfahren auch als lokal bezeichnet.

$$\begin{aligned} \mathrm{F_{LK}} = \sum_{\Psi} \left(\mathbf{I}_2(x+u, y+v) - \mathbf{I}_1(x,y)\right)^2 \approx \\ \sum_{\Psi} \left(\mathbf{I}_2(x,y) + u \cdot \frac{\partial \mathbf{I}_2}{\partial x} + v \cdot \frac{\partial \mathbf{I}_2}{\partial y} - \mathbf{I}_1(x,y)\right)^2 \end{aligned} \tag{3.15}$$

Im 2D Fall lassen sich die Parameter u und v als Verschiebungsvektor $\mathbf{h} = [u\ v]^T$ auffassen. Die Differenz $\mathbf{I}_t(x,y) - \mathbf{I}_{t+1}(x,y)$ kann als zeitlicher Versatz des Grauwertes betrachtet werden. Die Ableitungen von $\mathbf{I}(x,y)$ lassen sich mittels Differenzoperator definieren, so dass:

$$\begin{aligned} \mathrm{F}_{\mathrm{LK}}(\mathbf{h}) = \mathrm{F}_{\mathrm{LK}}(u,v) = \sum_{\Psi} (\mathbf{I}_x u + \mathbf{I}_y v + \mathbf{I}_t)^2 = \\ \sum_{\Psi} (\nabla \mathbf{I}^{\mathrm{T}} \cdot \mathbf{h} + \mathbf{I}_t)^2 \ \Rightarrow \ \min \end{aligned} \tag{3.16}$$

Um das Minimum zu bestimmen werden die Ableitungen gleich Null gesetzt:

$$\begin{aligned} \tfrac{\partial \mathrm{F}_{\mathrm{LK}}}{\partial u} &= 2 \sum_{\Psi} \mathbf{I}_x (\nabla \mathbf{I}^{\mathrm{T}} \cdot \mathbf{h} + \mathbf{I}_t) = 0\,, \\ \tfrac{\partial \mathrm{F}_{\mathrm{LK}}}{\partial v} &= 2 \sum_{\Psi} \mathbf{I}_y (\nabla \mathbf{I}^{\mathrm{T}} \cdot \mathbf{h} + \mathbf{I}_t) = 0 \end{aligned} \tag{3.17}$$

Gradientenmatrix $\mathbf{G}$ und Fehlervektor $\mathbf{b}$ lassen sich durch Umstellen der Summe herleiten. Folglich lässt sich der Verschiebungsvektor $\mathbf{h}$ durch Multiplikation mit der inversen Gradientenmatrix $\mathbf{G}^{-1}$ und Fehlervektor $\mathbf{b}$ bestimmen.

$$\begin{gathered} \mathbf{h} = \mathbf{G}^{-1} \cdot \mathbf{b} \\ \text{wobei} \\ \mathbf{G} = \sum_{\Psi} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}, \quad b = -\sum_{\Psi} \begin{bmatrix} I_x I_t \\ I_y I_t \end{bmatrix} \end{gathered} \tag{3.18}$$

Das beschriebene differentielle Verfahren nach Lucas-Kanade kann aufgrund seiner lokalen Sichtweise ebenfalls nur kleine Bewegungen präzise erfassen. Abhilfe schafft eine pyramidale Implementierung [Bou00], bei der durch Unterabtastung des Originalbildes das Signal auf verschiedenen Skalen ausgewertet wird, wodurch auch größere Bewegungen genau bestimmt werden können. In dieser Arbeit wird hierzu eine Hierarchie von drei Pyramiden und eine Fenstergröße mit einer lokalen Nachbarschaft von fünf Pixeln eingesetzt, was eine effiziente und robuste Berechnung des Optischen Flusses bei mimischer Bewegung ermöglicht.

3.3.2 Intensitätsorientierte Matching-Verfahren

Eine weitere Möglichkeit zur Ermittlung von Korrespondenzen zur Bewegungs- wie auch Stereoanalyse, bilden sogenannte Blockmatching-Verfahren, die auch als Korrelationsmethoden bezeichnet werden. Insbesondere zur binokularen Bildauswertung finden in dieser Arbeit Korrelationsmethoden Anwendung. Beim intensitätsorientierten Blockmatching werden meist flächenhafte Ausschnitte miteinander verglichen, welche unabhängig von Bildmerkmalen gewählt werden. Die Zuordnung der flächenhaften Ausschnitte wird als Matching bezeichnet. Das zugrunde liegende Prinzip beruht auf dem Vergleich eines Musters, das Referenzbereich genannt wird und normalerweise in Blöcken vorliegt, mit einem Suchbereich. Das Ziel ist es, die bestmögliche Übereinstimmung des Referenzbereichsblocks und eines Blocks innerhalb des Suchbereiches zu erhalten, um daraus den Verschiebungsvektor abzuleiten. Zur Verknüpfung der Blöcke wird ein Ähnlichkeitsmaß verwendet. Dieses liefert bei Übereinstimmung der Blöcke einen Extremwert, der je nach verwendetem Kriterium ein Maximum oder Minimum sein kann. Im Allgemeinen hängt es von der lokalen Struktur des Bildsignals ab, ob eine eindeutige Korrespondenzbeziehung ermittelt werden kann [AlH01] (Abbildung 3-6).

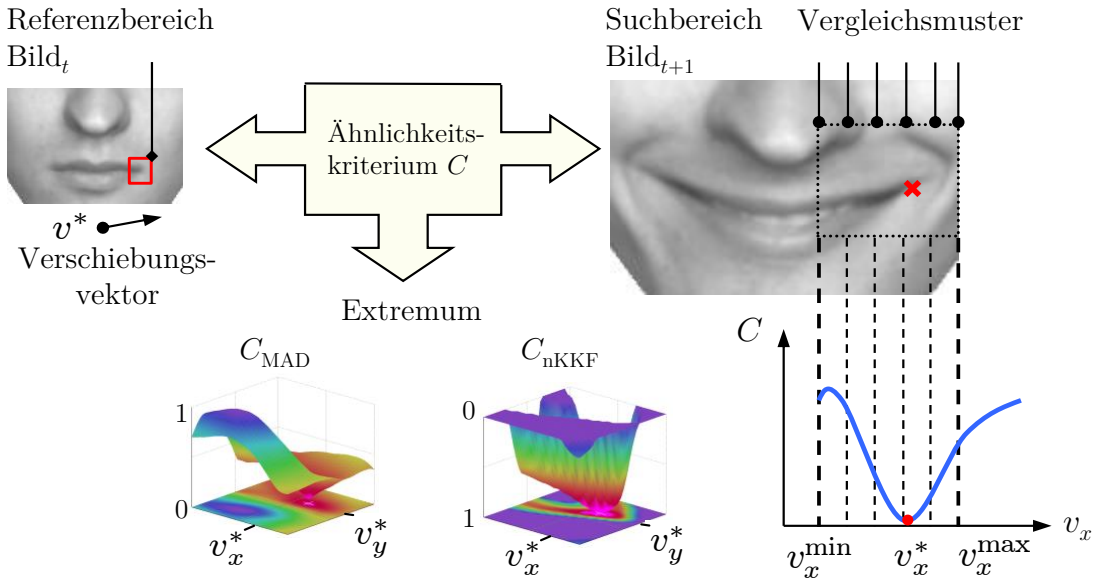


Abbildung 3-6: Blockmatching Methode mit Referenz- und Suchbereich, und Ähnlichkeitskriterien auf der Grundlage von MAD und nKKF.

Als Ähnlichkeitsmaße finden Korrelationsfunktionen, wie z.B. die Funktion der mittleren absoluten Differenz (MAD) (3.19) oder die normierte Kreuzkorrelationsfunktion (nKKF) (3.20) Anwendung.

$$\mathrm{MAD}(v_x, v_y) = \frac{1}{m_m \cdot m_n} \sum_{x=0}^{m_m-1} \sum_{y=0}^{m_n-1} |\mathbf{M}_{Ref}(x,y) - \mathbf{S}(x+v_x, y+v_y)| \tag{3.19}$$

$$\mathrm{nKKF}(v_x, v_y) = \frac{\sum_{x=0}^{m_m-1} \sum_{y=0}^{m_n-1} \mathbf{M}_{\mathrm{Ref}}(x,y) \cdot \mathbf{S}(x+v_x, y+v_y)}{\sqrt{\sum_{x=0}^{m_m-1} \sum_{y=0}^{m_n-1} \mathbf{M}^2_{\mathrm{Ref}}(x,y) \cdot \sum_{x=0}^{m_m-1} \sum_{y=0}^{m_n-1} \mathbf{S}^2(x+v_x, y+v_y)}} \tag{3.20}$$

mit Referenzbereich $\mathbf{M}_{Ref}$ und Suchbereich $\mathbf{S}$ sowie m_m und m_n als Blockdimension des Referenzbereichs.

Durch Einbeziehung des geometrischen Mittelwertes der lokalen Energien im Such- und Referenzbereich zeigt nKKF mehr Robustheit bei multiplikativen Helligkeitsschwankungen, ist dafür jedoch auch aufwendiger zu berechnen. Die MAD-Funktion liefert bei größtmöglicher Übereinstimmung ein Minimum als Extremwert, die nKKF ein Maximum [Mus85, Asc93].
Vorteilhaft bei derartigen Funktionen ist die einfache und vielseitige Möglichkeit zur Nutzung von Farbinformationen, bei der die Ähnlichkeitsfunktionen auf einzelnen Farbkanälen der Bilddaten arbeiten. Zu diesem Zweck werden die RGB-Werte einer meist linearen Farbraumtransformation zugeführt, welche in Fällen bestimmter Bildstörungen wie Helligkeitsschwankungen, eine Verbesserung des Signalverhaltens erzielen kann. Eine detaillierte Untersuchung dieser Thematik ist in [AlH01] zu finden. Wesentliche Nachteile des Blockmatchings sind der hohe Rechenaufwand, die feste Wahl des Blockrasters, das Fehlen eines eindeutigen (MAD) Minimums innerhalb homogener Gebiete und das Blendenproblem aufgrund rein lokaler Sicht.

3.4 Maschinelle Lernverfahren zur Klassifikation

Aus der Literatur ist eine Vielzahl von Verfahren zur Klassifikation von Messreihen bekannt, welche üblicherweise auf antrainierten Referenzdaten basieren. Die beiden im Moment gängigsten und am stärksten etablierten Ansätze sind künstliche neuronale Netze sowie Support Vector Machines (SVM). Insbesondere bei den

künstlichen neuronalen Netzen gibt es eine Vielzahl verschiedener Typen, welche sich vorrangig durch unterschiedliche Verbindungsarten und Netztopologien sowie Lernregeln unterscheiden und somit besondere Charakteristika und Eignungen für bestimmte Anwendungsgebiete besitzen. Insbesondere bei den neuronalen Netzen wird eine hinreichend große Menge an Lerndaten benötigt.

3.4.1 k-Nearest Neighbor Klassifikator

Der k-Nearest Neighbor (k-NN) Algorithmus gehört zu den einfachsten und populärsten Klassifikationsverfahren und ist besonders gut geeignet, wenn wenig bzw. kein Vorwissen über die Datenverteilung bekannt ist [Cov67, Mit97]. Hierbei wird k-NN zur Klassifikation von Objekten verwendet, die sich bezüglich eines Distanzmaßes am nächsten an den Trainingsbeispielen im Merkmalsraum befinden. Damit beruht das k-NN Verfahren auf einem instanzbasierten Lernen. Es wird auch als "Lazy Learner" bezeichnet, da der Teil des Lernens aus simplem Abspeichern annotierter Trainingsbeispiele besteht und die gesamte Berechnung bis zur Klassifikation aufgeschoben wird. Im Allgemeinen erzielt das k-NN Verfahren gute Klassifikationsergebnisse, wenn die Trainingsdaten repräsentativ und konsistent sind.
Eine der am häufigsten verwendeten Klassifikationsstrategien für ein durch einen Merkmalsvektor $\mathbf{x}$ beschriebenes Objekt beruht auf der einfachen Mehrheitsentscheidung [Mit97]. An dieser Mehrheitsentscheidung beteiligen sich die k zu $\mathbf{x}$ nächsten zuvor schon klassifizierten Objekte. Es existiert dabei eine Reihe von Distanzmaßen, wie etwa Euklidischer Abstand, Cosinus- und Manhattan-Distanz, die auch als Taxi- oder City-Blockdistanz bekannt ist. Ein durch Merkmalsvektor $\mathbf{x}$ beschriebenes Objekt wird der Klasse zugeordnet, die die größte Anzahl an Elementen in der Nachbarschaft hat, d.h. maximal k. Um keine unausgewogenen Ergebnisse zu erzeugen sollten die verwendeten Klassen eine in etwa gleich große Menge an Trainingsdaten aufweisen. Eine weitere Möglichkeit besteht darin, die Distanzen der k nächsten Nachbarn als Gewichtungsfaktor zu verwenden, so dass nahe Nachbarn stärker zur Klassifikation beitragen als entfernte.
Eine gewichtete Entscheidungsfunktion kann wie folgt repräsentiert werden:

$$f(\mathbf{x}_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} w_i \cdot \delta(v, f(\mathbf{x}_i)) \tag{3.21}$$

Mit w_i als distanzabhängiges Gewicht für den i-ten Nachbarn, δ als Entscheidungsfunktion und $f(\mathbf{x}_i)$ als zugehörige Klasse der i'ten Datenprobe. Eine Mehrheitsentscheidung kann mathematisch als

$$f(\mathbf{x}_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta\left(v, f(\mathbf{x}_i)\right) \tag{3.22}$$

definiert werden. Entscheidungsfunktion δ ist genau dann und nur dann wahr, wenn $f(\mathbf{x}_i)=v$. Eine Mehrheitsentscheidung kann mit einem unentschieden zwischen verschiedenen Klassen enden. Die beste Wahl des Parameters k hängt grundsätzlich vom Datenmaterial ab. Mit steigendem k sinkt die Rauschanfälligkeit, gleichzeitig verwischt jedoch die Grenze zwischen den Klassen. Eine gute Parameterwahl lässt sich durch heuristische Methoden wie beispielsweise Kreuzvalidierung bestimmen.
Die Nachteile des k-NN Klassifikators liegen neben der Empfindlichkeit gegenüber irrelevanten und redundanten Daten, welche alle gleichermaßen bei der Berechnung berücksichtigt werden, im hohen Rechenaufwand. Dieser kann jedoch durch Optimierungsschritte, z.B. durch binäre Suchbäume, reduziert werden. Vorteilhaft hingegen sind die Einfachheit des Verfahrens mit einer guten Interpretierbarkeit der Ergebnisse und die Tatsache, dass kein Training erforderlich ist. Im Hinblick auf Qualität und Performanz ist k-NN im Allgemeinen komplexeren Verfahren wie künstlichen neuronalen Netzen und Support Vector Machines klar unterlegen.

3.4.2 Multilayer Perceptron

Angelehnt an die Signalverarbeitung im Gehirn modellieren künstliche neuronale Netze (ANN - Artificial Neural Network) prinzipiell die Funktionsweise von Nervenzellen. Für die in dieser Arbeit durchgeführten Untersuchungen liegt aufgrund ihrer besonderen Eignung, der Fokus auf sogenannten Multilayer Perceptron (MLP) Netzen [Hay98, Kni07]. Diese stellen als mehrschichtige feedforward Netze eine spezielle Klasse von ANNs dar. MLPs bestehen dabei aus mehreren Verarbeitungsschichten, angefangen mit der Eingabeschicht. Genau hier werden merkmalsbasierte Daten in das Modell eingespeist.
Die Elemente der Ausgabeschicht repräsentieren hingegen das Klassifikationsergebnis. Zwischen der Ein- und Ausgangsschicht kann sich eine beliebige Anzahl versteckter Schichten mit einer beliebigen Menge versteckter Neurone befinden (s. Abbildung 3-7).

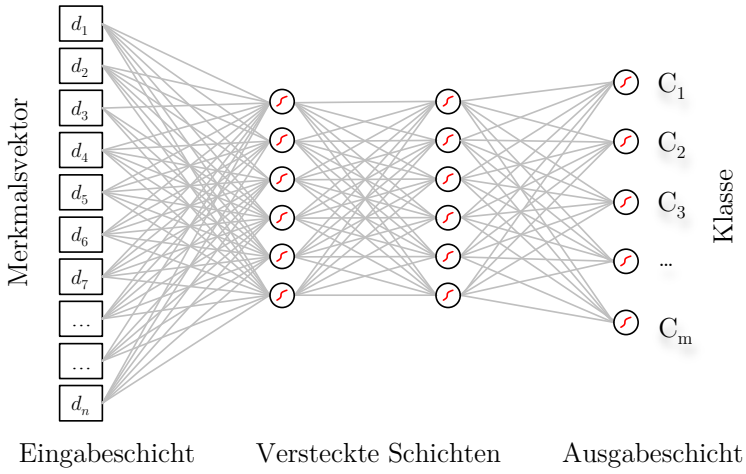


Abbildung 3-7: Struktur eines typischen voll vernetzten Multilayer Perceptrons.

Die Verbindungen und Gewichte zu und von diesen versteckten Neuronen entscheiden über die Performanz des Netzes. Beim überwachten Lernen werden MLPs mit Merkmalssätzen aus Lerndaten antrainiert, bei denen das gewünschte Klassifikationsergebnis bekannt ist. In dieser Arbeit finden insbesondere sogenannte Backpropagation Netzwerke mit einer sigmoiden Transferfunktion Verwendung. Während des Lernprozesses passt das Modell dazu die Gewichtungen in den verschiedenen Schichten an [Hay98].
Die Fähigkeit der MLPs zugrundeliegende Konzepte aus den gemessenen Daten abzuleiten sowie ihre hohe Verarbeitungsgeschwindigkeit beim Entscheidungsvorgang sind bestechende Vorteile dieses Verfahrens. Nachteilig sind hingegen die benötigten großen Mengen an Trainingsdaten, die nicht immer verfügbar sind.
Generell werden künstliche neuronale Netze zu Recht auch als “Black Boxes“ bezeichnet, da die in ihnen enthaltene Information nicht einfach zu interpretieren ist und somit nur schwer für die weitere Datenanalyse verwendet werden kann. Außerdem kann ein mathematischer Nachweis für die Korrektheit von ANNs bei der Anwendung nicht erbracht werden, obwohl diese empirisch erfolgreich eingesetzt werden [Kni07]. Ein Unter- und Übertrainieren mit Lerndaten kann sich als sehr problematisch erweisen. So kann es durch ein nicht ausreichendes Antrainieren zu Fehlklassifikationen kommen, da die Klassengrenzen zu ungenau oder falsch sind. Wenn sie hingegen zu sehr für eine spezielle Aufgabe angelernt werden und es zu

einer Überanpassung kommt, kann die Klassengrenze zu eng werden und die Generalisierungsfähigkeit[3] geht verloren.
Bei den MLP Netzen bieten sich zur Bestimmung geeigneter Parameter, beispielsweise Anzahl der versteckten Schichten bzw. Neurone, heuristische Methoden wie z.B. Kreuzvalidierung an. Auf diese Weise lassen sich mittels MLP für ein gegebenes Problem gute bis sehr gute Klassifikationsergebnisse erzielen.

3.4.3 Support Vector Machine

In der Anwendung liegt den Support Vector Machines (SVM) ein ähnliches Konzept wie den neuronalen Netzen zugrunde. Trainingsdaten, welche Merkmale aus einer Beobachtung enthalten, sollen in einem überwachten Lernvorgang als Grundwahrheit zum Antrainieren des Klassifikators dienen. Anstatt versteckte Schichten und anpassbare Gewichte zu verwenden wird bei den SVMs das Optimierungsproblem durch das gezielte Aufbauen einer hochdimensionalen Klassengrenze realisiert. Dabei wird der Randbereich zur Hyperebene zwischen den Klassen maximiert [Chri01, Suy03, Her03]. Aufgrund ihres Funktionsprinzips sind SVMs sehr gut dazu geeignet hochdimensionale Merkmalsräume zu verarbeiten. Grundsätzlich ist es das Ziel eine lineare binäre Entscheidungsfunktion $f(\mathbf{x})$ in den Merkmalsraum einzubringen (3.23), durch die üblicherweise ein hoher Generalisierungsgrad erreicht wird.

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}\mathbf{x} + \mathbf{b}) = \{-1,\ 1\} \tag{3.23}$$

mit $f(\mathbf{x})$ als zugehörige Klasse einer Stichprobe $\mathbf{x}$, Normalenvektor $\mathbf{w}$ und Bias $\mathbf{b}$.

Die zugrundeliegende Maximierung des Randes γ (3.24) entlang der klassentrennenden Hyperebene basiert auf dem Einsatz sogenannter Stützvektoren (Support Vectors). Wie in Abbildung 3-8 dargestellt lässt sich der Rand γ als minimaler Abstand der Hyperebene zu den Stützvektoren beschreiben.

$$\gamma = 2/\|\mathbf{w}\| \tag{3.24}$$

Der Abstand der Hyperebene vom Ursprung entlang des Normalenvektors wird dabei durch $\frac{\mathbf{b}}{||\mathbf{w}||}$ festgelegt. Die Einführung von Schlupfvariablen erlaubt es wei-

[3] Unter Generalisierung wird bei der Klassifikation die Fähigkeit zur Verallgemeinerung verstanden, die es ermöglicht unbekannte und vom Trainingsmaterial abweichende Merkmalssätze einer Klasse zuzuordnen.

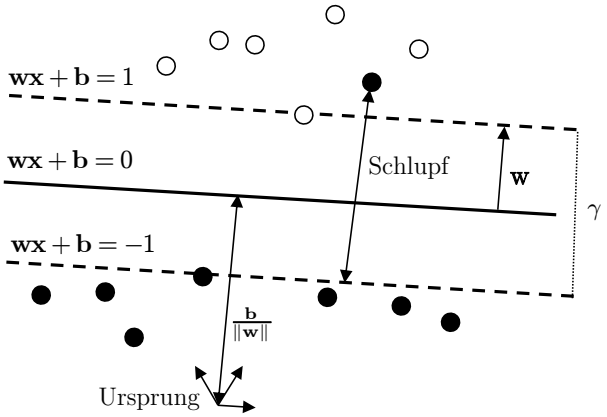


Abbildung 3-8: SVM, Randmaximierung mittels Stützvektoren.

terhin eine Überanpassung des Klassifikators zu vermeiden und die Anzahl der Stützvektoren zu senken, indem einzelne Messungen als falsch klassifiziert werden [Her03].
Da Trainingsdaten wegen möglicher Überlappungen und Messfehlern nicht immer linear trennbar sind, ist es erforderlich eine neue, linear trennbare Darstellung der Daten durch eine geeignete Transformation zu finden (3.25) (entsprechend Abbildung 3-9).

$$f(\mathbf{x}) = \sum_{i=1}^{m} w_i \Phi_i(\mathbf{x}) + \mathbf{b} \tag{3.25}$$

mit $\Phi : X \to F$ als nicht-lineare Transformation des Eingaberaums.

Eine zentrale Eigenschaft der SVM ist es, dass die Messdaten in einer sogenannten dualen Repräsentation vorliegen, woraus folgt, dass die Entscheidungsregel durch das Punktprodukt zwischen Trainings- und Testdaten formuliert werden

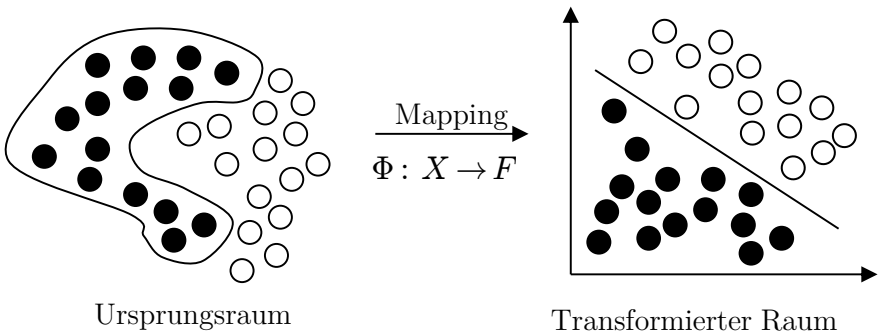


Abbildung 3-9: Transformation des Eingaberaums mittels Kernel Funktion in einen linear trennbaren höher dimensionalen Raum.

kann [Her03]. Diese Eigenschaft lässt sich folgendermaßen formulieren:

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i y_i \left(\Phi(x_i) \cdot \Phi(\mathbf{x})\right) + \mathbf{b} \tag{3.26}$$

mit α_i als Lagrange-Multiplikatoren und y_i als tatsächliche Klasse.

Durch den Einsatz von Kernel Funktionen, welche auf die Theorie der Integral Operatoren zurückgehen und das implizite Punktprodukt hat die Transformation in den höher dimensionalen Raum keinen Einfluss auf die Performanz des Lernvorgangs [Chri01, Suy03]. Die Kernel Funktionen definieren dabei die Berechnungsvorschrift für das Punktprodukt $\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})$ der Eingabevektoren. Ein großer Vorteil der Kernel Funktionen ist, dass die tatsächliche Berechnung der Merkmalsraumtransformation umgangen werden kann. Mathematisch lassen sich Kernel Funktionen folgendermaßen darstellen:

$$\mathrm{K}(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}) \tag{3.27}$$

Folgende Kernel Funktionen finden häufig Verwendung zur Transformation der Eingabedaten $\mathbf{x}$, $\mathbf{y}$ vom Ursprungsraum in den linear trennbaren Raum K($\mathbf{x}$, $\mathbf{y}$):

- Linear kernel : $\mathrm{K}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y})$
- RBF Gaussian kernel : $\mathrm{K}(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2}$
- Polynomial kernel : $\mathrm{K}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y})^d$
- Sigmoid kernel : $\mathrm{K}(\mathbf{x}, \mathbf{y}) = \tanh(\mathbf{k}\mathbf{x} \cdot \mathbf{y} + \mathbf{c})$

Da die meisten Klassifikationsprobleme mehr als nur zwei mögliche Klassen erfordern, beschäftigt sich ein aktuelles Forschungsgebiet der Mustererkennung damit, wie SVMs für Mehrklassenprobleme optimal eingesetzt werden können [Hsu02, Cha08]. Generell werden dabei zwei Ansätze verfolgt. Einerseits können alle Klassen in einem einzigen Optimierungsproblem betrachtet werden. Diese Herangehensweise hat jedoch den Nachteil, dass die Komplexität bei der Optimierung der Klassengrenzen erhöht und gleichzeitig die Skalierbarkeit eingeschränkt wird.
Der zweite Weg basiert auf einer binären Klassifikation vieler Hyperebenen mit anschließender Kombination in einem einzigen Klassifikator. Eine solche paarweise Verkopplung kann wiederum auf verschiedene Arten erfolgen, d.h. entweder einer-gegen-alle oder jeder-gegen-jeden. Beim ersten Ansatz werden n Hypothesen für n Klassen getroffen. Jeder i'ten Klasse wird eine 1 zugewiesen, während die restlichen Klassen als eine Klasse aufgefasst und auf -1 gesetzt werden. Dies ver-

einfacht das Optimierungsproblem bereits um ein vielfaches. Beim jeder-gegen-jeden Ansatz werden Trainingsdaten von jeweils genau zwei Klassen benötigt, um $n(n\text{-}1)/2$ Klassifikatoren anzulernen. Auch wenn die Anzahl der zu lernenden Hypothesen hier deutlich größer ist, erreicht diese Methode in vielen Situationen mehr Robustheit, da kleinere Optimierungsprobleme gelöst werden müssen.
Support Vector Machines haben eine Reihe von Vor- und Nachteilen. Wie die ANNs weisen sie eine hohe Generalisierungsfähigkeit auf. Darüber hinaus sind sie robuster bezüglich des Über- bzw. Untertrainierens. Weiterhin sind die Techniken zur Optimierung der SVMs mathematisch bewiesen. Ein Hauptnachteil besteht neben der verhältnismäßig hohen Laufzeit in der Auswahl vieler Parameter und Kernelfunktionen, die über die Leistungsfähigkeit des Verfahrens in der jeweiligen Anwendungssituation entscheiden. Idealerweise lässt sich eine gute Konfiguration durch Kreuzvalidierung bestimmen.

3.4.4 Selbstorganisierende Karten

Selbstorganisierende Karten (Self Organizing Maps - SOM), auch als Kohonen Map bezeichnet, wurden von Teuvo Kohonen 1981 eingeführt und stellen eine spezielle Klasse künstlicher neuronaler Netze dar, welche anders als MLP in einem unüberwachten Lernvorgang antrainiert werden. Ihre Hauptanwendung liegt darin aus einer Menge hochdimensionaler Merkmalsvektoren $\mathbf{v} \in \mathbb{R}^n$, eine meist zweidimensionale Repräsentation zu generieren [Koh97, Rit92]. SOMs sind somit hilfreich zur Auswertung und Visualisierung von Merkmalsräumen und stellen ein wichtiges Werkzeug im Data Mining dar. Neurobiologisch motiviert erhalten SOMs durch Verwendung von Nachbarschaftsfunktionen topologische Eigenschaften der Eingangsdaten. Dabei liegt die Idee zugrunde, dass analog zum menschlichen Kortex verschiedene Teile des Netzes ähnlich auf bestimmte Eingabemuster reagieren. Dies unterscheidet sie von klassischen künstlichen neuronalen Netzen.
Charakteristisch für SOMs ist eine Ausgabeschicht von Neuronen, die in einer quadratischen oder sechseckigen Gitterstruktur angeordnet sind. Dabei werden jedem Neuron n_i ein Referenzvektor $\mathbf{m}_i = [\mu_{i1}\ \mu_{i2}\ ...\ \mu_{in}]^{\mathrm{T}} \in \mathbb{R}^n$ und eine Gitterposition zugewiesen. Jeder Eingabevektor $\mathbf{x} = [\theta_1\, \theta_2\ ...\ \theta_n]^{\mathrm{T}} \in \mathbb{R}^n$ wird dabei mit den Referenzvektoren aller Neuronen n_j verglichen. Im einfachsten Fall wird hierzu der euklidische Abstand als Distanzmaß eingesetzt. Die Position des am besten passenden Neurons, der sogenannten Best Matching Unit (BMU), wird als Ort der größten Übereinstimmung ausgegeben. Bezüglich des eingesetzten Distanzma-

ßes hat dabei der Referenzvektor der BMU den geringsten Abstand zum Eingabevektor [Koh97].
Wie bei den meisten künstlichen Neuronalen Netzen arbeiten SOMs in zwei Stufen, d.h. anlernen und klassifizieren. Beim Anlernen aus einem Satz von Beispielen erfolgt die Aktivierung der Neuronen abhängig von der topologischen Distanz zur BMU (Abbildung 3-10(a)). Je nach Initialisierungsstrategie passt sich im Training das Netzwerk der Referenzvektoren der Verteilung der Eingabevektoren unterschiedlich schnell an (Beispiel in Abbildung 3-10(b)). Der Aktualisierungsschritt kann dabei entsprechend (3.28) gewählt werden.

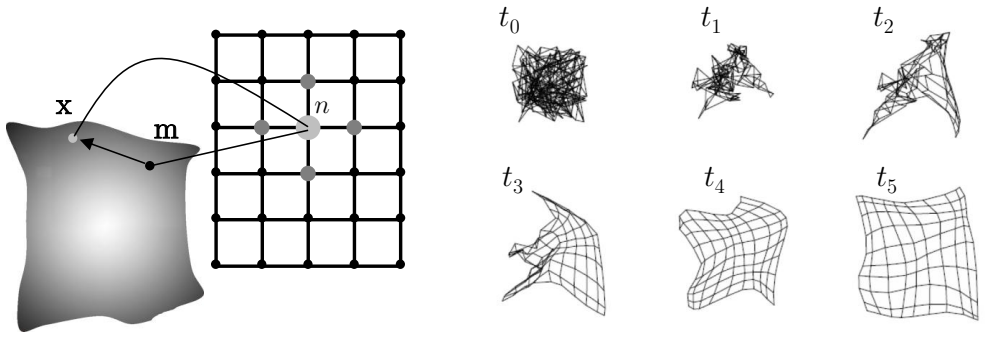
$$\mathbf{m}_i\,(t+1) = \mathbf{m}_i\,(t) + (\mathbf{x}(t) - \mathbf{m}_i(t))\,h_{ci}\,(t) \tag{3.28}$$

In Gleichung (3.28) stellt $h_{ci}(t)$ die Nachbarschaftsfunktion dar, deren Aufgabe es ist, den Lerneffekt in Abhängigkeit des Abstandes zur BMU abzuschwächen und weiterhin eine zeitabhängige Lernrate zu implementieren. Um die Konvergenz des Trainings sicherzustellen ist es nötig, dass $h_{ci}(t)$ gegen Null konvergiert, wenn t gegen Unendlich strebt. Auch wenn die Abstandsfunktion von h_{ci} unterschiedlich gewählt werden kann, kommt dabei häufig eine Gauß-Funktion zum Einsatz (3.29).

$$h_{ci} = \alpha(t) \exp\left(\frac{-\|\mathbf{r}_c - \mathbf{r}_i\|^2}{2\sigma^2(t)}\right) \tag{3.29}$$

wobei $\alpha(t)$ die Lernrate und $\sigma(t)$ die Nachbarschaftsbreite definiert, $\mathbf{r}_c$, $\mathbf{r}_i$ sind die 2D Ortsvektoren des betrachteten Neurons.

Bei der Klassifikation, dem sogenannten Mapping, wird ein Testvektor auf die



(a) (b)

Abbildung 3-10: Self Organizing Map, (a) Lernschritt, Eingabevektor $\mathbf{x}$ wird das Neuron n zugeordnet, dessen Referenzvektor $\mathbf{m}$ im Merkmalsraum am nächsten liegt, (b) iterative Anpassung einer SOM an einen 2D Merkmalsraum [Kru09].

zweidimensionale Karte abgebildet. Ein wesentlicher Effekt der SOMs ist, dass die Topologie der Eingabedaten erhalten bleibt und somit ursprünglich dicht beieinander liegende Merkmalsvektoren in der Karte auf ähnliche Positionen abgebildet werden. Folglich werden Cluster in den Merkmalsdaten auch in der SOM auftreten. Zur Darstellung dieser Cluster werden verschiedene Visualisierungen wie die sogenannten U-, P-, und U* Matrizen eingesetzt [Ult03]. Bei der am häufigsten verwendeten U-Matrix (unified distance matrix) werden die summierten Abstände der Referenzvektoren zu ihren Nachbarn visualisiert. Der Wert der U-Matrix für ein Neuron n_i ist definiert als:

$$u(n_i) = \sum_{n_j \in \mathbf{U}_i} d(n_i, n_j) \tag{3.30}$$

mit $\mathbf{U}_i$ als Menge aller Nachbarn von n_i.

Für gewöhnlich zeigen Merkmalscluster, die verschiedenen Klassen zugeordnet sind separate Regionen, die durch deutliche "Gräben" in der U-Matrixdarstellung erkannt werden können. Somit kann die zugrunde liegende Struktur der Daten analysiert werden und auf eine grundsätzliche Eignung von Merkmalsräumen zur Klassifikation geschlossen werden.

3.4.5 Kreuzvalidierung

Kreuzvalidierungsverfahren sind statistische Methoden, mit denen die Güte eines Vorhersagemodells nachgewiesen werden kann [Har05]. Damit kommen diese Verfahren häufig bei der Überprüfung neuer Erkennungsmethoden zum Einsatz.
Bei überwachten Lernverfahren wie z.B. k-NN, MLP und SVM wird diese Technik insbesondere zur empirischen Parameterbestimmung verwendet, durch die ein optimales Erkennungsergebnis ermöglicht wird.
Hierzu werden oft die folgenden Implementierungen verwendet:

Einfache k-fache Kreuzvalidierung
Bei dieser Methode wird das vorhandene Datenmaterial, das aus n Stichproben besteht in k ($k \leq n$) Teilmengen $\mathbf{T}_1,\ldots,\mathbf{T}_k$ zerlegt. Im Anschluss werden k Durchläufe durchgeführt, in denen jeweils eine Teilmenge $\mathbf{T}_i$ zum Testen und alle verbleibenden k-1 Teilmengen zum Trainieren des Klassifikators verwendet werden. Der Gesamtfehler bei der k-fachen Kreuzvalidierung errechnet sich aus den ermittelten k Einzelfehlerraten.

Stratifizierte Kreuzvalidierung
Bei einer ungefähr gleichen Verteilung der den k einzelnen Teilmengen zugerundeliegenden Sample-Klassen wird auch von einer stratifizierten k-fachen Kreuzvalidierung gesprochen. Diese hat den Vorteil, dass eine eventuell ungleiche Verteilung der Daten nicht zu einer falschen Schätzung des Gesamtfehlers führt.

Leave-One-Out Kreuzvalidierung
Eine Leave-One-Out Kreuzvalidierung entspricht dem Spezialfall einer k-fachen Kreuzvalidierung, bei der gilt k=n. Somit sind n Durchläufe erforderlich, bei denen sich die Gesamtfehlerquote als Mittelwert über alle Einzeltests berechnet.
Neben dem hohen Rechenaufwand hat diese Methode den Nachteil, dass eine Stratifizierung der Teilmengen nicht möglich ist, was in Extremfällen zu falschen Ergebnissen führen kann.

3.5 Hautfarbmodelle

Bei der Detektion von Gesichtern und zugehöriger Merkmale kann Farbinformation nutzbringend eingesetzt werden und bietet deutliche Vorzüge gegenüber grauwertbasierten Verfahren. Insbesondere ist eine Segmentierung von Hautfarbe invariant gegenüber Skalierung und Rotation. Grundsätzlich ist die Haut nicht durch eine einzige mehr oder weniger homogene Farbe beschreibbar, da Schattierungen durch verschiedene Lichteffekte wie z.B. die sogenannte Volumenstreuung, Unregelmäßigkeiten bzw. Unreinheiten, Rauschen, etc. das Erscheinungsbild beeinflussen. Somit wird eine Modellierung der Hautfarbe durch umgebungs- und personenspezifische Einflüsse erschwert. Hierzu gehören vorrangig:

- Hautfarbtyp: ethnisch bedingte Variationen der Hautfarbe,
- Beleuchtung: Farbton und Intensität einer Lichtquelle,
- Kameraparameter: Sensortyp, Belichtungsdauer, Farbkontrast,
- Umgebungssituation: Fremdobjekte mit Hautfarbcharakteristika.

Da aus den genannten Gründen die Haut einer Person im Bild Farbvariationen mit unterschiedlichen Häufigkeiten aufweist, werden zur Modellierung der Hautfarbe für gewöhnlich statistische Modelle verwendet, insbesondere Verteilungsfunktionen und Histogramme [LiS05]. Der dabei eingesetzte Farbraum wird meist von der Bildaufnahmesituation abhängig gemacht. Somit hängt die Qualität der Ergebnisse von den Referenzdaten ab, anhand derer das Hautfarbmodell erstellt

wird. Zudem spielt die Art der Modellierung eine wesentliche Rolle. In der Literatur finden sich dazu Anwendungen mit einer Vielzahl an Farbräumen.
Farbräume lassen sich prinzipiell in technisch-physikalische und wahrnehmungsorientierte Räume unterteilen. Der RGB-Farbraum leitet sich als bekanntester Vertreter technischer Farbräume aus der von Helmholtz begründeten trichromatischen oder Dreifarbentheorie ab [Gol07], welche die Aktivität der drei primären Klassen von lichtsensitiven Zellen (Zapfen) in der menschlichen Retina beschreibt. In der praktischen Umsetzung verwenden die technischen Farbräume zur Beschreibung einer Farbe eine Mischung aus drei Primärfarben. Unterschiede zwischen den Farbräumen bestehen in der Farbmischung und der Wahl der Primärfarben, was den Erfordernissen der Ausgabegeräte angepasst wird.
Die wahrnehmungsorientierten Farbräume leiten sich aus der Gegenfarbtheorie ab, welche auf physiologischer Ebene eine Dekorrelation der Farb- und Helligkeitsinformation durch eine besondere Verschaltung in den Neuronen der retinalen Ganglienzellen erklärt. Diese betrachten wie das natürliche Vorbild die Helligkeit getrennt von der Farbinformation. Hier sind z.B. der HSV (Hue, Saturation, Value) oder HUV-Farbraum zu nennen [Fol95].
Für Hautfarbmodelle kommen beide der zuvor genannten Kategorien von Farbräumen zum Einsatz. Häufig werden der normalisierte RGB-, bzw. HSV-Raum oder die aus der Übertragungstechnik bekannten Y-Modelle genutzt [LiS05].
Beim Einsatz statistischer Verteilungsmodelle zur Hautfarbmodellierung wird davon ausgegangen, dass den Farbwerten eine bekannte Wahrscheinlichkeitsverteilung zugrunde liegt. Häufig wird hier eine Normalverteilung angenommen, welche sich durch den Erwartungswert μ und Varianz σ^2 definiert (3.31) und besonders dann geeignet ist wenn die Verteilung nur einen Schwerpunkt aufweist.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.31}$$

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{3.32}$$

Für einen Beobachtungswert y ergibt sich dessen Wahrscheinlichkeit P durch:

$$P(y) = \frac{1}{\sqrt{2\Pi|\sigma^2|}} \cdot \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \tag{3.33}$$

Bei der Übertragung auf einen m-dimensionalen Vektorraum lassen sich x_i und y als m-dimensionale Vektoren auffassen. Die Varianz entspricht somit einer $m \times m$

dimensionalen Kovarianzmatrix Σ, entsprechend folgt aus den Gleichungen (3.31) und (3.32):

$$\Sigma = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^T (x_i - \mu) \tag{3.34}$$

$$P(y) = \frac{1}{(2\Pi)^{m/2}\sqrt{|\Sigma|}} \cdot \exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right) \tag{3.35}$$

Gleichung (3.35) stellt die sogenannte Mahalanobis-Distanz dar. Die Verteilung entspricht dabei einer Ellipse mit der größten Wahrscheinlichkeit im Zentrum, während diese nach außen hin kontinuierlich abfällt. Ausdehnungsrichtung und Steilheit leiten sich aus der Kovarianzmatrix ab.
Für Verteilungen, die mehr als einen Schwerpunkt aufweisen finden Gauß-Mixtur Modelle Anwendung. Hierbei werden die Eingangsdaten des Hautmodells durch einen Clusteralgorithmus, z.B. nach der Expectation-Maximization (EM) Methode in eine Anzahl von Partitionen unterteilt, wobei jeder der entstandenen Cluster eine eigene Gauß-Verteilung mit zugehöriger Kovarianzmatrix Σ_i und Mittelwert μ_i bildet. Der EM Algorithmus erzeugt des Weiteren einen normierten Gewichtungsfaktor Λ_i zu jedem der n Cluster. Eine Beobachtung y tritt dann mit einer Wahrscheinlichkeit $P(y)$ ein (3.36), die sich aus der Summe der gewichteten Einzelwahrscheinlichkeiten für jedes Cluster (3.37) ergibt.

$$P(y) = \sum_{i=1}^{n} \Lambda_i P_i(y|skin) \tag{3.36}$$

mit

$$P_i(y|skin) = \frac{1}{(2\Pi)^{m/2}\sqrt{|\Sigma_i|}} \cdot \exp\left(-\frac{1}{2}(y-\mu_i)^T \Sigma_i^{-1} (y-\mu_i)\right) \tag{3.37}$$

Die Anzahl der verwendeten Cluster ist generell applikationsabhängig, grundsätzlich steigt aber der Rechenaufwand mit deren Anzahl. Die Möglichkeit auch nicht triviale Verteilungen erfassen zu können, verhalfen den Gauß-Mixtur Modellen bei der farbbasierten Auswertung, insbesondere für die Mimik und Gestikanalyse zu verbreiteter Anwendung [Elm09].

Kapitel 4

Gesichtsmodell und Einführung dynamischer und geometrischer Merkmale

4.1 Motivation

Ein zentraler Gegenstand dieser Arbeit ist die Erfassung und Auswertung dynamischer und geometrischer Merkmalsdaten zum Zweck der automatisierten bildbasierten Mimikerkennung. In diesem Kapitel werden die Grundlagen zur Definition dieser Merkmale dargelegt, d.h. das verwendete Gesichtsmodell sowie relevante Merkmalsregionen und Merkmalspunkte motiviert und die konkreten Merkmale spezifiziert. Die Integration in die vorgeschlagene Systemstruktur unter Berücksichtigung zweier Ansätze (Gesichts- bzw. Merkmalsnormierung) wird im Folgekapitel erläutert.
Im vorgeschlagenen Konzept repräsentieren die dynamischen Merkmale durch Mimik verursachte kurzzeitige Veränderungen des Antlitzes, die sich durch flächenhafte Verschiebungen der Gesichtsoberfläche darstellen und bildbasiert mit Hilfe des Optischen Flusses, inklusive zeitlicher Filterung erfasst werden. Dynamische Merkmalsdaten ermöglichen eine schnelle und frühe Detektion von Bildänderungen was grundsätzlich zu einer Verbesserung der Erkennungsrate führt. Um eine effiziente und genaue Auswertung zu ermöglichen erfolgt die Erfassung dynamischer Merkmale nicht im gesamten Gesicht, sondern ausschließlich in sogenannten physiologisch motivierten Regionen (Abschnitt 4.3).
Anders als die dynamischen Merkmale lassen sich die geometrischen Merkmale aus einem einzigen Bild bestimmen und repräsentieren den Zustand der Mimik zu einem aktuellen Zeitpunkt t. Grundsätzlich beruhen die geometrischen Merkmale auf der Auswertung mimikrelevanter Merkmalspunkte (Abschnitt 4.4). Dabei

werden durch den Einsatz von Gesichtsmodellen und photogrammetrischen Techniken Maße wie Abstände und Winkel im dreidimensionalen Raum ausgewertet.
Während geometrische Merkmale zu jedem Zeitpunkt erfasst werden können, sind die dynamischen Merkmale aufgrund ihres differentiellen Charakters nur bei Änderungen der Mimik messbar. Insbesondere liegt dieser Arbeit die Hypothese zugrunde, dass durch eine integrierte Auswertung dynamischer und geometrischer Merkmale eine verbesserte Erkennungsleistung erzielt werden kann.

4.2 Gesichtsmodell durch Stereomessung

Die Extraktion dynamischer und geometrischer Merkmale beruht im vorgeschlagenen Konzept auf der Verwendung dreidimensionaler Modelle. Aus der Literatur ist eine Reihe von Techniken zur Erstellung flächenhafter Gesichtsmodelle bekannt. Sehr gut für die Synthese sind die sogenannten Morphable Models [Bla99] geeignet, welche heute vielfach Anwendung in der Visualisierung, etwa für Computerspiele finden [Sin09]. Bei dieser Technik werden durch Auswertung einer umfangreichen Datenbasis statistische Form- und Texturparameter erfasst und zugeordnet, so dass neue, nicht in der Trainingsmenge enthaltene Gesichter generiert werden können. Nachteilig ist an dieser Technik jedoch ein hoher manueller Aufwand bei der Parametrisierung.
Für die schnelle vollautomatische Erzeugung eines einfachen, starren Gesichtsmodells wurde daher in dieser Arbeit eine Technik entwickelt, die auf einer Stereomessung beruht und in einer Reihe von Verarbeitungsschritten, wie Gruppierung der Punkte zur Gesichtslokalisation in 3D, Triangulation und Glättung, die Oberfläche des Gesichts der Versuchsperson approximiert. Hierzu wird das Gesicht in der Frontalen bei neutraler Mimik initial mit einem Stereokamerasystem erfasst, welches simultan auch Farbbilddaten aufzeichnet. Das verwendete Stereokamerasystem kann dabei ein aktives oder auch passives sein, was die Handhabung deutlich vereinfacht.
Eine dauerhafte exakte 3D Erfassung des Gesichts, d.h. auch während des Auftretens verschiedener Gesichtsausdrücke ist jedoch nicht praktikabel, da dies die Verwendung von Markern bzw. eine Projektion von Lichtmustern erfordert, welche die Probanden stören würden.

4.2.1 3D-Gesichtslokalisation durch Clusterbildung

Die Erzeugung eines Gesichtsmodells erfordert im ersten Schritt eine Lokalisation des Gesichts in der 3D-Messpunktwolke der Stereoaufnahme. Da die zugrundeliegenden Daten dreidimensional sind, können herkömmliche auf Bildern basierende Verfahren zur Gesichtsdetektion (s. Abschnitt 2.3.1) nicht ohne weiteres eingesetzt werden. Aus diesem Grund wurde eine neue Methode entwickelt, die Gesichter in 3D lokalisiert. Die Methode eignet sich dabei sowohl für die dichten Punktmengen der aktiven Stereomessung als auch für die von Ausreißern und Löchern behafteten Punktmengen der passiven Messung. Dabei wird die Beobachtung genutzt, dass Oberflächen durch zusammenhängende Punktcluster abgebildet werden, fehlerhafte Messdaten dagegen häufig als geometrisch isolierte Punkte. Ferner wird die Farbinformation genutzt, die zu jedem 3D-Punkt durch Rückprojektion auf eines der Stereofarbbilder ermittelt wird. Da das Gesicht eine hinreichend kontinuierliche Farbgebung aufweist, kann aus der Kombination eines Farbähnlichkeitsmaßes und des euklidischen Abstandes ein Homogenitätskriterium definiert werden, auf dessen Grundlage die disjunkte Zerlegung einer Messpunktwolke $\mathbf{P}$ in eine Menge von Häufungsbereichen, sogenannten Clustern $\mathbf{C}_i$ erfolgt (4.1).

$$\mathbf{P} = \bigcup_i \mathbf{C}_i,\ \mathbf{C}_i = \{\mathbf{p}_1,\ \dots\ ,\ \mathbf{p}_{n_i}\},\ \mathbf{p}_j \in \mathbb{R}^3 \tag{4.1}$$

Alle Cluster $\mathbf{C}$ sind dabei disjunkt, d.h.

$$\forall i,j\ \mathbf{C}_i \cap \mathbf{C}_j = \emptyset \tag{4.2}$$

Zwei Punkte $\mathbf{p}_i$ und $\mathbf{p}_j$ sind ähnlich und gehören genau dann zum selben Cluster, wenn sie folgendes Homogenitätskriterium erfüllen, d.h. h=1.

$$h\left(\mathbf{p}_i, \mathbf{p}_j, d_{dist}, d_{col}\right) = \begin{cases} 1, \text{falls } \|\mathbf{p}_i - \mathbf{p}_j\| < d_{dist} \wedge \left\|\mathbf{p}_i^{col} - \mathbf{p}_j^{col}\right\| < d_{col} \\ 0, \text{ sonst} \end{cases} \tag{4.3}$$

mit d_{dist} als euklidischer Abstand, und Farbähnlichkeitsmaß d_{col}.

Die Schwelle d_{dist} repräsentiert den maximal gültigen Abstand zweier Punkte innerhalb eines Clusters und wird abhängig vom Skalierungsfaktor der Punktkoordinaten festgelegt. Die Schwelle der Farbähnlichkeit d_{col} wird in Abhängigkeit der spektralen Abbildungseigenschaften des Aufnahmesystems festgelegt. Zur Definition eines Farbähnlichkeitsmaßes können Hautfarbmodelle oder im einfachsten Fall auch der Abstand im RGB-Farbraum dienen. Die komplette Zerlegung einer

Punktwolke liefert stets ein eindeutiges Ergebnis und wird durch den Algorithmus in Anhang 8.3 formal beschrieben.
Eine Erweiterung des Homogenitätskriteriums ist möglich, hat sich bisher aber nicht als notwendig erwiesen. Der beschriebene Algorithmus verwendet eine Suchfunktion, welche alle Punkte ermittelt, die dem verwendeten Homogenitätskriterium genügen. Dies lässt sich effizient durch BSP- (Binary Space Partitioning) bzw. kd-Bäume realisieren, wodurch die Laufzeit des gesamten Algorithmus für n Punkte reduziert wird auf $O(n \log n)$ [Kle05, Fuc80].
Durch das beschriebene Verfahren kann die Tiefeninformation effektiv genutzt werden, um eine Vorsegmentierung der Kopfregion durchzuführen und Hintergrundbereiche zu eliminieren. Dazu wird für jedes Cluster $\mathbf{C}_i$ eine Merkmalsmenge $\mathbf{M}_i$ aufgestellt (4.4) und bezüglich Funktion (4.10) ausgewertet.

$$\mathbf{M}_i = \left\{ \bar{z}_i, \bar{b}_i, n_i \right\}, \bar{z}_i, \bar{b}_i \in [0,1], n_i \in \mathbb{N}, \tag{4.4}$$

mit $\bar{z}_i$ als Zentriertheit, normierter Blauanteil $\bar{b}$ und Punktanzahl n_i.

