

Visuelle Erkennung von Handposituren für einen interaktiven Gebärdensprachtutor

Von der Fakultät für Elektrotechnik und Informationstechnik
der Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Diplom-Ingenieur

Jörg Zieren

aus Neheim-Hüsten

Berichter: Universitätsprofessor Dr.-Ing. Karl-Friedrich Kraiss
Universitätsprofessor Dr. rer. nat. Thomas Seidl

Tag der mündlichen Prüfung: 6. Juli 2007

Diese Dissertation ist auf den Internetseiten der Hochschulbibliothek online verfügbar.

Vorwort

Diese Arbeit entstand im Rahmen meiner Tätigkeit am Lehrstuhl für Technische Informatik der RWTH Aachen. Mein besonderer Dank gilt Herrn Prof. Dr.-Ing. Karl-Friedrich Kraiss, dem Leiter des Lehrstuhls, für die Betreuung und Förderung der Arbeit, seine fachliche Unterstützung und viele kreative Freiheiten. Herrn Prof. Dr. rer. nat. Thomas Seidl danke ich für sein großes Interesse an der Arbeit und die Übernahme des zweiten Gutachtens.

Mit Freude erinnere ich mich an das sehr gute Arbeitsklima am Lehrstuhl, an viele interessante und motivierende Gespräche mit allen Kollegen und besonders Ulrich von Agris, Suat Akyol, Britta Bauer und Nico Hamacher.

Einen wesentlichen Beitrag zur Realisierung der Arbeit leisteten Diplomanden und Studenarbeiter sowie studentische und wissenschaftliche Mitarbeiter. Hervorheben möchte ich Thorsten Dick, mit dem ich in intensiver und engagierter Zusammenarbeit die Verbesserung der Welt im Allgemeinen und der Handpositurerkennung im Besonderen verfolgen konnte, sowie Christian Aengenheister, Johannes Ballé, Sebastian Boße, Nils Schademann und Nils Unger.

Wie immer stand mir meine Frau Nikola mit unendlicher Geduld, Hilfsbereitschaft und Verständnis zur Seite – und schreckte weder vor algorithmischen noch vor emotionalen Notlagen zurück. Ihr verdanke ich das stetige Vertrauen in ein Happy End.

Aachen, im Juli 2007

Jörg Zieren

Inhaltsverzeichnis

Abkürzungen	vii
Glossar	ix
Symbolverzeichnis	xiii
1 Einleitung	1
1.1 Motivation und Aufgabenstellung	2
1.2 Interaktives Tutorsystem für Gebärdensprache	4
1.3 Gliederung	8
2 Grundlagen eines interaktiven Gebärdensprachtutors	11
2.1 Gebärdensprache	11
2.1.1 Ausdrucksmittel	12
2.1.1.1 Handform	13
2.1.1.2 Handstellung	14
2.1.1.3 Handposition	15
2.1.1.4 Handbewegung	16
2.1.2 Phonologisches Modell	16
2.1.3 Abgrenzung zur Gestik	17
2.2 Videobasierte Erfassung von Gebärden und Gesten	18
2.2.1 Einordnung der Verfahren	19
2.2.2 Algorithmen zur Positurrekonstruktion	21
2.3 Leistungsfähigkeit existierender Systeme zur Positurrekonstruktion	23
2.3.1 Schwierigkeiten	23

2.3.2	Anforderungen	25
2.3.3	Systeme zur Positurrekonstruktion durch inverse Kinematik	26
2.3.4	Systeme zur Positurrekonstruktion durch Modellanpassung	33
2.3.5	Bewertung der Ansätze	42
2.4	Leistungsfähigkeit existierender Tutorsysteme für Gebärdensprache	44
2.5	Konzept des interaktiven Gebärdensprachtutors	46
2.5.1	Eingangsbilder	46
2.5.2	Hintergrundsubtraktion	47
2.5.3	Verwendete Merkmale	47
2.5.4	Merkmalsextraktion	51
2.5.5	Positurrekonstruktion	52
2.5.6	Positurenvergleich und Feedbackgenerierung	57
2.5.7	Zusammenfassung	58
3	Modellierung der Hand	61
3.1	Funktionale Anatomie	61
3.1.1	Gelenke und Freiheitsgrade	62
3.1.2	Einschränkungen	63
3.2	3D-Modell	64
3.2.1	Aufbau	64
3.2.2	Grafische Darstellung	65
3.2.3	Modellparameter	66
3.2.3.1	Handform	66
3.2.3.2	Handstellung	69
3.2.3.3	Handposition	70
3.2.3.4	Notation der Parameter	70
3.2.4	Einschränkungen	71
3.3	Vergleich von Posituren	71
3.3.1	Koordinatensystem und Fingerspitzenpositionen	72
3.3.2	Posituredifferenz	73
3.3.3	Handformdifferenz	74

3.3.4	Handstellungsdifferenz	75
3.3.5	Beispiele	76
4	Segmentierung und Merkmalsextraktion	81
4.1	Eingangsbilder	81
4.1.1	Notation	82
4.1.2	Farbrepräsentation	83
4.2	Hintergrundsubtraktion	84
4.2.1	Unimodale Modellierung	85
4.2.2	Multimodale Modellierung	89
4.3	Segmentierung der Marker	90
4.3.1	Mean-Shift-Clustering	90
4.3.2	Farbbasierte Regionenklassifikation	93
4.3.3	Topologische Regionenbeschreibung	95
4.3.4	Geometrische Merkmale	96
4.3.5	Randkurvenapproximation durch Ellipsen	99
4.4	Segmentierung des Gesichts	100
4.5	Zuordnung von Regionen zu Markern	102
4.5.1	Hypothesenraum	103
4.5.2	Viterbi-Pfadsuche	104
4.5.2.1	Komplexitätsvergleich	106
4.5.2.2	Beschleunigung durch Näherungen	106
4.5.3	Hypothesen und Bewertungskriterien für die Markerzuordnung . . .	106
4.5.3.1	Statische Wahrscheinlichkeiten	108
4.5.3.2	Dynamische Wahrscheinlichkeiten	109
4.5.3.3	Behandlung nicht sichtbarer Markierungen	110
4.6	Aufbau des Merkmalsvektors	110
4.6.1	Form und Stellung der dominanten Hand	110
4.6.2	Positionen beider Hände	114
5	Erkennung von Handposituren	115

5.1	Posituredatenbank	115
5.1.1	Quantisierung des Positurenraums	116
5.1.2	Segmentierung der synthetischen Posituransichten	118
5.1.3	Struktur der Einträge	119
5.1.4	Zugriffsverfahren	121
5.2	Merkmalsvergleich	122
5.2.1	Bestimmung der Rotation	123
5.2.2	Berechnung der Merkmalsdifferenz	123
5.2.2.1	Parametrische Merkmalsdifferenz	124
5.2.2.2	Topologische Merkmalsdifferenz	126
5.3	Positurrekonstruktion	130
5.3.1	Mehrdeutigkeitsauflösung	131
5.3.1.1	Statische Positurwahrscheinlichkeit	131
5.3.1.2	Dynamische Positurwahrscheinlichkeit	131
5.3.2	Positurenglättung	132
5.3.2.1	Handstellung	132
5.3.2.2	Handform	133
6	Feedbackgenerierung	135
6.1	Darstellung der Referenzgebärde	135
6.2	Gebärdenvergleich	136
6.2.1	Dynamische Zeitanpassung	136
6.2.2	Vergleichskriterien	139
6.2.2.1	Handposition	140
6.2.2.2	Bewegungsgeschwindigkeit	141
6.2.2.3	Handstellung	142
6.2.2.4	Handform	143
6.3	Präsentation des Feedbacks	143
7	Evaluation	151
7.1	Synthetische Posituransichten	152

7.1.1	Generierung der Eingangsdaten	152
7.1.2	Erkennungsleistung für unkorrelierte Einzelposituren	153
7.1.2.1	Evaluation der Posituredatenbank	154
7.1.2.2	Evaluation der Merkmalsdifferenz	155
7.1.2.3	Evaluation des Datenbankzugriffs	155
7.1.2.4	Beispiele rekonstruierter Posituren	156
7.1.2.5	Variation der Markergeometrie	158
7.1.3	Erkennungsleistung in Positurrensequenzen	159
7.2	Reale Posituransichten	159
7.2.1	Anpassung der Modellgeometrie	160
7.2.2	Eingabe der Positurparameter	161
7.2.3	Erkennungsleistung in Gebärden	164
7.2.4	Diskussion der Ergebnisse	167
7.2.4.1	Beispiele sehr gut rekonstruierter Posituren	168
7.2.4.2	Beispiele durchschnittlich gut rekonstruierter Posituren . . .	169
7.2.4.3	Analyse fehlerhafter Rekonstruktionen	170
7.2.4.4	Vergleich mit existierenden Systemen	174
8	Zusammenfassung und Ausblick	177
8.1	Zusammenfassung	177
8.2	Ausblick	180
A	Algorithmus zur multimodalen Hintergrundmodellierung	183
B	Viterbi-Algorithmus	187
C	Berechnung der Wahrscheinlichkeiten für Markerzuordnungshypothesen	189
D	Generierung synthetischer Testdaten	191
D.1	Erzeugen zufälliger Posituren	191
D.2	Erzeugen zufälliger Positurrensequenzen	192
D.3	Simulation des Bildrauschens	194

E	Analyse und Optimierung der Rekonstruktionsleistung	197
E.1	Optimale Positur bezüglich $\Delta_{\underline{P}}$	198
E.2	Optimale Positur bezüglich $\Delta_{\underline{F}}$ bei sequenzieller Suche	200
E.3	Optimale Positur bezüglich $\Delta_{\underline{F}}$ bei indexbasierter Suche	206
E.4	Positurrensequenzen	209
	Abbildungsverzeichnis	211
	Tabellenverzeichnis	217
	Literaturverzeichnis	219
	Lebenslauf des Verfassers	229

Abkürzungen

AAM	Active Appearance Model
ASM	Active Shape Model
CAD	Computer Aided Design
CIE	Commission Internationale de l'Eclairage
CMC	Karpometakarpalgelenk
COG	Center of Gravity
D	Daumen
DGS	Deutsche Gebärdensprache
DH	Dominante Hand
DIP	Distales Interphalangealgelenk
DOF	Degrees of Freedom
DTW	Dynamic Time Warping
EKF	Extended Kalman Filter
EM	Expectation-Maximization
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
IP	Interphalangealgelenk
KF	Kleiner Finger
MCP	Metakarpophalangealgelenk
MF	Mittelfinger
MRI	Magnetic Resonance Imaging
NDH	Nichtdominante Hand
PCA	Principal Components Analysis
PD	Phalanx Distalis
PDM	Point Distribution Model
PIP	Proximales Interphalangealgelenk
PM	Phalanx Media
PP	Phalanx Proximalis
RF	Ringfinger
RGB	Rot/Grün/Blau
RMS	Residual Mean Square
UKF	Unscented Kalman Filter

ABKÜRZUNGEN

VR	Virtuelle Realität
ZF	Zeigefinger

Glossar

Ansichtskugel

Gedachte Kugel um den Mittelpunkt der Hand. Durch Angabe eines Punktes auf der Oberfläche der Ansichtskugel lässt sich ein Ansichtswinkel beschreiben.

manuelle Ausdrucksmittel

Bezeichnet Handform, Handstellung, Handposition und Handbewegung, die in Gebärdensprachen den Hauptanteil der Information vermitteln.

nichtmanuelle Ausdrucksmittel

Bezeichnet Kopfhaltung, Körperhaltung, Gesichtsausdruck, Blickrichtung, Mundform, die in Gebärdensprachen zusätzlich zu den manuellen Ausdrucksmitteln zur Informationsübermittlung verwendet werden.

Avatar

Virtuelle Repräsentation eines Menschen.

bounding box

Kleinstes achsenparalleles Rechteck, das eine Bildregion vollständig einschließt.

Downhill-Simplex-Algorithmus

Iteratives numerisches Verfahren zur Optimierung nichtlinearer Funktionen.

Gebärdenraum

Bereich vor dem Oberkörper des Gebärdensprechers, in dem sich die Hände während des Gebärdens bewegen und der vom Gegenüber visuell erfasst werden kann, ohne den Blick vom Gesicht des Sprechers abzuwenden.

gimbal lock

Äquivalenz der Variation von Längengrad ϕ_{lon} und Rotation ϕ_{rot} bei Breitengraden von $\phi_{\text{lat}} = \pm 90^\circ$.

ground truth data

Fehlerfreie Beschreibung eines Sachverhalts, die zur Bewertung gemessener oder berechneter Daten verwendbar ist.

dominante Hand

Einhändige Gebärden werden nur mit der dominanten Hand ausgeführt, in beidhändigen

Gebärden transportiert sie den Hauptanteil der Information. Ergibt sich im Allgemeinen aus der Händigkeit des Gebärdensprechers.

nichtdominante Hand

Bleibt bei einhändigen Gebärden statisch, bei beidhändigen hat sie die Rolle eines Gegenparts oder spiegelt die Bewegung der dominanten Hand.

Handform

Form der Hand, die sich durch die Stellung der Gelenke in Fingern, Daumen und Handteller ergibt.

Handstellung

Orientierung der Hand im Raum, die sich aus der Stellung von Schulter-, Ellenbogen- und Handgelenk ergibt.

Handposition

Position der Hand im dreidimensionalen Raum.

Handbewegung

Zeitliche Variation von Handform, Handstellung und Handposition.

Handpositur

Fasst in der vorliegenden Arbeit Handform und Handstellung zusammen.

Hu-Momente

Von Hu in [Hu62] beschriebene statistische Momente, invariant gegenüber Translation, Rotation und Skalierung.

Kalman-Filter

Von R. E. Kalman in [Kal60] vorgestelltes rekursives Filter zur Schätzung des Zustands dynamischer Systeme aus rauschbehafteten Messwerten.

kinematische Kette

System aus starren Körpern, die durch Gelenke verbunden sind.

direkte Kinematik

Berechnung der Lage des letzten Elements einer kinematischen Kette aus den Stellungen ihrer Gelenke.

inverse Kinematik

Berechnung der Gelenkstellungen einer kinematischen Kette bei vorgegebener Lage des letzten Elements.

ansichtenbasierte Merkmale

Beschreiben Eigenschaften der visuellen Erscheinung eines Objekts, die in der Realität nicht notwendigerweise eine anschauliche Entsprechung haben. Ein Beispiel ist der Flächeninhalt einer Bildregion, die ein Objekt repräsentiert.

modellbasierte Merkmale

Beschreiben Eigenschaften oder Parameter eines Objekts oder eines Modells unabhängig von dessen visueller Erscheinung. Im Kontext der visuellen Handpositurerkennung sind Objekte drei- und deren Ansichten zweidimensional. Ein typisches modellbasiertes Merkmal ist der Beugungswinkel eines Gelenks.

Modellanpassung

Bezeichnet in dieser Arbeit den Vorgang der gezielten Variation der Modellparameter, beispielsweise der Gelenkstellungen, mit dem Ziel, Modell- und Realansicht zur Deckung zu bringen.

Monokularansicht

Ansicht einer Szene durch genau ein Objektiv.

Movement-Hold-Modell

Von Liddell und Johnson in [LJ89] beschriebenes Gebärdenmodell, das eine zeitliche Unterteilung in Bewegungs- bzw. Haltesegmente vornimmt und mit der Vokal-Konsonanten-Unterscheidung in Lautsprachen vergleichbar ist.

OpenGL

Abkürzung für Open Graphics Library, eine plattform- und sprachübergreifende 3D-Grafikbibliothek.

optische Achse

Gerade, die mit der Symmetrieachse des Objektivs übereinstimmt.

Palmarseite

Seite der Handinnenfläche

Phonem

Kleinste bedeutungsunterscheidende Einheit einer Sprache. In gesprochenen Sprachen repräsentiert ein Phonem einen Laut.

Positorenraum

Von den Parametern einer Positur, d. h. von Gelenkstellungen und Ansichtswinkeln aufgespannter Raum im mathematischen Sinne. Die Dimensionalität des Positorenraums entspricht der Summe der Freiheitsgrade in Handform und Handstellung.

Rendern

Erzeugung eines Bildes (z. B. einer Handansicht) auf Grundlage eines Modells.

Segmentierung

Unterteilung eines Bildes in Regionen, die ein Homogenitätskriterium erfüllen und/oder reale Objekte repräsentieren.

simulated annealing („simuliertes Ausglühen“)

Strategie zur Vermeidung von Nebenoptima in iterativen numerischen Optimierungsverfahren durch Überlagern der aktuellen Lösung mit einem Zufallsprozess abnehmender Amplitude.

Spline

Stückweise aus Polynomen zusammengesetzte Funktion zur Interpolation diskreter Daten.

stereo vision

Verwendung zweier Augen oder Kameras zur Verbesserten Erfassung räumlicher Tiefe.

Trajektorie

Position eines bewegten Objekts im Raum als Funktion der Zeit.

Vorbereitung, Ausführung, Rückführung

Aufeinanderfolgende Gebärdenphasen, in denen sich die Hände zum Ausführungsort bewegen, die Gebärde durchführen (Ausführung) und abschließend wieder in die Ruheposition zurückkehren (Rückführung). In Vorbereitungs- und Rückführungsphase werden keine Informationen übermittelt.

Symbolverzeichnis

Allgemeine Notation

x	Skalar
$ x $	Betrag
$\underline{\mathbf{x}}$	Vektor
$\ \underline{\mathbf{x}}\ $	euklidische Norm
$\mathbf{X}$	Matrix
	$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} = [x_{ij}]_{n \times m}$
$\mathbf{X}$	Menge
$ \mathbf{X} $	Kardinalität

Allgemeine Symbole

x, y, z	Ortskoordinaten
t	Zeitindex
T	Sequenzlänge
P	Wahrscheinlichkeit
η	Normalverteilung
μ	Verteilungsmittelwert
Σ	Kovarianzmatrix
σ	Standardabweichung
Θ	Schwellwert
ω	Gewichtungsfaktor
γ	Normierungskonstante
$E(\cdot)$	Erwartungswert
$\langle \dots \rangle$	Tupel, Sequenz

Segmentierung

<u>I</u>	Einzelbild
<u>C</u>	Farbwert
R, G, B	RGB-Farbkomponenten
L^*, u^*, v^*	CIE $L^*u^*v^*$ -Farbkomponenten
Vg	Vordergrund
Hg	Hintergrund
Δ_b	Helligkeitsdifferenz
Δ_c	chromatische Differenz
H	Histogramm
R	Bildregion
K	Marker-Kandidatenregionen

Merkmale

<u>F</u>	Merkmalsvektor
f	Merkmal
f'	normiertes Merkmal
a	Flächeninhalt
l	Umfang
$\underline{s} = \langle r, \varphi \rangle$	Schwerpunkt (Polarkoordinaten)
α	Orientierung
<u>e</u>	Ellipsen-Randkurvenpunkt
r_1, r_2	Ellipsenradien
$\hat{s}_G$	Scheitelpunkt der Gesichtsregion
b_G	Breite der Gesichtsregion
$\Delta_{\underline{F}}$	Merkmalsdifferenz
h	gerichtete Hausdorff-Distanz
H	ungerichtete Hausdorff-Distanz

Posituren

b	Beugungsgrad eines Finger-/Daumengelenks
<u>B</u>	Beugungsgrade aller Finger-/Daumengelenke bzw. Handform
$\phi_{\text{lat}}, \phi_{\text{lon}}, \phi_{\text{rot}}$	Breitengrad, Längengrad, Rotation um optische Achse
<u>Φ</u>	Ansichtswinkel bzw. Handstellung
<u>P</u>	Positur
$\Delta_{\underline{P}}$	Posituredifferenz

$\Delta_{\underline{\mathbf{B}}}$	Beugungsgrad- bzw. Handformdifferenz
$\Delta_{\underline{\Phi}}$	Ansichtswinkel- bzw. Handstellungsdifferenz
$\underline{\mathbf{c}}$	Finger-/Daumenspitzen-Koordinaten
Ψ	Abbildung von $\underline{\mathbf{B}}$ und $\underline{\Phi}$ auf Finger-/Daumenspitzen-Koordinaten
$\mathbf{P}$	Positurenmenge

Hypothesen

H	Hypothese
W	Pfad durch den Hypothesenraum
$\mathbf{W}$	Pfade gleicher Länge
w	Hypothesenindex

Dynamic Time Warping

b	Flags zum Vergleich von Handposition, -stellung und -form
$\mathbf{D}$	DTW-Matrix
Γ	Pfad durch die DTW-Matrix
δ	Positionsabweichung
Δ_v	Geschwindigkeitsabweichung

Kapitel 1

Einleitung

Gehörlose Menschen kommunizieren untereinander in Gebärdensprachen. Die Informationsvermittlung erfolgt dabei primär über die Hände, die hierzu verschiedene Formen, Stellungen und Positionen einnehmen. Jeder konkrete wie auch abstrakte Sachverhalt kann auf diese Weise visuell ausgedrückt werden. Gebärdensprachen sind vollwertige Sprachen, die sich wie Lautsprachen über Jahrhunderte entwickelt haben. Sie verfügen analog über eigene Grammatiken und sind national verschieden. Der Anteil Gehörloser an der Bevölkerung Europas liegt bei ca. 0,1% [Vri97].

Für die Mehrheit der Gehörlosen stellt Gebärdensprache die Muttersprache dar. Schriftsprache ist ihnen nur beschränkt zugänglich, da deren Erlernen in hohem Maße auf Kenntnis der Lautsprache beruht. So finden etwa die Codierung von Phonemen durch Buchstaben oder die Flexion von Worten in der Gebärdensprache keine Entsprechung. Viele Gehörlose erreichen daher lediglich Schriftsprachenkenntnisse auf Grundschulniveau und sind deshalb im Alltag auf die Gebärdensprache als Kommunikationsmittel angewiesen. Leider beherrschen jedoch nur wenige Hörende Gebärdensprache. Die Folge für die Gehörlosen sind Integrationsprobleme und starke Einschränkungen bezüglich eines unabhängigen, selbstbestimmten Lebens.

Seit einigen Jahren ist dieses Problem Gegenstand der Politik. Das 2002 verabschiedete Gleichstellungsgesetz anerkennt die Deutsche Gebärdensprache (DGS) als vollwertige Sprache. Gehörlose haben damit bei allen Amtsgeschäften ein Recht auf einen Gebärdendolmetscher. Auf europäischer Ebene existieren zahlreiche Programme wie etwa eAccessibility oder eInclusion¹, die die Integration sozial oder physisch benachteiligter Minderheiten in die moderne Informationsgesellschaft fördern.

Dem gesteigerten Interesse an Gebärdensprache steht ein Mangel an Dozenten gegenüber. Hieraus ergibt sich ein Markt für multimediale Lernprogramme, die das Selbststudium unterstützen. Allerdings sind diese im Unterschied zu vergleichbarer Software für Lautsprache nicht interaktiv. Dem Benutzer werden in Videoform Gebärden präsentiert, die er beobachtet und imitiert, um sie zu erlernen. Er erhält jedoch keine Rückmeldung hinsichtlich der Korrektheit seiner Gebärdenausführung. Im Hinblick auf einen effizienten Lernprozess stellt dies

¹Informationen hierzu sind unter http://europa.eu.int/information_society/policy/accessibility/ verfügbar.

ein Defizit dar, denn obwohl Gebärden für den Beobachter teilweise intuitiv verständlich sind, ist eine akkurate Reproduktion oft schwierig.

Um den Lernvorgang interaktiv zu gestalten, besteht der Wunsch nach einer automatischen computergestützten Analyse der Gebärdenausführung des Benutzers. Wenn Abweichungen gegenüber der Referenzausführung von einem Rechner erkannt werden, erschließen sich neue Möglichkeiten des Trainings. So ist neben einer quantitativen Fortschrittsbewertung insbesondere auch der gezielte Hinweis auf Fehler möglich. Ein automatischer Gebärdensprachtutor kann diese Informationen in Verbindung mit einem didaktischen Konzept ferner zur dynamischen Anpassung der Lerninhalte an das Verhalten des Benutzers auswerten.

Neben einem vereinfachten Zugang zur DGS für Hörende sind für ein solches System auch Anwendungen im Bereich der Rehabilitation denkbar. Bekannte Verfahren, die Patienten beim Wiedererlernen komplexer Bewegungsabläufe mittels rein reproduktiven Videofeeds unterstützen [TT97], lassen sich so um eine Möglichkeit des unbeaufsichtigten Trainings erweitern. Durch quantitative Rückmeldung der Lernfortschritte entsteht ein zusätzlicher Motivationsfaktor [Ker01].

1.1 Motivation und Aufgabenstellung

Gebärdensprachen verwenden Form, Stellung, Position und Bewegung der Hand als Ausdrucksmittel. Um eine vom Benutzer ausgeführte Gebärde mit der zu lernenden Referenzgebärde vergleichen und ggf. korrekatives Feedback geben zu können, muss ein interaktives Tutorsystem daher zunächst diese Komponenten in jedem Einzelbild des Gebärdenvideos ermitteln. Eine Zuordnung der Benutzergebärde zu diskreten Klassen, wie sie bei reiner Gebärdenspracherkennung geschieht, ist nicht ausreichend.

Die vorliegende Arbeit fasst Handform und -stellung unter dem Begriff der *Handpositur* zusammen. Gegenstand der Rekonstruktion sind darüber hinaus Handposition und Handbewegung, wobei letztere die zeitliche Änderung der übrigen Ausdrucksmittel darstellt. In englischsprachigen Texten wird diese Aufgabenstellung als *hand posture recognition* bezeichnet. Abbildung 1.1 zeigt beispielhaft einige Einzelbilder der Gebärde „Olympia“, die den Verlauf von Handpositur und -position über die Zeit veranschaulichen.

Obwohl ansichtenbasierte Verfahren gute Ergebnisse in der Gebärdenspracherkennung erzie-



Abbildung 1.1: Einzelbilder der Gebärde „Olympia“.

len [AG97, TSS02, Aky03, Bau04, ZK05], sind sie für den hier betrachteten Anwendungsfall ungeeignet. Ansichtenbasierte Merkmale der Hand haben, mit Ausnahme von Raumkoordinaten, keine dem Lernenden ersichtliche anschauliche Entsprechung. Zwar kann eine Übereinstimmung von Benutzer- und Referenzgebärde auf Basis derartiger Merkmale festgestellt werden, um jedoch im Falle einer Abweichung anschauliches korrekatives Feedback zu generieren, ist ein modellbasierter Ansatz erforderlich.

Die in der Literatur beschriebenen modellbasierten Verfahren zur Handpositurerkennung unterscheiden sich hinsichtlich der verwendeten Hardware, der Aufnahmebedingungen und der berücksichtigten manuellen Ausdrucksmittel. Ansätze mit zwei oder mehr Kameras extrahieren aus den Eingangsbildern direkt 3D-Informationen, beispielsweise mittels Korrespondenzanalyse [Nöl00, Reh95] oder volumetrischer Rekonstruktion [BHE⁺02]. Der zusätzliche Hardwareaufwand schränkt gleichzeitig die Bewegungsfreiheit der Hand ein, so dass diese Systeme keine Variation der Handposition erlauben. Meist verwenden sie zudem eine kontrollierte Aufnahmeumgebung, etwa indem die Hand in einen hölzernen Kasten eingeführt wird, der Kameras und Leuchtmittel enthält.

Bezüglich der Anforderungen an die zu verarbeitenden Eingangsdaten lassen sich die Verfahren weiter differenzieren. Einige Systeme verwenden optische Markierungen, die an verschiedenen Stellen der Hand befestigt werden und die Merkmalsextraktion unterstützen [CGH02, BHE⁺02, HOR99a]. Mit wenigen Ausnahmen [AS03b, FAK03, TvdM02] arbeiten sämtliche Systeme mit einfarbigem texturlosem Hintergrund. Meist muss sich die Hand in der Nähe der optischen Achse der Kamera befinden [AS03b, SS03, TvdM02], so dass keine translatorische Bewegung zulässig ist.

Sämtliche Systeme erkennen lediglich eine Untermenge der möglichen Handposituren, etwa solche, die in der Gebärdensprache häufig auftreten [FAK03], oder die anhand ihrer Kontur unterscheidbar sind [ISS04]. Oft müssen aus der Kameraperspektive alle Fingerspitzen bzw. Marker sichtbar sein [CGH02, Nöl00]. Darüber hinaus erfassen zahlreiche Methoden nur einen Teil der manuellen Ausdrucksmittel, häufig beispielsweise Form und Stellung, nicht jedoch Position und Bewegung.

Viele Arbeiten behandeln ausschließlich statische Eingangsbilder und nutzen die Kontinuität dynamischer Gebärden nicht aus. Das prinzipielle Problem der Mehrdeutigkeit der zweidimensionalen Handansicht, beispielhaft gezeigt in Abbildung 1.2, ist bisher in keiner Veröffentlichung methodisch berücksichtigt. Die bekannten Ansätze vermeiden dies meist durch eine manuelle Initialisierung, etwa indem die Hand zu Beginn eine vorgegebene Positur einnimmt. Somit ist die aktuelle Positur bekannt, und die Positurerkennung geschieht durch Anpassen der Modellparameter mit dem Ziel einer möglichst guten Übereinstimmung der resultierenden Modellansicht mit dem Eingangsbild. Dieses Vorgehen ist empfindlich gegen Störungen, da ein einmal fehlerhaft parametrisiertes Modell nicht mehr korrekt adaptiert werden kann [Der04]. In Anbetracht der unter realistischen Bedingungen stark verrauschten Eingangsdaten erscheint diese Strategie somit nicht Erfolg versprechend.

Es besteht daher Bedarf an einem Verfahren, das die in der Gebärdensprache verwendeten ma-

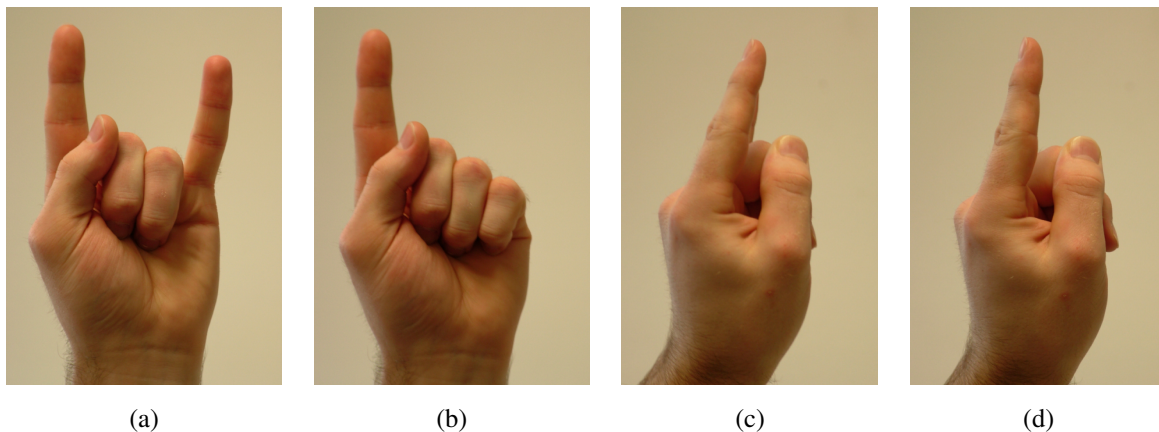


Abbildung 1.2: Verschiedene Handposituren (a, b) sind je nach Blickwinkel in einer monokularen Ansicht nicht unterscheidbar (c, d).

nuellen Ausdrucksmittel vollständig erfasst und gleichzeitig sämtliche Handposituren zulässt. Da die Segmentierung der unmarkierten bewegten Hand vor unkontrolliertem Hintergrund nach heutigem Wissensstand nur eingeschränkt möglich ist, erscheint für eine robuste Lösung die Verwendung optischer Markierungen angebracht. Ferner soll die Aufnahme mit nur einer Kamera erfolgen, um die Hardwareanforderungen zu reduzieren und den Aufnahmebereich nicht einzuschränken. Weil die Rekonstruktion eines dreidimensionalen verformbaren Objekts aus einer zweidimensionalen Ansicht im Allgemeinen ein unterbestimmtes Problem darstellt, folgen hieraus Mehrdeutigkeiten, die auf der Basis eines Einzelbildes prinzipiell nicht eliminiert werden können. Erst die Auswertung mehrerer zeitlich aufeinanderfolgender Ansichten erlaubt die Auflösung solcher Mehrdeutigkeiten. Aus diesem Grund sind Algorithmen zu entwerfen, die die Gesamtheit aller Beobachtungen zu jedem Zeitpunkt in die Ergebnisbildung einbeziehen. Die Entwicklung einer für diese Problemstellung geeigneten Lösung ist Gegenstand der vorliegenden Arbeit.

1.2 Interaktives Tutorsystem für Gebärdensprache

Ein interaktiver Gebärdensprachtutor verfügt über ein Vokabular, aus dem zunächst eine zu lernende Gebärde ausgewählt und dem Benutzer in Form einer Videoaufnahme eines menschlichen Gebärdensprechers präsentiert wird. Anschließend reproduziert der Anwender diese Vorgabe. Das System nimmt die Ausführung mit Hilfe einer monokularen Kamera auf und vergleicht sie mit der bekannten Referenzausführung. Im Falle einer Abweichung erhält der Lernende korrekatives Feedback, das eventuelle Fehler seiner Ausführung in der Gebärdenaufnahme visuell kennzeichnet und ihn so bei der Korrektur unterstützt. Zusätzlich bietet sich die Möglichkeit, die Lernfortschritte des Anwenders quantitativ zu erfassen und die Lerninhalte dynamisch anzupassen, so dass gut gelernte Gebärden seltener und Schwierigkeiten bereitende häufiger präsentiert werden. Abbildung 1.3 visualisiert diesen Ablauf.

Der Benutzer trägt farbig markierte Stoffhandschuhe, die einfach herzustellen sind und die

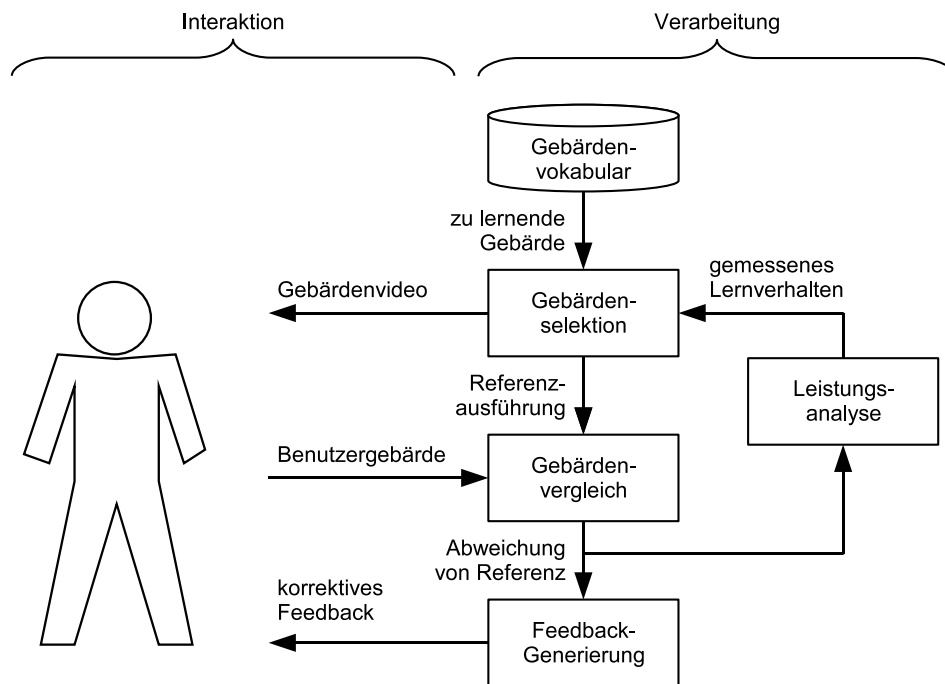


Abbildung 1.3: Anwendungsszenario für einen interaktiven Gebärdensprachtutor.

Bewegungsfreiheit der Hände nicht einschränken. Die Gestaltung der Markierungen hat eine möglichst hohe Varianz der zweidimensionalen Ansicht verschiedener Handposituren zum Ziel. Abbildung 1.4 zeigt hierzu ein Beispiel. Der gesamte von der Kamera erfasste Bildbereich ist für die Bewegung nutzbar.

Zur Rekonstruktion der Handpositur wird ein 3D-Handmodell eingesetzt, das dieselben Markierungen aufweist wie der Stoffhandschuh. Dieses Modell ist in Abbildung 1.5 gezeigt. Es kann jede anatomisch mögliche Handform annehmen und lässt sich aus beliebigem Winkel



Abbildung 1.4: Ausführung einer Gebärde mit farbig markiertem Stoffhandschuh.

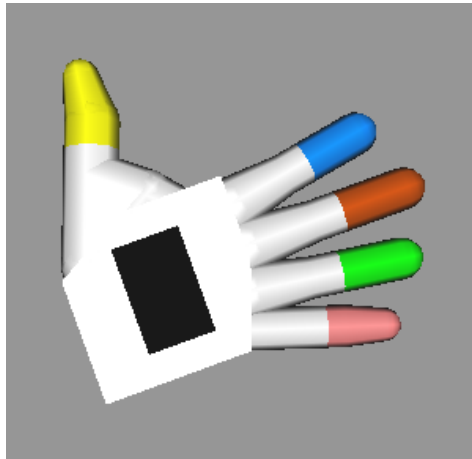


Abbildung 1.5: Das 3D-Handmodell, das zur Erkennung der Handpositur so parametrisiert und orientiert wird, dass synthetische und reale Ansicht übereinstimmen.

betrachten. Die Erkennungsaufgabe besteht in der Identifikation einer geeigneten Parametrisierung für dieses Modell sowie eines Betrachtungswinkels, so dass die resultierende Ansicht derjenigen der Benutzerhand entspricht.

Abbildung 1.6 stellt das realisierte Verfahren schematisch dar. Zunächst findet eine Trainingsphase statt, in der das 3D-Handmodell jede Positur annimmt, die später erkennbar sein soll. Durch Variation der Beugungswinkel von Fingern und Daumen sowie des Betrachtungswinkels ergeben sich bei geeigneter Quantisierung einige Millionen verschiedene Handansichten. In jeder einzelnen werden zweidimensionale geometrische Merkmale extrahiert und zusammen mit den der Ansicht zu Grunde liegenden Positurparametern in einer Datenbank gespeichert. Dieser einmalige Vorgang ist insbesondere aufgrund der Berechnung der 2D-Ansichten zeitintensiv und geschieht daher offline.

In der Anwendungsphase erfolgt die Aufnahme des Benutzers aus monokularer Frontalan-sicht. Die optische Achse der Kamera steht damit senkrecht zu der Ebene, in der translatorische Bewegungen am häufigsten auftreten, so dass diese optimal erfasst werden können.

Die Erkennung der Benutzergebärde beginnt mit der Lokalisation der einzelnen Markierungen. Aus diesen ergeben sich, analog zur Merkmalsextraktion in der Trainingsphase, zweidimensionale geometrische Merkmale, die die Ansicht der Hand beschreiben.

Diese Merkmale werden nun in der zuvor erstellten Datenbank gesucht. Aufgrund der Projektion in die zweidimensionale Bildebene ist jedoch prinzipiell kein eindeutiger Rückschluss auf die zu Grunde liegende dreidimensionale Konfiguration möglich. So können verschiedene Handposituren zu annähernd identischen 2D-Merkmalen führen. Darüber hinaus ist infolge von Bild- und Quantisierungsrauschen in den Merkmalen sowie der endlich genauen Diskretisierung des Positurenraums bei der Erstellung der Datenbank nicht damit zu rechnen, eine exakte Übereinstimmung zu finden. Es erfolgt daher ein unscharfer Vergleich, der $k > 1$ Positurkandidaten liefert. Die Suche verwendet ein effizientes Indexierungsverfahren, das die Anzahl der zu betrachtenden Datenbankeinträge im Vergleich zu einer linearen Suche um

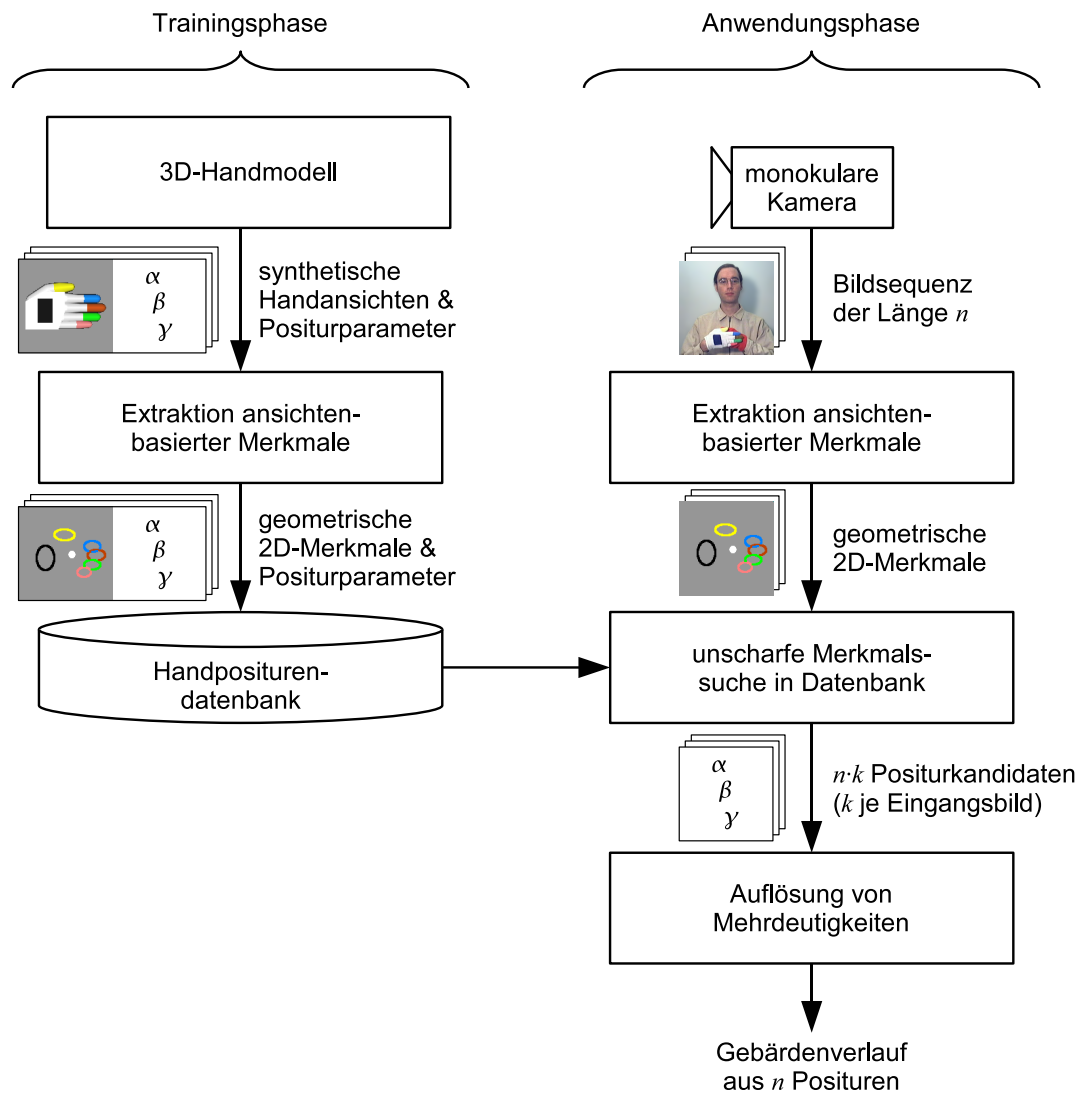


Abbildung 1.6: Die Erkennung der Handpositur erfolgt durch Vergleich der Eingangsbilder mit synthetisch generierten Handansichten, deren 2D-Merkmale während einer Trainingsphase in einer Datenbank gespeichert wurden. Nachdem für alle n Bilder der Eingangssequenz je k Positurkandidaten ermittelt sind, erfolgt die Auflösung der durch die Projektion in die Bildebene bedingten Mehrdeutigkeiten durch Selektion einer Positurensequenz der Länge n , die einer flüssigen kontinuierlichen Bewegung entspricht.

mehrere Größenordnungen reduziert.

Für jedes einzelne einer n Bilder umfassenden Eingangssequenz ergeben sich so jeweils k Kandidaten, die die beobachtete Markerkonfiguration erklären. Aufgrund der probleminhärenten Mehrdeutigkeit wird die Auswahl eines Gewinners nicht unmittelbar vollzogen, sondern geschieht erst nach Abschluss der Gebärde und unter Berücksichtigung des zeitlichen Verlaufs der Bewegung. Aus der Menge der $n \cdot k$ Kandidaten ist eine Sequenz der Länge n zu ermitteln, so dass einerseits in jedem Einzelbild eine möglichst große Ähnlichkeit zur Beobachtung besteht, andererseits aber auch eine flüssige Bewegung über die Zeit resultiert. Der

verfolgte Ansatz erlaubt die Auswertung aller verfügbaren Informationen vor der endgültigen Festlegung des Gebärdenverlaufs. Eine Initialisierung durch Einnehmen einer vorgegebenen Handpositur ist nicht erforderlich.

In der Gebärdensprache wird ein Großteil der Information lediglich von einer der beiden Hände übermittelt. Diese dominante Hand ergibt sich aus der Händigkeit des Sprechers. Die andere, nichtdominante Hand hat bei beidhändigen Gebärden die Rolle eines Gegenparts oder spiegelt die Bewegung der dominanten Hand [Boy95]. Bei einhändigen Gebärden bleibt sie unbewegt. Somit erscheint für die nichtdominante Hand aus Gründen der Laufzeiteffizienz eine vereinfachte Behandlung angebracht. Das System beschränkt sich hier auf die Bestimmung der Handposition unter Verwendung eines einfarbig markierten Handschuhs.

Die Zusammenstellung des Trainingskorpus erfordert für jede Gebärde eine Aufnahme mit unmarkierten Händen, die dem Lernenden präsentiert wird, sowie eine weitere mit Markierungen, die dem Tutorsystem als Referenz dient. Zur Bewertung der Benutzergebärde ist eine Kennzeichnung der drei Gebärdenphasen Vorbereitung, Ausführung und Rückführung, sowie die Differenzierung zwischen *movements* und *holds* (Bewegung und Verweilen) [Lid84] hilfreich. So liegt der Fokus während der Bewegung auf der Position und für die Dauer des Verweilens auf der Positur.

Der Vergleich von Benutzer- und Referenzgebärde beginnt mit einer zeitlichen Anpassung beider Signale, die zulässige Variationen in der Ausführungsgeschwindigkeit ausgleicht. Anschließend geschieht die Prüfung von Handposition und -bewegung. Stimmen diese mit der Vorgabe überein, werden Handform und -stellung verglichen. In beiden Fällen erfolgt bei Abweichungen eine visuelle Hervorhebung von Soll- und Istzustand unmittelbar in der Videoaufnahme des Benutzers. Der Lernende hat die Möglichkeit, dieses korrektive Feedback beliebig häufig und in variabler Geschwindigkeit zu betrachten.

Durch die Analyse der Gebärdenausführungen des Benutzers lässt sich eine Statistik erstellen, die eine quantitative Lernfortschrittsbewertung erlaubt. Ferner können diese Daten in das Tutorsystem zurückfließen, um gut gelernte Gebärden zu identifizieren und die Auswahl der präsentierten Gebärden auf die bisher ungelernten Teile des Vokabulars zu konzentrieren.

1.3 Gliederung

Im Folgenden wird die inhaltliche Struktur der Arbeit kurz vorgestellt und ein Überblick der einzelnen Kapitel gegeben.

Nach einer Einführung in die Gebärdensprache erfolgt in Kapitel 2 eine umfassende Darstellung des gegenwärtigen Kenntnisstandes in Form einer systematischen Analyse von insgesamt 18 Veröffentlichungen zur Handpositurerkennung sowie drei Tutorsystemen. Auf dieser Grundlage wird das Konzept des entwickelten interaktiven Gebärdensprachtutors erläutert.

Der weitere Aufbau der Arbeit orientiert sich an der Reihenfolge der Verarbeitung im Rechner. Kapitel 3 beschreibt die funktionale Anatomie der menschlichen Hand und das darauf

basierende dreidimensionale Modell, das die zentrale Komponente der Positurrekonstruktion bildet. Weiterhin werden Verfahren zum Vergleich zweier Posituren entworfen, die für mehrere Module der Verarbeitungskette wichtig sind.

Gegenstand von Kapitel 4 ist die Segmentierung der Eingangsbilder, die aus einer Hintergrundsubtraktion sowie der Bestimmung der Marker- und Gesichtsregionen besteht. Aus den geometrischen Merkmalen dieser Regionen ergibt sich nach geeigneter Normierung der zweidimensionale Merkmalsvektor als Ausgangspunkt für die in Kapitel 5 behandelte Positurerkennung.

Hier wird zunächst die Datenbank beschrieben, die etwa 100 Millionen Handposituren sowie deren Merkmale abdeckt. Mittels des anschließend eingeführten Merkmalsdistanzmaßes lassen sich zu jedem Eingangsbild mehrere visuell übereinstimmende Kandidatenposituren ermitteln, aus denen hinsichtlich des Gebärdenverlaufs zunächst multiple Hypothesen entstehen. Die nachfolgende Ermittlung der Gewinnerhypothese als wahrscheinlichste Positurenfolge schließt den Rekonstruktionsvorgang ab.

Kapitel 6 befasst sich mit der Generierung korrektiven Feedbacks. Dies umfasst die Repräsentation der Referenzgebärde, den Vergleich mit der Anwendergebärde sowie die im Falle einer Abweichung gegebenen textuellen und visuellen Korrekturhinweise.

Eine ausführliche Evaluation unter Verwendung synthetischer sowie realer Handansichten beinhaltet Kapitel 7. Ausgehend von einem geeigneten Genauigkeitsmaß werden verschiedene Komponenten des Systems untersucht, zahlreiche quantitative Auswertungen durchgeführt und die gemessenen Resultate diskutiert.

Zusammenfassung und Ausblick erfolgen in Kapitel 8. Details zu eingesetzten Algorithmen und Parametern sind im Anhang zu finden.

Kapitel 2

Grundlagen eines interaktiven Gebärdensprachtutors

Dieses Kapitel beschreibt die gebärdensprachlichen und technischen Grundlagen, die für den interaktiven Tutor von Bedeutung sind. Zunächst erfolgt in Abschnitt 2.1 eine Einführung in Struktur und Eigenschaften der Gebärdensprache, da diese die zur rechnerbasierten Verarbeitung angewandten Methoden maßgeblich bestimmen. Abschnitt 2.2 stellt kurz die Motivation zur videobasierten Erfassung von Gesten und Gebärden dar. Darauf folgt eine Analyse und tabellarische Gegenüberstellung bestehender Ansätze zur Positurrekonstruktion sowie eine Vorstellung existierender Gebärdensprachtutoren in den Abschnitten 2.3 bzw. 2.4. Schließlich entsteht in Abschnitt 2.5 das Konzept eines interaktiven Tutorsystems, dessen Hauptkomponente ein Verfahren zur Handpositurerkennung bildet, das für die Anwendung im Kontext der Gebärdensprache geeignet ist.

2.1 Gebärdensprache

Gebärdensprachen verwenden Gestik und Mimik zur visuellen Übermittlung von Informationen. Sie stellen ein vollwertiges Kommunikationsmedium dar und besitzen ein ähnlich komplexes Vokabular wie Lautsprachen. Analog zu diesen haben sie sich historisch entwickelt, sind national verschieden und verfügen über regionale Dialekte. Die Grammatik der Gebärdensprachen unterscheidet sich von der der Lautsprachen. So existieren etwa in der Deutschen Gebärdensprache keine Artikel, und der Satzbau folgt anderen Regeln als in der deutschen Lautsprache.

Im Unterschied zu Lautsprachen sind zahlreiche Aspekte der Gebärdensprachen wenig erforscht. Insbesondere existiert keine allgemein anerkannte und verbindliche Definition einzelner Gebärden, die etwa dem Duden vergleichbar wäre. Die Verschiedenheit der Dialekte ist daher relativ stark ausgeprägt. Häufig finden sich mehrere, zum Teil signifikant verschiedene Gebärden mit derselben Bedeutung. Derartige Variationen werden auch durch das Fehlen standardisierender nationaler Medien wie etwa Fernsehen oder Radio begünstigt. Aufgrund

von Kontextwissen ist Gebärdensprache für Menschen dennoch problemlos verständlich. Für die rechnerbasierte Erkennung ergeben sich allerdings erhebliche Schwierigkeiten, da dieses Kontextwissen nur sehr eingeschränkt modellierbar ist.

Im Allgemeinen entspricht eine Gebärde einem einzelnen Wort. Wie in der Lautsprache besteht ein Satz aus einer Abfolge mehrerer Einzelgebärden, die fließend ineinander übergehen. Dadurch entstehen Transitions- und Koartikulationseffekte, die dem Verschleifen gesprochener Worte vergleichbar sind.

Viele Gebärdensprachen sind heute offiziell anerkannt, wodurch Gehörlosen beispielsweise bei Amtsgeschäften das Recht auf einen Dolmetscher zugesichert wird. Der Anerkennungsprozess ist Bestandteil der Integration von Gehörlosen in die überwiegend aus Hörenden bestehende Gesellschaft. Er wird auf europäischer Ebene unter anderem durch Gesetze gefördert, verläuft jedoch national verschieden. Schweden hat beispielsweise bereits 1980 die Schwedische Gebärdensprache anerkannt. In Deutschland geschah dies für die DGS im Jahr 2002.

2.1.1 Ausdrucksmittel

Während Lautsprache als akustisches Signal sequenziell, d. h. als Folge einzelner Phoneme übertragen wird, kommen in der Gebärdensprache mehrere verschiedene visuelle Komponenten simultan zum Einsatz. Die größte Informationsmenge vermitteln die Hände. Form, Stellung, Position und Bewegung der Hand bilden die hierbei verwendeten manuellen Ausdrucksmittel.

Die beiden Hände haben in der Gebärde unterschiedliche Rollen. Während die *dominante* Hand die Gebärde bestimmt, bleibt die *nichtdominante* Hand bei einhändigen Gebärden unbewegt oder stellt bei zweihändigen einen Gegenpart dar, etwa indem sie die Bewegung der dominanten Hand spiegelt. Abbildung 2.1 zeigt dies am Beispiel der Gebärde „Buch“. Im



Abbildung 2.1: Die DGS-Gebärde „Buch“. Die nichtdominante (linke) Hand bewegt sich symmetrisch zur dominanten (rechten) Hand.

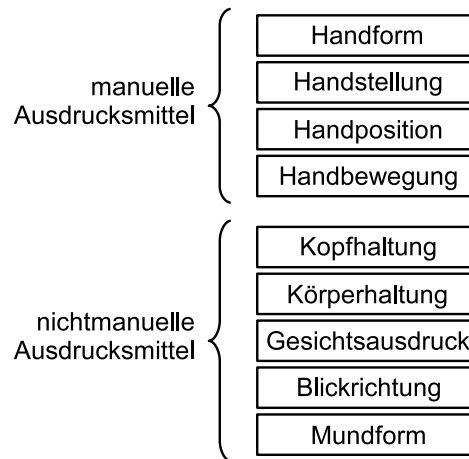


Abbildung 2.2: Ausdrucksmittel in der Gebärdensprache.

Allgemeinen bestimmt die Händigkeit des Sprechers, welche Hand die dominante ist.

Nichtmanuelle Ausdrucksmittel der Gebärdensprache sind Kopf- und Körperhaltung, Gesichtsausdruck, Blickrichtung und Mundform. Diese können beispielsweise differenzierende Funktion haben oder Adjektive und Adverbien ausdrücken. So unterscheiden sich etwa die DGS-Gebärden „Bruder“ und „Schwester“ nur in der Mundform, und „gelassen Auto fahren“ kombiniert die manuellen Komponenten der Gebärde „Auto fahren“ mit einem gelassenen Gesichtsausdruck.

Abbildung 2.2 zeigt eine Übersicht der manuellen und nichtmanuellen Ausdrucksmittel. Aufgrund ihrer Relevanz für die automatische Handstellungserkennung werden die manuellen Komponenten im Folgenden näher erläutert.

2.1.1.1 Handform

Die Handform ergibt sich durch die Stellung der Gelenke in Fingern und Daumen. Theoretisch resultieren aus den Möglichkeiten zur Flexion (Beugung) und Extension (Streckung) sowie Abduktion (Spreizen) und Adduktion (Schließen) insgesamt 21 Freiheitsgrade. Die zusätzliche Berücksichtigung der Mittelhand, die etwa zum Greifen einer Kugel ebenfalls leicht deformiert werden kann, erhöht diese Anzahl auf 23 [Stu92]. In der Praxis bewirken jedoch verschiedene Abhängigkeiten der Gelenke voneinander sowie Bewegungseinschränkungen durch das umgebende Gewebe eine deutliche Reduzierung der theoretischen Freiheitsgrade. So ist es beispielsweise den meisten Menschen nicht möglich, die Gelenke eines Fingers unabhängig voneinander zu bewegen, und eine Beugung des Mittelfingers verhindert üblicherweise die vollständige Streckung des Ringfingers.

Die zahlreichen Freiheitsgrade erlauben eine Vielzahl unterschiedlicher Handformen, die der Differenzierung zwischen verschiedenen Gebärden dienen. Abbildung 2.3 zeigt die Gebärden „essen“ und „rot“, die sich ausschließlich in der Handform unterscheiden, sonst jedoch gleich ausgeführt werden. Zwei derartige bedeutungsverschiedene Gebärden, die sich ledig-



Abbildung 2.3: Die Gebärden „essen“ und „rot“, die ein Minimalpaar hinsichtlich der Handform bilden.

lich in einem der in Abbildung 2.2 aufgelisteten Ausdrucksmittel unterscheiden, werden als Minimalpaar bezeichnet.

2.1.1.2 Handstellung

Die Ausrichtung der Hand in Bezug auf den Körper des Sprechers wird als Handstellung bezeichnet. Sie kann mit Hilfe von Schulter-, Ellenbogen- und Handgelenk variiert werden. Vom Standpunkt des Betrachters entspricht dies einer Rotation der Hand um ihren Schwerpunkt, die sich durch Breiten- und Längengrad auf einer imaginären Kugeloberfläche sowie eine Rotation um die Kameraachse angeben lässt und damit drei Freiheitsgrade besitzt. Für eine gegebene Handstellung existiert sprecherübergreifend im Allgemeinen genau eine anatomisch nahe liegende Konfiguration der relevanten Gelenke. Die Erfassung der Handstellung deckt somit implizit auch diese Gelenke ab.



Abbildung 2.4: Die Gebärden „du“ und „heute“ als Minimalpaar hinsichtlich der Handstellung.

Abbildung 2.4 zeigt die Gebärden „du“ und „heute“ als Beispiel eines Minimalpaars in Bezug auf die Handstellung.

2.1.1.3 Handposition

Die Handposition entspricht den Raumkoordinaten des Handmittelpunkts relativ zum Körper. Sie weist somit drei Freiheitsgrade auf. Beim Gebärden befinden sich die Hände überwiegend in einem begrenzten Bereich vor der Brust des Sprechers, der als Gebärdenraum bezeichnet wird [Boy95] und für beide Hände bequem erreichbar ist. Der Gebärdenraum ist in Abbildung 2.5 visualisiert. Sein Zentrum liegt unterhalb des Gesichts auf Brusthöhe, der Durchmesser beträgt etwa eine Armlänge. Für die videobasierte Gebärdenerkennung bestimmt der Gebärdenraum den Aufnahmebereich der Kamera.

Abbildung 2.6 zeigt die Gebärden „fragen“ und „Hunger“ als Minimalpaar bezüglich der Handposition.



Abbildung 2.5: Der als Gebärdenraum bezeichnete Bereich, in dem sich die Hände während des Sprechens befinden [Boy95].



(a) „fragen“

(b) „Hunger“

Abbildung 2.6: Die Gebärden „fragen“ und „Hunger“, die bezüglich der Handposition ein Minimalpaar bilden.

2.1.1.4 Handbewegung

Die Handbewegung ist definiert als die zeitliche Änderung der drei zuvor beschriebenen Ausdrucksmittel Form, Stellung und Position. Bei einhändigen Gebärden wird ausschließlich die dominante Hand bewegt. Bei zweihändigen Gebärden bewegt sich die nichtdominante Hand oft parallel oder spiegelsymmetrisch zur dominanten Hand.

Schnelle Handbewegungen erschweren für den Beobachter die Wahrnehmung von Handform und -stellung. Dies trifft sowohl auf das menschliche visuelle System zu, das bei bewegten Objekten weniger Details erfasst als bei statischen, als auch auf Kameras, die aufgrund endlicher Verschlussgeschwindigkeit Bewegungsunschärfe erzeugen, wenn das Motiv während der Blendenöffnung nicht annähernd stillsteht.

Einen Teilaspekt der Handbewegung stellt die Trajektorie dar, die die Handposition im Raum als Funktion der Zeit angibt. Experimente belegen, dass zahlreiche Gebärden allein aufgrund der Trajektorie, d. h. ohne Kenntnis von Form und Stellung, erkennbar sind [PKB90]. Abbildung 2.7 visualisiert die Trajektorien beider Hände für die Gebärde „bookings“ der Britischen Gebärdensprache. Abbildung 2.7a zeigt drei charakteristische Einzelbilder, die den Gebärdenverlauf veranschaulichen. Aus der Verbindung der Handmittelpunkte aller Einzelbilder ergeben sich die in Abbildung 2.7b gezeigten Kurven, denen als Referenz die Körperhaltung zu Beginn der Aufnahme unterlegt ist.



Abbildung 2.7: Die Gebärde „bookings“ der Britischen Gebärdensprache (a) sowie die entsprechenden Trajektorien der beiden Hände (b). Im Hintergrund ist als Referenz die Körperhaltung zu Beginn der Aufnahme gezeigt.

2.1.2 Phonologisches Modell

Liddell und Johnson beschreiben in [LJ89] ein strukturelles Gebärdensmodell, das sich auf die Handbewegung stützt, um Gebärden zeitlich in mehrere Abschnitte zu unterteilen. Es

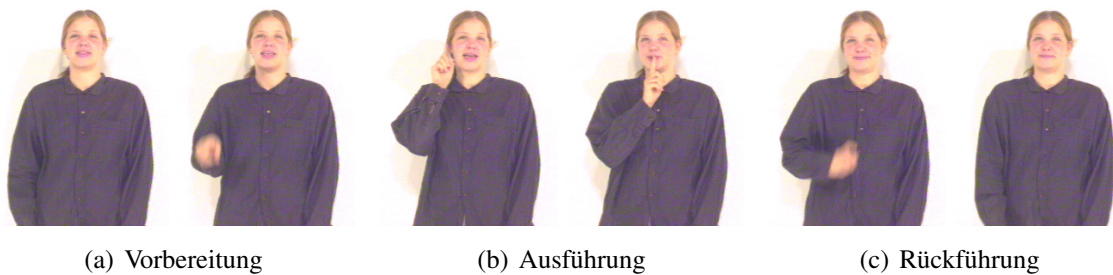


Abbildung 2.8: Die drei Phasen Vorbereitung, Ausführung und Rückführung am Beispiel der Gebärde „gehörlos“.

unterscheidet Segmente, in denen sich die Hand bewegt (*movement*), und solche, in denen sie stillsteht (*hold*), und wird entsprechend als *Movement-Hold-Modell* bezeichnet.

Parallel hierzu lässt sich eine Gebärde in die drei aufeinanderfolgenden Phasen Vorbereitung, Ausführung und Rückführung (*preparation*, *stroke* und *retraction*) gliedern. Zu Beginn werden die Hände von der Ruhestellung (z. B. seitlich am Körper) in den Gebärdenraum bewegt, wo die Ausführung der Gebärde geschieht. Danach nehmen die Hände wieder die Ruhestellung ein. Die Vorbereitungs- und die Rückführungsphase codieren keine Informationen und sind daher für die Interpretation der Gebärde ohne Bedeutung. Abbildung 2.8 veranschaulicht die drei Phasen am Beispiel der DGS-Gebärde „gehörlos“.

Folgen in einer Äußerung mehrere Gebärden aufeinander, so gehen deren Ausführungsphasen ineinander über. Die Bewegung der Hände nach Abschluss einer Gebärde an den Ausführungsort der nächsten Gebärde wird als Transition bezeichnet. Dabei beeinflussen die zeitlich benachbarten Gebärden einander, so dass die Ausführung je nach Abfolge leicht unterschiedlich sein kann. Dieser Effekt ist aus der Lautsprachforschung als Koartikulation bekannt.

2.1.3 Abgrenzung zur Gestik

Gebärdensprache weist neben zahlreichen Ähnlichkeiten einige wichtige Unterschiede zu Gestik im Allgemeinen auf. Um eine klare Abgrenzung zu ermöglichen, erfolgt an dieser Stelle ergänzend eine Definition des Begriffs der Gestik. In der Literatur finden sich hierzu verschiedene Vorschläge.

Für die vorliegende Arbeit ist es sinnvoll, Gestik als die allgemein verständliche Verwendung des Körpers zur visuellen Kommunikation zu definieren. Gebärdensprachen fallen nicht hierunter, da sie nur von einem eingeschränkten Personenkreis verstanden werden. Gestik wird hingegen unbewusst von allen Menschen erlernt und häufig auch unbewusst eingesetzt. Sie dient der Unterstützung der Lautsprache und kann diese in einigen Fällen, z. B. bei der Übermittlung einfacher Informationen wie „rechts“ oder „links“, auch ersetzen.

Gestik bedient sich somit denselben Ausdrucksmittel wie Gebärdensprache (vgl. Abbildung 2.2). Für die Verarbeitung im Rechner hat dies zur Folge, dass dieselben Verfahren einsetzbar sind. Allerdings stellt die Gebärdensprache einen festen Rahmen dar, da beispiels-



Abbildung 2.9: Die Geste und DGS-Gebärde „Uhrzeit“.

weise die Bedeutung einzelner Gebärden nicht frei bestimmt werden kann. Gesten sind hingegen meist weniger präzise definiert und gestatten daher bei der Systementwicklung eine höhere Flexibilität, etwa hinsichtlich der Auswahl des Vokabulars eines Erkennungssystems.

Einzelne Gesten existieren mit derselben Bedeutung auch in Gebärdensprachen. In der DGS ist die Gebärde „Uhrzeit“, die in Abbildung 2.9 dargestellt ist, ein Beispiel für eine solche Überschneidung.

2.2 Videobasierte Erfassung von Gebärden und Gesten

Die Natürlichkeit und Intuitivität gestenbasierter Kommunikation motiviert etwa seit Mitte der 80er Jahre Bestrebungen, diese Vorteile für die Bedienung technischer Geräte nutzbar zu machen. Während die Erfassung manueller Ausdrucksmittel zunächst mechanisch mittels Datenhandschuhen erfolgte, ermöglicht die Rechenleistung moderner Computer eine videobasierte und damit berührungslose Auswertung. Dies erhöht für den Anwender den Komfort und erlaubt gleichzeitig die Berücksichtigung nichtmanueller Ausdrucksmittel wie etwa des Gesichtsausdrucks. Darüber hinaus sind Kameras heute bereits Bestandteil vieler interaktiver Systeme, so dass eine Anschaffung zusätzlicher Hardware oft nicht erforderlich ist.

Neben der direkten Verwendung der visuell ermittelten Handform, -stellung, -position und -bewegung, beispielsweise für manipulative Aufgaben im Bereich der Virtuellen Realität (VR) oder der computergestützten Konstruktion (*computer aided design*, CAD), eröffnet sich die Möglichkeit der Interpretation dieser Daten durch den Rechner [ZC06]. Gestenerkennung stellt eine intuitive Möglichkeit zur Steuerung interaktiver Systeme dar. Gebärdenspracherkennung leistet dies speziell für Gehörlose und verfolgt daneben langfristig das Ziel, durch die automatische Übersetzung von Gebärden- in Laut- oder Schriftsprache die Kommunikation zwischen Gehörlosen und Hörenden zu fördern [ZCBK06, vAZC⁺07].

Interaktive Tutorsysteme stellen eine neue Anwendung der videobasierten Erfassung von Gebärdensprache dar. Im Unterschied zur Erkennungsaufgabe ist hier die präsentierte Gebärde bekannt, die Ausführung jedoch möglicherweise fehlerhaft. Die Aufgabe des Rechners liegt im Vergleich der Benutzergebärde mit einer gespeicherten Referenzausführung. Im Falle einer Abweichung wird der Benutzer hierauf hingewiesen und durch entsprechendes Feedback bei der Korrektur seiner Ausführung unterstützt. Daneben ist eine quantitative Leistungsana-

lyse möglich, die zur Motivation sowie zur dynamischen Anpassung der Lerninhalte genutzt werden kann.

2.2.1 Einordnung der Verfahren

Die verschiedenen Systeme, die zur Erfassung manueller Ausdrucksmittel eingesetzt werden, unterscheiden sich hinsichtlich ihrer Zielsetzung, der verarbeiteten Eingangsdaten sowie des zur Interpretation dieser Daten verfolgten Ansatzes. Abbildung 2.10 zeigt eine entsprechende schematische Einordnung. In den folgenden Abschnitten werden die einzelnen Kriterien diskutiert.

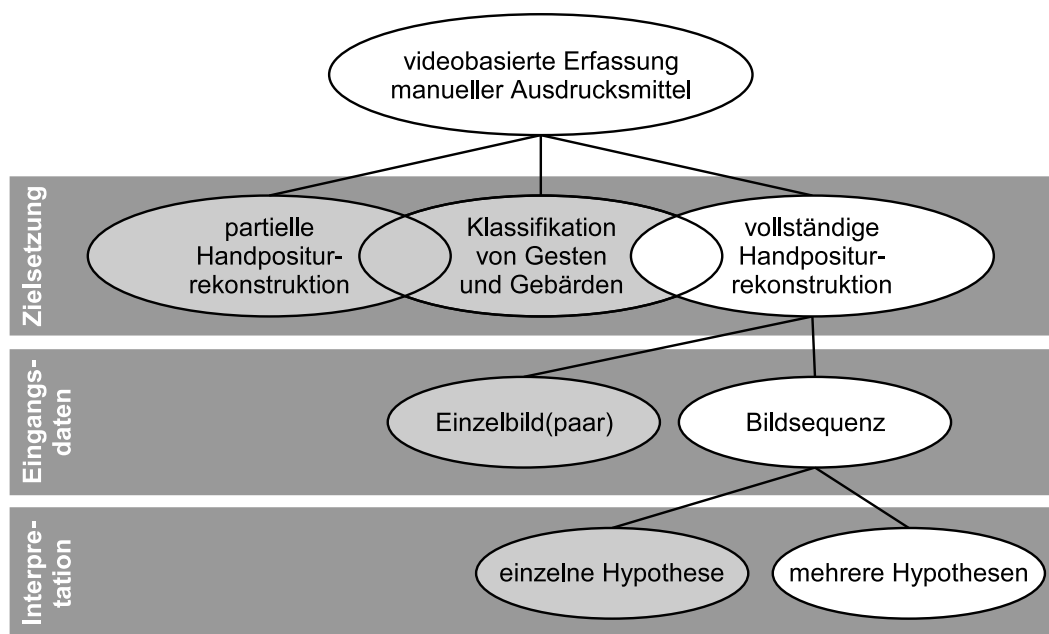


Abbildung 2.10: Schematische Übersicht der verschiedenen Verfahren zur videobasierten Erfassung manueller Ausdrucksmittel. Grau ausgefüllte Ansätze sind für die Verwendung in einem Tutorsystem ungeeignet und werden daher nicht weiter untergliedert.

Zielsetzung

Hinsichtlich der Zielsetzung ist zunächst zu unterscheiden, ob eine vollständige Rekonstruktion von Handform und -stellung erfolgt, oder ob sich die Rekonstruktion auf Teilaspekte wie etwa die Position des Handmittelpunkts oder einzelner Fingerspitzen beschränkt. Beide Vorgehensweisen erlauben eine Klassifikation bezüglich einer diskreten Gesten- oder Gebärdenmenge (Vokabular). Dies ist in Abbildung 2.10 durch Schnittmengen veranschaulicht.

Ansätze mit partieller erreichen gegenüber solchen mit vollständiger Positurrekonstruktion gegenwärtig höhere Erkennungsraten bei größeren Vokabularen. In [ZK05] werden 232 Gebärden mit einer Erkennungsrate von 99,3% klassifiziert, wobei neben der Handposition le-

diglich ansichtenbasierte 2D-Merkmale zum Einsatz kommen. Auf Basis derselben Merkmale ist auch ein personenunabhängiger Einsatz möglich [vASZK06].

Mittels vollständiger Positurrekonstruktion erreicht ein in [HOR99a] vorgestelltes System hingegen bei einer Vokabulargröße von nur 22 eine Erkennungsrate von 95,5%. Dieser deutliche Unterschied legt nahe, dass die Positurrekonstruktion die Aussagekraft der 2D-Merkmale nicht erhöht. Da keine Informationen gewonnen werden, die nicht bereits implizit vorhanden und für einen Klassifikator verwertbar sind, bedeutet der zusätzliche Verarbeitungsschritt lediglich eine Fehler- bzw. Rauschquelle. Aktuelle Arbeiten verwenden diesen Ansatz daher nicht [OR05].

Auch ohne Positurrekonstruktion kann eine Gebärdensklassifikation erfolgen [CT98, MG98, RKE98, ZKN05], allerdings sind die Vokabularen hier mit maximal 24 Elementen vergleichsweise klein. Gegenwärtig wird diese Strategie selten verfolgt.

Für den Einsatz in einem Gebärdentutor ist, wie eingangs beschrieben (s. Abschnitt 1.1), eine vollständige Positurrekonstruktion erforderlich, um aussagekräftiges korrekatives Feedback generieren zu können.

Eingangsdaten

Bezüglich der verarbeiteten Eingangsdaten lassen sich die Verfahren in zwei Gruppen unterteilen. Systeme, die lediglich ein statisches Einzelbild – bzw. Bildpaar bei Verwendung zweier Kameras (*stereo vision*) – analysieren, repräsentieren einen gedächtnislosen Ansatz. Sie werden zum Teil auch auf Bildsequenzen angewandt, betrachten diese jedoch als Folge unkorrelierter Einzelbilder. Im Unterschied hierzu ist bei der Verarbeitung von Bildsequenzen die Kontinuität der Bewegung und die daraus folgende Ähnlichkeit zeitlich benachbarter Posituren eine zentrale Komponente. Die Rekonstruktion erfolgt gedächtnisbehaftet, wobei die Positur im aktuellen Einzelbild die Grundlage für die Analyse des darauf folgenden Bildes ist.

Interpretation

Das Resultat der Positurrekonstruktion kann auf verschiedene Weise interpretiert werden. Systeme, die lediglich eine einzelne Hypothese berücksichtigen, liefern in jedem Einzelbild die jeweils wahrscheinlichste Positur als Ergebnis. Dieses Vorgehen hat den Nachteil, dass die folgenden Bilder bzw. die daraus extrahierten Informationen für die Rekonstruktion im aktuellen Bild keine Berücksichtigung finden. Darüber hinaus führen Mehrdeutigkeiten, die insbesondere bei Verwendung einer monokularen Ansicht unvermeidbar sind, zu einer hohen Fehlerwahrscheinlichkeit. Wenn die aktuelle Positurschätzung von der tatsächlichen Positur abweicht, besteht keine Möglichkeit der Korrektur, da sich die Systemzustand aufgrund der Annahme zeitlicher Kontinuität nur langsam ändern kann.

Eine Alternative besteht darin, pro Einzelbild nicht nur eine, sondern mehrere plausible Posituren zu ermitteln. Diese Positurstandidaten entsprechen in ihrer Ansicht den Eingangsda-

ten, können jedoch aufgrund der Mehrdeutigkeit der Problemstellung stark unterschiedlich sein. Jede mögliche Positurreihenfolge stellt eine Hypothese hinsichtlich der Interpretation der Eingangsdaten dar. Zusätzlich zur Positurerkennung sind daher geeignete Algorithmen erforderlich, die die wahrscheinlichste Hypothesenfolge bestimmen und somit das eindeutige Endergebnis der Rekonstruktion festlegen. Dieser Ansatz ermöglicht die Behandlung von Mehrdeutigkeiten in den Eingangsdaten. Rekonstruktionsfehler in einzelnen Bildern können durch Analyse der folgenden Bilder und Selektion einer besseren Hypothese korrigiert werden.

Die Betrachtung mehrerer Hypothesen ist lediglich bei der Verarbeitung von Bildsequenzen wirkungsvoll, da hier die zeitliche Korrelation der Bildfolge ein Auflösen von Mehrdeutigkeiten erlaubt.

2.2.2 Algorithmen zur Positurrekonstruktion

In der Literatur werden zwei verschiedene Klassen von Algorithmen unterschieden, die aus einem zweidimensionalen Eingangsbild bzw. einem Stereo-Bildpaar eine dreidimensionale Handpositur rekonstruieren. Diese sind im Folgenden kurz beschrieben.

Die Methode der **inversen Kinematik** berechnet Lage und Position der Elemente einer kinematischen Kette, d. h. mehrerer starrer, durch Gelenke miteinander verbundener Elemente, aus der Lage des letzten Elements. Entsprechende Verfahren, die im Bereich der Robotik zur Positionierung eines Endeffektors wie etwa eines Greifers verwendet werden, lassen sich auch auf die Bestimmung der Gelenkstellungen in Fingern, Daumen und Handgelenk anwenden. Hier bilden die Spitzen von Fingern und Daumen die Endelemente einer fünffach verzweigten kinematischen Kette. Aus ihrer Position wird die Handpositur in jedem Einzelbild rekonstruiert. In Abbildung 2.11 ist dieser Vorgang schematisch dargestellt.

Im Allgemeinen ist die Problemstellung bei inverser Kinematik unterbestimmt, jedoch ergeben sich im Fall der Hand aus anatomischen Faktoren mehrere Einschränkungen. So können etwa die Gelenke eines Fingers nicht unabhängig voneinander bewegt werden. Auf diese Weise lässt sich die Zahl der Freiheitsgrade reduzieren, bis eine Lösung zu finden ist.

Die Modellierung der Hand erfolgt in Form der zur Berechnung verwendeten Algorithmen. Ein visuell darstellbares Handmodell ist lediglich zur Veranschaulichung des Ergebnisses erforderlich. Hierarchisch stellt die inverse Kinematik ein *Bottom-Up*-Verfahren dar, da die zuerst bestimmten Fingerspitzenpositionen eine niedrigere Abstraktionsebene repräsentieren als die daraus ermittelte Handpositur. Die Verarbeitung der Eingangsdaten erfolgt prinzipiell einzelbildweise und ohne Berücksichtigung zeitlicher Korrelation.

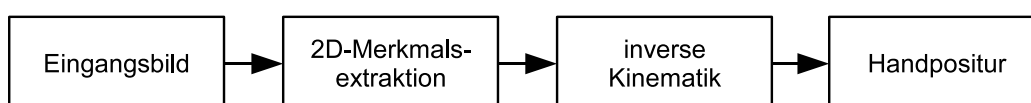


Abbildung 2.11: Schematische Darstellung der Handpositurrekonstruktion mittels inverser Kinematik.

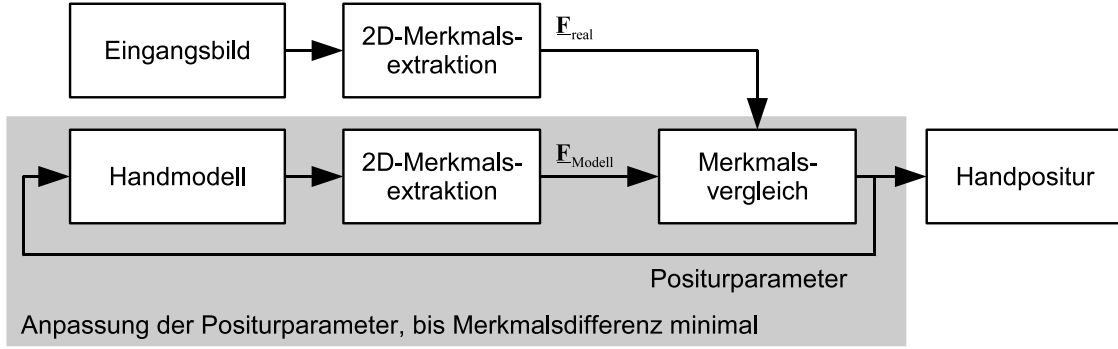


Abbildung 2.12: Schematische Darstellung der Handpositurrekonstruktion mittels Modellanpassung.

Der Ansatz der **Modellanpassung** bestimmt die Parameter eines dreidimensionalen Handmodells so, dass dessen zweidimensionale Darstellung möglichst gut mit den Eingangsdaten übereinstimmt. Der Grad der Übereinstimmung wird aus dem Vergleich der entsprechenden 2D-Merkmalvektoren $\underline{\mathbf{F}}_{\text{real}}$ und $\underline{\mathbf{F}}_{\text{Modell}}$ quantitativ berechnet. Das Handmodell muss alle Posituren, die das System erkennen soll, annehmen können. Abbildung 2.12 veranschaulicht dieses Verfahren.

Die Merkmale $\underline{\mathbf{F}}_{\text{real}}$ werden aus den Eingangsdaten berechnet. Der Merkmalsvektor $\underline{\mathbf{F}}_{\text{Modell}}$ ergibt sich analog aus der Modellansicht, die wiederum allein von der Positur $\underline{\mathbf{P}}$ des Modells abhängt, und kann daher als $\underline{\mathbf{F}}_{\text{Modell}}(\underline{\mathbf{P}})$ geschrieben werden. Oft wird das gesamte Bild bzw. der Ausschnitt, der die Hand enthält, mit einer geeigneten Funktion gefaltet oder pixelweise transformiert und als Merkmalsvektor verwendet.

Die Verschiedenheit der beiden Merkmalsvektoren wird durch eine skalare Merkmalsdifferenz $\Delta_{\underline{\mathbf{F}}}$ gemessen:

$$\Delta_{\underline{\mathbf{F}}}(\underline{\mathbf{F}}_{\text{real}}, \underline{\mathbf{F}}_{\text{Modell}}) \in \mathbb{R}_0^+ \quad \text{mit} \quad \Delta_{\underline{\mathbf{F}}}(\underline{\mathbf{F}}_{\text{real}}, \underline{\mathbf{F}}_{\text{Modell}}) = 0 \Leftrightarrow \underline{\mathbf{F}}_{\text{real}} = \underline{\mathbf{F}}_{\text{Modell}} \quad (2.1)$$

Damit lässt sich diese Methode als mehrdimensionale Optimierung auffassen, die die dem Eingangsbild ähnlichste Positur $\hat{\underline{\mathbf{P}}}$ gemäß

$$\hat{\underline{\mathbf{P}}} = \underset{\underline{\mathbf{P}}}{\operatorname{argmin}} \Delta_{\underline{\mathbf{F}}}(\underline{\mathbf{F}}_{\text{real}}, \underline{\mathbf{F}}_{\text{Modell}}(\underline{\mathbf{P}})) \quad (2.2)$$

bestimmt. Diese Gleichung entspricht dem grau hinterlegten Bereich in Abbildung 2.12. Da die Minimierung von einer Positur $\underline{\mathbf{P}}$ ausgeht und daraus entsprechende Merkmale $\underline{\mathbf{F}}_{\text{Modell}}$ berechnet, liegt ein *Top-Down*-Ansatz vor. Im Unterschied zur inversen Kinematik lässt sich die Modellanpassung auf die Verarbeitung von Bildsequenzen optimieren, indem durch Berücksichtigung von Nebenminima in (2.2) mehrere Hypothesen generiert werden.

2.3 Leistungsfähigkeit existierender Systeme zur Positurrekonstruktion

Die Erkennung der visuellen Ausdrucksmittel Handform, -stellung, -position und -bewegung findet neben der Gesten- und Gebärdenerkennung auch Anwendung in intuitiven Mensch-Maschine-Schnittstellen, etwa in den Bereichen der Virtuellen Realität und der computergenerierten Animation. Obwohl videobasierte Verfahren zur Handpositurerkennung seit etwa zwei Jahrzehnten Gegenstand der Forschung sind, liegt ihre Leistungsfähigkeit heute noch deutlich unterhalb der des menschlichen visuellen Systems. Eine Annäherung an dessen Leistung ist lediglich in Teilbereichen und unter einschränkenden Bedingungen möglich, die dem praktischen Einsatz oft entgegenstehen. Beispielsweise erfassen einige Systeme aus einem festen Blickwinkel Handformen, bei denen alle Fingerspitzen sichtbar sind, wohingegen der Mensch problemlos beliebige Handposituren unter nahezu allen Ansichtswinkeln erkennen kann.

Der gegenwärtige Kenntnisstand hinsichtlich der menschlichen Wahrnehmung und Verarbeitung optischer Reize reicht nicht aus, eine entsprechende Nachbildung zu implementieren [Pal99, Gol97]. Angesichts der sequenziellen Funktionsweise heutiger Computer erscheint es zudem fraglich, ob eine hochgradig parallele Verarbeitung, wie sie beim Sehprozess stattfindet [Gol97], prinzipiell realisierbar ist.

Aus diesen Gründen sind existierende Systeme zur Handpositurerkennung an die Funktionsweise des Rechners angepasst. Zur Umsetzung der in Abschnitt 2.2 skizzierten Rekonstruktionsmethoden wird eine Vielzahl verschiedener Strategien angewandt, die jeweils unterschiedliche Randbedingungen zur Folge haben. Einige wichtige Randbedingungen sind in Tabelle 2.1 aufgeführt. Die einzelnen Kategorien sind voneinander unabhängig und die entsprechenden Bedingungen somit beliebig kombinierbar.

Die Wahl der Randbedingungen bestimmt die Anwendungsmöglichkeiten eines Verfahrens und ist gleichzeitig maßgeblich für die auftretenden Schwierigkeiten, die in Abschnitt 2.3.1 kurz beschrieben werden. Abschnitt 2.3.2 umreißt die Anforderungen, die sich im Kontext der Gebärdensprache an die Positurrekonstruktion ergeben. Darauf folgt eine Diskussion existierender Systeme, unterteilt in Methoden der inversen Kinematik und der Modellanpassung, in den Abschnitten 2.3.3 bzw. 2.3.4. Alle Verfahren sind zusätzlich in Tabelle 2.2 (s. Seiten 27 bis 32) ausführlich beschrieben. Abschnitt 2.3.5 enthält eine Bewertung der Ansätze sowie einen tabellarischen Vergleich hinsichtlich der wichtigsten Systemeigenschaften.

2.3.1 Schwierigkeiten

Der durch die Projektion in die Bildebene bedingte Informationsverlust lässt sich auch bei Verwendung mehrerer Kameras nicht vollständig kompensieren. *Stereo-Vision*-Ansätze liefern zwar Tiefeninformationen und können somit einen Teil der verlorenen Informationen rekonstruieren, schränken dabei jedoch das Aktionsfeld der Hand ein und erfordern aufwändige

Tabelle 2.1: Wichtige Randbedingungen bei der visuellen Handpositurrekonstruktion.

Eingangsdaten	Kameras • 1 Kamera, monokulare Ansicht • 2 Kameras, binokulare Ansicht (<i>stereo vision</i>)
	Verarbeitungskontext • Einzelbild • Bildsequenz als Folge unkorrelierter Einzelbilder • korrelierte Bildsequenz
	Hintergrund • kontrolliert (einfarbig) • unkontrolliert
Funktionsweise	Merkmalsextraktion • Kontur und/oder Textur der unmarkierte Hand • optische Markierungen
	Initialisierung • automatisch oder nicht erforderlich (gedächtnisloses Verfahren) • manuell (Hand nimmt zu Beginn eine vorgegebene Positur ein)
Rekonstruktion	Rekonstruierte Ausdrucksmittel • Handform • Handstellung • Handposition
	Einschränkung des Positurreaums • Sichtbarkeit charakteristischer Punkte erforderlich • Ansichtswinkel beschränkt/konstant • keine Einschränkung

Spezialhardware.

Die Extraktion geeigneter Merkmale unter realistischen Bedingungen bereitet erhebliche Probleme. Unkontrollierte Beleuchtung, niedrige Bildauflösung und Bewegungsunschärfe durch endliche Verschlusszeiten erschweren Detektion und Segmentierung der Hand und führen zu starkem Rauschen in den berechneten Merkmalen. Darüber hinaus ist die Aussagekraft der aus einem zweidimensionalen Einzelbild extrahierten Merkmale hinsichtlich der dreidimensionalen Handkonfiguration beschränkt. Charakteristische Punkte, beispielsweise die Fingerspitzen, sind aufgrund von Selbstverdeckung häufig nicht sichtbar und können zur Positurerkennung daher nur eingeschränkt verwendet werden.

Mehrdeutigkeiten in der 2D-Ansicht können bei lokalen Optimierungsstrategien zu Nebenminima führen. Bei globalen Optimierungen bewegen sich die Unterschiede in der Zielfunktion

zwischen verschiedenen Kandidaten in der Größenordnung des Merkmalsrauschens, so dass das tatsächliche globale Minimum nicht zuverlässig identifizierbar ist.

Infolge der zahlreichen Freiheitsgrade der Hand ist der Positurreaum, d. h. der Raum aller anatomisch möglichen Posituren, hochdimensional. Die Verarbeitung in (annähernd) Echtzeit wird dadurch in verschiedener Hinsicht erschwert. Eine lineare Suche ist aus Geschwindigkeitsgründen selbst bei grober Quantisierung nicht praktikabel. Gleiches gilt für die Anwendung von Monte-Carlo-Strategien [Der04]. Das Reduzieren des Suchbereichs auf die Nachbarschaft der aktuellen Positur beschränkt die Möglichkeiten zur Fehlerkorrektur und führt daher zu Problemen bei ungenauen Merkmalsvektoren, die aus den oben genannten Gründen jedoch in der Praxis zu erwarten sind.

2.3.2 Anforderungen

Um eine Bewertung der anschließend vorgestellten Ansätze zu ermöglichen, ist zunächst eine kurze Definition der Anforderungen sinnvoll, die sich bei der Verwendung in einem Gebärdensprachtutor ergeben. Dazu sind im Folgenden zwei Kriterien beschrieben, die einfach und objektiv zu prüfen sind und die Menge der Methoden bereits stark einschränken. Für eine Anwendung in der Praxis ist die Erfüllung dieser Kriterien notwendig, jedoch nicht hinreichend. Insbesondere die Verarbeitungsgeschwindigkeit wird an dieser Stelle nicht berücksichtigt.

Rekonstruktion von Handform und Handstellung

Von den manuellen Ausdrucksmitteln Form, Stellung, Position und Bewegung müssen zumindest Handform und Handstellung – d. h. alle Gelenkstellungen jenseits des Handgelenks sowie der Ansichtswinkel – rekonstruiert werden. Zwar sind alle vier Komponenten in der Gebärdensprache von elementarer Bedeutung, jedoch ist die Bestimmung von Position und Bewegung vergleichsweise einfach und wird daher als nachträglich ergänzbar angenommen, selbst wenn sie in der Systembeschreibung keine Erwähnung findet.

Ausreichend feine Quantisierung des Positurreaums

Im Kontext eines Tutors ist die Rekonstruktion annähernd beliebiger Handformen und -stellungen eine grundlegende Voraussetzung. Eine Beschränkung auf übliche, in der Gebärdensprache häufig verwendete Handformen steht der gewünschten Erkennung von Ausführungsfehlern entgegen. Einige Systeme liefern zwar Form und Stellung der Hand, erkennen jedoch nur eine geringe Anzahl verschiedener Posituren.

Geeignet sind somit nur Ansätze, die eine parametrische Repräsentation der Handpositur mit ausreichend feiner Quantisierung der Parameter berechnen. Für diese Quantisierung lässt sich eine Untergrenze wie folgt abschätzen. Die Handstellung äußert sich aus der Beobachterperspektive durch eine Rotation der Hand um die Achsen eines dreidimensionalen Koordinaten-

systems, dessen Ursprung im Handgelenk liegt. Unterteilt man diese drei Winkel äußerst grob in 36° -Schritten und fordert zudem eine mindestens binäre Quantisierung des Beugungswinkels jedes Fingers sowie des Daumens, so ergibt sich als Untergrenze für die Mächtigkeit der Menge aller Ergebnisposituren eine Anzahl von $360^\circ \cdot 360^\circ \cdot 180^\circ \cdot (36^\circ)^{-3} \cdot 2^5 = 16000$ Posituren.¹ Verfahren, die diese Grenze nicht überschreiten, sind daher ungeeignet.

2.3.3 Systeme zur Positurrekonstruktion durch inverse Kinematik

Lee et al. [LK95] beschreiben ein Verfahren, das offline auf einem statischen Stereo-Bildpaar arbeitet und einen der frühesten Ansätze zur Positurerkennung darstellt. Inverse Kinematik und Modellanpassung werden alternierend angewandt, um ein Handmodell iterativ an die Eingangsdaten anzugleichen. Fünf farbige Marker an den Fingerspitzen sowie jeweils ein weiterer am Knöchel des kleinen Fingers und am Handgelenk bilden die Zielpunkte, die das Modell zu erreichen versucht. Zunächst erfolgt eine Ausrichtung am Handgelenk. Danach wird durch eine Drehung des Modells die Handstellung so angepasst, dass die übrigen Marker ihre jeweiligen Zielpunkte annähern. Hinsichtlich der Handform geschieht diese Annäherung anschließend mittels inverser Kinematik, wobei die Handstellung konstant bleibt. Die letzten beiden Schritte – Anpassung von Stellung und Form – werden alternierend ausgeführt, bis die Differenz zwischen Modell und Eingangsdaten eine Schwelle unterschreitet oder eine maximale Iterationsanzahl erreicht ist. Die Autoren gehen von der Sichtbarkeit aller Marker aus; zur Behandlung von Verdeckungen durch die Hand machen sie keine Angaben. Somit ist nur eine Untermenge aller Handposituren erkennbar. Eine Evaluation erfolgt qualitativ auf einigen Testaufnahmen.

Das von **Lathuilière et al. [LH99, LH00]** präsentierte System verwendet einen mit acht farbigen Markern versehenen Handschuh zur Merkmalsextraktion. Fünf Marker kennzeichnen die Fingerspitzen, drei bilden ein Dreieck auf dem Handrücken und müssen zur Positions- und Orientierungsschätzung in jedem Bild sichtbar sein. Die Positurrekonstruktion kann dank Restriktionen bezüglich der Unabhängigkeit der Beugung einzelner Gelenke eines Fingers rein analytisch erfolgen. Die Forderung nach der Sichtbarkeit der drei Handrückenmarkierungen bedeutet, dass sich der Beobachter in dem dem Handrücken zugewandten Halbraum befinden muss, was die erlaubten Ansichtswinkel um mindestens 50% reduziert. Sämtliche Posituren, in denen die Handinnenfläche sichtbar ist, können nicht erkannt werden. Die Autoren messen auf einer synthetischen Testsequenz mit 25 Einzelbildern eine mittlere Abweichung von 5° je Fingergelenk. Hierbei ist allerdings davon auszugehen, dass die dem Ansatz zu Grunde liegenden Abhängigkeitsannahmen Bestandteil des zur Generierung der Testdaten verwendeten Handmodells sind.

Die Arbeit von **Nölker [Nöl00]** ist wohl die am häufigsten zitierte deutschsprachige Publikation. Das beschriebene System erkennt Posituren, bei denen die Spitzen von Fingern und Daumen sichtbar sind. Die Hand befindet sich dabei in einem innen schwarzen Kasten, der

¹Überstreichen zwei der drei Winkel den Wertebereich $[0^\circ, 360^\circ)$, deckt der Wertebereich $[0^\circ, 180^\circ)$ für den dritten Winkel alle möglichen Ansichten ab.

Tabelle 2.2: Existierende Ansätze zur Handpositurerkennung.

	Lee et al. [LK95]	Lathuilière et al. [LH99, LH00]	Nölker [Nöl00]
Kameras	2	1	1 bzw. 2
Eingangsdaten	Einzelbildpaar	Bildsequenz ^b	Einzelbild/-paar
Hintergrund	kontrolliert	kontrolliert	kontrolliert
Beleuchtung	k. A.	kontrolliert	kontrolliert
Auflösung	k. A.	k. A.	192×144
Algorithmus	inverse Kinematik + Modellanpassung	inverse Kinematik	inverse Kinematik
Merkmale	7 Markerschwerpunkte	8 Markerschwerpunkte	2D- bzw. 3D-Position der Fingerspitzen (aus Textur)
Initialisierung	nicht erforderlich	nicht erforderlich	nicht erforderlich
Hypothesen	1	1	1
Restriktionen	alle Marker sichtbar	drei Marker auf Handrücken immer sichtbar	alle Fingerspitzen sichtbar, konstante Objektweite
Erfasste Ausdrucksmittel	Form ^a , Stellung ^a , Position	Form ^a , Stellung ^a , Position	Form ^a
Trainingsdaten	nicht erforderlich	nicht erforderlich	einige tausend Realbilder + synthetische Daten mittels direkter Kinematik
Testdaten	einige Realbilder	1 synthetische Sequenz	einige Realbilder
Leistung	qualitative Auswertung	mittlerer Fehler je Gelenk ca. 5°	qualitative Auswertung
Geschwindigkeit	k. A. (offline)	15 Hz (PowerMac 7500 mit 180-MHz- PowerPC-604e-CPU)	9,5 Hz ^c /1,5 Hz ^d (Pentium Pro, 200 MHz, 64 MB RAM)

^adurch Restriktionen eingeschränkt^bals Folge unkorrelierter Einzelbilder verarbeitet^c1 Kamera^d2 Kameras

Tabelle 2.2: Existierende Ansätze zur Handpositurerkennung (Fortsetzung).

	Guan et al. [GCH99, GCH01, CGH02]	Rehg et al. [RK93, RK94, RK95, Reh95]	Heap et al. [HH96a, HH96b]
Kameras	1	2	1
Eingangsdaten	Einzelbild	Bildsequenz	Bildsequenz
Hintergrund	kontrolliert	kontrolliert	kontrolliert
Beleuchtung	k. A.	kontrolliert	k. A.
Auflösung	k. A.	k. A.	k. A.
Algorithmus	inverse Kinematik	Modellanpassung	Modellanpassung
Merkmale	8 Markerschwerpunkte	Kontur ^b /Textur ^c	Kontur
Initialisierung	nicht erforderlich	vorgegebene Positur	vorgegebene Positur
Hypothesen	1	1	1
Restriktionen	alle Marker müssen sichtbar sein	keine/beschränkte Überlappung	keine Selbstverdeckung
Erfasste Ausdrucksmittel	Form ^a , Stellung ^a	Form ^{a,b} , Stellung ^{a,b} , Position ^b /Beugung des Zeigefingers ^c	Form ^{a,d} , Stellung ^{a,d} , Position
Trainingsdaten	1 Ansicht der Hand zur Kalibrierung	„ausreichend viele“ Ansichten der Hand ^c	8 3D-Handaufnahmen (MRI), 1 Person
Testdaten	6 Realbilder	zwei Realbildsequenzen (200 ^b /80 ^c Bilder)	k. A.
Leistung	qualitative Auswertung	qualitative Auswertung	qualitative Auswertung
Geschwindigkeit	k. A.	6,7 Hz (68040-CPU, 5 MB RAM, SG Indigo 2) ^b /k. A. (offline) ^c	10 Hz (SG Indigo, 134 MHz)

^adurch Restriktionen eingeschränkt^bkeine Überlappung^cbei Überlappung^dsofern durch lineare Transformation aus den Trainingsdaten zu berechnen

Tabelle 2.2: Existierende Ansätze zur Handpositurerkennung (Fortsetzung).

	Holden et al. [HOR99a]	Ouhaddi et al. [OH99]	Rosales et al. [RASS01]
Kameras	1	1	1
Eingangsdaten	Bildsequenz	Bildsequenz	Bildsequenz ^d
Hintergrund	kontrolliert	kontrolliert	kontrolliert
Beleuchtung	kontrolliert	k. A.	k. A.
Auflösung	256×196	k. A.	k. A.
Algorithmus	Modellanpassung	Modellanpassung	Modellanpassung
Merkmale	13 Markerschwerpunkte ^b	Kontur	7 Hu-Momente aus Kontur
Initialisierung	vorgegebene Positur	automatisch	nicht erforderlich
Hypothesen	1	1	1
Restriktionen	Stellung und Position konstant	Ansicht senkrecht auf Handinnenfläche, manuelle Segmentierung des Arms	keine Rotation in Bildebene
Erfasste Ausdrucks-mittel	Form	Form, Position ^a	Form, Stellung ^a , Position
Trainingsdaten	22 Gesten/Gebärden (teils statisch), konstante Handstellung	1 Ansicht der Hand zur Kalibrierung	8000 synthetische Posituren
Testdaten	22 Gesten/Gebärden (teils statisch), konstante Handstellung	einige Realbildsequenzen	ca. 300.000 synthetische Posituren, einige Realbildsequenzen
Leistung	Erkennungsrate ^c 95%	qualitative Auswertung	nicht näher definiertes Fehlermaß ^e /qualitative Auswertung ^f
Geschwindigkeit	k. A.	k. A.	k. A.

^adurch Restriktionen eingeschränkt^bbei Selbstverdeckung Extrapolation mittels Kalmanfilter^cerfasst keine systematischen Fehler, daher kein Rückschluss auf Rekonstruktionsleistung^dals Folge unkorrelierter Einzelbilder verarbeitet^esynthetische Testdaten^freale Testdaten

Tabelle 2.2: Existierende Ansätze zur Handpositurerkennung (Fortsetzung).

	Shimada et al. [SKSK00, SKS01, ISS04]	Stenger et al. [Ste01, SMC01b, SMC01a]	Stenger et al. [STTC06]
Kameras	1	1–2	1
Eingangsdaten	Bildsequenz	Bildsequenz	Bildsequenz
Hintergrund	kontrolliert	unkontrolliert	unkontrolliert
Beleuchtung	k. A.	k. A.	k. A.
Auflösung	k. A.	360×288	320×240
Algorithmus	Modellanpassung	Modellanpassung	Modellanpassung
Merkmale	Kontur	Kontur	Textur
Initialisierung	automatisch	vorgegebene Positur	automatisch
Hypothesen	5	1	1
Restriktionen	keine	Handform bis auf Daumen konstant	Ansichtswinkel auf Halbkugel
Erfasste Aus- drucksmittel	Form, Stellung	Form ^a , Stellung, Po- sition	Position sowie ent- weder Form (mit 3 DOF) oder Stellung ^a
Trainingsdaten	86016 synthetische Posituren, jeweils leichte Variation von Beugungs- und An- sichtswinkel	nicht erforderlich	nicht spezifizierte Anzahl Handformen mittels Datenhand- schuh
Testdaten	7 Realbildsequenzen zu je 100 Bildern	3 Realbildsequenzen zu je 80 Bildern	einige Realbildse- quenzen
Leistung	81,3% „korrekte“ Rekonstruktion (nicht näher definiert)	qualitative Auswer- tung	6,6–9,3 Pixel RMS für charakteristische Punkte ^b
Geschwindig- keit	10,8 Hz (40 Pentium- III-PCs, je 1-GHz- CPU/512 MB RAM)	3 Hz (Celeron, 433 MHz)	0,5 Hz (Pentium IV, 1 GHz) ^c /0,3 Hz (Pen- tium IV, 2,4 GHz) ^d

^adurch Restriktionen eingeschränkt^bDie Hand belegt etwa 5–10% der Bildfläche.^cbei Rekonstruktion von Position und Stellung^dbei Rekonstruktion von Position und Form (mit 2 DOF)

Tabelle 2.2: Existierende Ansätze zur Handpositurerkennung (Fortsetzung).

	Wu et al. [WLH01]	Vittrup et al. [VSM02]	Athitsos et al. [AS03b, AS03a]
Kameras	1	2	1
Eingangsdaten	Bildsequenz	Bildsequenz	Einzelbild
Hintergrund	kontrolliert	kontrolliert	unkontrolliert
Beleuchtung	k. A.	kontrolliert	k. A.
Auflösung	k. A.	k. A.	80×110 – 120×160
Algorithmus	Modellanpassung	Modellanpassung	Modellanpassung
Merkmale	Kontur	Kontur	Textur, Kontur ^f
Initialisierung	automatisch	vorgegebene Positur	nicht erforderlich
Hypothesen	1	1	1
Restriktionen	Ansicht senkrecht auf Handinnenfläche	Stellung und Position teilweise konstant	Form in Trainingsdaten enthalten
Erfasste Ausdrucksmittel	Form (7 DOF)	Form, Stellung ^a , Position ^a	Form ^a , Stellung
Trainingsdaten	30000 Handformen mittels Datenhandschuh	nicht erforderlich	107.328 synthetische Posituren (26 Formen)
Testdaten	1 Realbildsequenz (100 Bilder)	12 synthetische Sequenzen zu je 250 Bildern ^b , 6 davon mit konstanter Position	250/276 ^f Realbilder (verschiedene Personen) der 26 Trainingsformen
Leistung	qualitative Auswertung	Abweichung aller 23 Gelenke in Summe 11,8 cm ^c /87 cm ^d	13,6% ^{g,h} bzw. 8% ^{g,i} /21,7% ^{f,g}
Geschwindigkeit	k. A.	0,001 Hz	0,07 Hz/0,3 Hz ^f (Athlon 1,2 GHz, 300 MB/100 MB ^f RAM)

^adurch Restriktionen eingeschränkt^bDie Sequenzen bestehen jeweils aus einer linearen Interpolation zwischen 2 von 8 verschiedenen Handformen.^cbei konstanter Position^dbei variabler Position^eAuflösung der *bounding box*^fin [AS01, AS02a, AS02b] beschriebenes Vorgängersystem^gkorrekte Handform, Abweichung der Ansichtswinkel max. 30°^hperfekte Segmentierung, kein Hintergrundⁱ*Bounding-Box*-Segmentierung, unkontrollierter Hintergrund

Tabelle 2.2: Existierende Ansätze zur Handpositurerkennung (Fortsetzung).

	Fillbrandt et al. [Fil03, FAK03]	Lu et al. [LMSO03]	Metaxas et al. [MT05]
Kameras	1	1	1
Eingangsdaten	Bildsequenz	Bildsequenz	Bildsequenz
Hintergrund	unkontrolliert	kontrolliert	kontrolliert, teilw.
Beleuchtung	k. A.	kontrolliert	kontrolliert
Auflösung	k. A.	k. A.	k. A.
Algorithmus	Modellanpassung	Modellanpassung	Modellanpassung
Merkmale	Kontur, Textur	Textur	Textur
Initialisierung	vorgegebene Positur	automatisch	automatisch
Hypothesen	1	1	1
Restriktionen	nur in DGS verwen- dete Handformen, Ansicht nicht aus Fingerrichtung	keine Bewegung ver- deckter Finger ^d	letzten 15 Posituren müssen als Sequenz in Trainingsdaten enthalten sein ^e
Erfasste Aus- drucksmittel	Form ^a , Stellung ^a , Position	Form ^a , Stellung ^a , Position	Form ^a , Stellung ^a , Position
Trainingsdaten	Realbildsequenzen aller zu erkennenden Posituren ^b	nicht erforderlich	108 synthetische Se- quenzen aus 29 An- sichtswinkeln, je 15 Einzelbilder
Testdaten	einige Realbildse- quenzen	2 Realbildsequenzen	3 Realbildsequenzen
Leistung	Fehler 10% (Beugungsgrad) bzw. 10–20° (Ansichtswinkel) ^c	qualitative Auswer- tung	qualitative Auswer- tung
Geschwindig- keit	4 Hz (Pentium III, 1 GHz)	4 Hz (Pentium IV, 1 GHz)	k. A.

^adurch Restriktionen eingeschränkt^bIn diesen Sequenzen wurde die Positur für jedes Einzelbild manuell eingegeben.^cDer Fehler wurde visuell geschätzt und gilt für die Trainingsperson. Für nicht trainierte Personen wird der Fehler qualitativ als höher angegeben.^dDiese Restriktion ergibt sich aus der Forderung, dass die Zuordnung von Bildmerkmalen zu einzelnen Fingern aufgrund der Konfiguration im letzten Einzelbild zuverlässig schätzbar sein muss.^eDie Sequenzen werden als Ganzes erkannt, d. h. das System schätzt keine Einzelpositur, sondern eine Positurreihe.

Leuchtmittel und Kamerahardware enthält. Objektweite sowie Handstellung sind dadurch konstant, die Handinnenfläche ist dem Betrachter zugewandt. Das Verfahren arbeitet mit Intensitätsbildern. Zunächst geschieht die Lokalisation der Fingerspitzen basierend auf Texturinformationen mittels einer zweistufigen Anordnung neuronaler Netze, die hierzu mit einigen tausend realen Handansichten trainiert werden. Das System kann mit einer oder zwei Kameras arbeiten und ermittelt diese Positionen entsprechend in 2D bzw. 3D. Hieraus bestimmt ein neuronales PSOM-Netz (*parameterized self-organizing map* [Rit93]) die Gelenkstellungen. Die Trainingsdaten dieses Netzes können mit Hilfe der direkten Kinematik des implementierten Handmodells synthetisch generiert werden. Angaben hinsichtlich einer personenunabhängigen Nutzung des Systems finden sich nicht. Zur Evaluation werden einige Beispielbilder qualitativ ausgewertet. Da die Sichtbarkeit der Fingerspitzen auch aus der gewählten Perspektive senkrecht zur Handinnenfläche nicht gegeben ist, wenn die Finger stark gekrümmt oder ihre Spitzen vom Daumen verdeckt sind, kann dieses Verfahren somit nur eine Untermenge der Handformen rekonstruieren. Ferner sind Variationen der Handstellung prinzipiell nicht möglich.

Guan et al. [GCH99, GCH01, CGH02] stellen ein Verfahren zur Positurerkennung vor, das monokulare statische Einzelbilder verarbeitet. Die Spitzen von Fingern und Daumen, das Handgelenk sowie die Knöchel von Mittelfinger und Daumen werden farbig markiert, wobei die letzten drei Marker auf der Seite der Handinnenfläche angebracht sind. Durch einige vereinfachende Annahmen, die etwa die Unabhängigkeit der Gelenke voneinander einschränken, lassen sich die Gelenkstellungen aus den 2D-Koordinaten der acht Markierungen berechnen, solange diese komplett sichtbar sind. Somit sind Handformen mit moderater Beugung von Fingern und Daumen erkennbar, wenn die Handinnenfläche der Kamera zugewandt ist. Zu Beginn erfolgt eine Kalibrierung des Handmodells mittels einer Ansicht der Benutzerhand mit gestreckten Fingern. Die Evaluation des Ansatzes besteht aus sechs Realbildern und einer qualitativen Beurteilung des Ergebnisses. Informationen zu Beleuchtung, Bildauflösung, verwendeter Hardware und Geschwindigkeit werden nicht gegeben.

2.3.4 Systeme zur Positurrekonstruktion durch Modellanpassung

Einer der ersten Ansätze zur Rekonstruktion der Handpositur ist von **Rehg et al.** in [RK93, RK94, RK95, Reh95] vorgestellt. Zwei Kameras liefern Intensitätsbilder der Hand als Eingangsdaten. Zu Beginn muss die Hand eine vorgegebene Form annehmen. Die einzelnen Fingerglieder werden linienförmig approximiert. Solange sich die Finger einschließlich der Gelenke in den Eingangsbildern nicht überlappen, sind auf diese Weise 27 Freiheitsgrade messbar. Die Variation der Handform, insbesondere die Flexion der Finger, ist durch diese Randbedingung jedoch stark eingeschränkt. Eine alternative Version des Verfahrens verwendet zur Modellrepräsentation von Fingerspitzen und -gelenken lokale Templates, d. h. manuell in bei identischer Beleuchtung aufgenommenen Trainingsbildern markierte Bildausschnitte, die die entsprechenden Regionen zeigen. Diese werden gemäß der aktuellen Modellparameter in die Bildebenen der Kameras projiziert. Die resultierende Differenz zu den Eingangsbildern wird mittels Optimierung der Modellparameter in einem iterativen Prozess minimiert. Durch Über-

lagerung der Templates lassen sich Überlappungen einzelner Finger bei Einhaltung mehrerer Restriktionen behandeln, eine Anwendung dieser Strategie auf die gesamte Hand wird jedoch nicht gezeigt. Zur Evaluation dienen zwei Bildsequenzen. In der ersten Sequenz überlappen sich Fingerspitzen und -gelenke gemäß obiger Randbedingung nie, in der zweiten bewegt sich lediglich der Zeigefinger als vorderstes Element, alle anderen Gelenke bleiben unbewegt. Für die zweite Sequenz wurde die Anzahl der Freiheitsgrade auf neun reduziert, die die Konfiguration von Zeige- und Mittelfinger angeben. Eine personenunabhängige Anwendung ist nicht beschrieben.

Heap et al. [HH96a, HH96b] beschreiben eine der ersten auf PDMs (*point distribution model* [CTCG95]) basierenden Methoden zur Handformerkennung. Aus acht MRI-Aufnahmen (*magnetic resonance imaging*) einer Hand² wird ein dreidimensionales PDM generiert, das unter Berücksichtigung von Skalierung, Rotation und Translation die in einem monokularen Eingangsbild beobachtete Handansicht approximiert. Auf diese Weise sind Handformen erkennbar, die durch lineare Kombination aus den acht Trainingsansichten hervorgehen. Selbstverdeckungen können nicht behandelt werden. Das System leitet aus den PDM-Parametern keine parametrische Repräsentation der Handform wie etwa Gelenkstellungen ab.

Holden et al. [HOR99a] verwenden einen mit 13 ringförmigen Markern in sechs verschiedenen Farben gekennzeichneten Handschuh sowie ein Handmodell, das für eine gegebene Positur die Projektion der 2D-Koordinaten der Markierungen in die Bildebene ermöglicht. Anfangs nimmt die Hand eine vorgegebene Konfiguration ein. Diese passt das Verfahren mittels des in [Low91] beschriebenen iterativen Algorithmus an die Position der Markierungen in den folgenden Bildern an. Die Positionen nicht sichtbarer Marker werden aus dem aktuellen sowie den letzten fünf Parametersätzen mit abnehmender Gewichtung – ähnlich einem Kalmanfilter – extrapoliert. In Kombination mit einem adaptiven Fuzzy-Expertensystem, das in [HOR99b] beschrieben ist, wird das System zur Erkennung von 22 Gesten eingesetzt, die sich aus dem Übergang zwischen verschiedenen in der Australischen Gebärdensprache verwendeten Handformen ergeben. Die Kamera ist senkrecht auf die Handinnenfläche gerichtet und die Handstellung ist in 21 der Gesten konstant. Bei einer Vokabulargröße von 22 und jeweils einer Variation für Training und Test erreicht das Verfahren eine Erkennungsrate von 95% bzw. $\frac{21}{22}$. Angaben zum mittleren Fehler der Positurparameter finden sich in der Arbeit nicht. Die Extrapolation zur Verdeckungsbehandlung ist nach Aussage der Autoren nicht robust genug, um eine Änderung der Handstellung zu verarbeiten.

Das von **Ouhaddi et al. [OH99]** vorgestellte System verarbeitet monokulare Bildsequenzen der unmarkierten Hand vor monochromem Hintergrund. Nach einer Kalibrierung auf die Handgeometrie des Benutzers werden Form und Position aufgrund der Handkontur erkannt. Eine Hautfarbdetektion transformiert das Eingangsbild in eine binäre Maske, wobei der Arm des Benutzers ggf. manuell segmentiert und im Folgenden ignoriert wird. Durch Projektion des Handmodells entsteht ebenfalls eine binäre Maske, die mittels Simplex-Optimierung der Modellparameter an die erste Maske angenähert wird. Als Differenzmaß dient die Anzahl der

²Ein MRI-Scanner erzeugt dreidimensionale Aufnahmen des Körpers in Form einer Serie von 2D-Querschnitten. Im Unterschied zum Röntgenverfahren werden magnetische Felder zur Bildgebung verwendet.

in beiden Masken unterschiedlich gesetzten Pixel, normiert auf die Summe der gesetzten Pixel beider Masken, sowie die Chamfer-Distanz bezüglich der Randkurven.

Die gerichtete Chamfer-Distanz C_d zweier Punktmengen A und B ist nach [BTBW77] definiert zu

$$C_d(A, B) = \frac{1}{|A|} \sum_{\mathbf{a} \in A} \min_{\mathbf{b} \in B} \|\mathbf{a} - \mathbf{b}\| \quad (2.3)$$

Darin stehen $\mathbf{a}$ und $\mathbf{b}$ jeweils für ein 2D-Koordinatenpaar. Im Allgemeinen gilt $C_d(A, B) \neq C_d(B, A)$. Aus (2.3) ergibt sich die in A und B symmetrische ungerichtete Chamfer-Distanz C zu

$$C(A, B) = C_d(A, B) + C_d(B, A) \quad (2.4)$$

Das System gestattet keine Änderung der Handstellung, da die Aussagekraft der Kontur zu gering ist, um diese zu deuten. Die Autoren evaluieren den Ansatz qualitativ anhand einiger Realbildsequenzen, die dieser Voraussetzung genügen.

Einen monokularen Ansatz, der Handposituren in Ansichten der unmarkierten Hand rekonstruiert, stellen **Rosales et al. [RASS01]** vor. Jedes Einzelbild einer Bildsequenz wird ohne Berücksichtigung zeitlicher Zusammenhänge analysiert. Aus der Silhouette der Hand in der Bildebene, die als binäre Maske repräsentiert wird, ermittelt das System sieben Hu-Momente. Mit Hilfe eines synthetischen Handmodells werden diese Merkmale für zahlreiche Handkonfigurationen berechnet und zum Training von 30 neuronalen Netzen verwendet, die aus den Hu-Momenten die Positurparameter schätzen und jeweils auf einen Teil des Positurenraums spezialisiert sind. Die Autoren nennen diese Methode daher „Specialized Mappings Architecture“. Da bei der Verarbeitung realer Eingangsbilder die zu Grunde liegenden Positurparameter zunächst unbekannt sind, ist eine direkte Bestimmung des passenden Netzes nicht möglich. Stattdessen wird die Ausgabe sämtlicher Netze berechnet, so dass 30 Positurkandidaten resultieren. Diese bildet das System mit Hilfe des Handmodells und erneuter Merkmalsextraktion in den Merkmalsraum ab und selektiert diejenige Positur, deren Merkmale die geringste Mahalanobisdistanz zu den Eingangsmerkmalen aufweisen. Dieses Vorgehen beruht auf der Annahme, dass nur das geeignete Netz für einen gegebenen Merkmalsvektor $\mathbf{F}$ eine Positur $\tilde{\mathbf{P}}$ schätzt, deren Merkmale $\tilde{\mathbf{F}}$ wiederum $\mathbf{F}$ entsprechen. Das Training erfolgt mit 8000 synthetischen Posituren. Etwa 300.000 synthetische sowie einige Realbilder stellen die Testdatenmenge dar. Die Autoren messen einen mittleren quadratischen Fehler der Rekonstruktion „im Gelenkkonfigurationsraum“, geben jedoch weder eine Maßeinheit noch eine Berechnungsvorschrift an, so dass unklar bleibt, wie dieser Fehler zu interpretieren ist. Die Darstellung des Fehlers als Funktion des Ansichtswinkels zeigt eine deutliche Abhängigkeit. Weiterhin ergibt eine Analyse der Rekonstruktionsfehler, dass die betroffenen Posituren überwiegend aus der Amerikanischen Gebärdensprache stammen. Nähere Angaben zur Zusammensetzung der Test- und Trainingsdaten finden sich in der Arbeit nicht. Die Auswertung der Realansichten geschieht qualitativ. Rosales et al. nennen die geringe Aussagekraft der

Kontur und insbesondere der Hu-Momente als Fehlerursache. Eine Rekonstruktion beliebiger Posituren ist hiermit nach ihrer Aussage nicht möglich. Da Hu-Momente rotationsinvariant sind, kann eine Drehung der Hand in der Bildebene prinzipiell nicht erkannt werden.

Shimada et al. [SKSK00, SKS01, ISS04] präsentieren ein Verfahren, das aus einer monokular aufgenommenen Einzelbildsequenz eine Sequenz von Handformen und -stellungen extrahiert. Die Erkennung basiert dabei allein auf der Handkontur; eine Auswertung von Texturinformationen findet nicht statt. Mit Hilfe eines synthetischen 3D-Handmodells werden zunächst 1024 Handformen generiert und aus 84 verschiedenen Ansichtswinkeln betrachtet. Der Beugungsgrad jedes Fingers ist dabei mit vier Werten quantisiert ($4^5 = 1024$). Jede Kontur wird durch 256 äquidistante Landmarkenpunkte abgetastet. Mittels PCA (*principal components analysis*) reduzieren die Autoren die Dimensionalität dieses Merkmalsraums von 512 (x - und y -Koordinate jedes Landmarkenpunkts) auf 50. Innerhalb dieses 50-dimensionalen Eigenraums wird jede der $1024 \cdot 84 = 86016$ Konfigurationen durch ein PDM repräsentiert. Die nötigen Trainingsdaten für jedes PDM ergeben sich jeweils durch leichte Variation von Handform und Ansichtswinkel der entsprechenden Ausgangskonfiguration. Die PDMs erfordern eine Ausrichtung der Landmarkenpunkte der Kontur, die anhand der Extremwerte in Polarkoordinaten mit Ursprung im Flächenschwerpunkt erfolgt. Für eine gegebene Eingangskontur bestimmt das Verfahren auf Grundlage der Mahalanobisdistanz die am besten übereinstimmende Konfiguration.³ Im Offline-Betrieb umfasst der Suchraum hierbei alle 86016 PDMs, online berücksichtigt das System pro Einzelbild jeweils 2000 Nachbarn der fünf besten Kandidaten des letzten Bildes, d. h. 10000 Konfigurationen.

Die Verarbeitung geschieht auf einem Cluster aus 40 Pentium-III-CPUs mit 1 GHz Taktfrequenz und je 512 MB RAM und erreicht ca. 10,8 Hz für die Online-Variante. Die Testdatenmenge besteht aus sieben Sequenzen zu je 100 Einzelbildern. In 81,3% der Bilder ist die Positurerkennung „korrekt“, wobei die Autoren diese Qualität jedoch nicht näher definieren. Da aus vielen Ansichtswinkeln, insbesondere nahe der Handtellerebene, nur ein geringer Teil aller Handformen aufgrund ihrer Kontur unterscheidbar ist, hat die Verwendung dieses Merkmals eine signifikante Einschränkung des erkennbaren Positurenraums zur Folge. Beispielsweise haben eine Faust und ein ausgestreckter Zeigefinger aus der Richtung des Zeigefingers betrachtet dieselbe Kontur. Ferner können in einer seitlichen Ansicht der Hand aufgrund fehlender Tiefeninformationen die einzelnen Finger nicht unterschieden werden.

Die von **Stenger et al. in [SMC01a, SMC01b, Ste01]** beschriebene Methode kombiniert die Rekonstruktion von Handposition und -stellung mit der Erkennung einfacher Handformen, die sich mit nur einem Freiheitsgrad beschreiben lassen. Das Verfahren benötigt keine Marker. Es verarbeitet Bildsequenzen, in denen die Initialpositur vorgegeben ist. Ein synthetisches Handmodell in entsprechender Konfiguration wird auf Grundlage der Kontur an die Eingangsdaten angepasst. Diese Anpassung besteht in der Verschiebung einer Anzahl von Konturpixeln entlang der jeweiligen Normalen. Auf diese Weise entsteht ein Beobachtungsvektor, aus dem das

³Dieses Vorgehen ist in [ISS04] beschrieben. In [SKSK00, SKS01] verwenden die Autoren die Summe der quadrierten Abstände der Landmarkenpunkte, was jedoch weniger aussagekräftig ist. Die Anzahl der Handkonfigurationen ist mit ca. 16000 entsprechend kleiner.

System mit Hilfe eines UKF⁴ (*unscented Kalman filter*) die Handpositurparameter schätzt.

Das vorgestellte Verfahren erlaubt die Verwendung mehrerer Kameras, die Autoren machen jedoch keine Angaben hinsichtlich der erwarteten Vor- und Nachteile. Da die Aussagekraft der Kontur stark vom Blickwinkel abhängt, stellt sich die Frage nach einer Gewichtung der verschiedenen Ansichten, die die Arbeit jedoch nicht behandelt. Zur Evaluation dienen zwei monokular und eine binokular aufgenommene Realbildsequenz mit je 80 Bildern, in denen die Eingangspositur mit sechs bzw. sieben Freiheitsgraden variiert. Sechs Freiheitsgrade repräsentieren Position und Stellung, ein weiterer beschreibt die Beugung des Daumens. Die Beugung der vier Finger bleibt konstant und hat in den betrachteten Sequenzen eine hohe Varianz der Kontur zur Folge. Auf einem Rechner mit 433-MHz-Celeron-CPU beträgt die Verarbeitungsgeschwindigkeit 3 Hz.

In [STTC06] ersetzen Stenger et al. den UKF gegen einen probabilistischen Ansatz, der den Positurreaum hierarchisch partitioniert. Auf oberster Hierarchieebene erzeugt eine grobe Quantisierung eine Menge von Prototypen, die jeweils ein Teilvolumen repräsentieren. Jeder dieser Prototypen verfügt über Kinder, die das entsprechende Volumen feiner quantisieren. Mehrere Hierarchieebenen resultieren somit in einer Baumstruktur. Die Prototypen enthalten synthetisch generierte Handansichten, die für ein gegebenes Eingangsbild unter Verwendung der Chamfer-Distanz bezüglich Modellkontur und Kanten des Eingangsbildes (vgl. (2.3), (2.4)) sowie eines unimodalen parametrischen Hautfarbmodells probabilistisch bewertet werden. Diese Bewertung erfolgt zunächst auf oberster Ebene und wird nur für Prototypen mit guter Übereinstimmung verfeinert.

Dieses Vorgehen kombinieren die Autoren mit einer Modellierung der Dynamik der Handbewegung im Parameterraum durch einen Markov-Prozess. Zur Schätzung der Prozessparameter werden mit Hilfe eines Datenhandschuhs entsprechende Trainingsdaten erzeugt. Eine nähere Beschreibung der Trainingsdaten oder Angaben über die Natur der gemessenen zeitlichen Korrelation enthält die Arbeit nicht.

Neben einigen qualitativen Auswertungen umfasst die Evaluation des Systems drei Bildsequenzen mit fester Handform, in denen Stellung und Position variieren (sechs Freiheitsgrade), sowie eine weitere Sequenz, die eine Änderung der Handform mit zwei Freiheitsgraden zeigt. Bei einer Bildauflösung von 320×240 – die Hand belegt etwa 5–10% der Bildfläche – liegt der mittlere Pixelfehler hinsichtlich der Lokalisation einiger charakteristischer Punkte der Hand zwischen 6,6 und 9,3. Dieses Maß erfasst nicht die Abweichung senkrecht zur Bildebene, die in der verwendeten monokularen Ansicht jedoch vermutlich die Hauptkomponente darstellt. Ferner bleibt unklar, wie groß die Ähnlichkeit zwischen den Testsequenzen und den zur Parametrisierung des Markovprozesses benutzten Trainingsdaten ist. Die benötigte Rechenzeit pro Einzelbild liegt zwischen 2 Sekunden auf einem 1-GHz-Pentium-IV (konstante Handform) und 3 Sekunden auf einem 2,5-GHz-Pentium-IV (Handform mit zwei Freiheits-

⁴Der UKF ist eine nichtlineare Erweiterung des regulären Kalmanfilters. Er verwendet zur Prädiktion mehrere Punkte in der Nachbarschaft des aktuellen Zustands und umgeht damit die lokale Linearisierung, die der EKF (*extended Kalman filter*) darstellt. Auf diese Weise werden mehr Informationen ausgewertet und die Berechnung der Jacobi-Matrix vermieden.

graden).

Einen probabilistischen Ansatz beschreiben **Wu et al. [WLH01]**. Mit Hilfe eines Datenhandschuhs sammeln die Autoren zunächst etwa 30000 Gelenkkonfigurationen bei 20 Freiheitsgraden. Eine PCA reduziert die Dimensionalität dieser Daten auf sieben. Das Verfahren geht davon aus, dass alle Handformen im PCA-Raum auf einer Hypergeraden durch zwei von insgesamt 28 Grundhandformen liegen. In Verbindung mit der Annahme konstanter Handstellung ist der Positorenraum damit so weit komprimiert, dass ein Monte-Carlo-Ansatz realisierbar ist. Das System wertet Kantenmerkmale aus einer monokular aufgenommenen Eingangsbildsequenz sowie die Kontur eines einfachen Handmodells aus und bewertet anhand dieser Merkmale pro Einzelbild 100 Hypothesen. Die Auswahl dieser Hypothesen geschieht mittels *importance sampling* auf der Grundlage der Bewertung der Hypothesen des vorangegangenen Schritts. Die Evaluation des Verfahrens erfolgt qualitativ auf einer 100 Einzelbilder umfassenden Bildsequenz der unmarkierten Hand mit konstanter Stellung und Position. Die Kamera ist senkrecht auf die Handinnenfläche gerichtet. Angaben zur benötigten Rechenzeit oder zur Ähnlichkeit zwischen den für die PCA verwendeten Daten und der Testsequenz finden sich in der Arbeit nicht.

Vittrup et al. [VSM02, Vit02] untersuchen mehrere numerische Verfahren hinsichtlich ihrer Eignung zur Rekonstruktion von Handposituren in synthetisch generierten Stereo-Bildsequenzen. Ein Handmodell mit 9 Freiheitsgraden wird an die Eingangsdaten angepasst, wobei die Konturdifferenz als Optimierungskriterium dient. Diese Differenz berechnet sich wie in [OH99] durch die Erstellung je einer binären Maske für Eingangs- und Modellbild, Zählen der in beiden Masken unterschiedlich gesetzten Pixel und Division durch die Gesamtzahl der in beiden Masken gesetzten Pixel. Ausgehend von einer vorgegebenen Initialpositur werden die Modellparameter in jedem neuen Bild iterativ optimiert. Die Autoren vergleichen drei nachfolgend kurz skizzierte Optimierungsstrategien.

Der in [McC00] beschriebene Spider-Algorithmus fasst die Hand als kinematische Kette auf und betrachtet für jeden Freiheitsgrad den in der letzten Positur gewählten Parameter x sowie zwei Nachbarwerte $x \pm \Delta$, wobei die Schrittweite Δ in jeder Iteration abnimmt. Die Lösung wird mittels dynamischer Programmierung bestimmt. Das Downhill-Simplex-Verfahren stellt die zweite untersuchte Methode dar, die dritte ist eine Kombination aus Downhill Simplex und *simulated annealing* („simuliertes Ausglühen“), die durch gelegentliche Aufwärtsschritte (d. h. Verschlechterung der Modellanpassung) lokale Minima verlassen kann.

Zwölf synthetische Testsequenzen mit je 250 Einzelbildern stellen die Grundlage für die Evaluation der drei Optimierungsverfahren dar. Die Abweichung der Rekonstruktion von der Eingangspositur wird quantitativ berechnet als Summe der euklidischen Abstände der jeweiligen Gelenkpositionen im Raum. Finger- und Daumenspitzen, die die Enden der kinematischen Kette bilden und damit die größten Distanzen vermuten lassen, gehen in dieses Maß nicht ein. In sechs Testsequenzen sind Handstellung und -position konstant. Alle drei Verfahren erreichen unter dieser Randbedingung gute Ergebnisse, bei Variation von Stellung und Position ergeben sich jedoch mittlere Fehler von ca. 4 cm pro Gelenk. Als Ursache hierfür nennen die Autoren die geringe Aussagekraft der Handkontur bei Variation der Handstellung. Die

Verarbeitungsgeschwindigkeit liegt bei 40 Sekunden pro Einzelbild auf einem Pentium-III-Rechner mit 550 MHz Taktfrequenz. Die Konturberechnung geschieht dabei auf einer Matrox-Millennium-Grafikkarte.