Der Merkmalswahl liegen verschiedene Annahmen zugrunde. So ist die Wahrscheinlichkeit, dass ein Punktcluster $\mathbf{C}_i$ mit Schwerpunkt $\tilde{\mathbf{c}}_i$ zum Gesicht gehört höher, wenn sich dessen Projektion $k(\tilde{\mathbf{c}}_i)$ im Zentrum des Bildes befindet, was bereits bei der Kameraeinstellung sichergestellt wird. Zu diesem Zweck wird der normierte Abstand $\bar{z}_i$ (4.5) als Maß für die Zentriertheit verwendet. Der Abstand $\bar{z}_i$ wird durch Division durch den maximal möglichen Abstand d_{max} berechnet, der sich aus der Bilddimension ableitet (s. Abbildung 4-1(a)).

$$\bar{z}_i = \frac{1}{d_{max}} \left\| \tilde{\mathbf{i}} - k(\tilde{\mathbf{c}}_i) \right\|, \text{ Projektion } k \text{ entsprechend (3.5)} \tag{4.5}$$

mit

$$\tilde{\mathbf{i}} = \begin{bmatrix} w/2 \\ h/2 \end{bmatrix}, \; w, h \text{ als Breite und Höhe des Bildes} \tag{4.6}$$

und Schwerpunkt

$$\tilde{\mathbf{c}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{p}_j, \quad \mathbf{p}_j \in \mathbb{R}^3 \tag{4.7}$$

Weiterhin kann unter normalen Aufnahmebedingungen davon ausgegangen werden, dass das Gesicht eine vom Hintergrund hervorhebende Farbgebung aufweist. Die Hautfarbe des menschlichen Gesichts besitzt bei achromatischer Beleuchtung nur einen geringen Anteil im Blaukanal des RGB-Farbraumes [Nie05]. Hierzu

wird die gemittelte Farbe $\bar{F} = (\bar{R}, \bar{G}, \bar{B})$ ausgewertet, die sich aus der Summe aller n Farbwerte, der dem Cluster zugehörigen Punkte $\mathbf{p}_j$ errechnet (4.8).

$$\bar{F}_{RGB_i} = \left(\bar{R}, \bar{G}, \bar{B}\right) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{p}_{j_{rgb}} \tag{4.8}$$

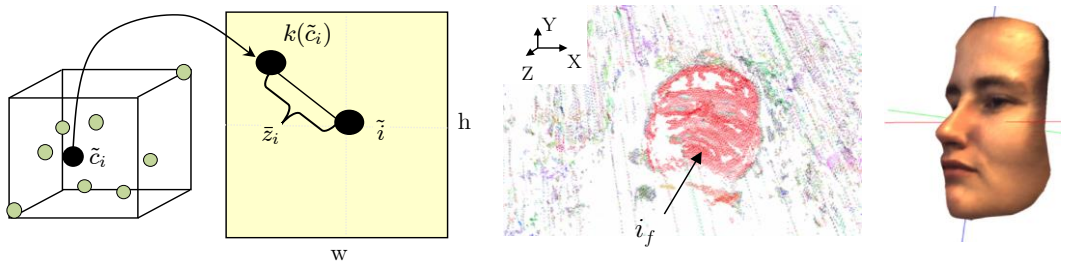
Die Intensität stellt eine Komponente des 3D-Farbraums dar. Durch Multiplikation des Farbvektors mit einem Skalar ändert sich diese. Durch die Farbnormierung über die Intensität werden alle Farben des 3D-Farbraums auf eine 2D-Farbebene reduziert. Um eine größere Invarianz bezüglich Intensitätsschwankungen zu erreichen, wird der normierte Blauanteil $\bar{b}$ in der Farbebene ausgewertet (4.9) [Jae05].

$$\bar{b} = \bar{B}/(\bar{R} + \bar{G} + \bar{B}), \bar{R}, \bar{G}, \bar{B} \in [0, 1] \tag{4.9}$$

In Testreihen mit hellhäutigen Personen mit verschiedenem Teint wurde für Punktcluster des Gesichts stets ein normierter Blauanteil von $\bar{b} < 0.25, \bar{b} \in [0, 1]$ ermittelt. Als drittes Merkmal eines Clusters wird die Punktanzahl n_i herangezogen. Mit Hilfe von Funktion $f_{fc}(i)$ (4.10) wird für jedes Cluster i der Gütewert dafür berechnet, das es das Gesicht repräsentiert. Zur Erfassung des Gesichts und dessen Rekonstruktion wird das Cluster i_f (4.11) aus der Gesamtmenge gewählt, für das die Funktion f_{fc} den kleinsten Betrag aufweist (z.B. Abbildung 4-1(b-c)). In empirischen Tests hat sich gezeigt, dass eine Gleichgewichtung (w_1=w_2=1) der gewählten Merkmale zu einer robusten Erfassung des Gesichts führt. Dabei hat sich weiterhin gezeigt, dass die Aufgabe mit weniger Merkmalen nicht zuverlässig lösbar ist.

$$f_{fc}(i) = (w_1 \cdot \bar{z}_i + w_2 \cdot \bar{b}_i)/n_i \in \mathbb{R}, \quad i, n_i \in \mathbb{N}, \ \bar{z}_i, \bar{b}_i, w_i \in \mathbb{R} \tag{4.10}$$

$$i_f = \arg\min_{i \in \mathbb{N}} f_{fc}(i) \tag{4.11}$$



(a) (b) (c)

Abbildung 4-1: Clusterverfahren zur 3D Gesichtsdetektion, (a) Zentriertheit $\bar{z}_i$, (b) Detektion des Clusters i_f, welches das Gesicht repräsentiert, (c) Rekonstruktion.

4.2.2 Modell-Rekonstruktion

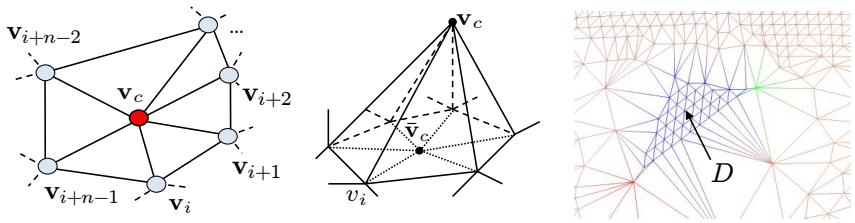
Ausgehend von den durch aktive bzw. passive Stereomessverfahren ermittelten und nachfolgend partitionierten 3D-Punktclustern, wurde ein automatisches Verfahren realisiert, welches mittels Delaunay-Triangulation [ORo98] und entsprechender Nachverarbeitung eine Dreiecksnetzrepräsentation der Gesichtsoberfläche erstellt. Hierbei wird das zuvor detektierte Gesicht in Sichtrichtung der Kameras trianguliert (s. Abbildung 4-2). Das somit resultierende, durch Gleichung (4.12) definierte Dreiecksnetz $\mathbf{S}$ wird durch eine Menge von n Vertices $\mathbf{v}_\mathrm{i}$ (triangulierte Messpunkte) und eine Liste von m Indices w_j repräsentiert, welche alle Dreiecke durch die entsprechenden Vertices bezeichnet. Dreiecksnetze sind aus der Computergraphik bekannt [Fol95].

$$\mathbf{S} = \{\mathbf{v}_1,\ ...,\ \mathbf{v}_n,\ w_1,\ ...,\ w_m\}, \mathbf{v}_i \in \mathbb{R}^3,\ w_j \in \mathbb{N} \tag{4.12}$$

Im Falle der passiven Stereodatenerfassung treten naturgemäß Messfehler auf. Dies äußert sich zum einen in Störungen der Oberfläche und zugehöriger Normalen, zum anderen in dünn besetzten Regionen des Dreiecksnetzes. Zum Ersetzen gestörter Vertices, wird die Nachbarschaftsinformation des Dreiecksnetzes genutzt. Berechnet man für jeden Vertex $\mathbf{v}_c$ des Dreiecksnetzes den Mittelwert $\bar{\mathbf{v}}_c$ aus allen n Nachbarn $\mathbf{v}_i$, so zeigen lediglich Ausreißer-Vertices eine deutliche Verschiebung $\|\mathbf{v}_c - \bar{\mathbf{v}}_c\| > t_{\max}$ und lassen sich somit leicht austauschen (Abbildung 4-3(a-b)).

(a) (b) (c) (d)

Abbildung 4-2: Modell Rekonstruktion, (a) gemessene Punktwolke, (b) Punktcluster des Gesichts, (c) Triangulation, (d) Modell mit Nachverarbeitung.



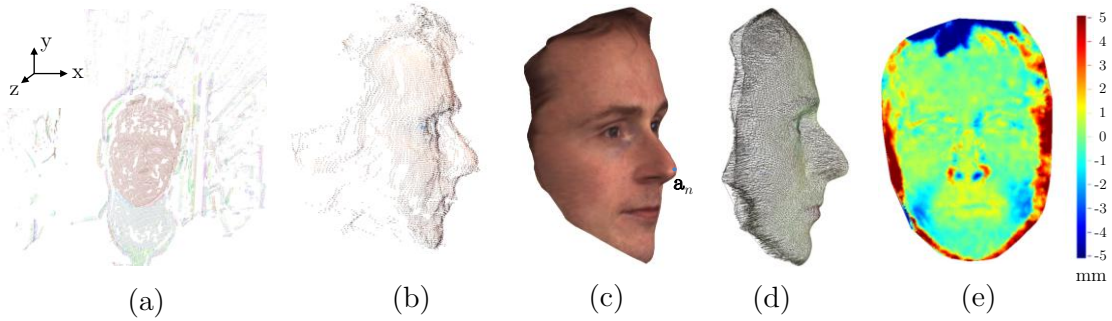
(a) (b) (c)

Abbildung 4-3: Nachverarbeitung, (a) Vertex Nachbarschaft, (b) Ausreißer $\mathbf{v}_c$ und Korrektur $\bar{\mathbf{v}}_c$, (c) Punktverdichtung für große Dreiecke D.

Zum Auffüllen dünn besetzter Stellen des Dreiecksnetzes, d.h. großen Dreiecken, existieren eine Reihe von Verfahren aus der Computergrafik, welche die umgebende Geometrie berücksichtigen, jedoch meist komplex und rechenintensiv sind. Da im vorliegenden Fall lediglich eine Approximation für die weitere Berechnung gefordert ist, wird hier folgende Problemlösung favorisiert. Für jedes Dreieck D, welches eine Fläche $a > a_{\max}$ aufweist, werden Vertices mit einem gleichmäßigen Abstand d_t innerhalb der Ebene des Dreiecks eingefügt (Abbildung 4-3(c)) wodurch eine dichte Oberflächenbeschreibung realisiert wird. In den durchgeführten Untersuchungen wurde eine durchschnittliche Anzahl von circa 1000 Dreiecken für ein 3D Modell verwendet.

Die Genauigkeit des rekonstruierten 3D Modells auf der Grundlage passiver Stereomessung mittels MAD-Korrespondenzbestimmung [Kle96], wurde durch einen Vergleich mit einem exakten Phasenshift-Verfahren ermittelt, welches durch die Projektion einer Sequenz von Lichtmustern eine hohe Auflösung erreicht [Alb98]. Hierzu wurde der Versatz zwischen den beiden Flächen bestimmt, nachdem diese durch einen ICP (Iterative Closest Point) Algorithmus bezüglich einer kleinsten-Fehlerquadrate-Metrik bestmöglich aneinander angenähert wurden (s. Abschnitt 5.1.1).

Bei der Prüfung der Genauigkeit wurden innerhalb des Gesichts, d.h. in Regionen wo eine Extraktion von Merkmalen stattfindet, Abweichungen gemessen, die stets unterhalb einer Schwelle von ± 2 mm lagen (Abbildung 4-4). In den Randbereichen ist diese Abweichung hingegen deutlich größer, was auf das sogenannte "foreshortening problem" zurückzuführen ist (verringerte Auflösung an Objekträndern). Dies verschlechtert wiederum die Messgenauigkeit, was jedoch für die Bestimmung relevanter Merkmale unerheblich ist, da diese nicht an den Rändern ermittelt werden.



(a) (b) (c) (d) (e)

Abbildung 4-4: Beispiel zur Genauigkeit bei der Rekonstruktion mittels passiver Stereomessung, (a) Zerlegung der Punktwolke, (b) Cluster des Gesichts, (c) Rekonstruktion $\mathbf{S}$ mit detektierter Nasenspitze $\mathbf{a}_n$, (d) genaue aktive Stereo Messung mittels Musterprojektion, (e) zugehörige Versatzkarte.

Mit der Erstellung des 3D Modells erfolgt durch Auswertung der Form die Bestimmung der Koordinate des Punktes der Nasenspitze $\mathbf{a}_n$, welcher zur Schätzung der Pose verwendet wird (Abschnitt 5.2.1). Dabei wird $\mathbf{a}_n$ prinzipiell als der am weitesten außen liegende Punkt vom Mittelpunkt des Modells $\mathbf{S}$ erfasst.

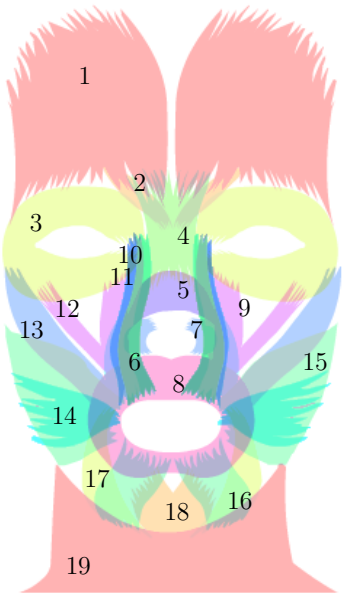
4.3 Physiologisch motivierte Regionen

Veränderungen der Mimik werden durch Muskelkontraktionen verursacht. Es gibt dabei eine Anzahl von 43 Muskeln, die grundlegend für die Entstehung der Gesichtsausdrücke verantwortlich sind [Ekm02]. Zu deren Erfassung schlug der Psychologe Paul Ekman das Facial Action Coding System (FACS) vor, mit dem es möglich ist, jeden nur denkbaren Gesichtsausdruck zu beschreiben. Da Mimik meist aus einer Kombination von verschiedenen Muskelaktivierungen resultiert, wird als Maßeinheit für das FACS nicht die Aktivität einzelner Muskeln selbst verwendet, sondern sogenannte Action Units (AUs). Insbesondere legen diese Einheiten Kontraktionen und Relaxationen von Gesichtsmuskeln bzw. Muskelgruppen fest.
Speziell ausgebildete FACS Coder sind in der Lage in zeitintensiver Arbeit Gesichtsausdrücke in einzelne Action Units manuell zu „zerlegen“. Auch wenn hier schon beachtliche Fortschritte erzielt wurden, sind automatische Systeme zur akkuraten Erkennung aller 64 AUs dem Menschen noch immer klar unterlegen. Generell kann bereits eine Untermenge an AUs eine geeignete Grundlage zur Klassifikation von Mimik mittels maschineller Lernverfahren darstellen.

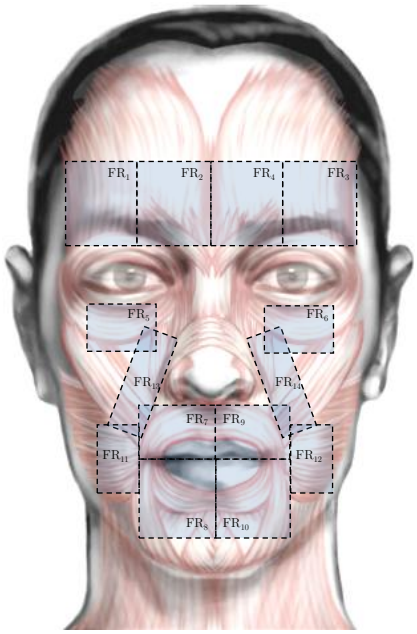
Da die explizite Bestimmung der Action Units einen erheblichen Mehraufwand bedeutet, wurde im Rahmen dieser Arbeit mit der Einführung physiologisch motivierter Regionen, welche sich ebenfalls an der Gesichtsmuskulatur orientieren und der Erkennung zugehöriger Bewegungsmerkmale dienen, ein anderer Weg beschritten.

Basierend auf empirischen Untersuchungen wurden dabei zur Erkennung von sechs verschiedenen Klassen emotional expressiver Mimik eine Menge $\mathbf{M}_{fr}$ (4.13) von 14 Regionen im Gesicht ermittelt, die im Folgenden auch als Flussregionen (FR) bezeichnet werden (Abbildung 4-5). Die Regionen werden hierzu auf der Grundlage von vier Ankerpunkten $\mathbf{a}_j \in \mathbb{R}^3$ durch Verhältnismaße definiert, d.h. linkes und rechtes Auge $\mathbf{a}_{le}$, $\mathbf{a}_{re}$ sowie Mundwinkel $\mathbf{a}_{lm}$, $\mathbf{a}_{rm}$. Flussregionen werden dem Gesichtsmodell $\mathbf{S}$ als begrenzende dreidimensionale Konturpunkte zugeordnet. Weiterhin werden für jede Region i sogenannte Bewegungsabtastpunkte $\mathbf{p}_{k,i} \in \mathbb{R}^3$ entlang eines Rasters auf dem Modell definiert (Abbildung 4-6(a)). Bewegungsabtastpunkte dienen der Erfassung dynamischer Merkmale.

$$\mathbf{M}_{fr} = \{FR_1, ..., FR_{14}\}, FR_i \in \left\{\mathbf{p}_1, ..., \mathbf{p}_{n_i}\right\}, \mathbf{p}_j \in \mathbb{R}^3 \tag{4.13}$$



1 Frontalis
2 Corrugator
3 Orbicularis Oculi
4 Procerus
5 Nasalis Transversa
6 Nasalis Alaris
7 Nose Expander
8 Orbicularis Oris
9 Caninus
10 Quadratus Labii Sup.
11 Own Elevator
12 Zygomaticus Min.
13 Zygomaticus Maj.
14 Risorius
15 Buccinatorius
16 Quadratus Labii Inf.
17 Triangularis
18 Mentalis
19 Platysma



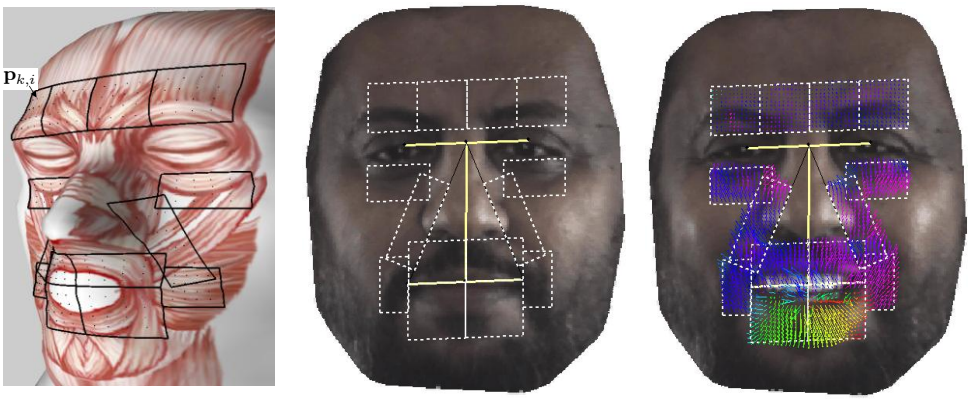
(a) (b)

Abbildung 4-5: Physiologisch motivierte Regionen, (a) vereinfachtes schematisches Modell der Gesichtsmuskulatur, Quelle [Flo10], (b) an der Muskulatur orientierte Definition sogenannter Flussregionen.

Flussregionen lassen sich somit für die beiden im Rahmen dieser Arbeit untersuchten Ansätze, d.h. Gesichtsnormierung (Abschnitt 5.1, Beispiel in Abbildung 4-6(b-c)) und Merkmalsnormierung (Abschnitt 5.2) zur Detektion dynamischer Merkmale sehr effektiv einsetzen. Die Verwendung von Flussregionen hat verschiedene Vorteile. Zum einen realisieren sie eine sinnvolle Reduzierung der Merkmalsdaten, d.h. der erfassten Bewegungsinformation. Eine Auswertung des Verschiebungsvektorfeldes für das gesamte Gesicht brächte hier eine Vielzahl von Mehrdeutigkeiten. Des Weiteren wird durch diese Reduzierung eine Weiterverarbeitung zur Merkmalsnormierung überhaupt erst ermöglicht (Abschnitt 5.2).

4.3.1 Allgemeine Definition der dynamischen Merkmale

Grundsätzlich gibt jedes dynamische Merkmal Auskunft über die Bewegung, die aufgrund von Mimikänderungen in einer physiologisch motivierten Region i zu verzeichnen ist. Dabei wird die Bewegung bildbasiert durch eine Menge von Verschiebungsvektoren (VV) erfasst. Die VV werden auf der Grundlage des pyramidalen Optischen Fluss Verfahrens nach Lucas-Kanade für jede der vierzehn Regionen berechnet [Luc81] (Abschnitt 3.3.1). Die Bestimmung der VV erfolgt nicht für alle Bildpunkte, sondern nur an definierten Stellen, welche durch die Bewegungsabtastpunkte des Gesichtsmodells festgelegt werden. Diese Punkte entspre-



(a) (b) (c)

Abbildung 4-6: Physiologisch motivierte Regionen II, (a) Assoziation der Flussregionen mit dem 3D Modell; jede Region i beinhaltet sogenannte Bewegungsabtastpunkte $\mathbf{p}_{k,i} \in \mathbb{R}^3$, (b) Projektion der Regionen auf das Bild (am Beispiel der Gesichtsnormierung), (c) Detektion des optischen Flusses (farbkodierte Bewegungsrichtung) auf der Grundlage physiologisch motivierter Regionen.

chen Kreuzungspunkten eines Rasters, dessen Zeilen- bzw. Spaltenabstand w_g abhängig von der Bildauflösung festgelegt werden (Abbildung 4-7).
Der in experimentellen Untersuchungen verwendete Abstand beträgt w_g=4 Pixel. Die Verwendung des Rasters führt zur Reduzierung einer Vielzahl sehr ähnlicher Vektoren und vermindert den Rechenaufwand deutlich. Für jede Region i wird somit zu einem gegebenen Zeitpunkt t eine Anzahl von m_i Vektoren $\mathbf{v}_{j,i}^t \in \mathbb{R}^2$ berechnet. Im Sinne einer zeitlichen Glättung erfolgt zum Verstärken der gemessenen Bewegung, bei gleichzeitiger Unterdrückung von Ausreißern, die Akkumulation $\tilde{\mathbf{v}}_{j,i}^t \in \mathbb{R}^2$ (4.14), d.h. Aufsummierung von n_{acc} Vektoren vorhergehender Zeitschritte (Abbildung 4-7(e)). Die Anzahl der Akkumulationen hängt dabei von der Bildfrequenz des Aufnahmesystems ab. In Untersuchungen hat sich n_{acc}=5 bei einer Bildrate von 25 Hz empirisch als zweckmäßig erwiesen.

$$\tilde{\mathbf{v}}_{j,i}^t = \frac{1}{n_{acc}} \sum_{k=1}^{n_{acc}} \mathbf{v}_{j,i}^{t-k+1},\ \mathbf{v}_{j,i} \in \mathbb{R}^2,\ t > n_{acc}\,, \tag{4.14}$$

mit n_{acc} als Anzahl der Akkumulationen.

Zur Reduktion des Datenaufkommens bei gleichzeitiger Erhöhung der Kompaktheit akkumulierter VV wird für jede Region i der gemittelte Vektor $\bar{\mathbf{v}}_i^t \in \mathbb{R}^2$ (4.15) bestimmt. Die dynamischen Merkmale werden generell durch die Menge der gemittelten Vektoren $\bar{\mathbf{v}}_i^t$ für alle n_i=14 Regionen zu einem bestimmten Zeitpunkt t repräsentiert (Abbildung 4-7(e)).

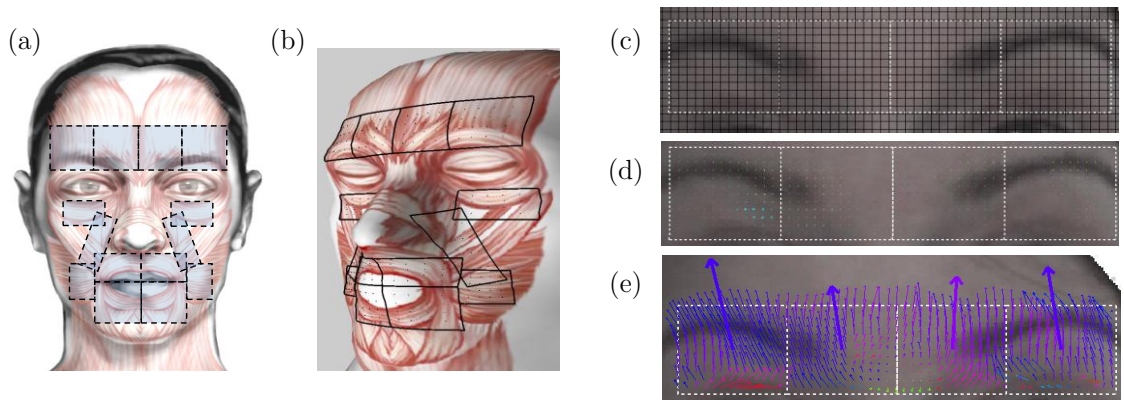


Abbildung 4-7: Ermittlung von Verschiebungsvektoren, (a, b) Flussregionen und Verknüpfung mit 3D Modell, (c) schematische Darstellung des zugrundeliegenden Rasters, (d) rastergestützte Berechnung des optischen Flusses und (e) Akkumulation. Die fett dargestellten Vektoren entsprechen den gemittelten Vektoren $\bar{\mathbf{v}}_i^t$ (4.15).

$$\bar{\mathbf{v}}_i^t = \frac{1}{m_i} \sum_{j=0}^{m_i} \tilde{\mathbf{v}}_{j,i}^t, \quad \bar{\mathbf{v}}_i^t, \tilde{\mathbf{v}}_{j,i}^t \in \mathbb{R}^2, \tag{4.15}$$

mit m_i als Anzahl der VV einer Region i.

Eine Nutzung dynamischer Merkmale zur Erkennung setzt eine minimale Aktivierung mimischer Bewegung, d.h. messbare Stärke der VV in den Flussregionen voraus. Nur wenn die Aktivierungsfunktion $v_{sum}(t)$ (4.16) als gemittelte Vektornorm über alle n_i=14 Regionen eine Schwelle v_{min} überschreitet, kann eine Klassifikation durchgeführt werden (Abbildung 4-8). Grundsätzlich ist außerhalb dieser Aktivierungsphase, wegen fehlender Merkmale, keine zuverlässige Erkennung auf der Grundlage dynamischer Merkmale möglich.

$$v_{sum}(t) = \sum_{i=1}^{n_i} \|\bar{\mathbf{v}}_i^t\|, \ \bar{\mathbf{v}}_i^t \in \mathbb{R}^2, \ \ n_i{=}14 \tag{4.16}$$

Während entsprechend (4.16) der Betrag der Verschiebungsvektoren über die Durchführbarkeit einer Klassifikation entscheidet, wird die Vektorrichtung als Informationsträger über die Tendenz des Gesichtsausdrucks zur eigentlichen Erkennung verwendet und zu diesem Zweck für alle n_i=14 Regionen im Merkmalsvektor für dynamische Merkmale $\mathbf{f}_{dyn}^t$ (4.17) zusammengefasst.

$$\mathbf{f}_{dyn}^t = (\measuredangle\bar{\mathbf{v}}_1^t \ \ldots \ \measuredangle\bar{\mathbf{v}}_{14}^t)^{\mathrm{T}}, \ \measuredangle\bar{\mathbf{v}}_i^t \in \mathbb{R} \tag{4.17}$$

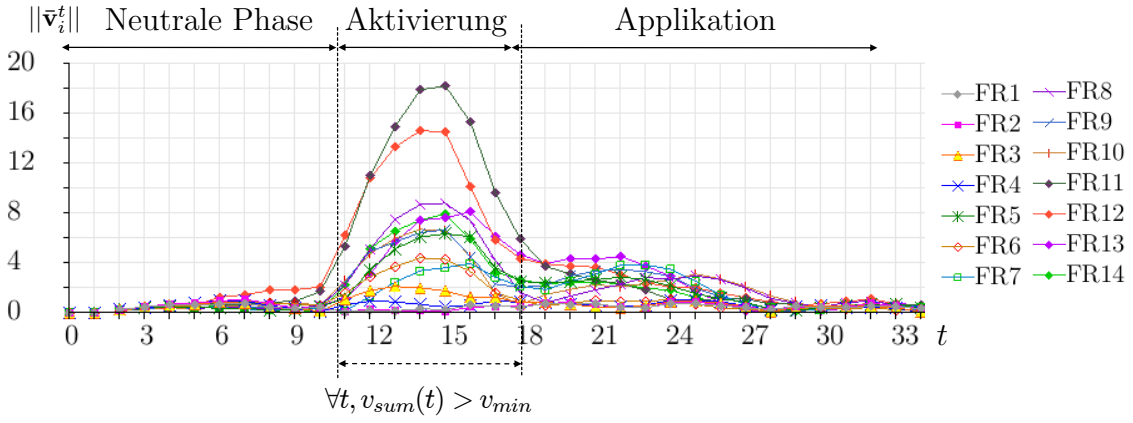
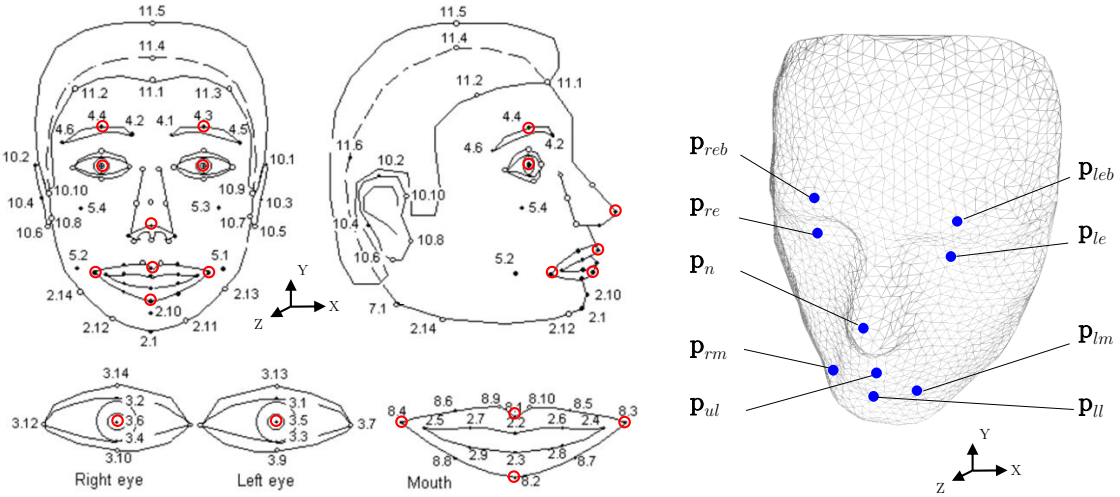


Abbildung 4-8: Phasen bei der Erfassung dynamischer Merkmale mit Überschreitung der Schwelle v_{min} in der Aktivierungsphase.

4.4 Mimikrelevante Merkmalspunkte

Bei der Gesichts- und Mimikerkennung ist die Verwendung von aussagekräftigen Punkten, etwa aus der Mund und Augenregion zur Merkmalsextraktion naheliegend und weit verbreitet. Im Allgemeinen dienen solche Punkte der Bestimmung von Abständen und zur Gesichtsnormierung [Soy07] bzw. werden zur Anpassung von Modellen verwendet [LiS05]. Die Besonderheit in dieser Arbeit liegt in der speziellen Verarbeitung der Merkmalspunkte mittels modellbasierter 3D Transformationen und Merkmalsnormierung, was zu einer neuartigen und qualitativ hochwertigen Erkennung führt.
Inspiriert wird die Nutzung einer Reihe charakteristischer 3D Merkmalspunkte durch das Face Animation Parameter (FAP) System, welches im Rahmen des MPEG-4 Standards zum Zweck der Animation von Gesichtern definiert wurde und auf der Variation sogenannter Face Definition Parameters (FDPs) beruht [Pan02, ISO01]. Diese beschreiben die Gesichtsform und aktuelle Mimik und basieren auf einer Menge von 88 Merkmalspunkten (Abbildung 4-9(a)).
Die im Rahmen dieser Arbeit durchgeführten Untersuchungen haben gezeigt, dass zum Zweck der Mimikerkennung, d.h. sechs Klassen expressiver Mimik sowie eine Klasse neutral, bereits eine Untermenge $\mathbf{P}_{fp}$ (4.18) von neun 3D Punkten des FDP Satzes (Abbildung 4-9(b)) und der darauf basierenden Definition geometrischer



(a) (b)

Abbildung 4-9: 3D-Merkmalspunkte, (a) MPEG4 Standard [ISO01] zur Beschreibung von Mimik und Gesichtsform sowie in Rot, die im vorgeschlagenen Konzept benutzte Untermenge $\mathbf{P}_{fp}$, (b) Gesichtsmodell $\mathbf{S}$ mit entsprechenden Merkmalspunkten.

Merkmale (Abschnitt 4.4.2) zu hervorragenden Ergebnissen führt.

$$\mathbf{P}_{fp} = \{\mathbf{p}_{le}, \mathbf{p}_{re}, \mathbf{p}_{leb}, \mathbf{p}_{reb}, \mathbf{p}_{lm}, \mathbf{p}_{rm}, \mathbf{p}_{ul}, \mathbf{p}_{ll}, \mathbf{p}_{n}\},\ \mathbf{p}_j \in \mathbb{R}^3, \tag{4.18}$$

mit linkem und rechtem Mittelpunkt der Augen $\mathbf{p}_{le}$, $\mathbf{p}_{re}$, Augenbrauen $\mathbf{p}_{leb}$, $\mathbf{p}_{reb}$ und Mundwinkel $\mathbf{p}_{lm}$, $\mathbf{p}_{rm}$, Mundober- und Unterkante $\mathbf{p}_{ul}$, $\mathbf{p}_{ll}$ und Nasenspitze $\mathbf{p}_{n}$.

Die Nasenspitze $\mathbf{p}_n$ wird ausschließlich zur Posebestimmung verwendet (Abschnitt 5.2.1). Die Bestimmung der Merkmale auf der Grundlage der 3D Merkmalspunkte $\mathbf{P}_{fp}$ erfolgt modellbasiert und setzt eine Detektion korrespondierender Merkmalspunkte im Bild voraus. Dies wird mittels Bildverarbeitungsmethoden realisiert.

4.4.1 Merkmalspunktdetektion im Bild

Im vorgeschlagenen Konzept zur Mimikerkennung wird eine Reihe von Techniken aus Bildverarbeitung und Mustererkennung zur Merkmals- und dabei insbesondere Merkmalspunktdetektion eingesetzt. Dabei nutzen diese zum Teil modellbasiert, Gradienten- und Farbinformation aus, mit dem Ziel die Bildkoordinaten $\mathbf{I}_{fp}$ (4.19) der entsprechenden Merkmalspunkte $\mathbf{P}_{fp}$ (4.18) in einem möglichst breiten Spektrum von Aufnahmesituationen detektieren zu können.

$$\mathbf{I}_{fp} = \{\mathbf{i}_{le}, \mathbf{i}_{re}, \mathbf{i}_{leb}, \mathbf{i}_{reb}, \mathbf{i}_{lm}, \mathbf{i}_{rm}, \mathbf{i}_{ul}, \mathbf{i}_{ll}, \mathbf{i}_{n}\},\ \mathbf{i}_j \in \mathbb{R}^2 \tag{4.19}$$

mit $\mathbf{i}_j$ analog zu $\mathbf{p}_j \in \mathbf{P}_{fp}$ (4.18).

Mit Ausnahme des Punktes $\mathbf{i}_n$ (Nasenspitze), welcher nicht zuverlässig durch Bildverarbeitungsmethoden bestimmt werden kann und somit eine gesonderte Prozedur mit zusätzlichem Kontextwissen erfordert (s. Abschnitt 5.2.1, Poseschätzung), werden alle anderen Punkte $\mathbf{I}_{fp}$ direkt erfasst.
Insbesondere wird für diese Aufgabe der von Viola und Jones vorgestellte Haar-like Feature basierte Detektor eingesetzt, der durch einen AdaBoost Algorithmus eine effiziente Detektion realisiert (s. Abschnitt 2.3.1, Abbildung 4-10).

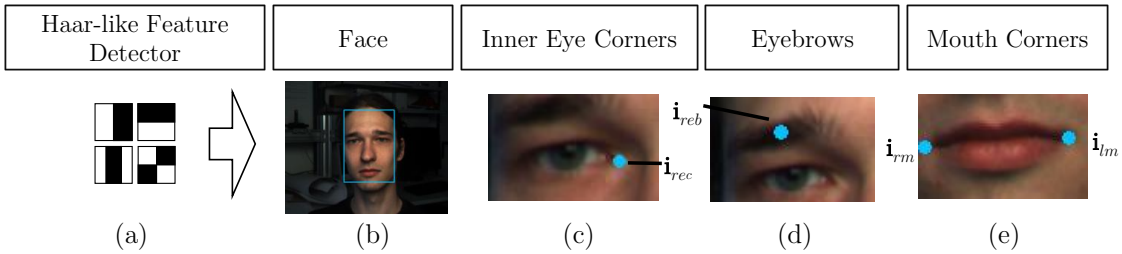


Abbildung 4-10: Detektion von Merkmalspunkten, (a) Haar-like features, (b) Gesichtsdetektion, (c) rechter innerer Augenpunkt $\mathbf{i}_{rec}$, (d) rechter Augenbrauenpunkt $\mathbf{i}_{reb}$, (e) linker und rechter Mundwinkel $\mathbf{i}_{lm}$ und $\mathbf{i}_{rm}$.

Durch den Gesichtsdetektor wird ein begrenzendes Rechteck für das Gesicht ermittelt und der Suchbereich für die verschiedenen Merkmalspunkte eingeschränkt. Analog zu [Pan08] wurden bei der Haar-like feature basierten Detektion folgende Detektorkaskaden für verschiedene Aufgaben eingesetzt:

- Gesichtsdetektion
- Merkmalspunktdetektion, Innenseite linkes und rechtes Auges ($\mathbf{i}_{lec}$, $\mathbf{i}_{rec}$)
- Merkmalspunktdetektion, Augenbrauen links und rechts ($\mathbf{i}_{leb}$, $\mathbf{i}_{reb}$)
- Merkmalspunktdetektion, Mundwinkel links und rechts ($\mathbf{i}_{le}$, $\mathbf{i}_{re}$)

Fehlerhafte Detektionen werden in einem Validierungsschritt erkannt und korrigiert, in dem die berechneten Punktkandidaten mit Schätzungen durch ein generisches Gesichtsmodell verglichen werden, welches sich im begrenzenden Rechteck des Gesichtsdetektors befindet.
Im Anschluss werden die Augenmittelpunkte $\mathbf{i}_{le}$, $\mathbf{i}_{re}$ aus den zuvor detektierten Punkten an der Augeninnenseite $\mathbf{i}_{lec}$, $\mathbf{i}_{rec}$ ermittelt. Die Mittelpunkte werden durch eine zur Gesichtsgröße proportionale Erweiterung der Achse zwischen beiden inneren Punkten bestimmt (Abbildung 4-11). Auf diese Weise können auch geschlossene Augen, beispielsweise durch Blinzeln, erfasst werden.
Die Detektion des Mundes mit entsprechenden Merkmalspunkten ist farbbasiert. Hierbei werden die unterschiedlichen Farbcharakteristika von Haut und Lippen zur Separierung genutzt. Für hellhäutige Personen bietet sich zur Unterscheidung eine Auswertung des normalisierten Grünkanals entsprechend (4.20) an, welcher eine klare Trennung ermöglicht [Jae05]. Bei sehr dunkler Hautfarbe sind bei der farbbasierten Unterscheidung von Haut und Lippen jedoch individuelle Adaptionen bei der Wahl der Farbraumtransformation erforderlich.

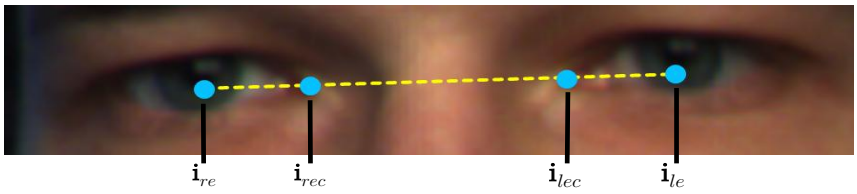


Abbildung 4-11: Beispielhafte Bestimmung der Augenmittelpunkte $\mathbf{i}_{le}$ und $\mathbf{i}_{re}$ auf der Grundlage der inneren Augenpunkte $\mathbf{i}_{lec}$ und $\mathbf{i}_{rec}$, welche durch den Haar-like feature basierten Punktdetektor erfasst werden.

$$g = G/(R+G+B),\ R, G,\ B \in RGB \text{ Farbraum} \tag{4.20}$$

Die zur Munddetektion untersuchte Bildregion (ROI - Region of Interest) wird durch die beiden Punkte der Mundwinkel festgelegt, die durch den Punktdetektor erfasst werden (Abbildung 4-12). Bei einer Betrachtung im normalisierten Grünkanal zeigt die Verteilung im Histogramm $\mathbf{H}$ der ROI eine Aufteilung in zwei Häufungsbereiche für die Lippen- und Hautpixel. Die Trennung der dem Histogramm zugeordneten Bildpunkte basiert in der verwendeten Methode auf einer Schwellwertfilterung. Die benutzte Schwelle t_{ng} entspricht dabei dem Minimum des Polynoms ξ, das mit Hilfe des Verfahrens der kleinesten Fehlerquadrate das Histogramm $\mathbf{H}$ approximiert (Abbildung 4-12(b)). Im Anschluss wird die ROI binarisiert und ein oder mehrere dem Mund zugehörige Blobs (binary large objects) werden erfasst. Morphologische Operationen wie Closing und konturbasierte Erosion [AlH03] werden auf diese Binärformen angewendet. Für den Fall eines geöffneten Mundes ist dies von Bedeutung, da die Zähne aufgrund anderer Farbcharakteristik nicht zum selben Teil des Histogramms wie die Lippen gehören.
Im nächsten Schritt wird mit $\mathbf{M}_c$ die Kontur der Binärmaske des Mundes ermittelt und für diese die sogenannte konvexe Hülle $ch(\mathbf{M}_c)$ als umschließender Polygonzug berechnet [Kle05] (Abbildung 4-12(c)). In der konvexen Hülle einer Punktmenge, liegt jede Verbindung zwischen zwei Punkten der Menge, innerhalb

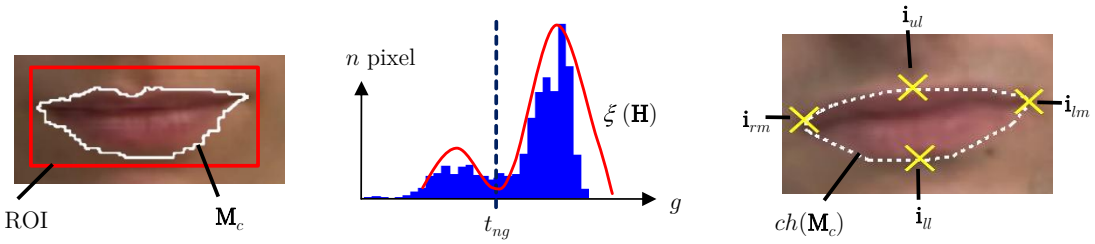


Abbildung 4-12: Merkmalspunktdetektion in der Mundregion, (a) Mund ROI mit Kontur $\mathbf{M}_c$, (b) ROI Histogramm $\mathbf{H}$ des normierten Grünkanals g und Fitting Polynom ξ mit Schwelle t_{ng}, (c) konvexe Hülle $ch(\mathbf{M}_c)$ der Mundkontur mit extrahierten Punkten.

des umschließenden konvexen Polygons. Zur effizienten Berechnung wird der Graham-Scan Algorithmus verwendet [ORo98].
Ausgehend von der konvexen Hülle werden die finalen Punkte der Mundwinkel $\mathbf{i}_{le}$, $\mathbf{i}_{re}$ und der oberen bzw. unteren Lippenkontur $\mathbf{i}_{ul}$, $\mathbf{i}_{ll}$ bestimmt. Hierzu werden parallel zur Augenachse die am weitesten außen bzw. in der Mitte der oberen und unteren Hälfte liegenden Punkte berechnet.

4.4.2 Allgemeine Definition geometrischer Merkmale

Bei der Auswertung von Gesichtsmerkmalen bietet der Übergang von 2D Bildmerkmalen zu einer 3D Merkmalsrepräsentation klare Vorteile. Während im Zweidimensionalen die aktuelle Orientierung eines untersuchten Objektes starken Einfluss auf die erfassten Bildmerkmale hat, etwa durch perspektivische Verkürzung bei Bewegung in Kamerarichtung und Rotation, so ermöglicht die 3D Merkmalsrepräsentation Unabhängigkeit bei Variationen der aktuellen Pose. Durch Auswertung der 3D Merkmalspunkte $\mathbf{P}_{fp}$ (4.18) wird diese Eigenschaft mit der Definition geometrischer Merkmale ausgenutzt. Die „Rohmerkmale“ werden in einem zehndimensionalen Merkmalsvektor $\mathbf{f}$ zusammengefasst, und stellen die Grundlage für die Normierung und Klassifikation dar (Abschnitt 5.2).

$$\mathbf{f} = (d_1 ... d_6 \; \alpha_1 ... \alpha_4)^{\mathrm{T}}, \; \mathbf{f} \in \mathbb{R}^{10}, \; d_i, \alpha_j \in \mathbb{R} \text{ nach (4.22)-(4.30)} \tag{4.21}$$

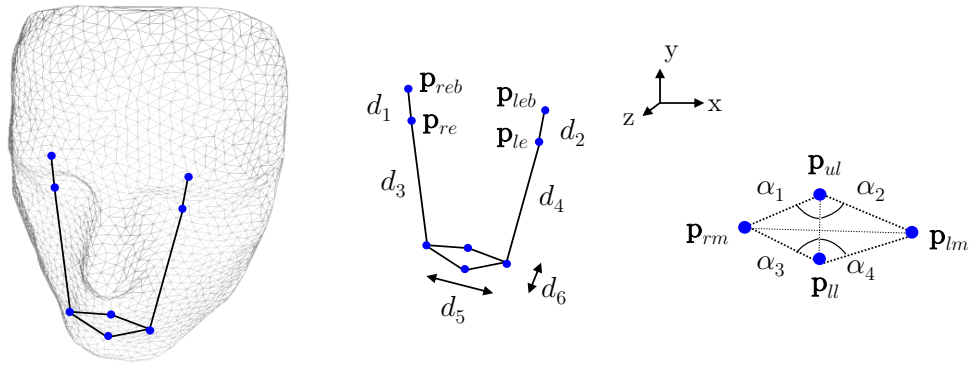
Diese Parameter umfassen sechs euklidische 3D Abstände d_i im Gesicht sowie vier Winkel α_j in der Mundregion (Abbildung 4-13), die in ihrer Gesamtheit charakteristisch für die verschiedenen Mimikklassen sind. Insbesondere wird durch die Abstände d_1 und d_2 das Anheben und Senken der Augenbrauen erfasst.
Die Abstände d_3 und d_4 zwischen den Mundwinkeln und Augenmittelpunkten detektieren Mundbewegungen. Des Weiteren wird durch d_5 und d_6 die Breite und Höhe des Mundes sowie die Winkel α_j als zusätzliche Formmerkmale bestimmt.

$$d_1 = \|\mathbf{p}_{reb} - \mathbf{p}_{re}\|, \tag{4.22}$$

$$d_2 = \|\mathbf{p}_{leb} - \mathbf{p}_{le}\|, \tag{4.23}$$

$$d_3 = \|\mathbf{p}_{re} - \mathbf{p}_{rm}\|, \tag{4.24}$$

$$d_4 = \|\mathbf{p}_{le} - \mathbf{p}_{lm}\|, \tag{4.25}$$



(a) (b) (c)

Abbildung 4-13: Definition geometrischer Merkmale, (a) Gesichtsmodell mit Merkmalspunkten $\mathbf{p}_i$ und (b) Abstände d_j sowie (c) Winkel α_k.

$$d_5 = \|\mathbf{p}_{rm} - \mathbf{p}_{lm}\|, \tag{4.26}$$

$$d_6 = \|\mathbf{p}_{ul} - \mathbf{p}_{ll}\|, \tag{4.27}$$

$$\alpha_1 = \arccos\left(\frac{\mathbf{v_1}\cdot\mathbf{v_2}}{\|\mathbf{v_1}\|\cdot\|\mathbf{v_2}\|}\right), \tag{4.28}$$

$$\alpha_2 = \arccos\left(\frac{\mathbf{v_2}\cdot\mathbf{v_3}}{\|\mathbf{v_2}\|\cdot\|\mathbf{v_3}\|}\right), \tag{4.29}$$

$$\alpha_3 = \arccos\left(\frac{-\mathbf{v_2}\cdot\mathbf{v_4}}{\|\mathbf{v_2}\|\cdot\|\mathbf{v_4}\|}\right), \tag{4.30}$$

$$\alpha_4 = \arccos\left(\frac{-\mathbf{v_2}\cdot\mathbf{v_5}}{\|\mathbf{v_2}\|\cdot\|\mathbf{v_5}\|}\right), \tag{4.31}$$

mit

$$\mathbf{v}_1 = \mathbf{p}_{rm} - \mathbf{p}_{ul},$$

$$\mathbf{v}_2 = \mathbf{p}_{ll} - \mathbf{p}_{ul},$$

$$\mathbf{v}_3 = \mathbf{p}_{lm} - \mathbf{p}_{ul},$$

$$\mathbf{v}_4 = \mathbf{p}_{rm} - \mathbf{p}_{ll},$$

$$\mathbf{v}_5 = \mathbf{p}_{lm} - \mathbf{p}_{ll}, \qquad \mathbf{v}_i,\ \mathbf{p}_j \in \mathbb{R}^3.$$

Die konkrete Berechnung der Abstände und Winkel erfordert zunächst die Ermittlung der 3D Merkmalspunkte $\mathbf{P}_{fp}$ sowie die Bestimmung der aktuellen Orien-

tierung des Gesichtsmodells und wird in Abschnitt 5.2.2 „Erfassung der geometrischen Merkmale“ erläutert.
Im folgenden Kapitel wird hierzu die konkrete Systemstruktur zur Mimikanalyse vorgestellt, für welche die in diesem Abschnitt beschriebene Definition dynamischer und geometrischer Merkmale als Grundlage dient. Insbesondere werden zwei verschiedene Herangehensweisen zur Merkmalsextraktion untersucht, zum einen die sogenannte Gesichtsnormierung (Abschnitt 5.1), bei der das Bild des Gesichts transformiert wird und zum anderen die Merkmalsnormierung, bei der lediglich erfasste Merkmale in eine andere Darstellung überführt werden (Abschnitt 5.2). Des Weiteren werden Wege zur Klassifikation beschrieben sowie eine Möglichkeit zur Integration geometrischer und dynamischer Merkmale im Sinne einer Fusion vorgeschlagen (Abschnitt 5.3).

Kapitel 5

Systemstruktur zur Mimikanalyse

In diesem Kapitel wird die vorgeschlagene Systemstruktur zur Mimikanalyse erläutert. Entsprechend Abbildung 5-1 wurden zur Realisierung der Merkmalsextraktion zwei Ansätze untersucht. Als erstes wurde die 3D gestützte Gesichtsnormierung betrachtet, bei der das Gesicht in eine einheitliche Frontaldarstellung transformiert wird, so dass die aktuelle Kopfpose keinen Einfluss mehr auf die Merkmalsextraktion hat. Dieses Verfahren basiert auf der Auswertung von Stereobildfolgen und erzielt somit eine verstärkte Robustheit durch Redundanz in den Beobachtungsdaten. Da die Gesichtsnormierung den ersten Meilenstein auf dem Weg zur Entwicklung der Systemstruktur zur Mimikanalyse darstellt, wurden zunächst nur dynamische Merkmale berücksichtigt.
Aufbauend auf den Ergebnissen wurde mit dem zweiten weniger rechenintensiven

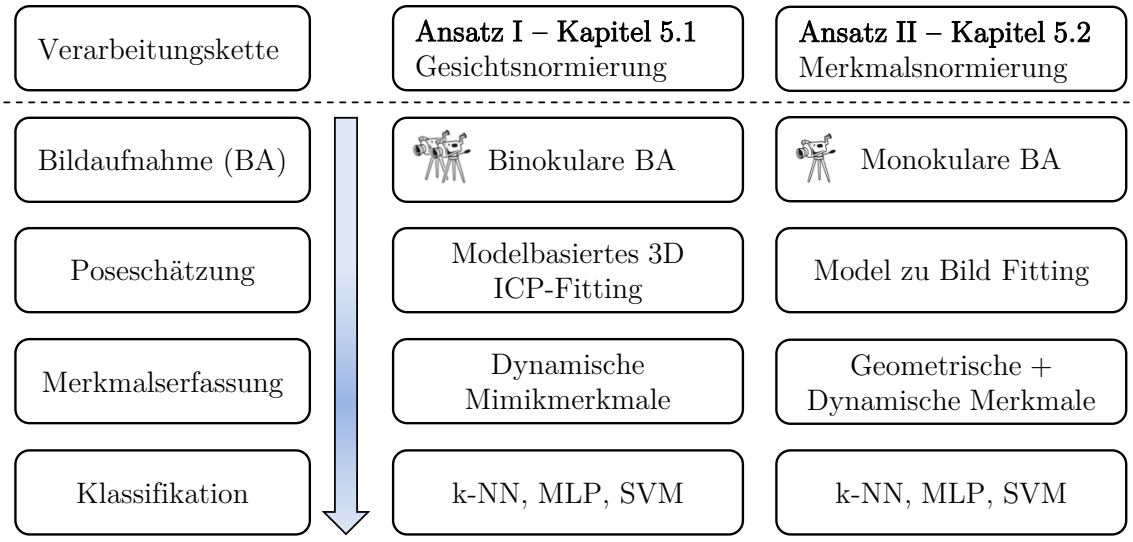


Abbildung 5-1: Globale Systemstruktur dieser Arbeit mit zwei untersuchten Ansätzen zur Mimikerkennung durch Auswertung dynamischer Merkmale im ersten Ansatz bzw. geometrischer und dynamischer Merkmale im zweiten Ansatz.

Ansatz mittels Monokamerasystem, der sogenannten Merkmalsnormierung, eine Weiterentwicklung vorgestellt, bei der nicht mehr das gesamte Gesicht transformiert wird, sondern nur die extrahierten Merkmale. Auf diese Weise wird eine hohe Performanz mit den Vorteilen der 3D Merkmalsextraktion verbunden. In der Folge dessen wurden umfangreiche Untersuchungen durchgeführt, bei denen die dynamischen Merkmale um geometrische Merkmale erweitert wurden.
Geometrische Merkmale lassen sich im Gegensatz zu den dynamischen Merkmalen nicht nur während einer Veränderung der Mimik erfassen, sondern jederzeit. Ein Weg zur integrierten Auswertung der beiden Merkmalsarten wird vorgeschlagen, durch den eine Verbesserung des Klassifikationsergebnisses erzielt werden kann.

5.1 Ansatz I – Mimikanalyse mittels Gesichtsnormierung

Die Idee der Gesichtsnormierung besteht darin, das gesamte Gesicht zur Merkmalsextraktion durch eine Reihe photogrammetrischer Techniken so zu transformieren, das es sich in einer Frontaldarstellung befindet. Es wird somit bezüglich der Pose normiert. Auf diese Weise wird die Merkmalsextraktion deutlich erleichtert, da nach der Normierung Änderungen im Bild nur noch auf Variationen der Mimik oder Beleuchtungssituation zurückzuführen sind. Zum anderen wird die Robustheit erhöht, indem Störungen durch Kopfbewegungen oder komplizierten Hintergrund eliminiert werden. In der untersuchten Methode erfolgt eine fortlaufende passive Stereomessung, durch die in jedem Zeitschritt eine Messpunktwolke $\mathbf{W}$ der Szene erfasst wird.

$$\mathbf{W} = \{\mathbf{p}_1, \ \dots \ , \mathbf{p}_n\}, \ \mathbf{p}_j \in \mathbb{R}^3 \tag{5.1}$$

Insbesondere beruht hierbei das Verfahren zur Stereoberechnung auf einem Korrelationsverfahren mittels MAD Funktion (Abschnitt 3.2). Nach Bestimmung der Orientierung des Kopfes durch Poseschätzung wird unter Nutzung des zuvor erstellten Gesichtsmodells $\mathbf{S}$ (4.12) eine Frontaldarstellung synthetisiert. Im Anschluss an die Gesichtsnormierung erfolgen die Detektion dynamischer Merkmale und die Klassifikation der Mimik.

Abbildung 5-2: Ansatz I – Mimikanalyse mittels Gesichtsnormierung.

5.1.1 Poseschätzung in binokularen Bildfolgen

Poseschätzung ist eine grundlegende Aufgabe im Fachgebiet Computer Vision. Generell wird die aktuelle Orientierung eines Modells durch einen Parametersatz beschrieben. Im Dreidimensionalen ist es daher normalerweise die Aufgabe, im Falle eines starren Modells drei Translations- und drei Rotationsparameter zu bestimmen, die im Folgenden als Zustandsvektor bzw. Posevektor $\mathbf{t}$ bezeichnet werden (5.2).

$$\mathbf{t}=(\mathrm{x\ y\ z}\ \omega\ \phi\ \kappa)^{\mathrm{T}},\ \mathbf{t}_i \in \mathbb{R} \tag{5.2}$$

Nach der Bestimmung des Zustandsvektors lässt sich durch Gleichung (5.3) die zugehörige homogene Transformationsmatrix $\mathbf{T}$ des Modells durch Multiplikation der Basismatrizen $\mathbf{E}$ für die aktuellen Translationen und Rotationen berechnen (s. Anhang 8.4).

$$\mathbf{T} = \mathbf{E}_{tx}(x) \cdot \mathbf{E}_{ty}(y) \cdot \mathbf{E}_{tz}(z) \cdot \mathbf{E}_{r\omega}(\omega) \cdot \mathbf{E}_{r\phi}(\phi) \cdot \mathbf{E}_{r\kappa}(\kappa),\ \mathbf{T} \in \mathbb{R}^{4\times4},\ \mathbf{E}_i \in \mathbb{R}^{m\times n} \tag{5.3}$$

Bekannt aus der optischen Vermessung ist die Bestimmung der Pose ein Optimierungsproblem, bei dem der Zustandsvektor für gewöhnlich iterativ bezüglich eines Fehlermaßes verbessert wird. Unterschiede bestehen hier in der Wahl des Fehlermaßes, der Art der korrespondierenden Merkmale zwischen Modell und Beobachtung sowie der Bestimmung der Korrespondenzen selbst. Im Falle von 3D Punktwolken und unbekannten Korrespondenzen werden häufig sogenannte Iterative Closest Point (ICP) Algorithmen eingesetzt [Rus01]. Auf diese Weise wird die Gesichtspose durch eine Ausrichtung des personenspezifischen Oberflächenmodells $\mathbf{S}$ an der mittels Stereoberechnung gewonnen Punktwolke $\mathbf{W}$ ermittelt.
Bei der Initialisierung des Zustandsvektors wird das Clusterverfahren zur 3D Gesichtslokalisation verwendet (Abschnitt 4.2.1). Es wird dabei der Schwerpunkt des Gesichtsclusters bestimmt und als Translationskomponente von $\mathbf{t}$ benutzt.

Das verwendete ICP Prinzip lässt sich formal wie folgt beschreiben:

- Sei $\mathbf{W}$ eine Menge von n Messpunkten $\mathbf{p}_i$ und $\mathbf{S}$ ein Oberflächenmodell bestehend aus m Vertices $\mathbf{a}_j$ und zugehörigen Normalen $\mathbf{b}_j$
- Sei $f_{cp}\,(\mathbf{W},\, \mathbf{a}_j)=\mathbf{p}_i$ der räumlich nächste Messpunkt zu einem Modellvertex $\mathbf{a}_j$
  1. Sei $\mathbf{t}^1$ die initiale Schätzung des Zustandsvektors
  2. Wiederhole für $k=1...k_{max}$ oder bis zum Erreichen einer Konvergenzschwelle:
     - Berechne die Menge aller korrespondierenden Punktpaare $\mathbf{C}^*$

       $\mathbf{C}^* = \bigcup_{i=1}^{m} \left\{ \left(\mathbf{a}_j(\mathbf{t}^k), f_{cp}\left(\mathbf{W}, \mathbf{a}_j(\mathbf{t}^k)\right)\right)\right\}$

- Berechne einen neuen Zustandsvektor $\mathbf{t}^{[k+1]}$, der die Fehlerfunktion $e(\mathbf{t})$ (5.4) bezüglich der Menge aller Punktpaare $\mathbf{C}^*$minimiert.

Zur Korrespondenzbestimmung zwischen den Modellvertices $\mathbf{a}_j$ und Messpunkten $\mathbf{p}_i$ wird eine Funktion f_{cp} eingesetzt, die auf der Grundlage eines kd-Baums die Suche des nächsten Nachbarn realisiert [Kle05, Ben75]. Ein kd-Baum ist eine raumunterteilende Datenstruktur zur effizienten Verwaltung von Punkt Suchanfragen bezüglich des nächstgelegenen Nachbarn im k-dimensionalen, im vorliegenden Fall dreidimensionalen Raum.
Die Fehlerfunktion $e(\mathbf{t})$ (5.4) repräsentiert die Güte des aktuellen Zustandsvektors $\mathbf{t}$. Der Gesamtfehler ergibt sich dabei aus der Summe der Abstände d_j zwischen den Modellvertices $\mathbf{a}_j \in \mathbb{R}^3$ zu der Ebene, die den räumlich nächstgelegenen Messpunkt $\mathbf{p}_i \in \mathbb{R}^3$ der Stereopunktwolke $\mathbf{W}$ enthält und dabei senkrecht zum Normalenvektor $\mathbf{b}_j$ des Modells $\mathbf{S}$ steht (Abbildung 5-3).

$$e(\mathbf{t}) = \sum_{j=1}^{m} (d_j(\mathbf{t}))^2 \rightarrow \min, \tag{5.4}$$

mit $d_j(\mathbf{t}) = (\mathbf{a}_j(\mathbf{t}) - \mathbf{p}_i) \cdot \mathbf{b}_j,\ \mathbf{t} \in \mathbb{R}^6,\ \mathbf{a}_j, \mathbf{b}_j, \mathbf{p}_i \in \mathbb{R}^3,\ d_j \in \mathbb{R}.$

Bezüglich der Fehlerfunktion $e(\mathbf{t})$ erfolgt die Optimierung des sechsdimensionalen Zustandsvektors $\mathbf{t}$ iterativ auf der Grundlage der Methode der kleinsten Fehlerquadrate. Da die Elementarmatrix für die Modellrotation Sinus und Kosinus Funktionen beinhaltet, ist das für den Minimierungsschritt aufzustellende Gleichungssystem nichtlinear. Daher wird mittels Taylorreihenapproximation eine Linearisierung durchgeführt, wobei die Reihe nach dem linearen Term abgeschnitten wird. Es werden hierzu die Modellkoordinaten nach den Komponenten des

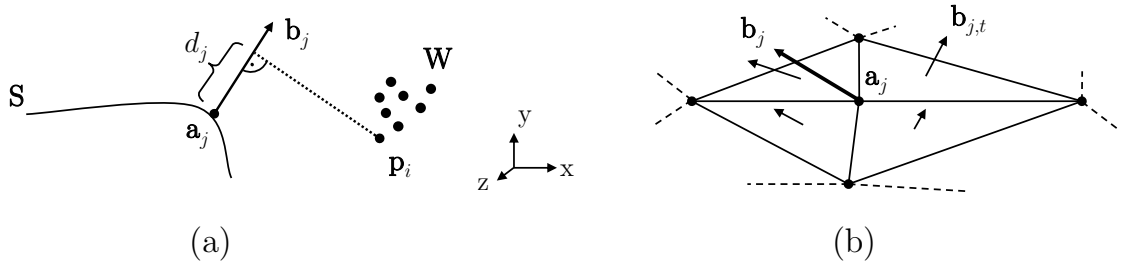


Abbildung 5-3: Modell Fitting, (a) Minimierung des senkrechten Abstandes d_j bei der Ausrichtung des Modells $\mathbf{S}$, (b) Definition der Normale $\mathbf{b}_j$ für jeden Vertex $\mathbf{a}_j$ durch Mittelung der Normalen $\mathbf{b}_{j,t}$ aller angrenzenden Dreiecke.

Zustandsvektors abgeleitet. Die Ableitungen $\partial \mathbf{a}_i / \partial \mathbf{t}$ werden analytisch berechnet, da dies im vorliegenden Fall von Translationen und Rotationen einfach zu realisieren ist. Für jeden 3D Modellpunkt $\mathbf{a}_i$ lassen sich somit drei Beobachtungsgleichungen aufstellen (5.5), was zu einem hochgradig überbestimmten Gleichungssystem führt. Aus dieser Überbestimmtheit folgt wiederum eine gegen Ausreißer robuste Schätzung des Zustandsvektors.

$$\mathbf{a}_i(\mathbf{t}) + \partial \mathbf{a}_i / \partial \mathbf{t} \cdot \Delta \mathbf{t} = \mathbf{p}_i \tag{5.5}$$

Die differentiellen Änderungen $\Delta \mathbf{t}$ des Zustandsvektors werden durch Lösung des Gleichungssystems mittels Ausgleichsrechnung bestimmt und zur iterativen Verbesserung von $\mathbf{t}$ benutzt. Das ICP Verfahren stoppt, falls $e(\mathbf{t})$ einen vorgegebenen Schwellwert unterschreitet oder eine bestimmte Anzahl an Iterationen erreicht wurde. Durch den somit bestimmten Zustandsvektor $\mathbf{t}$ des Modells wird die tatsächliche Orientierung des Gesichts zuverlässig erfasst (Abbildung 5-4).

5.1.2 3D-gestützte Normierung des Gesichts

Die Orientierung des Gesichts wird durch den Zustandsvektor $\mathbf{t}$ des Modells in Form von Translations- und Rotationsparametern erfasst. Durch Verwendung der aufgezeichneten Farbbilder und des 3D Modells ist es möglich auf der Grundlage der bekannten Orientierung eine standardisierte Frontaldarstellung des jeweiligen Gesichts zu erzeugen. Dieser Schritt basiert auf der Rasterisierung des Modells, bei dem dieses entsprechend eines Bildrasters abgetastet und somit in eine diskrete Pixeldarstellung überführt wird. Aus der Computergrafik ist eine Reihe von Techniken bekannt, um 3D Szenen und Objekte zu rasterisieren, z.B. durch Ray-

(a) (b)

Abbildung 5-4: Beispiel der stereobasierten ICP Poseschätzung, (a) ausgerichtetes Modell, (b) Projektion der aktuellen Pose auf das Bild.

casting oder durch OpenGL Rendering [Fol95], was in dieser Arbeit bevorzugt wurde, da es von aktuellen Grafikkarten hardwareseitig unterstützt wird und somit klare Geschwindigkeitsvorteile bietet.
Hierzu wird das Modell $\mathbf{S}$ in einem Vorverarbeitungsschritt aus frontaler Blickrichtung auf der Grundlage einer virtuellen Kamera $\mathbf{K}_{GL}$ abgetastet. Das verwendete Kameramodell entspricht dabei mit Ausnahme der Verzeichnungsparameter dem aus Abschnitt 3.1. Auf Grundlage der OpenGL Rasterisierung wird weiterhin die zur virtuellen Kamera korrespondierende Tiefenkarte $\mathbf{D}_f$ (5.6) berechnet.

$$\mathbf{D}_f = (df_{i,j})_{mxn}\,,\ df_{i,j} \in \mathbb{R} \tag{5.6}$$

Ausgehend von der Bildebene der Kamera $\mathbf{K}_{GL}$ repräsentiert die Matrix $\mathbf{D}_f$ Tiefenwerte $df_{i,j}$ der virtuellen Szene an allen Koordinaten $(i,\, j)$ (Abbildung 5-5). Praktisch bedeutet dies, dass ein Schnitttest zwischen einem Strahl ausgehend vom Projektionszentrum der Kamera zur Rasterkoordinate $(i,\, j)$ mit dem Modell in der Szene erfolgt (s. Abbildung 5-6).
Beim Auslesen der Tiefenkarte aus dem Speicher der Grafikkarte werden für alle Koordinaten $(i,\, j)$ korrespondierende 3D Punkte $\mathbf{p}_{f_{i,j}}$analog zur Transformation k^{-1} (3.6) bestimmt und zur Menge $\mathbf{P}_f$ zusammengefasst.

$$\mathbf{P}_f = \bigcup_{i,j=0}^{m,n} \mathbf{p}_{f_{i,j}} = k^{-1}\left([i\ j]^T\,,\ df_{i,j},\ \mathbf{K}_{GL}\right), \quad \mathbf{p}_{f_{i,j}} \in \mathbb{R}^3 \tag{5.7}$$

mit k^{-1} als Transformation von Bild- zu Weltkoordinaten entsprechend (3.6), Koordinate $(i,\, j)$, Tiefenwert $df_{i,j}$ und Kameramodell $\mathbf{K}_{GL}$.

Die Tiefenkarte $\mathbf{D}_f$ dient als Grundlage zur Berechnung des normierten Gesichts. Somit entspricht die Größe der Matrix $\mathbf{D}_f$ der Größe des normierten Gesichts. Für jeden Punkt $(i,\, j)$ der Matrix wird weiterhin unter Nutzung des Wissens über die aktuelle Orientierung $\mathbf{T}$ der zugehörige Farbwert aus den Bilddaten der Kamera $\mathbf{K}$ bestimmt. Hierzu wird die Bildkoordinate $\mathbf{i}_{i,j}$ (5.8) wie folgt für eine der beiden Stereokameras $\mathbf{K}_{1,2}$ ermittelt (Abbildung 5-5(c)).

$$\mathbf{i}_{i,j} = k\left(\mathbf{T}\cdot\mathbf{p}_{f_{i,j}},\ \mathbf{K}\right),\ \ \mathbf{T}\in\mathbb{R}^{4\times4},\ \mathbf{i}_{i,j}\in\mathbb{R}^2,\ \mathbf{p}_{f_{i,j}}\in\mathbb{R}^3 \tag{5.8}$$

mit k als Transformation von Welt- zu Bildkoordinaten entsprechend (3.5), Posematrix $\mathbf{T}$ (5.3) und Kameramodell $\mathbf{K}$.

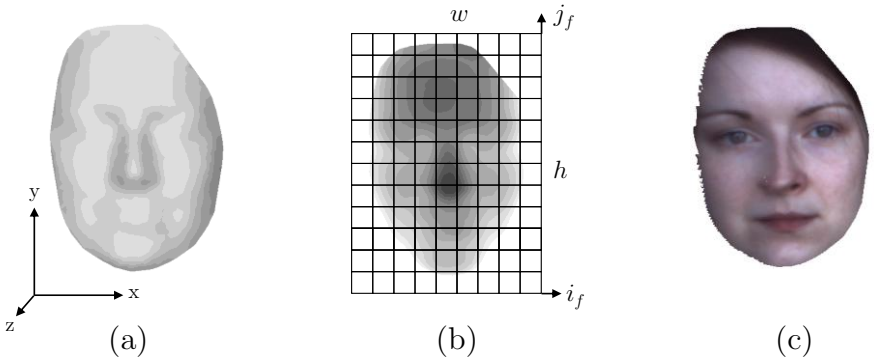


Abbildung 5-5: Rasterisierung, (a) Modell $\mathbf{S}$, (b) Raster mit Tiefenkarte $\mathbf{D}_f$, (c) Berechnung des Gesichts auf der Grundlage der Tiefenkarte.

Zur Erkennung und Korrektur von Selbstverdeckungen wird eine weitere Rasterisierung mit einer zugehörigen Tiefenkarte $\mathbf{D}_r$ (5.9) durchgeführt, bei der die aktuelle Modellpose $\mathbf{T}$ verwendet wird und die Parameter der realen Kamera $\mathbf{K}$ in der OpenGL Szene so nachgebildet werden, dass die virtuelle Bildebene mit dem aufgezeichneten Kamerabild übereinstimmt (Abbildung 5-6, Abbildung 5-7(a)).

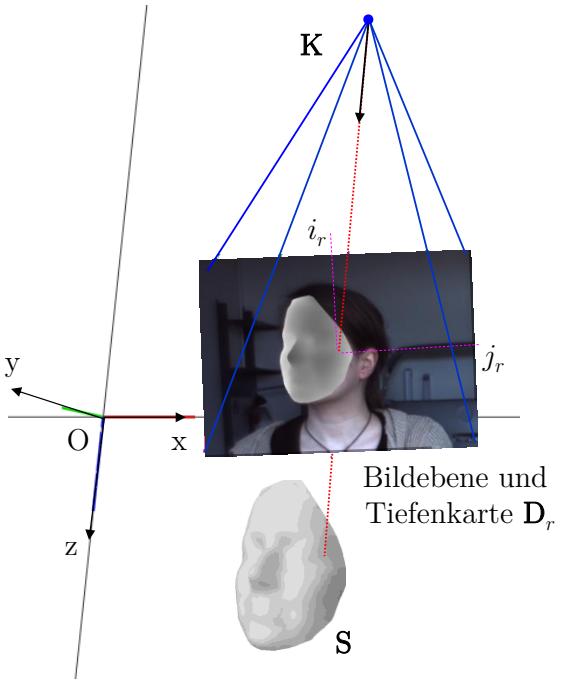
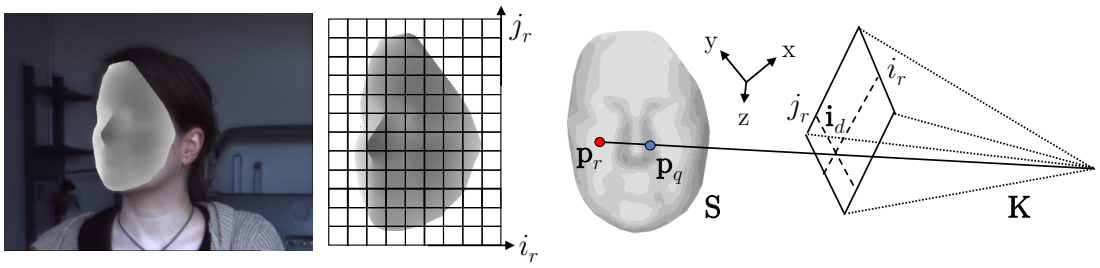


Abbildung 5-6: Szene mit Nachbildung der Kamera $\mathbf{K}$ in OpenGL und Modell $\mathbf{S}$ in aktueller Pose.



(a) (b)

Abbildung 5-7: Verdeckungstest, (a) Tiefenkarte des Modells in aktueller Pose und Rasterisierung $\mathbf{D}_r$, (b) Test für einen Sehstrahl durch Koordinate $\mathbf{i}_d$.

$$\mathbf{D}_r = (dr_{i,j})_{mxn} \tag{5.9}$$

Der Verdeckungstest erfordert die Berechnung zweier Transformationen. Im ersten Schritt werden alle Punkte der Menge $\mathbf{P}_f$ mit der aktuellen Posematrix $\mathbf{T}$ multipliziert. Das Ergebnis ist die Menge $\mathbf{P}_r$ aller abgetasteten Modellpunkte in der aktuellen Orientierung des Gesichts.

$$\mathbf{P}_r = \bigcup_i \mathbf{p}_{r_i} = \bigcup_i (\mathbf{T} \cdot \mathbf{p}_{f_i}), \quad \mathbf{p}_r \in \mathbb{R}^3, \mathbf{T} \in \mathbb{R}^{4\times4}, \ \mathbf{p}_{f_i} \in \mathbb{R}^3 \tag{5.10}$$

Im Anschluss wird für jeden Punkt $\mathbf{p}_r$ die Rasterkoordinate $\mathbf{i}_d = (i_r, j_r)$ auf der Grundlage des Kameramodells analog zu (3.5) bestimmt. Weiterhin wird der Punkt $\mathbf{p}_q$ (5.11) auf dem Modell berechnet, der bezüglich der Kamera $\mathbf{K}$ von der Koordinate $\mathbf{i}_d$ aus als erster vom Sehstrahl geschnitten wird (Abbildung 5-7(b)).

$$\mathbf{p}_q = k^{-1}(\mathbf{i}_d, \ dr_{i,j}, \ \mathbf{K}), \tag{5.11}$$

mit $\mathbf{p}_q \in \mathbb{R}^3$, $\mathbf{i}_d \in \mathbb{R}^2$, $dr_{i,j} \in \mathbb{R}$, Kamera $\mathbf{K}$ und k^{-1} entsprechend (3.6).

Nach der Bestimmung von $\mathbf{p}_r$ und $\mathbf{p}_q$ vereinfacht sich der Verdeckungstest zu einem Schwellwertvergleich. Somit ist im normierten Bild ein Punkt an der Rasterposition (i_f, j_f) verdeckt, wenn der euklidische Abstand d_{rq} zwischen den Punkten $\mathbf{p}_r$ und $\mathbf{p}_q$ größer als ein festgelegter Schwellwert ist. Wie in der Abbildung 5-7(b) zu sehen bedeutet dies, dass das ein Punkt genau dann verdeckt ist, wenn $\mathbf{p}_r$ und $\mathbf{p}_q$ nicht identisch sind.

$$d_{rq} = |\ \mathbf{p}_r - \mathbf{p}_q\ |, \ \mathbf{p}_r, \mathbf{p}_q \in \mathbb{R}^3 \tag{5.12}$$

Verdeckungen erscheinen als Löcher im normierten Bild. Kleine Löcher werden mit der Farbe umgebender Pixel gefüllt. Große Löcher hingegen können nur durch Hinzuziehen von Information aus weiteren Kameras beseitigt werden. Das entwickelte Framework ermöglicht hierbei grundsätzlich die Nutzung einer Multikameraanordnung.
Das Bild des normierten Gesichts stellt eine leistungsfähige Grundlage zur Merkmalsanalyse in Einzel- wie auch Sequenzaufnahmen dar. Merkmalsdetektion und Tracking werden durch die Tatsache, dass das Gesicht eine standardisierte Größe und Orientierung aufweist, stark vereinfacht (Abbildung 5-8).

5.1.3 Klassifikation nach Ansatz I

Grundsätzlich werden im Rahmen dieser Arbeit Mimikklassen betrachtet, die mit sechs Basisemotionen nach Ekman (Freude, Überraschung, Wut, Ekel, Angst, Trauer) assoziiert sind. Untersuchungen des Psychologen Paul Ekman haben ergeben, dass es für diese Emotionen eine personen-, ethnien- und kulturübergreifende Universalität gibt [Kel00]. Generell liegen jeder Mimik-Kategorie charakteristische Bewegungsmuster zugrunde, die durch die erfassten dynamischen Merkmale repräsentiert werden (Abbildung 5-9). Diese Muster eignen sich ausgezeichnet, um sie mit Hilfe maschineller Lernverfahren automatisch einer zugrundeliegenden Kategorie, d.h. Klasse zuzuordnen
Entsprechend der Festlegung physiologisch motivierter Regionen (Abschnitt 4.3) und der darauf aufbauenden allgemeinen Definition dynamischer Merkmale stellt die Gesichtsnormierung eine Möglichkeit zur Merkmalserfassung und anschließender Erkennung dar (Abbildung 5-10). Zu diesem Zweck werden die gemittelten Verschiebungsvektoren $\bar{\mathbf{v}}_i^t$ (4.15) für alle n_i=14 Regionen bestimmt. Entsprechend der Definition der Aktivierungsfunktion $v_{sum}(t)$ (4.16) ist eine Klassifikation nur bei einem Mindestgrad v_{min} an gemessener Bewegung sinnvoll. Insbesondere

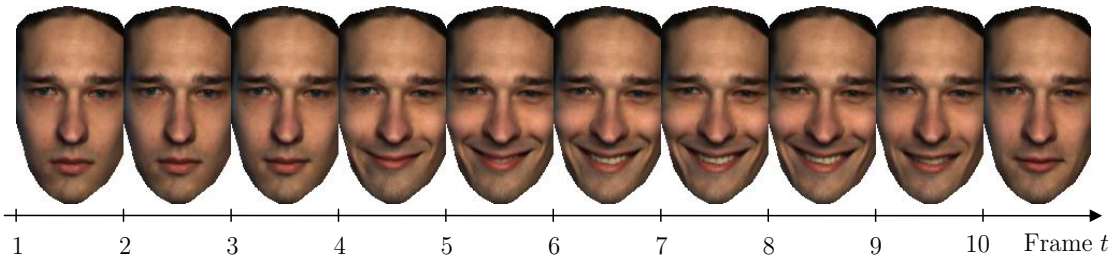


Abbildung 5-8: Beispielsequenz für die Gesichtsnormierung.

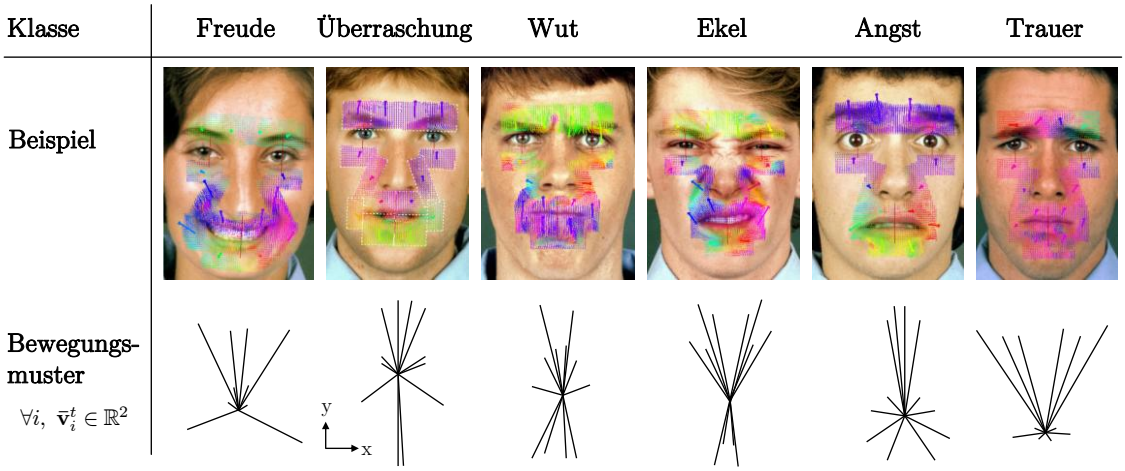


Abbildung 5-9: Beispiele typischer Bewegungsmuster, die auf der Grundlage physiologisch motivierter Regionen für die sechs betrachteten Mimikkategorien C_1-C_6 berechnet wurden. Dargestellt werden für jede Klasse die gemittelten Verschiebungsvektoren $\bar{\mathbf{v}}_i^t$ (4.15) aller n_i=14 Regionen. Diese Muster bilden die Grundlage zur Klassifikation.

wurden hierzu Untersuchungen mit maschinellen Lernverfahren wie Multilayer Perceptron, Support Vector Machines und k-Nearest Neighbor durchgeführt.

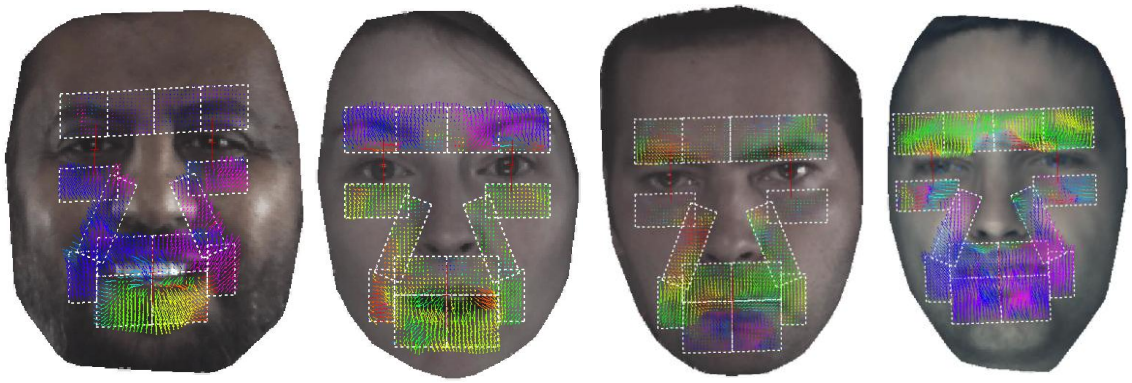


Abbildung 5-10: Beispiele zur Gesichtsnormierung und Extraktion dynamischer Merkmale $\mathbf{f}_{dyn}^t$ (4.17) sowie Erkennung verschiedener Mimikklassen.

5.1.4 Bewertung und Schlussfolgerungen

Experimentell validiert (Abschnitt 6.2.3) stellt die vorgeschlagene automatisierte Mimikanalyse auf der Grundlage der Gesichtsnormierung und der nachfolgenden Extraktion dynamischer Merkmale eine Möglichkeit zur Erkennung von Gesichtsausdrücken mit Fokus auf sechs prototypischen Basisemotionen dar. Hierbei wird durch Verwendung von Stereo- und Farbinformation das Gesicht des Nutzers automatisch und poseinvariant detektiert und die aktuelle Orientierung durch ICP Registrierung zuverlässig bestimmt. Durch den Einsatz personenspezifischer Gesichtsmodelle wird das Bild des Gesichts in eine normierte Darstellung überführt und auf diese Weise das Poseproblem überwunden (Abbildung 5-11).
Somit stellt die Gesichtsnormierung ein leistungsfähiges Werkzeug zur Merkmalsextraktion mit anschließendem Erkennungsschritt dar, was prinzipiell auch in anderen Applikationen Anwendung finden kann, z.B. zur Verifikation bei der Personenerkennung. Im vorgeschlagenen Ansatz zur Mimikerkennung bietet die Gesichtsnormierung eine solide Grundlage zur Verarbeitung dynamischer Merkmale, welche auf der Messung des Optischen Flusses in physiologisch motivierten Regionen basieren (s. Abschnitt 4.3, Abbildung 5-12). Eine Folge der Gesichtsnormierung ist eine weitgehende Personenunabhängigkeit extrahierter dynamischer Merkmale, welche die Klassifikation der Mimik begünstigt.

Zusammenfassend ist festzustellen, dass der Ansatz zur Mimikanalyse mittels Gesichtsnormierung folgende Vorteile bietet:

- Überwindung des Poseproblems durch Transformation des Gesichts in Frontaldarstellung.
- Keine Störungen bei der Bewegungserfassung durch Kopfbewegungen
- Robustheit durch Redundanz in den Stereo-Beobachtungsdaten
- Robustheit bei kompliziertem Hintergrund durch Stereomessung

Die Robustheitsvorteile der fortlaufenden Stereomessung bringen jedoch gleichermaßen den Nachteil eines hohen Rechenaufwandes, was die Anwendbarkeit erschwert. Im Rahmen dieser Arbeit wurde daher ein weiterer Ansatz untersucht, der die Technik der Gesichtsnormierung zur Merkmalsnormierung erweitert und somit einen reduzierten Berechnungsaufwand mit den Vorteilen der 3D Merkmalsextraktion verbindet. Desweiteren werden in dem erweiterten Ansatz nicht nur dynamische Merkmale berücksichtigt, sondern ebenso geometrische Merkmale, welche nicht den zeitlichen Kontext zur Änderungsmessung nutzen und somit das Klassifikationsergebnis grundlegend verbessern.

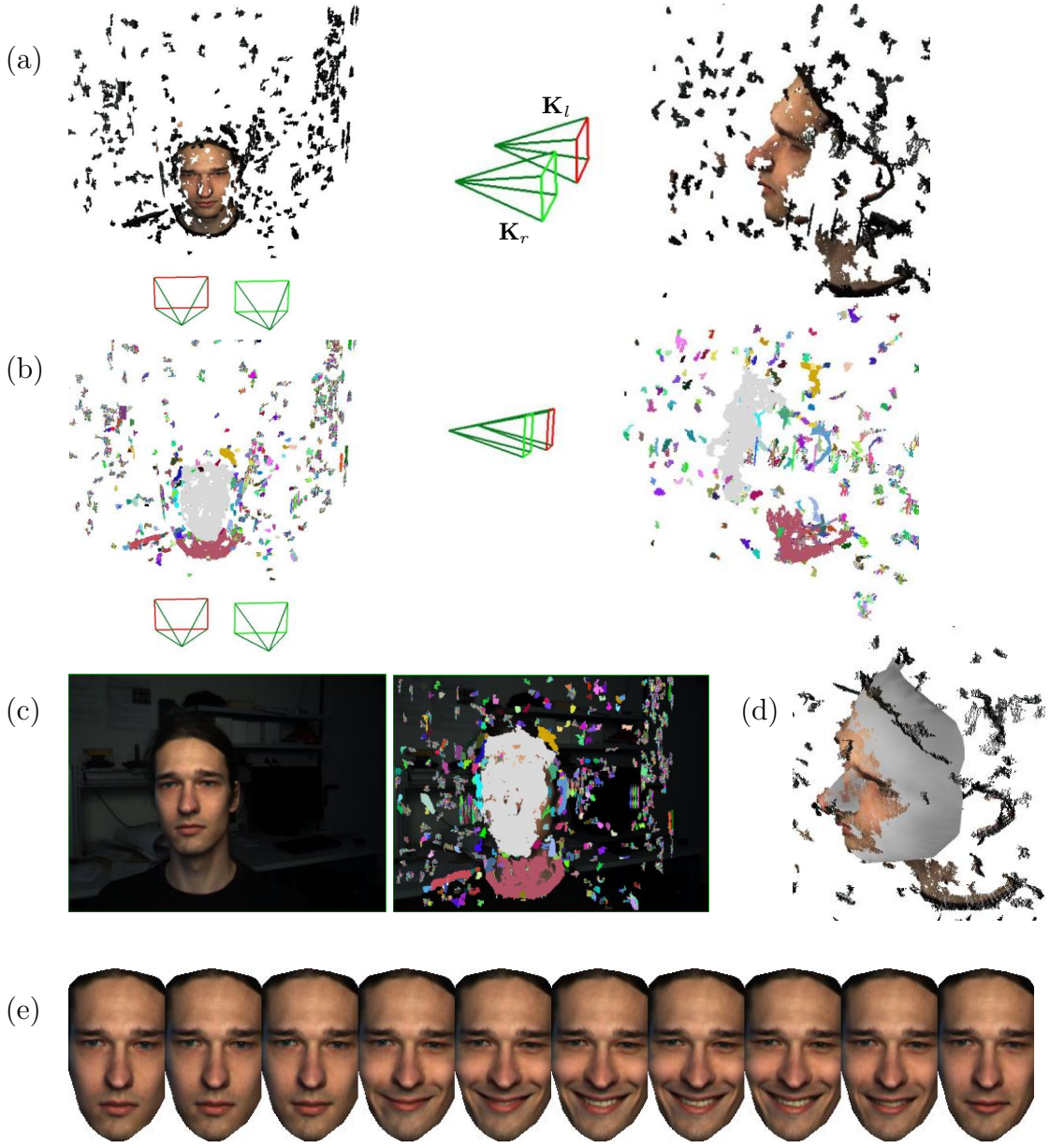


Abbildung 5-11: Verarbeitungsschritte nach dem Ansatz zur Gesichtsnormierung, (a) Punktwolke der Szene in Farbe, ermittelt durch ein passives Stereokamerasystem mit linker und rechter Kamera $\mathbf{K}_l/\mathbf{K}_r$, (b) Gesichtsdetektion durch Clusterbildung (hellgraues Cluster repräsentiert das Gesicht, s. Abschnitt 4.2.1), (c) Farbbild der rechten Kamera sowie Cluster, (d) Ergebnis des ICP Fittings von 3D Modell und Punktwolke, (e) Gesichtsnormierung.

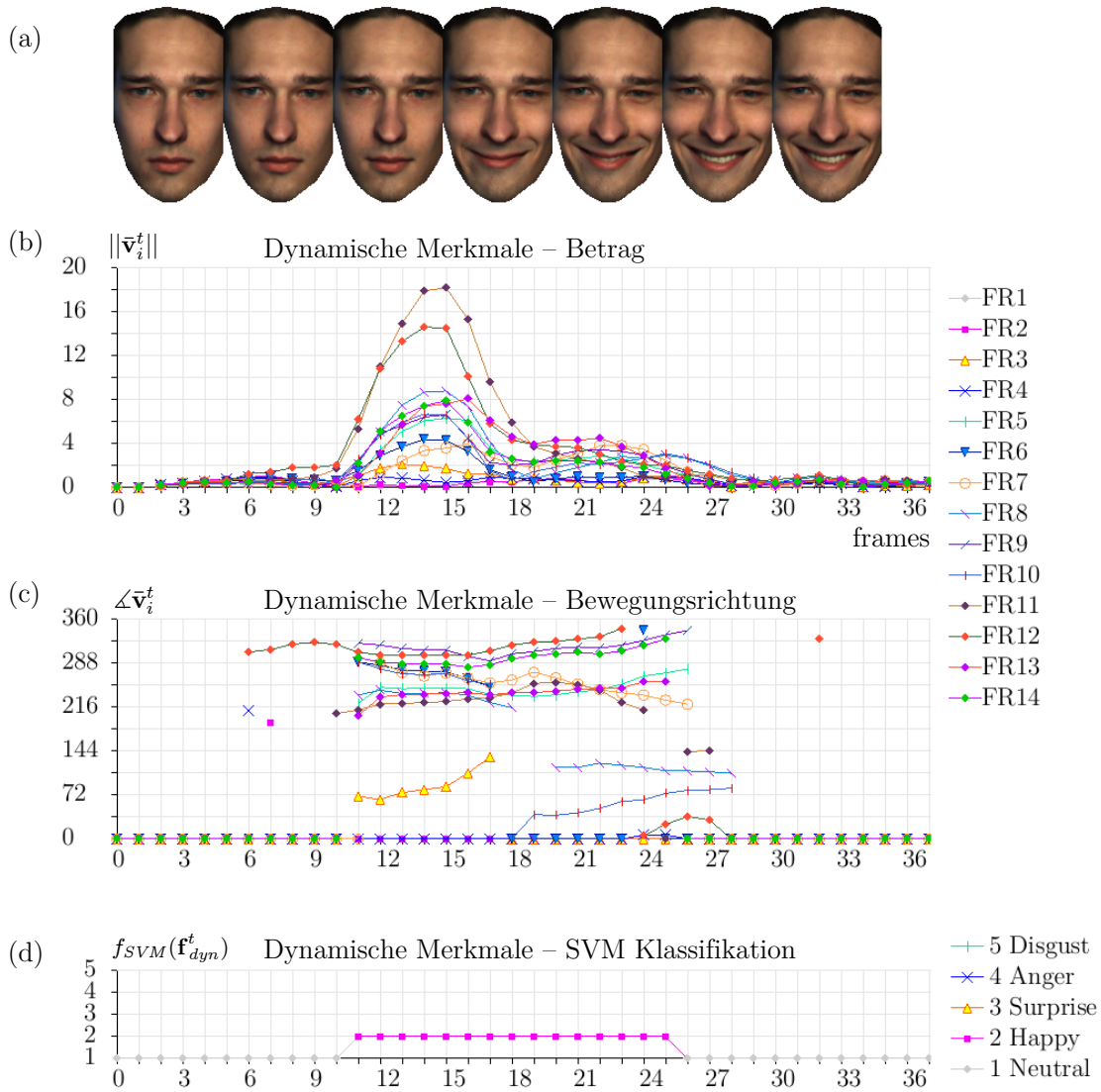


Abbildung 5-12: Merkmalsextraktion bei der Gesichtsnormierung, (a) Normierung, (b) Betrag $||\bar{\mathbf{v}}_i^t||$ der gemessenen Verschiebungsvektoren für jede Flussregion i, (c) zugehörige Richtungsinformation $\measuredangle\bar{\mathbf{v}}_i^t$, (d) SVM Klassifikation auf der Grundlage der Richtungsinformation.

5.2 Ansatz II – Mimikanalyse mittels Merkmalsnormierung

Die Idee der Merkmalsnormierung besteht darin, nicht das gesamte Gesicht zur Merkmalsextraktion durch eine Reihe photogrammetrischer Techniken zu transformieren, sondern lediglich die erfassten Merkmale. Auf diese Weise lassen sich nicht nur dynamische Merkmale einfach berücksichtigen (Abschnitt 4.3.1), welche ausschließlich zu Zeitpunkten bestimmt werden können an denen eine Änderung stattfindet, sondern ebenso geometrische Merkmale (Abschnitt 4.4.2), welche jederzeit messbar sind.
Zur Reduktion des Rechenaufwandes erfolgt hierbei eine Auswertung monokularer Farbbildsequenzen. Eine stereophotogrammetrische Aufnahme ist einzig in einem initialen Schritt zur Erzeugung des personenspezifischen Gesichtsmodells $\mathbf{S}$ (4.12) erforderlich. Im Anschluss an die Posebestimmung werden geometrische und dynamische Merkmale erfasst, normiert und einem Klassifikator zugeführt.

Abbildung 5-13: Ansatz II – Mimikanalyse mittels Merkmalsnormierung, nachfolgend werden die verwendeten Komponenten vorgestellt.

5.2.1 Poseschätzung in monokularen Bildfolgen

Vergleichbar zur Poseschätzung auf der Grundlage binokularer Messungen (Abschnitt 5.1.1), lässt sich durch Nutzung photogrammetrischer Verfahren die Modellorientierung mit Zustandsvektor $\mathbf{t}$ (5.2) bzw. Matrix $\mathbf{T}$ (5.3) ebenso aus monokularem Bildmaterial ermitteln. Hierfür wird eine Methode eingesetzt in der unter Berücksichtigung der projektiven Abbildung der verwendeten Kamera, der Zustandsvektor des Modells bestimmt wird, indem Ankerpunkte $\mathbf{a}_i$ des 3D Modells mit korrespondierenden Punkten $\mathbf{i}_j$ im Bild iterativ registriert werden.
Vergleichbar zur optischen Vermessung mittels Landmarken, wird für die Aufgabe zur Bestimmung der sechs Freiheitsgrade $\mathbf{t}=(\mathrm{x\ y\ z\ }\omega\ \phi\ \kappa)^{\mathrm{T}}$ eine Menge von mindestens drei korrespondierenden Punkten benötigt, um die erforderliche Anzahl an Beobachtungsgleichungen aufzustellen [Alb89].
Da die Mimikerkennung aus Gründen der Anwendbarkeit markerlos erfolgen soll, werden mit den Ankerpunkten virtuelle Landmarken eingesetzt. Die Schwierigkeit besteht darin, eine möglichst hohe Robustheit bei der Detektion dieser Punkte im

Bild zu erreichen, insbesondere bei veränderter Perspektive und Mimik. Weiterhin müssen die Punkte geeignet im Raum verteilt sein, um Fehler bei der Berechnung des Zustandsvektors durch lineare Abhängigkeiten zu vermeiden. Praktisch sind die einzigen Punkte, die jedes dieser Kriterien erfüllen, die beiden Mittelpunkte $\mathbf{i}_{le}$ und $\mathbf{i}_{re}$ der Augen. Auch im Falle geschlossener Augen können diese durch die Detektion der Augeninnenseite ermittelt werden (s. Abschnitt 4.4.1). Des Weiteren wird durch Nutzung von Modellinformation und Tracking die Nasenspitze $\mathbf{i}_n$ als dritter Punkt zur Berechnung der Pose verwendet. Für die Initialisierung und Reinitialisierung werden weiterhin die beiden Mundwinkel $\mathbf{i}_{lm}$ und $\mathbf{i}_{rm}$ benutzt. Es werden somit jeweils zwei Mengen von korrespondierenden Ankerpunkten $\mathbf{A}_\mathrm{j}/\mathbf{I}_\mathrm{j}$ (5.13) für das Modell und Bild definiert (Abbildung 5-14).

$$\begin{aligned} &\mathbf{A}_1 = \{\mathbf{a}_{re}, \mathbf{a}_{le}, \mathbf{a}_n\}, \quad \mathbf{A}_2 = \{\mathbf{a}_{re}, \mathbf{a}_{le}, \mathbf{a}_{rm}, \mathbf{a}_{lm}\}, \quad \mathbf{a}_j \in \mathbb{R}^3 \\ &\mathbf{I}_1 = \{\mathbf{i}_{re}, \mathbf{i}_{le}, \mathbf{i}_n\}, \quad \mathbf{I}_2 = \{\mathbf{i}_{re}, \mathbf{i}_{le}, \mathbf{i}_{rm}, \mathbf{i}_{lm}\}, \quad \mathbf{i}_j \in \mathbb{R}^2 \end{aligned} \tag{5.13}$$

Zur Initialisierung und Aktualisierung des Systems nach einer definierten Laufzeit ist eine möglichst frontale Kopfhaltung erforderlich. In der Zeitspanne dazwischen erfolgt ein Tracking der Merkmalspunkte mit gesonderter Behandlung der Augenpunkte, um Fehler durch Blinzeln zu vermeiden.
Zur Erhöhung der Robustheit wird beim Tracking für jeden der Merkmalspunkte ein Raster von 3x3 Punkten auf der Grundlage des pyramidalen Optischen Fluss Verfahrens nach Lucas-Kanade verfolgt (s. Abschnitt 3.3.1). In empirischen Tests hat sich dabei gezeigt, dass das Tracking auch bei Kopfrotationen, die aus der

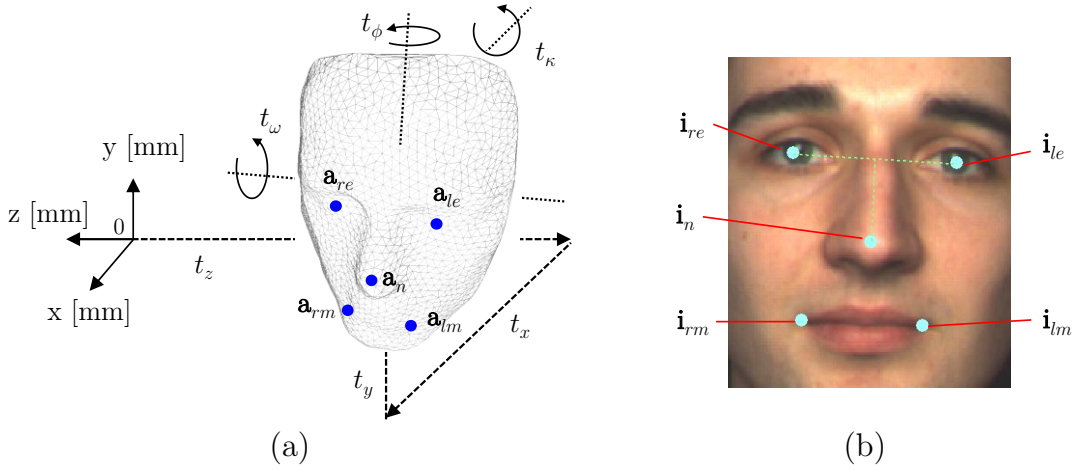


Abbildung 5-14: Posebestimmung in monokularen Bildern, (a) Modell mit Ankerpunkten $\mathbf{a}_j$, (b) korrespondierende Ankerpunkte im Bild $\mathbf{i}_j$.

Ebene führen, bis zu einer Stärke von $|t_\phi|,|t_\omega| \leq 25$ Grad stabil arbeitet.
Bei der Posebestimmung durch Approximation von 3D Modellen an Bildmaterial werden häufig Kanten oder markante Punkte verwendet [Wac97, Cal05]. Da in der vorliegenden Situation die Korrespondenzen zwischen Modell und Bilddaten bereits bekannt sind, kann die Modellapproximation unter Berücksichtigung der projektiven Abbildung des Bildaufnahmesystems direkt erfolgen. Hierbei repräsentiert die Fehlerfunktion $e(\mathbf{t})$ (5.14) die Güte des aktuellen Zustandsvektors $\mathbf{t}$, welcher in einem iterativen differentiellen Verfahren, basierend auf der Minimierung der kleinsten Fehlerquadrate, stückweise optimiert wird. In den durchgeführten experimentellen Untersuchungen hat sich dabei gezeigt, dass das verwendete Verfahren zumeist nach $j<5$ Iterationen konvergiert (Abbildung 5-15).

$$e(\mathbf{t}) = \sum_{j=1}^{n_c} ||\mathbf{i}_j - k(\mathbf{a}_j(\mathbf{t}))||^2 \rightarrow \min, \tag{5.14}$$

mit $e \in \mathbb{R}$, $\mathbf{i}_j \in \mathbb{R}^2$, $\mathbf{a}_j \in \mathbb{R}^3, \mathbf{t} \in \mathbb{R}^6$, k als projektive Abbildungsfunktion nach (3.5) und n_c als Anzahl korrespondierender Ankerpunkte. Für die Initialisierung und Reinitialisierung gilt n_c=4, sonst gilt n_c=3.

Anders als bei der ICP Approximation zwischen Modell und Messpunktwolke (Abschnitt 5.1.1), wird bei der Modell zu Bild Approximation die Projektion k beim Ermitteln der Differentiale berücksichtigt (5.15). Anschaulich betrachtet zeigen die Spaltenvektoren der Differentialmatrix $\delta k/\delta \mathbf{t}$ in die Richtung, in die sich ein in das Bild projizierter Modellpunkt $k(\mathbf{a}_j(\mathbf{t}))$ bewegen würde, wenn sich der zu der entsprechenden Spalte korrespondierende Freiheitsgrad um einen kleinen Betrag erhöhen würde.

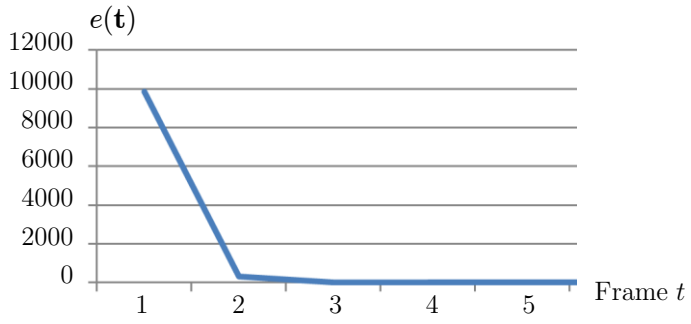


Abbildung 5-15: Beispiel zur Konvergenz der Fehlerfunktion $e(\mathbf{t})$ (5.14). Nach 5 Iterationen unterschreitet die Fehlerfunktion den Wert von $4.18 \cdot 10^{-13}$.

$$\delta k/\delta \mathbf{t} \cdot \Delta \mathbf{t} = \mathbf{i}_j - k(\mathbf{a}_j(\mathbf{t})) \tag{5.15}$$

Matrix $\delta k/\delta \mathbf{t}$ lässt sich aus dem Kameramodell herleiten [Wac97]. Folglich lassen sich nach (5.15) aus einer Anzahl von drei korrespondierenden Modell und Bildpunkten sechs Gleichungen zur Bestimmung des sechsdimensionalen Zustandsvektors $\mathbf{t}$ aufstellen. Aus dem somit hervorgehenden exakt bestimmten Gleichungssystem werden die differentiellen Änderungen $\Delta \mathbf{t}$ ermittelt und zu $\mathbf{t}$ addiert.
Zur Initialisierung und Reinitialisierung in längeren Bildfolgen erfolgt die Bestimmung der Pose auf der Grundlage der Ankerpunkte $\mathbf{I}_2$ und $\mathbf{A}_2$, d.h. Augen und Mundeckpunkten. Im Anschluss wird der Nasenpunkt des 3D Modells auf das Bild projiziert und bis zur Reinitialisierung mittels PLK-Tracker verfolgt [Luc81]. Zur Erhöhung der Stabilität, wird zum Tracking des Nasenpunktes ebenfalls ein Gitter von 3x3 Punkten verwendet. Während der Trackingphase erfolgt die Poseberechnung auf der Grundlage der Ankerpunkte $\mathbf{I}_1$ und $\mathbf{A}_1$, d.h. Augen- und Nasenpunkt. Im Anschluss an die Modellapproximation wird entsprechend (5.3) die zu $\mathbf{t}$ zugehörige Posematrix $\mathbf{T}$ verwendet, um das Gesichtsmodell $\mathbf{S}$ zur Erfassung geometrischer und dynamischer Merkmale in die aktuelle Orientierung zu überführen.

5.2.2 Erfassung der geometrischen Merkmale

Die Berechnung des zehndimensionalen Vektors $\mathbf{f}$ (4.21) für geometrische Roh-Merkmale, erfordert die Bestimmung der 3D Merkmalspunktmenge $\mathbf{P}_{fp}$ (4.18). Hierzu werden zunächst die entsprechenden Punkte $\mathbf{I}_{fp}$ im Bild durch BV Methoden (s. Abschnitt 4.4.1) ermittelt. Im Anschluss erfolgt mittels Funktion k^{-1} (5.16) die Projektion der Bildpunkte $\mathbf{I}_{fp}$ auf das Gesichtsmodell $\mathbf{S}$ unter Berücksichtigung der aktuellen Modellpose $\mathbf{t}$.

$$\mathbf{P}_{fp} = k^{-1}(\mathbf{I}_{fp}, d, \mathbf{K}) \text{ (analog zu (3.6))}, \tag{5.16}$$

mit $\mathbf{p}_j \in \mathbb{R}^3$, $\mathbf{i}_j \in \mathbb{R}^2$, $d \in \mathbb{R}$, Kameramodell $\mathbf{K}$.

Die für die Berechnung von k^{-1} erforderlichen Tiefenwerte d werden durch den Schnittest eines Sehstrahls der Kamera an der Pixelkoordinate $\mathbf{i}_j$ mit dem Gesichtsmodell $\mathbf{S}$ bestimmt. Tiefenwert d ist der Abstand zwischen der virtuellen Bildebene und dem Oberflächenmodell, welches entsprechend des Posevektors $\mathbf{t}$ ausgerichtet wird. Realisiert wird dieser Schnitttest durch eine sogenannte “binary space partitioning“ (BSP) Baumstruktur, die für das Modell $\mathbf{S}$ berechnet wird.

BSP-Bäume erzielen durch eine binäre Raumunterteilung eine logarithmische Laufzeit bei Schnittberechnungen und sind aus der Computergrafik bekannt [Fol95, Fuc80]. Ein Schnittstrahl wird somit nicht mehr gegen alle Dreiecke getestet, sondern rekursiv mit den Boxen des Baums und nur noch gegen eine Untermenge von Dreiecken in der letzten Hierarchieebene (Abbildung 5-16).
In den durchgeführten Untersuchungen wurde eine maximale Rekursionstiefe von n_i=5 und eine minimale Anzahl von n_t=50 Dreiecken je Box verwendet, bei einer durchschnittlichen Anzahl von circa 1000 Dreiecken für ein 3D Modell.