Ein datenbankbasiertes System zur Erkennung beliebiger Handformen und -stellungen ohne Verwendung von Markern beschreiben **Athitsos et al. [AS03b, AS03a]**. Das Verfahren analysiert statische Einzelbilder und arbeitet sowohl bei perfekter als auch bei grober Segmentierung der Hand in Form einer *bounding box*. Zur Erstellung der Datenbank nimmt ein synthetisches Handmodell 26 verschiedene Formen an, die jeweils aus 86 Ansichtswinkeln betrachtet und in je 48 Schritten rotiert werden, so dass insgesamt 107.328 Ansichten entstehen. Jede Ansicht bildet zusammen mit drei verschiedenen Merkmalsgruppen, die sich aus einem Kantenbild ergeben, einen Datenbankeintrag. Die Erkennung geschieht durch Berechnung der entsprechenden Merkmale für das Eingangsbild und Suche nach einer möglichst guten Übereinstimmung in der Datenbank.

Aus Effizienzgründen erfolgt diese Suche in zwei Stufen. Zunächst wird für jeden Datenbankeintrag die gerichtete Chamfer-Distanz (vgl. (2.3) und (2.4)) seines offline berechneten Kantenbildes zum Kantenbild der realen Hand bestimmt. Hierzu erfolgt eine Distanztransformation des Eingangs-Kantenbildes, die für jedes Pixel die Entfernung zur nächsten Kante speichert. Dadurch ist das Ergebnis der min-Operation in (2.3) direkt abrufbar und die Berechnung entsprechend schnell. Für den umgekehrten Fall, d. h. die Bestimmung der gerichteten Chamfer-Distanz des Eingangskantenbildes zu jedem Datenbankeintrag, ist diese Methode nicht anwendbar, da die Speicherung der Distanztransformationen sämtlicher Datenbankeinträge mehrere GB benötigt. Daher benutzen die Autoren eine Lipschitz-Einbettung [Bou85], um die Berechnung nicht im hochdimensionalen Kantenraum (in dem jedes Kantenpixel einer Dimension entspricht), sondern in einem Raum niedrigerer Dimensionalität durchführen zu können.

Die Lipschitz-Einbettung eines Raumes $\mathbb{R}^n$ in einen Raum $\mathbb{R}^k$ wird definiert durch die Auswahl von k beliebigen, paarweise verschiedenen Punkten $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_k \in \mathbb{R}^n$, wobei im Allgemeinen $k \ll n$ gilt. Für jeden Punkt $\underline{x} \in \mathbb{R}^n$ ergibt sich seine Einbettung $\Lambda(\underline{x}) \in \mathbb{R}^k$ zu

$$\Lambda(\underline{x}) = \begin{bmatrix} \|\underline{x} - \underline{x}_1\| \\ \|\underline{x} - \underline{x}_2\| \\ \vdots \\ \|\underline{x} - \underline{x}_k\| \end{bmatrix} \quad (2.5)$$

Dabei steht $\|\cdot\|$ für die verwendete Distanzmetrik. Dem Verfahren liegt die Annahme zu Grunde, dass zwei nahe beieinander liegende Punkte etwa den gleichen Abstand zu einem beliebigen dritten Punkt haben. Somit sind Punkte, die in $\mathbb{R}^n$ einen geringen Abstand haben, auch in $\mathbb{R}^k$ benachbart (jedoch nicht notwendigerweise umgekehrt). Nachbarschaft im eingebetteten Raum $\mathbb{R}^k$ ist daher eine notwendige, jedoch nicht hinreichende Bedingung für Nachbarschaft im Ursprungsraum $\mathbb{R}^n$.

Die erste Stufe der Suche liefert einige tausend Kandidaten aus der Datenbank. Für diese

berechnet das System in der zweiten Stufe die exakte Chamfer-Distanz, ein Kantenorientierungshistogramm sowie ein probabilistisches Übereinstimmungsmaß hinsichtlich aus den Kantenbildern extrahierter Linien. Die Autoren evaluieren das Verfahren mit 250 realen Handansichten aus beliebigen Ansichtswinkeln, die jeweils eine der 26 in der Datenbank enthaltenen Handformen zeigen. Die Ansichtswinkel wurden für diese Testdaten manuell geschätzt. Eine Positur gilt als korrekt erkannt, wenn die Handform übereinstimmt und die drei Ansichtswinkel jeweils nicht mehr als 30° abweichen. Damit existieren im Mittel für jede Positur 30,8 korrekte Datenbankeinträge. Bei perfekter Segmentierung der Hand im Eingangsbild wird eine Erkennungsrate von 13,6% erreicht, bei Bounding-Box-Segmentierung und Störobjekten im Hintergrund sinkt dieser Wert auf 8,0%. Die Verarbeitungsgeschwindigkeit auf einem 1,2-GHz-PC mit 300 MB RAM beträgt 15 Sekunden.

Zur Bewertung der gemessenen Rekonstruktionsgenauigkeit sind die gewählten Randbedingungen zu berücksichtigen. Das System analysiert Einzelbilder und kann daher keine zeitliche Korrelation nutzen. Ferner erfolgt der Test personenunabhängig und die Merkmale sind vertauscht. Bei vergleichbaren Veröffentlichungen ohne quantitative Evaluation ist daher nicht von höherer Leistung auszugehen.

In [AS01, AS02a, AS02b] präsentieren die Autoren einen Vorgänger dieses Systems, der anstelle der Linienmerkmale die Positionen der Fingerspitzen sowie sieben zentrale und sieben Hu-Momente verwendet. Diese Methode ist nur bei perfekter Segmentierung anwendbar und liefert für 276 Testbilder vier verschiedener Personen 21,7% Erkennungsrate bei 3–4 Sekunden Rechenzeit auf einem 1,2-GHz-PC mit 100 MB RAM.

Das Verfahren stellt eine interessante Anwendung einer Positurrendatenbank zur Positurrekonstruktion dar und verdeutlicht den Ressourcenbedarf, der in diesem Kontext aus der Verwendung hochdimensionaler Merkmale resultiert. Dem Einsatz zur Gebärdenerfassung stehen die hohe Rechenzeit sowie die geringe Anzahl erkannter Handformen entgegen. Bereits bei einer binären Quantisierung des Beugungswinkels jedes Fingers und des Daumens ergeben sich $2^5 = 32$ Formen. Diese Menge ist jedoch für die Repräsentation von Gebärden deutlich zu klein.

Fillbrandt et al. [Fil03, FAK03] stellen ein System zur Positurerkennung vor, das monokulare Bildsequenzen der unmarkierten Hand verarbeitet. Eine Menge von 15 in der Gebärdensprache häufig benutzten Handformen wird aus verschiedenen Blickwinkeln rekonstruiert, wobei der lokale Bildhintergrund entweder nicht hautfarben oder die Hand perfekt segmentiert sein muss. Kern des Verfahrens ist ein Netzwerk aus ca. 60 Handmodellen, die jeweils eine Handform und einen Blickwinkel repräsentieren. Hierfür werden AAMs (*active appearance model* [CTCG95]) verwendet, die die Silhouette sowie einen robusten lokalen Texturdeskriptor umfassen. Die Trainingsdaten zur Modellgenerierung bestehen aus Ansichten der Hand einer Person, wobei Handform und Ansichtswinkel manuell eingegeben wurden. Jedes Modell deckt dabei ausreichend Variationen ab, um einen Übergang zu benachbarten Modellen, d. h. Modellen mit geringfügig abweichender Handform oder Ansichtswinkel, zu ermöglichen. Diese Übergänge sind Bestandteil des Modellnetzwerks. Im Unterschied zu der in [CTCG95] beschriebenen Adaptionismethode geht in die Fehlerfunktion auch die Nachbar-

schaft der Kontur ein, um Anpassungsfehler – etwa durch zu kleine Skalierung des Modells – zu vermeiden.

Eine Evaluation des Ansatzes auf einigen Realbildsequenzen der trainierten Hand ergibt hinsichtlich des Fingerbeugungsgrades, der pro Finger durch einen skalaren Wert zwischen 0 und 1 angegeben wird, einen visuell geschätzten mittleren Fehler von 0,1. Der gleichermaßen geschätzte Fehler des rekonstruierten Ansichtswinkels beträgt 10–20°. Eine personenunabhängige Anwendung ist möglich, hat jedoch größere Rekonstruktionsfehler zur Folge. Die Verarbeitungsgeschwindigkeit auf einem 1-GHz-Pentium-III-PC ist 4 Hz. Das Verfahren ist für den Einsatz in der Gebärdenspracherkennung konzipiert, im Kontext eines Tutors reicht die Erkennung von 15 verschiedenen Handformen jedoch nicht aus. Ferner hängt die Menge der extrahierbaren Texturinformationen von der Anzahl der Pixel der Hand in der Bildebene und damit von der Handpositur ab. Ansichten aus Richtung der Finger oder des Handgelenks, die kaum aussagekräftige Textur aufweisen, sind im Modellnetzwerk des Systems nicht enthalten.

Lu et al. [LMSO03] beschreiben einen Ansatz, der unter verschiedenen Randbedingungen Handposituren in einer monokular aufgezeichneten Bildsequenz der unmarkierten Hand erkennen kann. Nach der Initialisierung mit einer vorgegebenen Startpositur, in der Finger und Daumen maximal gestreckt und abgespreizt sind, extrahiert das Verfahren Kontur- und Texturmerkmale sowie einige charakteristische Randkurvenpunkte und bestimmt deren Ableitung in Form der Differenz zu vorhergehenden Einzelbildern. Aus dieser Differenz wird die zu Grunde liegende Änderung der Modellparameter nach einer in [Met97] beschriebenen Methode, die die partielle Ableitung eines Bildpunktes nach den Modellparametern berechnet, geschätzt. Die projektionsbedingte Mehrdeutigkeit der Bildinformationen führt zu der Einschränkung, dass die Bewegung kontinuierlich sein muss. Verdeckungen der Finger sind zugelassen, die Autoren formulieren jedoch die Bedingung, dass die Fingerstellung währenddessen annähernd konstant bleibt. Parallel zur Positurrekonstruktion führt das System in den ersten 10–20 Einzelbildern eine Anpassung der Modellgeometrie an die Anatomie des aktuellen Nutzers durch, die auf dem Residuum der Modellanpassung beruht. Die Evaluation des Ansatzes besteht aus zwei Realbildsequenzen und bewertet die Modellanpassung qualitativ. Auf einem Pentium-IV-Prozessor mit 1 GHz Taktfrequenz erreicht die Implementierung eine Verarbeitungsgeschwindigkeit von 4 Hz.

In einer Weiterentwicklung des Ansatzes durch **Metaxas et al. [MT05]** ist die Annahme kontinuierlicher Bewegung in die Modellierung der Positurdynamik integriert. Die Autoren generieren mit einem synthetischen Handmodell 200 Sequenzen aus 29 Ansichtswinkeln, die jeweils 15 Einzelbilder enthalten. Diese Datenbank wird in 108 Cluster aufgeteilt, die im Mittel je ca. 54 Sequenzen repräsentieren. Zunächst erfolgt die Positurrekonstruktion wie in [LMSO03], d. h. mit einer vorgegebenen Initialpositur und ohne Zugriff auf die Datenbank. Die Autoren geben an, dass eine deutliche Flexion der Finger oder die Rotation der Hand um eine Achse in der Kameraebene bei diesem Verfahren zu Fehlern führen. Sie ermitteln daher offline empirisch einen Schwellwert für den Residualfehler der Positur und reinitialisieren das Modell bei dessen Überschreitung. Zum Zweck dieser Reinitialisierung identifiziert das

System zu jedem Zeitpunkt denjenigen Datenbankcluster, der bezüglich des Ansichtswinkels im aktuellen Einzelbild $\mathbf{I}(t)$ sowie der Anzahl ausgestreckter Finger in $\mathbf{I}(t)$ und $\mathbf{I}(t - 15)$ am besten zu den letzten 15 Eingangsbildern passt. Die Finger werden dabei auf Grundlage der Kontur mit einer einfachen Heuristik detektiert. Diesen Cluster durchsucht das System nach einer Sequenz, die mit den letzten 15 Positurschätzungen übereinstimmt. Die Suche berücksichtigt die Kontur sowie ein Orientierungshistogramm jeweils in allen 15 Einzelbildern. Um die damit verbundene Datenmenge zu reduzieren, bildet das System beide Deskriptoren auf je 15 skalare Werte ab.

Die Reinitialisierung mit einer Sequenz anstelle einer einzelnen Positur gestattet es, unmittelbar danach zu dem in [LMSO03] beschriebenen Rekonstruktionsverfahren zurückzuwechseln, da die benötigten Ableitungen für die Datenbanksequenz sofort berechenbar sind. In die Auswahl des Clusters, dem die Reinitialisierungssequenz entnommen wird, gehen nur die zwei diskreten Zeitpunkte t und $t - 15$ ein. Die Festlegung des Schwellwerts zur Einleitung einer Reinitialisierung stellt ein Signalentdeckungsproblem dar; Angaben zu der aus dem gewählten Schwellwert resultierenden Fehlerrate finden sich in der Arbeit nicht. Die Autoren evaluieren den Ansatz qualitativ mit drei Realbildsequenzen. Die 200 Datenbanksequenzen, insbesondere deren Ähnlichkeit zu diesen Testsequenzen, sind nicht näher beschrieben. Daher bleibt unklar, welche Handformen und -stellungen das System verarbeiten kann, die Datenbankgröße lässt jedoch eine unvollständige Abdeckung des Positurenraums vermuten.

2.3.5 Bewertung der Ansätze

In Tabelle 2.3 sind die vorgestellten Verfahren bezüglich ihrer Randbedingungen und des Rekonstruktionsumfangs miteinander verglichen. Zusätzlich finden sich die entsprechenden Angaben für das in der vorliegenden Arbeit entwickelte System.

Keiner der bisherigen Ansätze kann die Ausdrucksmittel der Gebärdensprache vollständig rekonstruieren. Viele Arbeiten gehen von einer konstanten und die Rekonstruktion vereinfachenden Handstellung aus, in der die Beugung der einzelnen Finger auf Basis der gewählten Merkmale deutlich erkennbar ist. Daneben wird häufig der Positurenraum eingeschränkt, um Mehrdeutigkeiten zu vermeiden.

Eine robuste Strategie zur Mehrdeutigkeitsauflösung und Fehlerkorrektur ist bisher nicht verfügbar. Die bestehenden Arbeiten setzen voraus, dass die Rekonstruktion in jedem Einzelbild zumindest annähernd korrekt oder die tatsächliche Positur in einer kleinen Kandidatenmenge enthalten ist.

Weiterhin stellen Art und Umfang der Evaluation ein signifikantes Defizit dar. Die Ursache hierfür liegt in der Schwierigkeit, geeignete Referenzdaten zu erzeugen, d. h. die korrekten Positurparameter in jedem Einzelbild manuell einzugeben. Die Mehrzahl der Autoren verzichtet hierauf und führt lediglich eine exemplarische Evaluation mit kleiner Testdatenmenge und qualitativer Beurteilung des Ergebnisses durch. Die tatsächliche Leistungsfähigkeit der präsentierten Ansätze lässt sich auf dieser Grundlage nicht bewerten.

Tabelle 2.3: Übersicht der vorgestellten Systeme zur Handpositurrekonstruktion. In der letzten Zeile ist zum Vergleich das in dieser Arbeit entwickelte System angegeben.

		Eingangsdaten			Funktionsweise				Rekonstruktion			Leistung	
Jahr		Kameras	korrelierte Bildsequenz	unkontrollierter Hintergrund	Algorithmus	Merkmale	automatische Initialisierung	Hypothesen	Handform	Handstellung	Handposition	Geschwindigkeit in Hz	quantitative Evaluation
Lee	'95	2	–	–	I/A	M	•	1	○	○	•	–	–
Lathuilière	'00	1	–	–	I	M	•	1	○	○	•	15	•
Nölker	'00	2	–	–	I	T	•	1	○	–	–	1,5	–
Guan	'02	1	–	–	I	M	•	1	○	○	–	–	–
Rehg	'95	2	•	–	A	K/T	–	1	○	○	○	6,7	–
Heap	'96	1	•	–	A	T	–	1	○	○	•	10	–
Holden	'99	1	•	–	A	M	–	1	•	–	–	–	–
Ouhaddi	'99	1	•	–	A	K	•	1	•	–	•	–	–
Rosales	'01	1	–	–	A	K	•	1	•	○	•	–	○
Shimada	'04	1	•	–	A	K	•	5	•	•	–	10,8	○
Stenger	'01	1	•	•	A	K+T	–	1	○	•	•	3	–
Stenger	'06	1	•	•	A	K+T	•	1	○	○	•	0,5	•
Wu	'01	1	•	–	A	K+T	•	1	•	–	–	–	–
Vittrup	'02	2	•	–	A	K	–	1	•	○	○	0,001	•
Athitsos	'03	1	–	•	A	T	•	1	○	•	–	0,07	•
Fillbrandt	'03	1	•	•	A	K+T	–	1	○	○	•	4	○
Lu	'03	1	•	–	A	T	•	1	○	○	•	4	–
Metaxas	'05	1	•	–	A	T	•	1	○	○	•	–	–
Zieren	'07	1	•	•	A	M	•	5	•	•	•	2,94	•

•: ja ○: teilweise/mit Einschränkungen –: nein/nicht verfügbar

I: Inverse Kinematik A: Modellanpassung

M: optische Marker K: Kontur T: Textur

Einige Arbeiten geben die Rekonstruktionsleistung quantitativ an, allerdings sind die verwendeten Testdaten entweder nicht spezifiziert [Fil03, FAK03, RASS01, SKSK00, SKS01, ISS04] oder umfassen nur einige wenige Posituren [LH99, LH00, STTC06, VSM02, Vit02].

Diese Messungen haben daher kaum Aussagekraft.

Zur Quantifizierung der Systemleistung kommen teilweise Maße zum Einsatz, die die Rekonstruktionsgenauigkeit nur unzureichend beschreiben und damit für eine Evaluation nicht geeignet sind. Während [STTC06] den Rekonstruktionsfehler in Pixeln angibt, fehlt in [RASS01] und [ISS04] jegliche Definition des verwendeten Abweichungsmaßes. Rückschlüsse auf die Genauigkeit der Rekonstruktion der Fingerbeugungs- und Ansichtswinkel sind hier nicht möglich.

Lediglich in zwei der vorgestellten Arbeiten ist die empirische Evaluation anschaulich und nachvollziehbar. Vittrup et al. [VSM02, Vit02] berechnen die mittlere Abweichung der Gelenkpositionen im Raum bei variabler Handstellung und -position zu ca. 4 cm. Dies zeigt deutlich die Defizite des präsentierten Ansatzes, die primär in der geringen Trennschärfe der verwendeten Merkmale liegen (s. Seite 38).

Das von Athitsos et al. [AS01, AS02a, AS02b, AS03b, AS03a] definierte binäre Trefferkriterium, das die Abweichung hinsichtlich Handform und -stellung beschränkt (vgl. S. 39), ermöglicht die Angabe einer skalaren Erkennungsrate, enthält jedoch erheblich weniger Informationen als etwa ein Histogramm der gemessenen Abweichungen. Obwohl die unter Optimalbedingungen erreichte Leistung von 13,6% in Anbetracht der Eingangsdaten (unkorrigierte Einzelbilder der unmarkierten Hand) aus technischer Sicht beachtlich ist, erlaubt sie keinen praktischen Einsatz des Verfahrens.

Für einen Gebärdensprachtutor ist eine robuste Erkennung aller manuellen Ausdrucksmittel erforderlich. Die Handpositur darf dabei in keiner Weise eingeschränkt werden. Durch Verfolgen multipler Hypothesen sind Mehrdeutigkeiten aufzulösen und Rekonstruktionsfehler zu korrigieren. Eine automatische Initialisierung ist hierfür Voraussetzung. Keine der existierenden Methoden erfüllt diese Anforderungen.

2.4 Leistungsfähigkeit existierender Tutorsysteme für Gebärdensprache

Bisher funktionieren kommerziell erhältliche elektronische Lernmaterialien wie [Kes02] noch unidirektional und geben dem Anwender kein Feedback bezüglich der von ihm ausgeführten Gebärde. Obwohl die Erkennung von Handposituren seit etwa 30 Jahren Gegenstand der Forschung ist⁵, wird die Anwendung der entwickelten Verfahren im Gebärdensprachunterricht erst seit wenigen Jahren untersucht. Methoden der Lautspracherkennung werden hingegen bereits in entsprechender Lernsoftware eingesetzt, um die Aussprache des Benutzers zu verbessern [Lan04].

Als Hauptursache für dieses Defizit kommen die hohen Kosten eines Datenhandschuhs sowie die nicht ausreichende Leistung bestehender bildbasierter Methoden zur Handposituren-erken-

⁵Der erste Datenhandschuh – nach einem seiner Erfinder „Sayre Glove“ genannt – wurde 1977 an der Universität von Illinois entwickelt.

nung in Betracht. In der Literatur finden sich drei Arbeiten zu interaktiven Gebärdensprach-tutorsystemen, von denen jedoch keine das Problem der videobasierten Positurrekonstruktion behandelt. Dennoch sollen diese Arbeiten im Folgenden kurz vorgestellt werden, um einen Einblick in alternative Ansätze zu geben.

Sagawa et al. [ST02] beschreiben einen interaktiven Tutor für Japanische Gebärdensprache, der zu lernende Gebärdensätze mittels eines Avatars präsentiert. Dies ermöglicht eine platzsparende Codierung und gestattet es dem Lernenden, die Gebärden aus beliebigem Blickwinkel zu betrachten. Zur Erfassung der Benutzerausführung kommen Datenhandschuhe zum Einsatz, die die Gelenkstellungen sowie die Position der Hand im Raum messen. Die so gewonnenen Merkmale werden mit denen der Referenzausführung zeitlich synchronisiert und verglichen. Angaben über die hierzu verwendeten Algorithmen machen die Autoren nicht. Das Feedback besteht aus einem skalaren Ähnlichkeitswert pro Gebärde sowie einer simultanen Wiedergabe von Benutzer- und Referenzausführung mit Hilfe des Avatars (eine Videoaufnahme des Benutzers steht nicht zur Verfügung). Informationen hinsichtlich der Art eventueller Fehler erhält der Anwender nicht. Das System unterscheidet keine Phasen innerhalb einer Gebärde.

Einen videobasierten Gebärdentutor stellen **Aran et al. [AKA05]** vor. Der Benutzer trägt einen einfarbigen Handschuh und wird von zwei Kameras aufgezeichnet, die den Gebärdenraum erfassen. Als Referenzpunkt dient die Gesichtsposition, die in einer Initialisierungsphase durch Bewegen der Hand vor dem Gesicht bestimmt und im weiteren Verlauf als konstant angenommen wird. Die Handposition im Raum misst das Verfahren mit drei Freiheitsgraden. Verschiedene Formen und Stellungen werden mittels eines nicht näher beschriebenen regelbasierten Verfahrens aufgrund geometrischer Konturmerkmale differenziert. Das Vokabular besteht aus sieben Gebärden in Türkischer Gebärdensprache, die jeweils durch ein entsprechend trainiertes HMM (*Hidden Markov Model*) repräsentiert sind. Hierbei berücksichtigen die Autoren ausschließlich die Trajektorie der Hand. Form und Stellung gehen nicht in die HMMs ein. Zur Feedbackgenerierung berechnet das System die Emissionswahrscheinlichkeit für jedes HMM. Eine Gebärde gilt genau dann als korrekt ausgeführt, wenn das entsprechende HMM die höchste Emissionswahrscheinlichkeit liefert. Diese Funktionalität entspricht einem Gebärdenspracherkenner.

Das Feedback des Systems ist binär und berücksichtigt lediglich den Verlauf der Handposition über die Zeit. Eine positive Rückmeldung bedeutet nicht notwendigerweise, dass die Trajektorie mit der Referenz übereinstimmt, sondern lediglich, dass sie ihr ähnlicher ist als den übrigen sechs Vokabularelementen. Hier wird deutlich, dass eine vollständigen Positurrekonstruktion erforderlich ist, um über eine binäre Bewertung hinaus korrekatives Feedback zu geben.

Das von **Henderson et al. [HLB⁺05, LHH⁺05]** beschriebene Lernsystem befindet sich gegenwärtig in der Entwicklung. Kinder im Alter von 6–8 Jahren sollen in einem einfachen Spiel durch vorgegebene Gebärdensprachäußerungen mit einer animierten Katze interagieren. Das geplante Vokabular umfasst knapp zehn Gebärdensätze. Bisher existiert lediglich die Interaktionskomponente des Systems, die Komponente zur Gebärdenerfassung ist noch nicht

implementiert. Um erste Experimente hinsichtlich Motivation und Akzeptanz durchführen zu können, verwenden die Autoren ein „Wizard of Oz“-Setup: Die Probanden glauben, mit dem Computer zu interagieren, tatsächlich wird dieser jedoch von einem verborgenen Menschen gesteuert.

2.5 Konzept des interaktiven Gebärdensprachtutors

In diesem Abschnitt wird das Konzept des interaktiven Gebärdensprachtutors entwickelt. Der Schwerpunkt liegt dabei auf der Positurrekonstruktion, die die komplexeste Komponente des Systems darstellt. Da die im letzten Abschnitt vorgestellten Verfahren die Ausdrucksmittel der Gebärdensprache nur eingeschränkt rekonstruieren und somit im Kontext eines Tutors nicht anwendbar sind, entsteht ein neuer Ansatz, der den Schwierigkeiten der Aufgabenstellung mit besser geeigneten Strategien begegnet.

Das resultierende Verfahren zur Handpositurerkennung zeichnet sich durch die Erfassung aller in der Gebärdensprache verwendeten manuellen Ausdrucksmittel aus und schränkt den Positurenraum in keiner Weise ein. Dadurch ist eine universelle Verwendung auch außerhalb des Tutors, etwa im Bereich der Virtuellen Realität, möglich. Eine Vorgängerversion dieser Rekonstruktionsstufe wurde in [DZK06] präsentiert.

2.5.1 Eingangsbilder

Hinsichtlich der Akquisition der Eingangsdaten stellt sich zunächst die Frage nach der Anzahl zu verwendender Kameras. Um eine Verarbeitung mit preiswerter und weit verbreiteter Standardhardware zu ermöglichen, soll die Aufnahme des Benutzers monokular erfolgen. Die Hardwareanforderungen des Systems beschränken sich somit auf eine reguläre Farbbildkamera, die in vielen PC-Systemen bereits vorhanden und ggf. kostengünstig nachzurüsten ist. Gegenüber *Stereo-Vision*-Ansätzen entfällt damit ferner eine Kalibrierung oder zeitliche Synchronisation der Kameras.

Der Bildausschnitt muss den gesamten Gebärdenraum einschließen, um die in der Gebärdensprache auftretenden Handbewegungen sowie das als Referenzpunkt benötigte Gesicht des Anwenders zu erfassen. Dadurch steht für die Hand lediglich ein geringer Teil der Bildauflösung zur Verfügung. Die eingesetzten Algorithmen sollen prinzipiell verschiedene Auflösungen unterstützen, als grober Richtwert können jedoch 100×100 Pixel für die bounding box der Hand angenommen werden. Abbildung 2.13 zeigt ein entsprechendes Beispiel.

Die Bildrate der Kamera muss ausreichend hoch sein, um auch schnelle Bewegungen angemessen aufzulösen und so eine ausreichende Menge an Informationen für die Rekonstruktion zur Verfügung zu stellen. In dieser Arbeit werden 25 Hz als Untergrenze angenommen, da nahezu alle aktuellen Kameras dies unterstützen. Analog zur Auflösung stellt dieser Wert keine Konstante dar, sondern kann zur Steigerung der Genauigkeit erhöht werden, ohne dass eine



Abbildung 2.13: Beispiel der für die Hand verfügbaren Bildauflösung. Die Auflösung des Bildausschnitts beträgt 100×100 Pixel. Für die Positurrekonstruktion wird die Hand mit optischen Markern versehen (s. Abschnitt 2.5.3).

Modifikation der verwendeten Algorithmen erforderlich ist.

2.5.2 Hintergrundsubtraktion

Um einen Einsatz unter realistischen Bedingungen zu ermöglichen, ist die Unabhängigkeit vom Bildhintergrund wünschenswert. Im Vergleich zu den in der Literatur überwiegenden Ansätzen, die Aufnahmen unter Laborbedingungen erfordern, erhöht sich dadurch der Nutzungskomfort für den Anwender. Daher werden auf Pixelebene Algorithmen zur Hintergrundmodellierung eingesetzt, die entsprechende Bildbereiche identifizieren und von der weiteren Verarbeitung ausschließen. Da diese Verfahren in der Praxis oft nicht vollkommen fehlerfrei arbeiten, ist es trotzdem erforderlich, Störobjekte auch in nachfolgenden Verarbeitungsstufen, d. h. auf höherer Abstraktionsebene, zu erkennen. Dies geschieht unter Verwendung von Domänenwissen, das anatomische Grundlagen sowie den Ablauf des Gebärdensprozesses umfasst, und ist in Abschnitt 2.5.4 beschrieben.

2.5.3 Verwendete Merkmale

Die aus den Eingangsbildern extrahierten 2D-Merkmale haben auf Struktur und Leistung des Systems entscheidenden Einfluss. Eine der wichtigsten Designentscheidungen ist daher die Auswahl geeigneter Merkmale, die eine möglichst robuste und genaue Positurerkennung ermöglichen. Hinsichtlich der Aussagekraft der Merkmale ergeben sich aus der Aufgabenstellung hohe Anforderungen. Die Rekonstruktion der in der Gebärdensprache verwendeten manuellen Ausdrucksmittel umfasst die Bestimmung der Beugungswinkel jedes einzelnen Fingergelenks zur Beschreibung der Handform sowie die Ermittlung der Ansichtswinkel zur Angabe der Handstellung.

Die Merkmalsauswahl muss die Beschaffenheit der Eingangsbilder bei den unter Realbedingungen auftretenden Störfaktoren berücksichtigen. In der Praxis stellen neben der endlichen Auflösung das Kamerarauschen sowie Bewegungsunschärfe wichtige Faktoren dar, die die Qualität der Eingangsdaten beeinflussen. Bei typischer Bürobeleuchtung erreichen einfache Bildsensoren, wie sie etwa in Webcams verwendet werden, die zur Vermeidung von Bewegungsunschärfe benötigten Verschlussgeschwindigkeiten nur durch elektronische Signalverstärkung. Dies führt jedoch zu erhöhtem Rauschen. Üblicherweise ist somit mindestens einer der beiden Störfaktoren bei Einsatz preiswerter Kameras stark ausgeprägt.

Da das Tutorsystem gerade mit einfacher Kamerahardware verwendbar sein soll, besteht eine wichtige Forderung darin, dass die Erkennungsleistung mit abnehmender Bildqualität nicht abrupt sinkt. Dies entspricht dem Wunsch nach Robustheit hinsichtlich der Qualität der Eingangsdaten. Zur Positurrekonstruktion sind daher Merkmale zu wählen, die von den beschriebenen degenerativen Einflüssen möglichst unabhängig sind.

Vor diesem Hintergrund erfolgt nun eine kurze Bewertung häufig verwendeter Merkmale sowie die Motivation der getroffenen Merkmalsauswahl.

Kontur der Hand

Aus der Handkontur ergeben sich mehrere Formmerkmale wie etwa Umfang und Flächeninhalt. Diese Merkmale sind robust, da die Hand – etwa im Vergleich zu einzelnen Texturerelementen – relativ groß und die Kontur damit gegenüber vielen degenerativen Einflüssen wenig anfällig ist. Die Kontur der gesamten Hand ist jedoch nicht aussagekräftig genug, um verschiedene Handposituren zuverlässig zu unterscheiden. Abbildung 2.14 zeigt zur Veranschaulichung drei stark verschiedene Handposituren, deren Konturdifferenzen im Bereich des unter Realbedingungen anzutreffenden Rauschens liegen und somit keine Differenzierung ermöglichen. Auch die in Abschnitt 2.3 betrachteten Arbeiten zeigen, dass die Handkontur lediglich bei bestimmten, konstanten Ansichtswinkeln – etwa senkrecht auf die Handinnenfläche – Rückschlüsse auf die Handform erlaubt.

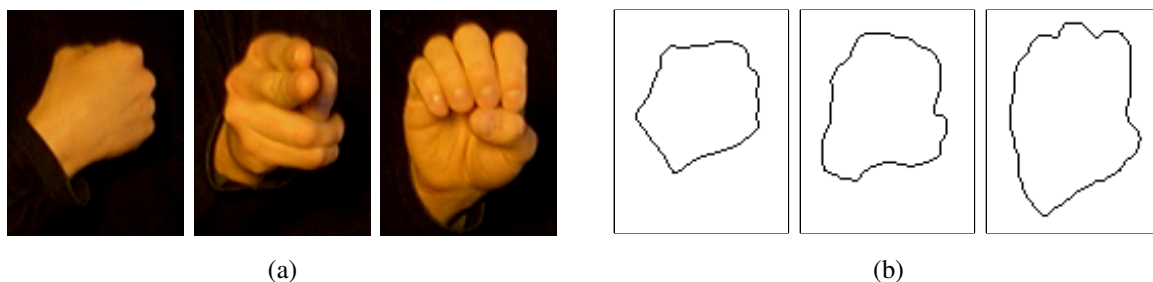


Abbildung 2.14: Drei stark verschiedene Handposituren (a), die die geringe Aussagekraft der Kontur (b) veranschaulichen.

Textur der Hand

Bei der in Abschnitt 2.5.1 beschriebenen Auflösung sind Texturmerkmale wie beispielsweise Kanten gegenüber den oben genannten Einflüssen – endliche Auflösung, Kamerarauschen und Bewegungsunschärfe – nicht robust, da sich ihre Abmessungen oft in der Größenordnung weniger Pixel bewegen. Abbildung 2.15 zeigt den Verlust von Texturinformation durch Bewegungsunschärfe am Beispiel einer Handpositur, die in Ruhestellung vergleichsweise stark ausgeprägte Kanten zwischen den ausgestreckten Fingern aufweist. Die Aufnahme erfolgte mit einer Kamera vom Typ Panasonic AG-DVX100B⁶ bei einer Verschlusszeit von 1/120 s. Beide Einzelbilder sind Teil einer Gebärdensequenz. Trotz der hohen Verschlussgeschwindigkeit enthält das Bild der bewegten Hand kaum Textur.

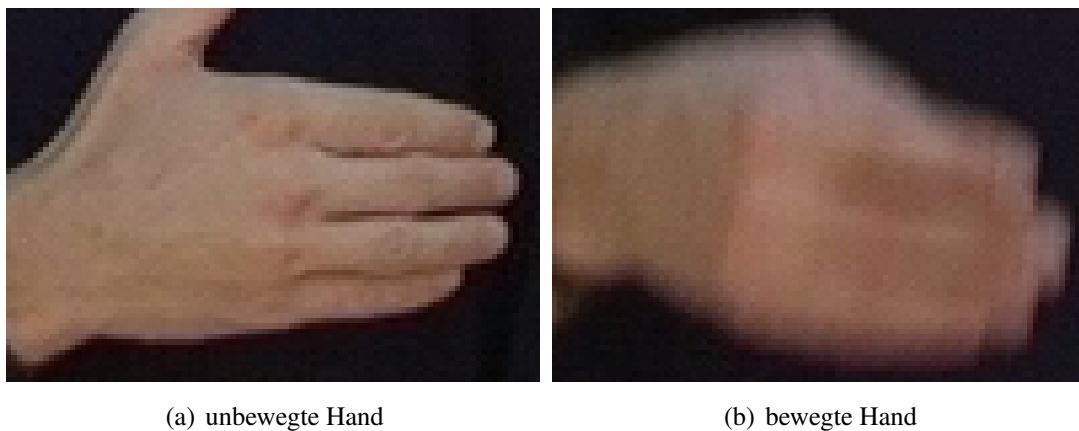


Abbildung 2.15: Auswirkung von Bewegungsunschärfe auf die unmarkierte Hand. Die Ausschnittauf Auflösung beträgt 100×70 Pixel, die Verschlusszeit 1/120 s.

Optische Marker

Die Verwendung optischer Marker erlaubt die robuste Extraktion aussagekräftiger Merkmale, die von den beschriebenen Störeinflüssen nur wenig beeinträchtigt werden. Durch Kennzeichnung von charakteristischen Punkten oder Teilen der Hand in verschiedenen Farben lassen sich in jedem Einzelbild erheblich mehr und gleichzeitig zuverlässigere Merkmale bestimmen. Nach einer farbbasierten Segmentierung können für jeden sichtbaren Marker geometrische Konturmerkmale wie etwa Position, Flächeninhalt und Hauptachsenorientierung berechnet werden. Falls ein Marker nicht sichtbar ist, stellt diese Tatsache ebenfalls eine wichtige Information dar, die die möglichen Posituren deutlich einschränkt. Insgesamt steht somit für die Positurrekonstruktion eine größere Menge an Informationen zur Verfügung, was angesichts der Schwierigkeit der Aufgabe sinnvoll ist.

In der vorliegenden Arbeit wird ein farbig markierter Stoffhandschuh verwendet, der die Bewegung von Hand und Fingern nicht beeinträchtigt und den Bedienkomfort daher kaum re-

⁶Diese Kamera wird als professionelle Ausrüstung für einen Preis von z. Z. etwa € 3500 angeboten.



Abbildung 2.16: Beispiel eines farbig markierten Stoffhandschuhs zur Verwendung mit dem Gebärdensprachtutor. Die Markergeometrie kann an die Aufgabenstellung angepasst werden.

duziert. Im Unterschied zu bisherigen Veröffentlichungen sind die zur Positurerkennung eingesetzten Algorithmen nicht auf eine bestimmte Markergeometrie festgelegt, sondern erlauben beliebige Anordnungen. Somit ist eine Optimierung für verschiedene Aufgabenstellungen möglich. Treten beispielsweise bestimmte Handformen und -stellungen häufig auf, können die Marker so angeordnet werden, dass ihre Merkmale in diesen Konfigurationen möglichst trennscharf sind.

Für den Einsatz im Kontext der Gebärdensprache ist die Markierung aller Finger und des Daumens erforderlich, da sich zahlreiche häufig verwendete Handformen lediglich hinsichtlich der Beugung eines Fingers oder der Daumenstellung unterscheiden. Um die Stellung der gesamten Hand zu erfassen, sind darüber hinaus eine oder mehrere weitere Markierungen auf dem Handrücken, der Handinnenfläche oder am Handgelenk sinnvoll. Abbildung 2.16 zeigt beispielhaft einen entsprechend markierten Stoffhandschuh.

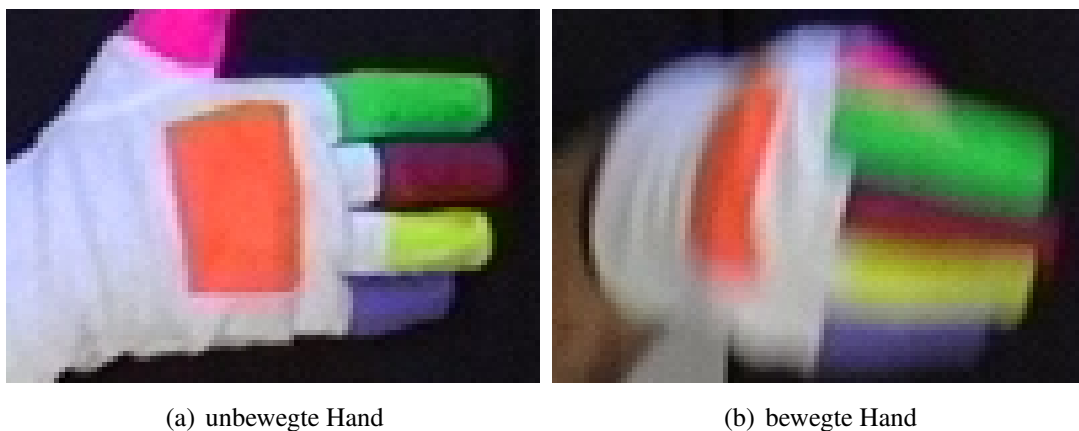


Abbildung 2.17: Auswirkung von Bewegungsunschärfe auf die markierte Hand. Wie in Abbildung 2.15 beträgt die Ausschnittaufösung 100×70 Pixel, die Verschlusszeit $1/120$ s (Kamera: Panasonic AG-DVX100B).

Im Vergleich zur unmarkierten Hand können auf diese Weise auch bei Bewegungsunschärfe, Kamerarauschen und vergleichsweise niedriger Auflösung stabile Merkmale bestimmt werden. Abbildung 2.17 verdeutlicht die Robustheit gegenüber Bewegungsunschärfe anhand derselben Positur, die in Abbildung 2.15 gezeigt ist. Mit Ausnahme des Stoffhandschuhs sind alle Aufnahmeparameter (Kamera, Verschlusszeit, Beleuchtung sowie die ausgeführte Gebärde, der das Einzelbild entnommen wurde) identisch. Das Bild veranschaulicht die Robustheit der optischen Marker, die bei gleicher Bewegungsgeschwindigkeit noch deutlich erkennbar und differenzierbar sind.

2.5.4 Merkmalsextraktion

Neben der Auswahl geeigneter Merkmale hat auch das zur Extraktion angewandte Verfahren entscheidenden Einfluss auf die Güte des resultierenden Merkmalsvektors. Unter Laborbedingungen sind die Anforderungen an die Merkmalsextraktion vergleichsweise niedrig. Insbesondere unter realistischen Aufnahmebedingungen stellt sie jedoch ein wichtiges Teilproblem der Positurrekonstruktion dar.

Das Tutorsystem soll vor unkontrollierten Hintergründen einsetzbar sein, um den Aufwand zur Bereitstellung geeigneter Aufnahmebedingungen zu reduzieren. Die Lokalisation der Marker über ihre Farben ist somit nicht notwendig eindeutig, da dieselben Farben auch im Bildhintergrund auftreten können. Abbildung 2.18 veranschaulicht dies anhand eines Beispiels.

Zur Behandlung der hieraus resultierenden Mehrdeutigkeiten wird, analog zur Positurrekonstruktion, der zeitliche Verlauf der Bewegung ausgewertet. In jedem Einzelbild ergeben sich zunächst mehrere Hypothesen hinsichtlich der Markerlokalisierung. Erst nach Ende der Sequenz, d. h. nach Analyse aller Einzelbilder, wird pro Einzelbild eine Hypothese so ausgewählt, dass sich eine flüssige Bewegung ergibt, deren Verlauf den Phasen der Gebärde – Vor-

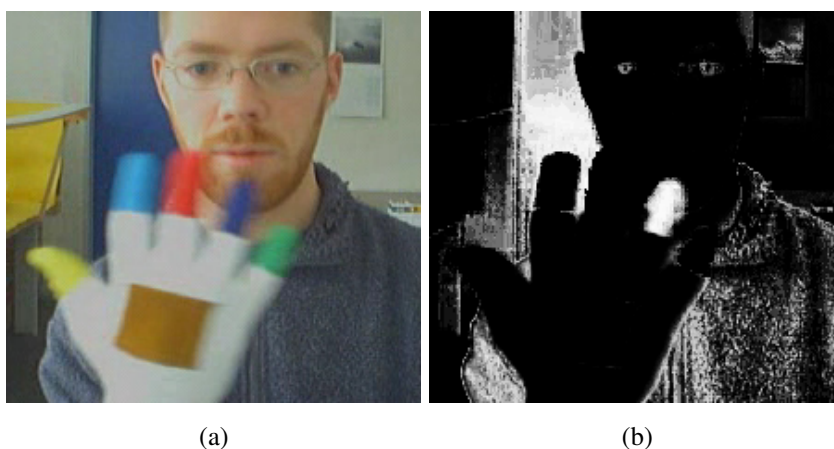


Abbildung 2.18: Die Farbe des Ringfingermarkers entspricht der Tür im Bildhintergrund (a). Die farbbasierte Markerdetektion liefert daher mehrere Kandidaten (helle Bereiche in (b)).

bereitung, Ausführung und Rückführung – entspricht. Dieser Ansatz des *multiple hypothesis tracking* wird in [ZK05] erfolgreich zur Verfolgung der unmarkierten Hand vor unkontrolliertem Hintergrund eingesetzt.

2.5.5 Positurrekonstruktion

Um den Anforderungen im Kontext des Gebärdensprachtutors zu genügen, sind beliebige Posituren aus beliebigen Ansichtswinkeln zu rekonstruieren. Dies schließt insbesondere Posituren ein, bei denen eine oder mehrere Markierungen durch die Hand selbst verdeckt sind. Weiterhin ist eine Auflösung von Mehrdeutigkeiten, die durch die Projektion in die 2D-Bildebene unvermeidbar sind, erforderlich.

Die Rekonstruktion aus beliebigen Ansichtswinkeln erlaubt auch eine beliebige Positionierung der Kamera. Dies erleichtert den Einsatz des Rekonstruktionsmoduls in Anwendungen außerhalb der Gebärdensprache, etwa im Bereich der Virtuellen Realität. Zur Erkennung von Greifbewegungen bietet sich hier die Aufnahme der Hand senkrecht von oben an.

Die wichtigste Entwicklungsentscheidung besteht in der Auswahl des Rekonstruktionsverfahrens. Ansätze, die auf inverser Kinematik beruhen, verwenden als Eingangsdaten die Positionen charakteristischer Punkte der Hand. Die Kontur der Marker liefert neben Ortskoordinaten jedoch noch weitere aussagekräftige Merkmale, beispielsweise Größe und Hauptachsenorientierung der sichtbaren Markerfläche. Um auch diese Informationen zu nutzen, geschieht die Rekonstruktion durch Anpassung eines Handmodells an die aus dem Eingangsbild extrahierten Merkmale (vgl. Abbildung 2.12).

Problembeschreibung

Mathematisch lässt sich die Erkennungsaufgabe im einfachsten Fall als Abbildung ψ_1 eines Merkmalsvektors $\underline{\mathbf{F}}$ auf eine Positur $\underline{\mathbf{P}}$ beschreiben:

$$\psi_1 : \underline{\mathbf{F}} \mapsto \underline{\mathbf{P}} \quad (2.6)$$

Erfolgt die Rekonstruktion ausgehend von der zuletzt erkannten Positur (d. h. gedächtnisbehaftet) durch schrittweises Optimieren einer Differenzfunktion, wird (2.6) erweitert zu

$$\psi_2 : \begin{bmatrix} \underline{\mathbf{F}}(t) \\ \underline{\mathbf{P}}(t-1) \end{bmatrix} \mapsto \underline{\mathbf{P}}(t) \quad (2.7)$$

Hierbei kann $\underline{\mathbf{P}}$ neben den manuellen Ausdrucksmitteln noch Zustandsinformationen wie etwa zeitliche Ableitungen beinhalten.

Weder (2.6) noch (2.7) beinhalten eine Möglichkeit zur Mehrdeutigkeitsauflösung. Weicht die rekonstruierte von der tatsächlichen Positur deutlich ab, hat dies nicht korrigierbare Rekon-

struktionsfehler zur Folge. Im Falle von (2.7) ist aufgrund der Rückkopplung von $\underline{\mathbf{P}}(t-1)$ ferner davon auszugehen, dass auch alle weiteren Posituren falsch erkannt werden.

Die Auflösung von Mehrdeutigkeiten in einem Einzelbild erfordert die Berücksichtigung des zeitlichen Zusammenhangs zwischen den Bildern einer Gebärdensequenz. In begrenztem Umfang kann dies durch Erweitern von (2.7) auf mehrere Positurkandidaten erreicht werden (vgl. [ISS04]):

$$\psi_3 : \begin{bmatrix} \underline{\mathbf{F}}(t) \\ \underline{\mathbf{P}}_1(t-1) \\ \underline{\mathbf{P}}_2(t-1) \\ \vdots \\ \underline{\mathbf{P}}_k(t-1) \end{bmatrix} \mapsto \begin{bmatrix} \underline{\mathbf{P}}_1(t) \\ \underline{\mathbf{P}}_2(t) \\ \vdots \\ \underline{\mathbf{P}}_k(t) \end{bmatrix} \quad (2.8)$$

Dieses Vorgehen berücksichtigt allerdings für einen Zeitpunkt t_0 nur die bis dahin extrahierten Merkmale, d. h. $[\underline{\mathbf{F}}(0) \ \underline{\mathbf{F}}(1) \ \dots \ \underline{\mathbf{F}}(t_0)]$, und damit lediglich einen Teil der in einer Sequenz enthaltenen Informationen. Um eine signifikante Steigerung der Rekonstruktionsleistung zu erreichen, ist ferner die Anzahl der Kandidaten $k \gg 1$ zu wählen. Dies ist bei Algorithmen mit einer Zeitkomplexität von $O(k)$ problematisch. Solange k klein gegen die Mächtigkeit der Positurenmenge (bei grober Quantisierung etwa 16000; s. Abschnitt 2.3.2) ist, besteht weiterhin die Gefahr der Konvergenz in Nebenoptima.

Um in jedem Einzelbild alle verfügbaren Informationen auszuwerten und dadurch die Rekonstruktionsleistung zu maximieren, sind die Eingangsdaten als Ganzes zu verarbeiten. Diese Strategie ist bisher in keiner Veröffentlichung zur Handpositurerkennung beschrieben. Bei einer Sequenzlänge von T Einzelbildern bedeutet dies

$$\psi_4 : [\underline{\mathbf{F}}(0) \ \underline{\mathbf{F}}(1) \ \dots \ \underline{\mathbf{F}}(T-1)] \mapsto [\underline{\mathbf{P}}(0) \ \underline{\mathbf{P}}(1) \ \dots \ \underline{\mathbf{P}}(T-1)] \quad (2.9)$$

Diese Abbildung erfolgt in mehreren Schritten, die zunächst formal beschrieben und anschließend implementierungsbezogen skizziert werden. Zunächst wird eine gedächtnislose Rekonstruktionsfunktion $\underline{\mathbf{R}}$ definiert, die pro Einzelbild k Positurkandidaten liefert:

$$\underline{\mathbf{R}}(\underline{\mathbf{F}}) = \begin{bmatrix} \underline{\mathbf{P}}_1 \\ \underline{\mathbf{P}}_2 \\ \vdots \\ \underline{\mathbf{P}}_k \end{bmatrix} \quad (2.10)$$

Für jeden Merkmalsvektor $\underline{\mathbf{F}}$ berücksichtigt $\underline{\mathbf{R}}$ alle möglichen Posituren $\underline{\mathbf{P}}$ und verhindert dadurch ein Konvergieren in einem Nebenoptimum. Weiterhin ist keine Initialisierung des Systems erforderlich, und Rekonstruktionsfehler haben keinen negativen Einfluss auf die weitere Verarbeitung.

Die Anwendung von (2.10) auf alle Merkmalsvektoren ergibt

$$\begin{aligned} [\underline{\mathbf{R}}(\underline{\mathbf{F}}(0)) \quad \underline{\mathbf{R}}(\underline{\mathbf{F}}(1)) \quad \dots \quad \underline{\mathbf{R}}(\underline{\mathbf{F}}(T-1))] &= \begin{bmatrix} \underline{\mathbf{P}}_1(0) & \underline{\mathbf{P}}_1(1) & \dots & \underline{\mathbf{P}}_1(T-1) \\ \underline{\mathbf{P}}_2(0) & \underline{\mathbf{P}}_2(1) & \dots & \underline{\mathbf{P}}_2(T-1) \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\mathbf{P}}_k(0) & \underline{\mathbf{P}}_k(1) & \dots & \underline{\mathbf{P}}_k(T-1) \end{bmatrix} \\ &= [\underline{\mathbf{P}}_{ij}]_{k \times T} \end{aligned} \quad (2.11)$$

In Verbindung mit einer Mehrdeutigkeitsauflösung $\underline{\mathbf{M}}$, die aus dieser Kandidatenmenge eine Positurrensequenz bestimmt,

$$\underline{\mathbf{M}}([\underline{\mathbf{P}}_{ij}]_{k \times T}) = [\underline{\mathbf{P}}(0) \quad \underline{\mathbf{P}}(1) \quad \dots \quad \underline{\mathbf{P}}(T-1)] \quad (2.12)$$

ergibt sich die in (2.9) beschriebenen Abbildung zu

$$\psi_4([\underline{\mathbf{F}}(0) \quad \underline{\mathbf{F}}(1) \quad \dots \quad \underline{\mathbf{F}}(T-1)]) = \underline{\mathbf{M}}([\underline{\mathbf{R}}(\underline{\mathbf{F}}(0)) \quad \underline{\mathbf{R}}(\underline{\mathbf{F}}(1)) \quad \dots \quad \underline{\mathbf{R}}(\underline{\mathbf{F}}(T-1))]) \quad (2.13)$$

Umsetzung

Die Realisierung der Rekonstruktionsfunktion $\underline{\mathbf{R}}$ basiert auf einem 3D-Modell der markierten Hand. Dieses Modell kann jede anatomisch mögliche Positur annehmen und gestattet die Berechnung derselben Merkmale, die für die Eingangsdaten vorliegen. Die Erkennungsaufgabe besteht nun in einer geeigneten Parametrisierung dieses Modells, so dass dessen 2D-Merkmale mit den aus den Eingangsdaten extrahierten möglichst gut übereinstimmen (vgl. (2.2)).

Um eine 2D-Modellansicht darzustellen und einen entsprechenden Merkmalsvektor zu berechnen, benötigt ein Standard-PC etwa 12 ms.⁷ In Echtzeit lässt sich auf diese Weise lediglich ein Bruchteil des Positurreaums berücksichtigen. Daher werden alle vom System zu erkennenden Posituren offline generiert und ihre Parameter $\underline{\mathbf{P}}$ zusammen mit den extrahierten Merkmalen $\underline{\mathbf{F}}$ in einer Datenbank gespeichert. So resultieren – je nach Quantisierung des Positurreaums – mehrere Millionen Einträge, die jeweils aus einem Paar $\langle \underline{\mathbf{P}}, \underline{\mathbf{F}}(\underline{\mathbf{P}}) \rangle$ bestehen.

Die Rekonstruktionsfunktion durchsucht diese Datenbank und ermittelt diejenigen k Posituren, für die eine geeignet zu definierende Merkmalsdifferenz $\Delta_{\underline{\mathbf{F}}}$ (siehe (2.1)) zu den Eingangsdaten möglichst niedrig ist. Für diese Suche wird die Datenbank indexiert, wobei zu berücksichtigen ist, dass die Merkmale rauschbehaftet sind und ein aus einem Realbild extrahierter Merkmalsvektor im Allgemeinen nicht exakt in der Datenbank enthalten ist. Die so

⁷Die Rechenzeit ist primär von CPU- und GPU-Leistung abhängig. Der angegebene Wert wurde mit einem AMD-Sempron-Rechner und einer ATI Radeon 9200-Grafikkarte (Taktfrequenz 1,6 GHz bzw. 250 MHz) ermittelt. Die Grafikkarte unterstützt die verwendete 3D-Bibliothek *OpenGL* direkt.

erreichte Rekonstruktionsleistung ist primär durch die Aussagekraft der Merkmalsdifferenz $\Delta_{\mathbf{F}}$ sowie die Indexierungs- bzw. Suchstrategie bestimmt.

Aufgrund der beschriebenen Mehrdeutigkeit enthält die Datenbank für nahezu alle möglichen Merkmalsvektoren $\mathbf{F}$ zahlreiche Einträge mit hoher Ähnlichkeit, jedoch deutlich verschiedenen Posituren $\mathbf{P}$. Daher ist k ausreichend groß zu wählen, um zu gewährleisten, dass mindestens einer der Kandidaten der tatsächlichen Positur entspricht. Auf diese Weise entstehen pro Einzelbild k Positurnhypthesen. Die Zeitkomplexität des gewählten Suchalgorithmus ist von k unabhängig, so dass große k an dieser Stelle keine Auswirkung auf die Laufzeit haben.

Dieses Ergebnis ist keine notwendige Folge der Verwendung einer Posituredatenbank, sondern trägt dem projektionsbedingten Informationsverlust Rechnung. Eine Wahl von $k = 1$ ist möglich, setzt jedoch die Leistung des Systems stark herab.

Eine Alternative besteht darin, die Datenbank nicht im Merkmals-, sondern im Positurreaum zu indexieren. Für das erste Einzelbild erfordert dies entweder eine vollständige Suche oder eine manuelle Initialisierung. Anschließend beschränkt sich die Suche auf die Nachbarschaft der aktuellen sowie ggf. einiger weiterer Posituren. Dieses Vorgehen entspricht (2.7) bzw. (2.8) und weist die dort genannten Nachteile auf. Es kommt daher im entwickelten System nicht zum Einsatz.

Zur Mehrdeutigkeitsauflösung, die pro Einzelbild einen der k Kandidaten auswählt, findet Domänenwissen hinsichtlich des Gebärdensprozesses Verwendung. Da der Tutor beliebige, insbesondere auch in der Gebärdensprache üblicherweise nicht auftretende Posituren bzw. Positurrensequenzen erkennen muss, stellt lediglich die zeitliche Kontinuität der Bewegung ein geeignetes Kriterium dar:

$$\mathbf{P}(t) \approx \mathbf{P}(t - 1) \quad (2.14)$$

Die Prüfung dieser Bedingung erfordert – analog zu $\Delta_{\mathbf{F}}$ – die Definition einer Posituredifferenz $\Delta_{\mathbf{P}}$ zu

$$\Delta_{\mathbf{P}}(\mathbf{P}_1, \mathbf{P}_2) \in \mathbb{R}_0^+ \quad \text{mit} \quad \Delta_{\mathbf{P}}(\mathbf{P}_1, \mathbf{P}_2) = 0 \Leftrightarrow \mathbf{P}_1 = \mathbf{P}_2 \quad (2.15)$$

Aus der Kandidatenmatrix $[\mathbf{P}_{ij}]_{k \times T}$ wird pro Spalte, d. h. pro Eingangsbild, je eine Positur so ausgewählt, dass die Summe der Differenzen aufeinanderfolgender Posituren sowie die Summe der entsprechenden Merkmalsdifferenzen möglichst gering sind. Im Unterschied zur $\mathbf{R}$ skaliert die Implementierung der Funktion $\mathbf{M}$ hinsichtlich der Laufzeit mit $O(Tk^2)$. Höhere Werte für k reduzieren daher die Geschwindigkeit, vergrößern jedoch den Betrachtungshorizont des Systems und damit die Wahrscheinlichkeit einer korrekten Erkennung.

Als Rekonstruktionsergebnis liegt schließlich eine kontinuierliche Positurrensequenz vor, die die Gebärde des Benutzers repräsentiert.

Bestimmung der Handposition im Raum

Der beschriebene Ansatz rekonstruiert Handform und -stellung im dreidimensionalen Raum, beschrieben durch die Lage der Finger- und Daumenglieder sowie des Handtellers relativ zueinander. Aufgrund der Beschränkung auf eine Kamera werden aus den Eingangsdaten jedoch keine Tiefeninformationen extrahiert und die Position der gesamten Hand daher nur in der xy -Ebene bestimmt. Zur Erfassung der Größenänderung der Marker infolge einer Bewegung entlang der z -Achse ist die räumliche Auflösung gängiger Kameras zu niedrig.

In der DGS hat die z -Achse die Funktion einer Zeitlinie. Bei einzelnen Gebärden und einfachen Sätzen ist der Informationsverlust daher gering. Für die Mehrzahl der Gebärden stellt die Beschränkung der Handposition auf die xy -Ebene eine gute Näherung dar.

Erweiterung auf beidhändige Gebärden

Bisher wurden die betrachteten Gebärden implizit als einhändig angenommen; die rekonstruierten Posituren beschreiben die dominante Hand. Für die nichtdominante Hand kann das Rekonstruktionsverfahren theoretisch ebenfalls angewendet werden. Da in der Praxis jedoch – wie im letzten Abschnitt beschrieben – prinzipiell ein Kompromiss zwischen Geschwindigkeit und Genauigkeit vorliegt, und die von der nichtdominanten Hand übermittelte Informationsmenge vergleichsweise gering ist, ist eine vereinfachende Behandlung sinnvoll.

In der DGS wird die nichtdominante Hand entweder symmetrisch zur dominanten Hand bewegt, oder sie bleibt in der Ausführungsphase statisch (s. Abschnitt 2.1.1). Daher wird ihre Positur in dieser Arbeit nicht vollständig rekonstruiert. Zur Feedbackgenerierung ist die Handposition bzw. die Trajektorie ausreichend. Für Merkmalsextraktion kommt entsprechend ein einfarbig markierter Handschuh zum Einsatz. Abbildung 2.19 zeigt beispielhaft die Markierung beider Hände.



Abbildung 2.19: Markierung der nichtdominanten (linken) sowie der dominanten (rechten) Hand.

Erweiterung auf mehrere Kameras

In der vorliegenden Arbeit erfolgt die Positurerkennung wie in Abschnitt 2.5.1 beschrieben aus monokularer Frontalansicht. Die Zuverlässigkeit der Rekonstruktion kann durch Verwendung mehrerer Kameras gesteigert werden, ohne die bestehenden Algorithmen zu modifizieren. Dazu sind zwei verschiedene Strategien möglich.

Nahe liegend ist die Zusammenfassung der aus n gleichzeitig aufgenommenen Einzelbildern extrahierten Merkmale in einem einzigen Merkmalsvektor:

$$\underline{\mathbf{F}} = \begin{bmatrix} \underline{\mathbf{F}}_1 \\ \underline{\mathbf{F}}_2 \\ \vdots \\ \underline{\mathbf{F}}_n \end{bmatrix} \quad (2.16)$$

Dieser Ansatz ist einfach umzusetzen, bewirkt jedoch eine Vergrößerung der Posituredatenbank entsprechend der zusätzlichen Merkmale. Weiterhin muss zum Zeitpunkt der Datenbankerstellung die Anordnung der Kameras bekannt sein, um entsprechende synthetische Ansichten zu generieren. In der Anwendung muss dieselbe Kameraanordnung verwendet werden.

Alternativ erfolgt für die Bilder jeder einzelnen Kamera zunächst die Bestimmung von k' Kandidaten wie beschrieben, d. h. unabhängig voneinander. Somit liegen bei n Kameras nk' Kandidaten vor. Für die Vereinigungsmenge dieser Posituren wird nun die Merkmalsdifferenz aus jeder Ansicht bestimmt. Dazu werden die Positurkandidaten je Kamera so rotiert, wie sie aus dem entsprechenden Blickwinkel erscheinen. Diejenigen k Posituren, für die die Summe dieser Merkmalsdifferenzen möglichst niedrig ist, d. h. die aus allen Blickwinkeln gut mit der beobachteten Positur übereinstimmen, bilden das Ergebnis der Rekonstruktionsfunktion $\underline{\mathbf{R}}$. Die weitere Verarbeitung erfolgt wie im monokularen Fall.

2.5.6 Positurenvergleich und Feedbackgenerierung

Um Anwender- und Referenzgebärde zu vergleichen, findet zunächst eine zeitliche Anpassung der beiden Positurensequenzen mittels *Dynamic Time Warping* (DTW) statt [MR81]. Dabei werden die Posituren einander so zugeordnet, dass die Abweichungen hinsichtlich der Handposition minimiert werden. Form und Stellung der Hand haben auf diesen Prozess keinen Einfluss. Dieses Vorgehen ist durch den hohen Informationsgehalt der Handposition motiviert [PKB90].

Aus der Folge der Handpositionen ergibt sich die Trajektorie sowie die Bewegungsgeschwindigkeit der Hand. Diese beiden Kriterien bilden die erste Stufe des Positurenvergleichs. Sie werden bei zweihändigen Gebärden für beide Hände geprüft.

Falls die gemessene Abweichung im Bereich der natürlichen Variabilität liegt, erfolgt anschließend für die dominante Hand der Vergleich von Handform und -stellung auf der Grund-

lage der in (2.15) zur Mehrdeutigkeitsauflösung eingeführten Posituredifferenz Δ_P .

Treten Abweichungen zwischen den verglichenen Gebärden auf, erfolgt eine entsprechende textuelle Rückmeldung sowie eine Visualisierung des Fehlers. Die Soll- und Ist-Trajektorien werden in Form von Polygonzügen angezeigt. Beim Vergleich von Handform und -stellung gestattet das 3D-Handmodell eine anschauliche Darstellung der Abweichung. Unabhängig hiervon kann der Lernende die zeitlich ausgerichteten Sequenzen parallel betrachten und manuell einzelbildweise vorwärts und rückwärts navigieren.

2.5.7 Zusammenfassung

Abschließend erfolgt eine kurze Übersicht des entwickelten Konzepts. In der Reihenfolge der Verarbeitung im Rechner werden die gewählten Methoden und die daraus resultierenden Randbedingungen zusammengefasst. Diese Reihenfolge bestimmt gleichzeitig die Gliederung der folgenden Kapitel. Abbildung 2.20 zeigt schematisch die in den Abschnitten 2.5.1 bis 2.5.5 behandelte Positurerkennung, während in Abbildung 2.21 Positurenvergleich und Feedbackgenerierung entsprechend Abschnitt 2.5.6 visualisiert sind.