Die Auswertung der Mimik auf der Grundlage geometrischer Merkmale erfolgt durch den Vergleich der aktuellen Messung mit dem zuvor bestimmten Neutralzustand. Hierzu ist es erforderlich das beobachtete Gesicht bei neutraler Mimik zu erfassen und den Merkmalsvektor $\mathbf{f}_{neutral}$ entsprechend (4.21) zu bestimmen. Es bietet sich an diesen Schritt initial bei der Aufnahme des Gesichtsmodells durchzuführen.
Der Vergleich zwischen einer aktuellen Messung $\mathbf{f}^t$ zum Zeitpunkt t und dem Neutralzustand $\mathbf{f}_{neutral}$ basiert auf der Bildung des Verhältnisses $\mathbf{f}^t_{ratio}$ (5.18). Zu diesem Zweck wird die komponentenweise Division zweier n dimensionaler Merkmalsvektoren $\mathbf{a}$ und $\mathbf{b}$ duch den Operator # wie folgt definiert:

$$\mathbf{a} \# \mathbf{b} = (a_1 \,/\, b_1 \;\; a_2 \,/\, b_2 \ldots a_\mathrm{n} \,/\, b_\mathrm{n}) \in \mathbb{R}^n,\ \mathbf{a},\ \mathbf{b} \in \mathbb{R}^n \tag{5.17}$$

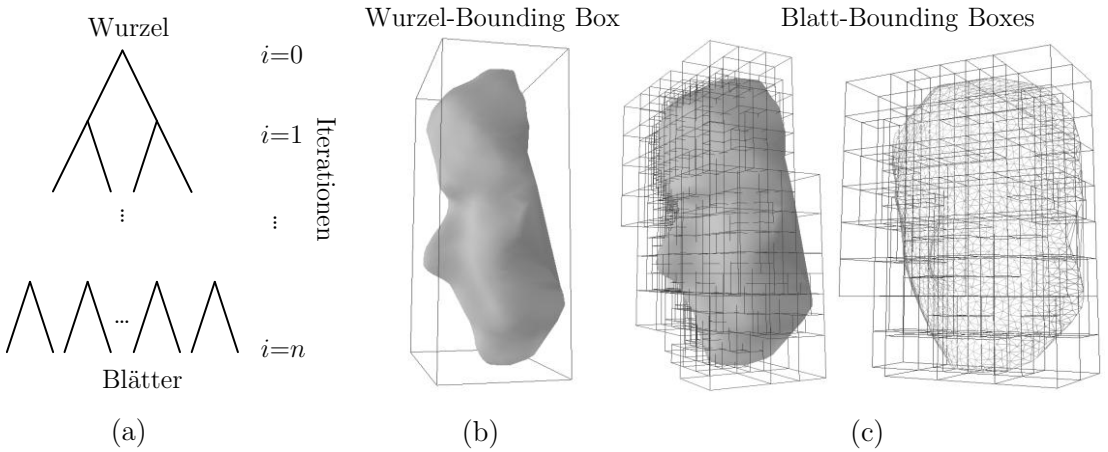


Abbildung 5-16: “Binary Space Partitioning” Baumstruktur (BSP-Tree), (a, b) Ausgehend von der Wurzel, die das gesamte 3D Modell mit allen Dreiecken enthält, werden für jede Ebene der Hierarchie rekursive Raumunterteilungen vorgenommen. (c) In der untersten Ebene enthalten die Blätter alle Dreiecke des Modells.

$$\mathbf{f}^t_{ratio} = \mathbf{f}^t \# \mathbf{f}_{neutral}, \quad \mathbf{f}^t_{ratio}, \mathbf{f}^t, \mathbf{f}_{neutral} \in \mathbb{R}^n \tag{5.18}$$

Es ist leicht ersichtlich, dass die Beträge der Abstände und Winkel, die zwischen den Merkmalspunkten erfasst werden, über die Menge der beobachteten Personen mitunter sehr verschieden sind. Ebenso variieren die Verhältnisse und Änderungen der einzelnen Merkmale. Dies führt zu einem Skalierungsproblem, da verschiedene Merkmale generell in unterschiedlichen Wertebereichen liegen. Um nicht die wesentliche Information über die Mimik durch Skalierungseffekte zu unterdrücken und somit das Klassifikationsergebnis zu verschlechtern, ist es erforderlich alle Messungen in einem Normierungsschritt in denselben Wertebereich zu überführen. Zu diesem Zweck wurde die Verteilung der Merkmalsverhältnisse in den benutzten Lerndaten, mit allen Klassen und Personen erfasst. Insbesondere wurden hierbei Mittelwert μ und Standardabweichung σ für jedes Element des Vektors $\mathbf{f}^t_{ratio}$ bestimmt. Die Normierung des Merkmalsvektors $\mathbf{f}^t_{geo}$ (5.20) erfolgt durch Berücksichtigung der Minimal- und Maximalwerte $\mathbf{c}_{min}$, $\mathbf{c}_{max}$ (5.19), welche sich über ein empirisch ermitteltes Konfidenzintervall von 2σ erstrecken [Kre05, Har05].

$$\begin{aligned} \mathbf{c}_{min} &= \mu \text{ - } 2\sigma, & \mathbf{c}_{min} &\in \mathbb{R}^n, \\ \mathbf{c}_{max} &= \mu + 2\sigma, & \mathbf{c}_{max} &\in \mathbb{R}^n, \end{aligned} \tag{5.19}$$

mit μ, $\sigma \in \mathbb{R}^n$ als Mittelwert und Standardabweichung.

$$\begin{gathered} \mathbf{f}^t_{geo} = (\mathbf{f}^t_{ratio} - \mathbf{c}_{min}) \# (\mathbf{c}_{max} - \mathbf{c}_{min}) = \\ (\mathbf{f}^t_{ratio} - \mathbf{c}_{min}) \# 4\sigma, \quad \mathbf{f}^t_{geo} \in \mathbb{R}^n \end{gathered} \tag{5.20}$$

5.2.3 Erfassung der dynamischen Merkmale

Auch für den Ansatz der Merkmalsnormierung erfolgt die Erfassung dynamischer Merkmale entsprechend der Festlegung physiologisch motivierter Regionen und der allgemeinen Definition dynamischer Merkmale. Unter Ausnutzung der aktuellen Pose des Modells, erfolgt die Bestimmung der Verschiebungsvektoren (VV) dabei auf der Grundlage der in das Bild projizierten Flussregionen. Hierzu werden VV an den Bewegungsabtastpunkten $\mathbf{p}_{k,i} \in \mathbb{R}^3$ im Bild berechnet, welche mit jeder Flussregion i verknüpft sind (Abschnitt 4.3, Abbildung 5-17(b)).

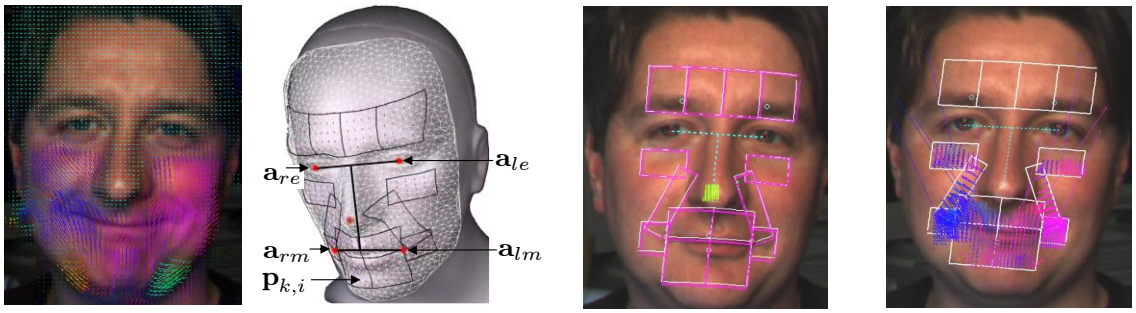
Die Schwierigkeit bei der Bestimmung auswertbarer VV liegt unter anderem in der Unterdrückung der globalen Kopfbewegung m_{head}, welche die Mimik relevante Bewegung m_{exp} zur Gesamtbewegung m_{total} überlagert. Zur Erfassung der globalen Kopfbewegung wird die Verschiebung der Randpunkte der einzelnen Flussregionen ausgewertet, dabei wird entsprechend der Relation (5.21) die globale Bewegung separat für jede Region unterdrückt und somit die relevante Bewegung bestimmt (Abbildung 5-17(c, d)).

$$m_{exp} = m_{total} - m_{head} \tag{5.21}$$

Die um die globale Kopfbewegung korrigierte und damit ausschließlich mit Mimik assoziierte Bewegungsinformation m_{exp} kann jedoch nicht sofort als Merkmal verwendet werden, da die Orientierung des Kopfes variieren kann und nicht zwingend frontal sein muss. Daraus folgt, dass die ermittelten Vektoren durch die Perspektive der Kamera beeinflusst werden. Um Vergleichbarkeit der VV zu erreichen ist zur Überwindung des Poseproblems somit ein weiterer Korrekturschritt erforderlich, welcher als Nullposetransformation z (5.22) bezeichnet wird. Durch diesen Schritt werden alle zu einem Zeitpunkt t erfassten VV $\mathbf{v}^t_{j,i} \in \mathbb{R}^2$ sämtlicher Regionen i in die sogenannte Nullpose überführt.

$$\hat{\mathbf{v}}^t_{j,i} = z(\mathbf{v}^t_{j,i}, \mathbf{E}, \mathbf{K}), \tag{5.22}$$

mit $\hat{\mathbf{v}}^t_{j,i}, \mathbf{v}^t_{j,i} \in \mathbb{R}^2$, Zeitpunkt t, Flussregion i, Kameramodell $\mathbf{K}$, Ebene $\mathbf{E}$.



(a) (b) (c) (d)

Abbildung 5-17: Bewegungserfassung bei der Merkmalsnormierung, (a) Gemessenes VV-Feld, (b) Gesichtsmodell mit 3D Flussregionen, welche auf der Grundlage von vier Ankerpunkten $\mathbf{a}_j \in \mathbb{R}^3$ bestimmt werden sowie Bewegungsabtastpunkte $\mathbf{p}_{k,i} \in \mathbb{R}^3$ für alle Regionen i, (c) Kopfbewegung, erkennbar durch VV an der Nase; die Subtraktion globaler Bewegung wird indirekt an den leeren Flussregionen verdeutlicht, (d) Erfassung relevanter Bewegung m_{exp} verursacht durch Mimik.

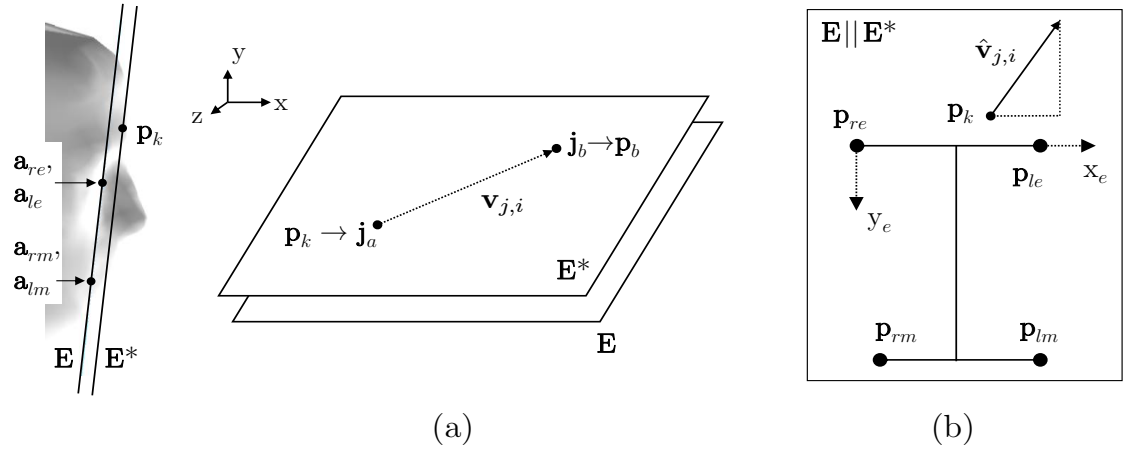


Abbildung 5-18: Überwindung der Poseabhängigkeit gemessener VV, (a) Bildprojektion der Bewegungsabtastpunkte $\mathbf{p}_k$, Bestimmung der VV und Rückprojektion auf Ebene E*, (b) VV in einheitlichem Koordinatensystem in Ebene E*.

Zur Realisierung der Transformation wird zunächst eine Ebene $\mathbf{E}$ definiert, welche die Ankerpunkte der Flussregionen, d.h. Augen und Mundwinkel ($\mathbf{a}_{re,}$ $\mathbf{a}_{le}$, $\mathbf{a}_{rm}$, $\mathbf{a}_{lm}$) in der aktuellen Orientierung bestmöglich approximiert (Abbildung 5-18(a)).
Sei $\mathbf{p}_k$ ein Bewegungsabtastpunkt einer beliebigen Flussregion, welcher sich in einer zu $\mathbf{E}$ parallelen Ebene $\mathbf{E}^*$ befindet (Abbildung 5-18). Ferner sei $\mathbf{j}_a$ die Projektion des Punktes ins Kamerabild und $\mathbf{v}_{j,i}^t \in \mathbb{R}^2$ der an dieser Position bestimmte VV, der zur Bildkoordinate $\mathbf{j}_b$ zeigt. Durch eine Schnittberechnung zwischen einem Sehstrahl der Kamera an der Subpixelkoordinate $\mathbf{j}_b$ und der Ebene $\mathbf{E}^*$ wird der VV $\hat{\mathbf{v}}_{j,i}^t \in \mathbb{R}^2$ ermittelt, der eine Projektion des Vektors $\mathbf{v}_{j,i}^t$ aus dem Bild in die Hilfsebene $\mathbf{E}^*$ realisiert. Zur Überwindung der Poseabhängigkeit wird der VV $\hat{\mathbf{v}}_{j,i}^t$ bezüglich eines zweidimensionalen, parallel zur Ebene $\mathbf{E}$ definierten Koordinatensystems ausgewertet, dessen Achsen durch die Ankerpunkte aufgespannt werden (Abbildung 5-18(b)).
Im Anschluss an die VV Transformation zu $\hat{\mathbf{v}}_{j,i}^t$ findet äquivalent zu Gleichung (4.14) und (4.15) eine Akkumulation $\tilde{\mathbf{v}}_{j,i}^t$ und nachfolgende Mittelung $\bar{\mathbf{v}}_i^t$ der berechneten VV statt, um die Wirkung von Ausreißern zu eliminieren bzw. zu reduzieren. Die Richtungsinformation der in allen 14 Regionen zu einem Zeitpunkt t berechneten Vektoren $\bar{\mathbf{v}}_i^t$ wird analog zur Definition des Merkmalsvektors für dynamische Merkmale $\mathbf{f}_{dyn}^t$ (4.17) zusammengefasst.

5.2.4 Klassifikation nach Ansatz II

Die durch den Ansatz der Merkmalsnormierung zu einem Zeitpunkt t erfassten hochdimensionalen Merkmalsvektoren $\mathbf{f}^t_{geo}$ und $\mathbf{f}^t_{dyn}$ eignen sich hervorragend, um durch maschinelle Lernverfahren zugrundeliegende Muster automatisch zu erkennen und diese verschiedenen Kategorien zuzuordnen. Ebenso wie beim Ansatz der Gesichtsnormierung setzen dabei die dynamischen Merkmale einen hinreichenden Aktivierungsgrad $v_{sum}(t)>v_{min}$ zur Klassifikation voraus, d.h. eine minimale erfassbare Bewegung, entsprechend (4.16). Hierzu wurden Untersuchungen mit drei gängigen überwachten Verfahren, d.h. k-Nearest Neighbor, Multilayer Perceptron und Support Vector Machines durchgeführt. Dabei hat das SVM Verfahren im Mittel die besten Resultate erzielt. Detaillierte Ergebnisse hierzu werden im Ergebniskapitel unter 6.1.2 und 6.2.2 dargelegt.
Die durch die vorgestellten Methoden erfassten Merkmale stellen den Ausgangspunkt zur Klassifikation dar. Aus diesem Grund wurden die Qualität bzw. Eignung der ermittelten Merkmalsräume als Lerndaten für automatische Entscheider mit Hilfe verschiedener Techniken der statistischen Datenanalyse und des maschinellen Lernens überprüft und nachgewiesen. Diesbezüglich werden umfangreiche Ergebnisse für geometrische und dynamische Merkmale im Ergebniskapitel in den Abschnitten 6.1.1 und 6.2.1 vorgestellt.
Durch verschiedene Techniken wurde dabei die Abgrenzung der Klassen in den hochdimensionalen Merkmalsräumen nachgewiesen.

5.2.5 Bewertung und Schlussfolgerungen

Die im Rahmen dieser Arbeit vorgeschlagene automatische bildbasierte Mimikanalyse mittels Merkmalsnormierung stellt neben der Gesichtsnormierung eine weitere neue Technik zur Erkennung von Mimik mit Fokus auf prototypischen Basisemotionen dar. Die durchgeführte experimentelle Validierung (Kapitel 6) bestätigt die erreichte Qualität und Zuverlässigkeit des entwickelten Verfahrens und ermöglicht weiterhin die Beantwortung, der in Kapitel 1.1 aufgeworfenen, sich aus der vorgeschlagenen Systemstruktur ableitenden grundlegenden Forschungsfragen dieser Arbeit.
Das vorgestellte Verfahren wird zur effizienten Erfassung und Analyse geometrischer und dynamischer Merkmale eingesetzt und stellt somit eine Erweiterung des zuvor beschriebenen Ansatzes zur Gesichtsnormierung dar. Das Verfahren nutzt dabei personenspezifische Gesichtsmodelle und monokulare Farbbildfolgen

zur Bestimmung der Kopfpose auf der Basis korrespondierender Ankerpunkte zwischen Modell und aktuellem Bild. Weiterhin beruht die Merkmalserfassung auf der modellgestützten 3D Transformation extrahierter Bildmerkmale, die durch BV Techniken bestimmt werden (Abbildung 5-19(a)). Insbesondere wird hierfür zu jedem Zeitpunkt t der Vektor $\mathbf{f}^t$ geometrischer Rohmerkmale bestimmt, welcher 3D Abstandsmaße und Winkel repräsentiert. Davon ausgehend wird das Verhältnis $\mathbf{f}^t_{ratio}$ zwischen derzeitigem und neutralem Gesicht berechnet und anschließend normiert. Das Ergebnis ist der geometrische Merkmalsvektor $\mathbf{f}^t_{geo}$ der direkt einem Klassifikator nach Wahl zugeführt wird (Abbildung 5-19(b-d), Abbildung 5-20).

Ergänzend hierzu wird Bewegungsinformation durch die Auswertung physiologisch motivierter Regionen erfasst. Gemäß der Definition dynamischer Merkmale $\mathbf{f}^t_{dyn}$ (4.17) werden für alle n_i=14 Regionen gemittelte Vektoren $\bar{\mathbf{v}}^t_i$ bestimmt (Abbildung 5-21). Bei der Berechnung werden Störeffekte durch globale Kopfbewegung und Perspektive unterdrückt.

Zusammenfassend ist festzustellen, dass der Ansatz zur Mimikanalyse mittels Merkmalsnormierung folgende Vorteile bietet:

- Reduzierter Berechnungsaufwand im Vergleich zur Gesichtsnormierung, da bei Merkmalsnormierung nur eine Auswertung monokularer Bilder erfolgt
- 3D Merkmalsextraktion, welche das Poseproblem überwindet
- Überwindung von Skalierungsproblemen bei der Klassifikation

Geometrische und dynamische Merkmale erfassen unterschiedliche Aspekte beim Auftreten von Mimik, was grundsätzlich der Steigerung der Erkennungsleistung dienen kann. Eine Möglichkeit zur integrierten Auswertung geometrischer und dynamischer Merkmale wird im nachfolgenden Abschnitt auf der Grundlage einer Fusion beschrieben.

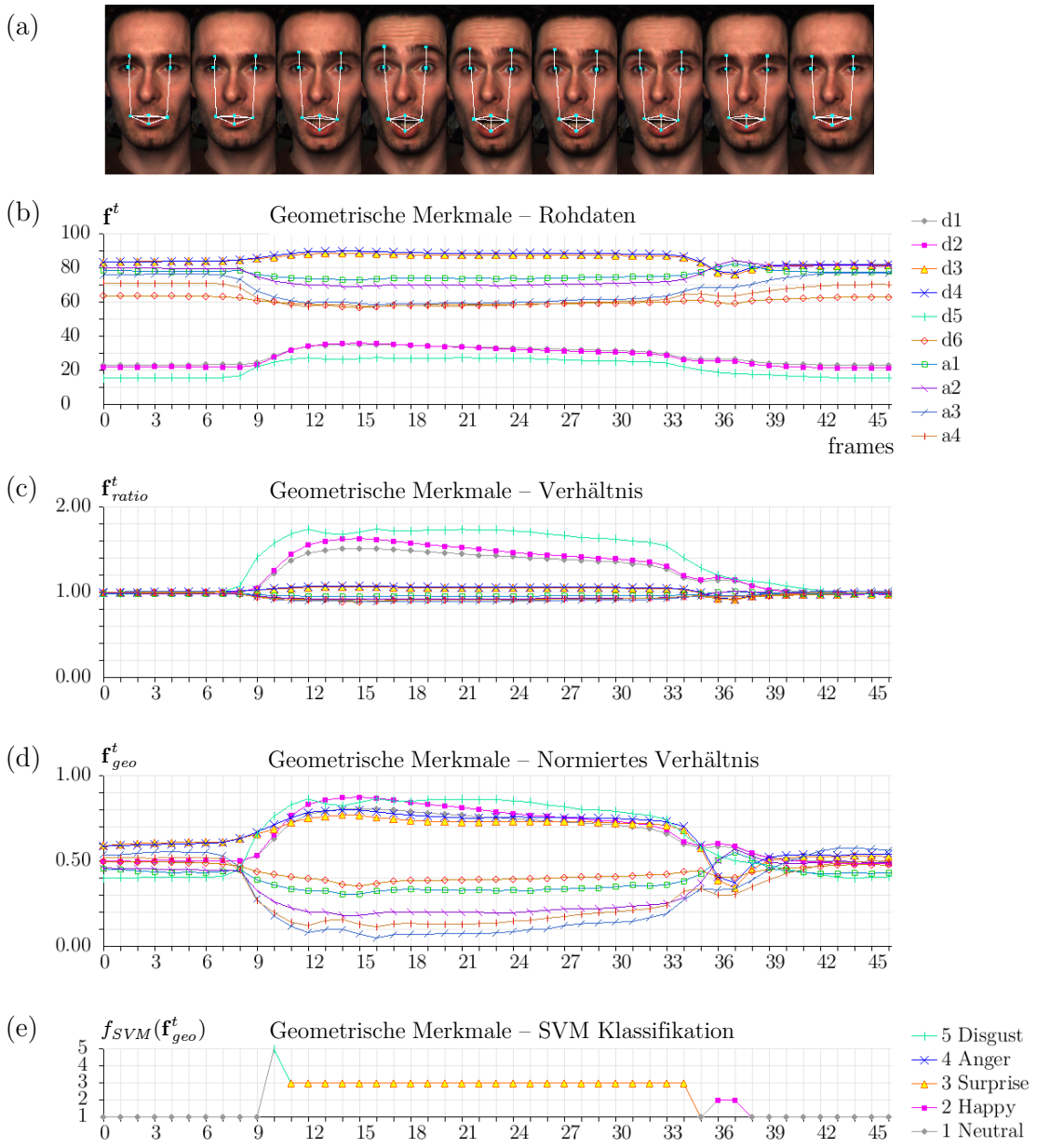


Abbildung 5-19: Beispielsequenz für die Erfassung geometrischer Merkmale nach dem Ansatz zur Merkmalsnormierung, (a) Projektion der gemessenen Rohmerkmale auf das Bild, (b) Vektor $\mathbf{f}^t$ der Rohmerkmale, bestehend aus euklidischen Abständen d_i und Winkeln α_j, (c) Verhältnis der Merkmalsvektoren zwischen neutraler und aktueller Mimik, (d) Normierung des Merkmalsvektors $\mathbf{f}^t_{geo}$, (e) zugehörige Klassifikation mittels SVM.

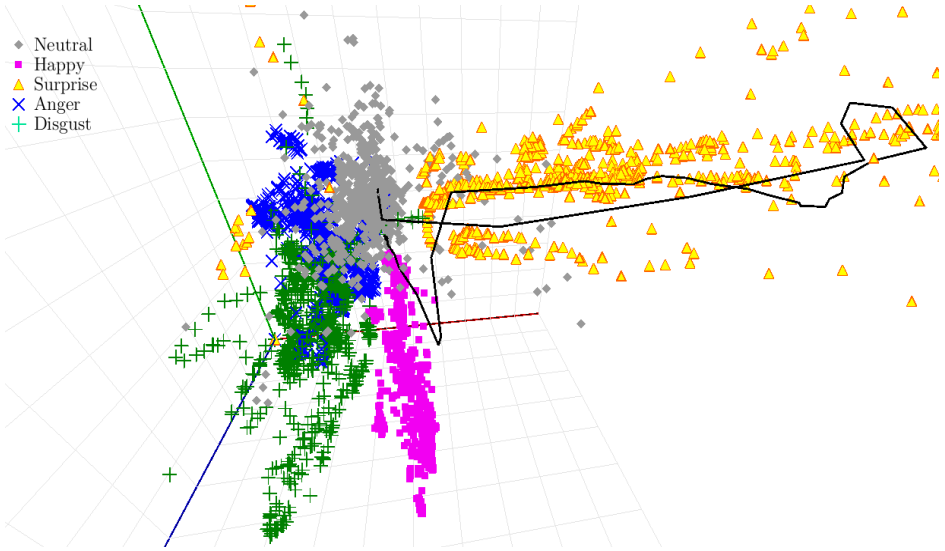


Abbildung 5-20: Pfad des Merkmalsvektors $\mathbf{f}_{geo}^{t}$ durch einen Unterraum des Merkmalsraumes (entspricht Beispiel aus Abbildung 5-19).

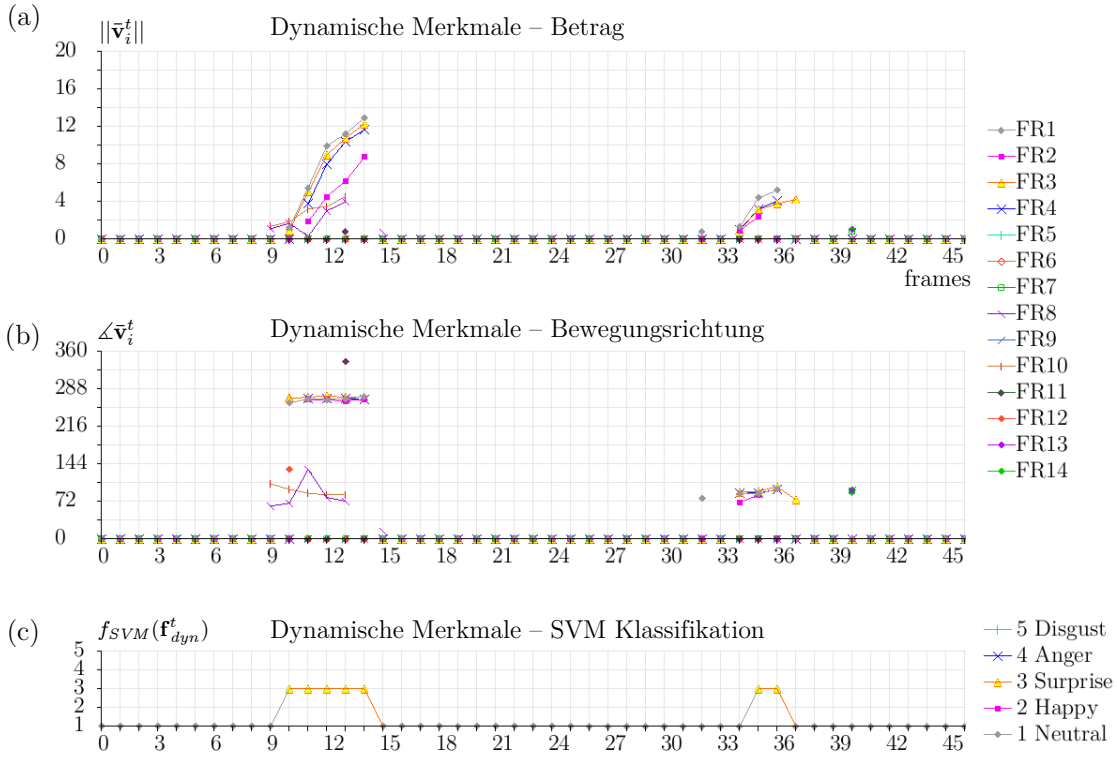


Abbildung 5-21: Erfassung dynamischer Merkmale (Beispiel aus Abbildung 5-19), (a, b) Länge und Richtung aller Verschiebungsvektoren $\bar{\mathbf{v}}_i^t$ für alle 14 Flussregionen $\mathrm{FR_i}$ (c) Klassifikation des dynamischen Merkmalsvektors.

5.3 Integration geometrischer und dynamischer Merkmale

Geometrische und dynamische Merkmale weisen jeweils vor- und nachteilige Eigenschaften auf, welche sich in der Gesamtheit jedoch grundsätzlich durch den vorgeschlagenen Integrationsansatz kompensieren lassen.

Zu den wichtigsten Vor- und Nachteilen der geometrischen Merkmale gehören:

- Kontinuierliche Merkmalsgewinnung möglich, keine Dynamik der Mimik erforderlich
- Kompensation globaler Kopfbewegung bzw. Orientierung nicht erforderlich
- Geringere Sensitivität des Merkmalsvektors aufgrund der kleineren Zahl an Messwerten, d.h. geringe Zahl an Merkmalspunkten im Gesicht

Vor- und Nachteile der dynamischen Merkmale:

- hohe Sensitivität aufgrund regionenbasierter Bewegungserfassung mit großer Anzahl an Messwerten bei Berechnung des Optischen Flusses
- Detektion mimikspezifischer Bewegungsinformation im gesamten Gesicht
- Keine Möglichkeit zur Merkmalsgewinnung ohne messbare Dynamik

Aufgrund der kleineren Anzahl an Beobachtungen, die den geometrischen Merkmalen zugrunde liegen, sind diese weniger sensitiv während einer Änderung. Das bedeutet, dass bei einem Wechsel des Gesichtsausdrucks, geometrische Merkmale aus der Sicht des Klassifikators zu Beginn ein falsches Bild der aktuellen Mimik wiedergeben können.
Die hohe Sensitivität dynamischer Merkmale ermöglicht hingegen eine frühzeitige Detektion der Dynamik und damit eine deutlich schnellere Erkennung der Mimik. Gleichzeitig gilt jedoch, dass sich dynamische Merkmale naturgemäß nur während einer Änderung erfassen lassen, geometrische Merkmale hingegen jederzeit.
Eine Möglichkeit zur Vereinigung der Vorteile beider Verfahren besteht in der Fusion der Klassifikationsergebnisse basierend auf geometrischen und dynamischen Merkmalsvektoren. Die Grundlagen hierzu liefert das Forschungsgebiet der Informationsfusion, welches die Theorie und Methoden der Erfassung, Verarbeitung und Integration von Daten aus verschieden Informationsquellen behandelt. Dabei ist es das Ziel der Fusion eine Kombination der Einzelinformationen zu einer integrierten und möglichst kohärenten Gesamtdarstellung zu erreichen. Generell kann dabei die betrachtete Information von der Sensor- und Signalebene bis zur Symbolebene reichen, welche das Klassifikationsergebnis repräsentiert. Das primäre Anliegen der Informationsfusion ist es, die Performanz des fusionierten

Modells hinsichtlich seiner Genauigkeit, Robustheit und Effizienz gegenüber den einzelnen Informationsquellen zu erhöhen [Rus07].
Generell finden sich Informationsfusionsansätze in zahlreichen Anwendungen, wie z.B. in der Medizin, für Mensch-Maschine-Schnittstellen, Fahrerassistenzsystemen, mobilen autonomen Systemen und bei der audio-visuellen Emotionserkennung [Man09, Wim08]. Durch die rasante Entwicklung der Sensorik und den zunehmenden Automatisierungsgrad, sowohl in der Industrie und Forschung als auch im Alltag wächst die Bedeutung der Informationsfusion, weshalb sich diese in den letzten Jahren zu einem wichtigen Zweig der Mustererkennung entwickelt hat. Dabei hat sich eine Reihe von Techniken zur Fusion unterschiedlicher Informationsquellen etabliert.

- Komplementäre Integration: Die Fusion verschiedenartiger Daten mit gleicher Nutzinformation wird zum Schließen von Informationslücken verwendet, z.B. Erfassung desselben Ereignisses durch verschiedene Sensoren.
- Konkurrierende Integration: Hierbei ist es das Ziel gleichartige Daten mit gleicher Nutzinformation zu fusionieren, um Unsicherheiten zu kompensieren, z.B. Mittelung von Bildern, die unter ähnlichen Bedingungen aufgenommen wurden.
- Kooperative Integration: Hierbei liegt die Nutzinformation über die zu erfassende Messgröße verteilt vor. Für eine vollständige Erfassung, müssen die Daten aller Informationsquellen aufgenommen, ausgewertet und integriert werden. Es handelt sich hierbei nicht um eine Fusion im engeren Sinne, sondern um eine Datenintegration. Ein Beispiel hierfür ist die die Integration von Stereobildern zur Bestimmung von Tiefenkarten.

Eine Kombination lässt sich auf verschiedenen Abstraktionsebenen durchführen.

- Frühe Fusion (Sensordatenfusion): Die Sensordaten werden auf der Signalebene direkt kombiniert, z.B. durch Zusammenfügen der Messvektoren. Voraussetzung hierfür ist die Synchronisation der Signale sowie eine Vergleichbarkeit der Daten durch geeignete Maße.
- Merkmalsfusion: Extrahierte Merkmale werden auf Merkmalsebene fusioniert, falls sich die Daten auf Signalebene nicht kohärent integrieren lassen.
- Späte Fusion (Entscheidungsfusion): Auf der Basis extrahierter Merkmale, welche separaten Entscheidern zugeführt werden, erfolgt auf der Symbolebene eine Fusion der Ergebnisse der Objekterkennung in Form von Klas-

senlabels mit zugeordneter Gewichtung, z.B. auf der Grundlage von a posteriori Wahrscheinlichkeitsschätzungen.

Bei der Integration geometrischer und dynamischer Merkmale ist sowohl die Sensordatenfusion als auch eine Merkmalsfusion nicht bzw. weniger geeignet, da zum einen die erfassten Rohdaten jeweils spezielle Berechnungen zur Korrektur bzw. Nachbehandlung erfordern (z.B. Nullposetransformation, Normierung) und zum anderen die beiden Merkmalsarten, aufgrund der kurzen Erscheinungsdauer dynamischer Merkmale, in ihrer Anzahl beträchtlich variieren. Es wird daher in dieser Arbeit eine Entscheidungsfusion vorgeschlagen, bei der die Gewichtung auf der Grundlage der Aktivierungsfunktion $v_{sum}(t)$ (4.16) dynamischer Merkmale realisiert wird.
Es werden hierzu zwei verschiedene Mimikzustände angenommen, welche im Folgenden als Applikations- und Aktivierungsphasen bezeichnet werden (Abbildung 5-22). Während der Aktivierung findet ein Wechsel der aktuellen Mimik statt, während der Applikation bleibt sie hingegen konstant.
Durch die Integration wird in Zeitabschnitten, in denen mittels dynamischer Merkmale deutliche Veränderungen erfasst werden, die die Aktivierungsschwelle v_{min} überschreiten, das Klassifikationsergebnis $f_{\mathrm{CLASS}}(\mathbf{f}^t_{dyn})$ bevorzugt. Somit erfolgt nach (5.23) beim Wechsel des Gesichtsausdrucks die Klassifikation auf der Grundlage dynamischer, sonst jedoch durch die Auswertung geometrischer Merkmale.

$$f_{\mathrm{CLASS}}(\mathbf{f}^t_{geo}, \mathbf{f}^t_{dyn}) = \alpha_1 \cdot f_{\mathrm{CLASS}}(\mathbf{f}^t_{geo}) + \alpha_2 \cdot f_{\mathrm{CLASS}}(\mathbf{f}^t_{dyn}), \tag{5.23}$$

$$\alpha_1 = \begin{cases} 1, \text{falls } v_{sum}(t) < v_{min}, \\ 0, \text{ sonst} \end{cases} \tag{5.24}$$

$$\alpha_2 = 1 - \alpha_1$$

mit f_{CLASS} als beliebiger Klassifikator (SVM, MLP, k-NN, etc.) für einen geometrischen $\mathbf{f}^t_{geo}$ bzw. dynamischen Merkmalsvektor $\mathbf{f}^t_{dyn}$.

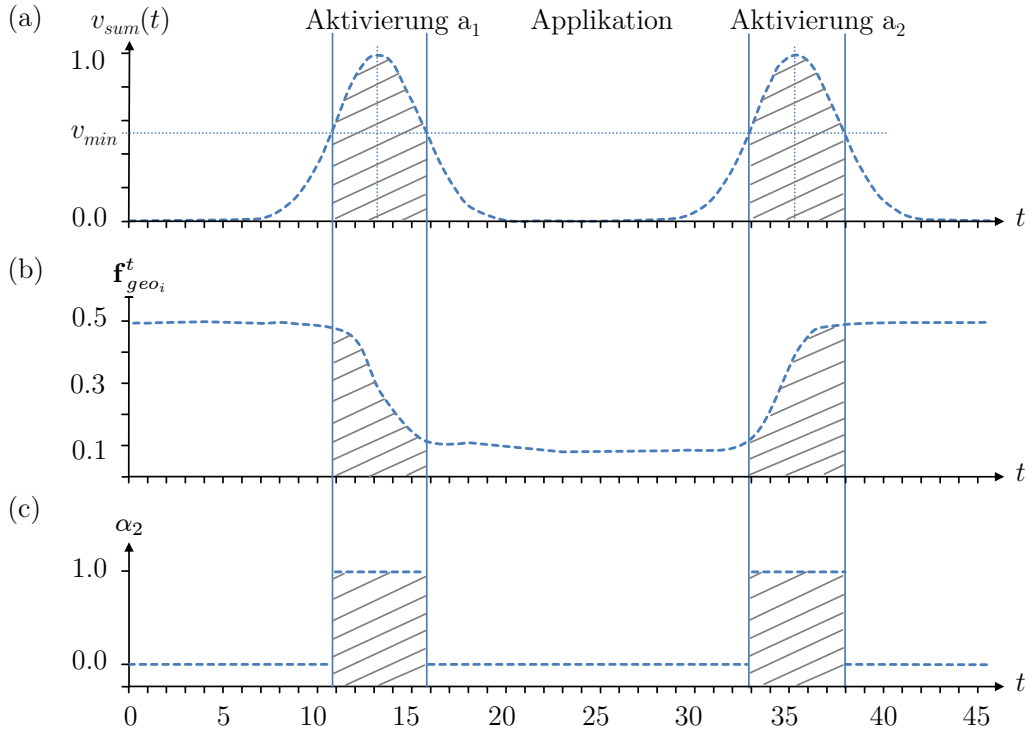


Abbildung 5-22: Prinzip der Fusion geometrischer und dynamischer Merkmale am Beispiel von Aktivierungs- und Applikationsphasen, (a) Aktivierungsfunktion $v_{sum}(t)$ für dynamische Merkmale, (b) gleichzeitiges geometrisches Merkmal $\mathbf{f}^t_{geo_i}$, (c) Gewicht α_i entscheidet über die Fusion der Klassifikationsergebnisse.

Empirische Untersuchungen haben gezeigt, dass auf der Grundlage der Fusion nach (5.23) unter gegebenen Randbedingungen eine deutliche Qualitätssteigerung gegenüber der einfachen Klassifikation durch geometrische Merkmale erzielt werden kann (Abschnitt 6.5, Beispiel in Abbildung 5-23).
Zu diesen Randbedingungen gehört, dass keine durch Sprechen verursachte Mimik vorliegt, da dies in der Mundregion unweigerlich zu schwer interpretierbarer Bewegungsinformation führt. Folglich wird eine Klassifikation auf der Grundlage dynamischer Merkmale erschwert bzw. verschlechtert.
In Untersuchungen hat sich des Weiteren gezeigt, dass für Mimikklassen, die durch einen geringen Erregungsgrad geprägt sind, wie z.B. Trauer, nur sehr wenig Bewegungsinformation erfasst werden kann und daher keine Verbesserung des Klassifikationsergebnisses durch Hinzuziehen dynamischer Merkmale möglich ist.

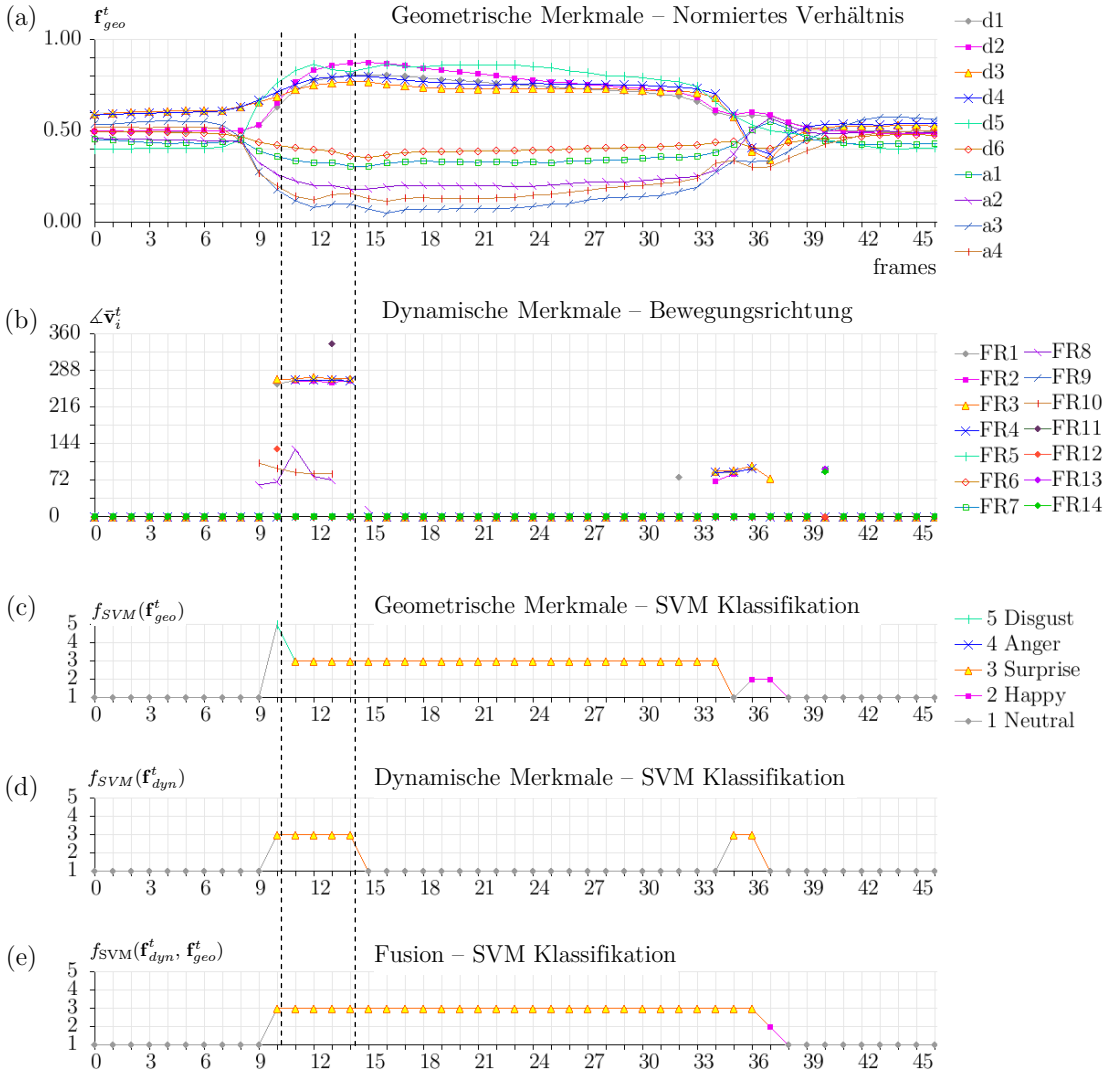


Abbildung 5-23: Beispiel zur Fusion durch Nutzung geometrischer und dynamischer Merkmale, (a) Sequenz geometrischer Merkmale $\mathbf{f}^t_{geo}$, (b) dynamische Merkmale, (c) SVM Klassifikation geometrisch ($f_{\mathrm{SVM}}(\mathbf{f}^t_{geo})$) mit Störung des Klassifikators im Zeitfenster der Mimikänderung, (d) Klassifikation dynamisch ($f_{\mathrm{SVM}}(\mathbf{f}^t_{dyn})$), (e) Verbesserung des Klassifikationsergebnisses durch Fusion nach (5.23).

Kapitel 6

Experimentelle Untersuchungen - Validierung der Systemstruktur

In diesem Kapitel werden Ergebnisse aus der Evaluation dreier Datenbanken mit emotionstypischer Mimik vorgestellt. Eine der Datenbanken wurde hierzu selbst aufgezeichnet. Die Resultate werden im Hinblick auf die in Abschnitt 1.1 gestellten Fragen diskutiert. Entsprechend der Übersicht in Abbildung 6-1 wurde hierzu eine Reihe von Untersuchungen, für die auf der Grundlage der vorgeschlagenen Systemstruktur erfassten geometrischen und dynamischen Merkmale, durchgeführt. Zur besseren Übersicht werden die beiden Merkmalstypen in Anhang 8.2 noch einmal aufgeführt. In den experimentellen Untersuchungen wurden die

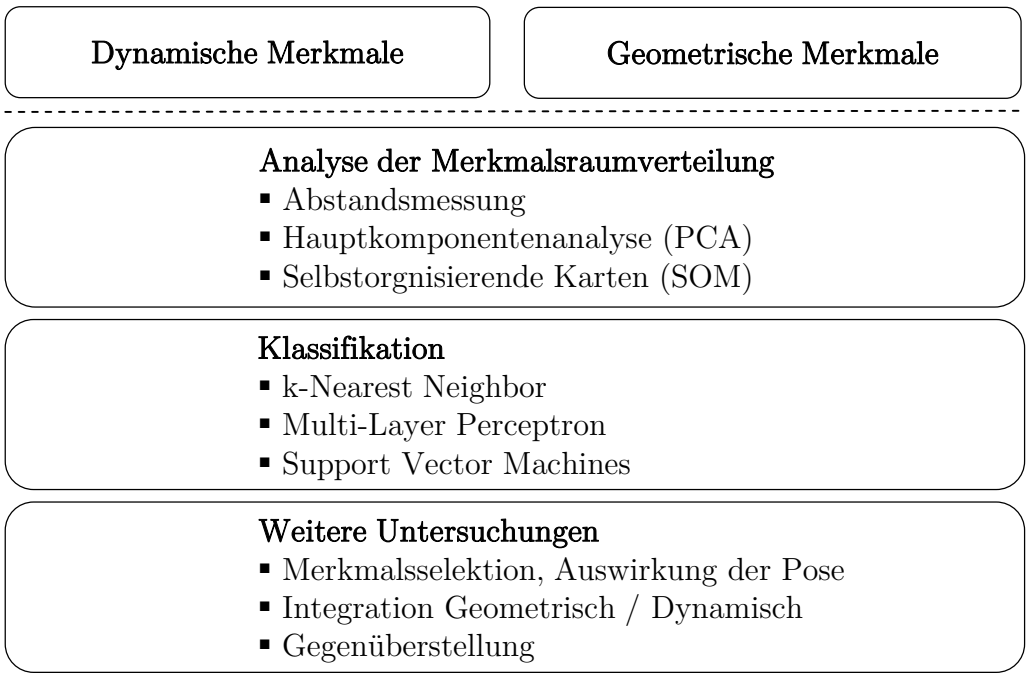


Abbildung 6-1: Übersicht der durchgeführten experimentellen Untersuchungen.

hochdimensionalen Merkmalsräume analysiert und deren Eignung zur Mustererkennung nachgewiesen. In den Ergebnissen wird der Informationsgehalt der Merkmale dargestellt und es wird auf Korrelationen der erfassten Daten und die Merkmalsselektion eingegangen. Die Untersuchungen haben dabei unter anderem die Personenunabhängigkeit der erfassten Merkmale gezeigt.
Die Bewertung der entwickelten Verfahren beruht auf der Berechnung der Erkennungsraten, welche mit Hilfe sogenannter Konfusionsmatrizen dargestellt werden. Die Ergebnisse werden dabei separat für die geometrischen und dynamischen Merkmale sowie deren Fusion vorgestellt (Abschnitt 6.5). Weiterführende Klassifikationsergebnisse werden in Anhang 8.1 gegeben.

Die im Rahmen dieser Arbeit entwickelten Verfahren erfordern die Aufzeichnung spezieller Daten, beispielsweise Parameter zur Kamerakalibrierung, 3D Gesichtsmodelle, Stereobilder aber auch Farbinformation, was die Nutzung fremder Datenbanken grundsätzlich einschränkt. Die Evaluation der vorgeschlagenen Systemstruktur zur Mimikanalyse erfolgte daher zum Teil auf der Grundlage hauseigener Datenbanken DB_{MD} und DB_{ULM} sowie durch Auswertung der Binghamton University BU-4DFE Database, kurz DB_{BU} [Yin08]. Die internen Datenbanken wurden in den Laboren des IESK aufgezeichnet bzw. durch die Arbeitsgruppe Prof. Traue der Universität Ulm freundlicherweise bereitgestellt. Jede dieser Datenbanken enthält vorrangig Frontalaufnahmen. Die Abhängigkeit der Merkmale und Klassifikationsergebnisse von der Pose wird in Abschnitt 6.4 erörtert.
Die Datenbank DB_{MD} enthält Bildfolgen von zwanzig männlichen Probanden. Es werden die Klassen Neutral (C_1), Freude (C_2), Überraschung (C_3), Wut (C_4) und Ekel (C_5) berücksichtigt, dabei handelt es sich um gespielte Mimik. Die Klassen Trauer und Angst sind aufgrund der schwierigen Auslösung und Darstellung nicht in der Datenbank enthalten. Für die Klasse Neutral wird nicht emotionstypische Mimik mit unspezifischen Gesichtsausdrücken verwendet. Insgesamt beinhaltet die Datenbank circa 7000 Einzelaufnahmen.
Die Datenbank DB_{BU} enthält 3D Farbbildfolgen von 101 Probanden, davon 58 weiblich und 43 männlich, mit einer Vielzahl ethnischer Abstammungen, darunter Asiaten, Schwarze, Lateinamerikaner und Weiße. Insgesamt umfasst die Datenbank mehr als 60.000 Einzelaufnahmen, während sieben Mimikklassen C_1-C_7 berücksichtigt werden. Die präsentierten Emotionen sind auch hier nicht authentisch, sondern professionell gespielt. Die 3D Daten wurden mit einem Laserscanner und simultan einer monokularen Farbbildkamera aufgezeichnet. Da keine Information über die Kamerakalibrierung vorliegt, ist eine besondere Vorverarbei-

tung erforderlich, um das entwickelte Verfahren zur Merkmalsnormierung anzuwenden. Zu diesem Zweck wurden alle 3D Sequenzen vor der Auswertung mit Hilfe eines OpenGL Renderings mit definierter Kamera $\mathbf{K}_{\mathrm{GL}}$ in einfache Farbbildfolgen transformiert (Abbildung 6-2). Dabei wurde ein Kameramodell analog zu (3.2) verwendet, die Parameter zur Verzeichnung wurden zur Vereinfachung auf null gesetzt.

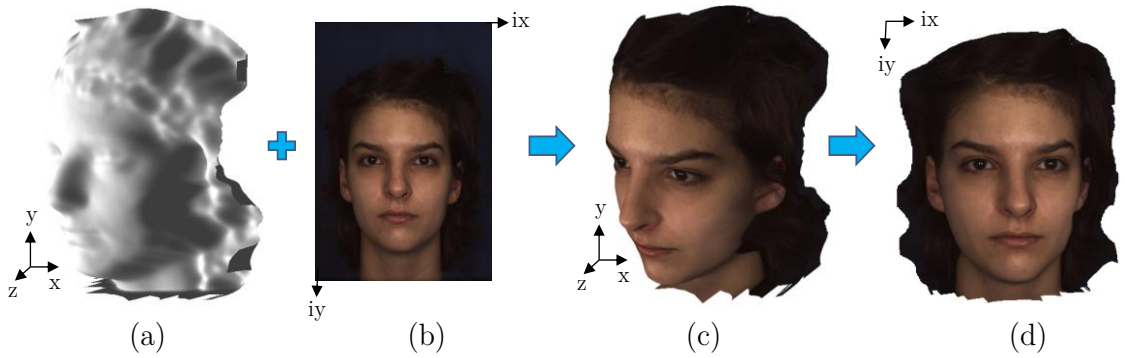


Abbildung 6-2: Erstellung von Farbbildfolgen aus der Datenbank $\mathrm{DB}_{\mathrm{BU}}$, (a) 3D Modell zum Zeitpunkt t, (b) zugehöriges Farbbild, (c) Berechnung des texturierten 3D Modells, (d) Rendering des Modells als 2D Bild mit Kamera $\mathbf{K}_{\mathrm{GL}}$.

Durch Auswertung der beiden Datenbanken $\mathrm{DB}_{\mathrm{MD}}$ und $\mathrm{DB}_{\mathrm{BU}}$ wurden auf der Grundlage der Merkmalsnormierung geometrische und dynamische Merkmale extrahiert, analysiert und zum trainieren bzw. klassifizieren verwendet. Dabei wurden stets andere Personen in den Lern- und Testdaten eingesetzt, um die Personenunabhängigkeit sicherzustellen. Die verschiedenen Verarbeitungsschritte der vorgeschlagenen Mimikerkennung laufen vollautomatisch für die Datenbank $\mathrm{DB}_{\mathrm{MD}}$. Da die BU-4DFE Datenbank Aufnahmen von Personen mit einer sehr großen Vielfalt an Hautfarben beinhaltet, erfolgt hier die erste Komponente der Merkmalsextraktion, d.h. die bildbasierte Merkmalspunktdetektion nicht vollautomatisch, sondern wird manuell unterstützt. Auf diese Weise wird sichergestellt, dass die Funktionalität und Qualität der vorgeschlagenen Systemkette bestätigt werden kann, unabhängig vom Schritt der Punktdetektion, welche ein eigenes Forschungsgebiet darstellt. Dies auch gilt für Besonderheiten wie Brille, Bart etc.
Zur Prüfung der Erkennungsleistung des Gesichtsnormierungsansatzes wird die Datenbank $\mathrm{DB}_{\mathrm{ULM}}$ ausgewertet, welche Sequenzen normierter Gesichter von 42 Probanden enthält (Abschnitt 6.2.3). Es werden dabei sechs Mimikklassen C_2-C_7 berücksichtigt.

6.1 Auswertung geometrischer Merkmale

Im Folgenden werden die Ergebnisse der Auswertung geometrischer Merkmale dargestellt. Dabei wird sowohl auf die Verteilung der Klassen in den Merkmalsräumen als auch auf die erreichten Klassifikationsergebnisse eingegangen.

6.1.1 Analyse der Merkmalsräume - geometrische Merkmale

Eine Untersuchung der Verteilung der erfassten Daten in den hochdimensionalen Merkmalsräumen ist zur Interpretation der Klassifikationsergebnisse entscheidend und damit auch zur Erklärung möglicher Fehler. Des Weiteren kann durch eine Analyse extrahierter Merkmale der Beitrag und die Bedeutung für die Erkennung nachgewiesen werden.
Insbesondere wurden hierzu die auf der Grundlage der Datenbanken $\mathrm{DB_{MD}}$ und $\mathrm{DB_{BU}}$ erfassten und normierten geometrischen Merkmale (s. Abschnitt 5.2.2) betrachtet. Zur Veranschaulichung der Verteilungen in den zehn- und höher dimensionalen Räumen wurden verschiedene Analysen durchgeführt, welche zum Teil auf eine gute, mäßige bzw. schlechte Trennung der Klassen hinweisen.

6.1.1.1 Merkmalsraumanalyse – Abstände der geometrischen Merkmale

In der ersten Untersuchung wurde der euklidische Abstand $||\mathbf{f}^{t}_{geo} - \mu^{\mathrm{C}_i}_{geo}||$ jedes Merkmalsvektors $\mathbf{f}^{t}_{geo} \in \mathbb{R}^{10}$ (4.17) zum Merkmalsmittelwert $\mu^{\mathrm{C}_i}_{geo}$ (6.1) der jeweiligen Klasse C_i bestimmt. Die Berechnung des euklidischen Abstands bietet sich an, um zu zeigen wie stark verschiedene Klassen voneinander separiert sind bzw. welche Klassen näher zueinander liegen. Dies ist ein Indiz für häufigere Verwechslungen bei der Klassifikation.

$$\mu^{\mathrm{C}_i}_{geo} = \tfrac{1}{n_{\mathrm{C}_i}} \sum_{j=1}^{n_{\mathrm{C}_i}} \mathbf{f}^{j,\,\mathrm{C}_i}_{geo}, \quad \mu^{\mathrm{C}_i}_{geo} \in \mathbb{R}^{10} \tag{6.1}$$

mit n_{C_i} als Anzahl aller Samples $\mathbf{f}^{j}_{geo}$, die zur Klasse C_i gehören.

Für die ermittelten geometrischen Merkmale der beiden Datenbanken $\mathrm{DB_{MD}}$ und $\mathrm{DB_{BU}}$ werden in Abbildung 6-3 und Abbildung 6-4 die gemessenen Abstände zwischen den nach Klassenzugehörigkeit sortierten Trainingsdaten und dem Schwerpunkt der jeweiligen Klasse dargestellt. Für alle betrachteten Klassen C_i der Datenbank $\mathrm{DB_{MD}}$, mit Ausnahme von C_5 (Ekel) ist dabei eine klare Abgrenzung zu den anderen Klassen zu erkennen, d.h. diese liegen ihrem jeweiligen Schwerpunkt $\mu^{\mathrm{C}_i}_{geo}$ deutlich näher als der Rest. Bei der Klasse C_5 liegen die Daten der Klasse

C_4 dem Schwerpunkt $\mu_{geo}^{C_5}$ ebenfalls recht nahe, weshalb Klasse C_4 ein Kandidat für Verwechslungen bei der Klassifikation ist (Abschnitt 6.1.2).
Die Auswertung der Datenbank DB_{BU} zeigt ebenfalls eine deutliche Separation der geometrischen Merkmale für die Klassen C_1 bis C_3, jedoch eine gewisse Nähe von C_7 zu C_4 sowie von C_4 zu C_5. Die Klassen C_6 und C_7 haben die schlechteste Trennung gegenüber den Merkmalen der anderen Klassen. Besonders die Daten der Klasse C_7 liegen C_6 nahe. Weiterhin zeigen die Klassen C_1 und C_4 räumliche Nähe zu C_7. Somit sind für die Klassen C_6 und C_7 die schlechtesten Klassifikationsergebnisse zu erwarten. Diese Vermutung wird durch die Berechnung der Konfusionsmatrizen bestätigt.

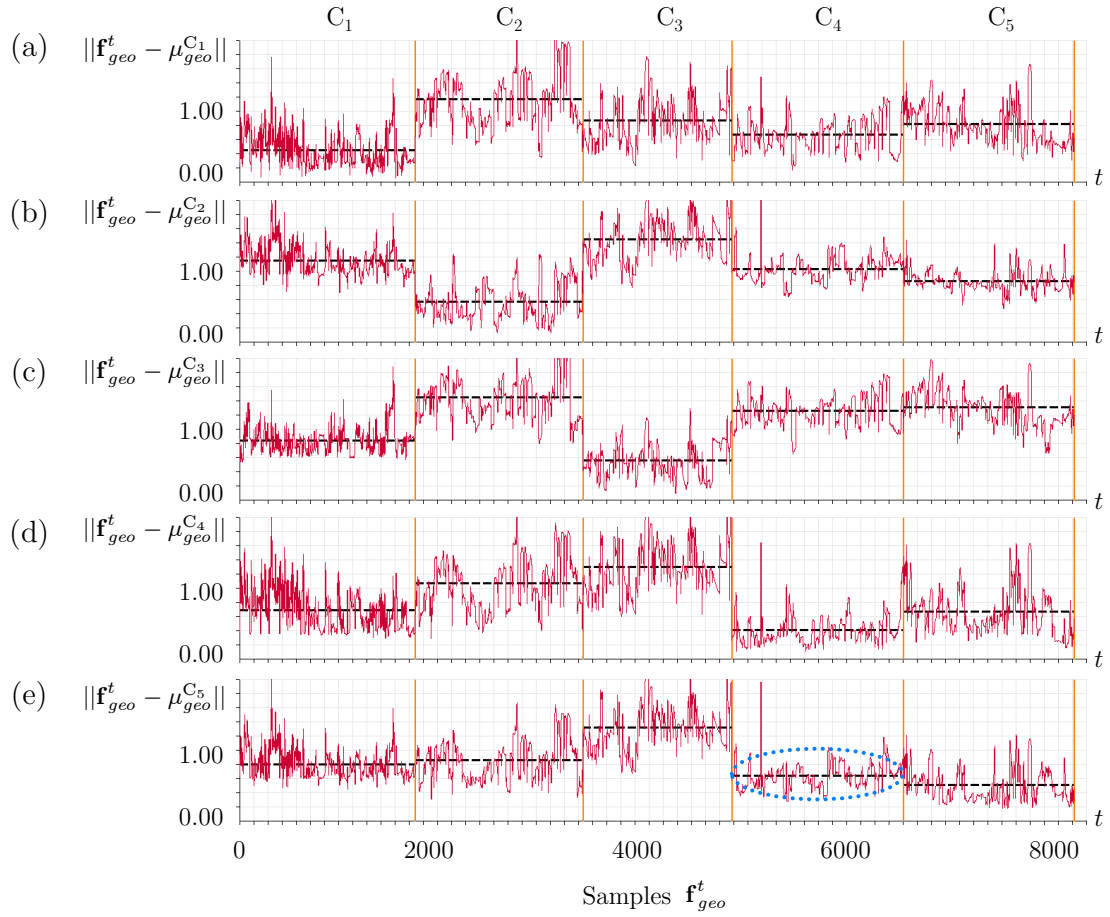


Abbildung 6-3: Analyse geometrischer Merkmale für DB_{MD}. Die gestrichelte Linie zeigt den mittleren euklidischen Abstand der Merkmalsvektoren verschiedener Klassen zum jeweiligen Zentrum im 10 dimensionalen Merkmalsraum, (a-d) die Klassen $C_{1\text{-}4}$ (Neutral, Freude, Überraschung, Wut) haben eine klare Abgrenzung zu den anderen Klas-

sen, (e) die Merkmale der Klasse C_4 liegen dem Zentrum der Klasse C_5 (Ekel) nahe, was Verwechslungen wahrscheinlich macht (blaue Ellipse).

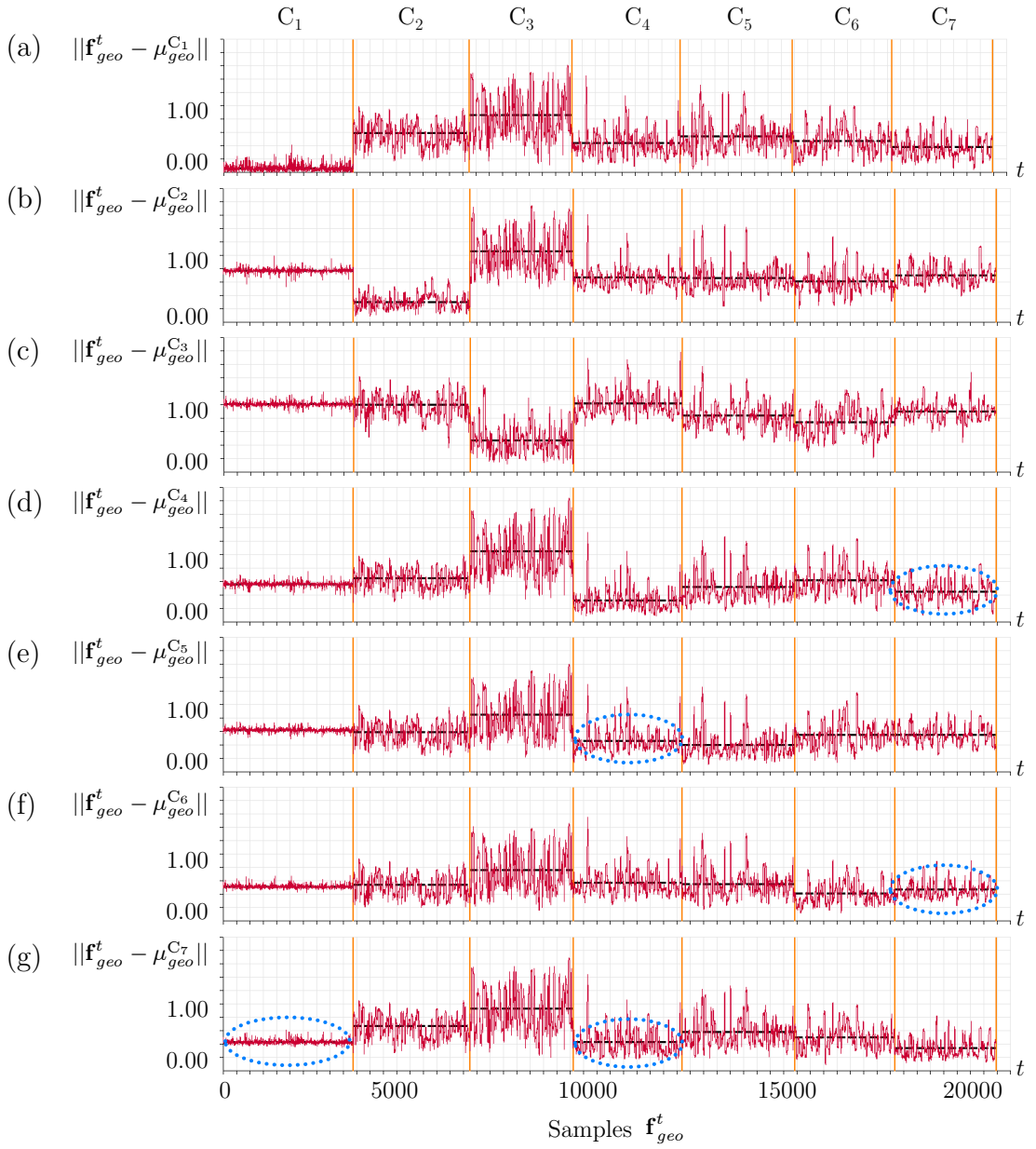


Abbildung 6-4: Analyse geometrischer Merkmale der Datenbank DB_{BU}, (a-c) eine deutliche Abgrenzung ist für die Klassen $C_{1\text{-}3}$ zu erkennen, (d) Klasse C_7 liegt C_4 nahe, (e) Klasse C_4 liegt nahe an C_5, (f) Klasse C_7 liegt C_6 nahe, (g) die Klassen C_1 und C_4 zeigen räumliche Nähe zu C_7.

6.1.1.2 Merkmalsraumanalyse – PCA der geometrischen Merkmale

Eine weitere Möglichkeit zur Veranschaulichung und Untersuchung des Merkmalsraumes besteht in der Dimensionsreduktion mittels Hauptkomponentenanalase (Principal Component Analysis, PCA). Das Ziel dieser Untersuchung besteht darin lineare Abhängigkeiten im Merkmalsraum festzustellen und diesen in eine neue niedriger dimensionale Darstellung zu überführen, in welcher der überwiegende Teil der enthaltenen Information durch eine kleinere Anzahl unkorrelierter Variablen repräsentiert wird. Bei der PCA wird das ursprünglich p dimensionale Koordinatensystem in ein q dimensionales ($q \leq p$) überführt, dessen Achsen (Hauptkomponenten) in die Richtungen der q größten Varianzen zeigen.
Die Untersuchung des Merkmalsraums mittels Hauptkomponentenanalyse zeigt, dass den zehn geometrischen Merkmalen starke lineare Abhängigkeiten zugrunde liegen. Für die Datenbanken DB_{MD} und DB_{BU} mit fünf bzw. sieben zugehörigen Klassen, decken die ersten drei Hauptkomponenten K_1, K_2, K_3 des zehndimensionalen Merkmalsraums mehr als 86 bzw. 75 Prozent der Gesamtvarianz ab.
In Abbildung 6-5 und Abbildung 6-6 werden die ersten drei Hauptkomponenten der transformierten Merkmalsräume dargestellt. Für die Daten der DB_{MD} wird deutlich, dass sich die Klassen C_1 bis C_4 klar abgrenzen, für C_5 jedoch Überlappungen mit C_4 bestehen.

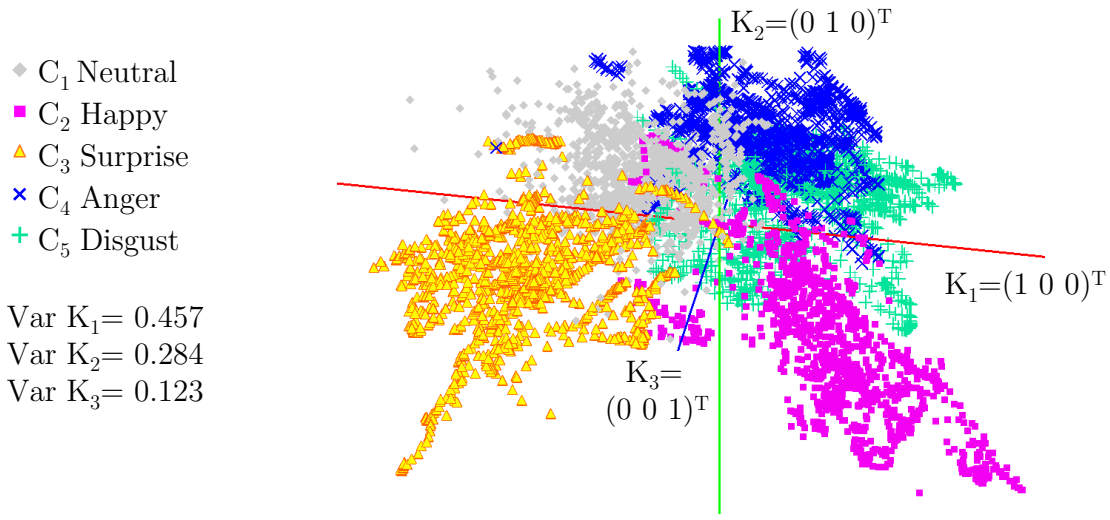


Abbildung 6-5: Dimensionsreduktion mittels PCA. Die Darstellung der ersten drei Hauptkomponenten K_i für die geometrischen Merkmale der Datenbank DB_{MD} zeigt bis auf Überschneidungen der Klassen C_5 mit C_2 und C_4 eine deutliche Klassentrennung.

Bei Betrachtung der ersten drei Hauptkomponenten des transformierten Merkmalsraumes der Datenbank DB_{BU} zeigt sich eine klare Trennung für die Klassen C_1 und C_2, für die Klasse C_3 (Überraschung) besteht jedoch eine Überlappung mit C_6 (Angst). Bei der Untersuchung des zugrunde liegenden Bildmaterials wurde ersichtlich, dass diese Überlappung durchweg aus einem überraschten Gesichtsausdruck resultiert, welcher von den Probanden als Angst präsentiert wurde. Bei der Inspektion des Merkmalsraumes werden weitere Überlappungen deutlich, etwa von C_7 mit C_4 sowie von C_5 und C_6 mit verschiedenen Klassen, was auf mögliche Ungenauigkeiten bei der Klassifikation hinweist.

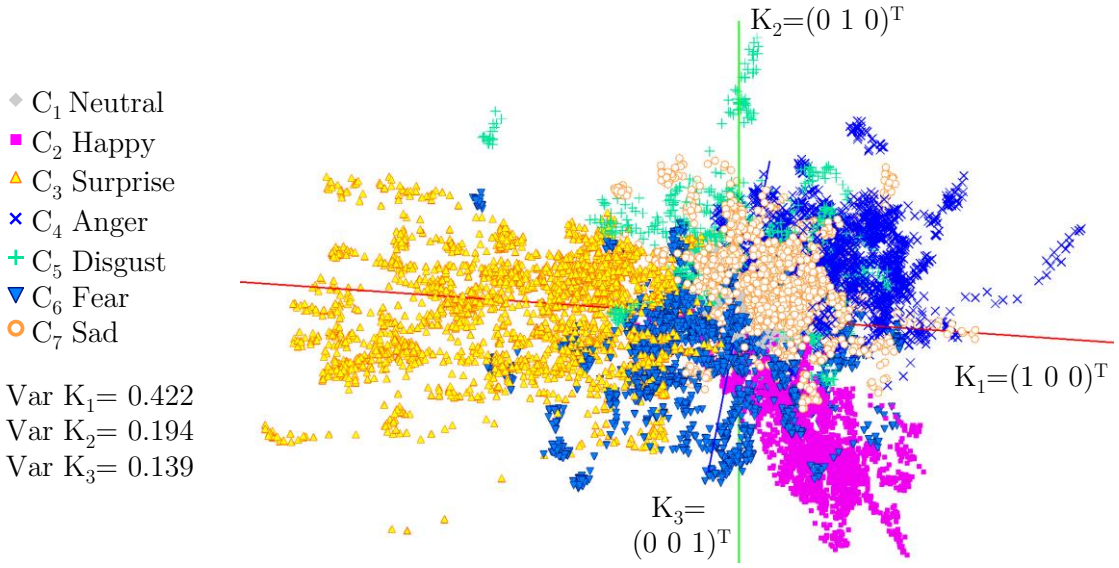


Abbildung 6-6: Dimensionsreduktion mittels PCA. Die Darstellung der ersten drei Hauptkomponenten K_i der geometrischen Merkmale bezüglich Datenbank DB_{BU} zeigt neben einer deutlichen Clusterbildung und Abgrenzung der Klassen C_2 und C_3 auch Überlappungen, z.B. von C_7 mit C_4, von C_6 mit C_3 und von C_5 mit weiteren Klassen.

6.1.1.3 Merkmalsraumanalyse – SOM für geometrische Merkmale

Beim dritten Verfahren zur Untersuchung des Merkmalsraumes wurden die geometrischen Merkmale zum Anlernen einer selbstorganisierenden Karte (SOM) verwendet (s. Abschnitt 3.4.4). Die in einem unüberwachten Lernverfahren ermittelte zweidimensionale SOM gibt grundsätzlich Aufschluss über eine den Daten zugrundeliegende Klasseneinteilung. Abbildung 6-7 stellt die U-Matrix Repräsen-

tation der SOM dar, welche für die geometrischen Merkmale der Datenbank DB_{MD} ermittelt wurde. Dabei zeigt sich die Bildung verschiedener Cluster, welche durch homogene helle Bereiche auffallen sowie Clustergrenzen, die als dunkle Begrenzungslinien erscheinen. Durch die zusätzliche Projektion der Lerndaten, von denen die Klassenzugehörigkeit als Grundwahrheit bekannt ist, wird ersichtlich, wie die Merkmalsvektoren der verschiedenen Klassen in einzelne Häufungsbereiche separiert werden.

In Abbildung 6-7 und Abbildung 6-8 wird jeweils die U-Matrix für die Merkmale der Datenbanken DB_{MD} und DB_{BU} dargestellt. Diese lassen partiell eine sehr deutliche Clusterbildung erkennen, was auf hohe Erkennungsraten bei der Klassifikation schließen lässt. Die meiste Verwechslung ist nach der visuellen Inspektion der SOM für die Klassen C_6 und C_7 zu erwarten, da diese keine auffällige Clusterbildung zeigen.

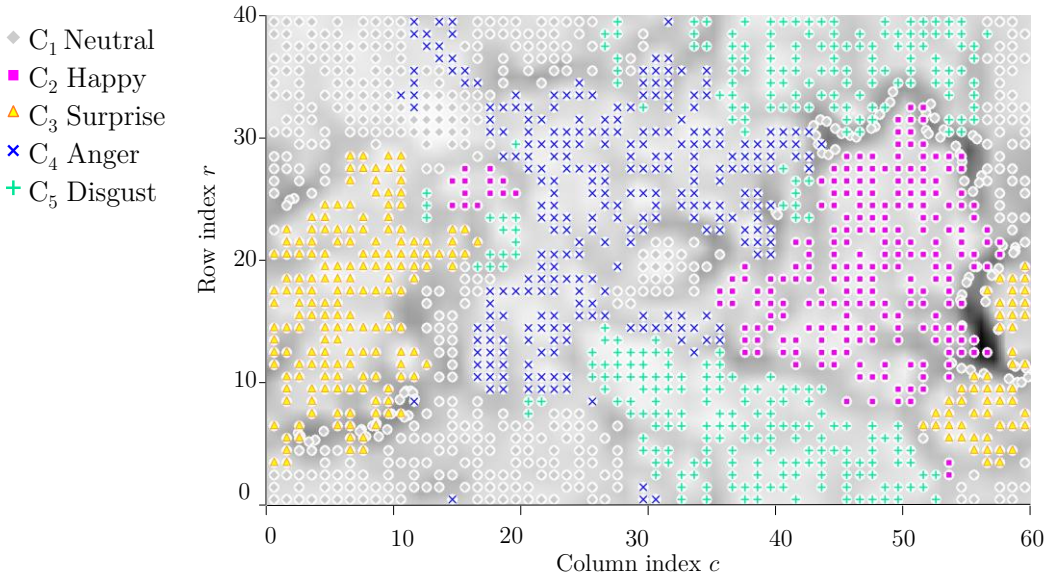


Abbildung 6-7: SOM, Darstellung der U-Matrix für die geometrischen Merkmale der Datenbank DB_{MD}, Häufungsbereiche für die projizierten Trainingsvektoren sind sehr ausgeprägt, Überlappungen hingegen kaum vorhanden.

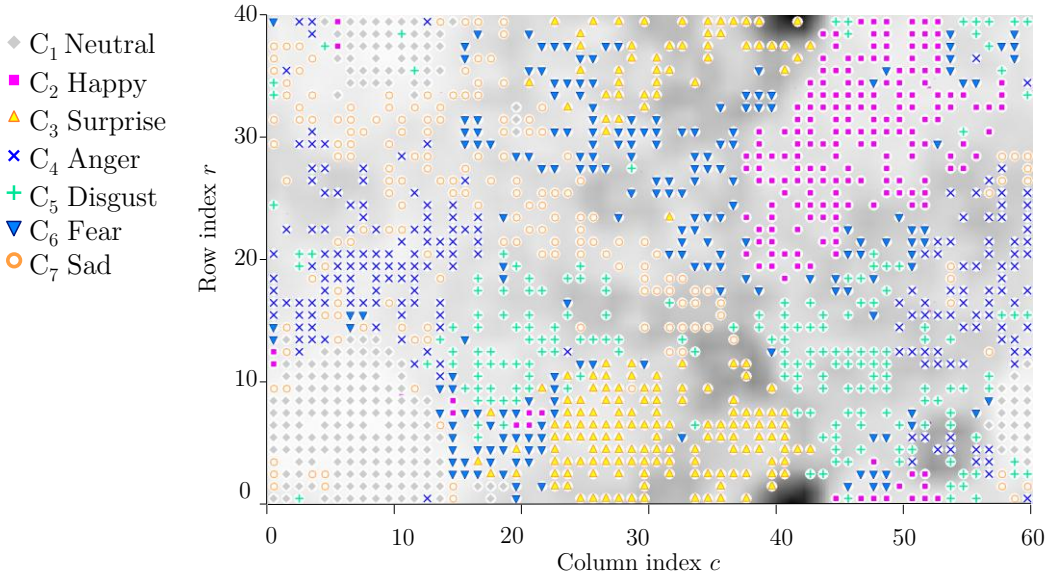


Abbildung 6-8: SOM, Darstellung der U-Matrix für die geometrischen Merkmale der Datenbank DB_{BU}, für die projizierten Trainingsvektoren sind Häufungsbereiche unterschiedlich deutlich zu erkennen, die kompaktesten Cluster sind für die Klassen C_1 bis C_3 zu sehen. Die anderen Klassen zeigen verschiedene Überlappungen.

6.1.2 Klassifikation auf der Grundlage geometrischer Merkmale

Die Bewertung der Qualität des Verfahrens zur Mimikerkennung beruht auf der Berechnung der Erkennungsraten, welche durch Auswertung von Testdaten mit bekannter Grundwahrheit ermittelt werden. Es werden dabei maschinelle Lernverfahren zur Klassifikation eingesetzt, der Fokus liegt hier auf k-NN, MLP und SVM Klassifikatoren (Abschnitt 3.4), die auf einer Menge von Lerndaten antrainiert werden. Die Parametrierung der Klassifikatoren wurde empirisch durch Kreuzvalidierung (Abschnitt 3.4.5) bestimmt und wie folgt festgelegt:

- k-NN: Parameter k=5,
- MLP: Sigmoide Transferfunktion, zwei versteckte Schichten,
- SVM: RBF Kernel.

Die Untersuchung mit 10-facher stratifizierter Kreuzvalidierung zeigte dabei für beide betrachteten Datenbanken DB_{MD} und DB_{BU} eine durchschnittliche Erkennungsrate (SVM) von 93 Prozent, bei einer Unterscheidung von 5 Klassen. Bei Berücksichtigung von 7 Klassen lag diese bei circa 81.5 Prozent.
Für eine detailliertere Betrachtung werden im Folgenden separate Klassifikationsergebnisse mit Hilfe sogenannter Konfusionsmatrizen dargestellt, welche durch die

Untersuchung gleichgroßer Lern- und Test-Teilmengen MD_{s1}/MD_{s2} bzw. BU_{s1}/BU_{s2} ermittelt wurden. Dabei enthalten die verschiedenen Merkmalsmengen stets Messungen unterschiedlicher Probanden. Tabelle 6-1 zeigt die Verwechslung bei der Klassifikation geometrischer Merkmale bezüglich Datenbank DB_{MD}.
Die Zeilen repräsentieren hierbei die tatsächlichen Klassen, die Spalten die Vorhersagen durch die Klassifikatoren. Die Summe jeder Zeile ergibt dabei stets 100 Prozent, was der Gesamtheit aller möglichen Klassen entspricht.

Klasse C_i vs. **Klassifikation** $P(C_j)$		$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$
C_1 - Neutral	k-NN	86.63	0.17	8.29	4.57	0.34
	MLP	93.40	0.00	6.60	0.00	0.00
	SVM	91.03	0.00	7.61	0.85	0.51
C_2 – Freude	k-NN	1.90	86.20	0.38	3.67	7.85
	MLP	4.94	91.39	0.00	1.27	2.41
	SVM	3.92	90.25	0.00	0.89	4.94
C_3 – Überraschung	k-NN	1.59	0.00	98.41	0.00	0.00
	MLP	7.63	0.00	92.37	0.00	0.00
	SVM	0.16	0.00	99.84	0.00	0.00
C_4 – Ärger	k-NN	7.46	0.00	0.00	86.78	5.76
	MLP	6.95	0.34	0.00	81.69	11.02
	SVM	0.51	0.00	0.00	95.93	3.56
C_5 – Ekel	k-NN	4.74	0.00	0.00	46.02	49.24
	MLP	0.61	1.68	3.82	32.87	61.01
	SVM	0.61	0.00	1.53	20.80	77.06

Tabelle 6-1: Konfusionsmatrix für geometrische Merkmale; tatsächliche Klasse C_i vs. Klassifikation $P(C_i)$ in Prozent, Training mit Datensatz MD_{s1}, Test von MD_{s2}.

Die höchsten Erkennungsraten wurden für den SVM Klassifikator erzielt, welche für die Klassen C_1 bis C_4 stets bei über 90 Prozent liegen. Für Klasse C_5 liegt die Erkennungsrate nur bei 77 Prozent, da eine erhebliche Verwechslung von über 20 Prozent mit Klasse C_4 vorliegt. Die im vorhergehenden Abschnitt dargestellten Untersuchungsergebnisse der Merkmalsräume, die für C_4 und C_5 Überlagerungen nachweisen, erklären diese Beobachtung (s. Abbildung 6-3(e)). Die geringste Performanz hatte erwartungsgemäß der k-NN Klassifikator, der aufgrund seiner Einfachheit bei der Klassentrennung, besonders bei der Erkennung von Klasse C_5, schlecht abgeschnitten hat.
Tabelle 6-2 zeigt die Konfusionsmatrix bei der Erkennung von sechs Klassen emotional expressiver Mimik sowie einer neutralen Klasse auf der Grundlage geomet-

rischer Merkmale, welche durch die Analyse der Datenbank DB_{BU} ermittelt wurden. Auch in Tabelle 6-2 zeigt der SVM Klassifikator die beste Erkennungsraten, welche für die Klassen C_1 bis C_3 bei über 92 Prozent liegen. Bei der Klasse C_4 (Ärger) kommt es hingegen häufig zu Verwechslungen mit Klasse C_7 (Trauer), bei C_5 mit Klasse C_4. Für die Klassen C_6 und C_7 liegen vielfältige Verwechslungen vor. Diese Fehler sind durch die Untersuchung der Verteilung im Merkmalsraum als sehr wahrscheinlich dargelegt worden.

Klasse C_i vs. **Klassifikation** $P(C_j)$		$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 - Neutral	k-NN	95.59	0.07	0.00	1.14	0.71	1.21	1.28
	MLP	98.72	0.00	0.00	0.64	0.00	0.21	0.43
	SVM	93.67	0.07	0.00	0.57	0.00	1.71	3.98
C_2 – Freude	k-NN	1.83	94.79	0.00	0.00	0.41	2.97	0.00
	MLP	2.10	87.96	0.14	0.00	0.07	9.74	0.00
	SVM	1.83	94.25	1.01	0.00	0.20	2.70	0.00
C_3 – Überraschung	k-NN	0.00	0.08	86.35	0.00	0.50	9.51	3.56
	MLP	0.17	0.00	91.81	0.00	0.17	5.87	1.99
	SVM	0.08	0.00	92.80	0.00	0.25	6.70	0.17
C_4 – Ärger	k-NN	5.08	0.15	0.00	62.05	8.55	0.92	23.25
	MLP	9.08	1.39	0.00	63.05	3.00	0.00	23.48
	SVM	2.23	2.54	0.00	68.51	3.62	0.69	22.40
C_5 – Ekel	k-NN	0.08	2.28	6.24	14.90	67.07	4.11	5.32
	MLP	8.59	1.52	4.18	10.80	72.55	2.36	0.00
	SVM	0.15	2.05	4.71	8.75	79.62	2.97	1.75
C_6 – Angst	k-NN	1.12	13.84	11.36	3.12	15.68	45.84	9.04
	MLP	2.80	9.28	11.60	2.96	14.88	52.08	6.40
	SVM	0.64	10.48	8.16	2.64	11.60	64.48	2.00
C_7 – Trauer	k-NN	8.13	1.74	0.41	14.01	11.44	6.22	58.04
	MLP	23.88	1.82	2.90	12.69	4.81	3.32	50.58
	SVM	10.45	1.91	0.00	11.53	4.81	5.22	66.09

Tabelle 6-2: Konfusionsmatrix für geometrische Merkmale; tatsächliche Klasse C_i vs. Klassifikation $P(C_i)$ in Prozent, Training mit Datensatz BU_{s1}, Test mit BU_{s2}.

6.1.3 Nachweis der Genauigkeit geometrischer Merkmale

Die Bestimmung geometrischer Merkmale beruht auf der Auswertung der Merkmalspunktmenge $\mathbf{P}_{fp}$ (4.18). Aufgrund technischer Beschränkungen können die 3D Punkte nicht unmittelbar erfasst werden, ohne die Versuchsperson durch Marker,

Musterprojektion o.Ä. zu stören. Daher erfolgt durch Nutzung von Pose und Modellinformation die Transformation k^{-1} (3.6) entsprechender Merkmalspunkte $\mathbf{I}_{fp}$ (4.19) aus dem Bild in die 3D Szene. Es wird dazu ein starres Modell des Gesichts verwendet, welches nicht die durch Mimik entstehenden Formvariationen berücksichtigt und somit die Frage nach der Genauigkeit erfasster Merkmale aufwirft.
In empirischen Untersuchungen erfolgte der Nachweis der hinreichenden Korrektheit geometrischer Merkmale durch Analyse der BU-4DFE Datenbank, welche eine Vielzahl von 3D Farbbildfolgen mit Mimik beinhaltet. Dabei wurden die Merkmale zum einen auf der Grundlage der tatsächlichen Oberfläche und zum anderen mit Hilfe des starren Modells ermittelt, normiert und zum Antrainieren und Testen verwendet. Abbildung 6-9 zeigt exemplarisch den Unterschied zwi-

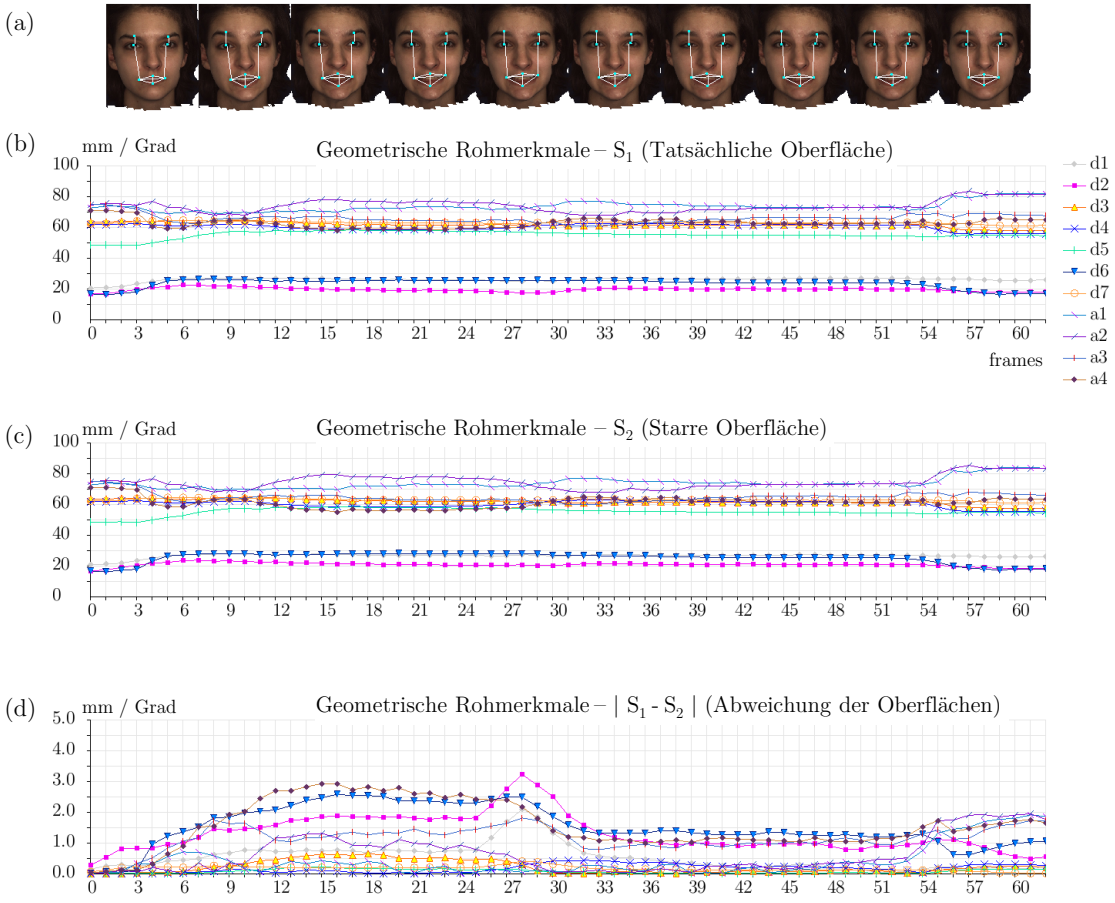


Abbildung 6-9: Beispiel zur Genauigkeit geometrischer Merkmale durch Auswertung der BU-4DFE Datenbank [Yin08]. (a) aktuelles Bild mit Merkmalspunkten, (b, c) geometrische Rohmerkmale des tatsächlichen und starren 3D Modells, (d) Versatz zwischen den Merkmalen in Millimeter bzw. Grad.

schen den Rohmerkmalen, welche auf der Grundlage des tatsächlichen sowie des starren Gesichtsmodells ermittelt wurden. Es konnte nachgewiesen werden, dass die Formvariationen auf die Berechnung des normierten Merkmalsvektors $\mathbf{f}_{geo}^{t}$ (5.20) und die nachfolgende Klassifikation nur einen untergeordneten Einfluss haben. Die bei der Untersuchung der Datenbank gemessenen Unterschiede liegen im Mittel bei circa 1.7mm für die Abstandsmerkmale d_i bzw. 3.5 Grad für die Winkelmerkmale α_j.

Durch eine 10-fache stratifizierte Kreuzvalidierung wurde für die SVM basierte Erkennung auf der Grundlage von Merkmalen, die auf der tatsächlichen Oberfläche beruhen, lediglich ein um durchschnittlich 1.3 Prozent besseres Ergebnis ermittelt. Entsprechend zeigt die Konfusionsmatrix bezüglich der Datensätze BU_{s1}/BU_{s2} in Tabelle 6-3 zwischen den Erkennungsraten der beiden Methoden S_1 und S_2 (echte vs. starre Oberfläche) nur geringe Unterschiede. Dies weist die hinreichende Genauigkeit der geometrischen Merkmale, unter Nutzung eines starren Gesichtsmodells nach.

Klasse C_i vs. **Klassifikation** $P(C_j)$		$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 - Neutral	S_1	94.31	0.07	0	0.92	0.14	0.85	3.7
	S_2	93.67	0.07	0	0.57	0	1.71	3.98
C_2 - Freude	S_1	1.83	94.39	1.01	0	0.14	2.64	0
	S_2	1.83	94.25	1.01	0	0.2	2.7	0
C_3 - Überraschung	S_1	0.08	0	90.24	0	0.17	9.02	0.5
	S_2	0.08	0	92.8	0	0.25	6.7	0.17
C_4 - Ärger	S_1	1.62	1.77	0	70.13	3.7	0.62	22.17
	S_2	2.23	2.54	0	68.51	3.62	0.69	22.4
C_5 - Ekel	S_1	0.53	2.05	4.71	10.8	78.17	2.36	1.37
	S_2	0.15	2.05	4.71	8.75	79.62	2.97	1.75
C_6 - Angst	S_1	0.88	11.52	6.8	2	10.88	63.04	4.88
	S_2	0.64	10.48	8.16	2.64	11.6	64.48	2
C_7 - Trauer	S_1	10.7	0.33	0	9.87	4.06	4.89	70.15
	S_2	10.45	1.91	0	11.53	4.81	3.22	68.09

Tabelle 6-3: Konfusionsmatrix; tatsächliche Klasse C_i vs. Klassifikation $P(C_j)$ in Prozent, SVM für geometrische Merkmale nach Methode S_1/S_2 (echte vs. starre Oberfläche), Training mit Datensatz BU_{s1}, Test von Datensatz BU_{s2}.

6.2 Auswertung dynamischer Merkmale

Analog zu Abschnitt 6.1 werden im Folgenden die Ergebnisse aus der Auswertung dynamischer Merkmale vorgestellt.

6.2.1 Analyse der Merkmalsräume - dynamische Merkmale

Entsprechend der Auswertung geometrischer Merkmale wurden die auf der Grundlage der Datenbanken DB_{MD} und DB_{BU} erfassten dynamischen Merkmale (s. Abschnitt 5.2.3) analysiert, um deren Verteilung im Merkmalsraum festzustellen und somit die Ergebnisse der Klassifikation einschätzen zu können.

6.2.1.1 Merkmalsraumanalyse – Abstände der dynamischen Merkmale

Zur Untersuchung der Verteilung dynamischer Merkmale wurden die im vorhergehenden Abschnitt 6.1.1.1 beschriebenen Methoden adaptiert. Insbesondere wurde der euklidische Abstand $||\mathbf{f}^t_{dyn} - \mu^{\mathrm{C_i}}_{dyn}||$ jedes Merkmalsvektors $\mathbf{f}^t_{dyn} \in \mathbb{R}^{14}$ (4.17) zum Merkmalsmittelwert $\mu^{\mathrm{C_i}}_{dyn}$ (6.2) der jeweiligen Klasse C_i bestimmt, um zu ermitteln, wie stark die betrachteten Klassen voneinander separiert sind bzw. wie nahe diese zueinander liegen, was ein grundsätzliches Indiz für Verwechslungen im Erkennungsschritt ist.

$$\mu^{\mathrm{C}_i}_{dyn} = \frac{1}{n_{\mathrm{C}_i}} \sum_{j=1}^{n_{\mathrm{C}_i}} \mathbf{f}^{j,\,\mathrm{C}_i}_{dyn}, \quad \mu^{\mathrm{C_i}}_{dyn} \in \mathbb{R}^{14} \tag{6.2}$$

mit n_{C_i} als Anzahl aller Samples $\mathbf{f}^j_{dyn}$, die zur Klasse C_i gehören.

Für die beiden Datenbanken DB_{MD} und DB_{BU} werden hierzu in Abbildung 6-10 und Abbildung 6-11 die Abstände zwischen den nach Klassenzugehörigkeit sortierten Trainingsdaten und dem Schwerpunkt der jeweiligen Klasse dargestellt. Nicht verwendet wird hier die neutrale Klasse C_1, da diese wegen der häufigen nicht möglichen Erfassbarkeit dynamischer Merkmale bei neutraler Mimik für die Erkennung ungeeignet ist. Für die Klassen C_2, C_3 und C_4, d.h. Freude, Überraschung und Wut, ist für beide Datenbanken eine klare Abgrenzung zu den anderen Klassen zu erkennen, d.h. diese liegen dem jeweiligen Schwerpunkt deutlich näher als der Rest. Eine solche Trennung ist für die Daten der Klassen C_5, C_6 und C_7, d.h. Ekel, Trauer und Angst nicht zu erkennen. Hier wurde durch die erfassten dynamischen Merkmale keine deutliche Abgrenzung erreicht, weshalb es bei der Klassifikation eher zu Verwechslungen kommt. Diese Beobachtung wird durch

die Berechnung der Klassifikationsgenauigkeit bestätigt. Die Ursachen hierfür liegen im nicht ausreichenden Informationsgehalt dynamischer Merkmale für diese Mimikklassen begründet (Abschnitt 6.2.2).
Bei der Auswertung der dynamischen Merkmale hat sich weiterhin gezeigt, dass für die Klasse C_7 (Trauer) weit weniger Bewegungsinformation erfasst werden konnte. Dies ist auf den geringeren Erregungsgrad der Probanden bei der Emotion Trauer zurückzuführen und lässt somit die Grenzen der dynamischen Merkmale erkennen.

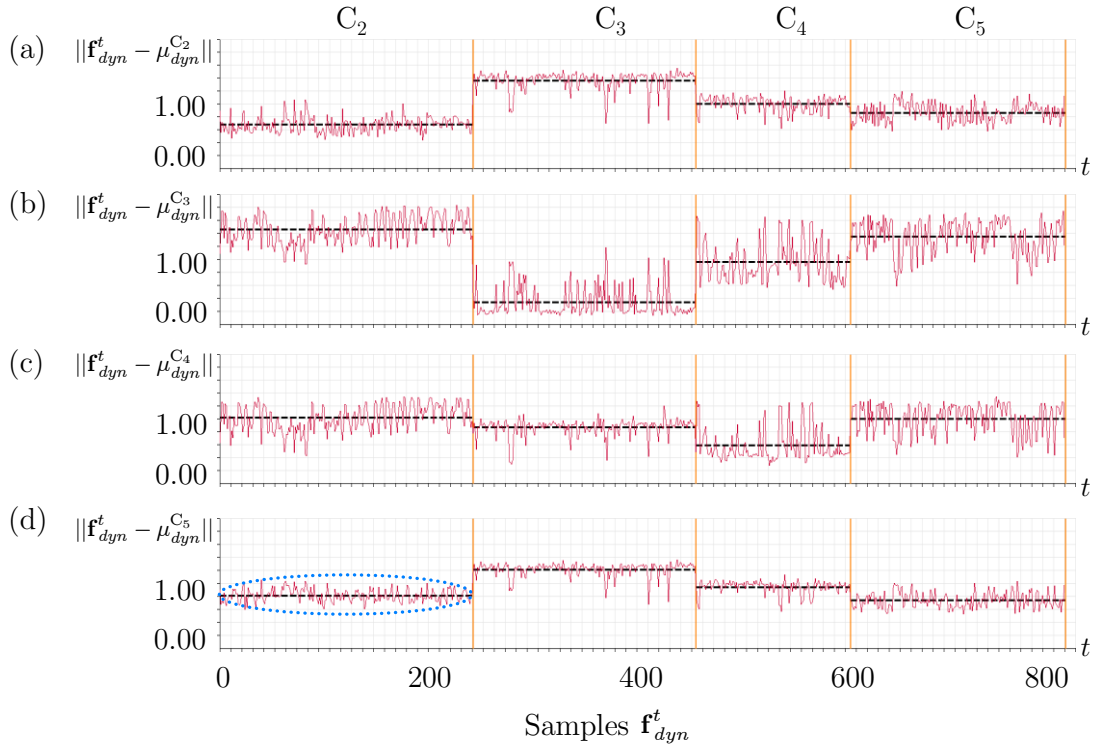


Abbildung 6-10: Analyse dynamischer Merkmale der Datenbank DB_{MD}. Die gestrichelte Linie stellt den durchschnittlichen euklidischen Abstand der Merkmalsvektoren verschiedener Klassen zum Zentrum im 14 dimensionalen Merkmalsraum dar. (a, b, c) Für die Klassen C_2, C_3, C_4, (Freude, Überraschung, Wut) ist eine klare Abgrenzung erkennbar, (d) Klasse C_5 (Ekel) hat eine deutlich schlechtere Trennung, Überschneidungen bestehen vorrangig mit der Klasse C_2.

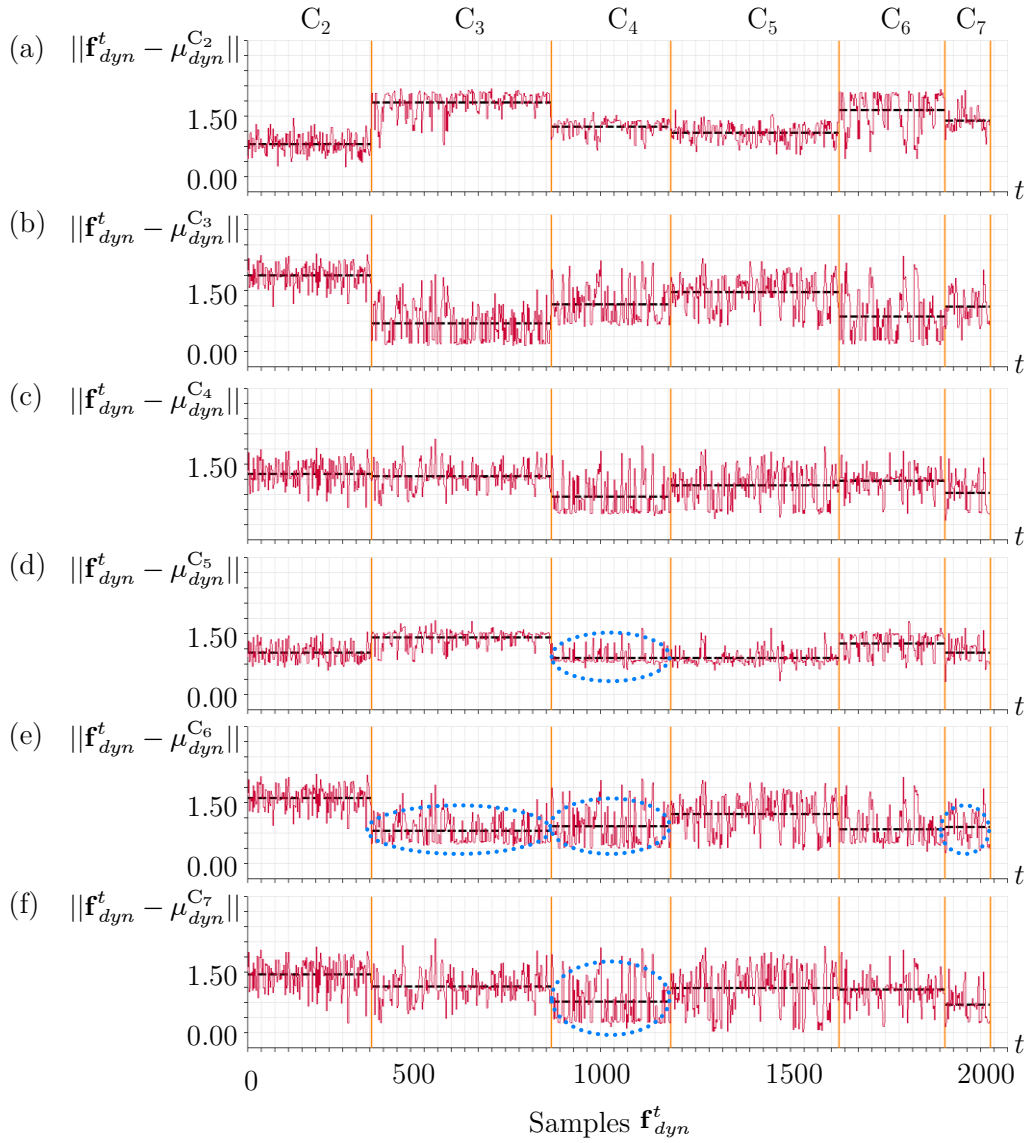


Abbildung 6-11: Analyse dynamischer Merkmale der Datenbank DB_{BU}, (a,b,c) eine deutliche Abgrenzung ist lediglich für die Klassen C_2, C_3 und C_4 zu erkennen, (d) Klasse C_5 zeigt vorrangig Überschneidungen mit C_4, (e) die Klasse C_6 zeigt vielfältige Überlappungen, (f) für C_7 (Trauer) wurden deutlich weniger dynamische Merkmale erfasst, Überschneidungen bestehen mit der Klasse C_4.

6.2.1.2 Merkmalsraumanalyse – PCA der dynamischen Merkmale

Die Untersuchung des Merkmalsraums mittels Hauptkomponentenanalyse zeigt, dass dem 14 dimensionalen dynamischen Merkmalsvektor ebenfalls starke lineare Abhängigkeiten zugrunde liegen. In Abbildung 6-12 und Abbildung 6-13 werden

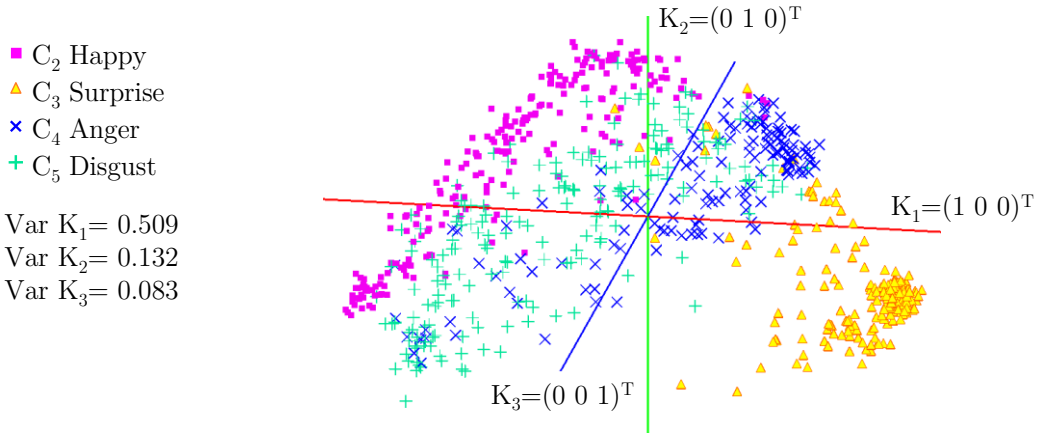


Abbildung 6-12: Dimensionsreduktion mittels PCA. Die Darstellung der ersten drei Hauptkomponenten K_i für die dynamischen Merkmale der Datenbank DB_{MD} zeigt eine klare Klassentrennung.

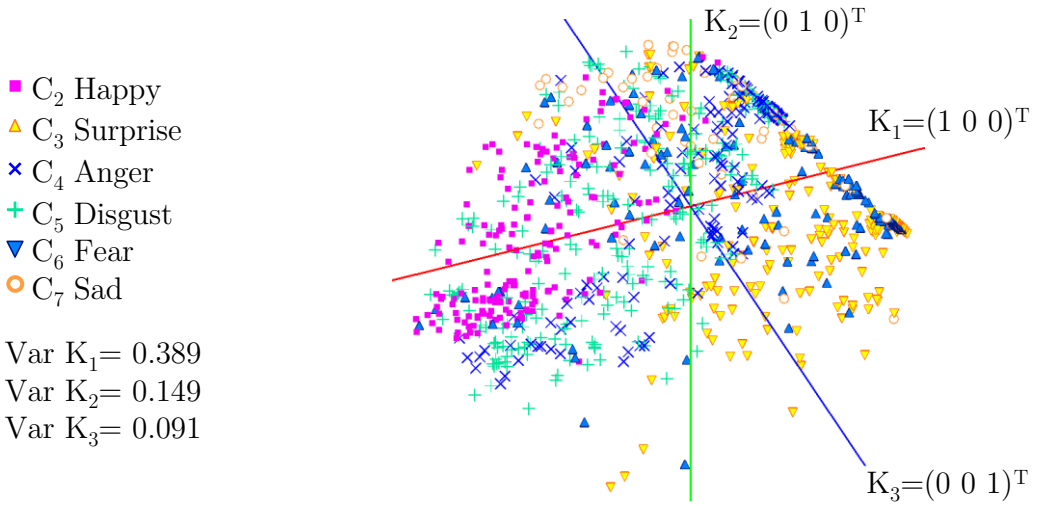


Abbildung 6-13: Dimensionsreduktion mittels PCA. Eine deutlich schlechtere Klassentrennung zeigen die ersten drei Hauptkomponenten K_i der dynamischen Merkmale für die Datenbank DB_{BU}.

die ersten drei Hauptkomponenten der reduzierten Merkmalsräume für dynamische Merkmale dargestellt.