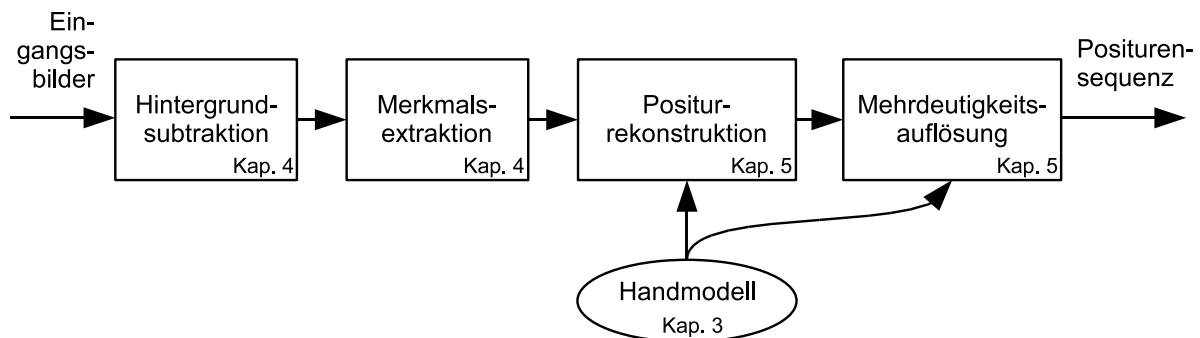


Abbildung 2.20: Verarbeitungskette zur Rekonstruktion einer Positurrensequenz. In den Verarbeitungsmodulen ist jeweils das Kapitel angegeben, das diese beschreibt.

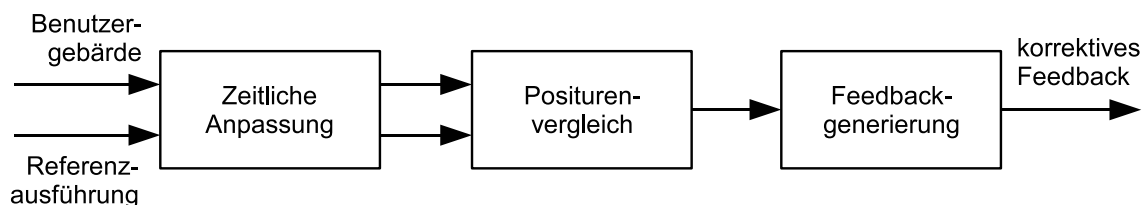


Abbildung 2.21: Schematische Darstellung des Vergleichs von Benutzer- und Referenzgebärden, die entsprechend Abbildung 2.20 erfasst wurden. Die verwendeten Algorithmen sind in Kapitel 6 erläutert.

Eingangsbilder

Die Aufnahme des Benutzers erfolgt monokular und mit einer Standard-Webcam bei einer Auflösung von ca. 100×100 Pixel für die Hand, wobei die Bildqualität durch Rauschen und/oder Bewegungsunschärfe beeinträchtigt sein kann. Der Bildausschnitt umfasst den gesamten Gebärdenraum, so dass translatorische Bewegungen in der Bildebene uneingeschränkt erfasst werden. Die Bildrate der Kamera beträgt mindestens 25 Hz. Der Bildhintergrund ist statisch, sonst jedoch unkontrolliert.

Hintergrundsubtraktion

Mit Hilfe eines Hintergrundmodells werden statische Bildbereiche auf Pixelebene erkannt und ausgeblendet. Für die eigentliche Verarbeitung erhöht sich die Verarbeitungsgeschwindigkeit, da nur noch bewegte Bereiche zu verarbeiten sind, und die Fehlerwahrscheinlichkeit sinkt.

Merkmalsextraktion

Zur Merkmalsextraktion wird die dominante Hand mit mehreren, die nichtdominante Hand mit einem farbigen Marker versehen. Das Verfahren verwendet multiple Hypothesen, um Mehrdeutigkeiten zu behandeln, die entstehen können, wenn die Markerfarben im Hintergrund enthalten sind. Die Bestimmung der Handposition geschieht innerhalb der Bildebene.

Positurrekonstruktion

Die Rekonstruktion erfolgt je Einzelbild durch Anpassung eines synthetischen Handmodells hinsichtlich der extrahierten 2D-Merkmale. Zur Beschleunigung der Ansichtengenerierung dieses Modells dient eine Datenbank mehrerer Millionen Posituren und jeweils zugehöriger 2D-Merkmale, die einmalig offline generiert wird. Das Ergebnis der Einzelpositurrekonstruktion ist eine Liste mehrerer Kandidaten und berücksichtigt somit die Mehrdeutigkeit der Problemstellung.

Mehrdeutigkeitsauflösung

Bildsequenzen werden als Ganzes verarbeitet, indem für jedes Einzelbild einer der zuvor ermittelten Positürkandidaten ausgewählt wird. Diese Auswahl erfolgt unter Berücksichtigung des zeitlichen Verlaufs, so dass eine flüssige Positurresequenz resultiert. Somit basiert die Positurselektion in jedem Einzelbild auf den aus der gesamten Sequenz extrahierten Informationen, wodurch die Rekonstruktionsgenauigkeit insbesondere bei verrauschten Merkmalen erhöht wird.

Rekonstruktionsleistung

Im Unterschied zu bestehenden Systemen wird der Positorenraum in keiner Weise eingeschränkt. Insbesondere sind Selbstverdeckungen der Marker sowie beliebige Ansichtswinkel möglich. Die Menge der zur Mehrdeutigkeitsauflösung und Fehlerkorrektur verfügbaren Informationen wird durch Analyse der gesamten Sequenz maximiert.

Positorenvergleich und Feedbackgenerierung

Der Vergleich von Anwender- und Referenzausführung basiert auf einer zeitlichen Anpassung bezüglich der Trajektorien. Sind diese hinreichend ähnlich, erfolgt anschließend der Vergleich von Handform und -stellung. Bei Abweichungen wird dem Lernenden textuell sowie grafisch korrekatives Feedback gegeben. Für die nichtdominante Hand beschränkt sich der Gebärdenvergleich auf die Position.

Kapitel 3

Modellierung der Hand

Grundlage für die Erkennung von Handposituren ist ein präzises dreidimensionales Modell der Hand. Da die Positurerkennung auf dem Vergleich von realer und Modellansicht beruht, ist eine möglichst große Ähnlichkeit beider Ansichten bei gleicher Positur das primäre Ziel der Modellierung. Dabei liegt der Schwerpunkt auf den Markerflächen, weil deren geometrische Eigenschaften für den Ansichtenvergleich verwendet werden. Um eine hohe Übereinstimmung zu erreichen, muss das Modell die Anatomie der Hand, insbesondere die Bewegungsmöglichkeiten der Finger und des Daumens, hinreichend genau nachbilden.

Neben der eigentlichen Positurrekonstruktion ermöglicht das Handmodell auch die Generierung synthetischer Eingangsdaten für vorgegebene Posituren bzw. Positurrensequenzen. Durch Vergleichen der Eingangsposituren mit dem Rekonstruktionsergebnis ist eine quantitative Leistungsmessung und damit eine Optimierung der verwendeten Algorithmen und Parameter möglich. Am Ende der Verarbeitungskette dient das Modell der Darstellung des Erkennungsergebnisses und der Visualisierung von Ausführungsfehlern des Anwenders.

Die beiden Abschnitte 3.1 und 3.2 beschreiben die funktionale Anatomie sowie die daraus abgeleitete Modellierung der Hand. In Abschnitt 3.3 schließt die Definition einer Metrik zum Vergleich zweier Handposituren an, die auf dem Handmodell beruht und an mehreren Stellen der Verarbeitungskette von Bedeutung ist.

3.1 Funktionale Anatomie

Die Hand ist der am stärksten verformbare Teil des menschlichen Skeletts. Sie besteht aus 27 Knochen, die zu acht Handwurzel-, fünf Mittelhand- und 14 Fingerknochen gruppiert werden [Sob93] und in Abbildung 3.1 gezeigt sind. Mittelhand und Handwurzel bilden zusammen die Handfläche. Jeder Finger besteht aus drei Gliedern (Phalangen):

- Phalanx proximalis (PP; Grundglied)
- Phalanx media (PM; Mittelglied)

- Phalanx distalis (PD; Endglied)

Der Daumen verfügt lediglich über Grund- und Endglied, das Mittelglied fehlt. Alle fünf Grundglieder reichen bis in die Handfläche und sind damit deutlich länger als Mittel- und Endglieder.

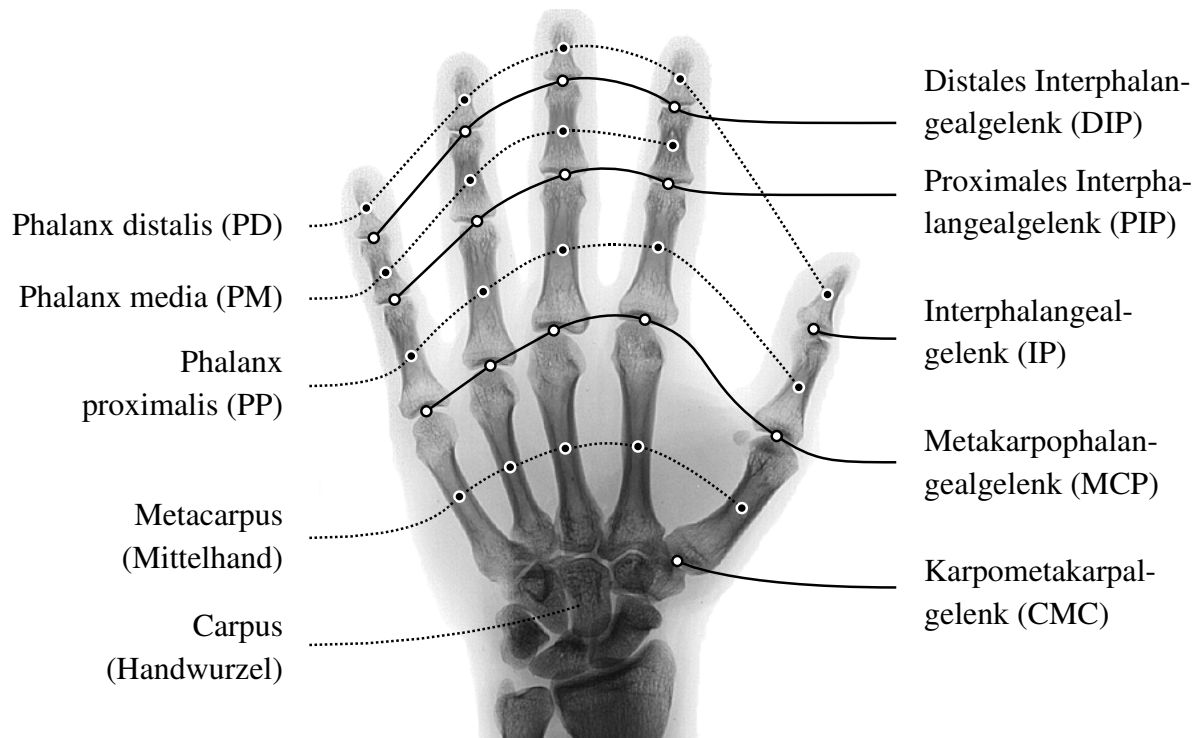


Abbildung 3.1: Röntgenaufnahme einer menschlichen Hand. Links ist die Bezeichnung der Knochen, rechts die der Gelenke angegeben. Am unteren Bildrand sind links Ulna (Elle) und rechts Radius (Speiche) zu sehen.

3.1.1 Gelenke und Freiheitsgrade

Zur Änderung ihrer Form verfügt die Hand über 28 Gelenke und 33 Einzelmuskeln [Sob93, Kap84]. Die Betrachtung dieser Gelenke und der daraus resultierenden Freiheitsgrade (*degrees of freedom*, DOF) hinsichtlich der Handform bildet eine zentrale Komponente der Modellierung.

Im Kontext der Gebärdensprache sind die Gelenke in Fingern und Daumen von Interesse, da nur diese eine von der Kamera registrierbare Formveränderung bewirken. Mittelhand und Handwurzel können als statisch angenommen werden, obwohl sie – etwa beim Greifen eines Balls – begrenzt verformbar sind.

Grund-, Mittel- und Endglied der vier Finger sind über zwei Scharniergelenke (proximales Interphalangealgelenk, PIP, bzw. distales Interphalangealgelenk, DIP) miteinander verbunden.

PIP und DIP ermöglichen Beugen (Flexion) und Strecken (Extension) und machen somit jeweils einen Freiheitsgrad aus. Das Grundgelenk (Metakarpophalangealgelenk, MCP) jedes Fingers ist ein Kugelgelenk, das neben Beugen und Strecken auch seitliches Spreizen (Abduktion) und Schließen (Adduktion) erlaubt und daher zwei Freiheitsgrade bedingt.

Der Daumen ist opponierbar – d. h. er lässt sich jedem Finger gegenüberstellen – und hat damit besonders bei Greifbewegungen eine wichtige Funktion. Trotz des nicht vorhandenen Mittiglieds verfügt er über mehr Freiheitsgrade als ein Finger. Das Interphalangealgelenk (IP) ist wie das DIP-Gelenk ein reines Scharniergelenk mit einem Freiheitsgrad. Im Daumengrundgelenk (MCP), mechanisch ein Ellipsoidgelenk, sind neben Flexion und Extension in deutlich geringerem Maß auch Adduktion, Abduktion und Rotation möglich. An der Handwurzel befindet sich das Karpometakarpalgelenk (CMC) mit den drei Freiheitsgraden der Flexion und Extension, Abduktion und Adduktion sowie Rotation, die als Pronation (Einwärtsdrehung) und Supination (Auswärtsdrehung) bezeichnet wird.

Insgesamt resultieren aus diesen Betrachtungen 23 Freiheitsgrade (vier je Finger, sieben für den Daumen). In der Literatur ist die Anzahl der Freiheitsgrade der menschlichen Hand überwiegend zwischen 21 und 29 angegeben. Diese Abweichungen entstehen dadurch, dass zahlreiche Autoren eine Beweglichkeit von wenigen Grad, wie sie etwa bei den Gelenken im Handteller auftritt, nicht als Freiheitsgrad werten.

3.1.2 Einschränkungen

Nicht alle Gelenke der Hand lassen sich unabhängig voneinander bewegen. Es bestehen verschiedene Kopplungen und Einschränkungen, die den zu betrachtenden Positorenraum reduzieren und daher zur Vereinfachung der Erkennungsaufgabe nutzbar sind.

DIP- und PIP-Gelenk verwenden dieselbe Sehne und können von den meisten Menschen nicht unabhängig voneinander bewegt werden. Weiterhin beeinflussen sich die vier MCP-Gelenke der Finger untereinander. Eine entsprechende Untersuchung [RF99] fand hier die stärkste Abhängigkeit bei der Streckung des Ringfingers, die ein Beugen von kleinem oder Mittelfinger signifikant beschränkt. Die Streckung des Mittelfingers hat hingegen den geringsten Einfluss und behindert die Beugung der übrigen Finger kaum. Quantitativ sind diese Effekte allerdings nicht personenunabhängig zu beschreiben.

Der Winkelbereich der Abduktion und Adduktion hängt vom Beugungsgrad der Fingergelenke ab. Er ist bei vollständiger Extension maximal und nimmt mit zunehmender Flexion bis auf annähernd Null ab. Darüber hinaus können benachbarte Finger die Bewegung einschränken; beispielsweise ist eine Abduktion des Ringfingers nur bei gleichzeitiger Abduktion des kleinen Fingers möglich.

3.2 3D-Modell

Dieser Abschnitt erläutert die dreidimensionale Modellierung der Hand. Zu Beginn wird der Aufbau des 3D-Modells sowie das Vorgehen zur grafischen Darstellung präsentiert. Daran schließt die Beschreibung der Positurparameter sowie die Umsetzung der im letzten Abschnitt skizzierten anatomischen Einschränkungen an.

3.2.1 Aufbau

Das entwickelte 3D-Modell bildet die in Abschnitt 3.1 beschriebene Anatomie der Hand mit Ausnahme der folgenden Vereinfachungen exakt nach:

- Bis auf das CMC-Gelenk wird der Handteller als statisch angenommen, da eine Veränderung seiner Form anatomisch nur sehr begrenzt möglich ist.
- Das MCP-Gelenk des Daumens wird als reines Scharniergelenk mit nur einem Freiheitsgrad (Flexion/Extension) betrachtet. Die Rotation überstreicht nur wenige Grad und ist damit im Eingangsbild kaum auflösbar. Gleiches gilt für Abduktion und Adduktion – die maximale Beugungswinkeldifferenz liegt hier im Bereich von lediglich ca. 10–20°.

Abbildung 3.2 zeigt schematisch das resultierende Handmodell. In Tabelle 3.1 sind die modellierten Gelenke mit der Anzahl ihrer Freiheitsgrade aufgelistet. Insgesamt ergeben sich für die Handform damit 20 DOF.

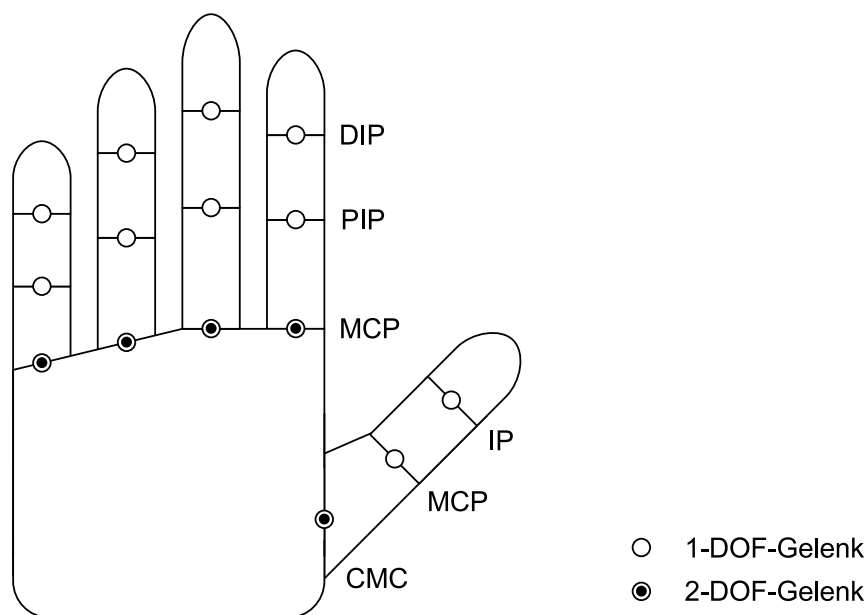


Abbildung 3.2: Schematische Darstellung des entwickelten Handmodells.

Tabelle 3.1: Gelenke des Handmodells und deren Freiheitsgrade.

	Gelenk	Abkürzung	DOF
Finger	Distales Interphalangealgelenk	DIP	1
	Proximales Interphalangealgelenk	PIP	1
	Metakarpophalangealgelenk	MCP	2
Daumen	Interphalangealgelenk	IP	1
	Metakarpophalangealgelenk	MCP	1
	Karpometakarpalgelenk	CMC	2

Obwohl die Positurrekonstruktion benutzerunabhängig erfolgen kann, ist es sinnvoll, die Handgeometrie, d. h. die Abmessungen von Fingern, Daumen und Handteller, variabel zu halten. Dadurch ist eine Optimierung der Erkennungsgenauigkeit für einzelne Anwender möglich. Tabelle 3.2 zeigt die verschiedenen Maße der Handgeometrie sowie beispielhafte Messwerte, die für eine Versuchsperson im visuellen Vergleich durch Überlagerung realer und synthetischer Ansichten bestimmt wurden.

Tabelle 3.2: Elemente des Handmodells, die die Handgeometrie bestimmen, sowie deren Abmessungen am Beispiel einer Versuchsperson.

Modellelement	Maße	Messwerte
Handteller	Breite $\times$ Höhe $\times$ Tiefe	8,75 $\times$ 9,25 $\times$ 3,00 cm
Daumen	PD / PP / Metakarpalknochen	2,50 / 3,00 / 4,50 cm
Zeigefinger	PD / PM / PP	1,75 / 2,50 / 3,25 cm
Mittelfinger	PD / PM / PP	2,00 / 3,00 / 3,75 cm
Ringfinger	PD / PM / PP	1,75 / 3,00 / 3,50 cm
Kleiner Finger	PD / PM / PP	2,00 / 2,00 / 3,25 cm

3.2.2 Grafische Darstellung

Zur grafischen Darstellung des Handmodells wird die plattform- und sprachübergreifende 3D-Grafikbibliothek *OpenGL* (*Open Graphics Library*) verwendet. Aktuelle Grafikkarten verfügen über eine hardwareseitige Beschleunigung bei der Ausführung von OpenGL-Kommandos, so dass das Rendern einer Positur auf einem Standard-PC (AMD-Sempron-CPU, 1,6 GHz, mit ATI Radeon 9200-Grafikkarte, 250 MHz) bei einer Auflösung von 256 $\times$ 256 ca. 12 ms benötigt.

Die Modellierung verwendet Kugeln für Gelenke und Fingerspitzen, Zylinder und Kegel-

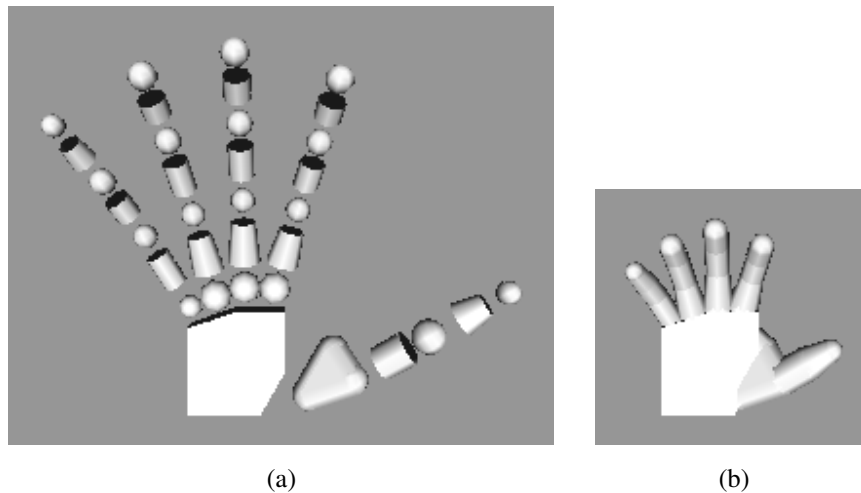


Abbildung 3.3: Palmaransicht (Sicht auf die Handinnenfläche) des Handmodells bei maximaler Abduktion der Finger: Explosionsansicht (a) und normale Ansicht (b).

stümpfe für die Fingerglieder und Polygone für die Handfläche und den Metakarpalknochen des Daumens. Abbildung 3.3 zeigt eine Explosionsansicht, in der die genannten Elemente erkennbar sind, sowie die normale Modellansicht.

Beliebige Markergeometrien lassen sich durch Einfärben der entsprechenden Modellelemente oder Hinzufügen farbiger Flächen leicht nachbilden. In Abbildung 3.4a/b ist eine reale und eine Modellansicht einer Markergeometrie zu sehen, bei der neben Fingern und Daumen auch der Handrücken markiert ist. Im Modell ist die letztgenannte Markierung durch ein entsprechend gefärbtes Rechteck repräsentiert. Die Fingermarker bedecken das PD- und das halbe PM-Glied. Abbildung 3.4c/d zeigt eine Geometrie, die an Stelle des Handrückensmarkers einen umlaufenden Handgelenkmarker verwendet. Die Fingermarker reichen etwa bis zur Mitte des PP-Glieds.

Da die Markergeometrie starken Einfluss auf die Genauigkeit der Positurrekonstruktion hat, erfolgt in Abschnitt 7.1.2.5 ein quantitativer Vergleich mehrerer Anordnungen. Aus Konsistenzgründen wird in den folgenden Abbildungen die in Abbildung 3.4a/b gezeigte Markergeometrie verwendet.

3.2.3 Modellparameter

Um eine vorgegebene Positur anzunehmen, muss das Handmodell entsprechend parametrisiert werden. Dies umfasst Form, Stellung und Position und ist im Folgenden erläutert.

3.2.3.1 Handform

Die Form der synthetischen Hand wird – neben der Modellgeometrie – durch die Stellungen der Gelenke bestimmt. Finger und Daumen enthalten jeweils zwei 1-DOF- sowie ein 2-DOF-

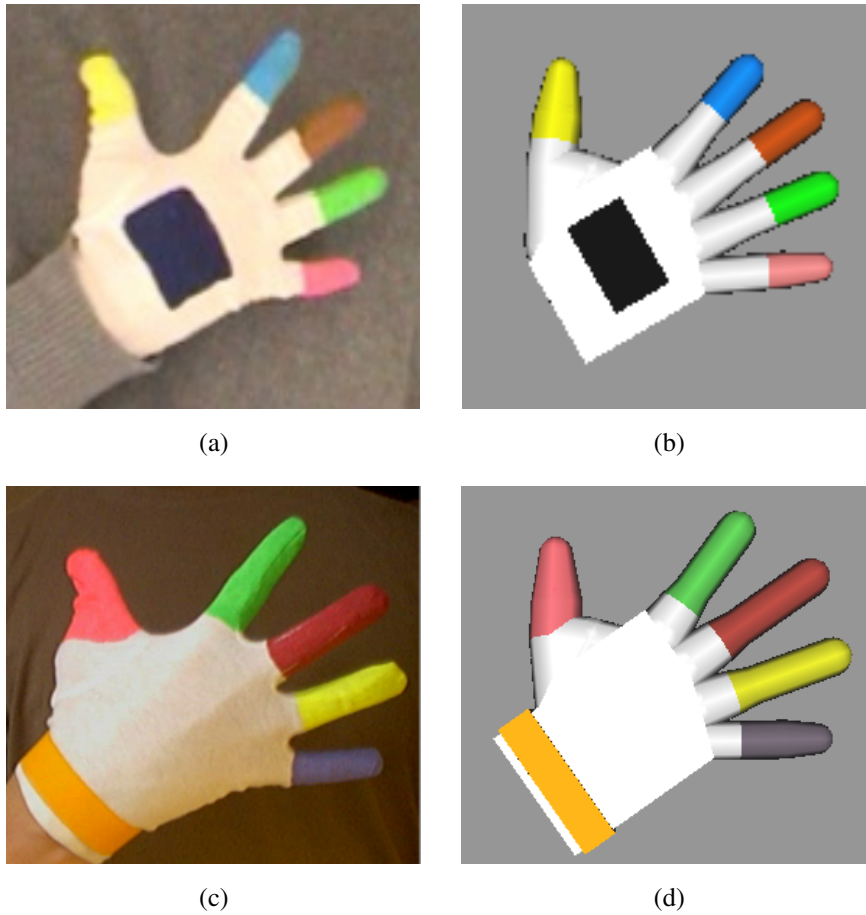


Abbildung 3.4: Modellierung einer Markergeometrie durch Einfärben von Finger- und Dau-menelementen sowie Hinzufügen zusätzlicher Markerflächen auf dem Handrücken (b) bzw. am Handgelenk (d).

Gelenk, so dass sich fünf vierdimensionale Parametervektoren ergeben:

$$\underline{\mathbf{B}}_D = \begin{bmatrix} b_{D,1} \\ b_{D,2} \\ b_{D,3} \\ b_{D,4} \end{bmatrix} \quad \underline{\mathbf{B}}_{ZF} = \begin{bmatrix} b_{ZF,1} \\ b_{ZF,2} \\ b_{ZF,3} \\ b_{ZF,4} \end{bmatrix} \quad \underline{\mathbf{B}}_{MF} = \begin{bmatrix} b_{MF,1} \\ b_{MF,2} \\ b_{MF,3} \\ b_{MF,4} \end{bmatrix} \quad \underline{\mathbf{B}}_{RF} = \begin{bmatrix} b_{RF,1} \\ b_{RF,2} \\ b_{RF,3} \\ b_{RF,4} \end{bmatrix} \quad \underline{\mathbf{B}}_{KF} = \begin{bmatrix} b_{KF,1} \\ b_{KF,2} \\ b_{KF,3} \\ b_{KF,4} \end{bmatrix} \quad (3.1)$$

Die skalaren Größen $b_{D,i}$, $b_{ZF,i}$, $b_{MF,i}$, $b_{RF,i}$ und $b_{KF,i}$ mit $i = 1, 2, 3, 4$ geben den Beugungs-grad an, wobei 0 für maximale Extension bzw. Adduktion und 1 für maximale Flexion bzw. Abduktion des Gelenks steht:

$$b_{D,i}, b_{ZF,i}, b_{MF,i}, b_{RF,i}, b_{KF,i} \in [0, 1] \quad i = 1, 2, 3, 4 \quad (3.2)$$

Um die anatomischen Einschränkungen der Gelenkbeweglichkeit zu modellieren, ist die Ab-bildung dieses Wertebereichs auf den Beugungswinkel nicht für alle Gelenke konstant, son-

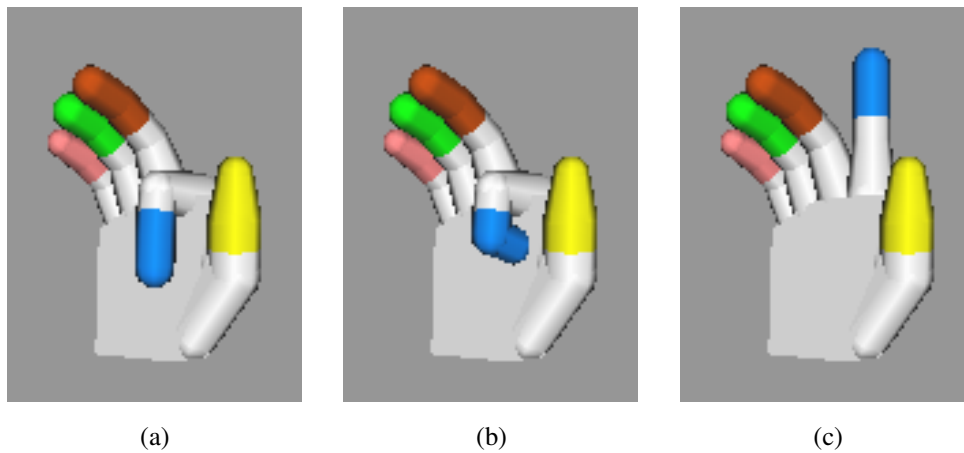


Abbildung 3.5: Wertebereich des Beugungswinkels des DIP-Gelenks in Abhängigkeit von PIP- und MCP-Gelenk. Sind diese vollständig gebeugt (a, b), wird das Parameterintervall $[0, 1]$ auf den Winkelbereich von 0° (a) bis 80° (b) abgebildet. Bei maximaler Streckung ist die Beugung des DIP-Gelenks konstant 0° (c).

dern wird jeweils anhand der Stellung der vorangehenden, d. h. in der kinematischen Kette näher am Handgelenk gelegenen Gelenke bestimmt. So entspricht etwa der Wertebereich $[0, 1]$ für das DIP-Gelenk bei vollständiger Beugung von PIP- und MCP-Gelenk dem Winkelbereich von ca. $[0^\circ, 80^\circ]^1$, während bei maximaler Extension in PIP- und MCP-Gelenk der Beugungswinkel des DIP-Gelenks konstant 0° ist. Dieses Beispiel ist in Abbildung 3.5 visualisiert. Die Beugungsabhängigkeit von Abduktion und Adduktion in den 2-DOF-Gelenken ist auf dieselbe Weise modelliert.

Durch Zusammenfassen sämtlicher Beugungsparameter lässt sich die Handform durch einen Vektor $\underline{\mathbf{B}} \in \mathbb{R}^{20}$ beschreiben:

$$\underline{\mathbf{B}} = \begin{bmatrix} \underline{\mathbf{B}}_{\text{D}} \\ \underline{\mathbf{B}}_{\text{ZF}} \\ \underline{\mathbf{B}}_{\text{MF}} \\ \underline{\mathbf{B}}_{\text{RF}} \\ \underline{\mathbf{B}}_{\text{KF}} \end{bmatrix} \quad (3.3)$$

Abbildung 3.6 zeigt exemplarisch verschiedene Beugungsgrade der Fingergelenke aus Palmar- und Seitenansicht.

¹Eine Beugung von 0° im DIP-Gelenk ist in dieser Konfiguration durch flaches Auflegen der Fingerendglieds auf den Handteller erreichbar. Da sich Objekte in OpenGL widerstandslos durchdringen, erfolgt die Modellierung direkt über den Beugungswinkel.

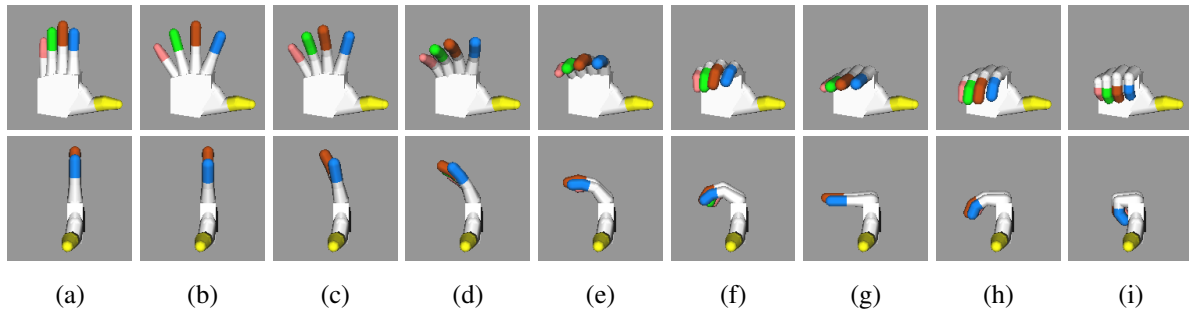


Abbildung 3.6: Verschiedene Beugungswinkel in Palmar- (oben) und Seitenansicht (unten).

3.2.3.2 Handstellung

Neben der Handform muss das Modell auch die Handstellung berücksichtigen. Bei konstanter Position ergeben sich aus den Bewegungsmöglichkeiten in Schulter-, Ellenbogen- und Handgelenk drei Freiheitsgrade. Diese lassen sich als Rotation der Hand um ihren Mittelpunkt, der als Mittelpunkt des Handtellers definiert wird, auffassen und durch drei Winkel angeben.

Zwei dieser Winkel beschreiben einen Punkt auf der Oberfläche einer imaginären Kugel, deren Mittelpunkt mit dem der Hand übereinstimmt. Die entsprechende Modellansicht ergibt sich durch den Blick von diesem Punkt in Richtung des Mittelpunkts, so dass beide Punkte auf der optischen Achse liegen. Diese Kugel wird als Ansichtskugel (*view sphere*) bezeichnet, die beiden Ansichtswinkel in Analogie zum Globus als Breitengrad (*latitude*) ϕ_{lat} und Längengrad (*longitude*) ϕ_{lon} . Der dritte Winkel ϕ_{rot} beschreibt eine Rotation der Hand bzw. – in entgegengesetzter Richtung – der Kamera um die optische Achse. Die drei Ansichtswinkel werden zu einem Vektor $\underline{\Phi} \in \mathbb{R}^3$ zusammengefasst, der die Handstellung charakterisiert:

$$\underline{\Phi} = \begin{bmatrix} \phi_{\text{lat}} \\ \phi_{\text{lon}} \\ \phi_{\text{rot}} \end{bmatrix} \quad (3.4)$$

Der Breitengrad ϕ_{lat} nimmt Werte von -90° („Nordpol“) bis $+90^\circ$ („Südpol“) an. Für Längengrad ϕ_{lon} und Rotation ϕ_{rot} beträgt der Wertebereich $[0^\circ, 360^\circ)$. In Abbildung 3.7 bis 3.9 sind einige Ansichten dargestellt, in denen jeweils einer der drei Winkel variiert wird, während die beiden anderen konstant bleiben.

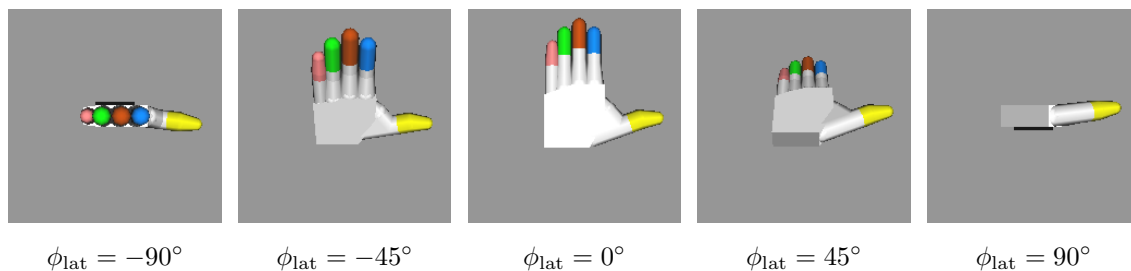


Abbildung 3.7: Verschiedene Breitengrade bei $\phi_{\text{lon}} = \phi_{\text{rot}} = 0^\circ$.

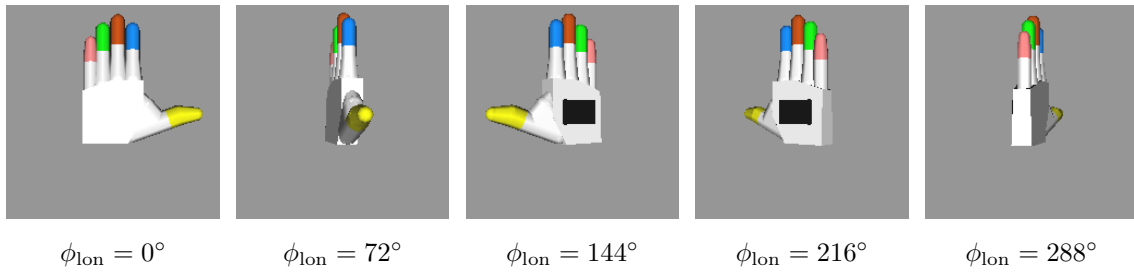


Abbildung 3.8: Verschiedene Längengrade bei $\phi_{lat} = \phi_{rot} = 0^\circ$.

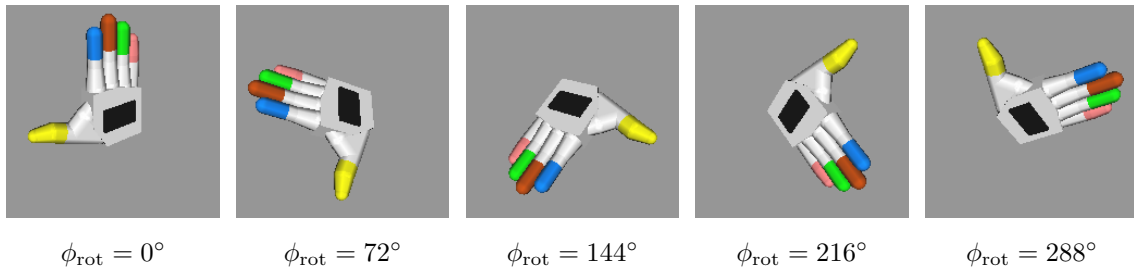


Abbildung 3.9: Verschiedene Rotationen bei $\phi_{lat} = -36^\circ$ und $\phi_{lon} = 144^\circ$.

Je geringer der Abstand eines Ansichtspunkts zu einem der beiden Pole der Ansichtskugel ist, desto mehr entspricht eine Änderung des Längengrads einer Rotation um die optische Achse, weil der Winkel zwischen beiden Drehachsen abnimmt. Für $\phi_{lat} = \pm 90^\circ$ sind beide Achsen identisch, so dass die Auswirkung einer Winkeländerung für Längengrad und Rotation gleich ist. Dieser Effekt wird als *gimbal lock* bezeichnet und ist beim Vergleich zweier Posituren zu berücksichtigen.

3.2.3.3 Handposition

Die Position des Handmodells im Bild kann mit drei Freiheitsgraden variiert werden, wobei eine Verschiebung entlang der z -Achse einer Skalierung entspricht. Allerdings macht die Rekonstruktionsstufe hiervon keinen Gebrauch, da die Positionsmessung ohne Modell einfacher und ebenso zuverlässig ist.

Im Rahmen der Positurrekonstruktion bleibt der Mittelpunkt des Handmodells daher fest, und die Handposition wird separat ermittelt. Lediglich zur Visualisierung des Erkennungsergebnisses erfolgt ein Rendern unter Berücksichtigung der Raumkoordinaten.

3.2.3.4 Notation der Parameter

Die Messung der Handposition ist vergleichsweise unkompliziert und bildet daher lediglich eine untergeordnete Komponente des Tutorsystems. Die Positurrekonstruktion besteht somit im Wesentlichen aus der Bestimmung von Handform $\underline{\mathbf{B}}$ und Handstellung $\underline{\Phi}$. Eine Positur $\underline{\mathbf{P}}$ im mathematischen Sinn wird als Tupel dieser beiden Komponenten definiert, um eine kompakte Notation zu erhalten:

$$\underline{\mathbf{P}} = \begin{bmatrix} \underline{\mathbf{B}} \\ \underline{\Phi} \end{bmatrix} \in \mathbb{R}^{23} \quad (3.5)$$

3.2.4 Einschränkungen

Der Positurvektor $\underline{\mathbf{P}}$ spannt den 23-dimensionalen Positurenraum auf, innerhalb dessen sich das Erkennungsergebnis bewegt. Trotz der in Abschnitt 3.2.3.1 beschriebenen variablen Interpretation der Beugungsgrade enthält dieser Raum noch Posituren, die in der Realität nicht auftreten können. Diese im Kontext der Handmodellierung zu identifizieren und auszublenden, erhöht die Leistungsfähigkeit der Rekonstruktion in zweierlei Hinsicht: Da weniger Posituren zu berücksichtigen sind, steigt die Verarbeitungsgeschwindigkeit, während gleichzeitig die Wahrscheinlichkeit für Rekonstruktionsfehler, die aus der Ermittlung einer visuell ähnlichen, anatomisch jedoch unmöglichen Positur resultieren, sinkt.

Im Positurenraum werden daher die folgenden Bereiche als unzulässig definiert:

1. Handformen, bei denen die Stellung der verschiedenen Finger aufgrund der in Abschnitt 3.1.2 beschriebenen Abhängigkeiten der Gelenke anatomisch nicht möglich ist. Diese Einschränkung betrifft die Beugungsparameter $\underline{\mathbf{B}}$ und lässt sich einfach regelbasiert modellieren.
2. Posituren, in denen die Hand aus der Richtung des Handgelenks betrachtet wird. Bei einer Frontalansicht sind derartige Handstellungen aufgrund der erforderlichen starken Beugung im Handgelenk schmerzhaft und treten daher beim Gebärden auch bei Lernenden nicht auf. Sie lassen sich durch Beschränkung des Breitengrades ϕ_{lat} vermeiden, was geometrisch betrachtet dem Ausschneiden eines Kegels am Südpol der Ansichtskugel entspricht.

Beispiele für Posituren dieser beiden Kategorien sind in Abbildung 3.10 zu sehen.

3.3 Vergleich von Posituren

In mehreren Modulen des Tutorsystems ist es erforderlich, die Verschiedenheit zweier Posituren zu messen. Ein Beispiel hierfür ist die Rekonstruktionsstufe, die eine „flüssige“ Positurrensequenz liefern soll, d. h. eine Sequenz, in der sich zeitlich aufeinanderfolgende Posituren wenig unterscheiden.

Zu diesem Zweck werden Differenzmaße benötigt, die die Unterschiede der dimensionslosen Beugungsgrade sowie die Abweichungen der Ansichtswinkel geeignet quantifizieren. Dabei darf sich der Gimbal-Lock-Effekt nicht auf das Ergebnis auswirken. Diese Anforderungen lassen sich erfüllen, indem die Positionen der Fingerspitzen im Raum relativ zu einem festen Bezugspunkt als Vergleichsgrößen herangezogen werden.

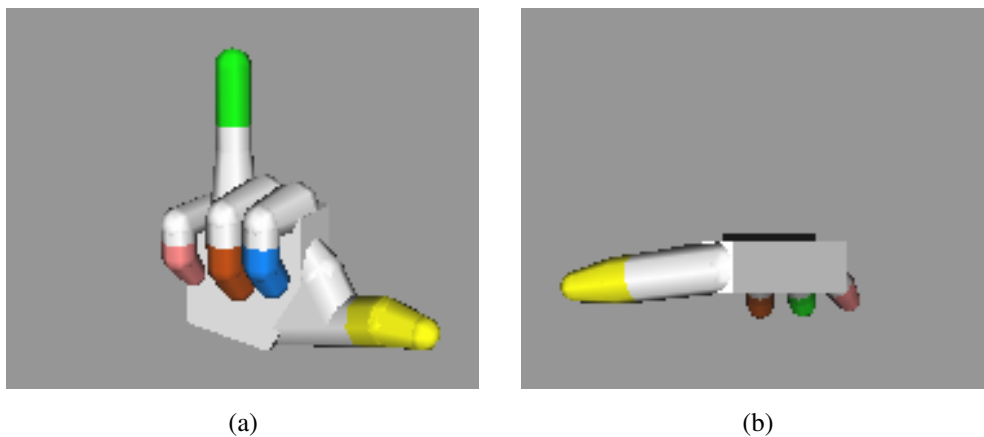


Abbildung 3.10: Handformen mit anatomisch unmöglichen Gelenkstellungen (a) sowie Handstellungen, die bei einer Frontalansicht nicht auftreten (b), werden aus dem Positorenraum ausgeblendet.

Abschnitt 3.3.1 definiert zunächst das Koordinatensystem des Handmodells, das die Grundlage der folgenden Berechnungen bildet. In Abschnitt 3.3.2 wird eine skalare Positorenendifferenz eingeführt, die Unterschiede hinsichtlich Handform und -stellung gemeinsam erfasst. Zwei weitere Maße, die eine getrennte Betrachtung von Form und Stellung erlauben, sind in den Abschnitten 3.3.3 bzw. 3.3.4 erläutert. Abschließend folgen in Abschnitt 3.3.5 mehrere Beispiele zur Veranschaulichung der verschiedenen Differenzmaße.

3.3.1 Koordinatensystem und Fingerspitzenpositionen

Das Handmodell verwendet ein Koordinatensystem, dessen Ursprung in der Mitte des Handtellers liegt. Die Orientierung der Achsen ist von der Handstellung unabhängig. In Abbildung 3.11 ist eine Handpositur mit überlagertem Koordinatensystem dargestellt.

Ausgehend von Handform und -stellung, beschrieben durch Beugungsgrade $\underline{B}$ bzw. Ansichtswinkel $\underline{\Phi}$, geschieht die Berechnung der 3D-Koordinaten der Fingerspitzen durch Hinterein-

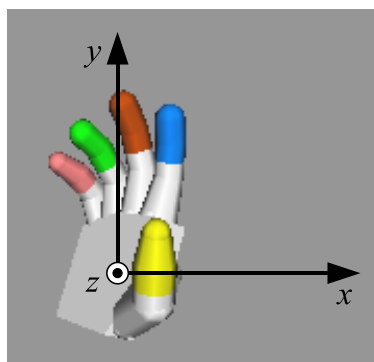


Abbildung 3.11: Handkoordinatensystem mit Ursprung im Mittelpunkt des Handtellers.

anderausführung von Translationen und Rotationen im Raum in derselben Weise, wie sie beim Zeichnen der Hand erfolgen. Mathematisch entspricht dies einer Abbildung Ψ , die aus $\underline{\mathbf{B}}$ und $\underline{\Phi}$ acht Punkte $\underline{\mathbf{c}}_i \in \mathbb{R}^3$ berechnet:

$$\Psi : \langle \underline{\mathbf{B}}, \underline{\Phi} \rangle \mapsto \langle \underline{\mathbf{c}}_1, \dots, \underline{\mathbf{c}}_8 \rangle \quad (3.6)$$

Dabei sind $\underline{\mathbf{c}}_1$ bis $\underline{\mathbf{c}}_5$ die Koordinaten der fünf Fingerspitzen, die somit die Handform charakterisieren. Die Punkte $\underline{\mathbf{c}}_6$ bis $\underline{\mathbf{c}}_8$ repräsentieren die Handstellung und befinden sich senkrecht über dem Handrücken, auf der Mittelachse der Hand in Richtung des Unterarms und seitlich des Handtellers gegenüber dem Daumen. Ihr Abstand vom Mittelpunkt des Handtellers entspricht der Länge des Mittelfingers. Abbildung 3.12 visualisiert alle acht Punkte.

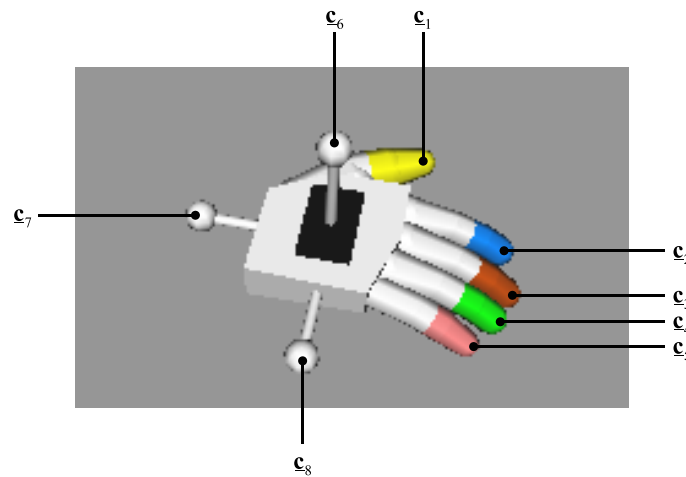


Abbildung 3.12: Lage der Punkte $\underline{\mathbf{c}}_1$ bis $\underline{\mathbf{c}}_8$, die zum Vergleich zweier Posituren verwendet werden.

Durch die Zentrierung des Handmodells im Koordinatenursprung ist die Abbildung von $\underline{\Phi}$ auf das 3-Tupel $\langle \underline{\mathbf{c}}_6, \underline{\mathbf{c}}_7, \underline{\mathbf{c}}_8 \rangle$ eindeutig. Je ähnlicher die beiden zu vergleichenden Ansichtswinkel sind, desto näher liegen die drei Punkte jeweils beieinander. Dies gilt insbesondere auch für Ansichten mit $\phi_{\text{lat}} \approx \pm 90^\circ$, die ähnliche oder identische Perspektiven darstellen, aufgrund des Gimbal-Lock-Effekts aber stark verschiedene Winkel ϕ_{lon} und ϕ_{rot} aufweisen können.

3.3.2 Posituredifferenz

Für zwei Posituren $\underline{\mathbf{P}}_1$ und $\underline{\mathbf{P}}_2$ mit den Handformen $\underline{\mathbf{B}}_1$ und $\underline{\mathbf{B}}_2$ sowie den Handstellungen $\underline{\Phi}_1$ und $\underline{\Phi}_2$ ergibt sich die skalare Posituredifferenz $\Delta_{\underline{\mathbf{P}}}$ aus den quadrierten euklidischen Abständen der durch Ψ erhaltenen Koordinaten. Dabei erfolgt eine Entkopplung von Form und Stellung, indem $\underline{\mathbf{P}}_2$ zum Vergleich der Handform ebenfalls unter dem Ansichtswinkel $\underline{\Phi}_1$ und lediglich zum Vergleich der Handstellung unter $\underline{\Phi}_2$ betrachtet wird:

$$\Psi(\underline{\mathbf{B}}_1, \underline{\Phi}_1) = \langle \underline{\mathbf{c}}_{1,1}, \dots, \underline{\mathbf{c}}_{1,8} \rangle \quad (3.7)$$

$$\Psi(\underline{\mathbf{B}}_2, \underline{\Phi}_2) = \langle \underline{\mathbf{c}}_{2,1}, \dots, \underline{\mathbf{c}}_{2,8} \rangle \quad (3.8)$$

$$\Psi(\underline{\mathbf{B}}_2, \underline{\Phi}_1) = \langle \underline{\mathbf{c}}'_{2,1}, \dots, \underline{\mathbf{c}}'_{2,8} \rangle \quad (3.9)$$

$$\Delta_{\underline{\mathbf{P}}}(\underline{\mathbf{P}}_1, \underline{\mathbf{P}}_2) = \underbrace{\sum_{i=1}^5 \|\underline{\mathbf{c}}_{1,i} - \underline{\mathbf{c}}'_{2,i}\|^2}_{\text{Formdifferenz}} + \underbrace{\sum_{i=6}^8 \|\underline{\mathbf{c}}_{1,i} - \underline{\mathbf{c}}_{2,i}\|^2}_{\text{Stellungsdifferenz}} \quad (3.10)$$

Auf diese Weise wird verhindert, dass sich Unterschiede in $\underline{\mathbf{B}}$ und $\underline{\Phi}$ hinsichtlich der Lage der Punkte $\underline{\mathbf{c}}_1$ bis $\underline{\mathbf{c}}_5$ gegenseitig kompensieren und so zu einer unangemessen niedrigen Formdifferenz führen.

Die Posituredifferenz nach (3.10) erfüllt die in (2.15) formulierte Bedingung

$$\Delta_{\underline{\mathbf{P}}}(\underline{\mathbf{P}}_1, \underline{\mathbf{P}}_2) = 0 \Leftrightarrow \underline{\mathbf{P}}_1 = \underline{\mathbf{P}}_2 \quad (3.11)$$

Sie ist ferner stabil, da kleine Änderungen in $\underline{\mathbf{B}}$ und/oder $\underline{\Phi}$ nur kleine Änderungen von $\Delta_{\underline{\mathbf{P}}}$ zur Folge haben. Aufgrund der Orthogonalität der Vektoren $\underline{\mathbf{c}}_6$, $\underline{\mathbf{c}}_7$ und $\underline{\mathbf{c}}_8$ sowie $\|\underline{\mathbf{c}}_6\| = \|\underline{\mathbf{c}}_7\| = \|\underline{\mathbf{c}}_8\|$ bewirkt eine Drehung der Hand um einen festen Winkel und eine beliebige Achse stets dieselbe Posituredifferenz.

3.3.3 Handformdifferenz

Der Vergleich zweier Handformen $\underline{\mathbf{B}}_1$ und $\underline{\mathbf{B}}_2$ beruht auf den euklidischen Distanzen der Daumen- und Fingerspitzen $\underline{\mathbf{c}}_1$ bis $\underline{\mathbf{c}}_5$. Im Unterschied zu $\Delta_{\underline{\mathbf{P}}}$ werden diese jedoch nicht quadriert, das Distanzmaß ist somit anschaulich nachvollziehbar. Zu beachten ist, dass ein Finger i (sowie der Daumen) in beiden Posituren zugleich verdeckt sein kann. Die zugehörigen Punkte $\underline{\mathbf{c}}_{1,i}$ und $\underline{\mathbf{c}}_{2,i}$ sind hier zwar definiert, eine eventuelle Abweichung ist für den Betrachter jedoch nicht sichtbar und somit für die Interpretation der Gebärde irrelevant. In diesem Fall wird die entsprechende Distanz zu Null angenommen. Abbildung 3.13 veranschaulicht die beschriebene Situation anhand eines Beispiels.

Wie auch bei der Posituredifferenz werden für den Vergleich der Handform dieselben Ansichtswinkel gesetzt, so dass Unterschiede hinsichtlich der Handstellung das Ergebnis nicht verfälschen (vgl. (3.7) und (3.9)). Damit resultiert zunächst die Definition der fünf Einzeldifferenzen hinsichtlich des Daumens und der vier Finger:

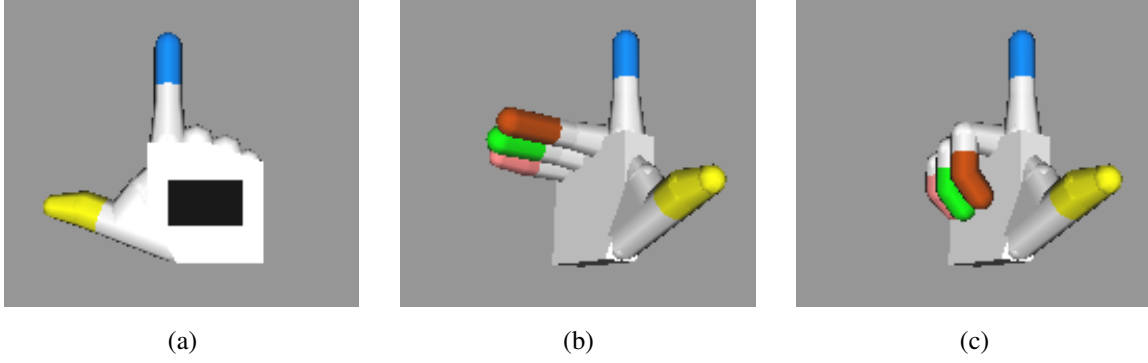


Abbildung 3.13: Aus der in (a) gezeigten Perspektive ist nicht erkennbar, ob die drei verdeckten Finger in den PIP- und DIP-Gelenken gestreckt (b) oder gebeugt (c) sind.

$$\Delta_{\underline{\mathbf{B}}, \mathbf{D}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2) = \begin{cases} 0 & \text{falls Daumen verdeckt} \\ \|\underline{\mathbf{c}}_{1,1} - \underline{\mathbf{c}}'_{2,1}\| & \text{sonst} \end{cases} \quad (3.12)$$

$$\vdots$$

$$\Delta_{\underline{\mathbf{B}}, \mathbf{KF}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2) = \begin{cases} 0 & \text{falls kleiner Finger verdeckt} \\ \|\underline{\mathbf{c}}_{1,5} - \underline{\mathbf{c}}'_{2,5}\| & \text{sonst} \end{cases} \quad (3.13)$$

$$\text{mit } \langle \underline{\mathbf{c}}_{1,1}, \dots, \underline{\mathbf{c}}_{1,5}, \dots \rangle = \Psi(\underline{\mathbf{B}}_1, \underline{\Phi}) \quad (3.14)$$

$$\langle \underline{\mathbf{c}}'_{2,1}, \dots, \underline{\mathbf{c}}'_{2,5}, \dots \rangle = \Psi(\underline{\mathbf{B}}_2, \underline{\Phi}) \quad (3.15)$$

Der Vektor $\underline{\Phi}$ hat auf die Abstände keinen Einfluss und ist daher beliebig wählbar.

Aus (3.12) bis (3.13) ergibt sich die über Finger und Daumen gemittelte Handformdifferenz $\bar{\Delta}_{\underline{\mathbf{B}}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2)$ zu

$$\bar{\Delta}_{\underline{\mathbf{B}}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2) = \frac{\Delta_{\underline{\mathbf{B}}, \mathbf{D}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2) + \dots + \Delta_{\underline{\mathbf{B}}, \mathbf{KF}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2)}{5} \quad (3.16)$$

Ergänzend wird ein weiteres Abweichungsmaß $\hat{\Delta}_{\underline{\mathbf{B}}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2)$ definiert als das Maximum der fünf Abstände $\Delta_{\underline{\mathbf{B}}, \mathbf{D}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2)$ bis $\Delta_{\underline{\mathbf{B}}, \mathbf{KF}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2)$:

$$\hat{\Delta}_{\underline{\mathbf{B}}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2) = \max \{ \Delta_{\underline{\mathbf{B}}, \mathbf{D}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2), \dots, \Delta_{\underline{\mathbf{B}}, \mathbf{KF}}(\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2) \} \quad (3.17)$$

Dieses Maß stellt sicher, dass Fehler bezüglich eines einzelnen Fingers nicht durch niedrige Abweichungen der übrigen Finger kompensiert werden.

3.3.4 Handstellungsdifferenz

Zur anschaulichen Quantifizierung der Handstellungsabweichung ist ein Winkelmaß geeignet, das die Orientierung der beiden Posituren relativ zueinander angibt. Aufgrund des Gimbal-

Lock-Effekts ergibt sich dieses Maß allerdings nicht direkt aus den Ansichtswinkeln, sondern ist wie folgt zu berechnen.

Zwei beliebige Handstellungen $\underline{\Phi}_1$ und $\underline{\Phi}_2$ – bzw. zwei beliebig orientierte Ansichtskugeln – lassen sich in jedem Fall durch eine einzige Rotation um eine geeignet gewählte Achse zur Deckung bringen. Im Handkoordinatensystem bedeutet dies, dass die durch $\underline{\Phi}_1$ und $\underline{\Phi}_2$ eindeutig definierten Punkte $\underline{c}_{1,6}$ bis $\underline{c}_{1,8}$ und $\underline{c}_{2,6}$ bis $\underline{c}_{2,8}$ paarweise aufeinander zu liegen kommen. Die Handstellungsdifferenz $\Delta_{\underline{\Phi}}(\underline{\Phi}_1, \underline{\Phi}_2)$ wird definiert als der Betrag des Drehwinkels dieser Rotation.

Mathematisch lässt sich die Rotation eines Punktes $\underline{p} \in \mathbb{R}^3$ um einen Einheitsvektor $\underline{x} = [x \ y \ z]^T$ um den Winkel α durch Multiplikation mit der Rotationsmatrix $\mathbf{R}$ darstellen:

$$\mathbf{R} = \begin{bmatrix} \cos \alpha + x^2(1 - \cos \alpha) & xy(1 - \cos \alpha) - z \sin \alpha & xz(1 - \cos \alpha) + y \sin \alpha \\ xy(1 - \cos \alpha) + z \sin \alpha & \cos \alpha + y^2(1 - \cos \alpha) & yz(1 - \cos \alpha) - x \sin \alpha \\ xz(1 - \cos \alpha) - y \sin \alpha & yz(1 - \cos \alpha) + x \sin \alpha & \cos \alpha + z^2(1 - \cos \alpha) \end{bmatrix} \quad (3.18)$$

Damit ergibt sich die Forderung

$$\underline{c}_{1,6} = \mathbf{R}\underline{c}_{2,6} \quad (3.19)$$

$$\underline{c}_{1,7} = \mathbf{R}\underline{c}_{2,7} \quad (3.20)$$

$$\underline{c}_{1,8} = \mathbf{R}\underline{c}_{2,8} \quad (3.21)$$

Durch zwei dieser drei Gleichungen sind die Rotationsparameter $\underline{x}$ und α eindeutig bestimmt. Die Definition der Handstellungsdifferenz ist somit

$$\Delta_{\underline{\Phi}}(\underline{\Phi}_1, \underline{\Phi}_2) = |\alpha| \quad \text{mit} \quad \alpha \in [-180^\circ, +180^\circ] \quad (3.22)$$

3.3.5 Beispiele

In den folgenden Abbildungen ist der Positurenvergleich anhand konkreter Beispiele veranschaulicht. Tabellen unter den Abbildungen geben jeweils die entsprechenden Differenzmaße an.

Zunächst verdeutlicht Abbildung 3.14 die Verwendung der euklidischen Abstände der Finger- und Daumenspitzen, d. h. der Punkte $\underline{c}_1$ bis $\underline{c}_5$, zum Vergleich der Handform gemäß Abschnitt 3.3.3. Abbildung 3.15 visualisiert die Berechnung der Handstellungsdifferenz durch Ausrichten der zu vergleichenden Posituren wie in Abschnitt 3.3.4 erläutert. In Abbildung 3.16 wird die Posituredifferenz anhand der Punkte $\underline{c}_1$ bis $\underline{c}_8$ gemessen (vgl. Abschnitt 3.3.3). Schließlich zeigt Abbildung 3.17 den Normalfall zweier leicht verschiedener Posituren und die daraus resultierenden typischen Werte der Differenzmaße.

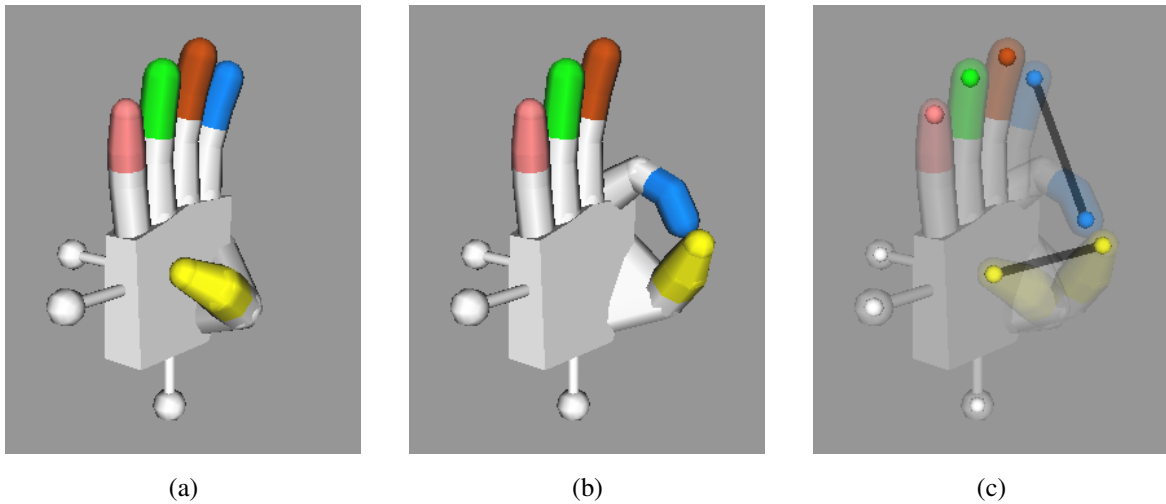


Abbildung 3.14: Berechnung der Handformdifferenz $\Delta_{\underline{B}}$ zweier Posituren (a, b) mit gleicher Handstellung. Die Handformen unterscheiden sich ausschließlich bezüglich Zeigefinger und Daumen, d. h. nur die Punkte $\underline{c}_1$ (Daumenspitze) und $\underline{c}_2$ (Zeigefingerspitze) sind verschieden, $\underline{c}_3$ bis $\underline{c}_8$ sind gleich (c). Zur Veranschaulichung sind der Differenzansicht (c) die Posituran-sichten (a, b) überlagert.

Tabelle 3.3: Differenzmaße der in Abbildung 3.14 verglichenen Posituren.