Für die Datenbank DB_{MD}, mit vier im Merkmalsraum zugeordneten Klassen, decken die ersten drei Hauptkomponenten K_1, K_2, K_3 mehr als 70 Prozent der Gesamtvarianz ab. Es ist gut zu erkennen, dass sich die Klassen C_2 und C_3 deutlich abgrenzen, C_4 jedoch Überlappungen mit C_5 aufweist, weshalb hier Verwechslun-

gen bei der Klassifikation zu erwarten sind. Zusätzlich ist eine Überlappung von C_5 mit C_2 zu erkennen, was ebenfalls in der Untersuchung mittels des euklidischen Abstandes der Merkmalsvektoren zum Klassenzentrum ermittelt wurde (Abbildung 6-10).
Da die Datenbank DB_{BU} die zusätzlichen Klassen C_6 und C_7 (Angst und Trauer) beinhaltet, welche weitere Informationen mit einbringen, decken hier die ersten drei Hauptkomponenten lediglich etwas mehr als 60 Prozent der Gesamtvarianz ab. Folglich wird in der Darstellung des transformierten Merkmalsraumes eine wesentlich umfangreichere Überlappung der Klassen deutlich (Abbildung 6-13). Auffällig sind spezifische Überlappungen von C_2 mit C_5, von C_4 mit C_5 sowie von C_6 mit C_3. Die Klassen C_4 bis C_7 zeigen keine auffällig getrennten Bereiche. Die beste Separation zeigt die Klasse C_3 (Überraschung), was aufgrund der visuellen Erscheinung dieser Mimik auch zu erwarten ist.

6.2.1.3 Merkmalsraumanalyse – SOM für dynamische Merkmale

Bei der Untersuchung der erfassten dynamischen Merkmale mittels SOM zeigt sich erwartungsgemäß ein ähnliches Bild. In Abbildung 6-14 und Abbildung 6-15 wird jeweils die U-Matrix für die Merkmale der Datenbanken DB_{MD} und DB_{BU} dargestellt. Der markanteste und am deutlichsten abgegrenzte Häufungsbereich ist dabei für die Klasse C_2 zu erkennen. Die anderen Klassen zeigen weniger kompakte aber dennoch erkennbare Ansammlungen.
Für die Merkmalsvektoren der Datenbank DB_{BU} ist hingegen durch die SOM eine deutlich schlechtere Separierung ermittelt worden (Abbildung 6-15).
Es zeigt sich, dass sich die sechs betrachteten Klassen durch die ermittelten Merkmale nur wenig voneinander abgrenzen, was die Ergebnisse der vorhergehenden Untersuchungen mittels des Abstandsmaßes und der Hauptkomponentenanalyse stützt. Häufungsbereiche sind hier partiell noch für die Klassen C_2 und C_3 erkennbar. Der Grund hierfür liegt in den nur schwach ausgeprägten dynamischen Merkmalen der zusätzlichen Klassen C_6 und C_7.

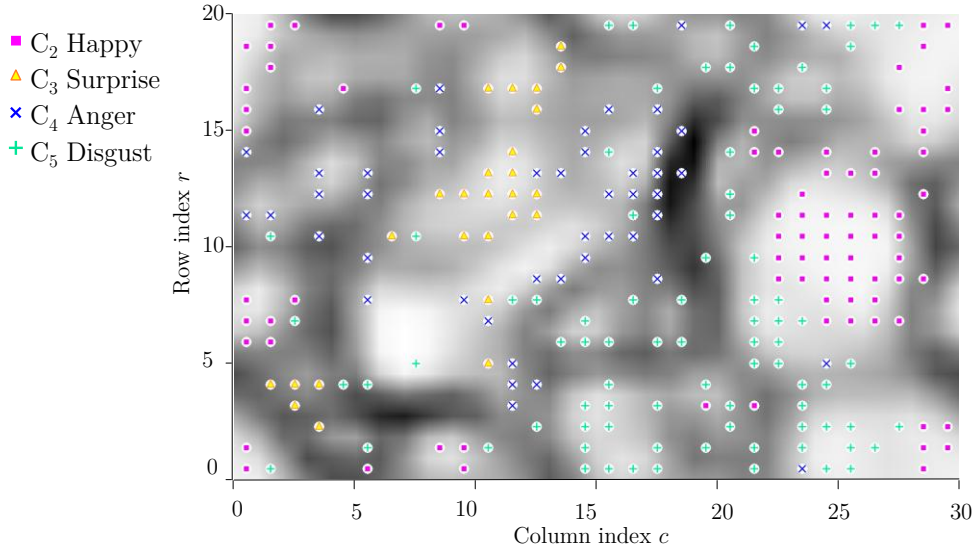


Abbildung 6-14: SOM, Darstellung der U-Matrix für die dynamischen Merkmale der Datenbank DB_{MD}. Häufungsbereiche sind für die projizierten Trainingsvektoren erkennbar.

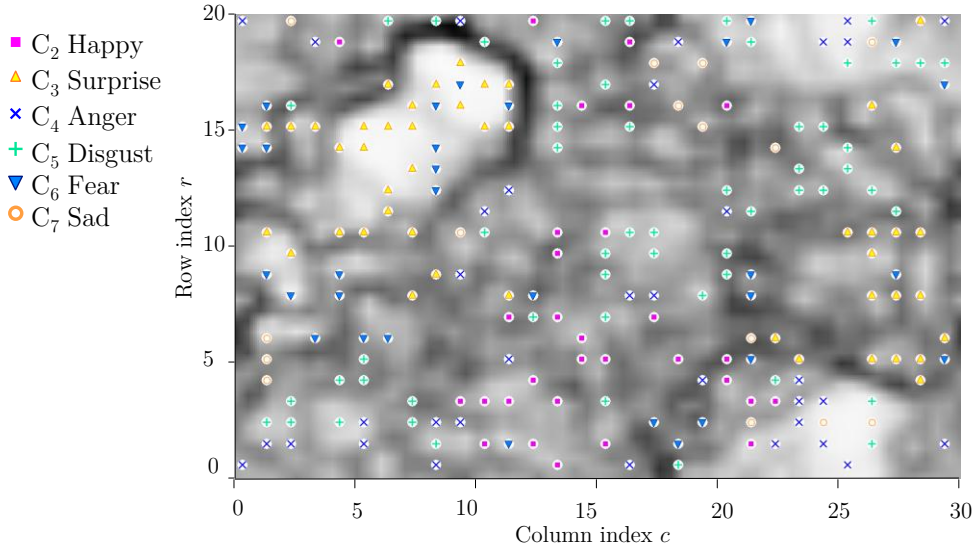


Abbildung 6-15: SOM, dynamische Merkmale, Datenbank DB_{BU}, Clusterbildung ist nur bedingt erkennbar, etwa für die Klassen C_2 und C_3.

6.2.2 Klassifikation auf der Grundlage dynamischer Merkmale

Die durch 10-fache stratifizierte Kreuzvalidierung ermittelten durchschnittlichen Erkennungsraten (SVM) liegen bei der Auswertung dynamischer Merkmale, bei einer Unterscheidung von 4 Klassen, für beide untersuchten Datenbanken DB_{MD} und DB_{BU} bei durchschnittlich 87 Prozent, bei Berücksichtigung von 6 Klassen lediglich bei circa 60 Prozent.
Eine Erkennung auf der Grundlage dynamischer Merkmale ist nur möglich, wenn Änderungen in den observierten Gesichtsregionen erfasst werden können, d.h. wenn eine Veränderung der Mimik stattfindet (siehe Aktivierungsphase, Abschnitt 4.3.1). Somit ist keine Aussage möglich, falls ein Gesichtsausdruck eine Zeit lang konstant bleibt, d.h. keine Veränderung mehr messbar ist. Tabelle 6-4 stellt die Klassifikationsergebnisse dar, die in der Aktivierungsphase, bei einer Überschreitung der Schwelle v_{min} durch die Aktivierungsfunktion $v_{sum}(t)$ (4.16) ermittelt wurden. Die dynamischen Merkmale der Datenbanken DB_{MD} und DB_{BU} wurden hierzu in gleichgroße Mengen MD_{d1}/MD_{d2} bzw. BU_{d1}/BU_{d2} unterteilt, wobei jeweils eine Menge zum Anlernen des Klassifikators verwendet wurde, die andere zum Testen. Dabei sind in den verschiedenen Merkmalsmengen durchweg Messungen unterschiedlicher Probanden enthalten. In der Konfusionsmatrix in Tabelle 6-4 sind blockweise Zeilen für die tatsächlichen Klassen C_i enthalten, welche als Grundwahrheit bekannt sind und in den Spalten $P(C_j)$ die Erkennungsrate auf der Grundlage des jeweiligen Klassifikators eingetragen. Beispielsweise wurde für die Testdaten der Klasse C_2 mittels des k-NN Verfahrens eine Erkennungsrate von 88.89 Prozent erzielt, während Verwechslungen mit anderen Klassen auftraten, z.B. mit C_5 in 9.52 Prozent aller Fälle.
Die Klassifikationsergebnisse für die Datenbank DB_{MD} zeigen, dass die Klassen C_2 und C_3 hohe Erkennungsraten von über 90 Prozent aufweisen, es jedoch bei den Klassen C_4 und C_5 häufiger zu Verwechslungen kommt, weshalb hier die Erkennungsrate nur bei ca. 80 Prozent liegt. Die im vorhergehenden Abschnitt dargestellten Untersuchungsergebnisse der Merkmalsräume, welche für C_4 und C_5 Überlagerungen nachweisen, bestätigen diese Beobachtung. Der SVM Klassifikator zeigt hier jedoch im Mittel die höchste Erkennungsrate und damit die größte Fähigkeit zur Generalisierung und Klassentrennung.

Klasse C_i vs. **Klassifikation** $P(C_j)$		$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$
C_2 – Freude	k-NN	88.89	0.00	1.59	9.52
	MLP	96.03	0.00	0.00	3.97
	SVM	93.65	0.00	0.00	6.35
C_3 – Überraschung	k-NN	0.00	98.96	0.00	1.04
	MLP	1.04	87.50	0.00	11.46
	SVM	2.08	93.75	2.08	2.08
C_4 – Ärger	k-NN	4.55	6.06	72.73	16.67
	MLP	4.55	7.58	43.94	43.94
	SVM	4.55	0.00	77.27	18.18
C_5 – Ekel	k-NN	20.00	0.00	13.33	66.67
	MLP	8.89	2.22	1.11	87.78
	SVM	13.33	0.00	8.89	77.78

Tabelle 6-4: Konfusionsmatrix für die Klassifikation mittels dynamischer Merkmale; tatsächliche Klasse C_i vs. Klassifikation $P(C_j)$ mittels k-NN, MLP und SVM in Prozent, Training mit Datensatz MD_{d1}, Test von Datensatz MD_{d2}.

Tabelle 6-5 stellt die Konfusionsmatrix für die Erkennung von sechs verschiedenen Klassen auf der Grundlage dynamischer Merkmale dar, welche durch die Analyse der Datenbank DB_{BU} ermittelt wurden. Wie bei der Analyse des Merkmalsraumes erläutert, ist die Unterscheidung von sechs Klassen ein deutlich schwierigeres Problem, insbesondere, da die dynamischen Merkmale der Klassen C_6 und C_7 nur eine schwache Ausprägung aufweisen. Der k-NN Klassifikator, welcher eine schlechtere Generalisierungsfähigkeit als MLP und SVM aufweist, zeigt hier fast durchweg ungenauere Ergebnisse. Die Klassen C_2 und C_3 lassen sich mittels SVM zu ca. 80 Prozent korrekt erkennen, die Klasse C_5 noch zu ca. 70 Prozent. Bei der Detektion der Klasse C_4 gibt es starke Verwechslungen mit Klasse C_5 und lediglich eine Erkennung von weniger als 60 Prozent. Die Klasse C_6 (Angst) wurde zum größten Teil mit Klasse C_3 (Überraschung) verwechselt. Es zeigt sich hier deutlich, dass eine Erkennung der Klasse C_6 auf der Grundlage dynamischer Merkmale nicht zweckmäßig ist. Zum Erkennen dieser Klasse sind somit zusätzliche Merkmale erforderlich. Das gleiche gilt für die Klasse C_7 (Trauer), welche sich bei der ausgewerteten Datenbank mit Hilfe dynamischer Merkmale aufgrund der schwachen Erregung der Versuchspersonen kaum erfassen ließ.

Klasse C_i vs. **Klassifikation** $P(C_j)$		$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_2 – Freude	k-NN	56.31	0.97	0.97	21.36	18.45	1.94
	MLP	73.79	4.85	2.91	18.45	0.00	0.00
	SVM	80.58	0.97	1.94	16.50	0.00	0.00
C_3 – Überraschung	k-NN	0.00	79.71	0.00	2.90	16.67	0.72
	MLP	4.35	94.20	0.00	1.45	0.00	0.00
	SVM	2.17	81.16	0.00	2.17	12.32	2.17
C_4 – Ärger	k-NN	5.41	0.00	35.14	35.14	5.41	18.92
	MLP	16.22	6.76	59.46	17.57	0.00	0.00
	SVM	4.05	6.76	44.59	39.19	0.00	5.41
C_5 – Ekel	k-NN	13.91	2.61	15.65	58.26	5.22	4.35
	MLP	33.04	2.61	13.91	50.43	0.00	0.00
	SVM	12.17	0.87	13.91	71.30	1.74	0.00
C_6 – Angst	k-NN	12.00	34.67	6.67	13.33	21.33	12.00
	MLP	12.00	62.67	4.00	21.33	0.00	0.00
	SVM	6.67	57.33	5.33	17.33	8.00	5.33
C_7 – Trauer	k-NN	4.35	8.87	34.78	26.09	0.00	26.09
	MLP	13.04	43.48	39.13	4.35	0.00	0.00
	SVM	0.00	4.35	30.43	4.35	8.70	52.17

Tabelle 6-5: Konfusionsmatrix für die Klassifikation mittels dynamischer Merkmale; tatsächliche Klasse C_i vs. Klassifikation $P(C_j)$ in Prozent, Training mit Datensatz BU_{d1}, Test von BU_{d2}. Die Vermischungen im Merkmalsraum (Abschnitt 6.2.1) erschweren die Klassifikation. Der SVM Klassifikator erzielt hier im Mittel die besten Ergebnisse.

Die Auswertung der BU Datenbank zeigt, dass die Schwäche bei der Erfassung dynamischer Merkmale in der begrenzten Möglichkeit liegt, sehr differenzierte und langsame Bewegungen zu detektieren, welche zur genauen Erkennung von Klassen wie Ärger, Angst und Trauer eingesetzt werden können. Die Stärke bei der Auswertung dynamischer Merkmale liegt jedoch in der Möglichkeit der frühen Erkennung von Mimikänderungen, was auf der Grundlage geometrischer Merkmale häufig zu Problemen führt. Bei diesen kommt es in der Anfangsphase schnell zu Fehlklassifikationen aufgrund nicht ausreichend ausgeprägter Merkmale. Daher stellen die dynamischen Merkmale im vorgeschlagenen Konzept eine Ergänzung zu den geometrischen Merkmalen dar, die wie in Abschnitt 6.1 dargestellt, eine höhere Erkennungsrate erzielen.

6.2.3 Klassifikation nach dem Ansatz zur Gesichtsnormierung

Das Verfahren zur Mimikanalyse mittels Gesichtsnormierung, welches ausschließlich auf der Auswertung dynamischer Merkmale beruht, kann aufgrund der erforderlichen Stereodaten sowie Information über die Kamerakalibrierung als Ganzes nur mit hauseigenen Aufnahmen überprüft werden. Die Aufzeichnung einer Datenbank mit sechs Basisemotionen war aufgrund der schwierigen Erzeugung der Emotionen Trauer und Angst nicht erfolgreich. Im Folgenden werden daher Erkennungsraten aus der Auswertung der Datenbank DB_{ULM} dargestellt, die Bildfolgen normierter Gesichter mit sechs Klassen C_2-C_7 beinhaltet (Abbildung 6-16). Auf diese Weise lässt sich das Prinzip der Auswertung normierter Gesichter auf der Grundlage physiologisch motivierter Regionen untersuchen. Da keine 3D-Modellinformation vorlag, erfolgte die Festlegung der 14 Regionen manuell.
Die durch 10-fache stratifizierte Kreuzvalidierung ermittelten durchschnittlichen Erkennungsraten (SVM) liegen bei einer Unterscheidung von 4 Klassen C_2-C_5 bei 86 Prozent, bei Unterscheidung aller 6 Klassen bei 80 Prozent und damit um gut 20 Prozent über dem Ergebnis der Auswertung dynamischer Merkmale mittels Merkmalsnormierung. Dies resultiert zum einen aus der Tatsache, dass das Datenmaterial DB_{ULM} deutlich expressivere Mimik für die Klassen Angst und Trauer beinhaltet und zum anderen daraus, dass bei der Erfassung dynamischer Merkmale aus bereits normierten Gesichtern keine Unterdrückung globaler Kopfbewegung erforderlich ist, wodurch die Sensitivität der Merkmale und somit auch die Erkennungsleistung gesteigert wird.
Eine detaillierte Darstellung der erzielten Erkennungsraten erfolgt aus Platzgründen in Anhang 8.1.1. Ein wichtiges Ergebnis dieser Auswertung ist der Nachweis des Funktionsprinzips der Erkennung mimisch präsentierter Basisemotionen auf der Grundlage normierter Bilder des Gesichts, unter Verwendung dynamischer Merkmale, welche in physiologisch motivierten Regionen erfasst werden. Als genereller Nachteil der Gesichtsnormierung hat sich hingegen die verhältnismäßig aufwendige Erzeugung normierter Bilder herausgestellt, was jedoch durch den erweiterten Ansatz zur Merkmalsnormierung kompensiert wurde.

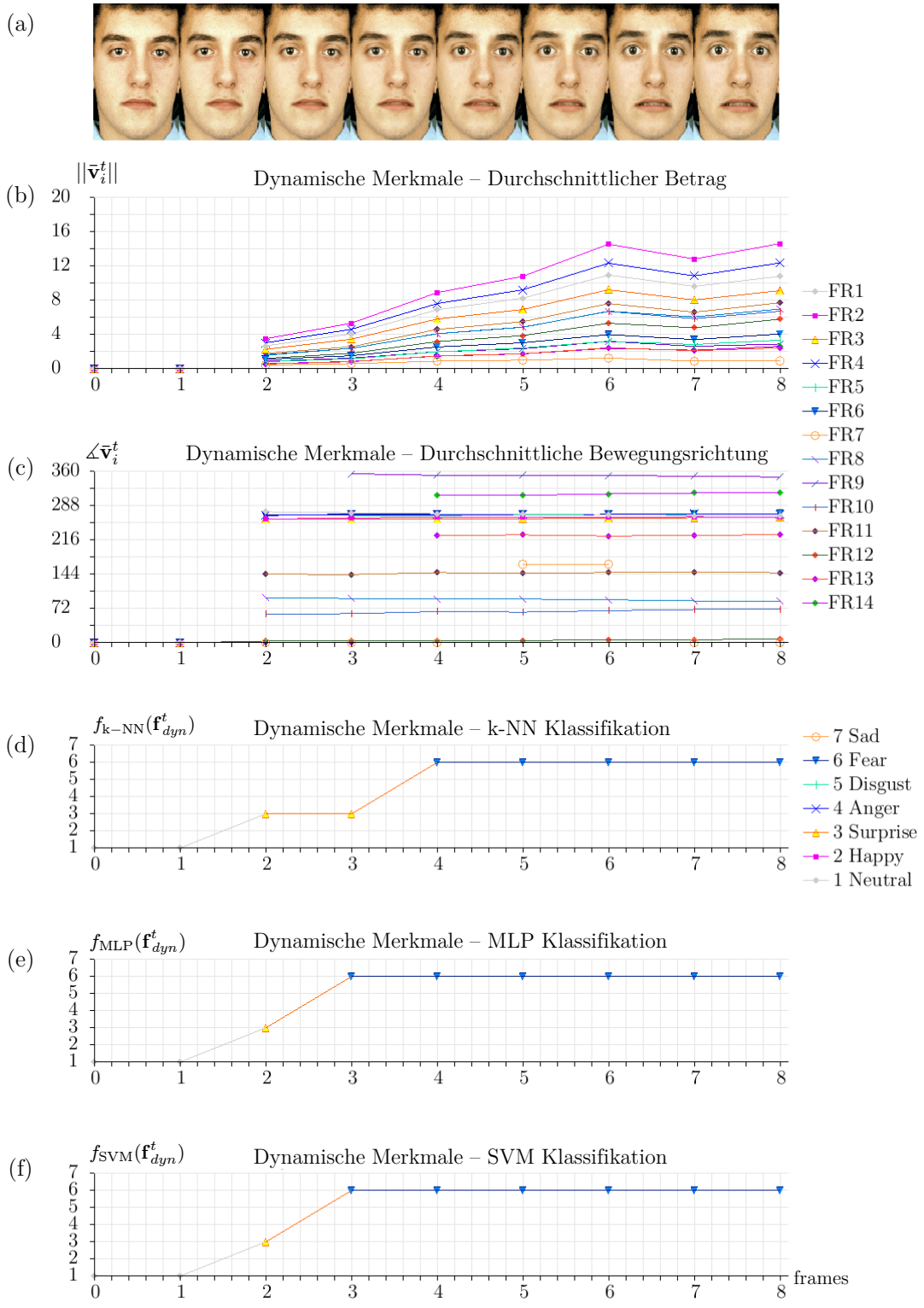


Abbildung 6-16: Beispielsequenz ULM_1 mit Mimikwechsel $C_1 \rightarrow C_6$, (a) Normiertes Bild, (b, c) dynamischer Merkmalsvektor, (d-f) Klassifikationsergebnisse.

6.3 Merkmalsselektion

Die Auswahl relevanterer Merkmale ermöglicht grundsätzlich das Erzeugen besserer Klassifikationsmodelle. Dabei wird durch die Merkmalsselektion die Leistungsfähigkeit des verwendeten maschinellen Lernverfahrens gesteigert, indem irrelevante und redundante Merkmale entfernt werden. Dies führt normalerweise neben einer Reduzierung der Dimension des Merkmalsraumes auch zu einem Geschwindigkeitsvorteil und einer besseren Interpretierbarkeit der Daten. Hierzu wird der Beitrag der Merkmale im Erkennungsprozess ermittelt, z.B. durch Kreuzvalidierung oder Bestimmung eines Merkmalsgütemaßes.
Die Untersuchung der Korrelation, d.h. der linearen Zusammenhänge zwischen den Komponenten der erfassten Merkmalsvektoren, zeigt erwartungsgemäß deutliche Redundanzen zwischen der linken und rechten Gesichtshälfte. Tabelle 6-6 und Tabelle 6-7 zeigen den Korrelationskoeffizienten r (6.3) für alle dynamischen und geometrischen Merkmale der Datenbank $\mathrm{DB_{MD}}$. Dabei wurden alle Klassen berücksichtigt. Der Korrelationskoeffizient stellt als dimensionsloses Maß den Grad des linearen Zusammenhangs zwischen den Merkmalskomponenten dar, wobei r=1 eine positive, r=-1 eine negative und r=0 eine nicht vorhandene lineare Abhängigkeit nachweist. Abhängigkeiten höherer Ordnung können dennoch bestehen. Zur Veranschaulichung der Ergebnisse sind diese jedoch nicht entscheidend.

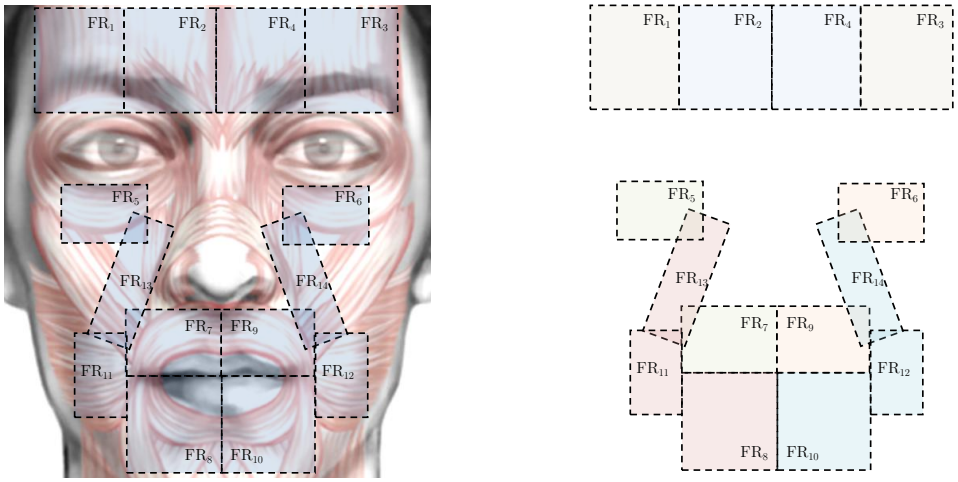
$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)\sigma_x\sigma_y}, \quad r \in [-1,\ +1] \tag{6.3}$$

mit $\bar{x}, \bar{y}$ als arithmetischer Mittelwert und σ_x, σ_y als Standardabweichung zweier Merkmale x, y. Dabei repräsentiert n die Anzahl der Merkmalsstichproben.

Die Korrelationsergebnisse zeigen, dass die gemessene Bewegung in den Regionen $\mathrm{FR_1/FR_2}$ vorrangig mit den achsensymmetrischen Regionen $\mathrm{FR_3/FR_4}$ in Zusammenhang steht. Ebenso korreliert die Bewegung in $\mathrm{FR_5/FR_7}$ mit $\mathrm{FR_6/FR_9}$ bzw. $\mathrm{FR_8/FR_{11}/FR_{13}}$ mit $\mathrm{FR_{10}/FR_{12}/FR_{14}}$. Somit zeigt sich, dass die Information über die Bewegung in der linken und rechten Gesichtshälfte prinzipiell das gleiche ausdrückt, zwischen den anderen Regionen jedoch nur eine geringe Abhängigkeit besteht. Eine Ausnahme bilden hier die Regionen $\mathrm{FR_1/FR_2}$ bzw. $\mathrm{FR_3/FR_4}$, die ebenfalls eine starke Abhängigkeit aufweisen, aber dennoch relevante zusätzliche Information beinhalten.

r	$\measuredangle\bar{\mathbf{v}}_1$	$\measuredangle\bar{\mathbf{v}}_2$	$\measuredangle\bar{\mathbf{v}}_3$	$\measuredangle\bar{\mathbf{v}}_4$	$\measuredangle\bar{\mathbf{v}}_5$	$\measuredangle\bar{\mathbf{v}}_6$	$\measuredangle\bar{\mathbf{v}}_7$	$\measuredangle\bar{\mathbf{v}}_8$	$\measuredangle\bar{\mathbf{v}}_9$	$\measuredangle\bar{\mathbf{v}}_{10}$	$\measuredangle\bar{\mathbf{v}}_{11}$	$\measuredangle\bar{\mathbf{v}}_{12}$	$\measuredangle\bar{\mathbf{v}}_{13}$	$\measuredangle\bar{\mathbf{v}}_{14}$
$\measuredangle\bar{\mathbf{v}}_1$	1	0.72	0.86	0.73	-.09	0.11	-.04	-.34	0.07	-.33	-.28	-.16	-.17	0.05
$\measuredangle\bar{\mathbf{v}}_2$		1	0.73	0.86	-.15	0.14	-.07	-.33	0.07	-.32	-.31	-.16	-.16	0.12
$\measuredangle\bar{\mathbf{v}}_3$			1	0.68	-.14	0.16	-.06	-.34	0.04	-.32	-.26	-.13	-.15	0.04
$\measuredangle\bar{\mathbf{v}}_4$				1	-.18	0.10	-.17	-.31	0.02	-.29	-.32	-.13	-.21	0.19
$\measuredangle\bar{\mathbf{v}}_5$					1	-.56	0.15	0.08	-.07	0.07	0.12	-.17	0.38	-.37
$\measuredangle\bar{\mathbf{v}}_6$						1	-.10	-.06	0.16	-.05	-.07	0.13	-.23	0.28
$\measuredangle\bar{\mathbf{v}}_7$							1	0.24	0.53	0.17	0.28	0.13	0.13	-.13
$\measuredangle\bar{\mathbf{v}}_8$								1	0.09	0.79	0.45	0.34	0.19	-.07
$\measuredangle\bar{\mathbf{v}}_9$									1	-.02	0.12	0.21	-.05	0.02
$\measuredangle\bar{\mathbf{v}}_{10}$										1	0.27	0.36	0.15	-.09
$\measuredangle\bar{\mathbf{v}}_{11}$											1	0.63	0.31	0.03
$\measuredangle\bar{\mathbf{v}}_{12}$												1	0.01	0.31
$\measuredangle\bar{\mathbf{v}}_{13}$													1	-.32
$\measuredangle\bar{\mathbf{v}}_{14}$														1

Tabelle 6-6: Korrelationskoeffizient r für die dynamischen Merkmale bzgl. Datenbank DB_{MD}. Die Ergebnisse der Datenbank DB_{BU} sind äquivalent. Deutlich erkennbar ist dabei die Symmetrie zwischen linker und rechter Gesichtshälfte (Abbildung 6-17).



(a) (b)

Abbildung 6-17: Flussregionen mit Gesichtsmuskulatur, Quelle [Flo10], (a) Definition der Regionen FR_i zur Veranschaulichung der Korrelation, (b) bei der Aktivierung zeigt sich die Symmetrie der linken und rechten Gesichtshälfte.

r	d_1	d_2	d_3	d_4	d_5	d_6	α_1	α_2	α_3	α_4
d_1	1	0.981	0.473	0.428	0.440	0.046	-0.347	-0.224	-0.096	-0.213
d_2		1	0.469	0.419	0.408	0.031	-0.320	-0.238	-0.114	-0.186
d_3			1	0.931	0.279	-0.578	-0.786	-0.726	0.118	0.109
d_4				1	0.277	-0.548	-0.762	-0.756	0.137	0.134
d_5					1	0.041	-0.321	-0.344	-0.696	-0.739
d_6						1	0.499	0.536	-0.011	-0.089
α_1							1	0.692	0.273	0.020
α_2								1	0.094	-0.216
α_3									1	0.685
α_4 -										1

Tabelle 6-7: Korrelationskoeffizient r für geometrische Merkmale bzgl. DB_{MD}, auch hier besteht eine deutliche Symmetrie zwischen linker und rechter Hälfte.

Für die geometrischen Merkmale bestehen entsprechende starke Korrelationen zwischen d_1 und d_2, d_3 und d_4, α_1 und α_2 sowie α_3 und α_4. Im Sinne der Merkmalsselektion werden üblicherweise korrelierte Merkmale entfernt. Die redundante Information der in beiden Gesichtshälften erfassten Merkmale hat jedoch Vorteile im Hinblick auf die Robustheit des gesamten Verfahrens sowie im Hinblick auf seine Erweiterbarkeit. Insbesondere sind hierbei folgende Punkte zu nennen:

- Auch bei unvollständiger Messung, z.B. bei Verdeckung einer Gesichtshälfte ist eine Klassifikation möglich, was somit die Zuverlässigkeit erhöht.
- Falls nur eine Gesichtshälfte ausreichend aktiviert wird, ist eine Erkennung häufig immer noch auf der Grundlage der restlichen Merkmale möglich.
- Eine interessante prospektive Anwendung stellt die Auswertung asymmetrischer Gesichtsausdrücke dar, z.B. für medizinische Zwecke [Ahl96].

Aus den genannten Gründen erfolgt die Erfassung der Merkmale in beiden Hälften des Gesichts. Die Notwendigkeit des Verwendens aller Merkmale ergibt sich weiterhin aus den Ergebnissen der Analyse des Datensatzes DB_{BU} mittels Kreuzvalidierung. Hierzu wurden vergleichend verschiedene Merkmalskombinationen zum gegenseitigen Lernen und Testen auf der Grundlage eines SVM Klassifikators untersucht. Die Auswertung des gesamten Datensatzes hat dabei gezeigt, dass die höchste Erkennungsrate nur bei Verwendung aller Merkmale erzielt wird.

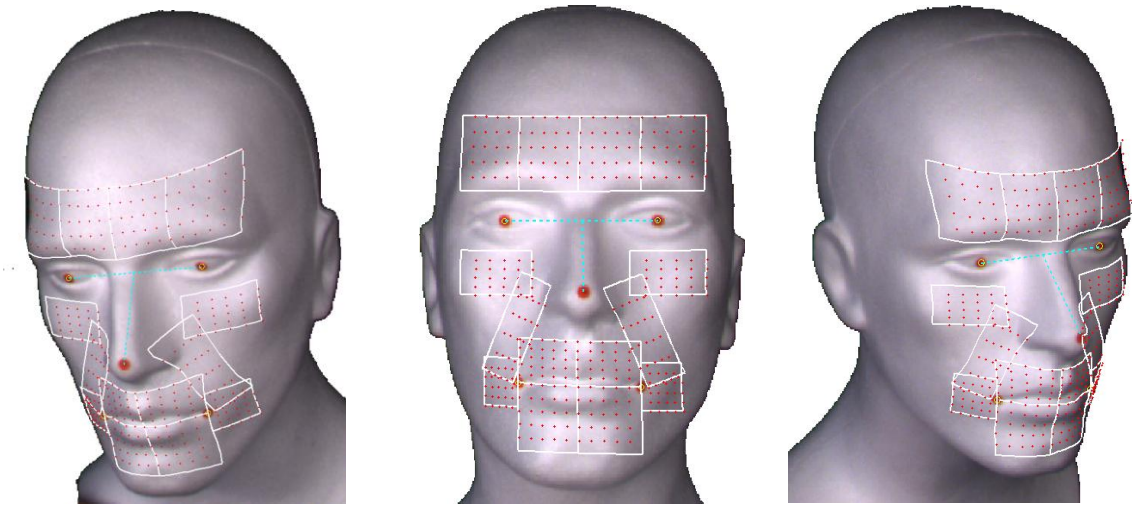
6.4 Auswirkungen der Pose auf Merkmale und Erkennung

Um die vorgeschlagene Systemstruktur zur Mimikanalyse auf ihren grundsätzlichen Nutzen und ihre Korrektheit zu prüfen, z.B. Verwendung geometrischer und dynamischer Merkmale, Merkmalsnormierung, etc., wurden bei der Auswertung der verschiedenen Datenbanken fast ausschließlich Frontalansichten verwendet.
Grundsätzlich bietet das Verfahren zur Auswertung von 3D Merkmalen, wie Abständen im Raum, Winkeln sowie mittels Nullposetransformation z (5.22) transformierten Verschiebungsvektoren eine gewisse Poseunabhängigkeit. Voraussetzung ist hierbei eine hinreichend korrekte Erfassung der 2D Merkmals- bzw. Ankerpunkte $\mathbf{I}_{fp}$ (4.19) und Bewegungsinformation im Bild, welche die Grundlage zur Pose- und Merkmalsberechnung darstellen. Prinzipiell wird die Detektion dieser Punkte durch auftretende Selbstverdeckungen limitiert, die durch Kopfrotation aus der Ebene entstehen und somit die Grenze für die maximal mögliche Pose festlegen. In Untersuchungen wurde diese bei ± 30 Grad für die Rotationsparameter ω, ϕ (s. Abbildung 5-14) ermittelt. Zur Überprüfung der Abhängigkeit gemessener Merkmale von der aktuellen Kopforientierung wurde die Abweichung extrahierter Merkmale bei konstanter Mimik, während einer bekannten Poseänderung, mit Hilfe eines Kopfmodells aus Kunststoff ermittelt (Abbildung 6-18/19).
Aus Abbildung 6-19 wird ersichtlich, wie sehr die Genauigkeit der gemessenen geometrischen Merkmale vom sogenannten "foreshortening problem", d.h. der durch die Perspektive verursachten Abstandsverkürzung und damit reduzierten Bildauflösung bei Rotationen aus der Ebene abhängt. Beim Auftreten von bis zu $\pm$ 25 Grad für die Rotationsparameter ω, ϕ, κ wurde in empirischen Untersuchungen eine Abweichung der gemessenen Werte zu den tatsächlichen von bis zu 25 Prozent ermittelt. Diese Ungenauigkeit kann nur durch das Hinzuziehen einer bzw. mehrerer Kameras reduziert werden, welche den durch die Perspektive verkürzten Bereich erfassen und somit eine genauere Messung ermöglichen.
Als weiteres Resultat dieser Untersuchung wird deutlich, wie sehr die Auswertung von 3D Merkmalen in Weltkoordinaten die rein 2D basierte Merkmalsmessung übertrifft, wie sie in anderen Verfahren praktiziert wird. Da bei einer solchen in Pixeln gemessen wird, entstehen durch translatorische Kopfbewegungen in Kamerarichtung schnell Variationen der gemessenen Merkmalswerte von 50 Prozent und mehr, was eine Normierung sehr fehleranfällig macht (Abbildung 6-19(c)).
Empirische Untersuchungen haben gezeigt, dass eine Klassifikation auf der Grundlage von Merkmalsvektoren, welche durch perspektivische Verkürzung be-

dingte Ungenauigkeiten enthalten, verhältnismäßig unproblematisch ist. Dennoch hat dies eine Verringerung der Erkennungsgenauigkeit zur Folge. Zur Simulation dieser Ungenauigkeit wurden exemplarisch die auf der Grundlage der Datenbank DB_{BU} erfassten Merkmale, um einen Betrag von ± 20 Prozent gestört und mittels SVM klassifiziert (Tabelle 6-8). Das Ergebnis ist eine noch immer akzeptable Erkennungsrate, bei einer durchschnittlichen Verschlechterung um 8 Prozent.

Klasse C_i vs. **Klassifikation** $P(C_j)$	$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 – Neutral	82.23	0.36	0.00	1.99	1.56	4.48	9.38
C_2 – Freude	1.08	91.75	1.22	0.07	1.56	3.72	0.61
C_3 – Überraschung	0.17	0.08	90.98	0.00	0.50	7.78	0.50
C_4 – Ärger	3.39	2.46	0.08	64.43	4.70	0.92	24.02
C_5 – Ekel	0.99	2.36	4.71	11.25	75.29	2.74	2.66
C_6 – Angst	1.28	10.56	8.00	2.00	14.88	59.12	4.16
C_7 – Trauer	8.62	2.16	1.58	13.10	5.39	7.13	62.02

Tabelle 6-8: Konfusionsmatrix (SVM) auf der Basis gestörter Merkmale $\mathbf{f}_{geo}^{t}$ (Random noise ± 20%), Klasse C_i vs. Klassifikation $P(C_j)$ in Prozent, vergleichend zeigt Tabelle 6-2 das Ergebnis für die nicht gestörten Merkmale.



$\{\omega, \phi, \kappa\}=\{21, -33, 12\}$ $\{\omega, \phi, \kappa\}=\{0, 0, 0\}$ $\{\omega, \phi, \kappa\}=\{22, 26, 24\}$

Abbildung 6-18: Modell zur Überprüfung der Abhängigkeit gemessener Merkmale von der aktuellen Kopfpose, dargestellt sind verschiedene Orientierungen.

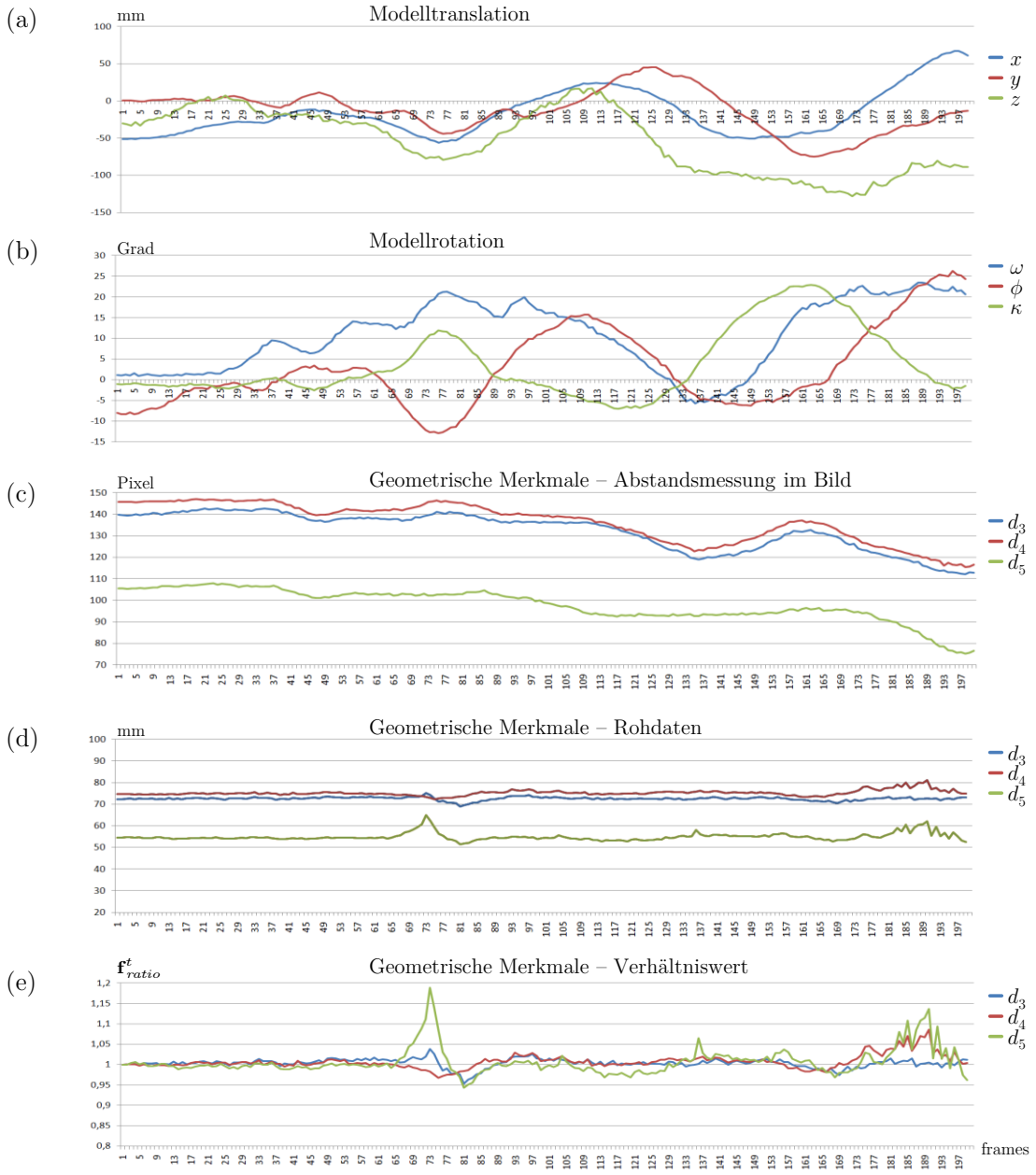


Abbildung 6-19: Beispielsequenz zur Abhängigkeit der geometrischen Merkmale von der Kopfpose, (a, b) Elemente des Posevektors $\mathbf{t}$ entsprechend (5.2), (c) Beispielparameter $d_{3\text{-}5}$, die starken Abweichungen der Abstände in Pixeln zeigen den Nachteil einer rein 2D basierten Merkmalserfassung, (d) Abstände d_3, d_4, d_5 in Weltkoordinaten, die Konstanz der Merkmale zeigt den Vorteil der 3D gestützten Auswertung, (e) normierte Merkmalsverhältnisse $\mathbf{f}^t_{ratio}$ (5.18).

6.5 Integration geometrischer und dynamischer Merkmale

Wie in Kapitel 5.3 dargelegt, weisen geometrische und dynamische Merkmale jeweils vor- als auch nachteilige Eigenschaften auf, welche sich jedoch durch eine Fusion unter bestimmten Randbedingungen kompensieren lassen. Zum einen kommt es während des Übergangs zwischen zwei Mimikzuständen bei der Klassifikation mittels geometrischer Merkmale häufig zu Fehlklassifikationen, was auf einen während der Übergangsphase wenig ausgeprägten Merkmalsvektor $\mathbf{f}_{geo}^{t}$ zurückzuführen ist. Weiterhin ist auf der Grundlage dynamischer Merkmale oft eine schnellere Erkennung der Mimik möglich. Dies setzt aber eine hinreichende Erfassbarkeit dynamischer Merkmale voraus, welche durch die Aktivierungsfunktion $v_{sum}(t)$ (4.16) nachgewiesen wird.
Durch die Integration der beiden Merkmalstypen lässt sich nach (5.23) eine Klassifikation abhängig von der Aktivierungsfunktion realisieren. Nachfolgend werden hierzu Beispielsequenzen dargestellt und im Anschluss interpretiert.
Die Beispielssequenz MD_1 in Abbildung 6-20 stellt zwei Übergänge (Neutral→Ärger→Freude) dar, wobei im zweiten Übergang eine kurzzeitige Fehlklassifikation auf der Grundlage des nicht stark genug ausgeprägten geometrischen Merkmalsvektors erfolgte (Verwechslung der Klassen Freude und Ekel). Die zugehörige Konfusionsmatrix in Tabelle 6-1 zeigt entsprechend für diese beiden Klassen die höchste Verwechslungsrate. Das Hinzuziehen der Klassifikation mittels dynamischer Merkmale behebt diesen Fehler. Die gleiche Situation besteht in den nachfolgenden Beispielsequenzen MD_2 und BU_1 (Abbildung 6-21/22).
Beispiel BU_2 in Abbildung 6-23 zeigt eine Sequenz mit einem Wechsel von Klasse Neutral zu Angst. Dabei zeigen sich sowohl für die Klassifikation mittels dynamischer als auch geometrischer Merkmale Verwechslungen mit der Klasse Überraschung. Dies resultiert, wie bei der Merkmalsanalyse dargestellt, aus der Nähe der beiden Klassen im Merkmalsraum. Für die Klasse Angst ist aufgrund der schlechten Erkennungsrate bei der Klassifikation durch dynamische Merkmale keine bzw. nur eine sehr geringe Verbesserung durch die integrierte Auswertung zu erwarten.
Zur Erkennung der Klasse Trauer sind dynamische Merkmale ebenfalls weniger geeignet. Die Beispielsequenz BU_3 in Abbildung 6-24 stellt den Übergang zwischen den Klassen Neutral zu Trauer dar. Hier wird deutlich, dass durch den geringen Erregungsgrad des Probanden nicht ausreichend dynamische Merkmale erfasst wurden, um eine Klassifikation durchzuführen. Die Klassifikation mittels geometrischer Merkmale verlief hingegen erfolgreich.

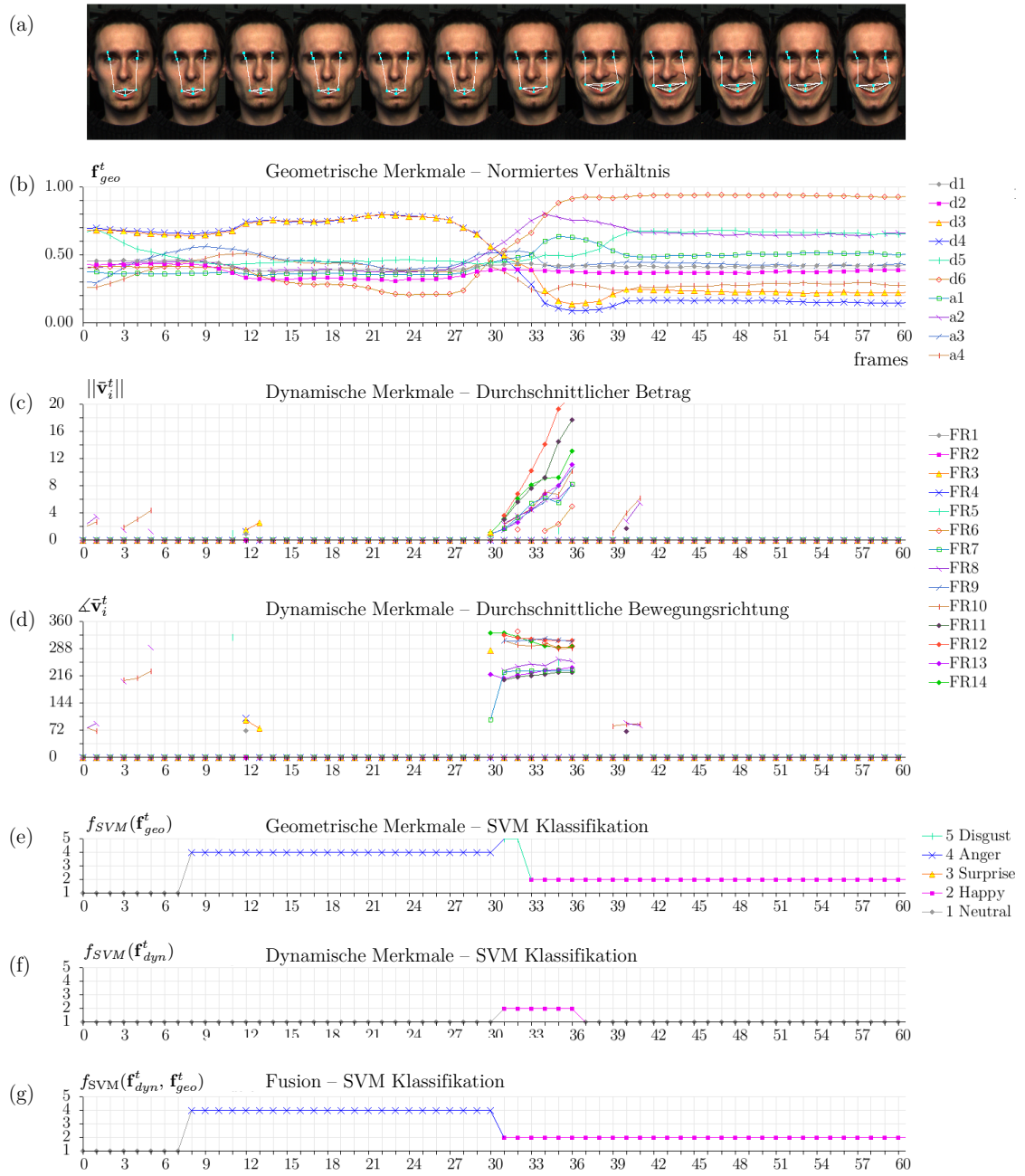


Abbildung 6-20: Beispielsequenz MD_1 mit Mimikwechsel $C_1 \rightarrow C_4 \rightarrow C_2$, (a) Projektion geometrischer Merkmale auf das aktuelle Bild, (b-d) erfasste geometrische und dynamische Merkmale, (e) Klassifikation $f_{\mathrm{SVM}}(\mathbf{f}^t_{geo})$ (geometrisch) mit Störung des Klassifikators im Zeitfenster der Mimikänderung, (f) Klassifikation $f_{\mathrm{SVM}}(\mathbf{f}^t_{dyn})$ (dynamisch), (g) die Fusion führt zur Beseitigung des Fehlers.

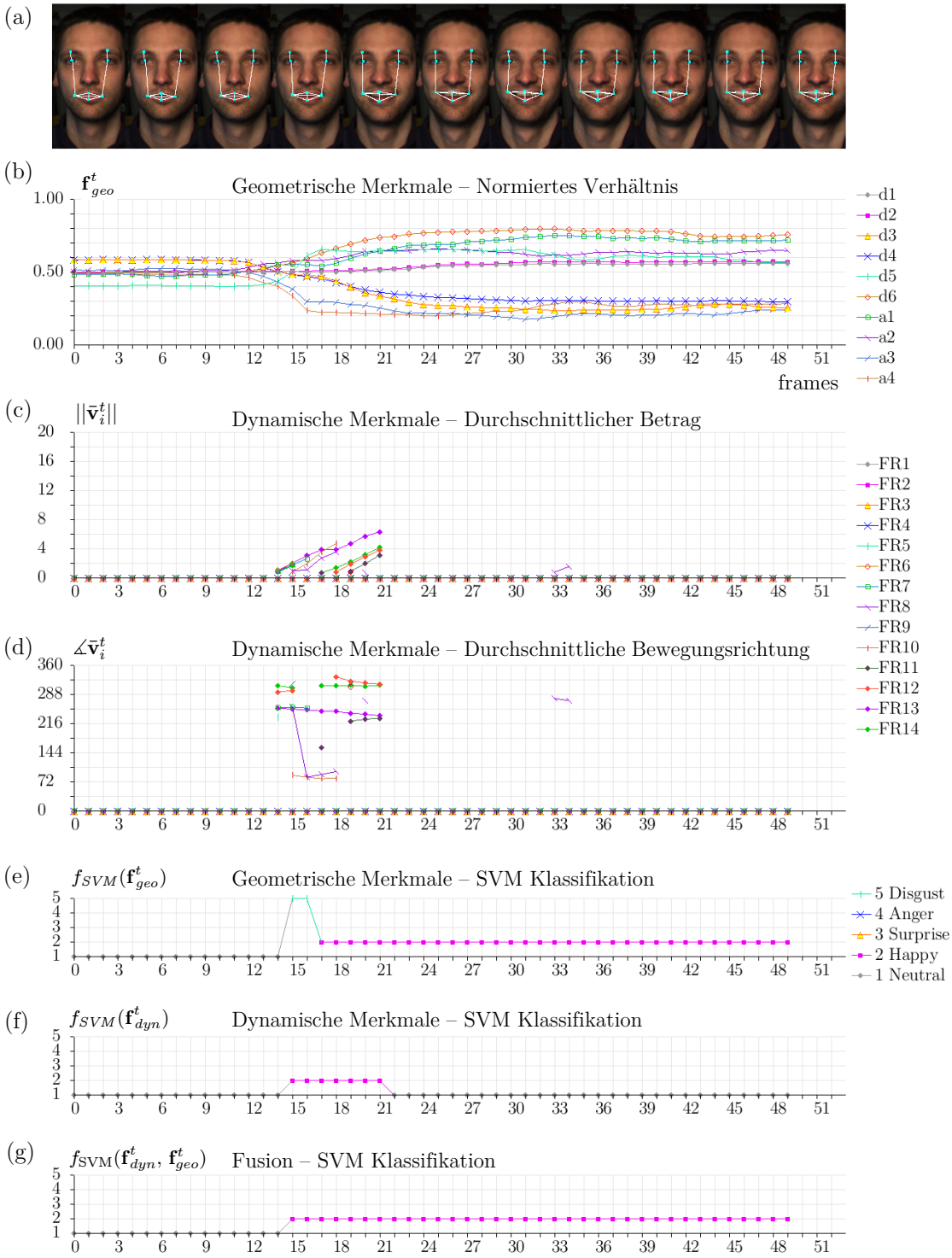


Abbildung 6-21: Beispielsequenz MD_2, Legende analog zu Abbildung 6-20, Mimikwechsel von $C_1 \rightarrow C_2$, typisch ist die Störung am Übergang.

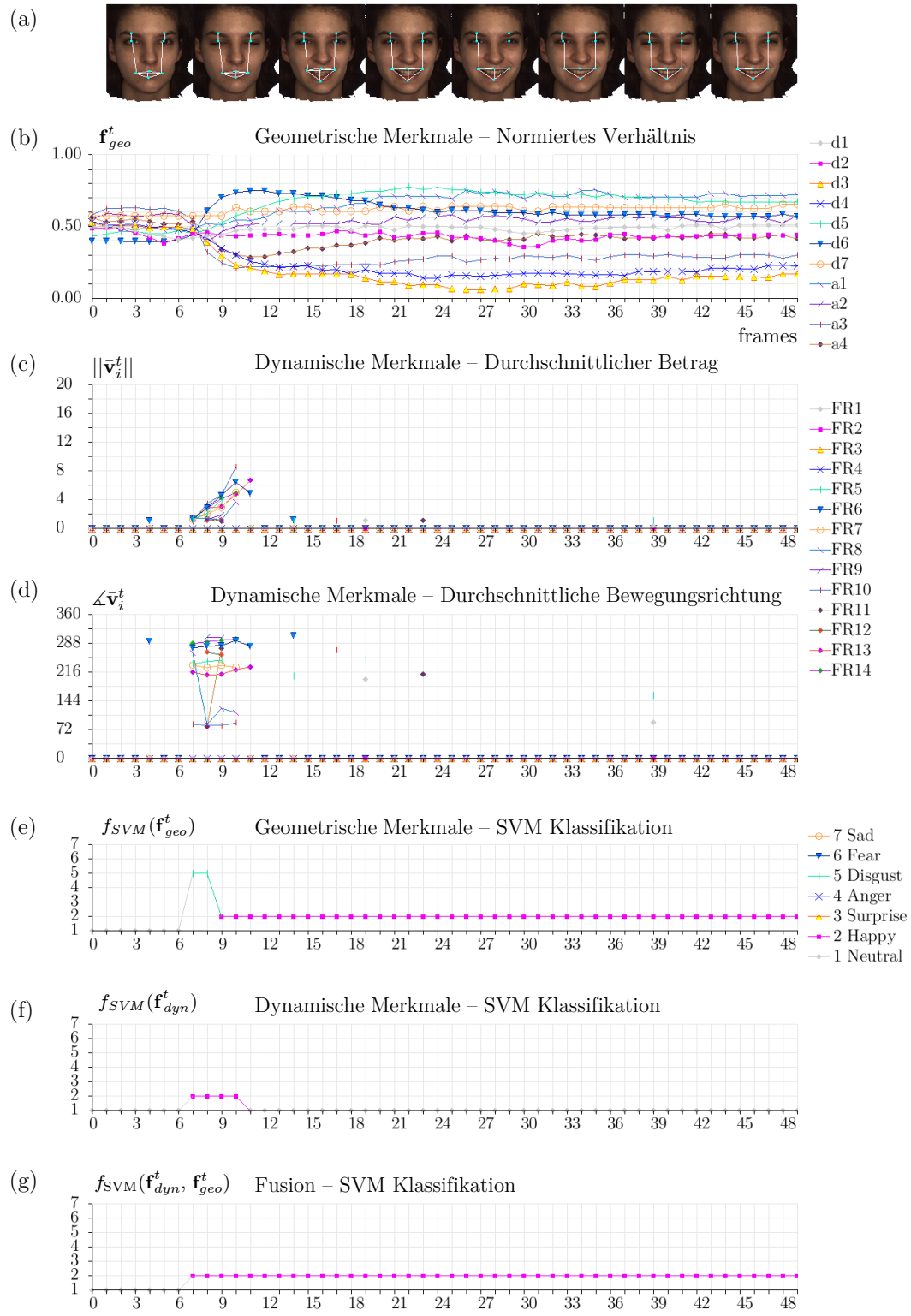


Abbildung 6-22: Beispielsequenz BU_1, Legende analog zu Abbildung 6-20, Mimikwechsel von $C_1 \rightarrow C_2$.

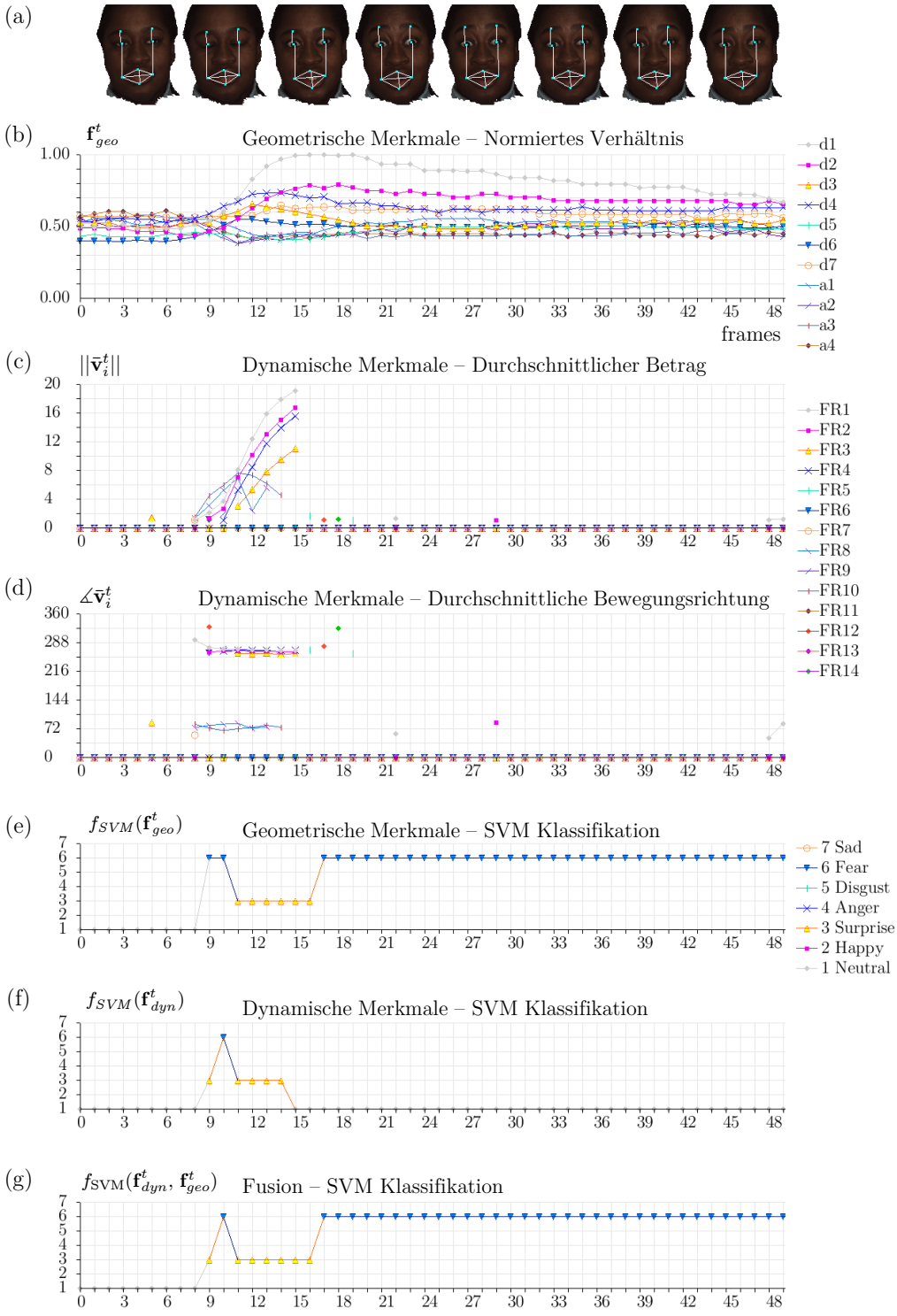


Abbildung 6-23: Beispielsequenz BU_2, mit Übergang von $C_1 \rightarrow C_6$, Verwechslung mit C_3 sowohl für dynamische als auch geometrische Merkmale. Dies ist auf die Merkmalsähnlichkeit für C_3 und C_6 zurückzuführen (Auswertung 6.2, 6.1).

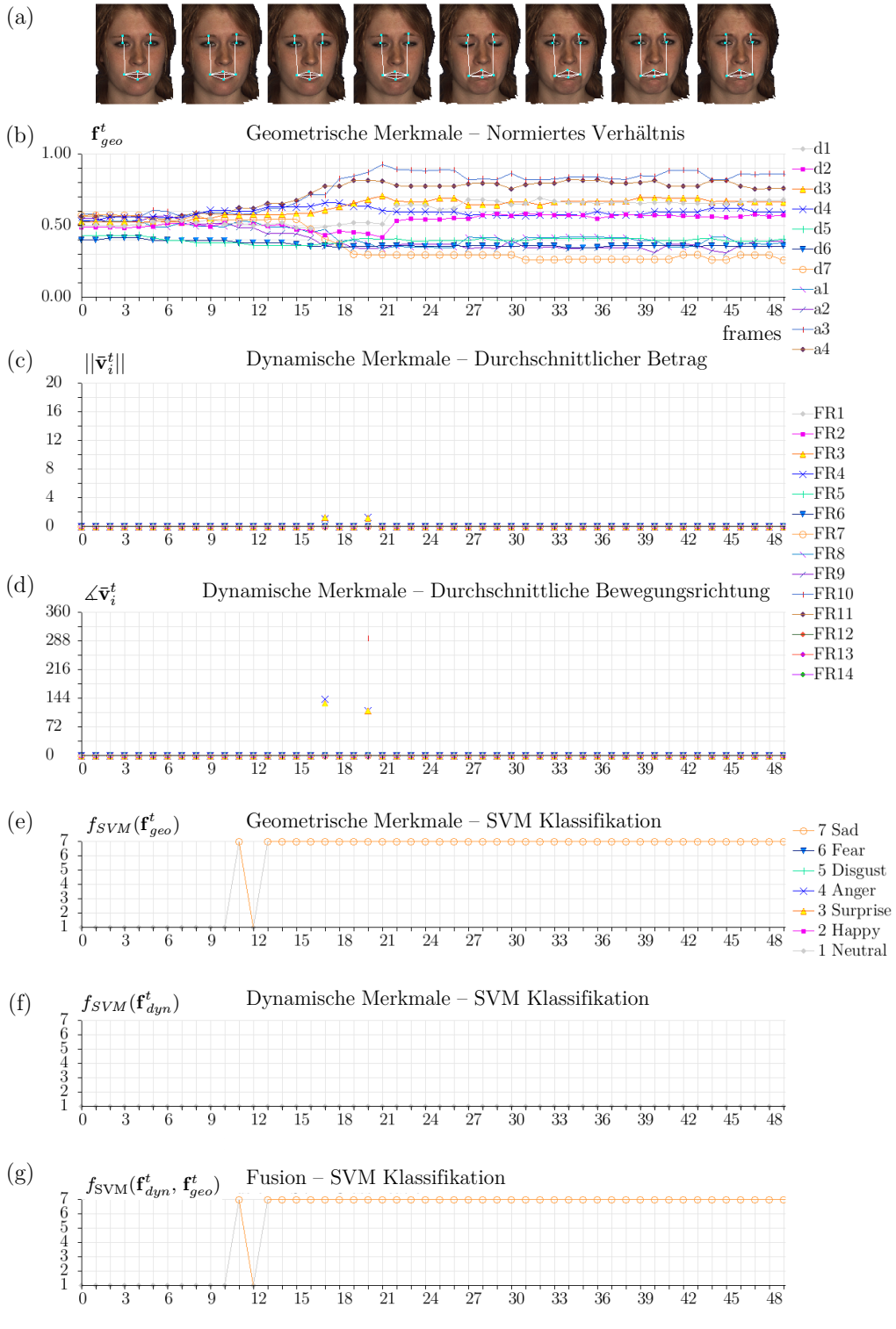


Abbildung 6-24: Beispielsequenz BU_3. Aufgrund der geringen Erregtheit der Versuchsperson werden zu wenig dynamische Merkmale ermittelt, somit kann auch keine Klassifikation mittels dynamischer Merkmale stattfinden.

6.6 Gegenüberstellung mit vergleichbaren Verfahren

Die mit der vorgeschlagenen Methode durch Kreuzvalidierung ermittelten Erkennungsraten mittels SVM liegen bei der Unterscheidung von 5 Klassen bei durchschnittlich 93 Prozent, bei Berücksichtigung von 7 Klassen bei fast 81.5 Prozent. Die Gegenüberstellung mit anderen Verfahren, welche ebenfalls 3D Merkmale einsetzen, zeigt dabei ein durchaus vergleichbares Abschneiden der vorgestellten Methode, jedoch bei reduziertem Aufwand und besserer Möglichkeit zur Automatisierung.
Das Verfahren nach Wang et al. [Wan06] wertet auf der Grundlage der 3D Datenbank BU-3DFE [Yin06] mit 60 Personen und sechs Klassen mimisch präsentierter Basisemotionen, die Variationen der Oberfläche mit Hilfe von 64 manuell selektierten Markerpunkten aus. Die Klassifikation beruht bei diesem Verfahren einzig auf 3D Formmerkmalen. Mittels linearer Diskriminanzanalyse (LDA) wurden dabei die höchsten Erkennungsraten erzielt. Tabelle 6-9 stellt die zugehörige Konfusionsmatrix dar.

Klasse C_i vs. **Klassifikation** $P(C_j)$	$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 - Neutral	-	-	-	-	-	-	-
C_2 – Freude	-	95.0	0.0s	0.0	0.8	3.8	0.4
C_3 – Überraschung	-	0.0	90.8	1.7	0.8	1.2	5.4
C_4 – Ärger	-	0.0	0.8	80.0	1.7	6.3	11.3
C_5 – Ekel	-	3.8	0.4	4.6	80.4	4.2	6.7
C_6 – Angst	-	12.5	2.1	0.0	2.5	75.0	7.9
C_7 – Trauer	-	0.0	5.8	8.3	2.5	2.9	80.4

Tabelle 6-9: Konfusionsmatrix in Prozent für das Verfahren nach [Wan06], basierend auf einer Klassifikation mittels LDA.

Die durchschnittliche Erkennungsrate liegt bei diesem Verfahren bei circa 84 Prozent. Die neutrale Klasse wurde hierbei nicht berücksichtigt, was die Erkennung grundsätzlich verbessert, da dies die Anzahl möglicher Verwechslungen verringert. Ungewöhnlich ist die untypische Vermischung zwischen den Klassen Überraschung und Trauer.
Das Verfahren nach Soyel&Demirel [Soy07] wertet auf der Grundlage von exakten 3D Modellen der BU-3DFE Datenbank Abstände zwischen 11 manuell selektier-

ten Merkmalspunkten aus und unterscheidet dabei sieben Mimikklassen. Zur Merkmalsnormierung wird hier die Breite des Gesichts verwendet, welche manuell aus der Frontalansicht bestimmt wird. Das Verfahren erzielt eine durchschnittliche Erkennungsrate von 91.3 Prozent. Tabelle 6-10 zeigt die Ergebnisse.
Die sehr gute Erkennung von über 90 Prozent für die Klassen Angst und Trauer sind hierbei außergewöhnlich, ebenso die Tatsache, dass für die Klasse Angst eine Verwechslung von 0 Prozent mit der Klasse Überraschung ermittelt wurde. Dies ist, bezugnehmend auf Ergebnisse anderer in der Literatur beschriebener Arbeiten, als auch mit den Ergebnissen dieser Arbeit, als schwer nachvollziehbar zu bewerten.
Trotz des durch Soyel&Demirel erreichten überragenden Ergebnisses; unter Berücksichtigung des Mehraufwandes der beiden genannten Verfahren, welche gewisse Ähnlichkeiten zu der hier beschriebenen Arbeit aufweisen, z.B. Merkmalspunkte und Abstände in 3D, ist der vorgeschlagene Ansatz auf der Grundlage geometrischer und dynamischer Merkmale aus zwei Gründen klar im Vorteil. Zum einen ist die fortlaufende exakte 3D Erfassung des Gesichts aufwendig, schwer zu realisieren und damit nicht wirklich praktikabel, weshalb in dieser Arbeit ein starres Modell vorgeschlagen wird, mit dem nachgewiesenermaßen, im Zusammenspiel mit der Merkmalsnormierung, vergleichbare Ergebnisse erzielt werden (Abschnitt 6.1.2). Zum anderen ist die automatische Detektion der beschriebenen acht Merkmalspunkte $\mathbf{I}_{fp}$ (4.19) im Bild mit Sicherheit robuster zu realisieren, als die teilweise deutlich größeren Punktmengen anderer Verfahren. Hinzu kommt, dass auf der Grundlage des beschriebenen Verfahrens dynamische Merkmale zur Verbesserung des Klassifikationsergebnisses beitragen können.

Klasse C_i vs. **Klassifikation** $P(C_j)$	$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 – Neutral	86.7	0.0	0.0	5.0	1.7	0.0	6.7
C_2 – Freude	0.0	95.0	0.0	0.0	1.7	3.3	0.0
C_3 – Überraschung	0.0	0.0	98.3	0.0	0.0	1.7	0.0
C_4 – Ärger	6.7	3.3	0.0	85.0	0.0	0.0	5.0
C_5 – Ekel	1.7	5.0	0.0	0.0	91.7	1.7	0.0
C_6 – Angst	1.7	3.3	0.0	1.7	0.0	91.7	1.7
C_7 – Trauer	3.7	0.0	0.0	3.7	1.9	0.0	90.7

Tabelle 6-10: Konfusionsmatrix in Prozent nach [Soy07], basierend auf einer Klassifikation von 3D Merkmalen durch künstliche neuronale Netze.

6.7 Diskussion

Die auf der Grundlage der beiden Datenbanken DB_{MD} und DB_{BU} durchgeführte Untersuchung der vorgeschlagenen Systemstruktur ermöglicht die Beantwortung der eingangs gestellten Fragen.

- Ermöglichen geometrische und dynamische Merkmale eine Erkennung und Unterscheidung der sechs gängigen Klassen emotional expressiver Mimik? Wo liegen hierbei die Stärken und Schwächen bzw. Grenzen?

Durch die Untersuchung der Merkmalsräume für geometrische und dynamische Merkmale wurde nachgewiesen, dass das in dieser Arbeit vorgeschlagene Verfahren grundsätzlich geeignet ist, Merkmale zur Unterscheidung von bis zu sechs Klassen emotional expressiver Mimik plus Neutral zu erzeugen. Dabei sind jedoch deutliche Unterschiede bei der Trennung der betrachteten Klassen zu erkennen. Insbesondere die Klassen C_1-C_3 zeigen sowohl für die geometrischen als auch dynamischen Merkmale in beiden untersuchten Datenbanken eine deutliche Separation und ermöglichen so eine akkurate Erkennung. Bei der Unterscheidung von fünf Klassen gilt dies ebenso für die Klassen C_4 und C_5, wenngleich hier eine geringere Erkennungsrate erzielt wurde. Durch Hinzunahme der Klassen Angst und Trauer kommt es beim sieben-Klassen-Szenario zu vermehrten Überlappungen im Merkmalsraum und somit zu einer reduzierten Erkennungsrate. Dabei ist die Ähnlichkeit der Merkmale oftmals auf die Gleichartigkeit der präsentierten Gesichtsausdrücke zurückzuführen, welche sich mitunter nur durch sehr kleine Variationen unterscheiden, wie z.B. zwischen den Klassen Angst und Überraschung. Da diese feinen Abweichungen nicht alle durch die derzeitig benutzten Merkmale erfasst werden, bestehen hier Überlappungen im Merkmalsraum. Folglich liegen in der exakten Bestimmung der Klassen C_6 und C_7, d.h. Angst und Trauer, die Grenzen des Verfahrens.

- Worin liegen Vor- und Nachteile der geometrischen und dynamischen Merkmale?

Zu den wichtigsten Vorteilen der dynamischen Merkmale gehören einerseits die Nutzung physiologisch motivierter Regionen, welche über das Gesicht verteilt zur Bewegungserfassung verwendet werden sowie eine hohe Sensitivität aufgrund der großen Anzahl an Messwerten, was generell eine schnellere Erkennung ermöglicht. Nach Subtraktion der globalen Kopfbewegung und Korrektur der Kopforientie-

rung repräsentieren die dynamischen Merkmale nur noch mimikrelevante Information. Dynamische Merkmale lassen sich naturgemäß nur während einer Veränderung der Mimik erfassen, was ihre Anwendung einschränkt. Dagegen haben die geometrischen Merkmale den Vorteil, dass sie zu jedem Zeitpunkt bestimmt werden können. Weiterhin ist hier keine Kompensation globaler Kopfbewegung bzw. Orientierung erforderlich. Da den geometrischen Merkmalen eine kleinere Anzahl an Messpunkten zugrunde liegt, sind diese weniger sensitiv während einer Änderung der Mimik. Dabei geben geometrischen Merkmale in der Übergangsphase aus der Sicht des Klassifikators zu Beginn ein mehrdeutiges Bild der aktuellen Mimik wieder, was zu Fehlklassifikationen führen kann.

- Führt die Integration geometrischer und dynamischer Merkmale zur Verbesserung bei der Mimikerkennung im Sinne der Klassifikationsergebnisse?

Mit Hilfe experimenteller Untersuchungen wurde gezeigt, dass die Integration geometrischer und dynamischer Merkmale, durch eine Fusion der Klassifikationsergebnisse, eine Verbesserung des Gesamtergebnisses erzielen kann. Wie bei der Analyse und Klassifikation dynamischer Merkmale dargelegt (s. Abschnitt 6.2), ist der Klassifikationserfolg dabei primär von der Separation der Klassen im Merkmalsraum abhängig sowie sekundär vom verwendeten Klassifikator. Als dritter Faktor kommt die erfolgreiche Erfassung dynamischer Merkmale hinzu, ohne die keine Erkennung erfolgen kann. Der Erfolg wird hierbei durch die Überschreitung der Aktivierungsschwelle v_{min} durch die Funktion $v_{sum}(t)$ signalisiert. Weiterhin ist auch die Fusion der Klassifikationsergebnisse nach (5.23) von der Überschreitung der Aktivierungsschwelle abhängig.

- In welchen Situationen ist eine solche Integration sinnvoll und wann nicht?

Wie in den experimentellen Untersuchungen gezeigt, ist die Zuhilfenahme dynamischer Merkmale sinnvoll, um Klassifikationsfehler aufgrund schlecht ausgeprägter geometrischer Merkmale in der Anfangsphase einer Mimikänderung zu korrigieren. Es hat sich hierbei jedoch auch gezeigt, dass für die Erkennung der Klassen Angst und Trauer keine Verbesserung erzielt werden konnte, aufgrund fehlender bzw. nicht ausreichend differenzierter dynamischer Merkmale. Daher ist eine Integration geometrischer und dynamischer Merkmale zur Erkennung dieser beiden Klassen nicht zweckmäßig.

- Führt die Nutzung starrer 3D Modelle zu hinreichend korrekten Merkmalen für die Mimikerkennung?

Die Erfassung geometrischer Merkmale beruht in dieser Arbeit auf der Verwendung eines starren Gesichtsmodells, welches nicht die tatsächliche, durch die aktuelle Mimik veränderte Oberfläche wiedergibt. Wie in Abschnitt 6.1.3 dargestellt, hat dies jedoch auf die Berechnung des normierten geometrischen Merkmalsvektors und die nachfolgende Klassifikation lediglich einen geringen Einfluss. Es wurde somit gezeigt, dass durch den vorgeschlagenen Ansatz zur Merkmalsberechnung keine aufwendige und nicht praktikable dauerhafte 3D Oberflächenmessung zur korrekten Klassifikation der Mimik erforderlich ist.

Kapitel 7

Schlussbetrachtungen

7.1 Zusammenfassung

Die vorliegende Arbeit befasst sich mit der Entwicklung eines neuen Ansatzes zur automatischen bildbasierten Mimikanalyse mit Fokus auf der Erkennung von sechs Basisemotionen (Freude, Überraschung, Ärger, Ekel, Angst und Trauer), für die der Psychologe Paul Ekman eine personen-, ethnien- und kulturübergreifende Universalität nachgewiesen hat.
Die hierzu durchgeführte Auswertung mono- und binokularer Farbbildfolgen basiert auf der Grundlage sogenannter geometrischer und dynamischer Merkmale. Die dynamischen Merkmale repräsentieren im vorgeschlagenen Konzept durch Mimik verursachte kurzzeitige Änderungen des Antlitzes, die sich durch Verschiebungen der Gesichtsoberfläche darstellen. Diese räumlich-zeitlichen Änderungen des Bildinhaltes werden auf der Grundlage des Optischen Flusses, inklusive temporaler Filterung erfasst. Wie in dieser Arbeit gezeigt, ermöglichen dynamische Merkmale eine schnelle und frühe Detektion von Bildänderungen was grundsätzlich zu einer Verbesserung bei der Erkennung führt. Um eine effiziente und genaue Auswertung zu realisieren, erfolgt die Erfassung dynamischer Merkmale nicht im gesamten Gesicht, sondern ausschließlich in sogenannten physiologisch motivierten Regionen, welche zur Selektion relevanter Bewegungsinformation verwendet werden.
Anders als die dynamischen Merkmale lassen sich geometrische Merkmale aus einem einzigen Bild bestimmen und repräsentieren den Zustand der Mimik zu einem aktuellen Zeitpunkt t. Die Merkmale beruhen dabei auf der Auswertung mimikrelevanter Merkmalspunkte. Hierzu werden durch den Einsatz von Gesichtsmodellen und photogrammetrischen Techniken geometrische Maße wie Abstände und Winkel in Weltkoordinaten berechnet. Bei deren Auswertung wird

immer Bezug zum Normalzustand, d.h. zur neutralen Mimik genommen. Diese wird im vorgeschlagenen Verfahren als bekannt vorausgesetzt und in einer initialen Bildaufnahme ermittelt. Während geometrische Merkmale zu jedem Zeitpunkt erfasst werden können, sind die dynamischen Merkmale aufgrund ihres differentiellen Charakters nur bei Änderungen der Mimik messbar.
Sowohl geometrische als dynamische auch Merkmale haben vor- und nachteilige Eigenschaften. Daher liegt dieser Arbeit die Hypothese zugrunde, dass durch eine integrierte Auswertung der beiden Merkmalsarten eine verbesserte Erkennungsrate erzielt werden kann. Zur Untersuchung dieses Forschungsgegenstandes wurde eine Systemstruktur vorgeschlagen, mit der eine Reihe grundsätzlicher Fragestellungen einhergeht. Diese wurden im Rahmen einer umfangreichen Validierung erörtert und die Möglichkeiten bzw. Stärken und Schwächen des Verfahrens im Ergebniskapitel dargestellt. Zur Bewertung der Erkennungsleistung bzw. -genauigkeit erfolgten in den Untersuchungen vergleichende Klassifikationen mit verschiedenen Entscheidern. Im Fokus standen dabei k-Nearest Neighbor, Multilayer Perceptron und Support Vector Machine.

Zur effektiven Erfassung und Auswertung geometrischer und dynamischer Merkmale wird in dieser Arbeit eine Systemstruktur vorgeschlagen, in der eine Reihe verschiedener Techniken aus Bildverarbeitung, Photogrammmetrie und Mustererkennung kombiniert wird. Insbesondere werden dabei personenspezifische Gesichtsmodelle zur Bestimmung von Mimikmerkmalen und aktueller Kopfpose verwendet. Die verwendeten Modelle beschreiben die Oberfläche des Gesichts der Versuchsperson in Form eines polygonalen Netzes und werden auf der Grundlage stereophotogrammetrischer Messungen erzeugt. Das Gesicht wird hierzu in einer 3D Messpunktwolke detektiert, indem diese in Cluster unterteilt wird. Im Anschluss erfolgt eine Vernetzung der Messpunkte durch Triangulation und eine Korrektur von Ausreißern, mit entsprechender Nachverarbeitung.
Zur Realisierung der Merkmalsextraktion werden zwei Ansätze, die sogenannte Gesichtsnormierung und als Erweiterung die Merkmalsnormierung, vorgestellt. Beim Ansatz der Gesichtsnormierung wird durch Auswertung von Stereobildfolgen und zugehörigen Punktwolken das Gesicht des Nutzers automatisch und unabhängig von der aktuellen Pose detektiert und die aktuelle Orientierung durch ICP Registrierung bestimmt. Durch den Einsatz personenspezifischer Gesichtsmodelle wird das Bild des Gesichts in eine normierte Darstellung überführt und auf diese Weise das Poseproblem für die anschließende Merkmalsauswertung überwunden. Die vorgeschlagene automatisierte Mimikanalyse auf der Grundlage der

Gesichtsnormierung und der nachfolgenden Merkmalsextraktion stellt eine Möglichkeit zur Erkennung von Gesichtsausdrücken mit Fokus auf sechs prototypischen Basisemotionen dar. Vorteilhaft ist dabei, dass aufgrund der Normierung keine Störungen bei der Bewegungserfassung durch Kopfbewegungen auftreten, was zur Bestimmung dichter Verschiebungsvektorfelder führt. Weiterhin besteht durch Redundanz in den Stereo-Beobachtungsdaten eine gewisse Unempfindlichkeit des Verfahrens gegenüber Störungen durch komplizierten Hintergrund und Rauscheinflüsse. Dennoch, der Robustheitsvorteil der fortlaufenden Stereomessung bringt gleichermaßen den Nachteil eines hohen Rechenaufwandes, was grundsätzlich die Anwendbarkeit erschwert. Im Rahmen dieser Arbeit wird daher ebenfalls der weiterentwickelte Ansatz der Merkmalsnormierung beschrieben.
Mit der Merkmalsnormierung wird eine Technik vorgestellt, die im Vergleich zur Gesichtsnormierung eine hohe Performanz mit den Vorteilen der 3D Merkmalsextraktion verbindet. Zur Reduktion des Rechenaufwandes erfolgt dabei eine Erfassung monokularer Farbbildsequenzen. Eine stereophotogrammetrische Aufnahme ist hierbei lediglich in einem initialen Schritt zur Erzeugung des Gesichtsmodells erforderlich. Durch Nutzung der Information über die Kopfpose, welche auf der Basis korrespondierender Ankerpunkte zwischen Modell und aktuellem Bild ermittelt wird, werden alle erfassten Merkmale aus dem Bild in eine 3D Repräsentation transformiert, normiert und schließlich einem Klassifikator zugeführt.
Indem stets unterschiedliche Personen in den Trainings und Testphasen verwendet wurden, konnte in der experimentellen Validierung die Personenunabhängigkeit der erfassten Mimikmerkmale gezeigt und die erreichte Qualität und Zuverlässigkeit des entwickelten Verfahrens nachgewiesen werden. Weiterhin ermöglicht die Validierung die Beantwortung der am Anfang der Arbeit aufgeworfenen, sich aus der vorgeschlagenen Systemstruktur ableitenden Fragen.
Die experimentellen Untersuchungen haben gezeigt, dass die angestrebten Zielstellungen dieser Forschungsarbeit mit der vorgeschlagenen Systemstruktur erreicht wurden. Es konnte gezeigt werden, dass eine Verbesserung bei der Mimikerkennung durch eine integrierte Auswertung geometrischer und dynamischer Merkmale möglich ist. Bestehende Grenzen wurden aufgezeigt.
Bei Vergleichen mit anderen aktuellen Verfahren wurden durch die vorgeschlagene Methode adäquate bzw. bessere Erkennungsraten, bei gleichzeitig geringerer Komplexität (Abschnitt 6.6) erzielt. Wie in den experimentellen Untersuchungen belegt, ermöglicht die vorgeschlagene 2D/3D Datenauswertung mit entsprechender Normierung bereits durch Verwendung eines verhältnismäßig kleinen Merkmalssatzes eine qualitativ hochwertige Erkennung expressiver Mimik. Andere Ver-

fahren erfordern hierzu deutlich aufwendigere Merkmale, was wiederum eine Automatisierung erschwert. Mit den erzielten Ergebnissen stellt diese Arbeit daher einen idealen Ausgangspunkt für weitere Entwicklungen und Experimente dar und leistet somit einen konstruktiven Beitrag zum aktuellen Stand der Forschung.

7.2 Ausblick

Auf der Grundlage des vorgestellten Verfahrens wurden bereits erste Erweiterungen realisiert und Untersuchungen durchgeführt, die über die Erkennung mimisch präsentierter Basisemotionen hinausgehen.
In zukünftigen Arbeiten bietet es sich an, zusätzliche Mimikklassen zu berücksichtigen. In empirischen Untersuchungen konnte gezeigt werden, dass auf der Grundlage des vorgestellten Verfahrens prinzipiell auch Gesichtsausdrücke von akut auftretenden Schmerzen erkannt werden können. Hierzu wurden Experimente im Rahmen einer Kooperation mit der Universitätsklinik für Anästhesiologie und Intensivtherapie (KAIT, Magdeburg) durchgeführt. Die bisher erzielten Ergebnisse sind vielversprechend und konnten bereits publiziert werden [Nie09].
Die vorgeschlagene Systemstruktur beinhaltet eine Vielzahl von Komponenten, die für ein robustes vollautomatisches Funktionieren des Gesamtsystems noch der Optimierung bedürfen. Insbesondere ist hier die Detektion der Merkmalspunkte im Bild zu nennen. Ein geeignetes Verfahren könnte dabei ein um zusätzliche Merkmale erweiterter kaskadenbasierter AdaBoost Gesichtsmerkmalspunktdetektor auf der Basis von Haar-Like Features darstellen.
Als ein weiterer Punkt zukünftiger Untersuchungen bietet sich eine vereinfachte Erstellung von Gesichtsmodellen auf der Grundlage generischer Modelle an, welche sich durch Frontal- und Seitenaufnahmen der Versuchsperson parametrieren lassen. Auf diese Weise ließe sich die je nach vorhandener Hardware eventuell mit Umständen verbundene Stereobildaufnahme zur Gesichtsmodellerzeugung ersetzen.
Für zukünftige Untersuchungen wird vermutlich neben der Erkennung der Mimik auch eine zugehörige Quantifizierung Forschungsgegenstand sein. In der Literatur ist dieser Punkt bisher nur wenig beschrieben worden. Das in dieser Arbeit vorgeschlagene Konzept stellt hierzu eine gute Grundlage dar. Das gleiche gilt für eine multimodale Datenauswertung, beispielsweise durch Hinzuziehen prosodischer Kontextinformationen.

Kapitel 8

Anhang

Ergänzend zu Kapitel 6 werden nachfolgend Klassifikationsergebnisse aus weiteren Untersuchungen gegeben. Darüber hinaus werden der in der vorgeschlagenen Systemstruktur benutzte Algorithmus zur Gruppierung der Messpunkte sowie Transformationsmatrizen dargestellt.

8.1 Klassifikationsergebnisse

Die in den folgenden zwei Unterpunkten präsentierten Ergebnisse basieren auf den Ansätzen zur Gesichts- und Merkmalsnormierung (Abschnitt 5.1, 5.2).

8.1.1 Klassifikation nach dem Ansatz zur Gesichtsnormierung

Die ermittelten Erkennungsraten nach dem Ansatz zur Gesichtsnormierung beruhen auf Analysen der Datenbank DB_{ULM}, welche hierzu in zwei Mengen UL_1 und UL_2 mit jeweils anderen Versuchspersonen aufgeteilt wurde. Zum besseren Vergleich wurden dabei wie bei der Analyse dynamischer Merkmale nach dem Ansatz zur Merkmalsnormierung (s. Abschnitt 6.2.2) Szenarien zur Unterscheidung von vier bzw. sechs Klassen berücksichtigt (Tabelle 8-1 bis Tabelle 8-4).
Die Erkennung der Klasse Neutral (C_1) ist durch Auswertung dynamischer Merkmale direkt nicht ohne weiteres möglich, da bei neutraler Mimik davon ausgegangen wird, dass die Versuchspersonen nur leichte Variationen im Gesichtsausdruck zeigen. Bei neutraler Mimik wird daher angenommen, dass die zur Klassifikation vorausgesetzte Aktivierungsschwelle v_{min} nicht durch die Aktivierungsfunktion $v_{sum}(t)$ (4.16) überschritten wird. Grundsätzlich setzt dies jedoch voraus, dass die Versuchspersonen nicht sprechen, das Gesicht verziehen, gähnen, etc., da so unbekannte Bewegungsmuster entstehen und es zu Fehlern bei der Erkennung kommt.

Klasse / Klassifikation UL_1 vs. UL_2		$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$
C_2 – Freude	k-NN	50.0	0.0	5.6	44.4
	MLP	27.8	0.0	5.6	66.7
	SVM	94.4	0.0	0.0	5.6
C_3 – Überraschung	k-NN	0.0	100.0	0.0	0.0
	MLP	0.0	100.0	0.0	0.0
	SVM	0.0	100.0	0.0	0.0
C_4 – Ärger	k-NN	0.0	0.0	90.0	10.0
	MLP	16.7	16.7	61.1	5.6
	SVM	0.0	0.0	83.3	16.7
C_5 – Ekel	k-NN	45.0	0.0	25.0	30.0
	MLP	50.0	0.0	25.0	25.0
	SVM	10.0	0.0	25.0	65.0

Tabelle 8-1: Konfusionsmatrix in Prozent; basierend auf dynamischen Merkmalen, berechnet mittels Gesichtsnormierung, k-NN, MLP und SVM Klassifikation, Unterscheidung von 4 Klassen, Training mit Datensatz UL_2, Test von UL_1.

Klasse / Klassifikation UL_1 vs. UL_2		$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_2 – Freude	k-NN	50.0	0.0	5.6	44.4	0.0	0.0
	MLP	50.0	5.6	5.6	38.9	0.0	0.0
	SVM	94.4	0.0	0.0	5.6	0.0	0.0
C_3 – Überraschung	k-NN	0.0	77.8	0.0	0.0	22.2	0.0
	MLP	0.0	83.3	16.7	0.0	0.0	0.0
	SVM	0.0	77.8	0.0	0.0	22.2	0.0
C_4 – Ärger	k-NN	0.0	0.0	100.0	0.0	0.0	0.0
	MLP	0.0	0.0	94.4	5.6	0.0	0.0
	SVM	0.0	0.0	83.3	16.7	0.0	0.0
C_5 – Ekel	k-NN	45.0	0.0	25.0	30.0	0.0	0.0
	MLP	0.0	0.0	50.0	50.0	0.0	0.0
	SVM	10.0	0.0	25.0	65.0	0.0	0.0
C_6 – Angst	k-NN	0.0	33.3	0.0	33.3	33.3	0.0
	MLP	0.0	0.0	33.3	66.7	0.0	0.0
	SVM	0.0	33.3	0.0	27.7	39.0	0.0
C_7 – Trauer	k-NN	0.0	20.0	0.0	0.0	0.0	80.0
	MLP	0.0	13.3	33.3	53.3	0.0	0.0
	SVM	0.0	13.3	0.0	0.0	0.0	86.7

Tabelle 8-2: Konfusionsmatrix; dynamische Merkmale, Gesichtsnormierung, k-NN, MLP, SVM Klassifikation, 6 Klassen, Training mit UL_2, Test von UL_1.

Klasse / Klassifikation UL_2 vs. UL_1		P(C_2)	P(C_3)	P(C_4)	P(C_5)
C_2 – Freude	k-NN	53.3	0.0	0.0	46.7
	MLP	53.3	0.0	0.0	46.7
	SVM	53.3	0.0	0.0	46.7
C_3 – Überraschung	k-NN	0.0	100.0	0.0	0.0
	MLP	0.0	100.0	0.0	0.0
	SVM	0.0	100.0	0.0	0.0
C_4 – Ärger	k-NN	0.0	0.0	76.5	23.5
	MLP	0.0	0.0	52.9	47.1
	SVM	0.0	0.0	76.5	23.5
C_5 – Ekel	k-NN	46.7	0.0	20.0	33.3
	MLP	46.7	0.0	0.0	53.3
	SVM	26.7	0.0	0.0	73.3

Tabelle 8-3: Konfusionsmatrix; dynamische Merkmale, Gesichtsnormierung, k-NN, MLP, SVM Klassifikation, 4 Klassen, Training mit UL_1, Test von UL_2.

Klasse / Klassifikation UL_2 vs. UL_1		P(C_2)	P(C_3)	P(C_4)	P(C_5)	P(C_6)	P(C_7)
C_2 – Freude	k-NN	53.3	0.0	0.0	46.7	0.0	0.0
	MLP	53.3	0.0	0.0	46.7	0.0	0.0
	SVM	53.3	0.0	0.0	46.7	0.0	0.0
C_3 – Überraschung	k-NN	0.0	100.0	0.0	0.0	0.0	0.0
	MLP	0.0	100.0	0.0	0.0	0.0	0.0
	SVM	0.0	100.0	0.0	0.0	0.0	0.0
C_4 – Ärger	k-NN	0.0	0.0	76.5	23.5	0.0	0.0
	MLP	0.0	5.9	0.0	94.1	0.0	0.0
	SVM	0.0	0.0	76.5	23.5	0.0	0.0
C_5 – Ekel	k-NN	46.7	0.0	20.0	33.3	0.0	0.0
	MLP	13.3	0.0	0.0	86.7	0.0	0.0
	SVM	46.7	0.0	0.0	53.3	0.0	0.0
C_6 – Angst	k-NN	0.0	56.3	0.0	0.0	43.4	0.0
	MLP	0.0	37.50	25.0	37.50	0.0	0.0
	SVM	0.0	37.5	0.0	0.0	62.50	0.0
C_7 – Trauer	k-NN	0.0	0.0	6.7	0.0	0.0	93.3
	MLP	73.3	0.0	26.7	0.0	0.0	0.0
	SVM	0.0	0.0	0.0	0.0	0.0	100.0

Tabelle 8-4: Konfusionsmatrix; dynamische Merkmale, Gesichtsnormierung, k-NN, MLP, SVM Klassifikation, 6 Klassen, Training mit UL_1, Test von UL_2.

Die Ergebnisse zeigen, dass auch in dieser Untersuchung der SVM Klassifikator durchschnittlich am besten abgeschnitten hat. Die verhältnismäßig guten Erkennungsraten der Klassen C_2, C_6 und C_7 zeigen, wie expressiv die Präsentation der gezeigten Emotionen in der untersuchten Datenbank erfolgte. Insgesamt zeigt sich bei der Auswertung mittels Gesichtsnormierung ein leicht besseres Abschneiden zur Analyse dynamischer Merkmale auf der Grundlage der Merkmalsnormierung. Dies ist auf die größere Sensitivität der Merkmale zurückzuführen, da die Auswertung normierter Gesichter keine Unterdrückung globaler Kopfbewegung erfordert.

8.1.2 Klassifikation nach dem Ansatz zur Merkmalsnormierung

Die hier dargestellten experimentellen Ergebnisse beruhen auf Analysen der Datenbank BU-4DFE. Hierzu wurden entsprechend Tabelle 8-5 verschiedene Teilmengen mit unterschiedlicher Probandenverteilung entnommen. Insbesondere wurden dazu zwei mit männlichen und weiblichen Probanden gemischte Datensätze D_1 und D_2 verwendet. Hingegen wurde auf Grundlage von D_3 (nur Frauen) und D_4 (nur Männer) die geschlechterübergreifende Universalität des vorgeschlagenen Verfahrens nachgewiesen.

Datensatz	Anzahl Probanden (W)	Anzahl Probanden (M)	Datensamples Dynamisch u. Geometrisch
D_1 - Mix 1	25	24	9693
D_2 - Mix 2	33	19	9941
D_3 - Frauen	58	-	11575
D_4 - Männer	-	43	8059

Tabelle 8-5: Untersuchte Datensätze der BU-4DFE Datenbank.

Nachfolgend werden Klassifikationsergebnisse (k-NN, MLP, SVM) für die Datensätze D_1 bis D_4 entsprechend Tabelle 8-5 gegeben. Dargestellt werden dabei die Konfusionsmatrizen bezüglich der Auswertung der Datensätze D_1 bis D_4 auf der Grundlage dynamischer und geometrischer Merkmale (Tabelle 8-6 bis Tabelle 8-17). Die erzielten Erkennungsraten entsprechen im Wesentlichen den in Kapitel 6 vorgestellten Ergebnissen. Die enthaltenen Verwechslungen lassen sich auf die in Abschnitt 6.1.1 und 6.2.1 beschriebenen Überlappungen im Merkmalsraum zurückführen. Die Ergebnisse zwischen den Datensätzen D_3/D_4 zeigen, dass keine spezifischen Unterschiede des Verfahrens bezüglich weiblicher bzw. männlicher Probanden vorliegen.

8.1.2.1 Konfusionsmatrix D_2 vs. D_1

Klasse / Klassifikation D_2 vs. D_1 / **k-NN**		P(C_1)	P(C_2)	P(C_3)	P(C_4)	P(C_5)	P(C_6)	P(C_7)
C_1 – Neutral	G	95.59	0.07	0.00	1.14	0.71	1.21	1.28
	D	-	-	-	-	-	-	-
C_2 - Freude	G	1.83	94.79	0.00	0.00	0.41	2.97	0.00
	D	-	56.31	0.97	0.97	21.36	18.45	1.94
C_3 - Überraschung	G	0.00	0.08	86.35	0.00	0.50	9.51	3.56
	D	-	0.00	79.71	0.00	2.90	16.67	0.72
C_4 – Ärger	G	5.08	0.15	0.00	62.05	8.55	0.92	23.25
	D	-	5.41	0.00	35.14	35.14	5.41	18.92
C_5 – Ekel	G	0.08	2.28	6.24	14.9	67.07	4.11	5.32
	D	-	13.91	2.61	15.65	58.26	5.22	4.35
C_6 – Angst	G	1.12	13.84	11.36	3.12	15.68	45.84	9.04
	D	-	12.00	34.67	6.67	13.33	21.33	12.00
C_7 – Trauer	G	8.13	1.74	0.41	14.01	11.44	6.22	58.04
	D	-	0.00	4.35	30.43	4.35	8.7	52.17

Tabelle 8-6: Konfusionsmatrix in Prozent; Merkmalsnormierung, k-NN für geometrische (G) und dynamische (D) Merkmale, Training mit D_1, Test von D_2.

Klasse / Klassifikation D_2 vs. D_1 / **MLP**		P(C_1)	P(C_2)	P(C_3)	P(C_4)	P(C_5)	P(C_6)	P(C_7)
C_1 - Neutral	G	98.72	0.00	0.00	0.64	0.00	0.21	0.43
	D	-	-	-	-	-	-	-
C_2 – Freude	G	2.10	87.96	0.14	0.00	0.07	9.74	0.00
	D	-	73.79	4.85	2.91	18.45	0.00	0.00
C_3 - Überraschung	G	0.17	0.00	91.81	0.00	0.17	5.87	1.99
	D	-	4.35	94.2	0.00	1.45	0.00	0.00
C_4 – Ärger	G	9.08	1.39	0.00	63.05	3.00	0.00	23.48
	D	-	16.22	6.76	59.46	17.57	0.00	0.00
C_5 – Ekel	G	8.59	1.52	4.18	10.80	72.55	2.36	0.00
	D	-	33.04	2.61	13.91	50.43	0.00	0.00
C_6 – Angst	G	2.80	9.28	11.60	2.96	14.88	52.08	6.40
	D	-	12.00	62.67	4.00	21.33	0.00	0.00
C_7 – Trauer	G	23.88	1.82	2.90	12.69	4.81	3.32	50.58
	D	-	13.04	43.48	39.13	4.35	0.00	0.00

Tabelle 8-7: Konfusionsmatrix in Prozent; Merkmalsnormierung, MLP für geometrische (G) und dynamische (D) Merkmale, Training mit D_1, Test von D_2.

Klasse / Klassifikation D_2 vs. D_1 / **SVM**		$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 – Neutral	G	93.67	0.07	0.00	0.57	0.00	1.71	3.98
	D	-	-	-	-	-	-	-
C_2 – Freude	G	1.83	94.25	1.01	0.00	0.20	2.70	0.00
	D	-	80.58	0.97	1.94	16.5	0.00	0.00
C_3 – Überraschung	G	0.08	0.00	92.8	0.00	0.25	6.70	0.17
	D	-	2.17	81.16	0.00	2.17	12.32	2.17
C_4 – Ärger	G	2.23	2.54	0.00	68.51	3.62	0.69	22.4
	D	-	4.05	6.76	44.59	39.19	0.00	5.41
C_5 – Ekel	G	0.15	2.05	4.71	8.75	79.62	2.97	1.75
	D	-	12.17	0.87	13.91	71.3	1.74	0.00
C_6 – Angst	G	0.64	10.48	8.16	2.64	11.6	64.48	2.00
	D	-	6.67	57.33	5.33	17.33	8.00	5.33
C_7 – Trauer	G	10.45	1.91	0.00	11.53	4.81	5.22	66.09
	D	-	4.35	8.70	34.78	26.09	0.00	26.09

Tabelle 8-8: Konfusionsmatrix in Prozent; Merkmalsnormierung, SVM für geometrische (G) und dynamische (D) Merkmale, Training mit D_1, Test von D_2.

8.1.2.2 Konfusionsmatrix D_1 vs. D_2

Klasse / Klassifikation D_1 vs. D_2 / **k-NN**		$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 – Neutral	G	97.96	0.00	0.00	0.93	0.17	0.12	0.81
	D	-	-	-	-	-	-	-
C_2 - Freude	G	0.91	85.73	0.15	1.67	2.05	5.39	4.10
	D	-	71.25	0.00	1.25	20.00	7.50	0.00
C_3 - Überraschung	G	1.98	1.11	78.35	0.00	2.30	12.21	4.04
	D	-	3.94	72.44	0.79	4.72	17.32	0.79
C_4 – Ärger	G	5.10	0.00	0.00	77.85	12.94	0.53	3.58
	D	-	7.84	1.96	42.16	28.43	11.76	7.84
C_5 – Ekel	G	2.73	0.50	2.52	9.50	71.08	5.25	8.42
	D	-	18.05	3.01	13.53	57.14	6.02	2.26
C_6 – Angst	G	20.96	13.65	9.48	3.83	8.52	40.52	3.04
	D	-	12.35	41.98	2.47	12.35	25.93	4.94
C_7 – Trauer	G	22.32	0.16	3.43	28.21	5.72	0.08	40.07
	D	-	4.55	2.27	34.09	13.64	20.45	25.00

Tabelle 8-9: Konfusionsmatrix in Prozent; Merkmalsnormierung, k-NN für geometrische (G) und dynamische (D) Merkmale, Training mit D_2, Test von D_1.

Klasse / Klassifikation D_1 vs. D_2 / **MLP**		$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 – Neutral	G	99.19	0.00	0.00	0.23	0.06	0.06	0.47
	D	-	-	-	-	-	-	-
C_2 - Freude	G	1.59	78.28	0.00	0.00	5.54	14.58	0.00
	D	-	91.25	0.00	1.25	7.50	0.00	0.00
C_3 - Überraschung	G	4.12	2.93	81.36	0.00	4.84	6.74	0.00
	D	-	6.30	81.89	4.72	7.09	0.00	0.00
C_4 – Ärger	G	7.91	0.00	0.00	77.47	10.43	0.30	3.88
	D	-	2.94	4.90	50.00	42.16	0.00	0.0
C_5 – Ekel	G	6.04	0.00	2.37	2.81	78.63	3.24	6.91
	D	-	13.53	3.01	18.8	64.66	0.00	0.00
C_6 – Angst	G	22.26	4.87	12.43	0.00	16.35	42.43	1.65
	D	-	22.22	54.32	14.81	8.64	0.00	0.00
C_7 – Trauer	G	36.63	0.00	0.00	24.12	1.72	0.65	36.88
	D	-	13.64	20.45	43.18	22.73	0.00	0.00

Tabelle 8-10: Konfusionsmatrix in Prozent; Merkmalsnormierung, MLP für geometrische (G) und dynamische (D) Merkmale, Training mit D_2, Test von D_1.

Klasse / Klassifikation D_1 vs. D_2 / **SVM**		$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 – Neutral	G	98.37	0.06	0.00	0.41	0.00	0.17	0.99
	D	-	-	-	-	-	-	-
C_2 - Freude	G	0.08	77.68	0.00	0.00	0.08	22.17	0.00
	D	-	82.5	0.00	6.25	11.25	0.00	0.00
C_3 - Überraschung	G	1.98	0.95	84.69	0.00	3.89	8.49	0.00
	D	-	1.57	76.38	15.75	2.36	3.94	0.00
C_4 – Ärger	G	3.27	0.00	0.00	84.09	8.83	0.53	3.27
	D	-	6.86	0.98	63.73	27.45	0.98	0.00
C_5 – Ekel	G	1.94	3.38	1.51	3.81	78.42	4.17	6.76
	D	-	13.53	1.5	31.58	52.63	0.75	0.00
C_6 – Angst	G	13.57	4.78	10.00	1.22	13.30	54.52	2.61
	D	-	17.28	59.26	8.64	7.41	7.41	0.00
C_7 – Trauer	G	28.29	0.00	0.00	26.08	1.64	0.08	43.91
	D	-	4.55	11.36	50.00	11.36	18.18	4.55

Tabelle 8-11: Konfusionsmatrix in Prozent; Merkmalsnormierung, SVM für geometrische (G) und dynamische (D) Merkmale, Training mit D_2, Test von D_1.

8.1.2.3 Konfusionsmatrix D_4 vs. D_3

Klasse / Klassifikation D_4 vs. D_3 / **k-NN**		P(C_1)	P(C_2)	P(C_3)	P(C_4)	P(C_5)	P(C_6)	P(C_7)
C_1 – Neutral	G	97.86	0.08	0.00	0.66	0.66	0.41	0.33
	D	-	-	-	-	-	-	-
C_2 - Freude	G	1.16	77.14	1.16	0.09	3.11	12.46	4.89
	D	-	65.85	0.00	0.00	19.51	14.63	0.00
C_3 - Überraschung	G	0.44	0.09	81.2	0.09	0.26	16.06	1.85
	D	-	0.00	83.48	0.87	0.00	14.78	0.87
C_4 – Ärger	G	4.72	1.31	0.00	63.37	11.8	0.00	18.79
	D	-	9.68	0.00	43.55	41.94	3.23	1.61
C_5 – Ekel	G	0.36	1.16	1.97	14.59	69.65	4.48	7.79
	D	-	16.67	2.38	16.67	53.57	4.76	5.95
C_6 – Angst	G	9.47	4.48	3.87	8.04	11.81	53.36	8.96
	D	-	5.56	53.70	3.70	12.96	20.37	3.0
C_7 – Trauer	G	19.16	0.00	0.41	29.25	6.69	3.4	41.09
	D	-	0.00	5.26	42.11	5.26	10.53	36.84

Tabelle 8-12: Konfusionsmatrix in Prozent; Merkmalsnormierung, k-NN für geometrische (G) und dynamische (D) Merkmale, Training mit D_3, Test von D_4.