$\Delta_{\underline{P}}$ [cm ²]	$\Delta_{\underline{\Phi}}$ [°]	$\Delta_{\underline{B},D}$	$\Delta_{\underline{B},ZF}$	$\Delta_{\underline{B},MF}$ [cm]	$\Delta_{\underline{B},RF}$	$\Delta_{\underline{B},KF}$	$\hat{\Delta}_{\underline{B}}$ [cm]	$\hat{\Delta}_{\underline{B}}$ [cm]
132,89	0,00	6,83	9,29	0,00	0,00	0,00	3,22	9,29

Die Posituredifferenz nach (3.10) korreliert gut mit der visuellen Bewertung eines menschlichen Beobachters. Sie ist robust und aussagekräftig genug, um die Ähnlichkeit zweier Posituren zuverlässig zu quantifizieren. Durch die Zusammenfassung von Form- und Stellungsabweichung zu einem skalaren Wert ist $\Delta_{\underline{P}}$ zwar nicht anschaulich nachvollziehbar, jedoch gut für Optimierungsaufgaben innerhalb der Positurrekonstruktion geeignet. Beispiele hierfür sind die Mehrdeutigkeitsauflösung (s. Abschnitt 5.3.1) sowie die Optimierung verschiedener Systemparameter (s. Abschnitt 7.1).

Im Unterschied hierzu sind $\Delta_{\underline{B}}$ und $\Delta_{\underline{\Phi}}$ intuitiv interpretierbar. Diese Maße kommen daher bei der Prüfung der Anwendergebärde in Kapitel 6 sowie zur Evaluation der Positurerkennung in Kapitel 7 zum Einsatz.

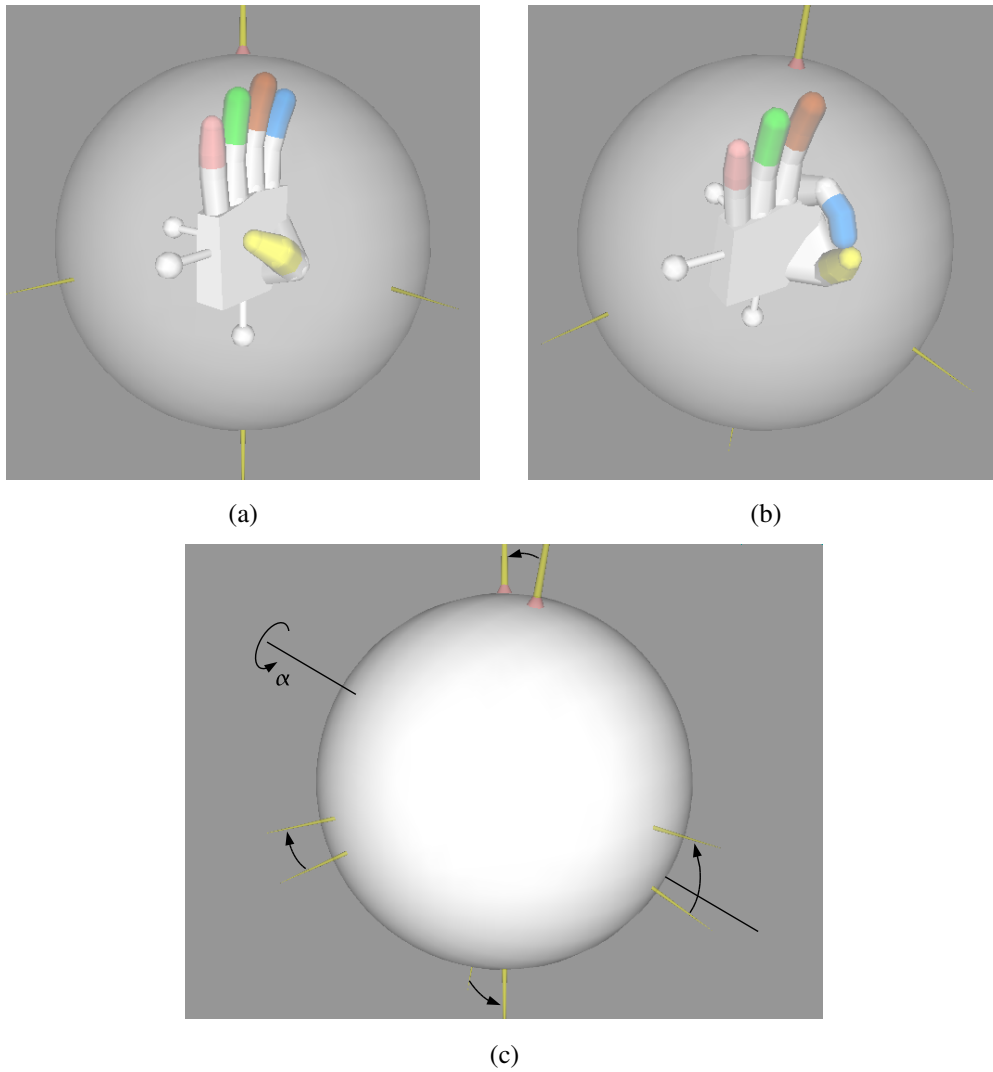


Abbildung 3.15: Berechnung der Handstellungsdifferenz $\Delta\Phi$ durch Ausrichten der zu vergleichenden Posituren. Den Posituren (a, b) ist die jeweilige Ansichtskugel überlagert, auf der die Durchstoßpunkte mit den Koordinatenachsen im unrotierten Zustand, d. h. für $\phi_{\text{lat}} = \phi_{\text{lon}} = \phi_{\text{rot}} = 0^\circ$, gekennzeichnet sind. Diese Punkte entsprechen dem „Nordpol“, dem „Südpol“ sowie vier äquidistanten Punkten auf dem „Äquator“. Durch Rotation der Ansichtskugel der zweiten Positur um eine gemäß Abschnitt 3.3.4 berechnete Achse um den Winkel α kommen alle Durchstoßpunkte zur Deckung (c), und die Handstellungsdifferenz ergibt sich zu $\Delta\Phi = |\alpha|$. Anschließend kann ein Vergleich der Handform wie in Abbildung 3.14 erfolgen.

Tabelle 3.4: Differenzmaße der in den Abbildung 3.15 verglichenen Posituren.

$\Delta_{\mathbf{P}}$ [cm ²]	Δ_{Φ} [°]	$\Delta_{\mathbf{B},D}$	$\Delta_{\mathbf{B},ZF}$	$\Delta_{\mathbf{B},MF}$ [cm]	$\Delta_{\mathbf{B},RF}$	$\Delta_{\mathbf{B},KF}$	$\bar{\Delta}_{\mathbf{B}}$ [cm]	$\hat{\Delta}_{\mathbf{B}}$ [cm]
164,18	26,13	6,83	9,29	0,00	0,00	0,00	3,22	9,29

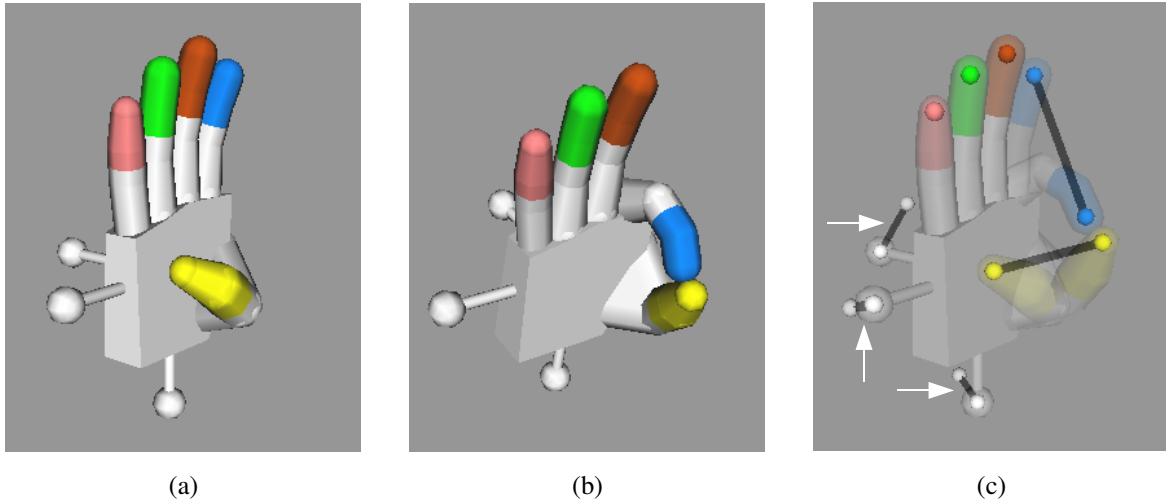


Abbildung 3.16: Berechnung der Posituredifferenz $\Delta_{\underline{P}}$ zweier Posituren (a, b) mit unterschiedlicher Form und Stellung. Zunächst erfolgt die Bestimmung von $\underline{c}_6$ bis $\underline{c}_8$ für beide Posituren sowie die Berechnung der jeweiligen Entfernungen (c; durch Pfeile gekennzeichnet). Anschließend werden die Posituren aneinander ausgerichtet, d. h. unter gleichem Ansichtswinkel betrachtet (c; überlagert dargestellt). Der weitere Vergleich entspricht Abbildung 3.14.

Tabelle 3.5: Differenzmaße der in den Abbildung 3.16 verglichenen Posituren.

$\Delta_{\underline{P}}$ [cm ²]	$\Delta_{\underline{\Phi}}$ [°]	$\Delta_{\underline{B},D}$	$\Delta_{\underline{B},ZF}$	$\Delta_{\underline{B},MF}$ [cm]	$\Delta_{\underline{B},RF}$	$\Delta_{\underline{B},KF}$	$\bar{\Delta}_{\underline{B}}$ [cm]	$\hat{\Delta}_{\underline{B}}$ [cm]
164,18	26,13	6,83	9,29	0,00	0,00	0,00	3,22	9,29

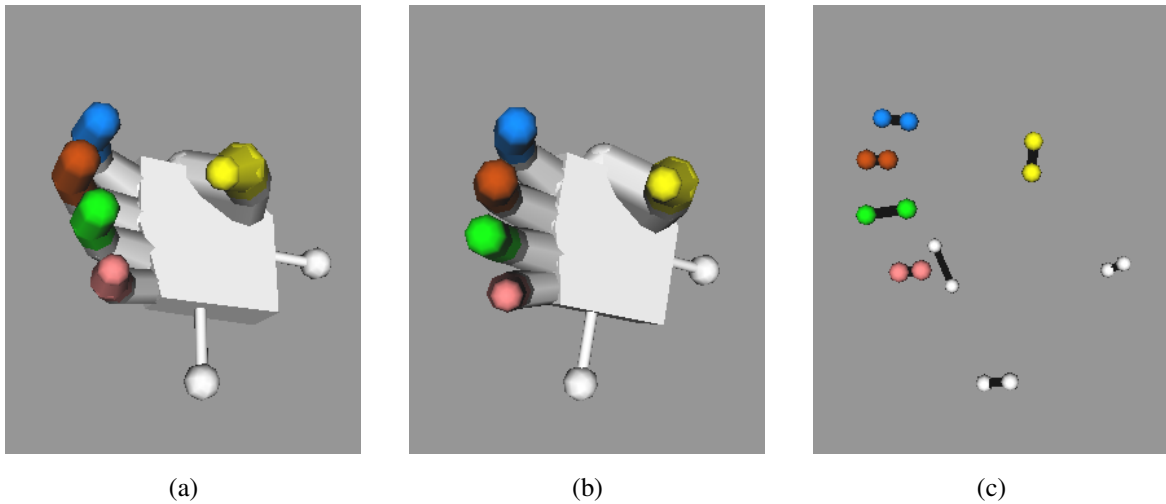


Abbildung 3.17: Vergleich zweier Posituren (a, b), die hinsichtlich Form und Stellung leicht verschieden sind. Dies ist der in der Verarbeitung am häufigsten auftretende Fall. Die euklidischen Abstände der Finger- und Daumenspitzen $\underline{c}_1$ bis $\underline{c}_5$ betragen etwa 1–2 cm (c), die Handstellungsdifferenz (in der Abbildung nicht eingezeichnet) liegt bei etwa 20° . Aus Gründen der Übersichtlichkeit wurde auf die Überlagerung der Posituransichten in (c) verzichtet.

Tabelle 3.6: Differenzmaße der in Abbildung 3.17 verglichenen Posituren.

$\Delta_{\underline{P}}$ [cm ²]	$\Delta_{\underline{\Phi}}$ [°]	$\Delta_{\underline{B},D}$	$\Delta_{\underline{B},ZF}$	$\Delta_{\underline{B},MF}$ [cm]	$\Delta_{\underline{B},RF}$	$\Delta_{\underline{B},KF}$	$\bar{\Delta}_{\underline{B}}$ [cm]	$\hat{\Delta}_{\underline{B}}$ [cm]
32,16	20,60	1,77	1,48	1,09	2,13	1,23	1,54	2,13

Kapitel 4

Segmentierung und Merkmalsextraktion

Die Aufgabe der Segmentierung besteht in der Unterteilung des Eingangsbildes in Regionen, die den im Bild repräsentierten realen Objekten entsprechen [SHB98]. Anschließend werden diese Regionen durch numerische Merkmale beschrieben und dadurch für die weitere Verarbeitung abstrahiert.

In der vorliegenden Arbeit erfolgt die Merkmalsextraktion für die auf den Handschuhen aufgebrachten Markierungen sowie das Gesicht des Gebärdenden, wobei jeweils unterschiedliche Verfahren zum Einsatz kommen. Alle übrigen Körperteile sowie Objekte im Bildhintergrund gehen nicht in die Analyse ein.

Abbildung 4.1 zeigt eine Übersicht der einzelnen Verarbeitungsschritte der Segmentierungs- und Merkmalsextraktionsstufe und gibt damit gleichzeitig den Inhalt dieses Kapitels an. Zunächst definiert Abschnitt 4.1 kurz die mathematische Beschreibung der Eingangsbilder. Abschnitt 4.2 hat die Hintergrundsubtraktion zum Gegenstand, die die Vorstufe der in Abschnitt 4.3 erläuterten Segmentierung der Markerregionen darstellt. Hier wird auch die Berechnung geometrischer Merkmale diskutiert. Danach skizziert Abschnitt 4.4 Segmentierung und Merkmalsextraktion für das Gesicht (linke Bildhälfte in Abbildung 4.1). Abschnitt 4.5 führt das Konzept der multiplen Hypothesen ein, das eine sichere Zuordnung der segmentierten Regionen zu den einzelnen Markern auch beim Vorhandensein gleichfarbiger Störobjekte erlaubt. Zuletzt befasst sich Abschnitt 4.6 mit der Normierung der Merkmale und dem Aufbau des Merkmalsvektors, der im folgenden Kapitel die Grundlage für die Handpositurerkennung bildet.

4.1 Eingangsbilder

Zur Akquisition der Bilddaten kommen preiswerte CCD-Kameras zum Einsatz, die über USB oder FireWire angeschlossen werden. Softwareseitig ist die aufgenommene Bildsequenz über die DirectShow-Schnittstelle des entsprechenden Kameratreibers direkt verfügbar.

Dieser Abschnitt führt kurz die zur Beschreibung der Eingangsbilder verwendete Notation ein

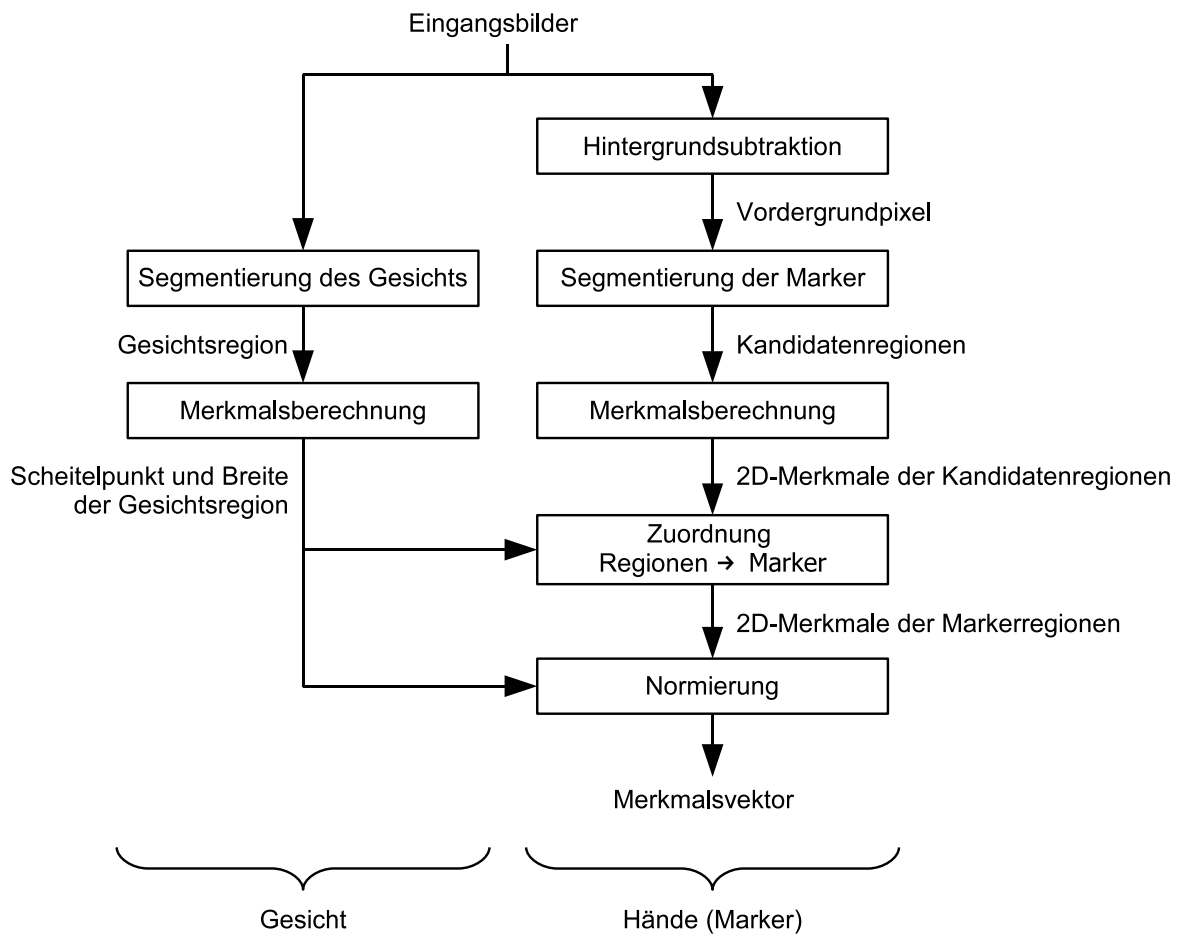


Abbildung 4.1: Schematische Darstellung von Segmentierung und Merkmalsextraktion.

und stellt verschiedene Systeme zur Farbdarstellung vor.

4.1.1 Notation

Ein einzelnes Bild mit einer Auflösung von N Zeilen und M Spalten lässt sich mathematisch als diskrete Funktion repräsentieren:

$$\underline{\mathbf{I}}(x, y) \quad \text{mit} \quad \begin{array}{l} x \in \{0, 1, \dots, M-1\} \\ y \in \{0, 1, \dots, N-1\} \end{array} \quad (4.1)$$

Das 2-Tupel $\langle x, y \rangle$ stellt Pixelkoordinaten mit dem Ursprung in der linken oberen Ecke dar. Der Wert der Funktion $\underline{\mathbf{I}}(x, y)$ beschreibt eine Eigenschaft des entsprechenden Pixels, beispielsweise seine Farbe, Intensität (Helligkeit), oder die Wahrscheinlichkeit, Teil eines gesuchten Objekts zu repräsentieren.

Die Beschreibung einer aus T Einzelbildern bestehenden Bildsequenz ergibt sich aus (4.1) durch Hinzufügen eines Zeitindex t :

$$\underline{\mathbf{I}}(x, y, t), \quad t = 0, 1, \dots, T - 1 \quad (4.2)$$

Der zeitliche Abstand zweier unmittelbar aufeinanderfolgender Bilder ist dabei der Kehrwert der bei der Aufnahme verwendeten Bildrate und beträgt bei 25 Hz entsprechend 40 ms. Für eine hinreichend genaue zeitliche Auflösung auch schneller Bewegungen stellt diese Bildrate das Minimum dar. Handelsübliche Webcams erreichen hierbei Auflösungen von 320×240 bis 640×480 .

Farbe wird in Abhängigkeit vom verwendeten Farbmodell als Tupel skalarer Werte repräsentiert; die Bildfunktion ist in diesem Fall vektorwertig. Bei Intensitäts- oder Wahrscheinlichkeitsbildern hingegen ist sie skalar und lässt sich daher – unabhängig von der beschriebenen Pixeleigenschaft – als Grauwertbild visualisieren. Wertdiskrete Bildfunktionen, die beispielsweise eine binäre pixelweise Klassifikation in Vordergrund und Hintergrund repräsentieren, werden als Maske bezeichnet, da sie häufig dem Zweck dienen, Bereiche des ursprünglichen Farbbildes (etwa den Bildhintergrund) auszublenden.

4.1.2 Farbrepräsentation

In der Computergrafik ist das RGB-Farbmodell weit verbreitet, da es der Funktionsweise von Bildsensoren und Monitoren entspricht. Es beschreibt eine Farbe $\underline{\mathbf{C}}$ als additive Mischung aus Rot-, Grün- und Blaukomponente:

$$\underline{\mathbf{C}}_{\text{RGB}} = \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (4.3)$$

Im RGB-Farbmodell wirkt sich die Intensität der Beleuchtung auf alle drei Kanäle aus. Diese Abhängigkeit ist für die Segmentierung nachteilig, da jede Helligkeitsänderung die Erscheinung der Objekte in $\underline{\mathbf{I}}$ verändert, ohne dass dabei Nutzinformationen übermittelt werden. So führt beispielsweise nichtdiffuse Beleuchtung dazu, dass die der Lichtquelle zugewandten Seiten eines Objekts in allen drei Farbkomponenten von den abgewandten Seiten abweichen und daher als separates Objekt erscheinen.

Durch eine geeignete Transformation lässt sich die Abhängigkeit von der Beleuchtungsintensität auf lediglich eine Farbkomponente beschränken. Diese kann dann in der weiteren Verarbeitung gesondert behandelt oder vollständig ignoriert werden.

Es existieren zahlreiche Farbmodelle, die diese Entkopplung anstreben. In der vorliegenden Arbeit kommt das 1976 von der CIE (*Commission Internationale de l'Eclairage*, franz. Internationale Beleuchtungskommission) vorgestellte CIE $L^*u^*v^*$ -Modell zum Einsatz, das neben einer Auskopplung der Helligkeitskomponente (L^* für *lightness*) auch eine Linearisierung bezüglich der Wahrnehmung approximiert: Die empfundene Verschiedenheit zweier Farben ist über den gesamten Wertebereich annähernd proportional zu deren numerischer Differenz.

Weiterhin definiert CIE $L^*u^*v^*$ Farben absolut, d. h. unter Berücksichtigung sämtlicher zur Reproduktion benötigter Elemente. Dies ist beim RGB-Farbmodell nicht gegeben, da die zur additiven Mischung benötigten Rot-, Grün- und Blaukomponenten selbst nicht spezifiziert sind und sich je nach Hardware unterscheiden können.

Die drei Farbkomponenten L^* , u^* und v^* berechnen sich aus R , G und B durch eine nicht-lineare Transformation sowie die Festlegung eines Weißpunkts. Dabei findet zunächst gemäß (4.4) eine lineare Transformation in den CIE-XYZ-Farbraum statt, der hier jedoch nicht näher betrachtet wird.

$$\underline{C}_{\text{XYZ}} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_{\text{D65}} = \begin{bmatrix} 0,412453 & 0,357580 & 0,180423 \\ 0,212671 & 0,715160 & 0,072169 \\ 0,019334 & 0,119193 & 0,950227 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (4.4)$$

Die Bezeichnung D65 steht für ein von der CIE definiertes Weiß, das etwa dem Mittagslicht in Europa entspricht. Für andere Weißdefinitionen ändern sich die Koeffizienten der Transformationsmatrix.

Aus den XYZ-Komponenten ergibt sich $\underline{C}_{\text{Luv}} = [L^* \ u^* \ v^*]^T$ zu

$$L^* = \begin{cases} 116 \sqrt[3]{\frac{Y}{Y_n}} - 16 & \text{für } \frac{Y}{Y_n} > 0,008856 \\ 903,3 \frac{Y}{Y_n} & \text{sonst} \end{cases} \quad (4.5)$$

$$u^* = 13L^*(u' - u'_n) \quad (4.6)$$

$$v^* = 13L^*(v' - v'_n) \quad (4.7)$$

$$\text{mit } u' = \begin{cases} \frac{4X}{X+15Y+3Z} & \text{für } X, Y, Z \neq 0 \\ 4 & \text{sonst} \end{cases} \quad (4.8)$$

$$v' = \begin{cases} \frac{9Y}{X+15Y+3Z} & \text{für } X, Y, Z \neq 0 \\ \frac{9}{15} & \text{sonst} \end{cases} \quad (4.9)$$

$$\text{und den Konstanten } Y_n = 1 \quad (4.10)$$

$$u'_n = 0,2009 \quad (4.11)$$

$$v'_n = 0,4610 \quad (4.12)$$

Die Kubikwurzel in (4.5) approximiert das logarithmische Helligkeitsempfinden des menschlichen Auges. Die konstant angenommenen Werte für Y_n , u'_n und v'_n beschreiben die verwendete Lichtquelle.

4.2 Hintergrundsubtraktion

Häufig enthält das Kamerabild neben den interessierenden Objekten im Vordergrund auch einen signifikanten Anteil Hintergrund. Dieser stellt für die Merkmalsextraktion, die eine

Beschreibung der Vordergrundobjekte zum Ziel hat, ein Hindernis dar. Hintergrundobjekte können beispielsweise mit Vordergrundobjekten verwechselt werden oder bei ähnlicher Farbe den Kontrast an den Objektkanten reduzieren. Hintergrundbereiche bereits in der ersten Verarbeitungsstufe zu erkennen und auszublenden, reduziert daher für alle folgenden Stufen die zu betrachtende Datenmenge sowie die Fehlerwahrscheinlichkeit. Dieses Vorgehen erfordert zunächst eine Modellierung des Hintergrunds. Weil zum Zeitpunkt der Segmentierung noch keine semantische Interpretation des Bildes vorliegt, arbeitet die Hintergrundmodellierung pixelbasiert und ohne Verwendung von Domänenwissen.

Trotz der in Abschnitt 4.1.2 beschriebenen Vorteile des CIE $L^*u^*v^*$ -Farbmodells findet hier das RGB-Farbsystem Verwendung. Die Wahrnehmungslinearität des CIE $L^*u^*v^*$ -Systems bietet bei der Hintergrundmodellierung keine Vorteile, und die Entkopplung der Helligkeitskomponente kann auch im RGB-Modell angenähert werden.

Im Folgenden ist von Bedeutung, dass das von einer Digitalkamera gelieferte Bild prinzipbedingt einen Rauschanteil enthält. Dieser wird u. a. durch thermoelektrische Effekte im Bildsensor verursacht und ist umso stärker, je geringer die Fläche eines einzelnen Pixels auf dem Sensor und je höher die Temperatur des Sensors ist. Ein Reduzieren des Lichteinfalls, beispielsweise durch Verkürzen der Belichtungszeit bei gleichbleibender Beleuchtung, erhöht den relativen Anteil des Rauschens am Bildsignal.

4.2.1 Unimodale Modellierung

Unimodale Verfahren modellieren die Wahrscheinlichkeitsverteilung der diskreten Farbwerte $\underline{\mathbf{C}}$ eines Pixels in einer Sequenz von Eingangsbildern statistisch auf Grundlage einer dreidimensionalen Normalverteilung im RGB-Farbraum:

$$P(\underline{\mathbf{C}}) = \eta(\underline{\mathbf{C}}, \underline{\mu}, \underline{\Sigma}) = \frac{1}{\sqrt{(2\pi)^3 \det \underline{\Sigma}}} \exp\left(-\frac{1}{2}(\underline{\mathbf{C}} - \underline{\mu})^T \underline{\Sigma}^{-1} (\underline{\mathbf{C}} - \underline{\mu})\right) \quad (4.13)$$

Dieser Ansatz ist durch die Tatsache motiviert, dass die Pixelfarbe auch bei statischem Bildinhalt nicht konstant ist, sondern zeitlichen Schwankungen unterliegt, die primär durch das oben beschriebene Kamerarauschen verursacht werden. Sind Beleuchtung, Kameraposition und Hintergrund konstant, modelliert (4.13) somit direkt diesen Rauschprozess.

Zur Schätzung der Verteilungsparameter $\underline{\mu}$ und $\underline{\Sigma}$, die Mittelwert bzw. Kovarianzmatrix der Farbwerte bezeichnen, existieren verschiedene Ansätze. Eine einfache Methode besteht darin, manuell alle Vordergrundobjekte aus dem Aufnahmebereich zu entfernen, so dass $\underline{\mu}$ und $\underline{\Sigma}$ für jedes Pixel unmittelbar gemessen werden können [WADP97]. Ändert sich die Beleuchtung oder der Hintergrund, muss die Messung wiederholt werden. Dieser Ansatz ist aufwändig und für den praktischen Einsatz daher ungeeignet.

Um eine dedizierte Vorbereitungsphase zu vermeiden, ist das Hintergrundmodell aus den Bildern der Eingangssequenz selbst zu berechnen. Hierzu ist die Bildsequenz komplett oder teilweise zwischenspeichern, bevor eine weitere Verarbeitung erfolgt. Für Endlossequenzen

eignet sich eine *Sliding-Window*-Verarbeitung, die die n zuletzt aufgenommenen Bilder – ggf. mit abnehmender Gewichtung – berücksichtigt.

Im Folgenden gibt n allgemein die Anzahl der zur Modellbildung herangezogenen Einzelbilder für endliche oder nichtendliche Bildsequenzen an. Ferner wird die Hintergrundfarbe eines Pixels definiert als die Farbe, die das Pixel in der überwiegenden Anzahl der n Bilder aufweist. Da die Bilder auch Vordergrundobjekte enthalten, die in diesem Zusammenhang statistische Ausreißer darstellen, ist der arithmetische Mittelwert keine gute Approximation von $\underline{\mu}$. Eine robustere Lösung besteht darin, unter den beobachteten Farbwerten $\{\underline{\mathbf{C}}(1), \underline{\mathbf{C}}(2), \dots, \underline{\mathbf{C}}(n)\}$ denjenigen auszuwählen, für den die Summe der euklidischen Abstände zu allen anderen Werten minimal ist [Bal04, ZK05, ZC06]:

$$\underline{\mu} = \underset{\underline{\mathbf{C}}(k)}{\operatorname{argmin}} \sum_{i=1}^n \|\underline{\mathbf{C}}(i) - \underline{\mathbf{C}}(k)\| \quad \text{mit } k \in \{1, 2, \dots, n\} \quad (4.14)$$

Diese Gleichung entspricht im Eindimensionalen dem Median. Die kanalweise Bestimmung des Medians,

$$\underline{\mu} = \begin{bmatrix} \operatorname{median}\{R(1), R(2), \dots, R(n)\} \\ \operatorname{median}\{G(1), G(2), \dots, G(n)\} \\ \operatorname{median}\{B(1), B(2), \dots, B(n)\} \end{bmatrix} \quad (4.15)$$

führt im Allgemeinen nicht zum gleichen Ergebnis, vermeidet jedoch die aufwändige Berechnung der $n(n-1)/2$ Abstände in (4.14) und stellt in der Praxis durchgängig eine gute Näherung dar [Bal04].

Nach der Schätzung des Farbmittelwerts $\underline{\mu}$ erlauben zwei empirisch motivierte Vereinfachungen die unmittelbare und effiziente Anwendung des Hintergrundmodells. Zunächst wird die Kovarianzmatrix Σ durch das Produkt der Einheitsmatrix mit einer skalaren Standardabweichung σ angenähert:

$$\Sigma = \mathbf{1} \cdot \sigma \quad (4.16)$$

Dies entspricht der Annahme, dass das Rauschen auf den drei Farbkanälen statistisch unabhängig, gleich stark ausgeprägt und unabhängig vom Farbwert ist.

Daneben wird die Berechnung der zur Klassifikation eines Farbwerts $\underline{\mathbf{C}}$ benötigten bedingten Hintergrundwahrscheinlichkeit $P(\text{Hg}|\underline{\mathbf{C}})$ vereinfacht. Mit den A-priori-Wahrscheinlichkeiten für Hintergrund und Vordergrund, $P(\text{Hg}) + P(\text{Vg}) = 1$, ergibt sich nach Bayes:

$$P(\text{Hg}|\underline{\mathbf{C}}) = \left(1 + \frac{P(\underline{\mathbf{C}}|\text{Vg})}{P(\underline{\mathbf{C}}|\text{Hg})} \cdot \frac{P(\text{Vg})}{P(\text{Hg})} \right)^{-1} \quad (4.17)$$

Als Hintergrund werden alle Pixel klassifiziert, für die (4.17) einen Schwellwert Θ_1 übersteigt. Da der Quotient aus $P(\text{Vg})$ und $P(\text{Hg})$ lediglich eine nichtlineare Skalierung (d. h.

eine streng monotone Abbildung) von $P(\text{Hg}|\underline{\mathbf{C}})$ bewirkt, kann er bei entsprechender Skalierung des Schwellwerts vereinfachend zu Eins angenommen werden.

Ohne Kenntnis der Vordergrundfarbverteilung besteht keine Grund, einzelne Farben häufiger als andere zu erwarten. Daher wird für $P(\underline{\mathbf{C}}|\text{Vg})$ vereinfachend eine Gleichverteilung angenommen, die wiederum eine nichtlineare Skalierung der rechten Seite von (4.17) zur Folge hat und somit implizit im Schwellwert ausgedrückt werden kann. Aus

$$\left(1 + \frac{P(\underline{\mathbf{C}}|\text{Vg})}{P(\underline{\mathbf{C}}|\text{Hg})} \cdot \frac{P(\text{Vg})}{P(\text{Hg})}\right)^{-1} > \Theta_1 \quad (4.18)$$

ergibt sich somit

$$\left(1 + \frac{1}{P(\underline{\mathbf{C}}|\text{Hg})}\right)^{-1} > \Theta_2 \quad (4.19)$$

und nach weiterer Vereinfachung

$$P(\underline{\mathbf{C}}|\text{Hg}) > \Theta_3 \quad (4.20)$$

wobei beide Seiten in (4.19) und (4.20) jeweils durch eine streng monotone Abbildung aus (4.18) bzw. (4.19) hervorgehen, die das Klassifikationsergebnis nicht beeinflusst. Mit (4.13) und (4.16) lässt sich $P(\underline{\mathbf{C}}|\text{Hg})$ als Funktion von $\underline{\mathbf{C}}$ schreiben:

$$\frac{1}{\sigma\sqrt{(2\pi)^3}} \exp\left(-\frac{1}{2} \frac{(\underline{\mathbf{C}} - \underline{\mu})^2}{\sigma^2}\right) > \Theta_3 \quad (4.21)$$

Zur Steigerung der Laufzeiteffizienz wird nicht jeder einzelne quadrierte Farbabstand $(\underline{\mathbf{C}} - \underline{\mu})^2$ transformiert, sondern einmalig die inverse Transformation auf den Schwellwert angewandt:

$$(\underline{\mathbf{C}} - \underline{\mu})^2 < -2\sigma^2 \ln\left(\Theta_3 \sigma \sqrt{(2\pi)^3}\right) = \Theta^2 \quad (4.22)$$

Somit erfolgt die unimodale Hintergrundmodellierung durch Schätzung der Hintergrundfarbe gemäß (4.14) und Festlegung eines geeigneten Schwellwerts Θ bezüglich des euklidischen Abstands von dieser Farbe. Im Farbraum definiert (4.22), wie in Abbildung 4.2 gezeigt, eine Kugel mit dem Radius Θ , die beispielsweise für $\Theta = 2,5\sigma$ ca. 98% der Hintergrundverteilung enthält.

Die in (4.22) berechnete Differenz berücksichtigt gleichermaßen Unterschiede in Chromatizität und Helligkeit. Da Helligkeitsschwankungen häufig von Schatten verursacht werden und nicht auf eine Objektbewegung zurückgehen, ist es vorteilhaft, diese beiden Komponenten der Farbdifferenz getrennt zu betrachten und bezüglich der Helligkeit größere Schwankungen zuzulassen. Dadurch sinkt die Anzahl der aufgrund von Schatten fälschlicherweise als Vordergrund klassifizierten Hintergrundpixel.

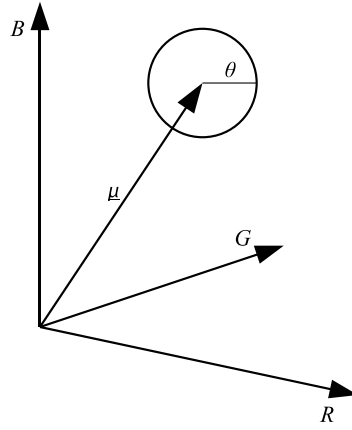


Abbildung 4.2: Hintergrundfarbmodell eines einzelnen Pixels. Der Farbwert $\underline{\mu}$ gibt den Mittelpunkt einer Kugel mit dem Radius Θ an, die den Bereich der Hintergrundfarbe definiert.

Helligkeitsänderungen lassen sich im RGB-Farbmodell durch eine einfache Skalarmultiplikation approximieren. Farb- und Helligkeitsdifferenz Δ_c bzw. Δ_b ergeben sich daher durch Zerlegen des Farbvektors $\underline{C}$ in eine Parallel- und eine Orthogonalkomponente zu $\underline{\mu}$. Mit der Projektion von $\underline{C}$ auf $\underline{\mu}$,

$$p = \frac{\underline{C} \cdot \underline{\mu}}{\|\underline{\mu}\|} \quad (4.23)$$

berechnen sich Δ_c und Δ_b somit zu

$$\Delta_c = \|\underline{C}\|^2 - p^2 \quad (4.24)$$

$$\Delta_b = p - \|\underline{\mu}\| \quad (4.25)$$

Durch Festlegen zweier Schwellwerte Θ_c und Θ_b mit $\Theta_c < \Theta_b$ entsteht ein zylindrisches Volumen im RGB-Raum, das die dem Hintergrundfarbwert $\underline{\mu}$ zuzuordnenden Hintergrundfarben definiert. In Abbildung 4.3 ist dieses Volumen dargestellt.

Abbildung 4.4 zeigt die Anwendung des unimodalen Hintergrundmodells auf eine Sequenz von 54 Bildern. Neben der gebärdenden Person ist in jedem Bild mindestens eine weitere Person im Hintergrund sichtbar (a). Das Hintergrundmodell (b) visualisiert für jedes Pixel dessen Hintergrundfarbwert $\underline{\mu}$ entsprechend (4.14). Es ist zu erkennen, dass die Modellierung die statischen Bildbereiche präzise erfasst, obwohl keines der ausgewerteten Einzelbilder frei von Störobjekten ist. Die Anwendung des Modells auf ein Beispielbild (c) der Sequenz ergibt eine binäre Maske (d), in der Vordergrundbereiche weiß gekennzeichnet sind. Dies beinhaltet auch die Person in der linken Bildhälfte, die sich zwar räumlich im Hintergrund befindet, aufgrund ihrer Bewegung aber ein Vordergrund- bzw. Störobjekt darstellt.

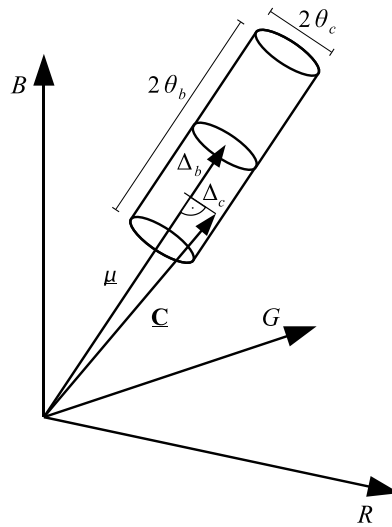


Abbildung 4.3: Hintergrundfarbmodell mit Schattenerkennung. Der zu $\underline{\mu}$ parallele Zylinder definiert die als Hintergrund klassifizierten Farbwerte.



(a) Eingangssequenz



(b) Hintergrundmodell

(c) Beispielbild

(d) Segmentierung

Abbildung 4.4: Vier der 54 Einzelbilder der Eingangssequenz (a), das Hintergrundmodell nach (4.14) (b) sowie ein Beispielbild aus der Sequenz (c) und dessen Segmentierung gemäß (4.22) (d).

4.2.2 Multimodale Modellierung

Die im letzten Abschnitt beschriebene Modellierung ist zuverlässig, wenn der Hintergrund seine Farbe nicht ändert und in mehr als 50% der Bilder sichtbar ist. Wechselt ein Hintergrundpixel jedoch zwischen verschiedenen Farben, ist ein unimodales Modell nicht geeignet. Beispiele für solche dynamischen Hintergründe sind grüne Blätter, die sich vor einer weißen Wand bewegen, sowie flimmernde CRT-Monitore.

Ein multimodales Hintergrundmodell ersetzt die unimodale Gaußverteilung aus (4.13) gegen eine Gaußsche Mischverteilung (*gaussian mixture model*, GMM). Somit ergibt sich für die Wahrscheinlichkeitsverteilung

$$P(\underline{\mathbf{C}}) = \sum_{k=1}^K \omega_k \eta(\underline{\mathbf{C}}, \underline{\mu}_k, \underline{\Sigma}_k) \quad (4.26)$$

wobei ω_k der Gewichtung der k -ten Gaußkomponente entspricht. Da $\sum_{\underline{\mathbf{C}}} P(\underline{\mathbf{C}}) = 1$ gelten muss, folgt für die Gewichte

$$\sum_{\underline{\mathbf{C}}} P(\underline{\mathbf{C}}) = \sum_{k=1}^K \omega_k \sum_{\underline{\mathbf{C}}} \eta(\underline{\mathbf{C}}, \underline{\mu}_k, \underline{\Sigma}_k) = \sum_{k=1}^K \omega_k = 1 \quad (4.27)$$

Nicht alle Verteilungskomponenten repräsentieren notwendigerweise Hintergrundfarben. Vordergrundobjekte, die ausreichend lange sichtbar sind, gehen ebenfalls in die Modellierung ein und werden später heuristisch von den Hintergrundkomponenten getrennt. Der vollständige Algorithmus zur multimodalen Hintergrundmodellierung findet sich in Anhang A.

4.3 Segmentierung der Marker

Nach der Hintergrundsubtraktion liegt eine binäre Maske vor, die die bewegten Bildbereiche enthält. Diese werden mit Hilfe des Mean-Shift-Algorithmus weiter partitioniert und die so erhaltenen Regionen aufgrund ihrer Farbe entweder als Markerkandidat oder als Hintergrundobjekt klassifiziert. Dieser Vorgang ist Gegenstand der beiden Abschnitte 4.3.1 und 4.3.2.

Danach erfolgt für alle Kandidatenregionen die Berechnung zweidimensionaler geometrischer Merkmale. Hierfür ist zunächst eine topologische Regionenbeschreibung erforderlich, die in Abschnitt 4.3.3 eingeführt wird. Nach der in Abschnitt 4.3.4 erläuterten Merkmalsberechnung beschreibt Abschnitt 4.3.5 eine elliptische Approximation der Regionen, die für die in Kapitel 5 behandelte Positurerkennung vorteilhaft ist.

4.3.1 Mean-Shift-Clustering

Der Mean-Shift-Algorithmus segmentiert ein Eingangsbild durch Clustern sämtlicher Pixel in einem Merkmalsraum [CM02]. Die Pixel eines Clusters bilden jeweils eine Region im Bild. Für jedes Pixel wird im Merkmalsraum ein lokaler Gradient – der Mean-Shift-Vektor – berechnet. Diesen Gradienten verfolgt das Verfahren so lange, bis ein lokales Maximum und damit ein Modalwert der Verteilung erreicht ist. Jeder Modalwert repräsentiert ein Cluster bzw. eine Bildregion.

Zunächst wird die wertdiskrete Verteilung H der Elemente (d. h. Pixel) $\underline{\mathbf{x}}_i$ im Merkmalsraum durch eine Kerneldichteschätzung in eine wertkontinuierliche, stetig differenzierbare Verteilung

lung p umgewandelt. Dies geschieht durch Multiplikation jedes Elements mit einer noch zu definierenden Kernelfunktion K . Sei n die Anzahl der Elemente, d die Dimensionalität des Merkmalsraums und h ein Bandbreitenparameter der Kernelfunktion, so ergibt sich die Kerndichteschätzung zu

$$p(\underline{\mathbf{x}}) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\underline{\mathbf{x}} - \underline{\mathbf{x}}_i}{h}\right) \quad (4.28)$$

Für die Modalwerte der Verteilung gilt $\nabla p(\underline{\mathbf{x}}) = 0$. Das Mean-Shift-Verfahren ermittelt diese Punkte mit Hilfe der Ableitungen der Kernelfunktion. Eine Berechnung von (4.28) ist nicht erforderlich.

Geeignet sind Kernel der Form

$$K(\underline{\mathbf{x}}) = \alpha k(\|\underline{\mathbf{x}}\|^2) \quad (4.29)$$

mit einer Funktion $k(x)$, die das „Profil“ eines einzelnen Elements festlegt, und einer Normalisierungskonstanten α . Das Epanechnikovprofil

$$k_E(x) = \begin{cases} 1 - x & \text{für } x \leq 1 \\ 0 & \text{sonst} \end{cases} \quad (4.30)$$

nähert die reale Verteilung für $n \rightarrow \infty$ im Sinne eines mittleren quadratischen Fehlers optimal an. Häufig wird jedoch das mathematisch einfacher zu handhabende Profil

$$k_N(x) = \exp\left(-\frac{1}{2}x\right) \quad (4.31)$$

verwendet, das den Gaußkernel

$$K_N(\underline{\mathbf{x}}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2} \|\underline{\mathbf{x}}\|^2\right) \quad (4.32)$$

ergibt. Abbildung 4.5 zeigt exemplarisch die Anwendung dieses Kernels auf eine eindimensionale Verteilung sowie den Einfluss der Bandbreite h auf das Ergebnis der Clusters.

Mit der Ableitung der Profilkfunktion $g(x) = -k'(x)$ und einer weiteren Normalisierungskonstanten β ergibt sich ein neuer Kernel $G(\underline{\mathbf{x}})$ zu

$$G(\underline{\mathbf{x}}) = \beta g(\|\underline{\mathbf{x}}\|^2) \quad (4.33)$$

Der Mittelwert der Elemente der Verteilung, gewichtet mit dem Kernel $G(\underline{\mathbf{x}})$, zentriert im Punkt $\underline{\mathbf{x}}$, ist

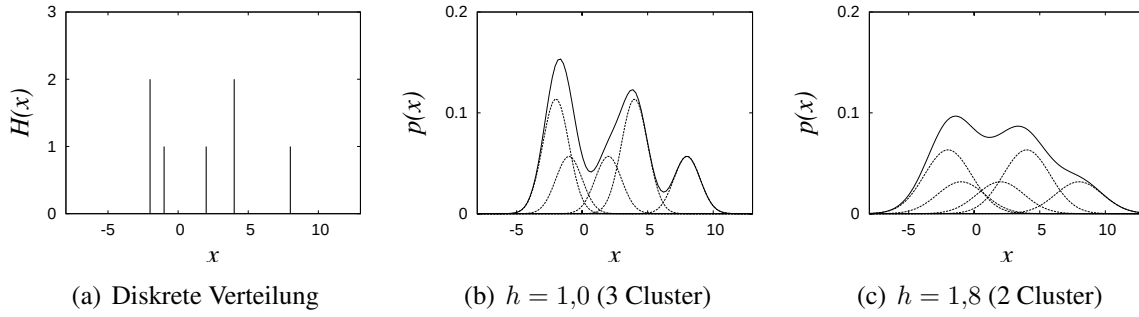


Abbildung 4.5: Diskrete Verteilung (a) und kontinuierliche Approximation mittels Gaußkernel gemäß (4.32) (b, c). Zur Veranschaulichung sind gestrichelt die einzelnen Summanden von $p(x)$ dargestellt. Für die Modalwerte ist die Ableitung $p'(x) = 0$.

$$\underline{\mu}_{h,G}(\underline{\mathbf{x}}) = \frac{\sum_{i=1}^n \underline{\mathbf{x}}_i g\left(\left\|\frac{\underline{\mathbf{x}} - \underline{\mathbf{x}}_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{\underline{\mathbf{x}} - \underline{\mathbf{x}}_i}{h}\right\|^2\right)} \quad (4.34)$$

Hieraus wird der Mean-Shift-Vektor $\underline{\mathbf{m}}_{h,G}(\underline{\mathbf{x}})$ berechnet, dessen Richtung der maximalen Steigung der Verteilung entspricht:

$$\underline{\mathbf{m}}_{h,G}(\underline{\mathbf{x}}) = \underline{\mu}_{h,G}(\underline{\mathbf{x}}) - \underline{\mathbf{x}} \quad (4.35)$$

Für einen gegebenen Punkt $\underline{\mathbf{x}}$ im Merkmalsraum approximiert das Mean-Shift-Verfahren den zugehörigen Modalwert $\underline{\mathbf{d}}$ durch fortgesetzte Addition des Mean-Shift-Vektors:

$$\underline{\mathbf{d}}^{(0)} = \underline{\mathbf{x}} \quad (4.36)$$

$$\underline{\mathbf{d}}^{(i)} = \underline{\mathbf{d}}^{(i-1)} + \underline{\mathbf{m}}_{h,G}(\underline{\mathbf{d}}^{(i-1)}) = \underline{\mu}_{h,G}(\underline{\mathbf{d}}^{(i-1)}) \quad i = 1, 2, \dots \quad (4.37)$$

Diese Reihe konvergiert für jedes konvexe und monoton fallende Profil $k(x)$ [CM02].

Zur Farbbildsegmentierung ist ein Merkmalsraum geeignet, der Farbe $\underline{\mathbf{C}}$ und Position $\underline{\mathbf{p}}$ eines Pixels kombiniert [CM02]. Aufgrund der unterschiedlichen Natur dieser beiden Merkmale werden diese im Kernel entsprechend separiert:

$$K_{h_p, h_C} = \frac{\gamma}{h_p^2 h_C^3} k\left(\left\|\frac{\underline{\mathbf{p}}}{h_p}\right\|^2\right) k\left(\left\|\frac{\underline{\mathbf{C}}}{h_C}\right\|^2\right) \quad (4.38)$$

Darin ist γ erneut eine Normierungskonstante. Die Multiplikation der beiden Profildfunktionen entspricht aufgrund (4.31) der Addition der Abweichungen bezüglich Position und Farbe, jeweils normiert auf die entsprechende Bandbreite h_p bzw. h_C . Der verwendete Algorithmus benutzt das in Abschnitt 4.1.2 beschriebene CIE $L^*u^*v^*$ -Farbmodell. Da für die Segmentierung der Markerflächen keine feinen Strukturen aufzulösen sind, werden vergleichsweise große

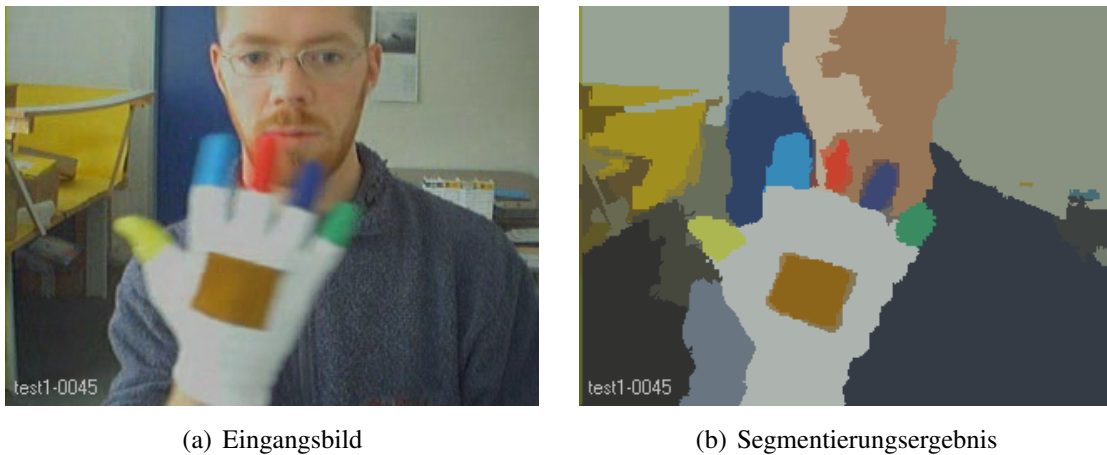


Abbildung 4.6: Anwendung des Mean-Shift-Segmentierungsalgorithmus auf ein Farbbild (a). Die erhaltenen Regionen sind in der Farbe des jeweils zugehörigen Modalwerts dargestellt (b). Die Bandbreiten wurden empirisch zu $h_C = 15$ und $h_p = 17$ bestimmt (bei einem Wertebereich von $\{0, \dots, 255\}$) und sind für beliebige Aufnahmesituationen geeignet.

Bandbreiten gewählt. Abbildung 4.6 zeigt ein Beispiel für die Anwendung des Mean-Shift-Algorithmus. Die Markerregionen sind trotz deutlicher Bewegungsunschärfe korrekt segmentiert.

4.3.2 Farbbasierte Regionenklassifikation

Nach Anwendung des Mean-Shift-Algorithmus ist das Eingangsbild vollständig in Regionen unterteilt, die jeweils einen Cluster darstellen und dem entsprechenden Modalwert – d. h. Farbwert – zugeordnet sind. Im nächsten Schritt sind diejenigen Regionen zu identifizieren, deren Farbe einer der Markerfarben entspricht. Alle übrigen Regionen werden nicht weiter verarbeitet.

Zur Beschreibung der Markerfarben dienen Histogramme im RGB-Farbraum, die vor der Anwendung des Systems manuell erstellt werden. Hierzu ist es ausreichend, die Markerregionen in 5–10 nicht aufeinanderfolgenden Einzelbildern einer Kalibrierungssequenz zu maskieren, die unter ähnlichen Beleuchtungsverhältnissen aufgezeichnet wurde (s. Abbildung 4.7).

Ergebnis dieses Schritts sind pro Marker i zwei Histogramme, $H_i(\underline{C})$ und $H_{-i}(\underline{C})$, die die Häufigkeit der Farbwerte $\underline{C} = [R \ G \ B]^T$ in den Objekt- bzw. Nicht-Objektpixeln angeben. Die zur Berechnung eines repräsentativen Histogramms benötigte Anzahl Farbwerte lässt sich durch Extrapolation gering halten, indem das Histogramm mit einer dreidimensionalen Gaußfunktion gefaltet wird.

Die Farbwerte sind mit je 8 Bit quantisiert, so dass $R, G, B \in \{0, \dots, 255\}$ gilt. Bei Verwendung von 4 Byte zur Darstellung der Häufigkeiten wären für jedes Histogramm $256^3 \cdot 4 \text{ Byte} = 64 \text{ MB}$ erforderlich. Um diesen Speicherplatzbedarf zu reduzieren, werden die Farbkanäle in den Histogrammen lediglich mit je 5 Bit quantisiert. Damit sind pro Histogramm nur



Abbildung 4.7: Manuelle Maskierung der Markerregionen in einer Kalibrierungssequenz.

$(2^5)^3 \cdot 4 \text{ Byte} = 128 \text{ kB}$ nötig.

Mit der Anzahl der Einträge $n_i = \sum_{\underline{\mathbf{C}}} H_i(\underline{\mathbf{C}})$ und $n_{\neg i} = \sum_{\underline{\mathbf{C}}} H_{\neg i}(\underline{\mathbf{C}})$ ergeben sich die bedingten Wahrscheinlichkeiten $P(\underline{\mathbf{C}}|i)$ und $P(\underline{\mathbf{C}}|\neg i)$ zu

$$P(\underline{\mathbf{C}}|i) = \frac{H_i(\underline{\mathbf{C}})}{n_i} \quad (4.39)$$

$$P(\underline{\mathbf{C}}|\neg i) = \frac{H_{\neg i}(\underline{\mathbf{C}})}{n_{\neg i}} \quad (4.40)$$

und nach Bayes

$$P(i|\underline{\mathbf{C}}) = \left(1 + \frac{P(\underline{\mathbf{C}}|\neg i)}{P(\underline{\mathbf{C}}|i)} \cdot \frac{P(\neg i)}{P(i)} \right)^{-1} \quad (4.41)$$

Eine gegebene Region R mit der Farbe $\underline{\mathbf{C}}$ wird als Markerfarbe i klassifiziert, wenn (4.41) für i maximal ist und gleichzeitig einen Schwellwert Θ_i übersteigt. Existiert kein i , das diese Bedingungen erfüllt, gilt die Region als Hintergrund. Somit ergibt sich für jede Markerfarbe i eine (u. U. leere) Kandidatenmenge K_i zu

$$K_i = \left\{ R \mid \operatorname{argmax}_j P(j|\underline{\mathbf{C}}) = i \wedge P(i|\underline{\mathbf{C}}) \geq \Theta_i \right\} \quad (4.42)$$

Vereinfachend werden die A-priori-Wahrscheinlichkeiten $P(i)$ und $P(\neg i)$ für alle Marker gleich angenommen. Analog zur Hintergrundmodellierung beeinflusst der Quotient der beiden Werte in (4.41) nicht das Klassifikationsergebnis, sondern lediglich den Schwellwert Θ (vgl. Abschnitt 4.2.1), so dass $P(i)/P(\neg i) = 1$ gesetzt werden kann.

Die Bedingung in (4.42) lässt sich effizient als Lookup-Tabelle implementieren, die jedem Farbwert den entsprechenden Markerindex i bzw. -1 für Nicht-Markerfarben zuord-

net. Im Vergleich zur direkten Anwendung dieser Bedingung auf Pixelebene und anschließender Regionenbildung ist die hier beschriebene Vorgehensweise auf Basis der Mean-Shift-Segmentierung deutlich robuster. Die Varianz des Modalwerts einer Region ist niedriger als die einzelner Pixel, so dass eine höhere Trennschärfe erreichbar ist, d. h. weniger Fehlklassifikationen auftreten.

4.3.3 Topologische Regionenbeschreibung

Nach Ermittlung der Kandidatenmengen K_i erfordert die weitere Verarbeitung eine topologische Beschreibung der darin enthaltenen Regionen. Diese bildet die Grundlage für die im nächsten Abschnitt erläuterte Merkmalsberechnung.

Eine Bildregion ist eine Menge zusammenhängender Pixel, die sich im Allgemeinen bezüglich mindestens einer Eigenschaft – hier der Farbe – gleichen oder ähneln. Die Eigenschaft „zusammenhängend“ kann auf verschiedene Weise interpretiert werden. Definiert man zwei Pixel genau dann als zusammenhängend, wenn sie unmittelbar neben- oder übereinander liegen, ergibt sich die in Abbildung 4.8a gezeigte 4er-Nachbarschaft. Sind auch schräge Verbindungen zugelassen, resultiert die 8er-Nachbarschaft (Abbildung 4.8b).

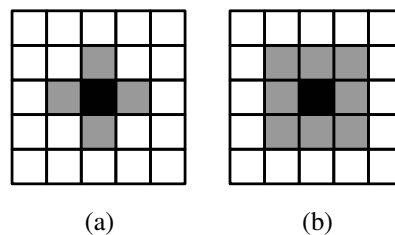


Abbildung 4.8: Referenzpixel (schwarz) mit 4er-Nachbarschaft (a) und 8er-Nachbarschaft (b).

Der in Abschnitt 4.3.1 beschriebene Mean-Shift-Algorithmus verwendet eine 8er-Nachbarschaft und liefert Regionen, deren Pixel jeweils demselben Cluster (d. h. demselben Modalwert) zugeordnet sind. Eine Kandidatenmenge K_i enthält möglicherweise benachbarte Regionen mit unterschiedlichen Modalwerten, die jedoch demselben Markerindex zugeordnet sind. Diese Regionen werden vor der weiteren Verarbeitung zusammengefügt.

Zur Repräsentation von Regionen im Rechner sind verschiedene Datenstrukturen verwendbar, die jeweils unterschiedliche Vor- und Nachteile aufweisen. Im Folgenden werden diese kurz gegenübergestellt.

Ist für gegebene Koordinaten lauffzeiteffizient die zugehörige Region zu ermitteln, so eignet sich eine Matrix, deren Zeilen- und Spaltenzahl mit der Bildauflösung übereinstimmt und die für jedes Pixel den Index der entsprechenden Region enthält (*labeled mask*). Aus dieser Darstellung sind jedoch geometrische Eigenschaften wie etwa der Umfang einer Region nicht direkt zu berechnen. Dies gilt auch für das nahe liegende Speichern aller Pixel einer Region in einer linearen Liste.

Eine weitere Möglichkeit besteht darin, eine Region als Liste ihrer Randpixel zu repräsentieren. Ein Pixel $p \in R$ ist genau dann ein Randpunkt einer Region, wenn es am Rand des Bildes liegt (d. h. $x = 0 \vee x = M - 1 \vee y = 0 \vee y = N - 1$ bei einer Auflösung von $M \times N$) oder wenn in seiner 4er-Nachbarschaft mindestens ein Pixel $q \notin R$ existiert. Aus dieser Darstellung lassen sich zahlreiche geometrische Merkmale mit deutlich weniger Rechenoperationen bestimmen als aus einer Liste aller Pixel der Region. Darüber hinaus wird weniger Speicherplatz benötigt, da die Anzahl der Randpixel deutlich kleiner ist als die Gesamtzahl der Pixel. Das Testen eines beliebigen Pixels p hinsichtlich der Bedingung $p \in R$ ist zwar aufwändiger als in den beiden erstgenannten Fällen, wird jedoch für die in dieser Arbeit angewandten Algorithmen nicht benötigt.

Abbildung 4.9 zeigt beispielhaft eine Region und deren Randpixel. Zur Berechnung geometrischer Merkmale ist zunächst eine mathematisch konsistente Interpretation der durch die Region beschriebenen geometrischen Fläche erforderlich. Im Folgenden ist diese Fläche definiert als das Polygon, das die Mittelpunkte der Randpixel der Region bilden (Abbildung 4.9c). Hierdurch werden Fehler vermieden, die aus der häufig verwendeten und einfach zu implementierenden, jedoch inkonsistenten Definition des Flächeninhalts als Gesamtzahl der Pixel der Region und der Randkurvenlänge als Umfang des beschriebenen Polygons entstehen.

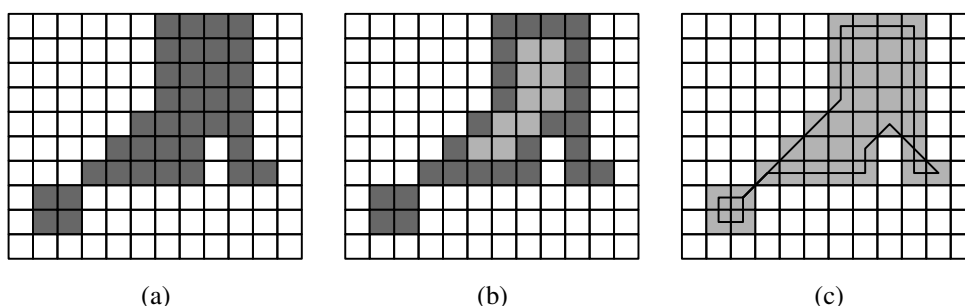


Abbildung 4.9: Beispiel einer Bildregion (a), repräsentiert als Liste ihrer Randpixel (b) sowie als Polygon (c). Hellgraue Pixel in (b) und (c) dienen dem Vergleich mit (a) und sind nicht in der Repräsentation enthalten.

4.3.4 Geometrische Merkmale

Die Berechnung geometrischer Merkmale einer Bildregion stellt in der Verarbeitungskette einen wichtigen Abstraktionsschritt dar. Die vorliegende Arbeit verwendet mehrere skalare Merkmale, die die Form der segmentierten Regionen beschreiben und damit einen Vergleich erlauben (s. Abschnitt 5.2). Der Inhalt der Regionen wird nicht berücksichtigt, da die verwendeten optischen Marker keine Textur aufweisen.¹

Im Folgenden sind die verwendeten Merkmale kurz erläutert. In Tab. 4.1 ist die Abhängigkeit jedes Merkmals von Translation und Rotation der Region in der Bildebene sowie Skalierung, d. h. Objektweite bzw. Auflösung der Kamera, angegeben. Diese Eigenschaften sind für den

¹Darüber hinaus sind Texturinformationen nicht robust gegen Bewegungsunschärfe, vgl. Abschnitt 2.5.3.

praktischen Einsatz der Handpositurerkennung wichtig, da der Nutzer sich nicht immer an derselben Stelle im Bild befindet und unterschiedliche Entfernungen von der Kamera einnehmen kann. Die erforderliche Invarianz ist durch eine geeignete Normierung erreichbar, die in Abschnitt 4.6.1 beschrieben wird.

Umfang

Der Umfang l einer Region ergibt sich trivial durch Summation über die Randpixel, wobei Pixel in 4er-Nachbarschaft einen Abstand von 1 und schräg aneinander grenzende einen Abstand von $\sqrt{2}$ haben. Der resultierende Wert ist vergleichsweise stark rauschbehaftet, da sich Störungen der Kontur fast ausschließlich als Steigerung auswirken und sich nicht, wie etwa beim Flächeninhalt, gegenseitig ausgleichen.