Klasse / Klassifikation D_4 vs. D_3 / **MLP**		P(C_1)	P(C_2)	P(C_3)	P(C_4)	P(C_5)	P(C_6)	P(C_7)
C_1 - Neutral	G	99.84	0.00	0.00	0.00	0.08	0.08	0.00
	D	-	-	-	-	-	-	-
C_2 – Freude	G	0.18	60.85	0.71	0.00	8.19	30.07	0.00
	D	-	85.37	0.00	4.88	9.76	0.00	0.00
C_3 - Überraschung	G	2.29	3.44	85.79	0.00	5.65	1.68	1.15
	D	-	0.00	95.65	3.48	0.87	0.00	0.00
C_4 – Ärger	G	10.49	0.00	0.00	71.59	9.53	0.44	7.95
	D	-	0.00	0.00	45.16	54.84	0.00	0.00
C_5 – Ekel	G	11.19	0.00	3.13	0.36	80.57	0.09	4.66
	D	-	15.48	1.19	35.71	47.62	0.00	0.00
C_6 – Angst	G	17.41	4.79	9.37	0.20	25.97	41.45	0.81
	D	-	11.11	74.07	5.56	9.26	0.00	0.00
C_7 – Trauer	G	37.9	0.00	0.31	16.17	2.16	5.36	38.11
	D	-	10.53	10.53	63.16	15.79	0.00	0.00

Tabelle 8-13: Konfusionsmatrix in Prozent; Merkmalsnormierung, MLP für geometrische (G) und dynamische (D) Merkmale, Training mit D_3, Test von D_4.

Klasse / Klassifikation D_4 vs. D_3 / **SVM**		$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 – Neutral	G	96.62	0.08	0.00	0.41	0.00	0.16	2.72
	D	-	-	-	-	-	-	-
C_2 – Freude	G	0.00	71.26	0.09	0.00	3.11	25.53	0.00
	D	-	82.93	0.00	7.32	9.76	0.00	0.00
C_3 – Überraschung	G	0.09	0.00	87.73	0.00	3.44	8.21	0.53
	D	-	0.00	95.65	2.61	0.00	0.87	0.87
C_4 – Ärger	G	1.92	1.49	0.00	83.3	4.81	0.00	8.48
	D	-	6.45	1.61	54.84	35.48	0.00	1.61
C_5 – Ekel	G	0.18	1.61	2.42	0.90	83.62	6.54	4.74
	D	-	15.48	1.19	32.14	51.19	0.00	0.00
C_6 – Angst	G	6.52	1.43	5.4	2.24	14.56	67.11	2.75
	D	-	12.96	70.37	5.56	7.41	3.70	0.00
C_7 – Trauer	G	21.73	0.00	3.71	23.07	0.51	2.47	48.51
	D	-	15.79	0.00	47.37	5.26	15.79	15.79

Tabelle 8-14: Konfusionsmatrix in Prozent; Merkmalsnormierung, SVM für geometrische (G) und dynamische (D) Merkmale, Training mit D_3, Test von D_4.

8.1.2.4 Konfusionsmatrix D_3 vs. D_4

Klasse / Klassifikation D_3 vs. D_4 / **k-NN**		$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 – Neutral	G	96.55	0.16	0.05	1.10	0.31	0.47	1.36
	D	-	-	-	-	-	-	-
C_2 - Freude	G	1.79	94.98	0.00	1.26	0.30	1.67	0.00
	D	-	47.89	0.70	2.11	33.80	14.08	1.41
C_3 - Überraschung	G	1.05	2.02	88.78	0.00	0.90	6.66	0.60
	D	-	3.33	74.67	2.00	4.67	14.67	0.67
C_4 – Ärger	G	5.11	0.2	0.00	69.57	9.67	5.72	9.73
	D	-	1.75	0.00	37.72	41.23	7.02	12.28
C_5 – Ekel	G	2.46	2.77	4.03	10.39	70.53	6.11	3.72
	D	-	8.54	0.00	20.12	63.41	6.10	1.83
C_6 – Angst	G	7.40	23.27	14.74	2.19	10.01	39.92	2.47
	D	-	9.80	40.20	2.94	19.61	22.55	4.90
C_7 – Trauer	G	9.81	1.44	0.07	22.91	8.64	9.19	47.94
	D	-	0.00	12.50	8.33	22.92	12.50	43.75

Tabelle 8-15: Konfusionsmatrix in Prozent; Merkmalsnormierung, k-NN für geometrische (G) und dynamische (D) Merkmale, Training mit D_4, Test von D_3.

Klasse / Klassifikation D_3 vs. D_4 / **MLP**		$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 – Neutral	G	98.85	0.10	0.00	0.89	0.10	0.05	0.00
	D	-	-	-	-	-	-	-
C_2 – Freude	G	1.73	97.01	0.00	0.00	0.06	1.20	0.00
	D	-	71.13	0.70	0.70	27.46	0.00	0.00
C_3 - Überraschung	G	4.11	0.15	90.2	0.00	0.00	5.31	0.22
	D	-	6.00	81.33	3.33	9.33	0.00	0.00
C_4 – Ärger	G	4.63	2.04	0.00	84.55	7.76	0.07	0.95
	D	-	3.51	3.51	44.74	48.25	0.00	0.00
C_5 – Ekel	G	4.47	4.53	5.35	16.18	65.37	3.90	0.19
	D	-	7.93	2.44	14.02	75.61	0.00	0.00
C_6 – Angst	G	12.91	17.07	10.44	1.13	10.08	43.51	4.87
	D	-	16.67	54.90	8.82	19.61	0.00	0.00
C_7 – Trauer	G	26.34	1.58	0.00	20.03	7.27	4.39	40.4
	D	-	22.92	16.67	31.25	29.17	0.00	0.00

Tabelle 8-16: Konfusionsmatrix in Prozent; Merkmalsnormierung, MLP für geometrische (G) und dynamische (D) Merkmale, Training mit D_4, Test von D_3.

Klasse / Klassifikation D_3 vs. D_4 / **SVM**		$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$	$P(C_6)$	$P(C_7)$
C_1 – Neutral	G	95.35	0.10	0.00	0.58	0.10	1.67	2.20
	D	-	-	-	-	-	-	-
C_2 - Freude	G	1.61	97.37	0.06	0.00	0.00	0.96	0.00
	D	-	66.20	0.00	3.52	28.17	0.00	2.11
C_3 - Überraschung	G	1.27	0	90.5	0.00	0.15	8.08	0.00
	D	-	2.67	69.33	5.33	9.33	13.33	0.00
C_4 – Ärger	G	3.27	1.09	0.00	71.41	9.39	1.70	13.14
	D	-	5.26	0.00	45.61	46.49	0.00	2.63
C_5 – Ekel	G	1.32	3.84	3.40	10.26	74.24	6.68	0.25
	D	-	8.54	0.00	14.02	75.00	1.83	0.61
C_6 – Angst	G	3.03	13.89	11.5	0.00	11.00	54.58	5.99
	D	-	7.84	51.96	7.84	20.59	9.80	1.96
C_7 – Trauer	G	8.44	1.58	0.14	18.04	2.19	12.83	56.79
	D	-	0.00	8.33	12.50	50.00	16.67	12.50

Tabelle 8-17: Konfusionsmatrix in Prozent; Merkmalsnormierung, SVM für geometrische (G) und dynamische (D) Merkmale, Training mit D_4, Test von D_3.

8.2 Verwendete Merkmale zur Mimikerkennung

Der Übersicht halber werden im Folgenden die in dieser Arbeit verwendeten dynamischen und geometrischen Rohmerkmale zur Erkennung der Mimik noch einmal kompakt dargestellt (Abbildung 8-1 und 8-2). Dynamische Merkmale werden dabei durch eine Messung des optischen Flusses innerhalb markanter Gesichtsregionen bestimmt, welche modellgestützt entlang eines Rasters festgelegt werden. Wie in Abschnitt 4.3.1 beschrieben, werden dazu 14 gemittelte Vektoren $\bar{\mathbf{v}}_i^t$ (4.15) berechnet.
Geometrische Merkmale repräsentieren hingegen zehn räumliche Parameter, d.h. Abstände und Winkel in 3D (Abschnitt 4.4.2).

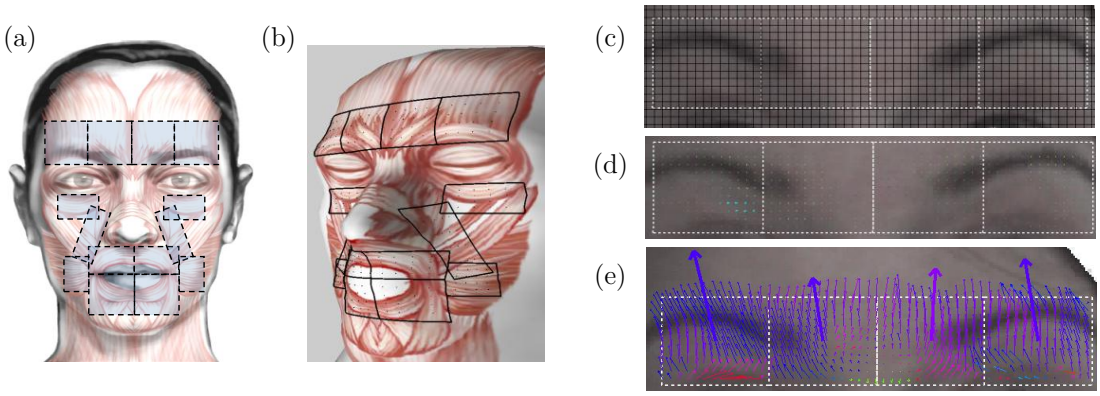
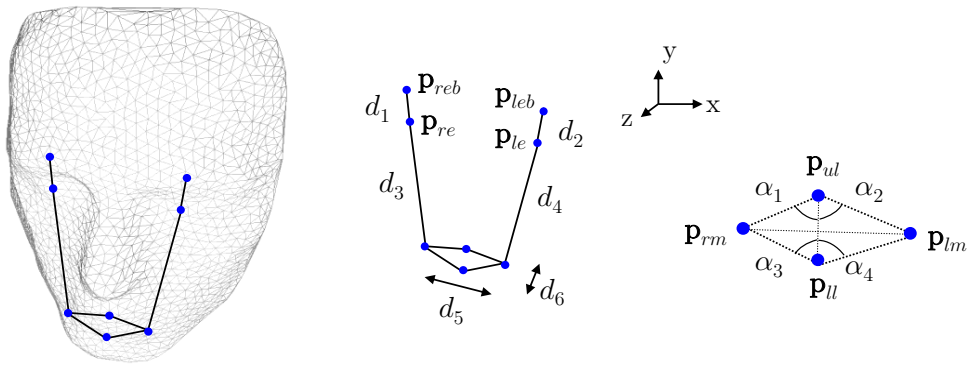


Abbildung 8-1: Dynamische Merkmale, (a, b) Flussregionen und Verknüpfung mit 3D Modell, (c) schematische Darstellung des zugrundeliegenden Rasters, (d) rastergestützte Berechnung des optischen Flusses und (e) Akkumulation und Mittelung $\bar{\mathbf{v}}_i^t$.



(a) (b) (c)

Abbildung 8-2: Definition geometrischer Merkmale, (a) Gesichtsmodell mit Merkmalspunkten $\mathbf{p}_i$ und (b) Abstände d_j sowie (c) Winkel α_k.

8.3 Pseudocode zur Gruppierung von Messpunkten

```
h(p_i, p_j, d_dist, d_color)                ; Homogenitätskriterium (Abstand, Farbe)
P := {p_0,..., p_n-1}                       ; Initiale Punktmenge
A := {∅}                                    ; Stack der aktiven Punkte
j := 0                                      ; Cluster ID

while (P ≠ ∅)                               ; solange P noch Punkte enthält
{
        p ∈ P   P := P \ p                  ; entnehme nächsten Punkt p aus P
        p set ID := j                       ; p erhält Cluster ID j
        A push p                            ; addiere p zum Stack A der aktiven Punkte

        while (A ≠ ∅)                       ; solange A noch Punkte enthält
        {
                a := A pop                  ; hole Punkt a von Stack A

; finde Menge F aller Punkte in P, die das Homogenitätskriterium h bzgl. a erfüllen

                F := find (h, P, a)         ; Untermenge F von P erfüllt h bzgl. a

                if (F ≠ ∅)                  ; es wurden Punkte gefunden
                {
                        ∀f ∈ F, f set ID := j   ; setze Cluster ID für alle Punkte in F
                        P := P \ F          ; entferne alle Punkte der Menge F aus P
                        A push F            ; addiere Punkte der Menge F zu Stack A
                }
        }
        j :=j +1                            ; beginne neues Cluster mit ID j
}
```

Die Suchfunktion “find“ lässt dabei sich effizient durch einen binären Suchbaum (BSP-Tree) realisieren [Kle05, Fuc80].

8.4 Homogene Transformationsmatrizen

In Abschnitt 5.1.1 und 5.2.1 werden die verwendeten Verfahren zur Schätzung der Pose erläutert. Dabei werden die Elementarmatrizen nach (5.3) zur Berechnung der aktuellen Modellposematrix $\mathbf{T}$ verwendet. Die inversen Rotationsmatrizen ergeben sich unmittelbar aus der transponierten Matrix.

8.4.1 Elementare Rotationsmatrizen

Operation	Matrix $\mathbf{E}(\alpha)$	Matrix $(\mathbf{E}(\alpha))^{-1}$
$\mathbf{E}_{rx}(\alpha)$ bzw. $(\mathbf{E}_{rx}(\alpha))^{-1}$ Rotation um X-Achse	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) & 0 \\ 0 & \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) & 0 \\ 0 & -\sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
$\mathbf{E}_{ry}(\alpha)$ bzw. $(\mathbf{E}_{ry}(\alpha))^{-1}$ Rotation um Y-Achse	$\begin{bmatrix} \cos(\alpha) & 0 & \sin(\alpha) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\alpha) & 0 & \cos(\alpha) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} \cos(\alpha) & 0 & -\sin(\alpha) & 0 \\ 0 & 1 & 0 & 0 \\ \sin(\alpha) & 0 & \cos(\alpha) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
$\mathbf{E}_{rz}(\alpha)$ bzw. $(\mathbf{E}_{rz}(\alpha))^{-1}$ Rotation um Z-Achse	$\begin{bmatrix} \cos(\alpha) & -\sin(\alpha) & 0 & 0 \\ \sin(\alpha) & \cos(\alpha) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} \cos(\alpha) & \sin(\alpha) & 0 & 0 \\ -\sin(\alpha) & \cos(\alpha) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

8.4.2 Elementare Translationsmatrizen

Operation	Matrix $\mathbf{E}(d)$	Matrix $(\mathbf{E}(d))^{-1}$
$\mathbf{E}_{tx}(d)$ bzw. $(\mathbf{E}_{tx}(d))^{-1}$ Translation X-Achse	$\begin{bmatrix} 1 & 0 & 0 & d \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & -d \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
$\mathbf{E}_{ty}(d)$ bzw. $(\mathbf{E}_{ty}(d))^{-1}$ Translation Y-Achse	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & d \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -d \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
$\mathbf{E}_{tz}(d)$ bzw. $(\mathbf{E}_{tz}(d))^{-1}$ Translation Z-Achse	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & d \\ 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -d \\ 0 & 0 & 0 & 1 \end{bmatrix}$

Bibliographie

Veröffentlichungen -

Journal Publikationen und Zeitschriftenbeiträge

1. **Niese, R., Al-Hamadi, A. & Michaelis, B. 2010.** *Emotion Recognition based on 2D-3D Facial Feature Extraction from Color Image Sequences*. Journal of Multimedia, in print, 2010.

2. **Niese, R., Al-Hamadi, A., Panning, A., Brammen, D. G., Ebmeyer, U. & Michaelis, B. 2009.** *Towards Pain Recognition in Post-Operative Phases Using 3D-based Features From Video and Support Vector Machines*. International Journal of Digital Content Technology and its Applications (JDCTA), AICIT, Vol. 3(4), pp. 21-33, 2009.

3. **Panning, A., Al-Hamadi, A., Niese, R. & Michaelis, B. 2008.** *Facial expression recognition based on Haar-like feature detection*. MAIK Nauka/Interperiodica (Springer Science), Pattern Recognition and Image Analysis, Vol. 18(3), pp. 447-452, 2008.

4. **Niese, R., Al-Hamadi, A. & Michaelis, B. 2007.** *A Novel Method for 3D Face Detection and Normalization*. Journal of Multimedia, ISSN: 1796-2048, Vol. 2(5), pp. 1-12, 2007.

5. **Niese, R., Al-Hamadi, A. & Michaelis, B. 2007.** *Nearest Neighbor Classification for Emotion Recognition in Stereo Image Sequences*. Transactions on Electronics and Signal Processing, ISAST, ISSN: 1797-2329, Vol. 1(1), pp. 88-94, 2007.

6. **Al-Hamadi, A., Niese, R. & Michaelis, B. 2006.** *A fast and robust approach for the segmentation of moving objects*. Journal of Computer Vision and Graphics, ISSN: 1381-6446, Vol. 32(1), pp. 13-19, 2006.

7. **Al-Hamadi, A., Niese, R. & Michaelis, B. 2007.** *Multi-Object Tracking In Video Using A Trisection Paradigm*. Special Issues of the International Journal of Pattern Recognition and Image Analysis, ISSN: 1054-6618, Vol. 17(4), pp. 493-507, 2007.

8. **Al-Hamadi, A., Niese, R. & Michaelis, B. 2006.** *Feature-Based Correspondence Analysis in Color Image Sequences*. Journal of Computer Vision and Graphics, ISSN: 1381-6446, Vol. 32(1), pp. 179-189, 2006.

9. **Al-Hamadi, A. K., Niese, R., Panning, A. & Michaelis, B. 2006.** *Toward robust face analysis method of non-cooperative persons in stereo color image sequences*. International Journal of Machine Graphics & Vision, Vol. 15(3), pp. 245-254, 2006.

10. **Heinzel, A., Bermpohl, F., Niese, R., Pfennig, A., Pascual-Leone, A., Schlaug, G. & Northoff, G. 2005.** *How do we modulate our emotions? Parametric fMRI reveals cortical midline structures as regions specifically involved in the processing of emotional valences*. Cognitive Brain Research, Elsevier, Vol. 25(1), pp. 348-358, 2005.

11. **Northoff, G., Heinzel, A., Bermpohl, F., Niese, R., Pfennig, A., Pascual-Leone, A. & Schlaug, G. 2004.** *Reciprocal modulation and attenuation in the prefrontal cortex: An fMRI study on emotional-cognitive interaction*. Human Brain Mapping, Wiley, Vol. 21(3), pp. 202-212, 2004.

Konferenzartikel

12. **Krell, G., Niese, R. & Michaelis, B. 2009.** *Facial Expression Recognition with Multichannel Deconvolution*. International Conference on Advances in Pattern Recognition, Kolkata, pp. 413-416, 2009.

13. **Niese, R., Al-Hamadi, A., Aziz, F. & Michaelis, B. 2008.** *Robust Facial Expression Recognition Based on 3-d Supported Feature Extraction and SVM Classification*. 8th IEEE International Conference on Face and Gesture Recognition (FG'08), Amsterdam, pp. 1-7, 2008.

14. **Al-Hamadi, A., Niese, R., Pathan, S. S. & Michaelis, B. 2008.** *Geometric and Optical Flow Based Method for Facial Expression Recognition in Color Image Sequences*. International Conference on Computer Vision and Graphics (ICCVG'08), Nov. 10-12, Warsaw, Poland, pp. 228-238, 2008.

15. **Niese, R., Al-Hamadi, A., Panning, A. & Michaelis, B. 2007.** *Real-Time Capable Method for Facial Expression Recognition in Color and Stereo Vision*. Computational Science and Its Applications (ICCSA'07), Kuala Lumpur, Malaysia, pp. 397-408, 2007.

16. **Al-Hamadi, A., Homberg, U., Niese, R. & Michaelis, B. 2007.** *Efficient tracking approach of multiple interacting objects using data association*. The 22nd International IEEE Symposium on Computer and Information Sciences (ISCIS'07), Ankara, Turkey, pp. 2007.

17. **Panning, A., Al-Hamadi, A., Niese, R. & Michaelis, B. 2006.** *Facial Expression Recognition Approach based on Haar-Like Feature Detection*. 7th Open German/Russian Workshop on Pattern Recognition and Image Understanding, Etlingen, pp. 502-512, 2006.

18. **Niese, R., Al-Hamadi, A. & Michaelis, B. 2006.** *A Stereo and Color-based Method for Face Pose Estimation and Facial Feature Extraction*. 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, pp. 299-302, 2006.

19. **Kuhn, R., Niese, R., Calow, R. & Michaelis, B. 2006.** *Schnelle berührungslose Bestimmung der Oberflächenstruktur langrunder Körper*. Geoinformatik und Erdbeobachtung - Vorträge der 26. Wissenschaftlich-Technischen Jahrestagung der DGPF, Berlin, pp. 257–264, 2006.

20. **Al-Hamadi, A., Panning, A., Niese, R., Pathan, S. S. & Michaelis, B. 2006.** *A Model-based Image Analysis Method for Extraction and Tracking of Facial Features in Video Sequences*. The 4th International Multi-Conference on Computer Science and Information Technology (CSIT'06), Amman, pp. 502-512, 2006.

21. **Niese, R., Calow, R., Al-Hamadi, A. & Michaelis, B. 2005.** *Automatische 3D-Gesichtserfassung in Farbstereosequenzen*. 8. Anwendungsbezogener Workshop zur Erfassung, Modellierung, Verarbeitung und Auswertung von 3D-Daten, Berlin, pp. 53-60, 2005.

22. **Michaelis, B. & Niese, R. 2005.** *Emotionsbewertung durch 3D Gesichtserfassung. Workshop zu Methoden und Verfahren zur Entwicklung, Bewertung und Anwendung von Systemen zur Mensch-Maschine-Interaktion*, Fraunhofer Institut für Fabrikbetrieb und Automatisierung, Magdeburg, pp. 17-26, 2005.

23. **Al-Hamadi, A., Niese, R. & Michaelis, B. 2005.** *Hierarchische merkmalsbasierte Bewegungsanalysemethode für Videosequenzen*. Proc. 9th German Workshop on Colour Image Processing (FarbBV '03), Esslingen, pp. 2005.

24. **Al-Hamadi, A., Niese, R. & Michaelis, B. 2004.** *Feature-based Correspondence Analysis in Color Image Sequences*. International Conference on Computer Vision and Graphics (ICCVG'04), September 22-24, Warsaw, Poland, pp. 179-187, 2004.

25. **Al-Hamadi, A., Niese, R. & Michaelis, B. 2004.** *A Fast and Robust Approach for the Segmentation of Moving Objects*. International Conference on Computer Vision and Graphics (ICCVG'04), September 22-24, Warsaw, Poland, pp. 13-20, 2004.

26. **Al-Hamadi, A., Niese, R. & Michaelis, B. 2004.** *Multi-Object Tracking In Video Using A Trisection Paradigm*. 7th International Conference on Pattern Recognition and Image Analysis (PRIA'04), St. Petersburg, pp. 599-602, 2004.

27. **Al-Hamadi, A., Niese, R. & Michaelis, B. 2003.** *Towards Robust segmentation and tracking of moving objects in video sequences*. 3rd IEEE-EURASIP Symposium on Image and Signal Processing and Analysis (ISPA'03), September 18-20, 2003, Rome, Italy, pp. 645-651, 2003.

28. **Al-Hamadi, A., Niese, R. & Michaelis, B. 2003.** *A robust approach for contour extraction and tracking of moving objects in video sequences*. International Conference on Signal processing, pattern recognition and applications (Greece June 30 - July 2, 2003), proceedings: Acta Press, pp. 336-341, 2003.

29. **Al-Hamadi, A., Niese, R. & Michaelis, B. 2003.** *Another Paradigm for the Solution of the Correspondence Problem in Motion Analysis*. 8th iberoamerican congress (CIARP'03), Havana, Cuba, pp. 95-103, 2003.

Literatur

[Ahl96] **Ahlrichs**, **U., Paulus**, **D. & Wolf**, **S. 1996**. *Objektivierung der Beurteilung von Gesichtsasymmetrien durch Bildanalyse*. Workshop Bildverarbeitung für die Medizin, pp. 125-130, 1996.

[Alb89] **Albertz**, **J. & Kreiling**, **W. 1989**. *Photogrammetrisches Taschenbuch*. (ed. 4), Herbert Wichmann Verlag GmbH, 1989.

[Alb98] **Albrecht**, **P. & Michaelis**, **B. 1998**. *Stereo Photogrammetry with Improved Spatial Resolution*. International Conference on Pattern Recognition (ICPR'98), pp. 845-849, IEEE Computer Society, 1998.

[AlH01] **Al-Hamadi**, **A. 2001**. Verbesserte Störsicherheit bei der Bewegungsanalyse in monokularen Farbbildsequenzen durch adaptive Farbraumtransformationen. Otto-von-Guericke-Universität Magdeburg, Dissertation, 2001.

[AlH03] **Al-Hamadi**, **A., Michaelis**, **B. & Niese**, **R. 2003**. *A robust approach for contour extraction and tracking of moving objectsin video sequences*. International Conference on Signal Processing, Pattern Recognition, pp. 336-341, 2003.

[AlH06a] **Al-Hamadi**, **A., Niese**, **R. & Michaelis**, **B. 2006**. *Feature-based Correspondence Analysis in Color Image Sequences*. Journal of Computer Vision and Graphics, Vol. 2(5), pp. 179-187, 2006.

[AlH06b] **Al-Hamadi**, **A., Panning**, **A., Niese**, **R. & Michaelis**, **B. 2006**. *A Model-based Image Analysis Method for Extraction and Tracking of Facial Features in Video Sequences*. The 4th International Multi-Conference on Computer Science and Information Technology (CSIT'06), Amman, pp. 502-512, 2006.

[Amb05] **Ambadar**, **Z., Schooler**, **J. & Cohn**, **J. F. 2005**. *Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions*. Psychological Science Journal, Vol. 16(1), pp. 403-410, 2005.

[Asc93] **Aschwanden**, **P. F. 1993**. *Experimenteller Vergleich von Korrelationskriterien in der Bildanalyse*. Dissertation, Techn. Wiss. ETH Zürich, Nr. 10196, doi:10.3929/ethz-a-000904762, 1993.

[Bar06] **Bartlett**, **M. S., Littlewort**, **G., Frank**, **M. G., Lainscsek**, **C., Fasel**, **I. R. & Movellan**, **J. R. 2006**. *Fully Automatic Facial Action Recognition in Spontaneous Behavior*. IEEE International Conference on Face and Gesture Recognition (FG'06), pp. 223-230, 2006.

[Bar98] **Bartlett**, **M. S. 1998**. *Face Image Analysis by Unsupervised Learning and Redundancy Reduction*. University of California, San Diego, Phd Thesis, 1998.

[Ben75] **Bentley, J. L. 1975**. *Multidimensional binary search trees used for associative searching*. Commun. ACM, Vol. 18(9), pp. 509-517, 1975.

[Bla99] **Blanz, V. & Vetter, T. 1999**. *A Morphable Model for the Synthesis of 3D Faces*. SIGGRAPH, pp. 187-194, 1999.

[Boe05] **Böttcher, R. A. 2005**. *Flow in Computerspielen*. Otto-von-Guericke-Universität Magdeburg, Diplomarbeit, 2005.

[Bou00] **Bouguet, J. 2000**. Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm. Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm , Intel Corporation Microprocessor Research Labs, 2000.

[Bra06] **Brahnam, S., Chuang, C., Shih, F. Y. & Slack, M. R. 2006**. *Machine recognition and representation of neonatal facial displays of acute pain*. Acute Pain, Vol. 8(3), pp. 146-146, 2006.

[Cal05] **Calow, R. 2005**. *Markerlose Ganganalyse mit einem Multikamerasystem*. Otto-von-Guericke-Universität Magdeburg, Dissertation, 2005.

[Cha06] **Chang, Y., Hu, C., Feris, R. & Turk, M. 2006**. *Manifold based analysis of facial expression*. Image Vision Comput., Vol. 24(6), pp. 605-614, 2006.

[Cha08] **Chang, C. & Lin, C. 2008**. *LIBSVM: a library for support vector machines*. www.csie.ntu.edu.tw/~cjlin/libsvm, 2008.

[Chri01] **Cristianini, N. & Shawe-Taylor, J. 2001**. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. (ed. 1), Cambridge University Press, ISBN-13: 978-0521780193, 2001.

[CMU10] **CMU, 2010**. *FACS - Facial Action Coding System*. http://www-2.cs.cmu.edu/afs/cs/project/face/www/facs.htm, 2010.

[Coh01] **Cohn, J., Kanade, T., Moriyama, T., Ambadar, Z., Xiao, J., Gao, J. & Imamura, H. 2001**. *A Comparative Study of Alternative FACS Coding Algorithms*. Robotics Institute, Pittsburgh, PA: 2001.

[Coh04] **Cohn, J. F. & Schmidt, K. L. 2004**. *The timing of facial motion in posed and spontaneous smiles*. J. Wavelets, Multi-resolution & Information Processing, Vol. 2(1), pp. 1-12, 2004.

[Coh08] **Cohn, J. F. & Ekman, P. 2008**. *Measuring facial action*. In J. A. Harrigan, R. Rosenthal & K. R. Scherer (Editor), *Handbook of nonverbal behavior research methods in the affective sciences*, pp. 9-64, Oxford Universtiy Press, 2008.

[Cov67] **Cover, T. & Hart, P. 1967**. *Nearest Neighbor Pattern Classification*. IEEE Transactions on Information Theory, Vol. 13(1), pp. 21-27, 1967.

[Cro84] **Crow**, **F. C. 1984**. *Summed-area tables for texture mapping*. SIGGRAPH '84: Proceedings of the 11th annual conference on Computer graphics and interactive techniques, pp. 207-212, ACM, 1984.

[Del07] **Delac**, **K. & Grgic**, **M. 2007**. *Face Recognition*. (ed. 1), I-Tech Education and Publishing, ISBN 978-3-902613-03-5, 2007.

[Dig10] **Digimask**, **2010**. *http://www.digimask.com,* Digimask — The Face of the Future, 2010.

[Ekm02] **Ekman**, **P., Friesen**, **W. V. & Hager**, **J. C. 2002**. *Facial Action Coding System. A Human Face*. (ed. 2), Salt Lake City, Research Nexus eBook, 2002.

[Ekm03] **Ekman**, **P. 2003**. *Darwin, deception, and facial expression*. Ann. N. Y. Acad. Sci., Vol. 1000(1), pp. 1, 2003.

[Ekm05] **Ekman**, **P. & Rosenberg**, **E. L. 2005**. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford Univ. Press, ISBN-13: 978-0195179644, 2005.

[Elm09] **Elmezain**, **M., Al-Hamadi**, **A. & Michaelis**, **B. 2009**. *A Novel System for Automatic Hand Gesture Spotting and Recognition in Stereo Color Image Sequences*. The 17-th International Conference on Computer Graphics, Visualization, pp. 89-96, Springer Verlag, 2009.

[Fac10] **2010**. Face Recognition Homepage, Databases. Face Recognition Homepage, Databases , http://www.face-rec.org/databases/, 2010.

[Fas04] **Fasel**, **B., Monay**, **F. & Gatica-Perez**, **D. 2004**. *Latent semantic analysis of facial action codes for automatic facial expression recognition*. Multimedia Information Retrieval, pp. 181-188, 2004.

[Fel99] **Fellenz**, **W. A., Taylor**, **J. G., Tsapatsoulis**, **N. & Kollias**, **S. 1999**. *Comparing Template-based, Feature-based and Supervised Classification of Facial Expressions from Static Images*. Proceedings of Circuits, Systems, Communications and Computers (CSCC'99), pp. 5331-5336, 1999.

[Fle01] **Fleuret**, **F. & Geman**, **D. 2001**. *Coarse-to-Fine Face Detection*. Int. J. Comput. Vision, Vol. 41(1), pp. 85-107, 2001.

[Flo10] **Flores**, **V. C. 2010**. *ARTNATOMY/ARTNATOMIA*. www.artnatomia.net, 2010.

[Fol95] **Foley**, **J., Dam**, **A. v., Feiner**, **S. & Hughes**, **J. 1995**. *Computer Graphics: Principles and Practice*. (ed. 2), Pearson, ISBN-13: 978-8131705056, 1995.

[Fuc80] **Fuchs**, **H., Kedem**, **Z. M. & Naylor**, **B. F. 1980**. *On visible surface generation by a priori tree structures*. SIGGRAPH'80, pp. 124-133, 1980.

[Gol07] **Goldstein**, **B. E. 2007**. *Wahrnehmungspsychologie*. (ed. 2.), Spektrum-Akademischer Vlg, ISBN-13: 978-3827410832, 2007.

[Gro08] **Gross**, **R., Matthews**, **I., Cohn**, **J. F., Kanade**, **T. & Baker**, **S. 2008**. *Multi-PIE*. IEEE International Conference on Automatic Face and Gesture Recognition (FG'08), pp. 1-8, 2008.

[Har05] **Hartung**, **J., Elpelt**, **B. & Klösener**, **K. 2005**. *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. (ed. 14), Oldenbourg Wissenschaftsverlag, ISBN: 3-486-57890-1, 2005.

[Hay98] **Haykin**, **S. 1998**. *Neural Networks: A Comprehensive Foundation*. (ed. 3), Prentice Hall, ISBN-13: 978-0131471399, 1998.

[Hei05] **Heinzel**, **A., Bermpohl**, **F., Niese**, **R., Pfennig**, **A., Pascual-Leone**, **A., Schlaug**, **G. & Northoff**, **G. 2005**. How do we modulate our emotions? Parametric fMRI reveals cortical midline structures as regions specifically involved in the processing of emotional valences. Cognitive Brain Research, Vol. 25(1), pp. 348-358, 2005.

[Hei07] **Heisele**, **B., Serre**, **T. & Poggio**, **T. 2007**. *A Component-based Framework for Face Detection and Identification*. International Journal of Computer Vision, Vol. 74(1), pp. 167-181, 2007.

[Her03] **Herbrich**, **R. 2003**. Learning Kernel Classifiers: Theory and Algorithms (Adaptive Computation and Machine Learning). (ed. 1), Mit Press, ISBN-13: 978-0262083065, 2003.

[Hor81] **Horn**, **B. K. & Schunck**, **B. G. 1981**. *Determining Optical Flow*. Artificial Intelligence, Vol. 17(1), pp. 185-203, 1981.

[Hor96] **Horprasert**, **T., Yacoob**, **Y. & Davis**, **L. S. 1996**. *Computing 3-D head orientation from a monocular image sequence*. 2nd International Conference on Automatic Face and Gesture Recognition (FG'96), pp. 242-247, 1996.

[Hsu02] **Hsu**, **C. & Lin**, **C. 2002**. *A comparison of methods for multiclass support vector machines*. IEEE transactions on neural networks, Vol. 13(2), pp. 415-425.

[HsuAJ02] **Hsu**, **R., Abdel-Mottaleb**, **M. & Jain**, **A. K. 2002**. *Face Detection in Color Images*. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 24(1), pp. 696-706, 2002.

[Hua07] **Huang**, **C., Ai**, **H., Li**, **Y. & Lao**, **S. 2007**. *High-Performance Rotation Invariant Multiview Face Detection*. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 29(1), pp. 671-686, 2007.

[ISO01] **ISO**, **2001**. ISO/IEC 14496-2:2001. Information technology -- Coding of audio-visual objects -- Part 2: Visual (formal name). 2001.

[Jae05] **Jähne, B. 2005**. *Digitale Bildverarbeitung*. (ed. 6), Springer Verlag, ISBN-13: 978-3540249993, 2005.

[Kan00] **Kanade, T., Tian, Y. l. & Cohn, J. F. 2000**. *Comprehensive Database for Facial Expression Analysis*. IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), pp. 46-53, 2000.

[Kel00] **Keltner, D. & Ekman, P. 2000**. *Facial expression of emotion*. In M. Lewis & J. M. Haviland-Jones (Editor), *Handbook of Emotions*, pp. 236–249, The Guilford Press, 2000.

[Kha06] **Khan, M. M., Ingleby, M. & Ward, R. D. 2006**. Automated Facial Expression Classification and affect interpretation using infrared measurement of facial skin temperature variations. ACM Trans. Auton. Adapt. Syst., Vol. 1(1), pp. 91-113, 2006.

[Kim08] **Kim, B., Ban, S. & Lee, M. 2008**. *Improving AdaBoost Based Face Detection Using Face-Color Preferable Selective Attention*. IDEAL '08: Proceedings of the 9th International Conference on Intelligent Data Engineering and Automated Learning, pp. 88-95, Springer-Verlag, 2008.

[Kle05] **Klein, R. 2005**. *Algorithmische Geometrie*. (ed. 2), Springer-Verlag, ISBN 3-540-20956-5, 2005.

[Kle96] **Klette, R., Koschan, A. & Schlüns, K. 1996**. *Computer Vision*. (ed. 1), Vieweg Technik Verlag, ISBN 3-528-06625-3, 1996.

[Kni07] **Knieling, S. 2007**. *Einführung in die Modellierung künstlich neuronaler Netzwerke*. (ed. 1), WiKu-Verlag, ISBN-13: 978-3865531926, 2007.

[Koh97] **Kohonen, T. 1997**. *Self-Organizing Maps*. (ed. 3), Springer, ISBN: 3-540-67921-9, 1997.

[Kre05] **Krengel, U. 2005**. *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. (ed. 8), Vieweg+Teubner, ISBN-13: 978-3834800633, 2005.

[Kru09] **Kruse, R. 2009**. *Neural Networks*. Skript zur Vorlesung, Self-Organizing Maps, pp. Chapter 6, 2009.

[Lie98] **Lien, J. J., Kanade, T., Cohn, J. F. & Li, C. 1998**. *Subtly Different Facial Expression Recognition and Expression Intensity Estimation*. CVPR, pp. 853-859, 1998.

[LiS05] **Li, S. Z. & Jain, A. K. 2005**. *Handbook of Face Recognition*. (ed. 1), Springer Verlag, ISBN-13: 978-0387405957, 2005.

[Luc81] **Lucas, B. D. & Kanade, T. 1981**. *An Iterative Image Registration Technique with an Application to Stereo Vision*. Proc. of 7th International Joint Conference on Artificial Intelligence, pp. 674-679, 1981.

[Luh03] **Luhmann**, **T. 2003**. *Nahbereichsphotogrammetrie. Grundlagen, Methoden und Anwendungen*. (ed. 2), Wichmann Verlag, ISBN-10: 3879073988, 2003.

[Mal00] **Mallot**, **H. A. 2000**. *Sehen und die Verarbeitung visueller Informationen*. (ed. 2), Vieweg Verlagsgesellschaft, ISBN-13: 978-3528156596, 2000.

[Mal05] **Malassiotis**, **S. & Strintzis**, **M. G. 2005**. *Robust real-time 3D head pose estimation from range data*. Pattern Recognition, Vol. 38(8), pp. 1153-1165, 2005.

[Man09] **Mansoorizadeh**, **M. & Charkari**, **N. M. 2009**. *Multimodal information fusion application to human emotion recognition from face and speech*. Multimedia Tools and Applications (Springer Science + Business Media, LLC 2009), Vol. 1(1), pp. 1-21, 2009.

[MaW05] **Ma**, **R. & Wang**, **J. 2005**. Automatic Facial Expression Recognition Using Linear and Nonlinear Holistic Spatial Analysis. ACII, pp. 144-151, 2005.

[McG04] **McGlone**, **C., Mikhail**, **E. & Bethe**, **J. 2004**. *Manual of Photogrammetry*. (ed. 5), ASPRS, ISBN: 1-57083-071-1, 2004.

[Mec99] **Mecke**, **R. 1999**. Grauwertbasierte Bewegungsschätzung in monokularen Bildsequenzen unter besonderer Berücksichtigung bildspezifischer Störungen. Otto-von-Guericke-Universität Magdeburg, Dissertation, ISBN 3-8265-6712-9, 1999.

[Mia06] **Mian**, **A. S., B.**, **M. & Owens**, **R. A. 2006**. *Automatic 3D Face Detection, Normalization and Recognition*. Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06), pp. 735-742, 2006.

[Mit97] **Mitchell**, **T. M. 1997**. *Machine Learning*. (ed. 1), Mcgraw Hill Book Co., ISBN-13: 978-0070428072, 1997.

[Mor10] **Morency**, **L., Whitehill**, **J. & Movellan**, **J. R. 2010**. *Monocular head pose estimation using generalized adaptive view-based appearance model*. Image Vision Comput., Vol. 28(5), pp. 754-761, 2010.

[Mur09] **Murphy-Chutorian**, **E. & Trivedi**, **M. M. 2009**. *Head Pose Estimation in Computer Vision: A Survey*. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 31pp. 607-626, 2009.

[Mus85] **Musmann**, **H., Pirsch**, **P. & Grallert**, **H. 1985**. *Advances in picture coding*. Proceedings of the IEEE, Vol. 73(4), pp. 523-548, 1985.

[Nie05] **Niese**, **R., Calow**, **R., Al-Hamadi**, **A. & Michaelis**, **B. 2005**. *Automatische 3D-Gesichtserfassung in Farbstereosequenzen*. 8. Anwendungsbezogener Workshop zur Erfassung, Modellierung, Verarbeitung und Auswertung von 3D-Daten, Berlin, pp. 53-60, 2005.

[Nie07a] **Niese**, **R., Al-Hamadi**, **A. & Michaelis**, **B. 2007**. *A Novel Method for 3D Face Detection and Normalization*. Journal of Multimedia, Vol. 2(5), pp. 1-12, 2007.

[Nie07b] **Niese**, **R., Al-Hamadi**, **A. & Michaelis**, **B. 2007**. *Nearest Neighbor Classification for Emotion Recognition in Stereo Image Sequences*. Transactions on Electronics and Signal Processing, ISAST, (ISSN 1797-2329), Vol. 1(1), pp. 88-94, 2007.

[Nie09] **Niese**, **R., Al-Hamadi**, **A., Panning**, **A., Brammen**, **D. G., Ebmeyer**, **U. & Michaelis**, **B. 2009**. *Towards Pain Recognition in Post-Operative Phases Using 3D-based Features From Video and Support Vector Machines*. International Journal of Digital Content Technology and its Applications (JDCTA), Vol. 3(4), pp. 21-33, 2009.

[Oer08] **Oertel**, **K. 2008**. *Emotionsforschung im Auto*. Systeme|Automotive 07-08, Carl-Hanser Verlag, Vol. pp. 71-73, 2008.

[ORo98] **O'Rourke**, **J. 1998**. *Computational geometry in C*. (ed. 2), Cambridge University Press, ISBN-13: 978-0521649766, 1998.

[Ots98] **Otsuka**, **T. & Ohya**, **J. 1998**. *Spotting Segments Displaying Facial Expression from Image Sequences Using HMM*. Proc. Int'l Conf. Automatic Face and Gesture Recognition (FG'98), pp. 442-447, 1998.

[Pan02] **Pandzic**, **I. S. & Forchheimer**, **R. 2002**. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. (ed. 1), John Wiley & Sons, Ltd, ISBN: 0-470-84465-5, 2002.

[Pan04] **Pantic**, **M. & M.**, **L. J. 2004**. Case-based reasoning for user-profiled recognition of emotions from face images. ICME, pp. 391-394, 2004.

[Pan06] **Pantic**, **M. & Patras**, **I. 2006**. Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences. IEEE Trans. Systems, Man, and Cybernetics, Part B, Vol. 36(1), pp. 433-449, 2006.

[Pan07] **Pantic**, **M. & Bartlett**, **M. S. 2007**. *Machine Analysis of Facial Expressions*. In & K. Delac & M. Grgic (Editor), *Face Recognition*, pp. 377-416, I-Tech Education and Publishing, 2007.

[Pan08] **Panning**, **A., Al-Hamadi**, **A., Niese**, **R. & Michaelis**, **B. 2008**. *Facial expression recognition based on Haar-like feature detection*. Pattern Recognition and Image Analysis, Vol. 18(3), pp. 447-452, 2008.

[Pan09] **Pantic**, **M. 2009**. *Facial Expression Recognition*. In S. Z. Li (Editor), *Encyclopedia of Biometrics*, pp. 400-406, ISBN 978-0-387-73003-5, 2009.

[Pap98] **Papageorgiou**, **C., Oren**, **M. & Poggio**, **T. 1998**. *A General Framework for Object Detection*. Sixth International Conference on Computer Vision (ICCV), pp. 555-562, 1998.

[Par08] **Park**, **U., Tong**, **Y. & Jain**, **A. K. 2008**. *Face recognition with temporal invariance: A 3D aging model*. IEEE FG2008, 8th Int'l Conf. on Automatic Face and Gesture Recognition, Amsterdam, ISBN: 978-1-4244-2153-4, pp. 1-7, 2008.

[Pat09] **Patras**, **I. 2009**. *Face Pose Analysis*. In S. Z. Li (Editor), *Encyclopedia of Biometrics*, pp. 324-329, ISBN 978-0-387-73003-5, 2009.

[Rit92] **Ritter**, **H., Martinetz**, **T. & Schulten**, **K. 1992**. *Neural Computation and Self-Organizing Maps: An Introduction*. (ed. 1), World Scientific Pub Co, ISBN-13: 978-9812381514, 1992.

[Rot09] **Rothkrantz**, **L., Datcu**, **D. & Wiggers**, **P. 2009**. *FACS-coding of facial expressions*. CompSysTech '09: Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing, pp. 1-6, ACM, 2009.

[Rus01] **Rusinkiewicz**, **S. & Levoy**, **M. 2001**. *Efficient Variants of the ICP Algorithm*. Proc. of the 3rd Int. Conf. on 3D Digital Imaging & Modeling, pp. 145-152, 2001.

[Rus07] **Ruser**, **H. & León**, **F. P. 2007**. *Informationsfusion - Eine Übersicht*. Technisches Messen, Vol. 74pp. 93-102, 2007.

[Sch04] **Schneiderman**, **H. & Kanade**, **T. 2004**. *Object Detection Using the Statistics of Parts*. International Journal of Computer Vision, Vol. 56(1), pp. 151-177, 2004.

[Sha09] **Shan**, **C., Gong**, **S. & McOwan**, **P. W. 2009**. *Facial expression recognition based on Local Binary Patterns: A comprehensive study*. Image Vision Comput., Vol. 27(6), pp. 803-816, 2009.

[Sim02] **Sim**, **T., Baker**, **S. & Bsat**, **M. 2002**. *The CMU Pose, Illumination, and Expression (PIE) Database*. FGR, pp. 53-58, 2002.

[Sin09] **SingularInversions**, **2009**. *FaceGen Modeller 3.2.. FaceGen Modeller 3.2.* , http://www.facegen.com, 2009.

[Soy07] **Soyel**, **H. & Demirel**, **H. 2007**. *Facial Expression Recognition Using 3D Facial Feature Distances*. Springer LNCS, Volume 4633/2007, pp. 831-838, 2007.

[Sun98] **Sung**, **K. K. & Poggio**, **T. 1998**. *Example-Based Learning for View-Based Human Face Detection*. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 20(1), pp. 39-51, 1998.

[Suy03] **Suykens**, **J. A., Gestel**, **T. V. & Brabanter**, **J. D. 2003**. *Least Squares Support Vector Machines*. (ed. 1), World Scientific Pub Co, ISBN-13: 978-9812381514, 2003.

[Tai08] **Taini**, **M., Zhao**, **G., Li**, **S. Z. & Pietikäinen**, **M. 2008**. *Facial expression recognition from near-infrared video sequences*. Proc. 19th International Conference on Pattern Recognition (ICPR'08), pp. 1-4, 2008.

[Tia05] **Tian, Y., Cohn, J. F. & Kanade, T. 2005**. *Facial expression analysis*. In & A. K. Jain & S. Z. Li (Editor), *Handbook of Face Recognition*, pp. 247–276, Springer, 2005.

[Tor07] **Torre, F. D., Campoy, J., Ambadar, Z. & Cohn, J. F. 2007**. *Temporal Segmentation of Facial Behavior*. ICCV, pp. 1-8, 2007.

[Tsi06] **Tsiamyrtzis, P., Dowdall, J., Shastri, D., Pavlidis, I., Frank, M. G. & Ekman, P. 2007**. *Imaging Facial Physiology for the Detection of Deceit*. International Journal of Computer Vision, Vol. 71(2), pp. 197-214, 2007.

[Ult03] **Ultsch, A. 2003**. *Maps for the Visualization of high-dimensional Data Spaces*. Workshop on Self-Organizing Maps (WSOM03), Kyushu, Japan, pp. 225-230, 2003.

[Vac04] **Vacchetti, L., Lepetit, V. & Fua, P. 2004**. *Stable Real-Time 3D Tracking Using Online and Offline Information*. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 26(10), pp. 1385-1391, 2004.

[Val06a] **Valstar, M. & Pantic, M. 2006**. *Fully Automatic Facial Action Unit Detection and Temporal Analysis*. IEEE Int'l Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW'06), pp. 149, 2006.

[Val06b] **Valstar, M. F., Pantic, M., Ambadar, Z. & Cohn, J. F. 2006**. *Spontaneous vs. posed facial behavior: automatic analysis of brow actions*. ICMI, pp. 162-170, 2006.

[Vio04] **Viola, P. A. & Jones, M. J. 2004**. *Robust Real-Time Face Detection*. International Journal of Computer Vision, Vol. 57(1), pp. 137-154, 2004.

[Wac97] **Wachter, S. 1997**. *Verfolgung von Personen in monokularen Bildfolgen*. Universität Karlsruhe, Dissertation, 1997.

[Wan06] **Wang, J., Yin, L., Wei, X. & Sun, Y. 2006**. *3D Facial Expression Recognition Based on Primitive Surface Feature Distribution*. IEEE Int'l. Conference on Computer Vision and Pattern Recognition, CVPR06, pp. 1399-1406, 2006.

[Wei06] **Weidenbacher, U., Layher, G., Bayerl, P. & Neumann, H. 2006**. *Detection of Head Pose and Gaze Direction for Human-Computer Interaction*. Perception and Interactive Technologies, International Tutorial and Research Workshop (PIT'06), Proceedings, Kloster Irsee, Germany, June 19-21, pp. 9-19, 2006.

[Wei08] **Weidenbacher, U. & Neumann, H. 2008**. *Unsupervised Learning of Head Pose through Spike-Timing Dependent Plasticity*. Perception in Multimodal Dialogue Systems, 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems (PIT'08), Proceedings, Kloster Irsee, Germany, June 16-18, pp. 123-131, 2008.

[Wes03] **Westman, H. 2003**. *Computer graphics: defining the literal meaning of images*. SIGGRAPH Comput. Graph., Vol. 37(4), pp. 3-3, 2003.

[Wim08] **Wimmer**, **M., Schuller**, **B., Arsic**, **D., Rigoll**, **G. & Radig**, **B. 2008**. *Low-Level Fusion of Audio, Video Feature for Multi-Modal Emotion Recognition*. VISAPP, pp. 145-151, 2008.

[Wu08] **Wu**, **J., Brubaker**, **S. C., Mullin**, **M. D. & Rehg**, **J. M. 2008**. *Fast Asymmetric Learning for Cascade Face Detection*. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 30(3), pp. 369-382, 2008.

[Xia02] **Xiao**, **J., Kanade**, **T. & Cohn**, **J. F. 2002**. *Robust Full-Motion Recovery of Head by Dynamic Templates and Re-registration Techniques*. Int'l Journal of Imaging Systems and Technology, Vol. 13(1), pp. 85-94, 2002.

[Xia09] **Xiaohua**, **L., Lam**, **K., Lansun**, **S. & Jiliu**, **Z. 2009**. *Face detection using simplified Gabor features and hierarchical regions in a cascade of classifiers*. Pattern Recogn. Lett., Vol. 30(8), pp. 717-728, 2009.

[Yac94] **Yacoob**, **Y. & Davis**, **L. 1994**. *Recognizing Facial Expressions by Spatio-Temporal Analysis*. International Conference on Pattern Recognition (ICPR'94), pp. 747-749, Computer Society Press, 1994.

[Yam08] **Yamauchi**, **Y., Fujiyoshi**, **H., Hwang**, **B. & Kanade**, **T. 2008**. *People detection based on co-occurrence of appearance and spatiotemporal features*. 19th International Conference on Pattern Recognition (ICPR'08), pp. 1-4, 2008.

[Yan02] **Yang**, **M., Kriegman**, **D. J. & Ahuja**, **N. 2002**. *Detecting Faces in Images: A Survey*. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 24(1), pp. 34-58, 2002.

[Yan09] **Yang**, **M. 2009**. *Face Detection*. In S. Z. Li (Editor), *Encyclopedia of Biometrics, ISBN 978-0-387-73003-5*, pp. 303-308, 2009.

[Yin06] **Yin**, **L., Wei**, **X., Sun**, **Y., Wang**, **J. & Rosato**, **M. J. 2006**. *A 3D Facial Expression Database For Facial Behavior Research*. FG '06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, pp. 211-216, IEEE Computer Society, 2006.

[Yin08] **Yin**, **L., Chen**, **X., Sun**, **Y., Worm**, **T. & Reale**, **M. 2008**. *A high-resolution 3D dynamic facial expression database*. IEEE International Conference on Automatic Face and Gesture Recognition (FG'08), pp. 1-6, 2008.

[Zen09] **Zeng**, **Z., Pantic**, **M., Roisman**, **G. I. & Huang**, **T. S. 2009**. *A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions*. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 31(1), pp. 39-58, 2009.

[Zha05] **Zhang**, **Y. & Ji**, **Q. 2005**. *Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences*. IEEE Trans. Pattern Analysis & Machine Intelligence, Vol. 27(5), pp. 699-714, 2005.

[Zha06] **Zhang**, **Z. & Zhang**, **J. 2006**. *Driver Fatigue Detection Based Intelligent Vehicle Control*. International Conference on Pattern Recognition, pp. 1262-1265, IEEE Computer Society, 2006.

Lebenslauf

Name	Robert Niese
Geburtsdatum	04.02.1977 in Halberstadt
Staatsbürgerschaft	deutsch
Familienstand	verheiratet, ein Kind
Wohnort	W.-Kobelt-Str. 9, 39108 Magdeburg
Schulbildung	1995 Abitur, Martineum Halberstadt
Wehrdienst	1995-1996
Studium	1997 - 2004 Otto-von-Guericke Universität Magdeburg ▪ Diplomstudiengang Computervisualistik ▪ Abschluss mit Auszeichnung
Praktika	2001 - Harvard University ▪ Harvard Medical School, Dept. of Neurology, Boston, USA 2004 - North Western Medical Physics, Manchester, UK ▪ Christies Hospital Cancer Centre, Dept. of Radiotherapy
Berufliche Tätigkeiten	09/2004 - 08/2007 Anstellung bei InnoMed e.V. Magdeburg, Netzwerk für Neuromedizintechnik ▪ Qualifizierung im Bereich Medizintechnik, ZENIT Magdeburg ▪ Forschungstätigkeit an der Otto-von-Guericke Universität Schwerpunkt Bildverarbeitung, Multikamerasysteme, Computer Vision und Mustererkennung seit 09/2007 Anstellung an der O.v.G. Universität Magdeburg ▪ Projektarbeit Bernstein Gruppe, SFB/TRR 62

Magdeburg, den 31. Mai 2010