Flächeninhalt, Schwerpunkt und Momente 2. Ordnung

Für Polygone lassen sich Flächeninhalt a , Schwerpunkt $\underline{s} = \langle x_s, y_s \rangle$ und beliebige Momente $\nu_{p,q}$ effizient und subpixelgenau aus den Koordinaten der Eckpunkte berechnen [Ste96]. Im Folgenden bezeichnen x_i und y_i die Koordinaten des i -ten Eckpunkts und n die Anzahl der Eckpunkte, wobei $x_0 = x_n$ und $y_0 = y_n$ definiert ist, um ein geschlossenes Polygon zu erhalten. Normierte Momente sind mit $\alpha_{p,q}$ bezeichnet, zentrale Momente mit $\mu_{p,q}$. Gleichungen für Momente beliebiger Ordnung sind in [Ste96] zu finden.

$$\nu_{0,0} = a = \frac{1}{2} \sum_{i=1}^n x_{i-1}y_i - x_iy_{i-1} \quad (4.43)$$

$$\alpha_{1,0} = x_s = \frac{1}{6a} \sum_{i=1}^n (x_{i-1}y_i - x_iy_{i-1})(x_{i-1} + x_i) \quad (4.44)$$

$$\alpha_{0,1} = y_s = \frac{1}{6a} \sum_{i=1}^n (x_{i-1}y_i - x_iy_{i-1})(y_{i-1} + y_i) \quad (4.45)$$

$$\alpha_{2,0} = \frac{1}{12a} \sum_{i=1}^n (x_{i-1}y_i - x_iy_{i-1})(x_{i-1}^2 + x_{i-1}x_i + x_i^2) \quad (4.46)$$

$$\alpha_{1,1} = \frac{1}{24a} \sum_{i=1}^n (x_{i-1}y_i - x_iy_{i-1})(2x_{i-1}y_{i-1} + x_{i-1}y_i + x_iy_{i-1} + 2x_iy_i) \quad (4.47)$$

$$\alpha_{2,0} = \frac{1}{12a} \sum_{i=1}^n (x_{i-1}y_i - x_iy_{i-1})(y_{i-1}^2 + y_{i-1}y_i + y_i^2) \quad (4.48)$$

$$\mu_{2,0} = \alpha_{2,0} - \alpha_{1,0}^2 \quad (4.49)$$

$$\mu_{1,1} = \alpha_{1,1} - \alpha_{1,0}\alpha_{0,1} \quad (4.50)$$

$$\mu_{0,2} = \alpha_{0,2} - \alpha_{0,1}^2 \quad (4.51)$$

Die Momente zweiter Ordnung ($p + q = 2$) werden insbesondere zur Berechnung weiterer Formmerkmale wie unten beschrieben verwendet.

Trägheitsmomente

Betrachtet man die Bildpunkte einer Region als massebehaftet, so existiert dazu eine Achse mit minimalem Trägheitsmoment, um die sich die Region mit dem geringsten Energieaufwand drehen lässt. Die Trägheitsmomente j_1 und j_2 entlang dieser Hauptachse und senkrecht dazu sind

$$j_1 = \frac{\mu_{2,0} + \mu_{0,2}}{2a} + \sqrt{\left(\frac{\mu_{2,0} - \mu_{0,2}}{2a}\right)^2 + \left(\frac{\mu_{1,1}}{a}\right)^2} \quad (4.52)$$

$$j_2 = \frac{\mu_{2,0} + \mu_{0,2}}{2a} - \sqrt{\left(\frac{\mu_{2,0} - \mu_{0,2}}{2a}\right)^2 + \left(\frac{\mu_{1,1}}{a}\right)^2} \quad (4.53)$$

Exzentrizität

Ein Maß für die Exzentrizität einer Region ist nach [Jai89]

$$e = \frac{(\mu_{2,0} - \mu_{0,2})^2 + 4\mu_{1,1}}{a} \quad (4.54)$$

Dieser Wert ist für Kreise gleich Null und steigt für längliche Formen. Eine intuitive Definition ist weiterhin der Quotient der Trägheitsmomente j_1 und j_2 :

$$e = \frac{j_1}{j_2} \quad (4.55)$$

Nach (4.55) ist e für kreisförmige Flächen gleich Eins und nimmt für längliche Formen zu.

Orientierung

Aus der intuitiven Definition der Orientierung α einer Fläche als Orientierung der Hauptträgheitsachse ergibt sich

$$\alpha = \frac{1}{2} \arctan\left(\frac{2\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}}\right) \quad (4.56)$$

Die Orientierung ist periodisch mit 180° , was in vielen Fällen eine besondere Behandlung erfordert. Die Aussagekraft der Orientierung ist umso höher, je größer j_1 gegen j_2 ist. Flächen mit $j_1 \approx j_2$ weisen keine ausgeprägte Orientierung auf, was zu annähernd zufälligen Werten für α führt.

Tabelle 4.1: Abhängigkeit der vorgestellten geometrischen Merkmale von Translation, Rotation und Skalierung der betrachteten Fläche.

Merkmal	Invarianz gegen		
	Translation	Rotation	Skalierung
l	•	•	—
a	•	•	—
$\underline{s}$	—	—	—
j_1, j_2	•	•	—
e nach (4.54)	•	•	—
e nach (4.55)	•	•	•
α	•	—	•
c	•	•	•

Kompaktheit

Aus Umfang und Flächeninhalt berechnet sich die Kompaktheit zu

$$c = \frac{4\pi a}{l^2} \quad (4.57)$$

Kompakte Formen haben einen großen Flächeninhalt bei gleichzeitig niedrigem Umfang. Die kompaktste Fläche ist ein Kreis ($c = 1$), während für längliche Regionen oder solche mit zerklüfteter Randkurve $c \rightarrow 0$ gilt.

4.3.5 Randkurvenapproximation durch Ellipsen

Für die Handpositurerkennung erweist sich die Repräsentation von Bildregionen in Form eines Polygons (s. Abbildung 4.9c) aus zwei Gründen als zu aufwändig. Zunächst ist der Speicherplatzbedarf sehr hoch, da jeder Eckpunkt des Polygons berücksichtigt werden muss und selbst bei niedrigen Bildauflösungen häufig Polygone mit über 100 Eckpunkten entstehen. Darüber hinaus erfordern einfache geometrische Operationen wie etwa die Berechnung der Entfernung eines Punktes von der Region das Traversieren der gesamten Randkurve.

Eine kompaktere und besser handhabbare Repräsentation ergibt sich durch Approximation der Randkurve mittels einer Ellipse. Da die zweidimensionalen Abbildungen der Markerflächen nahezu immer konvex sind, ist diese Näherung hinreichend genau. Die Beschreibung der Ellipse erfolgt parametrisch in Form der Radien r_1 und r_2 , der Orientierung α und des Schwerpunkts $\underline{s}$ wie in Abbildung 4.10 gezeigt.

Grundlage zur Bestimmung der Ellipsenparameter bilden die geometrischen Merkmale der Region (vgl. Abschnitte 4.3.3 und 4.3.4). Da der Flächeninhalt a ein sehr aussagekräftiges

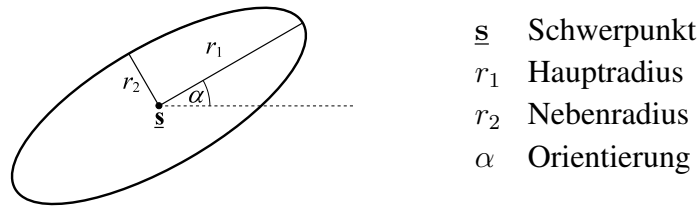


Abbildung 4.10: Parametrische Beschreibung einer beliebig positionierten und orientierten Ellipse.

und robustes Merkmal darstellt, wird zunächst gefordert, dass die elliptische Näherung diesen beibehält:

$$\pi r_1 r_2 = a \quad (4.58)$$

Schwerpunkt $\underline{s}$ und Orientierung α ergeben sich gemäß (4.44), (4.45) und (4.56) und werden unverändert übernommen.

Hinsichtlich der Wahl der Radien verbleibt mit (4.58) ein Freiheitsgrad. Um die tatsächliche Randkurve möglichst gut anzunähern, wird die Fläche der „Fehlerregion“ $(R \cup R') \setminus (R \cap R')$ minimiert. Dabei steht R für die polygonförmige und R' für die elliptische Region. Abbildung 4.11 visualisiert diese Fläche. In Abbildung 4.12 ist die Anwendung auf Markerregionen gezeigt.

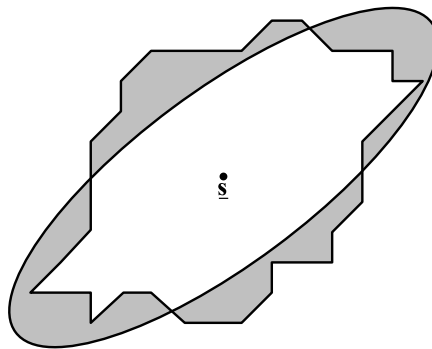


Abbildung 4.11: Bestimmung der Ellipsenradien durch Minimieren der grau eingezeichnete Fläche unter Einhaltung von (4.58). Zur Veranschaulichung sind die Radien hier nicht optimal gewählt.

4.4 Segmentierung des Gesichts

Unabhängig von der Segmentierung der Marker erfolgt auch die Lokalisation des Gesichts, das den Bezugspunkt zur Interpretation der Handposition bildet. Da der Benutzer sich während des Gebärdens kaum bewegt, ist eine einfache Heuristik ausreichend, die hautfarbene Regionen im oberen mittleren Bildbereich identifiziert und nach gesichtsähnlichen Formen

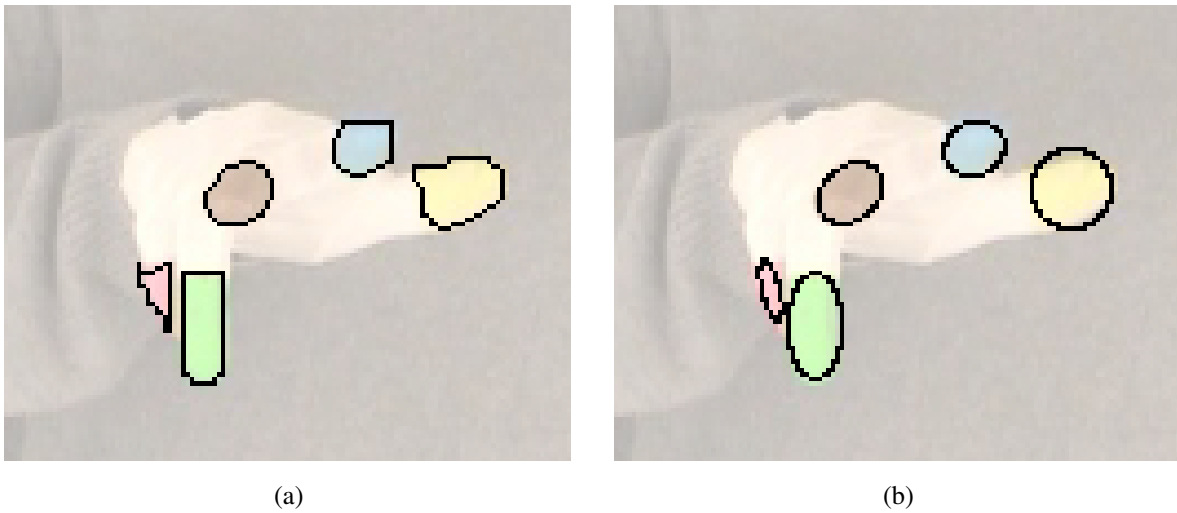


Abbildung 4.12: Randkurven der Markerregionen (a) und approximierende Ellipsen mit gleichem Flächeninhalt (b).

sucht. Dies erfolgt während der Vorbereitungs- und der Rückführungsphase, in denen sich Gesicht und Hände im Bild nicht überlappen und die Segmentierung somit vergleichsweise einfach ist. Ein alternatives, jedoch personenabhängiges Verfahren, das ein ASM (*active shape model*) einsetzt und das Gesicht dadurch auch bei teilweiser Verdeckung segmentieren kann, ist in [AZ02] beschrieben.

Zur Hautfarbdetektion dienen die in [JR98] veröffentlichten Histogramme von Hautfarbe und Nicht-Hautfarbe. Die Beschreibung der erwarteten, d. h. durchschnittlichen Gesichtsform und -position erfolgt auf Grundlage der im letzten Abschnitt erläuterten Merkmale und entsprechender empirisch bestimmter Mittelwerte und Varianzen. Als Gesicht wird diejenige Kandidatenregion ausgewählt, deren Merkmale die geringste Mahalanobisdistanz zu dieser Durch-

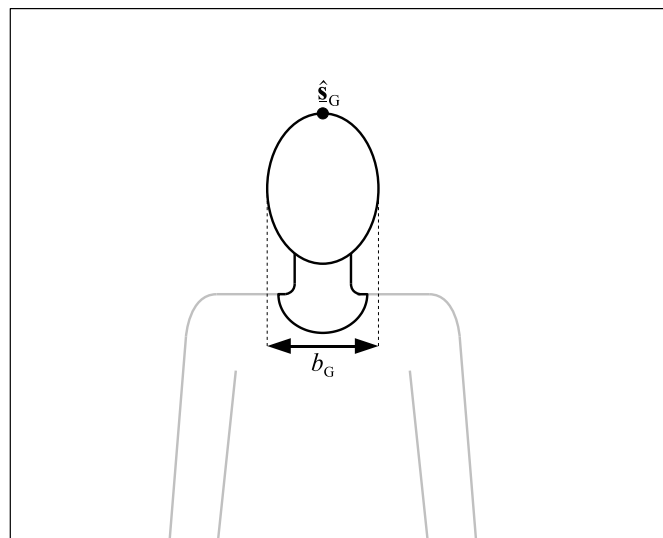


Abbildung 4.13: Scheitelpunkt $\hat{s}_G$ und Breite b_G der Gesichtsregion.

schnittsform aufweisen. Einzelheiten hierzu finden sich in [ZK05]. Die so erhaltene Region schließt häufig den Hals und ggf. den Ausschnitt ein, die ebenfalls hautfarben sind.

Nach der Lokalisation werden der Scheitelpunkt $\hat{s}_G$ sowie die Breite b_G der Gesichtsregion wie in Abbildung 4.13 dargestellt bestimmt, da diese für die weitere Verarbeitung von Bedeutung sind (s. Abschnitte 4.5.3 und 4.6.2). Im Gegensatz zu Mittelpunkt, Fläche oder Höhe der hautfarbenen Gesichtsregion bleiben $\hat{s}_G$ und b_G von Bärten oder tief ausgeschnittener Kleidung unbeeinflusst und stellen daher robuste Formmerkmale dar.

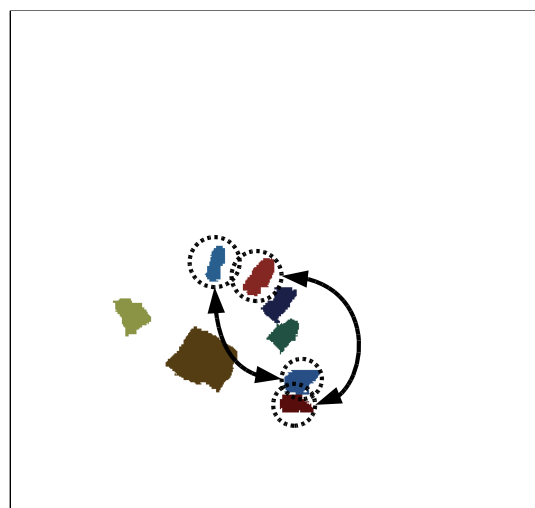
4.5 Zuordnung von Regionen zu Markern

Die Merkmalsextraktion liefert für jede Markerfarbe im Allgemeinen mehrere Regionen, wenn weitere Objekte dieser Farbe im Bild enthalten sind und durch die Hintergrundsubtraktion nicht vollständig eliminiert werden konnten. Dieser Fall ist exemplarisch in Abbildung 4.14 gezeigt.

Eine einfach zu realisierende und daher häufig angewandte Strategie zur Behandlung derartiger Mehrdeutigkeiten besteht darin, auf Informationen aus einer oder mehreren vergangenen Beobachtungen zurückzugreifen. Die aktuelle Szene wird so interpretiert, dass sich eine möglichst plausible Fortsetzung ergibt, d. h. die Abweichung von einer Prädiktion minimiert wird. In zahlreichen Arbeiten geschieht diese Prädiktion mit Hilfe eines Kalman-Filters. Dieser Algorithmus ist weit verbreitet, da er lauffeit- und speichereffizient implementierbar ist und einzelne Beobachtungen in Abhängigkeit von ihrer Zuverlässigkeit unterschiedlich gewichten



(a) Eingangsbild



(b) Segmentierungsergebnis

Abbildung 4.14: Mehrdeutigkeiten bei der Detektion der farbigen Markierungen (a). Die Mean-Shift-Segmentierung liefert je zwei Regionen der Farben Hellblau und Rot (b), so dass im Folgenden Domänenwissen nötig ist, um zu entscheiden, welche dieser Regionen den gesuchten Markierungen entsprechen.

kann [WB04, ZUA02].

Das Verfolgen einer einzelnen Hypothese hat jedoch den Nachteil, dass für die Interpretation einer Beobachtung zum Zeitpunkt t_0 nur Informationen aus früheren Bildern $\mathbf{I}(t)$ mit $t < t_0$ verwertbar sind. Spätere Beobachtungen, die aufgrund ihrer zeitlichen Proximität ebenfalls eine hohe Aussagekraft haben, werden nicht berücksichtigt.

Darüber hinaus kann eine einzelne Fehlinterpretation die Prädiktion erheblich stören und so eine korrekte Verarbeitung der folgenden Daten verhindern. Dies trifft insbesondere bei hochdimensionalen Lösungsräumen zu, wenn der Suchbereich aus Effizienzgründen auf die Umgebung der aktuellen Prädiktion beschränkt wird. Zur robusten Verarbeitung mehrdeutiger Daten sind derartige Ansätze somit ungeeignet.

Aus diesem Grund werden in der vorliegenden Arbeit zur Markerdetektion mehrere Hypothesen parallel verfolgt [ZK04, ZK05]. Damit berücksichtigt die Interpretation jedes Einzelbildes sowohl aus vorangegangenen als auch aus nachfolgenden Bildern gewonnene Informationen. Eine Fehlerkorrektur ist jederzeit, auch retrospektiv, durch Auswahl einer entsprechenden Hypothese möglich. Erst nach der Analyse sämtlicher Eingangsdaten erfolgt die Festlegung einer endgültigen Ergebnishypothese. Vergleichbare Verfahren, die den Hypothesenraum allerdings einschränken, sind in der Radartechnik bekannt [BS87, Bro98].

Die folgenden Abschnitte 4.5.1 und 4.5.2 beschreiben allgemein die mathematischen Grundlagen zur Behandlung multipler Hypothesen sowie Algorithmen zur effizienten Ermittlung der plausibelsten Hypothese. Abschnitt 4.5.3 geht konkret auf die Anwendung zur Zuordnung von Regionen zu Markern ein. Das Konzept der multiplen Hypothesen wird später in Abschnitt 5.3.1 aufgegriffen und für die Mehrdeutigkeitsauflösung bei der Positurrekonstruktion verwendet.

4.5.1 Hypothesenraum

Die Möglichkeiten zur Interpretation eines mehrdeutigen zeitdiskreten Vorgangs lassen sich grafisch als Trellisdiagramm (engl. *trellis*: Gitter) darstellen (s. Abbildung 4.15). Zu jedem Zeitpunkt t existieren $n(t) \geq 1$ verschiedene Hypothesen, von denen genau eine als richtig oder die Realität besser als alle anderen Hypothesen approximierend angenommen wird. Von allen Hypothesen zum Zeitpunkt t ist jeweils ein Übergang zu jeder Hypothese des Zeitpunkts $t + 1$ möglich. Eine Sequenz von Hypothesen bildet einen Pfad durch das Trellisdiagramm. In Abbildung 4.15 ist exemplarisch ein beliebiger Pfad hervorgehoben.

Jeder Pfad repräsentiert eine mögliche Interpretation des beobachteten Vorgangs. Insgesamt existieren $\prod_{t=0}^{T-1} n(t)$ verschiedene Pfade. Beispielsweise ergibt eine dreisekündige und mit einer Bildrate von 25 Hz aufgenommene Gebärde bereits bei 10 Hypothesen pro Bild 10^{75} Pfade. Diese Komplexität stellt den Nachteil des Verfolgens multipler Hypothesen dar und muss durch effiziente Algorithmen geeignet reduziert werden, um eine Verarbeitung in annähernd Echtzeit zu erreichen.

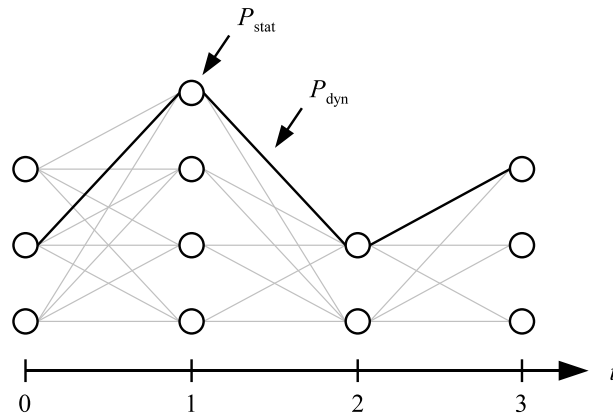


Abbildung 4.15: Trellisdiagramm des Hypothesenraums. Kreise symbolisieren Hypothesen, Verbindungslinien stellen Übergänge zwischen zeitlich benachbarten Hypothesen dar. Exemplarisch ist ein beliebiger Pfad hervorgehoben. Die jeder Hypothese und jedem Übergang zugeordneten Wahrscheinlichkeiten P_{stat} bzw. P_{dyn} sind in Abschnitt 4.5.2 erläutert.

4.5.2 Viterbi-Pfadsuche

Gesucht ist die Hypothesenfolge, die den realen Vorgang optimal beschreibt. Bei einer wertdiskreten Messung wertkontinuierlicher Größen, wie sie etwa ein Kamerabild darstellt, bedeutet „optimal“ hier die Minimierung der Differenz zwischen Hypothesen und Realität. Da die tatsächlich vorliegende Konfiguration unbekannt ist, muss diese Minimierung auf Wahrscheinlichkeiten basieren, die die Plausibilität der einzelnen Pfadelemente quantitativ schätzen. Die Güte dieser Schätzung ist entscheidend für die Leistung des Systems.

Jeder einzelnen Hypothese $H_i(t)$ des Zeitindex t lässt sich zunächst eine Wahrscheinlichkeit zuordnen, die unabhängig vom zeitlichen Zusammenhang ist, d. h. von möglichen Vorgängern und Nachfolgern nicht beeinflusst wird. Diese statische Wahrscheinlichkeit wird mit $P_{\text{stat},i}(t)$ bezeichnet.

Komplementär ergibt sich eine dynamische Wahrscheinlichkeit $P_{\text{dyn},ij}(t)$ für den Übergang zwischen zwei Hypothesen $H_i(t-1)$ und $H_j(t)$. Von der statischen Wahrscheinlichkeit der Hypothesen i und j selbst ist P_{dyn} unabhängig.

Die konkrete Berechnung von P_{stat} und P_{dyn} ist anwendungsspezifisch und wird für die Markerdetektion sowie die Positurrekonstruktion in den Abschnitten 4.5.3 bzw. 5.3.1 beschrieben. Sinnvoll ist für beide Werte grundsätzlich die Betrachtung mehrerer Kriterien, weil so Unsicherheiten hinsichtlich eines Kriteriums durch andere ausgeglichen werden können.

Ein Pfad W durch den Hypothesenraum, der für einen Zeitindex t die Hypothese mit dem Index $w(t)$ enthält, lässt sich schreiben als Folge

$$W = \langle w(0), w(1), \dots, w(T-1) \rangle \quad (4.59)$$

Die Gesamtwahrscheinlichkeit dieses Pfades ist das Produkt der statischen und dynamischen

Wahrscheinlichkeiten seiner Hypothesen und der entsprechenden Übergänge:

$$P(W) = P_{\text{stat},w(0)}(0) \prod_{t=1}^{T-1} P_{\text{dyn},w(t-1)w(t)}(t) \cdot P_{\text{stat},w(t)}(t) \quad (4.60)$$

Die Lösung des eigentlichen Problems besteht nun darin, für jeden Zeitindex sämtliche Hypothesen zu erstellen, alle statischen und dynamischen Wahrscheinlichkeiten zu berechnen und den wahrscheinlichsten Pfad $\widehat{W}$ zu bestimmen:

$$\widehat{W} = \underset{W}{\operatorname{argmax}} P(W) \quad (4.61)$$

Mit Hilfe des Viterbi-Algorithmus kann (4.61) effizient gelöst werden, falls die zur Berechnung von P_{stat} und P_{dyn} angewandten Algorithmen gedächtnislos sind. Für P_{stat} wurde dies mit der Unabhängigkeit von Vorgänger- und Nachfolgehypothese bereits explizit gefordert. Für P_{dyn} wären theoretisch gedächtnisbehaftete stochastische Zustandsschätzer wie etwa der Kalman-Filter geeignet, um beispielsweise die Trajektorie eines Objekts vorherzusagen oder den aktuellen Geschwindigkeitsvektor zu bestimmen. Die Anwendung derartiger Verfahren gestattet den Einsatz des Viterbi-Algorithmus jedoch nicht und führt zu einer nicht handhabbaren Zeitkomplexität, die Abschnitt 4.5.2.1 anhand eines Beispiels veranschaulicht.

Der Viterbi-Algorithmus reduziert den benötigten Rechenaufwand auf einen Bruchteil. Für jede Hypothese $H_i(t)$ muss von allen einlaufenden Pfaden $W_i(t)$ nur genau der weiterverfolgt werden, der $H_i(t)$ mit der höchsten Wahrscheinlichkeit erreicht:

$$\widehat{W}_i(t) = \underset{W_i(t)}{\operatorname{argmax}} \left(P_{\text{stat},w(0)}(0) \prod_{t'=1}^t P_{\text{dyn},w(t'-1)w(t')}(t') \cdot P_{\text{stat},w(t')}(t') \right) \quad (4.62)$$

Dabei steht $W_i(t)$ für einen Pfad, der mit der Hypothese $H_i(t)$ endet und somit die Länge $t+1$ hat:

$$W_i(t) = \langle w(0), w(1), \dots, w(t-1), i \rangle \quad (4.63)$$

Alle anderen Pfade zu $H_i(t)$ können (4.61) nicht lösen: Für jede beliebige Fortsetzung $\widetilde{W} = \langle w(t+1), w(t+2), \dots, w(T-1) \rangle$ ergäbe sich eine niedrigere Gesamtwahrscheinlichkeit als für $\widehat{W}_i(t)$ mit derselben Fortsetzung, da $P(\widetilde{W})$ von $W_i(t)$ – d. h. von der Vorgeschichte – unabhängig ist. Bei einer gedächtnisbehafteten Wahrscheinlichkeitsberechnung wäre diese Unabhängigkeit nicht gewährleistet.

In [RJ86] ist eine rekursive Variante des Viterbi-Algorithmus beschrieben. Eine nichtrekursive Umsetzung findet sich in Anhang B.

4.5.2.1 Komplexitätsvergleich

Wird die Anzahl der Hypothesen n pro Zeitindex vereinfachend als konstant angenommen, so sind im allgemeinen Fall, d. h. wenn die Wahrscheinlichkeitsberechnung abhängig von der Vorgeschichte des betrachteten Pfades ist, $\sum_{t=0}^{T-1} n^{t+1}$ statische und $\sum_{t=1}^{T-1} n^{t+1}$ dynamische Wahrscheinlichkeiten zu berechnen. Die Zeitkomplexität ist somit $O(n^T)$. Benötigt die Implementierung pro Wahrscheinlichkeitswert 10^{-15} Sekunden, so erfordert das eingangs angeführte Beispiel einer dreisekündigen Sequenz mit $n = 10$ eine Rechenzeit von $3,5 \cdot 10^{52}$ Jahren.

Die Unabhängigkeit von der Vorgeschichte des Pfades reduziert dies auf Tn für P_{stat} und $(T-1)n^2$ für P_{dyn} , so dass sich eine Zeitkomplexität von $O(Tn^2)$ ergibt. Durch Anwendung des Viterbi-Algorithmus wird darüber hinaus die Anzahl der gleichzeitig zu verwaltenden Pfade von n^T auf n reduziert. Für das obige Beispiel ergibt sich eine Rechenzeit von $8,15 \cdot 10^{-12}$ Sekunden.

4.5.2.2 Beschleunigung durch Näherungen

Der Viterbi-Algorithmus findet in jedem Fall den wahrscheinlichsten aller n^T möglichen Pfade. Bei Verzicht auf diese Garantie ist eine weitere Beschleunigung erreichbar.

Zunächst besteht die Möglichkeit, unwahrscheinliche Hypothesen bereits vor Beginn der Pfadsuche zu löschen. Das entsprechende Kriterium kann beispielsweise eine Obergrenze n_{max} für die Anzahl $n(t)$ der Hypothesen pro Zeitindex oder eine maximale Wahrscheinlichkeitsdifferenz bezüglich der plausibelsten Hypothese sein.

Die in Anhang B präsentierte nichtrekursive Umsetzung des Algorithmus erlaubt darüber hinaus auf einfache Weise ein vorzeitiges Abbrechen wenig aussichtsreicher Pfade. Nach Schritt 3 können aus der Menge $W(t+1)$ beliebig viele Elemente entfernt werden. Geeignete Strategien sind, analog zum zuvor beschriebenen Löschen einzelner Hypothesen, die Festlegung einer maximalen Wahrscheinlichkeitsdifferenz zum gegenwärtig besten Pfad sowie einer Obergrenze $n_{W,\text{max}}$ für die Gesamtzahl $|W(t+1)|$ der zum Zeitpunkt $t+1$ betrachteten Pfade (dies wird als *beam search* bezeichnet).

4.5.3 Hypothesen und Bewertungskriterien für die Markerzuordnung

Bei der Zuordnung von Regionen zu Markern sind pro Marker i drei verschiedene Kategorien von Hypothesen möglich:

1. Marker i ist nicht sichtbar; alle Kandidatenregionen K_i sind Störobjekte. Diese Kategorie enthält immer genau eine Hypothese.
2. Marker i entspricht genau einer Kandidatenregion $R_j \in K_i$; alle übrigen Regionen $K_i \setminus R_j$ sind Störobjekte. Diese Kategorie enthält $|K_i|$ Hypothesen.

3. Marker i entspricht mehreren nicht zusammenhängenden Kandidatenregionen² aus K_i .

Die dritte Kategorie enthält $\sum_{j=2}^{|K_i|} \binom{|K_i|}{j}$ Hypothesen und ist damit vergleichsweise komplex. Da sie bei der verwendeten Markergeometrie in der Praxis äußerst selten auftritt, wird sie vernachlässigt, um die Verarbeitung zu beschleunigen. Damit ist die Gesamtzahl der Hypothesen pro Marker

$$n_i = 1 + |K_i| \quad (4.64)$$

und die Anzahl der Hypothesen insgesamt

$$n = \prod_i n_i \quad (4.65)$$

Zur Berechnung der statischen und dynamischen Wahrscheinlichkeiten kommt ein heuristisches Verfahren zum Einsatz. Die Abweichung mehrerer geometrischer Maße von empirisch ermittelten Erwartungswerten wird mit einer stetigen nichtlinearen Bewertungsfunktion gewichtet und zu einem Gesamtwert verrechnet.

Hierzu eignet sich die Sigmoidfunktion

$$P = \gamma \left(1 - \frac{1}{1 + \exp\left(\frac{x_{0,5} - x}{\beta}\right)} \right) \quad (4.66)$$

Darin steht x für die Abweichung vom Standardwert, und γ ist eine Normalisierungskonstante, die so gewählt wird, dass sich die Wahrscheinlichkeiten P aller Hypothesen bzw. Hypothesenübergänge zu Eins addieren.³ Für $x = x_{0,5}$ ist $P = 0,5\gamma$. Der Wert β bestimmt die Steilheit und erlaubt dadurch die Beschreibung von „harten“ Kriterien (steiler Abfall von P) und „weichen“ Kriterien (flacher Abfall von P).

Abbildung 4.16 visualisiert die Sigmoidfunktion. Die Werte $x_{0,9}$ und $x_{0,1}$ markieren die Stellen, an denen $P = 0,9\gamma$ bzw. $P = 0,1\gamma$ beträgt. Sie ergeben sich aus (4.66) zu

$$x_{0,1} = x_{0,5} + \beta \ln 9 \quad (4.67)$$

$$x_{0,9} = x_{0,5} - \beta \ln 9 \quad (4.68)$$

und werden im Folgenden an Stelle von $x_{0,5}$ und β verwendet, da sie die anschauliche Entsprechung „sehr plausibel“ bzw. „sehr unwahrscheinlich“ haben

²Zusammenhängende Regionen in K_i sind bereits vereinigt (vgl. Abschnitt 4.3.3).

³ Da die Konstante γ gemäß (4.60) in allen Pfaden enthalten ist, hat sie keinen Einfluss auf die Bestimmung des Gewinnerpfads nach (4.61). In der Implementierung wird daher vereinfachend $\gamma = 1$ gesetzt.

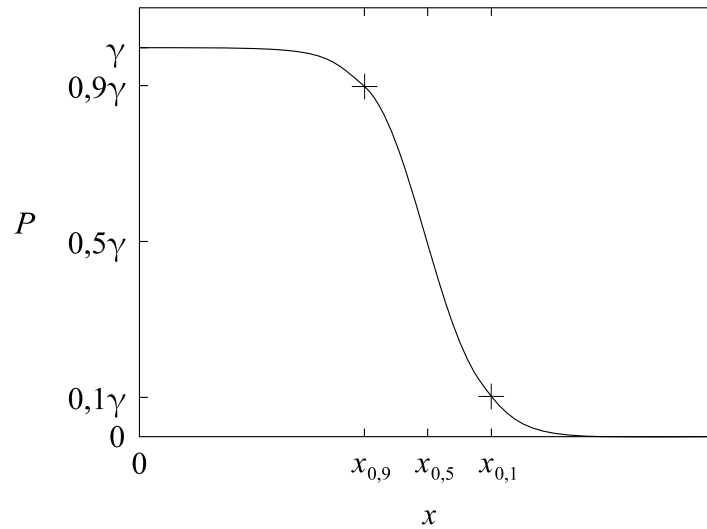


Abbildung 4.16: Sigmoidfunktion zur Abbildung von Merkmalen bzw. Merkmalsdifferenzen x auf statistische Wahrscheinlichkeiten P . Es gilt $P(x_{0,9}) = 0,9\gamma$ und $P(x_{0,1}) = 0,1\gamma$.

Die Berechnung der statischen und dynamischen Wahrscheinlichkeiten besteht nun in der Auswahl der zu bewertenden Merkmale sowie der Festlegung der Parameter $x_{0,9}$ und $x_{0,1}$. Dies ist im Folgenden qualitativ erläutert; quantitative Angaben finden sich in Anhang C. Sowohl P_{stat} als auch P_{dyn} ergeben sich als Produkt der Funktionswerte von (4.66) für jeden einzelnen Marker und für jedes Kriterium.

4.5.3.1 Statische Wahrscheinlichkeiten

Die statische Hypothesenwahrscheinlichkeit P_{stat} berücksichtigt den Flächeninhalt der Markerregionen, ihre Entfernung vom gemeinsamen Schwerpunkt und die Abweichung der Pixelfarben vom Hintergrundmodell:

$$P_{\text{stat}} = P_{\text{stat},a} \cdot P_{\text{stat},r} \cdot P_{\text{stat},\mu} \quad (4.69)$$

Flächeninhalt Durch (Selbst-)Verdeckung kann die Markerfläche a beliebig klein und für nicht sichtbare Marker auch Null werden, so dass keine Untergrenze für plausible Werte existiert. In Abhängigkeit von der Gesichtsbreite b_G (s. Abschnitt 4.4) kann jedoch eine Grenze $x_{0,9}$ geschätzt werden, bei deren Überschreitung die Hypothese wesentlich unwahrscheinlicher wird. Dabei ist die physikalische Größe der jeweiligen Markerfläche zu berücksichtigen, die etwa bei der in Abbildung 3.4 (S. 67) gezeigten Geometrie für Handrücken- und Fingermarker unterschiedlich ist. Der Wert $x_{0,1}$ wird geeignet oberhalb von $x_{0,9}$ gewählt.

Entfernung vom gemeinsamen Schwerpunkt Analog zum Flächeninhalt existiert auch für r keine untere Grenze, da jeder Marker beliebig nah am gemeinsamen Schwerpunkt $\underline{s}_{\text{DH}}$

liegen kann. Viele Falschzuordnungen zu Störobjekten im Hintergrund resultieren hingegen in großem r . Geeignete Grenzen $x_{0,9}$ und $x_{0,1}$, die wiederum relativ zu b_G gewählt werden, vermeiden die entsprechenden Hypothesen.

Abweichung vom Hintergrundmodell Die Hintergrundsubtraktion liefert eine binäre Maske, die die weitere Verarbeitung auf Bereiche beschränkt, deren Abweichung vom Hintergrundmodell einen Schwellwert übersteigt (vgl. Abschnitt 4.2). Sinnvoll ist darüber hinaus, den Grad dieser Abweichung in die Berechnung der statischen Zuordnungswahrscheinlichkeit einzubeziehen. Regionen, die sich deutlich vom Hintergrund abheben, werden dadurch bevorzugt gegenüber solchen, die nur knapp oberhalb des Schwellwerts liegen und möglicherweise Störungen darstellen. Der letztgenannte Fall ist häufig die Folge einer automatischen kamera-seitigen Anpassung von Belichtung oder Weißabgleich während der Aufnahme.

4.5.3.2 Dynamische Wahrscheinlichkeiten

In die Berechnung der dynamischen Wahrscheinlichkeit fließen Fläche a und Position $\underline{s}$ der Markerregionen ein:

$$P_{\text{dyn}} = P_{\text{dyn},a} \cdot P_{\text{dyn},s} \quad (4.70)$$

Da zwischen zwei aufeinanderfolgenden Eingangsbildern lediglich 0,04 Sekunden vergehen, ändern sich beide Merkmale nur langsam. Grundsätzlich kann daher der zum Zeitpunkt $t - 1$ gemessene Wert als Prädiktion für den Zeitpunkt t fungieren. Diese Annahme ist mit der Forderung nach Gedächtnislosigkeit (vgl. Abschnitt 4.5.2) vereinbar und hat sich in der Praxis bewährt. Sie hat weiterhin den Vorteil, Schwingungen und die Rückkopplung von Fehlern zu vermeiden.

Flächenänderung Analog zu $P_{\text{stat},a}$ bewertet $P_{\text{dyn},a}$ die Flächenänderung $|\Delta a|$ relativ zu b_G . Die Wahrscheinlichkeit des Hypothesenübergangs sinkt mit steigendem $|\Delta a|$, wobei $x_{0,9}$ und $x_{0,1}$ hinreichend groß zu wählen sind, um die durch die Änderung von Handform und -stellung bedingten Schwankungen zuzulassen.

Zu Beginn und am Ende der Sequenz erscheinen die Hände am unteren Bildrand bzw. verlassen dort den Aufnahmebereich. In diesen beiden Phasen und in diesem Bildbereich wird das Vorzeichen der Flächendifferenz beachtet und Zu- bzw. Abnahme der Markerfläche durch entsprechend größere Werte $x_{0,9}$ und $x_{0,1}$ gestattet.

Bewegung Ein aussagekräftiges Kriterium hinsichtlich der Plausibilität eines Hypothesenübergangs ist die Bewegung des Schwerpunkts der Markerregion. Je weiter die Schwerpunkte derselben Region in zwei aufeinanderfolgenden Hypothesen voneinander entfernt sind, desto unwahrscheinlicher ist der Übergang. Wie auch hinsichtlich der Flächenänderung werden

für $x_{0,9}$ und $x_{0,1}$ ausreichend große Werte verwendet, so dass die im Rahmen des Gebärdens stattfindende Bewegung plausibel bleibt.

4.5.3.3 Behandlung nicht sichtbarer Markierungen

Für nicht sichtbare Markerregionen sind mit Ausnahme der Fläche $a = 0$ keine sinnvollen Merkmale definierbar. Daher können in diesem Fall nicht alle der oben genannten Kriterien ausgewertet werden. Die verdeckten Regionen müssen jedoch trotzdem einen Beitrag zur Gesamtwahrscheinlichkeit liefern, der etwa dem der sichtbaren Regionen entspricht. Andernfalls würden Hypothesen begünstigt, die möglichst wenige oder gar keine sichtbaren Markerregionen enthalten. Aus diesem Grund werden für $P_{\text{stat},r}$, $P_{\text{stat},\mu}$ und $P_{\text{dyn},r}$ jeweils empirisch bestimmte Konstanten verwendet, wenn eine Berechnung aufgrund fehlender Merkmale nicht möglich ist.

4.6 Aufbau des Merkmalsvektors

Der Merkmalsvektor $\underline{\mathbf{F}}$ hat eine kompakte und abstrahierende Repräsentation eines Eingangsbildes $\underline{\mathbf{I}}$ zum Ziel und muss die in der Gebärdensprache verwendeten Ausdrucksmittel der Handform, -stellung und -position umfassen. Er lässt sich dementsprechend in folgende Komponenten aufteilen:

$$\underline{\mathbf{F}} = \begin{bmatrix} \underline{\mathbf{F}}_{\text{DH}} \\ \underline{\mathbf{s}}_{\text{DH}} \\ \underline{\mathbf{s}}_{\text{NDH}} \end{bmatrix} \quad (4.71)$$

Dabei hängt $\underline{\mathbf{F}}_{\text{DH}}$ von Form und Stellung der dominanten Hand ab, $\underline{\mathbf{s}}_{\text{DH}}$ und $\underline{\mathbf{s}}_{\text{NDH}}$ stehen für die Position (d. h. den Schwerpunkt) der dominanten bzw. nichtdominanten Hand. Im Folgenden ist die Berechnung dieser Elemente aus den zuvor ermittelten Regionen erläutert.

4.6.1 Form und Stellung der dominanten Hand

Die Beschreibung der Ansicht der dominanten Hand mittels $\underline{\mathbf{F}}_{\text{DH}}$ bildet das zweidimensionale Komplement zur Spezifikation der 3D-Positur $\underline{\mathbf{P}}$ durch $\underline{\mathbf{B}}$ und $\underline{\Phi}$ (vgl. (3.5)). Sie verwendet geeignete geometrische Merkmale, um die Erscheinung der Markerflächen zu charakterisieren. Anschließend besteht die Erkennungsaufgabe in der Abbildung einer Merkmalsvektorsequenz $\underline{\mathbf{F}}_{\text{DH}}(t)$ auf eine Positurresequenz $\underline{\mathbf{P}}(t)$ (vgl. (2.9)).

Der Aufbau des Merkmalsvektors ist von zentraler Bedeutung für das gesamte Erkennungssystem. Neben der obligatorischen Forderung nach möglichst hoher Aussagekraft müssen die Merkmale weiteren Anforderungen genügen, die nachfolgend skizziert sind. Anschließend wird die Berechnung und Normierung der Merkmale beschrieben.

Anforderungen

Da die Rekonstruktion die Abfrage einer Posituredatenbank umfasst, die einige Millionen Einträge der Form $\langle \underline{\mathbf{F}}_{\text{DH}}, \underline{\mathbf{P}} \rangle$ enthält (s. Abschnitt 5.1), ist eines der wichtigsten Ziele ein möglichst geringer Speicherplatzbedarf. Lässt sich die Datenbank vollständig im Arbeitsspeicher puffern, wird die Zugriffsgeschwindigkeit im Vergleich zum Lesen von der Festplatte signifikant erhöht.

Um den Positurenraum ausreichend fein zu quantisieren, muss die Datenbank etwa 2,5 Millionen Einträge enthalten (s. Abschnitt 5.1.1). Bei einer Speichergröße von 512 MB stehen somit pro Eintrag ca. 200 Byte zur Verfügung. Die Positurkomponenten $\underline{\mathbf{B}}$ und $\underline{\Phi}$ nehmen diskrete Werte an und lassen sich daher indizieren: Für die Beugungsgrade der Gelenke genügt pro Finger ein Byte, während die Ansichtswinkel zweckmäßig als 4-Byte-Integer codiert werden. Damit verbleiben bei sechs Markern je etwa 32 Byte für die Merkmale.

Eine weitere Anforderung ist die Invarianz von $\underline{\mathbf{F}}_{\text{DH}}$ gegen Translation und Skalierung, d. h. die Unabhängigkeit von der Kameraauflösung sowie der Position der Hand im Raum. Diese Eigenschaft ist erforderlich, um unabhängig von Auflösung und Objektweite mit einer einzigen Datenbank arbeiten zu können.

Merkmalsberechnung

Zunächst werden die im Eingangsbild segmentierten Markerregionen wie in Abschnitt 4.3.5 erläutert mittels Ellipsen approximiert. Damit liegen Schwerpunkt $\underline{\mathbf{s}}_i$, Orientierung α_i sowie Radien $r_{1,i}$ und $r_{2,i}$ für jede Ellipse i vor. Ferner ist der Flächeninhalt a_i aus (4.58) bekannt. Für den Index i gilt $1 \leq i \leq n$, wobei n die Anzahl der verwendeten Marker angibt und größer sein kann als die Anzahl der im aktuellen Bild tatsächlich sichtbaren bzw. segmentierten Markerregionen. Die Merkmale nicht sichtbarer Regionen werden gleich Null gesetzt; sie haben damit keinen Einfluss auf die folgenden Berechnungen, ohne dass eine explizite Fallunterscheidung nötig ist.

Als Hilfsgröße dient der gemeinsame Schwerpunkt der Markerflächen der dominanten Hand

$$\underline{\mathbf{s}}_{\text{DH}} = \frac{\sum_{i=1}^n \underline{\mathbf{s}}_i a_i}{\sum_{i=1}^n a_i} \quad (4.72)$$

Jeder Ellipsenschwerpunkt $\underline{\mathbf{s}}_i$ wird relativ zu $\underline{\mathbf{s}}_{\text{DH}}$ und in Polarkoordinaten $\langle r_i, \varphi_i \rangle$ angegeben. Damit sind die Merkmale translationsinvariant. Abbildung 4.17 visualisiert die bisher erläuterten Schritte.

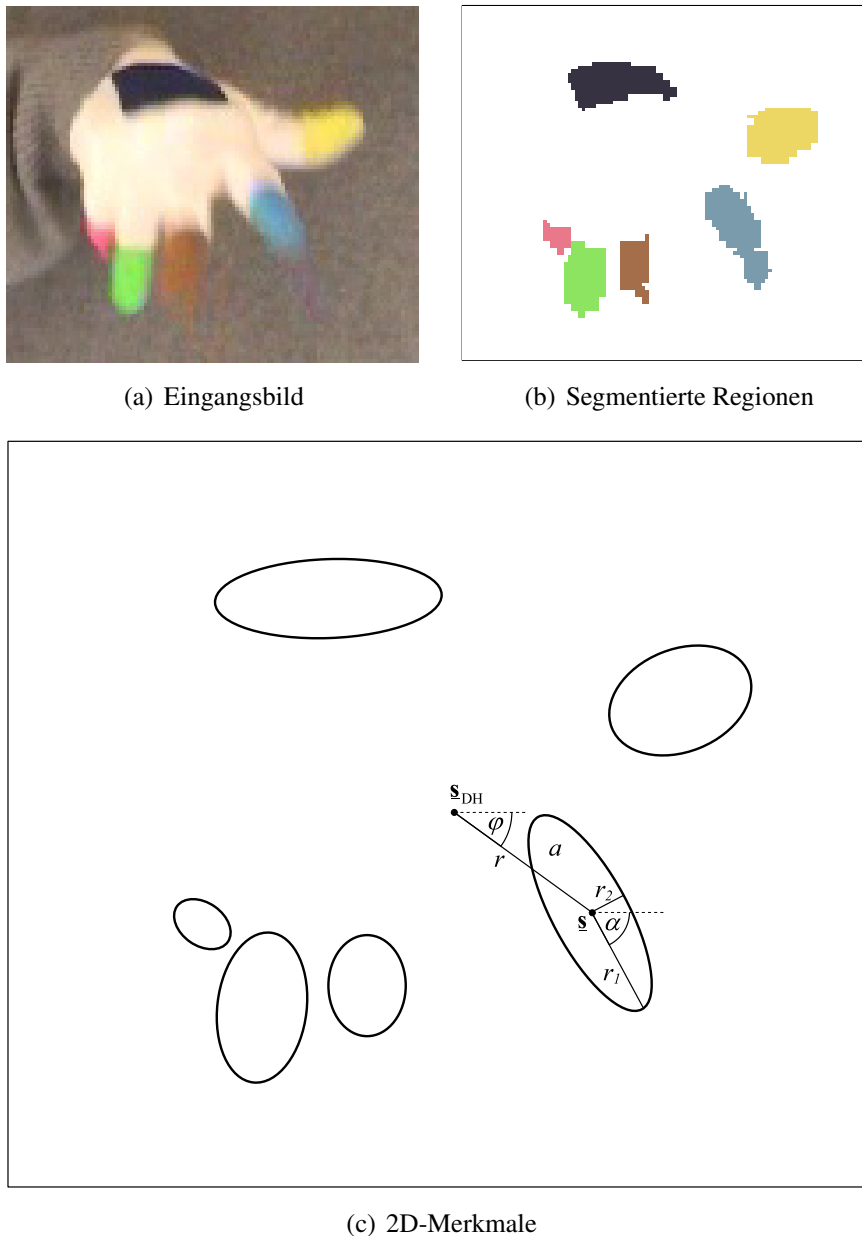


Abbildung 4.17: Berechnung der 2D-Merkmale. Jede Markerregionen wird mittels einer Ellipse approximiert, die durch ihre Position $\underline{s} = \langle r, \varphi \rangle$ relativ zum gemeinsamen Schwerpunkt $\underline{s}_{DH}$ aller Marker der dominanten Hand, Orientierung α sowie die Radien r_1 und r_2 charakterisiert ist. Zur Berechnung von $\underline{s}_{DH}$ und zur Normierung der Merkmale dient die Fläche a . Der Index i ist aus Gründen der Übersichtlichkeit ausgelassen.

Normierung

Skalierungsinvarianz ergibt sich aus der Normierung aller Entfernungen auf die Quadratwurzel der Summe aller Flächen:

$$r'_i = \frac{r_i}{\gamma} \quad (4.73)$$

$$r'_{1,i} = \frac{r_{1,i}}{\gamma} \quad (4.74)$$

$$r'_{2,i} = \frac{r_{2,i}}{\gamma} \quad (4.75)$$

mit der Normierungskonstanten

$$\gamma = \sqrt{\sum_{i=1}^n a_i} \quad (4.76)$$

Die Flächen a_i werden hierdurch auf γ^2 normiert:

$$a'_i = \pi r'_{1,i} r'_{2,i} = \frac{\pi r_{1,i} r_{2,i}}{\gamma^2} = \frac{a_i}{\gamma^2} \quad (4.77)$$

Sie ergänzen sich in Summe zu Eins:

$$\sum_{i=1}^n a'_i = \frac{\sum_{i=1}^n a_i}{\gamma^2} = 1 \quad (4.78)$$

Die translations- und skalierungsinvarianten Merkmale $\underline{\mathbf{F}}_i$ einer Ellipse i sind somit

$$\underline{\mathbf{F}}_i = \begin{bmatrix} a'_i \\ r'_i \\ \varphi_i \\ r'_{1,i} \\ r'_{2,i} \\ \alpha_i \end{bmatrix} \quad (4.79)$$

Der Merkmalsvektor $\underline{\mathbf{F}}_{\text{DH}}$, der alle n Marker der dominanten Hand beschreibt, ergibt sich schließlich zu

$$\underline{\mathbf{F}}_{\text{DH}} = \begin{bmatrix} \underline{\mathbf{F}}_1 \\ \vdots \\ \underline{\mathbf{F}}_n \end{bmatrix} \quad (4.80)$$

4.6.2 Positionen beider Hände

Die Beschreibung der Handposition greift auf die in Abschnitt 4.4 dargestellten Merkmale der Gesichtsregion zurück. Der Scheitelpunkt $\hat{s}_G$ bildet den Ursprung des Handkoordinatensystems, so dass $\underline{s}_{DH}$ und $\underline{s}_{NDH}$ unabhängig von der Position des Gebärdenden im Bild sind. Um darüber hinaus den Einfluss der Bildauflösung bzw. der Entfernung des Benutzers von der Kamera zu eliminieren, erfolgt eine Normierung auf die Gesichtsbreite b_G . Abbildung 4.18 zeigt das resultierende Koordinatensystem.

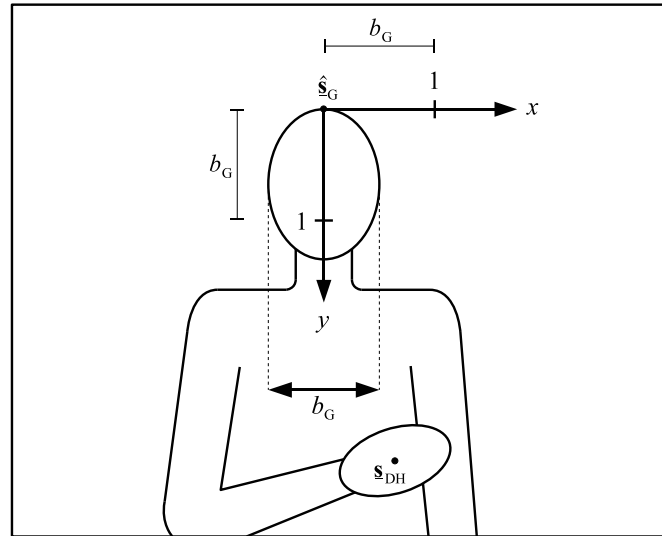


Abbildung 4.18: Die Handposition wird relativ zum Scheitelpunkt $\hat{s}_G$ des Gesichts angegeben und auf die Gesichtsbreite b_G normiert.

Die Position der dominanten Hand liegt in Form des gemeinsamen Schwerpunkts der Markerflächen $\underline{s}_{DH}$ vor. Analog ergibt sich die Position der nichtdominanten Hand $\underline{s}_{NDH}$, wobei hier lediglich eine Markierung existiert. Zusammen mit den Form und Stellung beschreibenden Merkmalen $\underline{F}_{DH}$ ist der Merkmalsvektor gemäß (4.71) somit vollständig bestimmt.

Kapitel 5

Erkennung von Handposituren

Nach der Anwendung der im vorangehenden Kapitel beschriebenen Verfahren zur Segmentierung und Merkmalsextraktion liegt für eine zu verarbeitende Gebärdensequenz eine Folge zweidimensionaler Merkmalsvektoren $\underline{\mathbf{F}}_{\text{DH}}(t)$ vor. Hieraus wird die gesuchte dreidimensionale Positurresequenz wie nachfolgend beschrieben rekonstruiert.

Abschnitt 5.1 motiviert zunächst die Verwendung der offline generierten Positurrendatenbank vor dem Hintergrund der Laufzeitkomplexität des Systems. Neben dem Inhalt der Datenbank und der Struktur der einzelnen Einträge wird ein Zugriffsverfahren vorgestellt, das für jeden Merkmalsvektor $\underline{\mathbf{F}}_{\text{DH}}(t)$ eine Menge von Datenbankposituren $\tilde{\mathbf{P}}$ mit numerisch ähnlichen Merkmalen liefert und für die Anwendung in annähernd Echtzeit ausreichend schnell ist.

Für die Merkmale der Posituren $\tilde{\mathbf{P}}$ wird eine topologische Distanz zu den extrahierten Merkmalen berechnet. Die hierzu eingesetzten Algorithmen quantifizieren die Ähnlichkeit zweier Merkmalsvektoren und bilden eine zentrale Komponente der Positurerkennung. In jedem Einzelbild stellen die k bezüglich ihrer Merkmale ähnlichsten Posituren die Kandidatenmenge $\hat{\mathbf{P}}$ und damit das Ergebnis der Positurrekonstruktion dar. Diesen Prozess hat Abschnitt 5.2 zum Gegenstand.

Die k Positurrkandidaten je Einzelbild erlauben hinsichtlich der Interpretation der Gesamtsequenz eine Vielzahl von Positurrenfolgen. Um hieraus die wahrscheinlichste zu ermitteln, kommt der bereits in Abschnitt 4.5 behandelte Viterbi-Algorithmus zum Einsatz. Anschließend wird die resultierende Positurrenfolge geglättet, um eine flüssige Bewegung zu erhalten und das durch die endliche Größe der Positurrendatenbank bedingte „Quantisierungsrauschen“ zu kompensieren. Diese beiden Verarbeitungsschritte schließen die Positurerkennung ab und werden in Abschnitt 5.3 erläutert.

5.1 Positurrendatenbank

Der Ansatz der Positurerkennung durch Modellanpassung erfordert zur Erkennung einer Positur $\mathbf{P}_{\text{real}}$ mit den Merkmalen $\underline{\mathbf{F}}_{\text{DH,real}}$ das Rendern der 2D-Ansichten einer Menge von Kan-

didatenposituren $\underline{P}_{\text{Modell}}$, um diejenige zu identifizieren, deren Merkmale $\underline{F}_{\text{Modell}}$ möglichst nahe bei $\underline{F}_{\text{real}}$ liegen (vgl. (2.2)). Dies setzt voraus, dass die getestete Kandidatenmenge die Positur $\underline{P}_{\text{real}}$ bzw. eine hinreichend ähnliche Positur enthält.

Erfolgt das Rendern der Posituransichten online, lässt sich die Kandidatenmenge zwar dynamisch anpassen, aber die Anzahl der prüfbaren Posituren ist aufgrund der benötigten Rechenzeit sehr gering. Selbst wenn die Eingangspositur des letzten Einzelbildes bekannt ist, muss für jedes Gelenk und für die drei Ansichtswinkel je mindestens ein weiterer Wert betrachtet werden, um eine Bewegung erkennen zu können. Unter der vereinfachenden Annahme nur eines Freiheitsgrades für Finger und Daumen ergeben sich so bereits $2^{5+3} = 256$ Posituren, für deren Darstellung ein Standard-PC etwa drei Sekunden benötigt. Eine automatische Initialisierung oder das Verfolgen mehrerer Hypothesen sind auf diese Weise nicht praktikabel.

Aus diesem Grund werden die Modellansichten in der vorliegenden Arbeit offline gerendert und die zu Grunde liegenden Positurparameter $\underline{P}$ sowie die resultierenden Merkmale $\underline{F}_{\text{DH}}$ als Tupel in einer Datenbank gespeichert. Das Durchsuchen dieser Datenbank nach Einträgen, deren Merkmale denen der Eingangspositur ähnlich sind, ist ausreichend schnell, um in jedem Einzelbild etwa $2,5 \cdot 10^6$ Posituren zu berücksichtigen. Dies ermöglicht eine automatische Initialisierung und das parallele Verfolgen nahezu beliebig vieler Hypothesen.

Die in der Datenbank enthaltenen Posituren bilden das „Vokabular“ des Erkennungssystems. Während für die Verarbeitung korrekt ausgeführter Gebärden eine Beschränkung auf die in der entsprechenden Gebärdensprache verwendeten Handformen¹ (einschließlich der Übergänge zwischen diesen) sinnvoll wäre, erfordert das Tutorsystem die vollständige und gleichmäßige Abdeckung des Positurenraums, um jede mögliche Fehlausführung zu erkennen.

Abschnitt 5.1.1 beschreibt die Quantisierung des Positurenraums, die den Inhalt der Posituredatenbank bestimmt. Zur Berechnung der Merkmale der Datenbankposituren ist die Segmentierung der synthetischen Posituransichten erforderlich, auf die Abschnitt 5.1.2 eingeht. Anschließend erfolgt die Darstellung der Struktur eines einzelnen Datenbankeintrags in Abschnitt 5.1.3. Den zur Suche in der Datenbank entwickelten Algorithmus erläutert schließlich Abschnitt 5.1.4.

5.1.1 Quantisierung des Positurenraums

Die Diskretisierung des von den Positurparametern $\underline{B}$ und $\underline{\Phi}$ aufgespannten 23-dimensionalen Positurenraums hat zum Ziel, für jede beliebige Eingangspositur eine nahezu identische Datenbankpositur verfügbar zu machen. Gleichzeitig soll die Datenbank keine anatomisch unmöglichen Posituren enthalten (vgl. Abschnitt 3.2.4), da diese unnötig Speicherplatz belegen und lediglich zu Erkennungsfehlern führen können.

Um die Menge der Ansichtswinkel $\underline{\Phi}$ zu bestimmen, werden zunächst 105 Punkte so auf der

¹Je nach Autor und Gebärdensprache wird die Anzahl der auftretenden Handformen in der Literatur zwischen 20 und 150 geschätzt [Bec97].

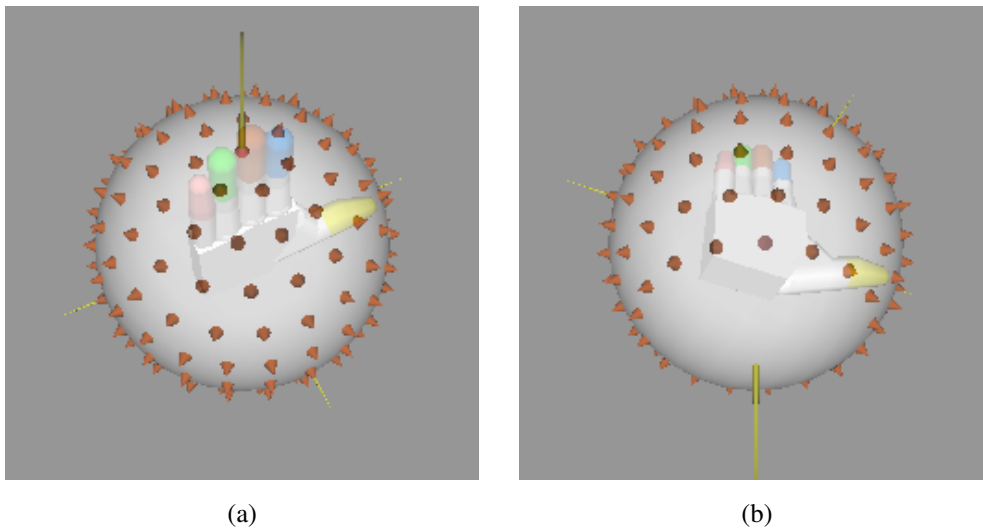


Abbildung 5.1: Verteilung der Ansichtspunkte auf der „nördlichen“ (a) und „südlichen“ (b) Hälfte der Ansichtskugel. Da Ansichten aus Richtung des Handgelenks nicht Bestandteil der Posituredatenbank sind, existieren in der Umgebung des Südpols keine Ansichtspunkte.

Oberfläche der Ansichtskugel verteilt, dass ihre Dichte möglichst konstant ist.² Diese Punkte definieren die diskreten Ansichtswinkel $\langle \phi_{\text{lat}}, \phi_{\text{lon}} \rangle$. Um Perspektiven aus der Richtung des Handgelenks, die in der Praxis nicht auftreten (vgl. Abschnitt 3.2.4), auszuschließen, sind Ansichten mit $\phi_{\text{lat}} > 65^\circ$ nicht zugelassen. Abbildung 5.1 visualisiert die resultierenden Ansichtspunkte.

Da die Rotation einer Positur in der Bildebene lediglich eine Rotation ihrer Merkmale um denselben Winkel bewirkt, lässt sich der Positurenraum in der Datenbank um diese Dimension reduzieren. Für alle Posituren wird $\phi_{\text{rot}} = 0$ gesetzt und die jeweils passende Rotation im Rahmen des Merkmalsvergleichs online bestimmt. Dies ist in Abschnitt 5.2.1 erläutert.

Zur Beschreibung der Handform wird für jeden Finger und den Daumen jeweils eine Liste mehrerer Gelenkstellungen explizit vorgegeben. Tabelle 5.1 zeigt die gewählten Beugungswinkel, aus denen zunächst $7 \cdot 9 \cdot 8 \cdot 9 \cdot 9 = 40824$ Kombinationen resultieren. Hiervon verbleiben nach Ausschluss anatomisch unmöglicher Handformen noch 23352 Kombinationen.

Die in der Tabelle enthaltenen Fingerbeugungswinkel entsprechen den in Abbildung 3.6 (s. Seite 69) gezeigten Ansichten. Die sieben Daumenstellungen sind in Abbildung 5.2 veranschaulicht.

Für jeden der 105 Ansichtspunkte enthält die Datenbank alle 23352 möglichen Handformen, so dass insgesamt 2.451.960 Posituren resultieren. Dabei ist zu berücksichtigen, dass die Rotation der Posituren in der Bildebene online erfolgt. Bei einer Auflösung von 10° stehen daher etwa $88 \cdot 10^6$ verschiedene Posituren sowie deren Merkmale zur Verfügung.

²Dieses Problem ist nichttrivial und wird näherungsweise mit Hilfe eines heuristischen Algorithmus gelöst.

Tabelle 5.1: Beugungswinkel der Finger und des Daumens. Jedes 4-Tupel untereinander stehender Werte beschreibt eine Finger-/Daumenstellung. Die Anzahl der möglichen Kombinationen (d. h. Handformen) beträgt $7 \cdot 9 \cdot 8 \cdot 9 \cdot 9 = 40824$.

	Gelenke	Beugungswinkel [°]								Stellungen
Daumen	CMC ^a	0	80	70	35	75	45	42		7
	CMC ^b	0	10	35	50	100	80	135		
	MCP	0	0	11	13	0	4	39		
	IP	0	0	39	18	0	39	28		
Zeigefinger	MCP ^a	0	-20	-16	-14	-11	-7	0	-1	9
	MCP ^b	0	0	9	18	36	54	90	86	
	PIP	0	0	0	14	36	50	0	36	
	DIP	0	0	0	8	21	27	0	35	
Mittelfinger	MCP ^a	0	0	0	0	0	0	0	0	8
	MCP ^b	0	14	27	41	54	90	86	90	
	PIP	0	9	23	36	54	0	36	90	
	DIP	0	5	13	28	40	0	35	36	
Ringfinger	MCP ^a	0	16	8	7	6	4	0	0	9
	MCP ^b	0	0	14	27	41	50	90	86	
	PIP	0	0	9	23	36	68	0	36	
	DIP	0	0	5	13	28	40	0	35	
Kleiner Finger	MCP ^a	0	36	23	19	15	11	0	1	9
	MCP ^b	0	0	14	27	41	54	90	86	
	PIP	0	0	9	23	36	59	0	36	
	DIP	0	0	5	13	28	40	0	35	

^aAdduktion/Abduktion

^bFlexion/Extension

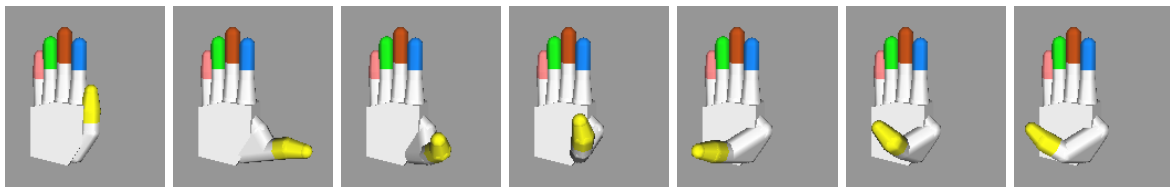


Abbildung 5.2: Ansicht der in Tabelle 5.1 beschriebenen Daumenstellungen.

5.1.2 Segmentierung der synthetischen Posituransichten

Analog zu der in Kapitel 4 beschriebenen Verarbeitung realer Ansichten erfordert die Berechnung der 2D-Merkmale $\underline{F}_{DH}$ der synthetischen Posituransichten zunächst eine Segmen-

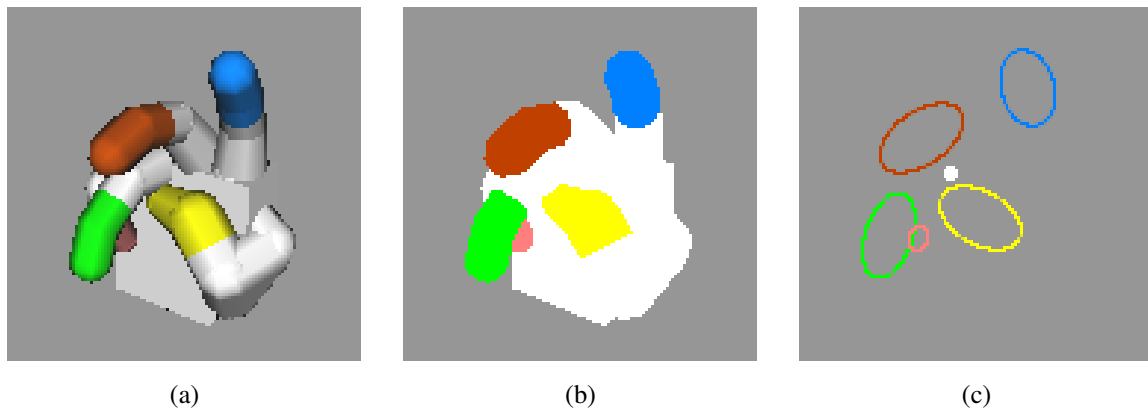


Abbildung 5.3: Synthetische Ansicht derselben Positur bei aktivierter (a) bzw. deaktivierter (b) Beleuchtung. Die Segmentierung der Markerregionen sowie die anschließende Approximation der Randkurven durch Ellipsen gemäß Abschnitt 4.3.5 liefert das in (c) gezeigte Ergebnis. Der weiße Punkt in der Bildmitte kennzeichnet den gemeinsamen Markerschwerpunkt s_{DH} . Die Ausschnittauflösung beträgt 120×120 Pixel.

tierung. Diese ist jedoch vergleichsweise trivial und erfolgt grundsätzlich fehlerfrei.

Die verwendete Grafikbibliothek *OpenGL* erlaubt das Deaktivieren des Beleuchtungseffekts, der die bisher in allen Abbildungen des Handmodells sichtbaren Schatten und Farbverläufe erzeugt. Dadurch erhalten alle Marker sowie die unmarkierten Bereiche des Handschuhs jeweils exakt den als Grundfarbe vorgegebenen Farbwert. Für den menschlichen Beobachter ist die Interpretation einer solchen Ansicht aufgrund der fehlenden Tiefeninformationen schwieriger, die Segmentierungsaufgabe reduziert sich hingegen auf den Vergleich jedes Pixels mit den Markerfarben sowie die anschließende Bestimmung der Randkurven der resultierenden Regionen. In Abbildung 5.3 ist eine synthetische Posituransicht bei aktivierter und bei deaktivierter Beleuchtung sowie das Ergebnis der Segmentierung visualisiert.

Aufgrund der Skalierungsunabhängigkeit des Merkmalsvektors ist die Bildauflösung der synthetischen Handansicht frei wählbar. Sie beträgt in der vorliegenden Arbeit 256×256 , wobei die Hand maximal etwa 13% dieser Fläche belegt. Ein Erhöhen der Auflösung hat auf $\underline{F}_{DH}$ keinen nennenswerten Einfluss, verlangsamt das Rendern der Ansicht jedoch deutlich.

5.1.3 Struktur der Einträge

Um die spätere Verarbeitung der Datenbankeinträge zu beschleunigen, geschieht die Speicherung in einer Form, die nach dem Einlesen möglichst direkt verwendbar ist, d. h. keine weiteren Transformationen erfordert. Darüber hinaus soll der pro Eintrag benötigte Speicherplatz so gering sein, dass die Datenbank vollständig im Arbeitsspeicher gehalten werden kann.

Für den Vergleich zweier Merkmalsvektoren, der in Abschnitt 5.2 behandelt wird, ist eine Abtastung der elliptischen Randkurven der Markerregionen notwendig. Dies erfolgt mit je acht näherungsweise gleichverteilten Punkten $\langle \underline{e}_1, \dots, \underline{e}_8 \rangle$. Vier dieser Punkte ergeben sich

aus den Schnittpunkten mit Haupt- und Nebenachse, weitere vier werden so gewählt, dass ihre Projektion auf die Hauptachse den Abstand

$$x = \frac{r_1}{2} + \frac{r_2}{4} \quad (5.1)$$

vom Mittelpunkt $\underline{s}$ hat. Die Entfernung von der Hauptachse selbst beträgt mit der Ellipsengleichung

$$y = r_2 \sqrt{1 - \left(\frac{1}{2} + \frac{r_2}{4r_1} \right)^2} \quad (5.2)$$

In Abbildung 5.4 ist die Randkurvendiskretisierung für verschiedene Radienverhältnisse veranschaulicht.

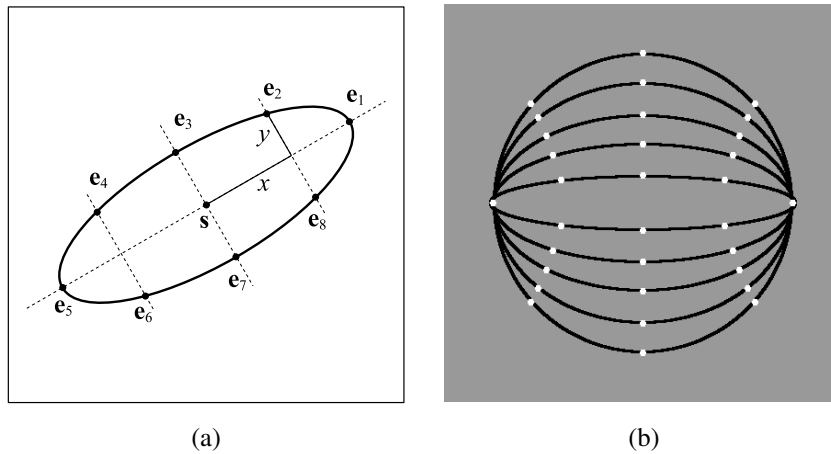


Abbildung 5.4: Diskretisierung der Ellipsenrandkurve mit acht Punkten unter Verwendung von (5.1) und (5.2) (a) sowie Beispiele für verschiedene Radienverhältnisse (b).

Pro Marker enthält jeder Datenbankeintrag in Polarkoordinaten die Punkte $\underline{s}' = \langle r', \varphi \rangle$ sowie $\underline{e}_1$ bis $\underline{e}_3$ relativ zu $\underline{s}'$. Auf diese Weise lassen sich $\underline{e}_4$ bis $\underline{e}_8$ durch Spiegelung an den Achsen mit wenigen Operationen berechnen. Die Quantisierung der Polarkoordinaten verwendet 4 Byte für den Radius und 1 Byte für den Azimut. Für $\underline{e}_3$ ist die Speicherung des Azimut nicht notwendig, da er sich aus dem des Punktes $\underline{e}_1$ durch Addition von 90° ergibt.

Weiterhin wird der von mehreren Algorithmen benötigte Flächeninhalt a' gespeichert, um die Multiplikation $\pi r'_1 r'_2$ nicht zur Laufzeit ausführen zu müssen. Die Quantisierung erfolgt hier mit 1 Byte.

Da die Positurparameter $\langle \phi_{\text{lat}}, \phi_{\text{lon}} \rangle$ sowie die Beugungswinkel der Gelenke abzählbar sind, lassen sie sich indizieren und über eine Lookup-Tabelle rekonstruieren. Die entsprechenden Indizes werden mit 4 Byte bzw. je Finger 1 Byte quantisiert.

Tabelle 5.2 fasst die Felder eines Datenbankeintrags zusammen und gibt den jeweils benötigten Speicherplatz in der Datenbankdatei sowie im Arbeitsspeicher an. Hierbei auftretende

Tabelle 5.2: Struktur eines Datenbankeintrags. In der letzten Zeile ist die Gesamtlänge bei Verwendung von sechs Markern angegeben.

	Feld	Speicherplatz	
		Datenbankdatei	Arbeitsspeicher
Positurparameter	Ansichtswinkel $\langle \phi_{\text{lat}}, \phi_{\text{lon}} \rangle$	4 Byte	4 Byte
	20 Beugungswinkel	5 Byte	8 Byte
Merkmale je Marker	Flächeninhalt a'	1 Byte	4 Byte
	Schwerpunkt $\underline{s}' = \langle r', \varphi \rangle$	5 Byte	8 Byte
	Randkurvenpunkt $\underline{e}_1$	5 Byte	8 Byte
	Randkurvenpunkt $\underline{e}_2$	5 Byte	8 Byte
	Randkurvenpunkt $\underline{e}_3$	4 Byte	4 Byte
Gesamt		129 Byte	204 Byte

Unterschiede entstehen durch die verwendete Hardwarearchitektur, in der das kürzeste Fließkommaformat 4 Byte und Felder ein Vielfaches von 4 Byte belegen. Für nicht sichtbare Marker werden alle Merkmale zu Null gesetzt; die Länge des Eintrags bleibt konstant.

Die Größe der Posituredatenbank ist somit $2.335.200 \cdot 129 \text{ Byte} \approx 302 \text{ MB}$ bzw. im Arbeitsspeicher $2.335.200 \cdot 204 \text{ Byte} \approx 477 \text{ MB}$.

5.1.4 Zugriffsverfahren

Die Posituredatenbank besteht aus einer „flachen Datei“ (*flat file*), die ausschließlich die in Tabelle 5.2 beschriebenen Einträge in unmittelbarer Folge enthält, sowie der Lookup-Tabellen der Positurparameter und einer nachfolgend beschriebenen Indexstruktur. Aufgrund der konstanten Eintragslänge lässt sich die Position jedes Eintrags in der Datenbankdatei direkt berechnen.

Die Positurrekonstruktion beinhaltet die Aufgabe, in jedem Einzelbild eine Menge von Positurkandidaten $\hat{P}$ zu liefern, deren Merkmale den aus dem Eingangsbild extrahierten möglichst ähnlich sind. Zur Quantifizierung dieser Ähnlichkeit dient die im nächsten Abschnitt beschriebene Merkmalsdifferenz $\Delta_{\underline{F}}$. Allerdings ist Berechnung von $\Delta_{\underline{F}}$ für sämtliche Datenbankeinträge aufgrund der resultierenden Zeitkomplexität nicht praktikabel. Daher kommt ein im Folgenden dargestelltes Zugriffsverfahren zum Einsatz, das zunächst eine Obermenge $\tilde{P} \supseteq \hat{P}$ identifiziert, für die einige aussagekräftige Merkmale numerisch nahe bei den gesuchten liegen. Diese Eigenschaft ist notwendig, jedoch nicht hinreichend, für eine niedrige Merkmalsdifferenz. Für diese Positurenmenge $\tilde{P}$ erfolgt die tatsächliche Berechnung von $\Delta_{\underline{F}}$ und die anschließende Selektion der k besten Einträge $\hat{P}$.

Der Parameter k ist innerhalb einer Gebärde konstant und gibt die Anzahl der parallel verfolg-

ten Hypothesen an. Falls $|\tilde{P}| \leq k$, so gilt $\hat{P} = \tilde{P}$. Dies wirkt sich auf die folgende Verarbeitung jedoch nicht aus – die Merkmalsdifferenz $\Delta_{\mathbf{F}}$ wird auch in diesem Fall für alle Elemente in $\hat{P}$ berechnet, weil sie für die anschließende Mehrdeutigkeitsauflösung erforderlich ist (s. Abschnitt 5.3.1).

Da die Datenbank nach ihrer Erstellung nicht mehr verändert wird, ist eine statische Indexstruktur für den Zugriff geeignet. Die Einträge werden daher mit Hilfe eines höhenbalancierten Baums nach ihren Merkmalen sortiert [Sch05]. Zunächst erfolgt die Auswahl m einzelner Merkmale f_i , für die die oben beschriebene numerische Ähnlichkeit zu fordern ist. Hierfür eignen sich die normierte Markerfläche a' sowie der normierte Radius r' . Beginnend mit dem Wurzelknoten, der der gesamten Datenbank entspricht, werden die Einträge jedes Knotens auf der i -ten Baumebene nach dem entsprechenden Merkmal f_i sortiert und in n_i disjunkte Partitionen mit jeweils gleich vielen Elementen unterteilt. Der resultierende Baum hat somit die Höhe $h = m + 1$.

Jedes Blatt des Baums enthält im Allgemeinen mehrere Einträge, für die $\Delta_{\mathbf{F}}$ anschließend explizit zu berechnen ist. Da die extrahierten Merkmale verrauscht sind, ist es jedoch nicht ausreichend, lediglich das exakt passende Blatt zu identifizieren. Vielmehr muss die Suche in benachbarte Teilbäume verzweigen, wenn ein Merkmal nahe an den Intervallgrenzen liegt. Um diese Nähe zu quantifizieren, wird für jedes Intervall die Varianz der enthaltenen Einträge bestimmt und zusätzlich gespeichert. Auf diese Weise liefert ein Datenbankzugriff in den meisten Fällen die Kandidaten mehrerer Blätter, wobei die Anzahl der Kandidaten pro Blatt und damit die Mächtigkeit der Menge $\tilde{P}$ von den Verzweigungsgraden n_i abhängt.

Die Festlegung der Merkmale f_i und der Verzweigungsgrade n_i hat wesentlichen Einfluss auf die Leistung der Positurerkennung. In Abschnitt 7.1.2.3 erfolgt eine entsprechende Evaluation.

5.2 Merkmalsvergleich

Für die wie im letzten Abschnitt beschrieben ermittelte Positurenmenge $\tilde{P}$ wird die Differenz $\Delta_{\mathbf{F}}$ zu den extrahierten Merkmalen berechnet, um die k besten Kandidaten $\hat{P}$ zu finden. Aufgrund der Mehrdeutigkeit der 2D-Ansicht lässt sich nicht verhindern, dass sowohl $\tilde{P}$ als auch $\hat{P}$ Posituren enthalten, die trotz ähnlicher Merkmale keine oder nur geringe Ähnlichkeit zu der beobachteten Positur haben. Dieses Problem wird im nächsten Verarbeitungsschritt behandelt (s. Abschnitt 5.3.1).

Innerhalb der Positurenmenge $\tilde{P}$ sollte $\Delta_{\mathbf{F}}$ möglichst gut mit $\Delta_{\mathbf{P}}$ korrelieren. Bei stark verschiedenen Merkmalsvektoren ist dies hingegen nicht erforderlich, da diese innerhalb von $\tilde{P}$ nicht auftreten.

Die Zusammenstellung einer Kandidatenmenge $\hat{P}$ erfolgt für jedes Einzelbild, d. h. 25 mal pro Sekunde. Sie beansprucht den größten Anteil der gesamten Laufzeit, so dass die Geschwindigkeit der Positurrekonstruktion primär durch die Geschwindigkeit der Merkmalsdifferenz-

berechnung bestimmt ist und zeitaufwändige Algorithmen entsprechend zu vermeiden sind.

Im Folgenden steht $\underline{\mathbf{F}}_{\text{DH}}^{\text{ex}}$ für die aus dem Eingangsbild extrahierten Merkmale der zu rekonstruierenden Positur, während $\underline{\mathbf{F}}_{\text{DH}}^{\text{db}}$ einen aus der Posituredatenbank gelesenen Merkmalsvektor kennzeichnet.

5.2.1 Bestimmung der Rotation

Da die Posituren in der Datenbank mit $\phi_{\text{rot}} = 0$ gespeichert sind, ist vor dem eigentlichen Vergleich die Rotation zu ermitteln, die die Merkmale $\underline{\mathbf{F}}_{\text{DH}}^{\text{db}}$ mit $\underline{\mathbf{F}}_{\text{DH}}^{\text{ex}}$ optimal zur Deckung bringt. Dazu werden die beiden Merkmalsvektoren zunächst am jeweiligen gemeinsamen Schwerpunkt aller Marker ausgerichtet, so dass $\underline{\mathbf{s}}_{\text{DH}}^{\text{db}} = \underline{\mathbf{s}}_{\text{DH}}^{\text{ex}}$ gilt. Dieser Punkt bildet im Folgenden das Rotationszentrum.

Zur Bestimmung des Rotationswinkels dienen Schwerpunkte und Flächeninhalte der einzelnen Markerregionen. Während $\underline{\mathbf{F}}_{\text{DH}}^{\text{ex}}$ konstant bleibt, berechnen sich die rotierten Schwerpunkte der aus der Datenbank gelesenen Merkmale zu

$$\underline{\mathbf{s}}_i^{\text{db}}(\phi_{\text{rot}}) = \langle r_i^{\text{db}}, \varphi_i^{\text{db}} + \phi_{\text{rot}} \rangle \quad (5.3)$$

Der gesuchte Winkel $\hat{\phi}_{\text{rot}}$ ergibt sich durch Minimieren der quadrierten euklidischen Abstände der Schwerpunkte $\underline{\mathbf{s}}_i' = \langle r_i', \varphi_i' \rangle$, jeweils gewichtet mit dem Produkt der normierten Flächeninhalte:

$$\hat{\phi}_{\text{rot}} = \underset{\phi}{\operatorname{argmin}} \sum_{i=1}^n a_i^{\text{ex}} a_i^{\text{db}} \left\| \underline{\mathbf{s}}_i^{\text{ex}} - \underline{\mathbf{s}}_i^{\text{db}}(\phi) \right\|^2 \quad (5.4)$$

Die Gewichtung erhöht den Einfluss großer, d. h. deutlich sichtbarer Regionen, während nicht sichtbare Regionen mit $a' = 0$ ignoriert werden. Der Winkel $\hat{\phi}_{\text{rot}}$ liefert keinen Beitrag zu $\Delta_{\underline{\mathbf{F}}}$, sondern schafft erst die Voraussetzung für den folgenden Vergleich. Abbildung 5.5 visualisiert die Rotation der Modellmerkmale anhand eines Beispiels. Im Folgenden wird davon ausgegangen, dass $\underline{\mathbf{F}}_{\text{DH}}^{\text{db}}$ gemäß (5.4) rotiert ist.

5.2.2 Berechnung der Merkmalsdifferenz

Nach der Rotation der Modellmerkmale liegen zwei Merkmalsvektoren vor, die aufgrund der Normierung direkt vergleichbar sind. Zwei Ansätze zur Berechnung von $\Delta_{\underline{\mathbf{F}}}$ sind in diesem Abschnitt diskutiert. Eine quantitative Evaluation beider Verfahren findet sich in [Dic06].

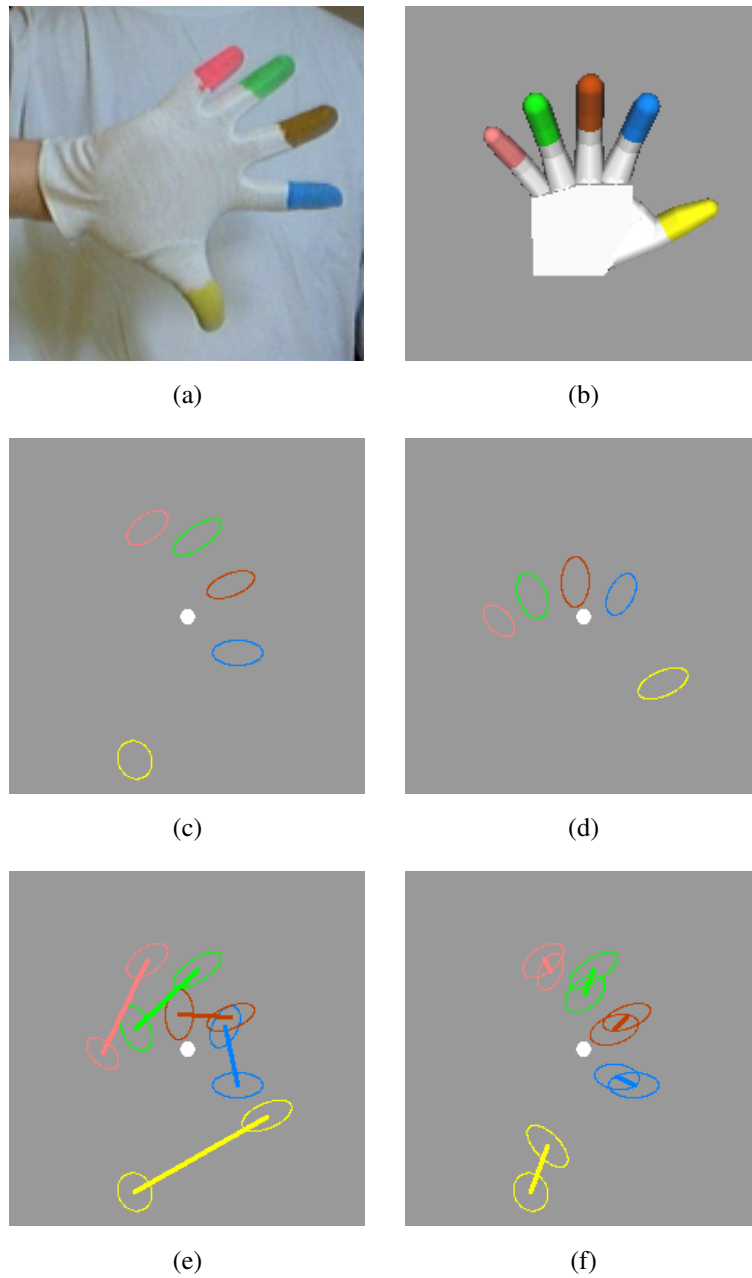


Abbildung 5.5: Eingangspositur (a) und Modellpositur (b) sowie entsprechende Merkmale (c bzw. d). Zur Veranschaulichung sind die Handformen nicht maximal ähnlich. Vor der Berechnung der Merkmalsdifferenz $\Delta_{\mathbf{F}}$ erfolgt zunächst ein Ausrichten am jeweiligen gemeinsamen Schwerpunkt $\underline{s}_{\text{DH}}$ (e) – in Weiß dargestellt – und eine Rotation zur Minimierung von (5.4). Im gezeigten Beispiel ergibt sich $\hat{\phi}_{\text{rot}} = 290^\circ$ (f).

5.2.2.1 Parametrische Merkmalsdifferenz

Nahe liegend scheint zunächst die Berechnung der Mahalanobisdistanz im Merkmalsraum [Sch96]:

$$\Delta_{\mathbf{F}}(\mathbf{F}_{\text{DH}}^{\text{ex}}, \mathbf{F}_{\text{DH}}^{\text{db}}) = (\mathbf{F}_{\text{DH}}^{\text{ex}} - \mathbf{F}_{\text{DH}}^{\text{db}})^T \Sigma^{-1} (\mathbf{F}_{\text{DH}}^{\text{ex}} - \mathbf{F}_{\text{DH}}^{\text{db}}) \quad (5.5)$$

Darin steht Σ für die Kovarianzmatrix des Merkmalsvektors $\mathbf{F}_{\text{DH}}^{\text{db}} = \mathbf{F}_{\text{DH}}(\underline{\mathbf{P}})$, gemessen in einer Umgebung ϵ der Kandidatenpositur $\underline{\mathbf{P}}$. Diesem Ansatz stehen mehrere Schwierigkeiten entgegen, die nachfolgend erläutert sind.

Die Umgebung ϵ enthält im kontinuierlichen Positorenraum unendlich viele Positoren, so dass in der Praxis eine Diskretisierung erforderlich ist. Um Σ zuverlässig zu ermitteln, sind möglichst viele Positoren erforderlich. Selbst bei Betrachtung von lediglich zwei benachbarten Beugungsgraden pro Finger und insgesamt vier Nachbaransichten ergeben sich $3^5 \cdot 4 = 972$ Positoren, die für jede Datenbankpositur zu rendern sind. Die hierfür aufzuwendende Rechenzeit beträgt auf einem Standard-PC, der pro Positur 12 ms benötigt, bei einer Datenbankgröße von $2,5 \cdot 10^6$ Positoren etwa ein Jahr.

Aus diesem Grund muss auf die Generierung zusätzlicher Positoren verzichtet und die ϵ -Umgebung der Positur $\underline{\mathbf{P}}$ durch die Menge der in der Datenbank enthaltenen nächsten Nachbarn $N(\underline{\mathbf{P}})$ approximiert werden:

$$N(\underline{\mathbf{P}}) = \{\underline{\mathbf{P}}' \mid \underline{\mathbf{P}}' \in P^{\text{db}} \wedge \Delta_{\underline{\mathbf{P}}}(\underline{\mathbf{P}}, \underline{\mathbf{P}}') < \theta\} \quad (5.6)$$

wobei θ ein geeignet zu wählender Schwellwert ist und P^{db} die Gesamtmenge der Datenbankposituren bezeichnet. Der Rechenaufwand sinkt hierdurch erheblich, da die Merkmale $\mathbf{F}_{\text{DH}}(\underline{\mathbf{P}}') \forall \underline{\mathbf{P}}' \in P^{\text{db}}$ bereits vorliegen. Da Positoren aus $N(\underline{\mathbf{P}})$ für realistische Schwellwerte θ größere Unterschiede aufweisen als Positoren aus ϵ , resultieren jedoch höhere Varianzen, die

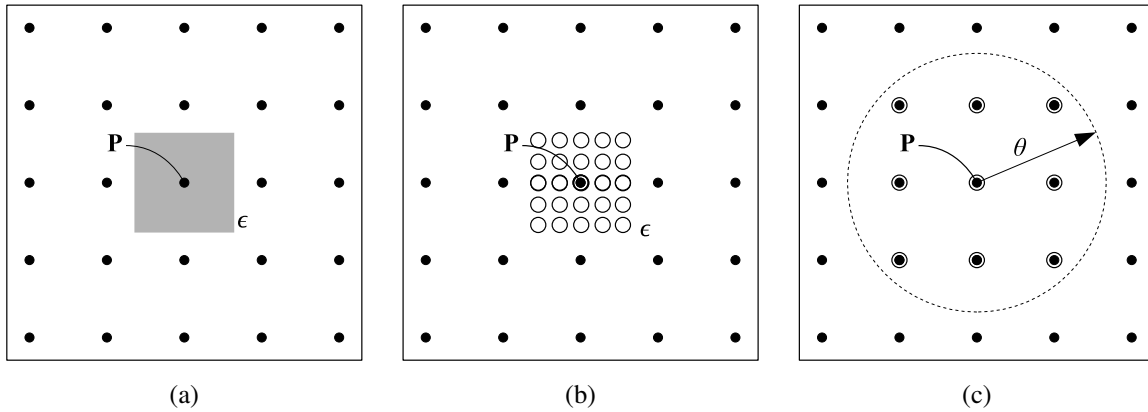


Abbildung 5.6: Auswahl von Posituren zur Berechnung der Kovarianzmatrix Σ . Punkte symbolisieren Positoren aus P^{db} , Kreise die zur Bestimmung von Σ verwendeten Positoren. Im kontinuierlichen Fall (a) resultiert Σ aus der Betrachtung einer Umgebung ϵ um die Kandidatenpositur $\underline{\mathbf{P}}$, die unendlich viele Positoren enthält. Für die praktische Implementierung ist eine Diskretisierung von ϵ erforderlich (b), mit vertretbarem Aufwand lässt sich jedoch lediglich eine Approximation durch die Betrachtung benachbarter Positoren $\underline{\mathbf{P}}'$ mit $\Delta_{\underline{\mathbf{P}}}(\underline{\mathbf{P}}, \underline{\mathbf{P}}') < \theta$ realisieren (c).

bezüglich $\underline{\mathbf{P}}$ weniger spezifisch sind.

Abbildung 5.6 visualisiert die ϵ -Umgebung in allgemeiner und diskreter Form sowie die Approximation nach (5.6).

5.2.2.2 Topologische Merkmalsdifferenz

Ein topologischer Vergleich der Randkurven der Markerregionen umgeht dieses Problem und ist daher für die Implementierung von $\Delta_{\underline{\mathbf{F}}}$ besser geeignet. Grundlage hierfür bildet die ungerichtete Hausdorff-Distanz H , die die Verschiedenheit zweier Kurven A und B quantifiziert:

$$H(A, B) = \max \{h(A, B), h(B, A)\} \quad (5.7)$$

$$h(A, B) = \max_{\mathbf{a} \in A} \left(\min_{\mathbf{b} \in B} \|\mathbf{a} - \mathbf{b}\| \right) \quad (5.8)$$

Darin steht $h(A, B)$ für die gerichtete Hausdorff-Distanz. Sie lässt sich anschaulich interpretieren als der Radius des größten Kreises, der um einen beliebig wählbaren Punkt der Kurve A gezogen werden kann, ohne B zu schneiden oder einzuschließen. Abbildung 5.7 visualisiert dies am Beispiel zweier Ellipsen.

Auf Basis der in $\underline{\mathbf{F}}_{\text{DH}}$ enthaltenen Ellipsenradien und -mittelpunkte ist theoretisch eine exakte Bestimmung von H möglich. Aus Geschwindigkeitsgründen wird jedoch die in Abbildung 5.4 gezeigte Randkurvendiskretisierung verwendet, so dass lediglich 64 diskrete Abstände zu berechnen sind. Im Unterschied zur Mahalanobisdistanz sind hierzu keine Metadaten – wie etwa Varianzen – erforderlich. Abbildung 5.8 zeigt die Hausdorff-Distanzen der eingangs betrachteten Merkmalsvektoren (s. Abbildung 5.5).

Für die Merkmalsdifferenz werden die ungerichteten Hausdorff-Distanzen quadriert, um deutliche Abweichungen stärker zu gewichten als geringe. Mit der normierten parametrischen

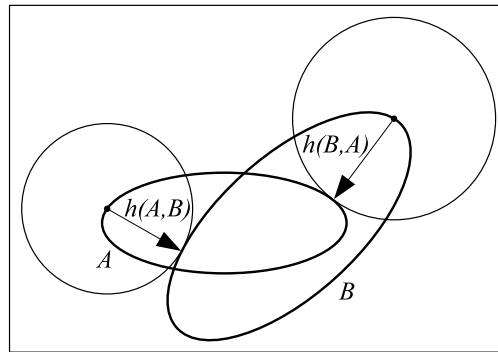


Abbildung 5.7: Hausdorff-Distanz zweier Ellipsen A und B . Für die gerichteten Distanzen gilt in diesem Beispiel $h(A, B) \approx 0,8h(B, A)$. Die ungerichtete Hausdorff-Distanz ergibt sich zu $H(A, B) = \max \{h(A, B), h(B, A)\} = h(B, A)$.

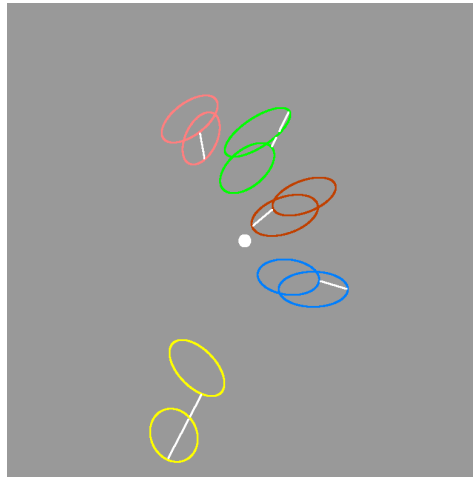


Abbildung 5.8: Anwendung der Hausdorff-Distanz zum Vergleich zweier Merkmalsvektoren. Die weißen Linien verbinden die relevanten Punkte der verglichenen Regionen, d. h. die Punkte $\underline{a}$ und $\underline{b}$, die die Extremwerte in (5.8) darstellen.

Regionenbeschreibung in Form der Merkmale $\underline{\mathbf{F}}_i$ (vgl. (4.79)) ergibt sich die topologische Definition der Merkmalsdifferenz $\Delta_{\underline{\mathbf{F}}}$ zu

$$\Delta_{\underline{\mathbf{F}}}(\underline{\mathbf{F}}_{\text{DH}}^{\text{ex}}, \underline{\mathbf{F}}_{\text{DH}}^{\text{db}}) = \sum_i H^2(\underline{\mathbf{F}}_i^{\text{ex}}, \underline{\mathbf{F}}_i^{\text{db}}) \quad (5.9)$$

Bisher wurde implizit vorausgesetzt, dass alle Marker sichtbar sind. Ist ein Marker i sowohl in $\underline{\mathbf{F}}_{\text{DH}}^{\text{ex}}$ als auch in $\underline{\mathbf{F}}_{\text{DH}}^{\text{db}}$ verdeckt, ergibt sich sein Beitrag zu (5.9) zu Null. Fehlt der Marker in genau einer der beiden Ansichten, wird mit Hilfe des Handmodells eine Region generiert, deren Hausdorff-Distanz zur sichtbaren Markerregion die Tatsache, dass der Marker nicht sichtbar ist, geeignet bewertet. Dies bedeutet insbesondere Stabilität, d. h. dass $\Delta_{\underline{\mathbf{F}}}$ sich nicht sprunghaft ändert, wenn eine Markerregion mit kleiner Fläche sichtbar oder unsichtbar wird. Verdeckte Regionen, die aus anatomischen Gründen eine wesentliche Änderung der Handpositur erfordern, um sichtbar bzw. unsichtbar zu werden, sollten hingegen entsprechend mehr zu $\Delta_{\underline{\mathbf{F}}}$ beitragen.

Abbildung 5.9 visualisiert die Situation einer einseitig fehlenden Markierung. Im Folgenden ist zu unterscheiden, ob der Merkmalsvektor, in dem die Markierung fehlt, aus der Posituren-datenbank gelesen wurde und die zugehörige Handpositur somit bekannt ist, oder ob er ein Eingangsbild repräsentiert, d. h. die entsprechende Handpositur nicht vorliegt.

Fehlende Markierung im Datenbankeintrag Wenn der aus der Datenbank gelesene Merkmalsvektor $\underline{\mathbf{F}}_{\text{DH}}^{\text{db}}$ die nicht sichtbare Markierung aufweist (im Beispiel Abbildung 5.9b), ist die zugehörige Handpositur bekannt. Somit lässt sich die eigentliche Position der nicht sichtbaren Markierung im Bild berechnen durch Rendern einer Positur, in der alle Komponenten des Handmodells bis auf die betroffene Markierung fehlen (d. h. „durchsichtig“ sind), und anschließendes Segmentieren der nun nicht mehr verdeckten Markierung.

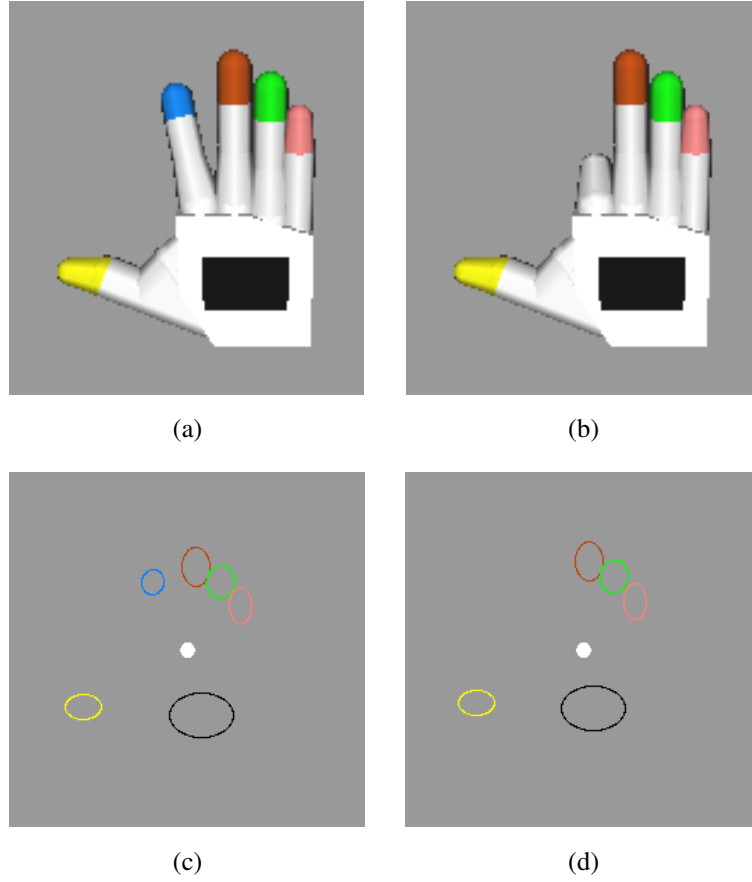


Abbildung 5.9: Zwei Posituren, in denen der Zeigefingermarker sichtbar (a) bzw. unsichtbar (b) ist, sowie deren Merkmale (c, d). Die Berechnung der Merkmalsdifferenz $\Delta_{\mathbf{F}}$ erfordert eine spezielle Behandlung der einseitig fehlenden Zeigefingermerkmale.

Auf diese Weise ergibt sich eine Region i mit den Merkmalen $\tilde{\mathbf{F}}_i^{\text{db}}$ gemäß (4.79). Für die Merkmalsdifferenz wird diese Region als punktförmig angenommen, so dass mit Ausnahme der Schwerpunktskoordinaten $\tilde{\mathbf{s}}_i^{\text{db}} = \langle \tilde{r}_i^{\text{db}}, \tilde{\varphi}_i^{\text{db}} \rangle$ alle Merkmale Null sind:

$$\tilde{\mathbf{F}}_i^{\text{db}} = \begin{bmatrix} 0 \\ \tilde{r}_i^{\text{db}} \\ \tilde{\varphi}_i^{\text{db}} \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (5.10)$$

Aufgrund von $a_i^{\text{db}} = 0$ bleibt die Normierungskonstante γ^{db} unbeeinflusst und die Normierung weiterhin gültig. Die durch $\tilde{\mathbf{F}}_i^{\text{db}}$ beschriebene Ellipse ist entartet, dies wirkt sich auf Berechnung und Aussagekraft der Hausdorff-Distanz jedoch nicht aus. Abbildung 5.10 visualisiert die resultierende Merkmalsdifferenz für das in Abbildung 5.9 gezeigte Beispiel.

Die beschriebene Vorgehensweise ist neben der eingangs geforderten Stabilität auch trenn-

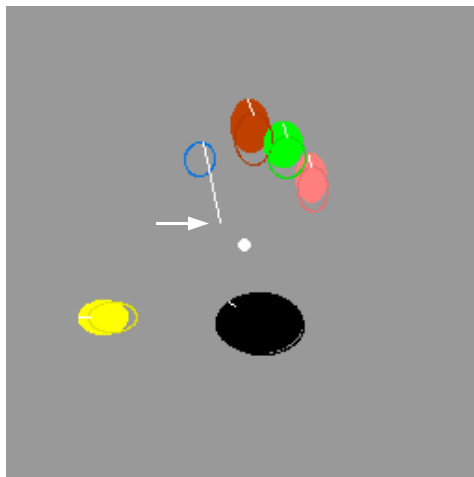


Abbildung 5.10: Merkmalsdifferenz bei einseitig fehlender Markierung im aus der Datenbank gelesenen Merkmalsvektor (vgl. Abbildung 5.9c/d). Der Schwerpunkt $\tilde{\mathbf{s}}^{\text{db}}$ der verdeckten Zeigefingermarkierung befindet sich links oberhalb des gemeinsamen Schwerpunkts $\mathbf{s}_{\text{DH}}$ und ist mit einem Pfeil gekennzeichnet. An den Distanzlinien ist erkennbar, dass hieraus der größte Beitrag zu $\Delta_{\mathbf{F}}$ resultiert. Durch die Verschiebung des gemeinsamen Schwerpunkts sind die sichtbaren Markierungen nicht exakt deckungsgleich und tragen ebenfalls zu $\Delta_{\mathbf{F}}$ bei.

scharf, da durch $r_{1,i}^{\text{db}} = r_{2,i}^{\text{db}} = 0$ in jedem Fall eine positiver Differenzbeitrag resultiert, selbst wenn die Merkmalsvektoren mit Ausnahme der fehlenden Markierung i identisch sind und darüber hinaus $\tilde{\mathbf{F}}_i^{\text{db}}$ mit den aus dem Eingangsbild extrahierten Merkmalen $\mathbf{F}_i^{\text{ex}}$ übereinstimmen sollte.

Fehlende Markierung im Eingangsbild Falls der aus dem Eingangsbild berechnete Merkmalsvektor $\mathbf{F}_{\text{DH}}^{\text{ex}}$ die nicht sichtbare Markierung i aufweist, ist die zugehörige Handpositur noch unbekannt und die oben beschriebene Strategie somit nicht anwendbar. In diesem Fall wird untersucht, wie weit sich die in der Datenbank gespeicherte Positur $\mathbf{P}$, mit deren Merkmalen $\mathbf{F}_{\text{DH}}^{\text{db}}$ der Vergleich erfolgt, ändern müsste, um die Markierung i zu verdecken und darin $\mathbf{F}_{\text{DH}}^{\text{ex}}$ zu entsprechen.

Für die in $\mathbf{F}_{\text{DH}}^{\text{db}}$ sichtbare Markierung i wird diejenige Positur $\mathbf{Q}$ bestimmt, die sich lediglich in den Beugungsgraden $\mathbf{B}_i$ des entsprechenden Fingers von $\mathbf{P}$ unterscheidet und die Fläche a'_i minimiert. Falls sich hieraus keine eindeutige Lösung ergibt, wird $\mathbf{Q}$ so gewählt, dass ferner der euklidische Abstand der Markerschwerpunkte $\mathbf{s}'_i$ zwischen $\mathbf{P}$ und $\mathbf{Q}$ minimal ist. Die zweite Bedingung kommt insbesondere dann zum Tragen, wenn a'_i für $\mathbf{Q}$ Null, d. h. die Markierung i in $\mathbf{Q}$ verdeckt ist. Die entsprechenden Koordinaten $\tilde{\mathbf{s}}'_i$ sind in diesem Fall wie im letzten Abschnitt beschrieben zu ermitteln. Abbildung 5.11 veranschaulicht das Prinzip anhand eines Beispiels.

Hieraus resultiert eine anatomisch begründete Schätzung bezüglich der im Sinne niedriger Differenz $\Delta_{\mathbf{F}}$ günstigsten Position der verdeckten Markerfläche i in $\mathbf{F}_{\text{DH}}^{\text{ex}}$. Wie im zuvor diskutierten Fall enthält $\tilde{\mathbf{F}}_i^{\text{ex}}$ lediglich den Schwerpunkt dieser Fläche, die übrigen Merkmale

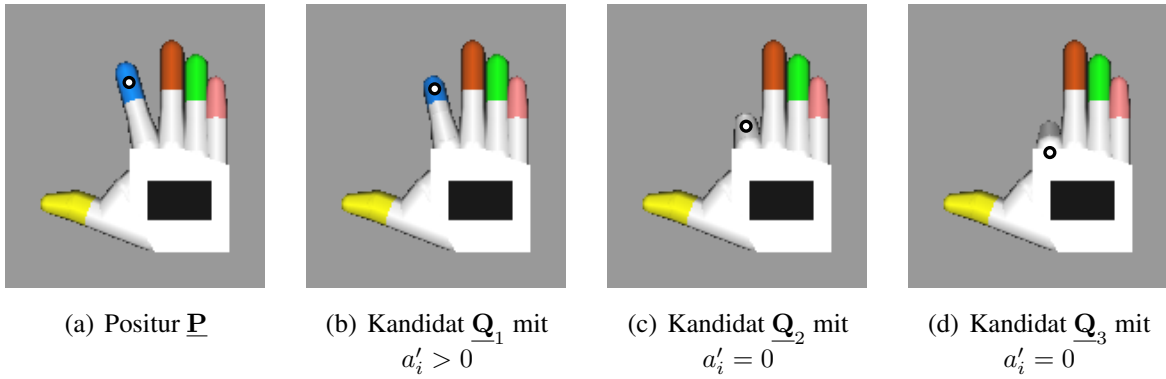


Abbildung 5.11: Bestimmung einer Positur $\underline{Q}$, die sich von $\underline{P}$ (a) lediglich in der Beugung des Zeigefingers unterscheidet. Unter den Kandidaten (b, c, d) wird derjenige gewählt, dessen Fläche a'_i minimal ist (c, d) und für den der Markerschwerpunkt $\tilde{s}'_i$ (gekennzeichnet durch einen weißen Punkt) möglichst nah an dem für $\underline{P}$ ermittelten liegt, d. h. in diesem Beispiel $\underline{Q}_2$ (c).

sind Null. Anschließend erfolgt die Berechnung der Hausdorff-Distanz wie oben beschrieben.

Eine Markierung, die nicht am Finger, sondern beispielsweise auf dem Handrücken oder am Handgelenk angebracht ist, kann durch Variation der Beugungsgrade $\underline{B}$ nicht beeinflusst werden. Entsprechen die einseitig fehlenden Merkmale $\underline{F}_i^{\text{ex}}$ einer solchen Markierung, wird daher in Abwandlung der obigen Methode der Schwerpunkt $\underline{s}_i^{\text{ex}}$ direkt aus $\underline{F}_{\text{DH}}^{\text{db}}$ übernommen und alle anderen Merkmale zu Null gesetzt.

5.3 Positurrekonstruktion

Mit Hilfe der bisher beschriebenen Algorithmen bestimmt das Tutorsystem eine Menge von Positirkandidaten $\hat{\underline{P}}(t)$ für jedes Einzelbild der Gebärde, d. h. jeden Merkmalsvektor $\underline{F}_{\text{DH}}(t)$, indem es die Posituredatenbank nach potenziell ähnlichen Merkmalen durchsucht und aus der resultierenden Menge $\tilde{\underline{P}}(t)$ die k Posituren mit niedrigster Merkmalsdifferenz $\Delta_{\underline{F}}$ ermittelt.

Damit ergibt sich – analog zur Zuordnung von Regionen zu Markern (vgl. Abschnitt 4.5) – ein Hypothesenraum, der alle auf Grundlage der einzelnen Kandidatenmengen $\hat{\underline{P}}(t)$ möglichen Positurreihenfolgen umfasst (s. Abbildung 4.15 auf Seite 104). Zur Bestimmung der wahrscheinlichsten Positurreihenfolge dient erneut der in Abschnitt 4.5.2 beschriebene Viterbi-Algorithmus. Auf diese Weise werden Mehrdeutigkeiten in $\underline{F}_{\text{DH}}$ aufgelöst, die möglicherweise zu unpassenden Posituren in der Kandidatenmengen $\hat{\underline{P}}$ geführt haben.

5.3.1 Mehrdeutigkeitsauflösung

Im Folgenden sind die Bewertungskriterien erläutert, aus denen sich die vom Viterbi-Algorithmus benötigten statischen und dynamischen Hypothesenwahrscheinlichkeiten P_{stat} bzw. P_{dyn} im Fall der Positurrekonstruktion ergeben. Hierbei bezeichnen $\underline{\mathbf{F}}_{\text{DH},i}^{\text{db}}(t)$ und $\underline{\mathbf{P}}_i^{\text{db}}(t)$ die Merkmale bzw. die Positurparameter des i -ten Elements der Kandidatenmenge $\hat{\mathbf{P}}(t)$, und $\underline{\mathbf{F}}_{\text{DH}}^{\text{ex}}(t)$ steht für die aus dem Eingangsbild extrahierten Merkmale.

In die Hypothesenbewertung fließen die Merkmals- sowie die Posituredifferenz der Kandidaten zu den extrahierten Merkmalen bzw. zur Vorgängerpositur ein. In beiden Fällen führen niedrige Werte zu hohen Wahrscheinlichkeiten, so dass die Gewinnersequenz Posituren enthält, die sowohl die beobachteten Merkmale erklären, als auch eine plausible, d. h. flüssige Bewegung ergeben. Die Anwendung von Domänenwissen hinsichtlich der Gebärdenproduktion wäre bei geübten Gebärdensprechern sinnvoll, ist jedoch für das Tutorsystem im Hinblick auf die Erkennung von Ausführungsfehlern nicht angemessen.

5.3.1.1 Statische Positurwahrscheinlichkeit

Die Wahrscheinlichkeit P_{stat} , die eine Hypothese i unabhängig von Vorgänger und Nachfolger bewertet, wird aus der Merkmalsdifferenz berechnet:

$$P_{\text{stat}} = \gamma_{\text{stat}} \exp(-\beta \Delta_{\mathbf{F}}(\underline{\mathbf{F}}_{\text{DH},i}^{\text{db}}(t), \underline{\mathbf{F}}_{\text{DH}}^{\text{ex}}(t))) \quad (5.11)$$

Dabei ist γ_{stat} so zu wählen, dass sich die Wahrscheinlichkeiten zu Eins ergänzen.³ Der konstante Gewichtungsfaktor $\beta \in [0, 1]$ ist im nächsten Abschnitt erläutert.

Die Verwendung der Exponentialfunktion ermöglicht, die Maximierung der Pfadwahrscheinlichkeit in (4.61) durch die Maximierung der Exponentensumme zu ersetzen, deren Berechnung schneller und numerisch stabiler ist. Die Wahl der Basis e ist dabei willkürlich und folgt der Konvention in der Sprachverarbeitung [Jel97].

5.3.1.2 Dynamische Positurwahrscheinlichkeit

Die Bewertung eines Übergangs zwischen zwei Posituren $\underline{\mathbf{P}}_i^{\text{db}}(t-1) \in \hat{\mathbf{P}}(t-1)$ und $\underline{\mathbf{P}}_j^{\text{db}}(t) \in \hat{\mathbf{P}}(t)$ basiert auf der Differenz dieser Posituren. Eine Berechnung analog zu (5.11) würde allerdings Positurenfolgen begünstigen, in denen die Handform überwiegend statisch bleibt. Dieser Effekt tritt verstärkt bei besonders unspezifischen Merkmalsvektoren auf, die beispielsweise bei wenigen sichtbaren Markierungen entstehen, weil hier die Aussagekraft der Merkmalsdifferenz vergleichsweise gering ist.

³Wie in Gleichung (4.66) hat die Normierungskonstante γ_{stat} auf die Bestimmung des Gewinnerpfads keinen Einfluss und wird in der Implementierung daher vereinfachend zu Eins angenommen.

Daher wird bei der Abbildung von $\Delta_{\underline{\mathbf{P}}}$ auf P_{dyn} eine empirisch bestimmte plausible Differenz $\Delta_{\underline{\mathbf{P}}}^*$ berücksichtigt, deren Unterschreiten $\Delta_{\underline{\mathbf{P}}}$ nicht weiter reduziert:

$$P_{\text{dyn}} = \gamma_{\text{dyn}} \exp\left(-(1 - \beta) \max\left\{\Delta_{\underline{\mathbf{P}}}(\underline{\mathbf{P}}_i^{\text{db}}(t-1), \underline{\mathbf{P}}_j^{\text{db}}(t)) - \Delta_{\underline{\mathbf{P}}}^*, 0\right\}\right) \quad (5.12)$$

Der Faktor $\beta \in [0, 1]$ gewichtet P_{stat} gegenüber P_{dyn} (vgl. (5.11)). Er kompensiert in erster Linie die unterschiedlichen Größenordnungen der Differenzmaße $\Delta_{\underline{\mathbf{F}}}$ und $\Delta_{\underline{\mathbf{P}}}$. Darüber hinaus ist es sinnvoll, β bei zuverlässigen Merkmalen $\underline{\mathbf{F}}_{\text{DH}}$ größer zu wählen, um den entsprechend höheren Informationsgehalt der statischen Positurwahrscheinlichkeit zu berücksichtigen. Sind die Merkmale hingegen stark verrauscht, kann der Fokus der Hypothesenbewertung durch Reduzieren von β auf die Flüssigkeit der Gebärdenbewegung (d. h. auf das höhere Wissen) gelegt werden.

5.3.2 Positurenglättung

Das Ergebnis der Mehrdeutigkeitsauflösung ist eine Sequenz von Posituren $\underline{\mathbf{P}}(t) \in P^{\text{db}}$, die idealerweise maximale Ähnlichkeit zu den Eingangsposituren aufweisen. Durch die Endlichkeit der Positurenmenge P^{db} existieren jedoch im Allgemeinen für jede Eingangspositur mehrere annähernd gleich ähnliche Datenbankposituren. Da die extrahierten Merkmale rauschbehaftet sind, hat die Auswahl unter diesen geeigneten Posituren teilweise Zufallscharakter.

Dieser Effekt kann wiederum als ein den Positurparametern überlagerter Rauschprozess aufgefasst werden, der sich in Form eines „Zitterns“ bzw. „Springens“ zwischen zeitlich benachbarten Posituren äußert. Er lässt sich durch Glättung des Verlaufs der Ansichtswinkel $\underline{\Phi}(t)$ und Beugungsgrade $\underline{\mathbf{B}}(t)$ reduzieren; das Erkennungsergebnis ist dadurch sowohl visuell ansprechender als auch genauer im Sinne minimaler Differenz zur Eingangspositur. Ferner können einzelne Erkennungsfehler auf diese Weise durch Ausnutzen der bei einer Bildrate von 25 Hz gegebenen Kontinuität der Gebärdenbewegung kompensiert werden.

Für Ansichtswinkel und Beugungsgrade kommen unterschiedliche Strategien zum Einsatz, die in den nächsten beiden Abschnitten erläutert sind.

5.3.2.1 Handstellung

Bei der Glättung der Ansichtswinkel $\underline{\Phi}$ ist der gimbal lock zu beachten, der bei unabhängiger Verarbeitung von ϕ_{lat} , ϕ_{lon} und ϕ_{rot} in Polnähe zu großen Abweichungen zwischen aufeinander folgenden Ansichten führen würde. Aus diesem Grund wird die in Abschnitt 3.3.4 definierte Handstellungsdifferenz $\Delta_{\underline{\Phi}}$ eingesetzt, die vom gimbal lock unbeeinflusst ist.

Die Glättung der Handstellungsfolge $\underline{\Phi}(t)$ geschieht durch Anpassen jeder Ansicht an ihre zeitlichen Nachbarn. Hierzu wird die gewichtete Summe der Abweichungen minimiert:

$$\hat{\underline{\Phi}}(t) = \underset{\underline{\Phi}'}{\operatorname{argmin}} \sum_{i=-n}^n \omega_i \Delta_{\underline{\Phi}}(\underline{\Phi}', \underline{\Phi}(t+i)) \quad (5.13)$$

Die Gewichte ω_i sind mit einem eindimensionalen Faltungskern der Länge $2n+1$ vergleichbar und steuern den Grad der Glättung. Geeignete Werte sind

$$n = 2 \quad (5.14)$$

$$\langle \omega_{-2}, \omega_{-1}, \omega_0, \omega_1, \omega_2 \rangle = \langle 1, 2, 4, 2, 1 \rangle \quad (5.15)$$

In Abbildung 5.12 ist der Verlauf der Ansichtswinkel ϕ_{lat} , ϕ_{lon} und ϕ_{rot} vor und nach der Glättung visualisiert.

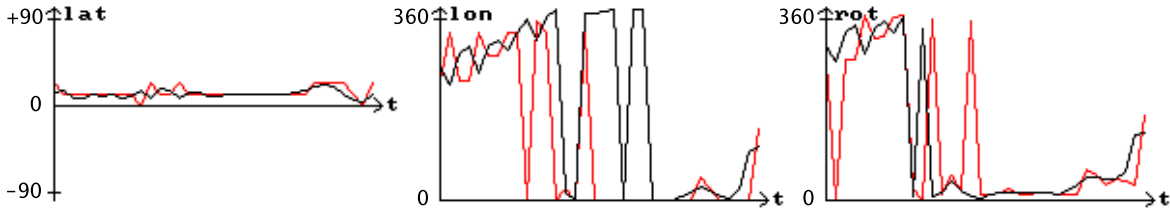


Abbildung 5.12: Glättung der Ansichtswinkel ϕ_{lat} , ϕ_{lon} und ϕ_{rot} . Die geglätteten Kurven sind schwarz dargestellt. Bedingt durch den gimbal lock verlaufen ϕ_{lon} und ϕ_{rot} auch nach der Verarbeitung an mehreren Stellen unstetig. Dieser Effekt ist rein numerisch, d. h. die entsprechenden Ansichten sind trotz stark verschiedener Winkel sehr ähnlich.

5.3.2.2 Handform

Die Glättung des Handformverlaufs $\underline{\mathbf{B}}(t)$ besteht in der Glättung der einzelnen skalaren Beugungsgrade. Anatomisch unmögliche Handformen sind dabei durch die in Abschnitt 3.2.3.1 beschriebene, von der Beugung der übrigen Fingergelenke abhängige Interpretation implizit ausgeschlossen, so dass lediglich die Beschränkung auf den Wertebereich $[0, 1]$ sicher zu stellen ist.

Mit der Merkmalsdifferenz $\Delta_{\underline{\mathbf{F}}}$ liegt für jeden Messwert (d. h. jede Handform $\underline{\mathbf{B}}(t)$) ein Gütemaß vor. Die Glättung geschieht durch Auswahl $n+1$ zuverlässiger Stützstellen und kubische Spline-Interpolation in den resultierenden n Intervallen [BS91]. Zunächst werden die Handformen $\underline{\mathbf{B}}(t_0 = 0)$ (Beginn der Vorbereitungsphase) und $\underline{\mathbf{B}}(t_n = T - 1)$ (Ende der Rückführung) unabhängig von $\Delta_{\underline{\mathbf{F}}}$ als Stützstellen definiert. Anschließend erfolgt iterativ die Selektion weiterer Zeitpunkte t_i aus einem Intervall $[t_{i-1} + a, t_{i-1} + b]$, die sich durch möglichst niedrige Merkmalsdifferenz auszeichnen:

$$t_i = \underset{a \leq t - t_{i-1} \leq b}{\operatorname{argmin}} \Delta_{\underline{\mathbf{F}}}(\underline{\mathbf{F}}_{\text{DH}}^{\text{db}}(t), \underline{\mathbf{F}}_{\text{DH}}^{\text{ex}}(t)) \quad (5.16)$$

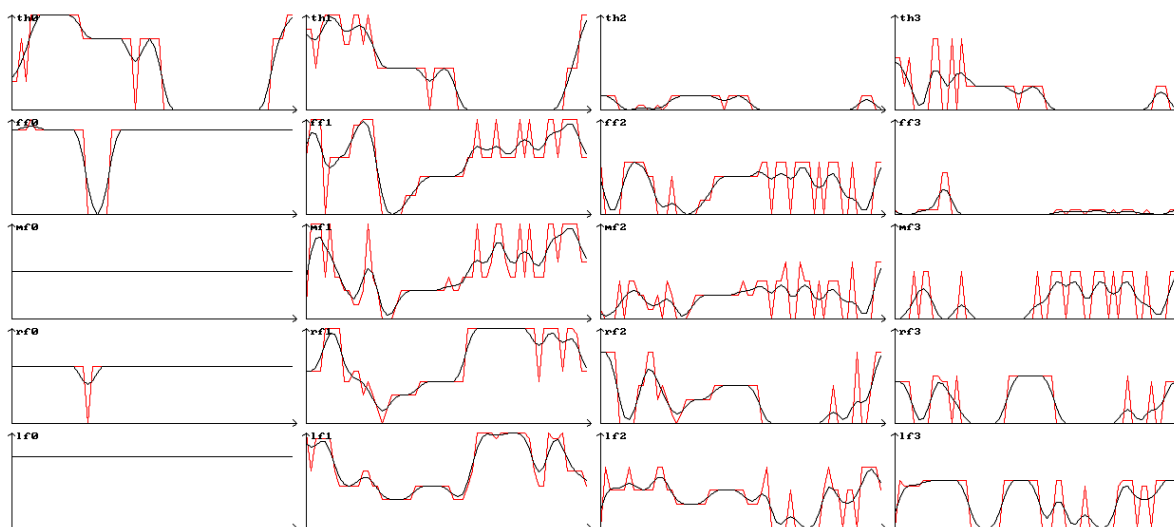


Abbildung 5.13: Glättung der 20 Beugungsgrade $\underline{B}(t)$ durch Interpolation mit kubischen Splines. Die Intervallgrenzen zur Bestimmung der Stützstellen wurden zu $a = 1$ und $b = 4$ gewählt. Die fünf Zeilen entsprechen Daumen, Zeigefinger, Mittelfinger, Ringfinger und kleinem Finger (von oben nach unten). In der ersten Spalte ist die Abduktion/Adduktion im CMC-Gelenk (Daumen) bzw. MCP-Gelenk (Finger) dargestellt, in der zweiten bis vierten Spalte die Flexion/Extension im CMC-/MCP-/IP-Gelenk (Daumen) bzw. im MCP-/PIP-/DIP-Gelenk (Finger). Auf der Abszisse ist die Zeit $t \in \{0, \dots, 59\}$ aufgetragen, die Ordinate stellt den Wertebereich $[0, 1]$ dar.

Bei einer Bildrate von 25 Hz sind Intervallgrenzen von $a = 1$ und $b = 4$ sinnvoll. Nach Festlegung der Stützstellen ergeben sich die $4n$ Parameter der n kubischen Spline-Polynome aus den folgenden Randbedingungen:

1. Der Spline soll durch die $n + 1$ Stützstellen verlaufen ($2n$ Bedingungen).
2. Steigung und Krümmung des Splines an den $n - 1$ Übergängen soll stetig sein sein ($2n - 2$ Bedingungen).
3. Die Krümmung des Splines an den Stützstellen t_0 und t_{n-1} soll gleich Null sein (2 Bedingungen). Diese Form wird als *natürliche Splines* bezeichnet.

Abbildung 5.13 zeigt die Glättung der Beugungsgrade anhand eines Beispiels.

Kapitel 6

Feedbackgenerierung

Zur Bewertung der Anwendergebärde hinsichtlich ihrer Korrektheit erfolgt ein mehrstufiger Vergleich mit der Referenzausführung, der die einzelnen Ausdrucksmittel der Gebärdensprache berücksichtigt. Hierfür ist neben den beiden Positurenfolgen auch die Kenntnis der einzelnen Phasen der Gebärde erforderlich, denn nicht alle Ausdrucksmittel sind zu jedem Zeitpunkt relevant. Die entsprechenden Metainformationen sind in Abschnitt 6.1 aufgeführt.

Da die Ausführungsgeschwindigkeit der Anwender- und der Referenzgebärde im Allgemeinen leicht unterschiedlich ist, erfolgt vor dem eigentlichen Gebärdenvergleich auf Positurebene eine zeitliche Ausrichtung der beiden Positurensequenzen. Dabei wird jedes Einzelbild einer Sequenz jeweils einem oder mehreren Einzelbildern der anderen Sequenz zugeordnet und somit die beim Vergleich der manuellen Ausdrucksmittel zu Grunde liegenden Positурparameter festgelegt. Dieser Prozess berechnet auch die Geschwindigkeit der Anwendergebärde relativ zur Referenz, die ebenfalls ein Vergleichskriterium darstellt. Abschnitt 6.2 beschreibt die eingesetzten Algorithmen und erläutert, in welcher Weise die verschiedenen manuellen Ausdrucksmittel in den Vergleich eingehen.

Abschließend sind in Abschnitt 6.3 mehrere Beispiele korrektiven Feedbacks gezeigt, die alle möglichen Fehlerarten abdecken.

Im Folgenden bezeichnen $\underline{\Phi}(t)$ und $\underline{B}(t)$ die gemäß Abschnitt 5.3 geglätteten Handstellungen bzw. -formen. Die Superskripte A und R kennzeichnen die Anwender- bzw. die Referenzgebärde, sofern diese zu unterscheiden sind. Für eine nicht sichtbare Hand wird eine Ruheposition unmittelbar unterhalb der unteren Bildkante angenommen, deren x -Koordinate in der gemäß Abschnitt 4.6.2 normierten Darstellung $-\frac{3}{2}$ für die rechte und $+\frac{3}{2}$ für die linke Hand beträgt (dies entspricht etwa der jeweiligen Schulter). Form und Stellung der nicht sichtbaren Hand sind undefiniert.

6.1 Darstellung der Referenzgebärde

Nach Abschluss der Positurrekonstruktion liegen die folgenden Gebärdendaten vor:

- Position $\underline{s}_{\text{DH}}(t)$ der dominanten Hand
- Stellung $\underline{\Phi}(t)$ der dominanten Hand
- Form $\underline{B}(t)$ der dominanten Hand
- Position $\underline{s}_{\text{NDH}}(t)$ der nichtdominanten Hand (bei zweihändigen Gebärden)

Nicht alle dieser Komponenten sind in jeder Gebärdenphase von Bedeutung. Beispielsweise transportiert die Handform während der Vorbereitung und der Rückführung keine Informationen. Bei der Mehrzahl der Gebärden sind in der Ausführungsphase Form, Stellung und Position wichtig, bei Zahlengebärden jedoch lediglich Form und Stellung.

Um einen aussagekräftigen Vergleich mit der Anwendergebärde durchzuführen, sind daher zusätzlich Metainformationen hinsichtlich der Relevanz der jeweiligen Komponenten erforderlich. Diese werden für die Referenzgebärde einmalig manuell eingegeben und sind durch die in Abschnitt 6.2.1 erläuterte dynamische Zeitanpassung auf die Anwendergebärde übertragbar. Die Darstellung erfolgt in Form dreier binärer Funktionen (Flags), wobei Form und Stellung zusammengefasst sind:

$$b_{\underline{s},\text{DH}}^{\text{R}}(t^{\text{R}}) = \begin{cases} 1 & \text{Handposition } \underline{s}_{\text{DH}}^{\text{R}}(t^{\text{R}}) \text{ ist relevant} \\ 0 & \text{Handposition } \underline{s}_{\text{DH}}^{\text{R}}(t^{\text{R}}) \text{ ist nicht relevant} \end{cases} \quad (6.1)$$

$$b_{\underline{P},\text{DH}}^{\text{R}}(t^{\text{R}}) = \begin{cases} 1 & \text{Handstellung } \underline{\Phi}^{\text{R}}(t^{\text{R}}) \text{ und -form } \underline{B}^{\text{R}}(t^{\text{R}}) \text{ sind relevant} \\ 0 & \text{Handstellung } \underline{\Phi}^{\text{R}}(t^{\text{R}}) \text{ und -form } \underline{B}^{\text{R}}(t^{\text{R}}) \text{ sind nicht relevant} \end{cases} \quad (6.2)$$

$$b_{\underline{s},\text{NDH}}^{\text{R}}(t^{\text{R}}) = \begin{cases} 1 & \text{Handposition } \underline{s}_{\text{NDH}}^{\text{R}}(t^{\text{R}}) \text{ ist relevant} \\ 0 & \text{Handposition } \underline{s}_{\text{NDH}}^{\text{R}}(t^{\text{R}}) \text{ ist nicht relevant} \end{cases} \quad (6.3)$$

6.2 Gebärdenvergleich

Der Vergleich zweier Gebärden setzt zunächst eine zeitliche Anpassung voraus, um unvermeidbare Variationen in der Ausführungsgeschwindigkeit zu kompensieren. Diese treten beim Gebärden regelmäßig auf und stellen keinen Fehler dar, solange sie sich in üblichen Grenzen bewegen. Anschließend geschieht der eigentliche Vergleich in mehreren Stufen und unter Berücksichtigung der Flags $b_{\underline{s},\text{DH}}^{\text{R}}$, $b_{\underline{P},\text{DH}}^{\text{R}}$ und $b_{\underline{s},\text{NDH}}^{\text{R}}$. In den beiden folgenden Abschnitten werden diese zwei Verarbeitungsschritte erläutert.

6.2.1 Dynamische Zeitanpassung

Die zeitliche Anpassung der beiden zu vergleichenden Gebärden besteht in einer Zuordnung der einander entsprechenden Zeitindizes t^{A} und t^{R} , die sich als Folge von Paaren $\langle t_i^{\text{A}}, t_i^{\text{R}} \rangle$ darstellen lässt. Anfang und Ende bilden dabei $\langle 0, 0 \rangle$ bzw. $\langle T^{\text{A}} - 1, T^{\text{R}} - 1 \rangle$. Eventuell zu Beginn oder am Schluss der Gebärden vorhandene Einzelbilder, in denen beide Hände nicht

sichtbar sind, haben für den Vergleich keine Bedeutung und werden vor der Verarbeitung entfernt.

Aufgrund des in [PKB90] empirisch nachgewiesenen hohen Informationsgehalts der Handposition¹ bildet diese die Basis für die zeitliche Ausrichtung. Die Berechnung der Korrespondenzen geschieht mittels dynamischer Zeitanpassung (Dynamic Time Warping, DTW). Hauptkomponente ist hierbei ein Abstandsmaß $\Delta_{\underline{s}}$, das die Verschiedenheit der Handpositionen quantifiziert:

$$\Delta_{\underline{s}}(t^A, t^R) = \delta_{\text{DH}}(t^A, t^R) + \delta_{\text{NDH}}(t^A, t^R) \quad (6.4)$$

Darin bezeichnen δ_{DH} und δ_{NDH} die Anteile der dominanten bzw. der nichtdominanten Hand:

$$\delta_{\text{DH}}(t^A, t^R) = \begin{cases} \|\underline{s}_{\text{DH}}^A(t^A) - \underline{s}_{\text{DH}}^R(t^R)\|^2 & \text{für } b_{\underline{s}, \text{DH}}^R(t^R) = 1 \\ \epsilon & \text{für } b_{\underline{s}, \text{DH}}^R(t^R) = 0 \end{cases} \quad (6.5)$$

$$\delta_{\text{NDH}}(t^A, t^R) = \begin{cases} \|\underline{s}_{\text{NDH}}^A(t^A) - \underline{s}_{\text{NDH}}^R(t^R)\|^2 & \text{für } b_{\underline{s}, \text{NDH}}^R(t^R) = 1 \\ \epsilon & \text{für } b_{\underline{s}, \text{NDH}}^R(t^R) = 0 \end{cases} \quad (6.6)$$

Die Konstante ϵ steht für einen möglichst kleinen positiven Wert und wird auf Seite 138 erläutert. Für einhändige Gebärden ist $b_{\underline{s}, \text{NDH}}^R(t^R) = 0 \forall t^R$ und (6.6) somit ohne Einfluss.

Aus den Werten $\Delta_{\underline{s}}(t^A, t^R)$ lässt sich eine Matrix $\mathbf{D}$ konstruieren, in der jedes Element $d_{i,j}$ eine Zuordnung der Zeitindizes $j - 1$ der Anwendergebärde und $i - 1$ der Referenzgebärde repräsentiert und den Wert des entsprechenden Abstands $\Delta_{\underline{s}}(j - 1, i - 1)$ hat:

$$\mathbf{D} = [d_{ij}]_{T^R \times T^A} \quad \text{mit} \quad d_{ij} = \Delta_{\underline{s}}(j - 1, i - 1) \quad (6.7)$$

In Abbildung 6.1 ist diese Matrix vertikal gespiegelt dargestellt, so dass das Element $d_{1,1}$ auf den Ursprung fällt. Die gesuchte Ausrichtung der beiden Gebärden entspricht einem Pfad von $\langle 0, 0 \rangle$ nach $\langle T^A - 1, T^R - 1 \rangle$.

Um Rückwärtsbewegungen in der Zeit zu vermeiden, muss dieser Pfad monoton steigen, d. h. $t_i^A \leq t_j^A \wedge t_i^R \leq t_j^R$ für $i < j$. Da weder in der Anwender- noch in der Referenzgebärde Zeitindizes übersprungen werden sollen, ergeben sich drei mögliche Schritte, die in Abbildung 6.2 gezeigt sind. Jedem Schritt ist ein Gewicht ω zugeordnet, mit welchem der dem Zielpunkt zugeordnete Abstand δ multipliziert wird. Diese Gewichte sind so gewählt, dass sie eine möglichst geringe Verzerrung der Zuordnung, d. h. den Fall $t_{i+1}^A = t_i^A + 1 \wedge t_{i+1}^R = t_i^R + 1$, fördern:

¹In einer abgedunkelten Versuchsanordnung konnten gehörlose Probanden zahlreiche Gebärden einzig über die Handposition identifizieren, die durch an den Handgelenken befestigte Lämpchen erkennbar war.

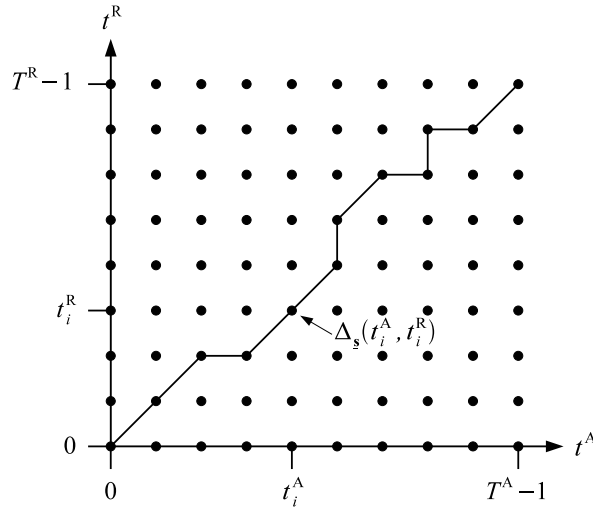


Abbildung 6.1: Matrix aller möglichen 2-Tupel $\langle t^A, t^R \rangle$. Jedem Punkt ist die entsprechende Distanz δ zugeordnet. Exemplarisch ist ein beliebiger Pfad eingezeichnet, der eine Folge von Paaren $\langle t_i^A, t_i^R \rangle$ darstellt.

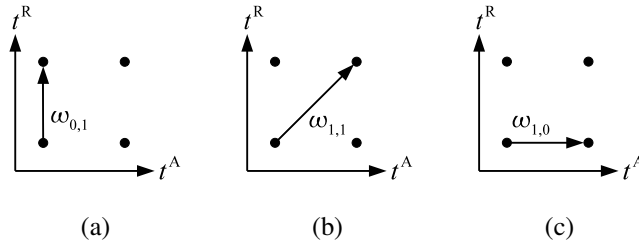


Abbildung 6.2: Mögliche Schritte auf dem Pfad von $\langle 0, 0 \rangle$ nach $\langle T^A - 1, T^R - 1 \rangle$. Wenn die Ausführungsgeschwindigkeit der Anwendergebärde höher ist als die der Referenzgebärde, ergibt sich (a), d. h. zwei Zeitindizes der Referenzgebärde werden einem Zeitindex der Anwendergebärde zugeordnet. Der umgekehrte Fall entspricht (c). Bei gleicher Geschwindigkeit resultiert der Normalfall (b). Jedem Schritt ist ein Gewicht ω zugeordnet.

$$\omega_{1,0} = 2 \quad (6.8)$$

$$\omega_{1,1} = 1 \quad (6.9)$$

$$\omega_{0,1} = 2 \quad (6.10)$$

An dieser Stelle wird die Bedeutung der Konstanten ϵ in (6.5) und (6.6) deutlich. Mit $\epsilon > 0$ ist gewährleistet, dass der durch das niedrigste Gewicht $\omega_{1,1}$ präferierte Diagonalschritt auch in Phasen bevorzugt wird, in denen die Handposition nicht bedeutungsrelevant ist und daher $b_{\mathbf{s},\text{DH}}^R = b_{\mathbf{s},\text{NDH}}^R = 0$ gilt. In der praktischen Implementierung ist ϵ gerade so groß zu wählen, dass numerische Effekte wie etwa Rundungsfehler ohne Einfluss bleiben.

Ein Pfad Γ , der aus L Schritten besteht und somit $L + 1$ Matrixpunkte verknüpft, wird als Folge von $L + 1$ Paaren $\Gamma_i = \langle t_i^A, t_i^R \rangle$ geschrieben:

$$\Gamma = \langle \Gamma_0, \dots, \Gamma_i, \dots, \Gamma_L \rangle \quad (6.11)$$

$$\text{mit } \Gamma_0 = \langle 0, 0 \rangle \quad (6.12)$$

$$\Gamma_i = \langle t_i^A, t_i^R \rangle \quad (6.13)$$

$$\Gamma_L = \langle T^A - 1, T^R - 1 \rangle \quad (6.14)$$

Für jeden Pfad Γ lässt sich eine Gesamtdistanz $\Delta_\Gamma(\Gamma)$ berechnen, die die Plausibilität der entsprechenden Zuordnung der Zeitindizes quantifiziert:

$$\Delta_\Gamma(\Gamma) = \frac{1}{T^A T^R} \sum_{i=0}^L \omega_i \Delta_{\underline{s}}(t_i^A, t_i^R) \quad (6.15)$$

Dabei ist ω_i das Gewicht des Schrittes von Γ_{i-1} nach Γ_i gemäß (6.8) bis (6.10). Für den ersten Punkt des Pfades gilt $\omega_0 = 1$. Die Normierung auf $T^A T^R$ beeinflusst die Anpassung nicht, erleichtert aber den Vergleich der numerischen Werte verschiedener Gebärdenpaare.

Die gesuchte zeitliche Anpassung der Gebärdenverläufe entspricht dem hinsichtlich (6.15) optimalen Pfad:

$$\hat{\Gamma} = \underset{\Gamma}{\operatorname{argmin}} \Delta_\Gamma(\Gamma) \quad (6.16)$$

Zur Bestimmung dieses Pfades wird erneut der Viterbi-Algorithmus (vgl. Abschnitt 4.5.2) angewandt. Dabei entspricht die Minimierung der Pfaddistanz $\Delta_\Gamma(\Gamma)$ der Maximierung der Pfadwahrscheinlichkeit $P(W)$ nach (4.60). Anders als bisher sind jedoch Start- und Endzustand vorgegeben und Übergänge im Suchraum nur gemäß der in Abbildung 6.2 gezeigten Schritte erlaubt. Abbildung 6.3 veranschaulicht die dynamische Zeitanpassung bei korrekter sowie bei zu schnell ausgeführter Anwendergebärde.

6.2.2 Vergleichskriterien

Nach der dynamischen Zeitanpassung kann für jedes Pfadelement $\hat{\Gamma}_i$ ein direkter Vergleich der einander zugeordneten Handposituren erfolgen. Darüber hinaus gibt die Steigung des Pfades $\hat{\Gamma}$ in der Umgebung des i -ten Punktes an, ob der Anwender die Gebärde in dieser Phase mit der richtigen, d. h. der Referenzgebärde entsprechenden Geschwindigkeit ausgeführt hat.

Im Einzelnen berücksichtigt der Positurenvergleich die folgenden Kriterien:

1. Abweichungen der Handposition δ_{DH} und δ_{NDH}
2. Abweichung der Bewegungsgeschwindigkeit Δ_v
3. Abweichung der Handstellung $\Delta_{\underline{\Phi}}$
4. Abweichung der Handform $\hat{\Delta}_{\underline{B}}$

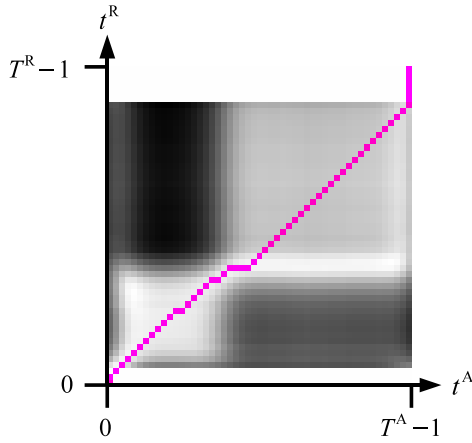
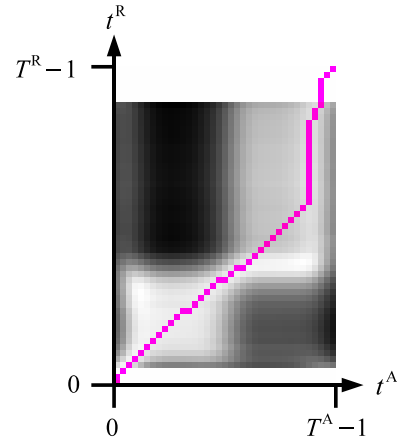
(a) $\Delta_\Gamma(\Gamma) = 0,12624$, $T^A = 51$, $T^R = 53$ (b) $\Delta_\Gamma(\Gamma) = 0,14705$, $T^A = 37$, $T^R = 53$

Abbildung 6.3: Beispiel der dynamischen Zeitanpassung durch Berechnung eines Pfades durch die DTW-Matrix $\mathbf{D}$. Dunkle Werte entsprechen hohen Einzeldifferenzen δ (logarithmische Skalierung, normiert auf $\max \delta$). Die weißen Bereiche am oberen und unteren Rand entstehen durch $b_{\mathbf{s},\text{DH}}^R = b_{\mathbf{s},\text{NDH}}^R = 0$ zu Beginn und am Ende der Referenzgebärde. Bei korrekter Ausführung ergibt sich ein überwiegend diagonalen Pfad (a). Stimmt die Anwendergebärde hinsichtlich der Ausführungsgeschwindigkeit nicht mit der Referenzgebärde überein, weicht der Pfad von dieser Ideallinie ab. Im Beispiel (b) war die Ausführung in der zweiten Hälfte zu schnell.

Die vier Abschnitte 6.2.2.1 bis 6.2.2.4 erläutern diese Vergleichskriterien und führen das Abstandsmaß Δ_v ein. Abbildung 6.4 zeigt ein Ablaufdiagramm des gesamten Vergleichs.

6.2.2.1 Handposition

Die Verifikation der Handposition verwendet die bereits vorliegenden Entfernung der Hand-schwerpunkte gemäß (6.5) und (6.6). Das Kriterium ist erfüllt, wenn die Position jeder Hand nicht wesentlich von der Referenzgebärde abweicht oder in der aktuellen Gebärdenphase nicht relevant ist:

$$\delta_{\text{DH}}(t_i^A, t_i^R) \leq \theta_{\mathbf{s},\text{DH}}^2 \quad \vee \quad b_{\mathbf{s},\text{DH}}^R(t_i^R) = 0 \quad (6.17)$$

$$\delta_{\text{NDH}}(t_i^A, t_i^R) \leq \theta_{\mathbf{s},\text{NDH}}^2 \quad \vee \quad b_{\mathbf{s},\text{NDH}}^R(t_i^R) = 0 \quad (6.18)$$

Dabei geben die Schwellwerte $\theta_{\mathbf{s},\text{DH}}$ und $\theta_{\mathbf{s},\text{NDH}}$ den maximalen normierten euklidischen Abstand der Schwerpunkte in der zweidimensionalen Bildebene an.

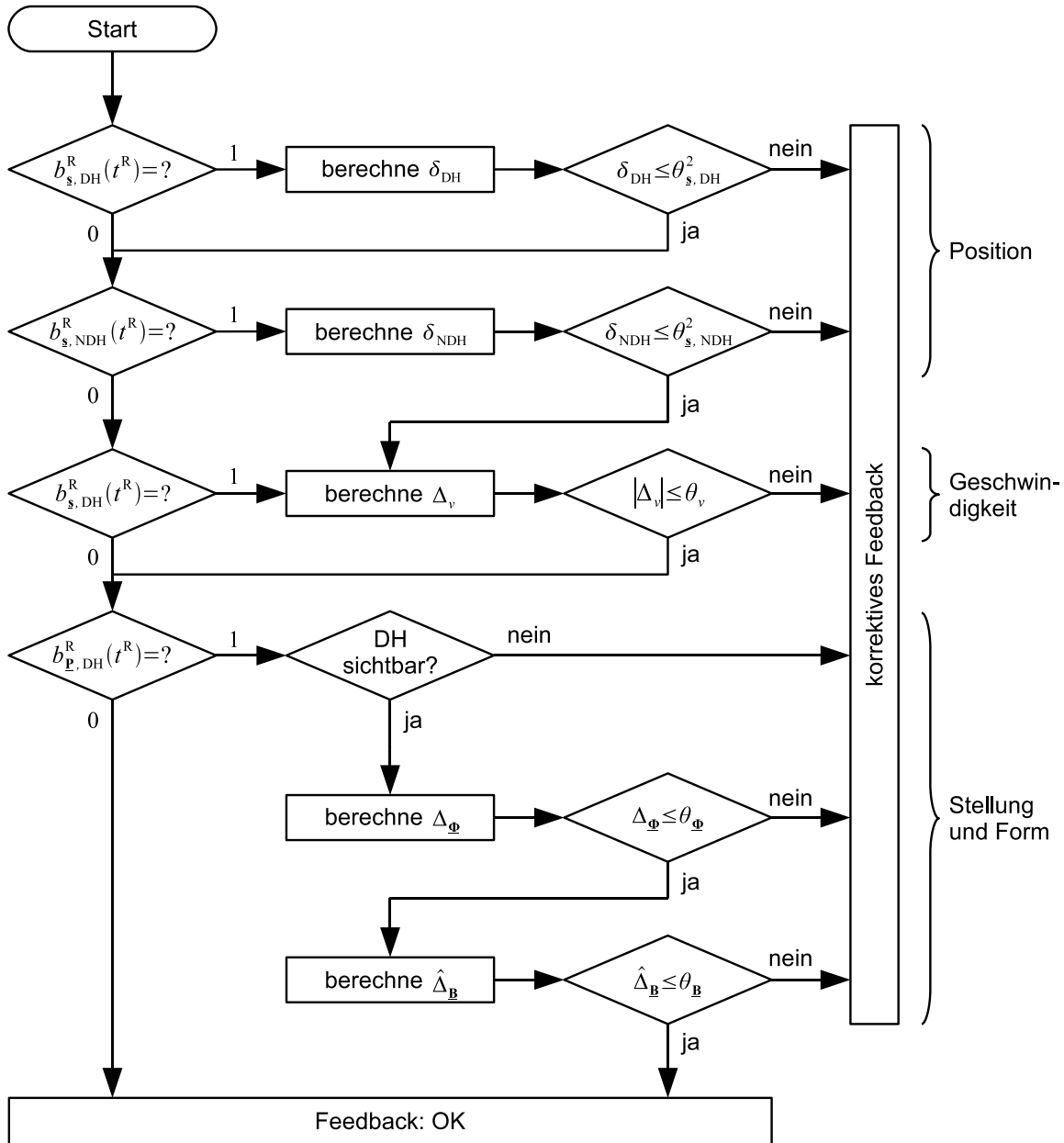


Abbildung 6.4: In Abhängigkeit von den Flags $b_{s,DH}^R$, $b_{s,NDH}^R$ und $b_{p,DH}^R$ erfolgt die Prüfung der Kriterien in der genannten Reihenfolge, indem jeweils mit einem Schwellwert θ verglichen wird, der je nach Gebärde unterschiedlich sein kann. Überschreitet die Gebärde keinen der Schwellwerte, bestätigt das System die korrekte Ausführung. Andernfalls erhält der Anwender entsprechendes korrekatives Feedback wie in Abschnitt 6.3 beschrieben.

6.2.2.2 Bewegungsgeschwindigkeit

Die Bewegungsgeschwindigkeit $v(i)$ entspricht der lokalen Steigung des Pfades $\hat{\Gamma}$ und berücksichtigt damit sowohl die Position der dominanten als auch – bei zweihändigen Gebärden – der nichtdominanten Hand. Sie wird als Steigungswinkel angegeben und berechnet sich zu

$$v(i) = \begin{cases} \arctan\left(\frac{\Delta t^A}{\Delta t^R}\right) & \text{für } \Delta t^R \neq 0 \\ 90^\circ & \text{sonst} \end{cases} \quad (6.19)$$

mit

$$\Delta t^A = t_{i_1}^A - t_{i_0}^A \quad (6.20)$$

$$\Delta t^R = t_{i_1}^R - t_{i_0}^R \quad (6.21)$$

Die Indizes i_0 und i_1 beschreiben den Pfadabschnitt, über dem die Geschwindigkeit gemittelt wird. Sie haben vom betrachteten Punkt i den Abstand Δi , sofern sie nicht durch die Pfadlänge beschränkt werden:

$$i_0 = \max\{i - \Delta i, 0\} \quad (6.22)$$

$$i_1 = \min\{i + \Delta i, L\} \quad (6.23)$$

$$(6.24)$$

Bei der verwendeten Bildrate von 25 Hz ist $\Delta i = 3$ sinnvoll, so dass $\max\{\Delta t^A, \Delta t^R\}$ etwa einer Viertelsekunde entspricht. Die Sollgeschwindigkeit ergibt sich bei gleichmäßigem Fortschreiten in t^A und t^R und entspricht somit einem Steigungswinkel von 45° . Die Definition der Abweichung Δ_v ist daher

$$\Delta_v(i) = v(i) - 45^\circ \quad (6.25)$$

Die Ausführungsgeschwindigkeit der Anwendergebärde gilt als korrekt, wenn die Abweichung von der Sollgeschwindigkeit unterhalb eines Schwellwerts θ_v liegt oder beide Handpositionen zum Zeitpunkt i ohne Bedeutung sind:

$$|\Delta_v(i)| \leq \theta_v \vee (b_{\underline{s}, \text{DH}}^R(t_i^R) = 0 \wedge b_{\overline{s}, \text{DH}}^R(t_i^R) = 0) \quad (6.26)$$

Ist (6.26) nicht erfüllt, gibt das Vorzeichen der Differenz Δ_v an, ob der Anwender zu schnell ($\Delta_v > 0$) oder zu langsam ($\Delta_v < 0$) gebärdet hat. Durch die Angabe als Winkel ist Δ_v in beiden Fällen betragsmäßig gleich groß und damit besser zu interpretieren.

6.2.2.3 Handstellung

Sind die ersten beiden Kriterien erfüllt und darüber hinaus die Positur der dominanten Hand durch $b_{\underline{\text{P}}, \text{DH}}^R(t_i^R) = 1$ als relevant gekennzeichnet, erfolgt zunächst der Vergleich der Handstellungen. Hierfür muss die dominante Hand in der Anwendergebärde zum betrachteten Zeitpunkt i sichtbar sein – andernfalls wird der Anwender entsprechend informiert und der Gebärdenvergleich beendet.

Der Handstellungsvergleich verwendet das in Abschnitt 3.3.4 eingeführte Differenzmaß Δ_{Φ} . Die Stellung wird als korrekt gewertet, wenn die Hand im Vergleich zur Referenz um nicht mehr als θ_{Φ} rotiert oder die Handpositur irrelevant ist:

$$\Delta_{\Phi}(\Phi^A(t_i^A), \Phi^R(t_i^R)) \leq \theta_{\Phi} \quad \vee \quad b_{\underline{\mathbf{P}}, \text{DH}}^R(t_i^R) = 0 \quad (6.27)$$

6.2.2.4 Handform

Zur Kontrolle der Handform kommt das in Abschnitt 3.3.3 definierte Differenzmaß $\hat{\Delta}_{\mathbf{B}}$ zum Einsatz, das den maximalen euklidischen Abstand der Finger- und Daumenspitzen angibt. Analog zur Handstellung gilt die Ausführung des Anwenders als korrekt, wenn $\hat{\Delta}_{\mathbf{B}}$ einen Schwellwert $\theta_{\mathbf{B}}$ nicht überschreitet oder die Handpositur für den Zeitindex i nicht von Bedeutung ist:

$$\hat{\Delta}_{\mathbf{B}}(\mathbf{B}^A(t_i^A), \mathbf{B}^R(t_i^R)) \leq \theta_{\mathbf{B}} \quad \vee \quad b_{\underline{\mathbf{P}}, \text{DH}}^R(t_i^R) = 0 \quad (6.28)$$

6.3 Präsentation des Feedbacks

Im Falle eines Ausführungsfehlers soll das Tutorsystem dem Anwender korrekatives Feedback geben, das die Abweichung von der Referenzgebärde hervorhebt und die richtige Ausführung

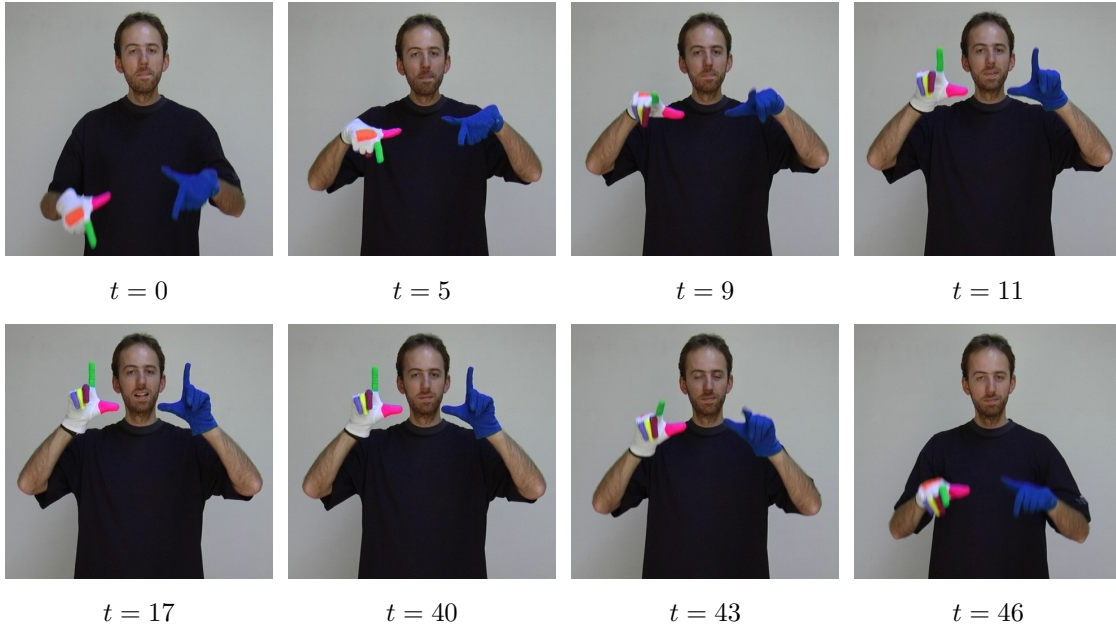


Abbildung 6.5: Die Gebärde „Bildtelefon“, aufgezzeichnet mit 25 Hz. Die Ausführungsphase entspricht $5 \leq t \leq 43$. In diesem Intervall gilt $b_{\underline{\mathbf{s}}, \text{DH}}^R(t) = b_{\underline{\mathbf{s}}, \text{NDH}}^R(t) = b_{\underline{\mathbf{P}}, \text{DH}}^R(t) = 1$, außerhalb sind alle drei Flags gleich Null.

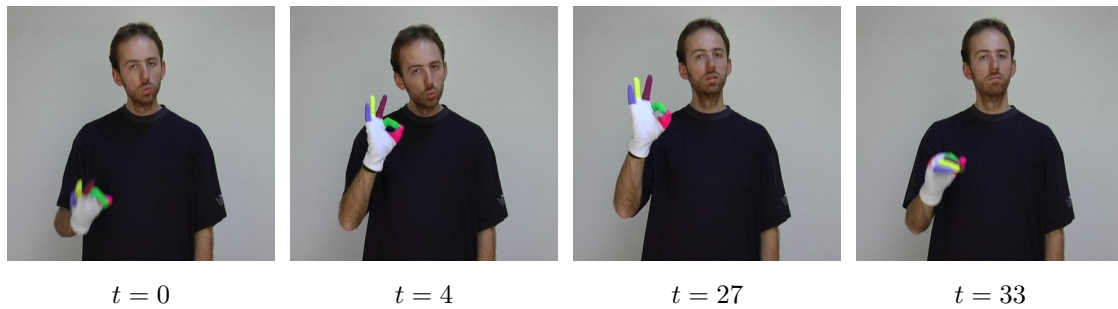


Abbildung 6.6: Die Gebärde „gut“, aufgezeichnet mit 25 Hz. Die Ausführungsphase entspricht $4 \leq t \leq 27$. In diesem Intervall gilt $b_{\underline{P},DH}^R(t) = 1$, außerhalb ist $b_{\underline{P},DH}^R(t) = 0$. Da die Gebärde einhändig und ferner die Position der dominanten Hand von untergeordneter Bedeutung ist, sind $b_{\underline{S},DH}^R(t) = b_{\underline{S},NDH}^R(t) = 0 \forall t$.

verdeutlicht. Auf die Angabe der im letzten Abschnitt diskutierten Abstandsmaße wird hierbei verzichtet, da diese für den Benutzer wenig transparent sind. Auch ohne Quantifizierung ist die Rückmeldung des Tutors klar verständlich.

Zur Visualisierung des Feedbacks dienen im Folgenden beispielhaft die Gebärden „Bildtelefon“ und „gut“, die in Abbildung 6.5 bzw. 6.6 gezeigt sind. Im Unterschied zu den bisherigen Abbildungen wird eine Markergeometrie verwendet, in der etwa 85% der Finger bedeckt sind. Diese Anordnung hat sich in der Evaluation (s. Abschnitt 7.1.2.5) als zuverlässiger erwiesen.

Bei falscher Handposition erfolgt ein Feedback wie in Abbildung 6.7 am Beispiel der Gebärde „Bildtelefon“ veranschaulicht. In der linken Bildhälfte ist das Einzelbild t_i^A der Anwendergebärde zu sehen, rechts das diesem zugeordnete Einzelbild t_i^R der Referenzgebärde. Pfeile verdeutlichen die Positionsabweichung der Hände und geben die korrekte Position an.

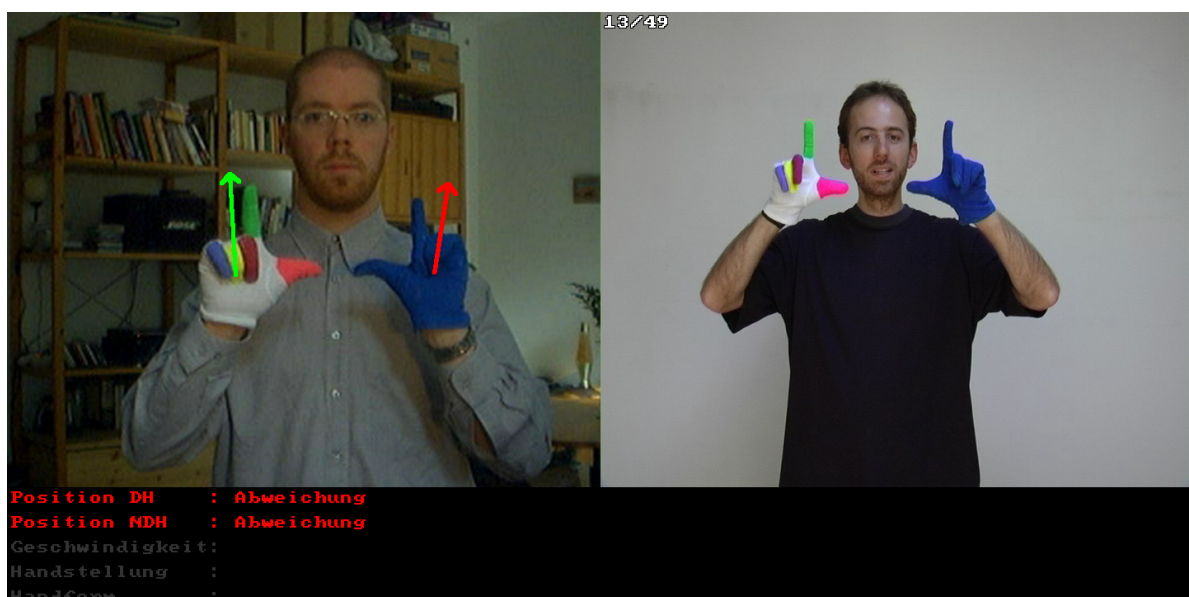


Abbildung 6.7: Korrekatives Feedback bei abweichender Handposition am Beispiel der Gebärde „Bildtelefon“.

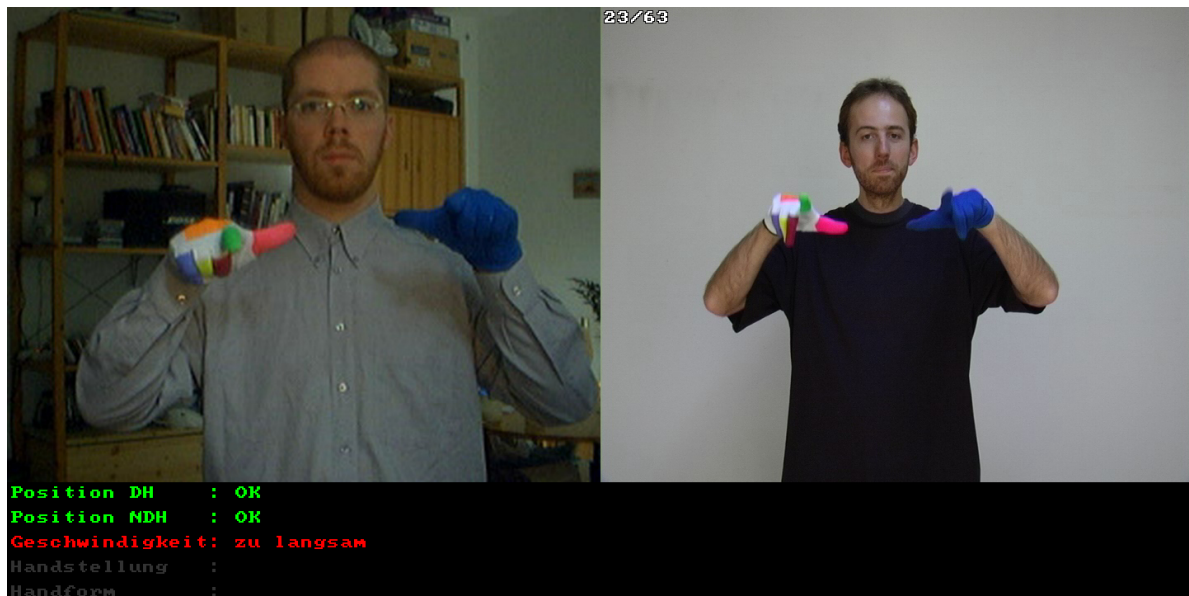


Abbildung 6.8: Korrektives Feedback bei abweichender Ausführungsgeschwindigkeit am Beispiel der Gebärde „Bildtelefon“. Die Anwendergebärde wurde in dieser Phase zu langsam ausgeführt.

Abbildung 6.8 zeigt die Rückmeldung des Tutors bei korrekter Handposition, aber abweichender Ausführungsgeschwindigkeit. Die Anzeige erfolgt in diesem Fall rein textuell.

Bei abweichender Handstellung blendet das Tutorsystem die in der Anwendergebärde erkannte Handpositur wie in Abbildung 6.9 gezeigt ein. Zusätzlich erhält der Benutzer einen

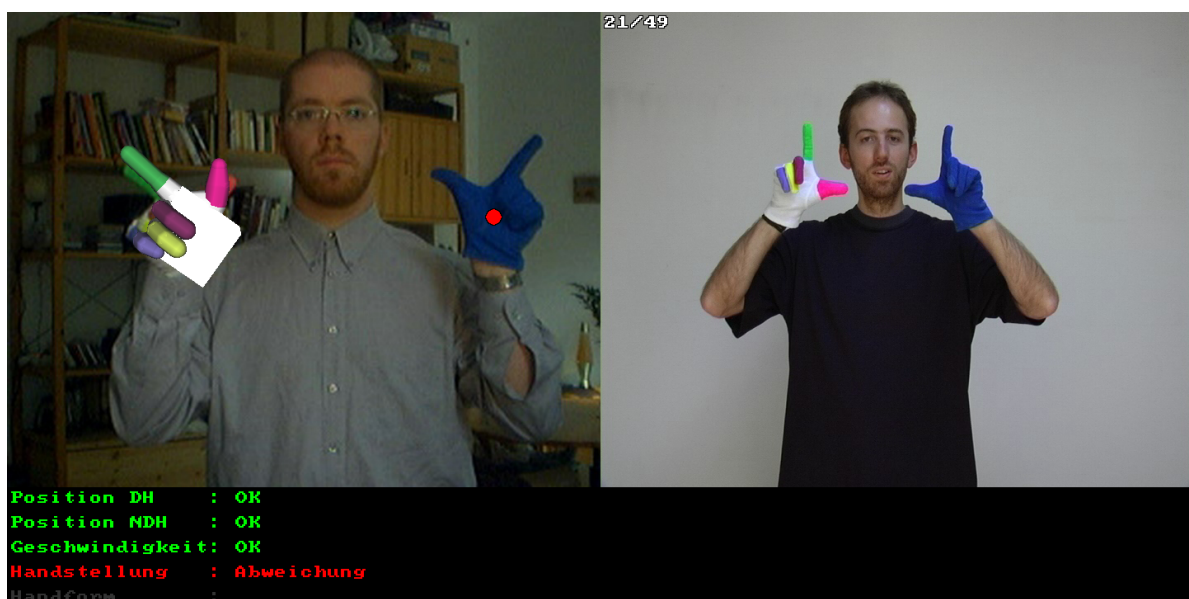


Abbildung 6.9: Korrektives Feedback bei abweichender Handstellung am Beispiel der Gebärde „Bildtelefon“. Im Vergleich zur Referenzpositur ist die Anwenderpositur etwa 60° nach außen rotiert.

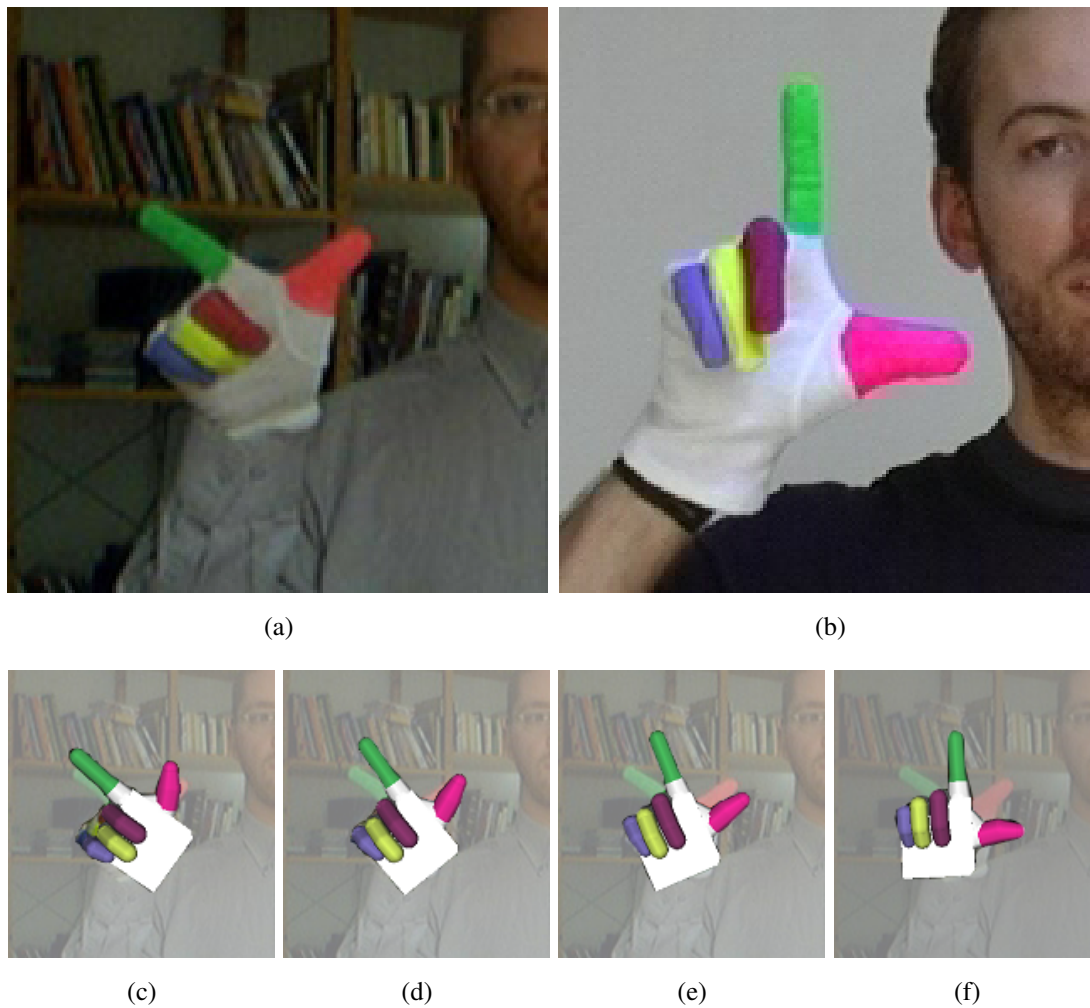


Abbildung 6.10: Korrektives Feedback bei abweichender Handstellung. In (a) ist die Eingangspositur gezeigt, in (b) die Ansicht der Referenzpositur des Gebärdensprechers aus der rechten Bildhälfte in Abbildung 6.9. Die erkannte abweichende Positur wird auf Tastendruck in die Referenzpositur bewegt (c–f).

Hinweis in textueller Form. Auf Tastendruck wird die fehlerhafte Handpositur mit wählbarer Geschwindigkeit in die Referenzpositur bewegt. Im betrachteten Beispiel ist die dominante Hand des Anwenders in der Bildebene um etwa 60° auswärts rotiert, während für die Referenzpositur $\phi_{\text{rot}} \approx 0$ gilt. Abbildung 6.10 visualisiert die Rotation von der fehlerhaften in die korrekte Handstellung.

Eine von der Referenz abweichende Handform wird dem Lernenden auf dieselbe Weise zurückgemeldet. Für das Beispiel der Gebärde „gut“ ist in Abbildung 6.11 das Feedback des Systems zu sehen. Abbildung 6.12 veranschaulicht die Bewegung der fehlerhaften Handpositur des Anwenders in die korrekte Positur der Referenzausführung.

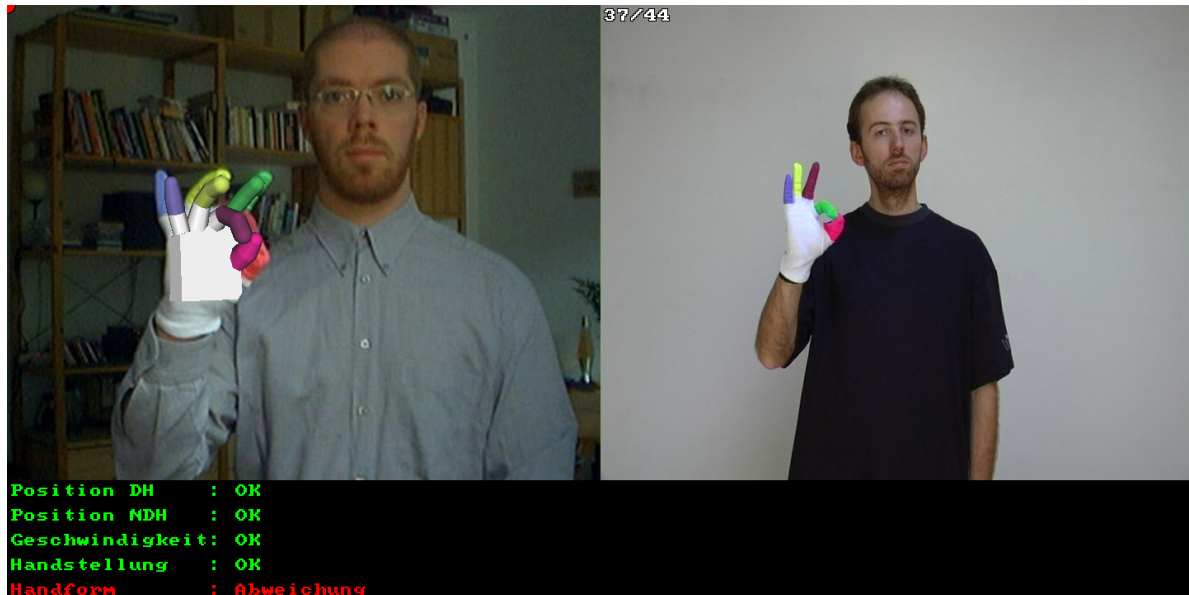


Abbildung 6.11: Korrekatives Feedback bei abweichender Handform am Beispiel der Gebärde „gut“. In der korrekten Ausführung ist der Daumen dem Zeigefinger opponiert, in der Ausführung des Anwenders jedoch dem Mittelfinger.

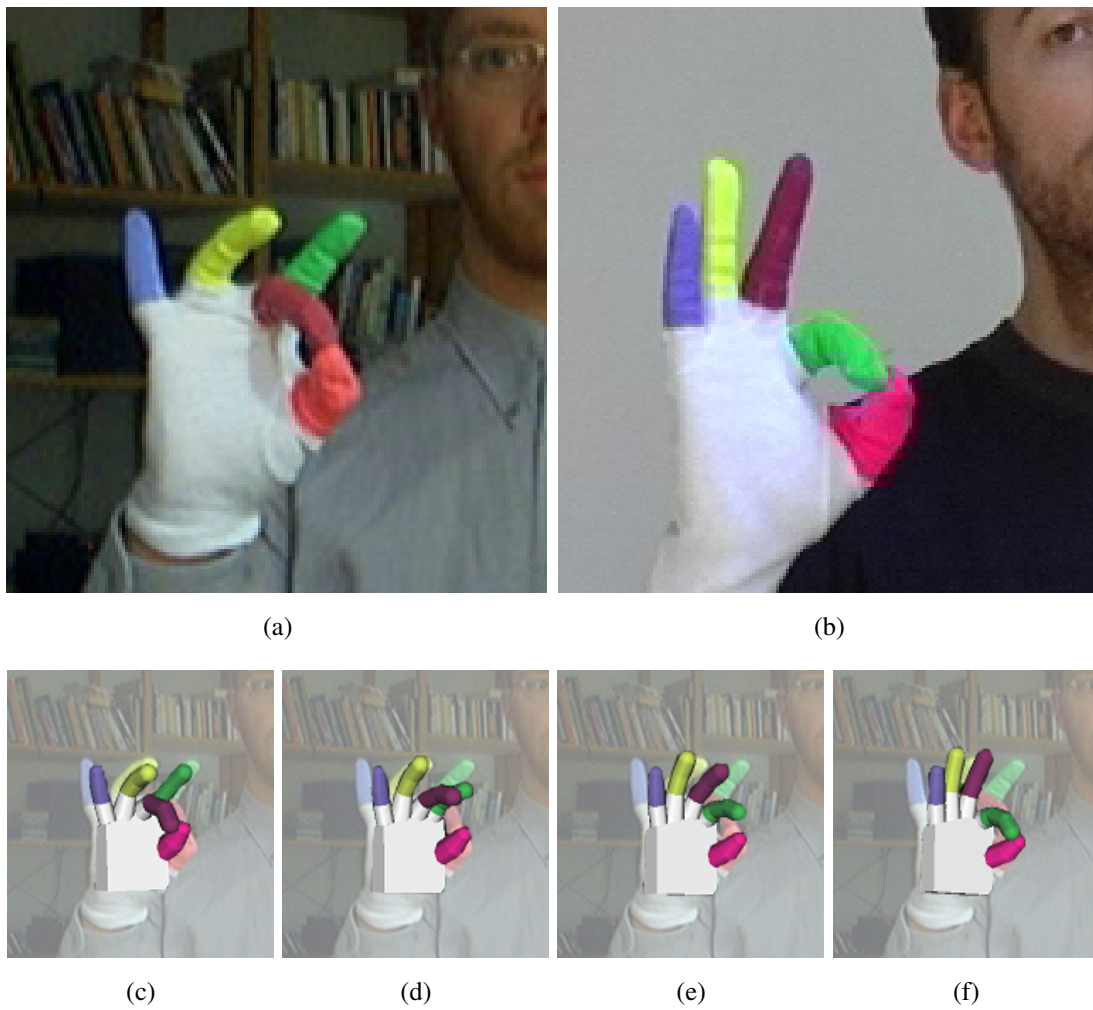


Abbildung 6.12: Korrektives Feedback bei abweichender Handform (analog zu Abbildung 6.10).

Hat der Anwender die Gebärde in jeder Hinsicht korrekt ausgeführt, erfolgt eine textuelle Bestätigung gemäß Abbildung 6.13. Darin sind die vier erläuterten Vergleichskriterien Handposition, Bewegungsgeschwindigkeit, Handstellung und Handform explizit aufgeführt.

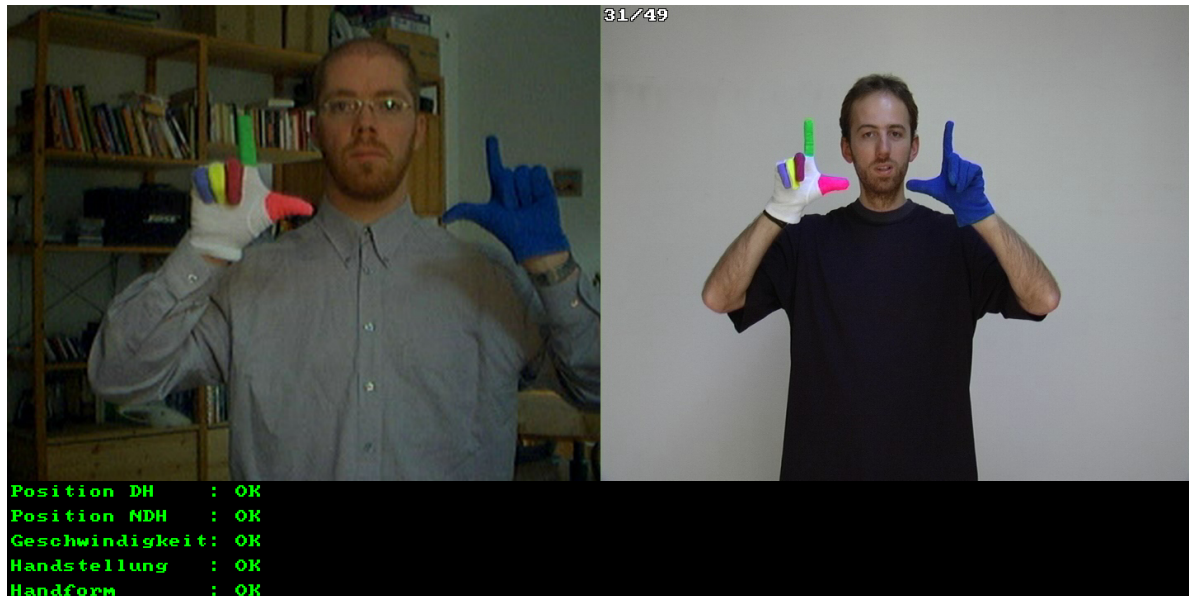


Abbildung 6.13: Feedback bei vollständig korrekter Gebärdenausführung.

Kapitel 7

Evaluation

Im Folgenden wird die Leistungsfähigkeit der Handpositurerkennung analysiert und der Einfluss verschiedener Parameter untersucht. Aufgrund der hohen Dimensionalität des Positurreaums ($\mathbf{P} \in \mathbb{R}^{23}$; vgl. Abschnitt 3.2.3) sind hierzu große Mengen von Testdaten erforderlich. Beispielsweise hat eine Messung mit 1000 zufällig ausgewählten Posituren noch Stichprobencharakter und kann bei mehrmaliger Durchführung deutlich verschiedene Ergebnisse liefern. Im Fall einer Gebärde besteht zwischen den Posituren eine starke Korrelation, so dass die für eine aussagekräftige Evaluation benötigte Datenmenge weiter ansteigt.

Das Handmodell bietet die Möglichkeit, synthetische Testdaten im benötigten Umfang mit geringem Aufwand zu erzeugen. Da die vom System zu rekonstruierenden Positurparameter hierbei in jedem Einzelbild bekannt sind, ist durch den Vergleich mit dem Rekonstruktionsergebnis eine quantitative Evaluation möglich. Insbesondere kann die Abhängigkeit der Erkennungsleistung von einzelnen Faktoren wie etwa Markergeometrie, Bildqualität oder Hypothesenanzahl erfasst werden. Die Evaluation mit synthetischen Positursichten ist in Abschnitt 7.1 beschrieben.

Um die Genauigkeit der Rekonstruktion in realen Aufnahmen quantitativ zu bewerten, müssen die korrekten Posituren manuell eingegeben werden (*ground truth data*). Hierzu wird das Handmodell in das Eingangsbild eingeblendet und seine Parameter so angepasst, dass die resultierende Ansicht der realen Positur maximal ähnlich ist. Ein geübter Bediener benötigt für diesen Vorgang etwa 1–3 Minuten pro Positur, d. h. ca. 1–2 Stunden pro Gebärde. Daher steht für die Evaluation eine entsprechend kleinere Datenmenge als im Fall synthetischer Eingangsdaten zur Verfügung. In Abschnitt 7.2 sind die für diese Testdaten ermittelten Ergebnisse präsentiert.

Gegenstand der Analyse sind die beiden Ausdrucksmittel Handform und Handstellung, da diese den Schwerpunkt der vorliegenden Arbeit bilden. Die Rekonstruktion der Handposition ist ausschließlich von der Segmentierung der Eingangsbilder (vgl. Abschnitte 4.2 und 4.3) sowie der anschließenden Regionenzuordnung (vgl. Abschnitt 4.5) abhängig und erfolgt in typischen Büroumgebungen ohne nennenswerte Fehler.¹

¹Eine Evaluation verschiedener Segmentierungsalgorithmen im Kontext der Objektklassifikation findet sich

Auf Grundlage der anschaulichen Differenzmaße $\Delta_{\underline{B}}$ und $\Delta_{\underline{\Phi}}$ wird die Rekonstruktionsgenauigkeit bezüglich Handform und Handstellung getrennt quantifiziert. Daneben kommt die Posituredifferenz $\Delta_{\underline{P}}$ zum Einsatz, die beide Ausdrucksmittel in einem skalaren Wert zusammenfasst und damit eine relative Einschätzung der Erkennungsleistung bei Variation der Systemparameter erleichtert. Alle drei Maße sind in Abschnitt 3.3 ausführlich erläutert.

7.1 Synthetische Posituransichten

Mit Hilfe synthetischer Posituransichten lässt sich nicht nur das Gesamtsystem evaluieren, sondern insbesondere auch der Einfluss verschiedener Komponenten separat untersuchen. Auf diese Weise ist auch die Bestimmung geeigneter Werte für verschiedene Systemparameter möglich. Da eine vollständige Abdeckung aller anatomisch zulässigen Handformen und -stellungen aus Komplexitätsgründen nicht praktikabel ist, werden sämtliche verwendeten Posituren bzw. Positurensequenzen zufällig generiert. Systematische Effekte, die die Analyse verfälschen und die Gefahr einer Überanpassung des Systems bergen, sind damit weitgehend ausgeschlossen.

In Abschnitt 7.1.1 ist die Erzeugung der synthetischen Testdaten erläutert. Anschließend analysiert Abschnitt 7.1.2 verschiedene Komponenten des Erkennungssystems und ihren Einfluss auf die Rekonstruktionsgenauigkeit für unkorrelierte Einzelposituren. Auf die Verarbeitung zeitlich korrelierter Positurensequenzen geht Abschnitt 7.1.3 ein.

7.1.1 Generierung der Eingangsdaten

Zur Evaluation mit synthetischen Eingangsdaten werden einzelne, unkorrelierte Posituren sowie korrelierte Positurensequenzen – d. h. flüssige Bewegungen – zufällig generiert. Dies erfolgt unter Ausnutzen des gesamten anatomisch möglichen Positurenraums. Eine Beschränkung auf in der DGS häufige Posituren findet bewusst nicht statt, da die Positurparameter der Anwendergebärde infolge fehlerhafter Ausführung auch außerhalb dieser Untermenge liegen können. Anhang D beschreibt die verwendeten Zufallsalgorithmen im Detail. Beispiele einzelner Zufallsposituren sowie eine zufällige Positurensequenz sind in den Abbildungen 7.1 und 7.2 zu sehen.

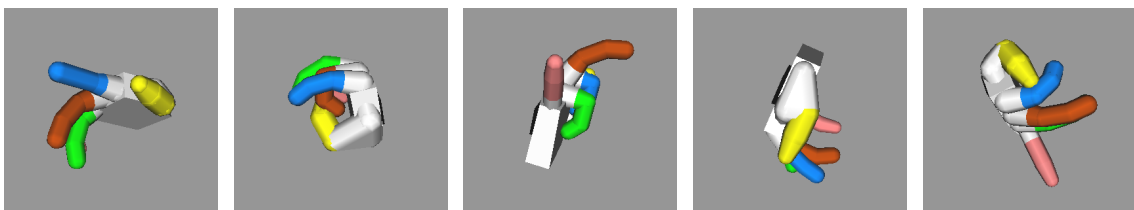
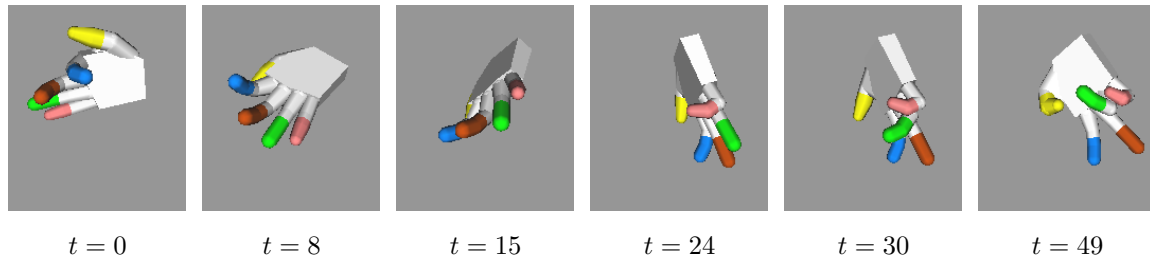


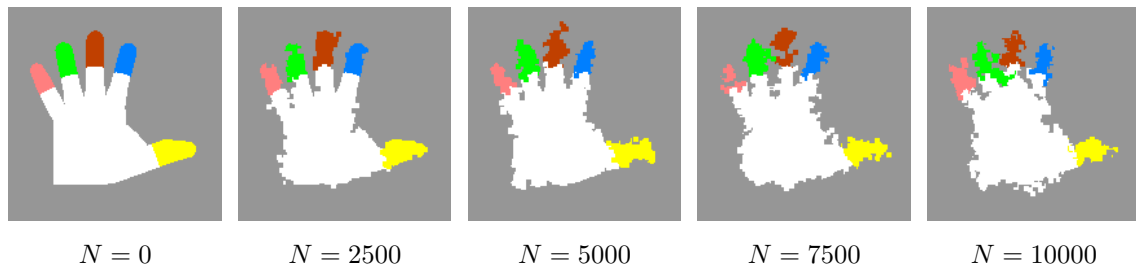
Abbildung 7.1: Zufällig generierte Einzelposituren.

in [Alv04].

Abbildung 7.2: Zufällig generierte Positursenfolge der Länge $T = 50$.

Um die prinzipielle Übertragbarkeit der Simulation auf die praktische Anwendung sicherzustellen, sollten die synthetischen Positursichten mit realen Eingangsdaten vergleichbar sein. Aus diesem Grund wird die Qualität der synthetischen Bilder durch einen stochastischen Rauschprozess variabler Intensität reduziert. Dieser Rauschprozess sowie seine Auswirkung auf den Merkmalsvektor $\underline{E}_{DH}$ sind ebenfalls in Anhang D beschrieben. Die Intensität des Rauschens wird in Iterationen $N \in \{0, 2500, 5000, 7500, 10000\}$ angegeben und ist in Abbildung 7.3 veranschaulicht.

Wie bereits in Abschnitt 5.1.2 erwähnt, beträgt die Auflösung der synthetischen Positursichten 256×256 . Die Hand belegt maximal ca. 13% der Bildfläche, d. h. etwa 8500 der insgesamt 65536 Pixel.

Abbildung 7.3: Auswirkung des Rauschprozesses bei verschiedenen Iterationsanzahlen N .

7.1.2 Erkennungsleistung für unkorrelierte Einzelposituren

Im Folgenden werden die Komponenten des Erkennungssystems nacheinander in die Evaluation einbezogen, so dass ihr jeweiliger Einfluss deutlich wird. Abschnitt 7.1.2.1 betrachtet ausschließlich die Positursdatenbank und ignoriert jegliche Ungenauigkeiten, die sich im Rahmen der Extraktion oder der Interpretation der Merkmale ergeben können. Hieraus ergibt sich die bei der gewählten Diskretisierung des Positursraums (vgl. Abschnitt 5.1.1) maximal erreichbare Rekonstruktionsgenauigkeit in unkorrelierten Einzelbildern.

Die tatsächliche Aussagekraft des Merkmalsvektors, insbesondere in Verbindung mit dem in der Realität auftretenden Bildrauschen, geht ab Abschnitt 7.1.2.2 in die Betrachtung ein. Hier wird zunächst die Trennschärfe der Merkmalsdifferenz untersucht und die Eignung der Hausdorff-Distanz nachgewiesen. Danach behandelt Abschnitt 7.1.2.3 das beschleunigte Datenbank-Zugriffsverfahren, das auf Merkmalsvektor und Merkmalsdifferenz aufbaut.

In den drei letztgenannten Abschnitten steigt die Schwierigkeit der Erkennungsaufgabe schrittweise an, bis schließlich der praktische Anwendungsfall erreicht ist. Um Vergleichbarkeit zu gewährleisten, basieren alle Messungen auf einer konstanten Testdatenmenge P_{in} , die 1000 zufällige Posituren enthält. In Abschnitt 7.1.2.4 werden für verschiedene Beispielposituren aus P_{in} die jeweiligen Erkennungsergebnisse gezeigt und die erreichte Genauigkeit quantitativ angegeben, so dass der Einfluss der einzelnen Verarbeitungsstufen anschaulich nachvollziehbar ist.

Abschnitt 7.1.2.5 vergleicht mehrere Markergeometrien und motiviert die zur Evaluation mit realen Posituransichten verwendete Anordnung.

7.1.2.1 Evaluation der Posituredatenbank

Eine Grundvoraussetzung für die Erkennung beliebiger Handformen und -stellungen besteht darin, dass die Posituredatenbank für sämtliche anatomisch möglichen Eingangsposituren eine oder mehrere annähernd identische Kandidatenposituren liefern kann. Für eine Eingangspositur $\underline{P}_{in}$ lässt sich die bezüglich der Posituredifferenz Δ_P optimale Kandidatenpositur

$$\underline{P}_{opt, \Delta_P} = \underset{\underline{P} \in P_{db}}{\operatorname{argmin}} \Delta_P(\underline{P}, \underline{P}_{in}) \quad (7.1)$$

mittels sequenzieller Suche berechnen. Aufgrund der Diskretisierung des Positurenraums ist $\underline{P}_{in}$ im Allgemeinen nicht exakt in der Datenbank enthalten, so dass $\Delta_P(\underline{P}_{opt, \Delta_P}, \underline{P}_{in}) > 0$ gilt.

Durch Berechnung der minimalen Posituredifferenz gemäß (7.1) sowie der Handform- und -stellungsabweichungen Δ_B und Δ_Φ der entsprechenden Posituren $\underline{P}_{opt, \Delta_P}$ für die Testposituren P_{in} ergibt sich eine über alle Finger gemittelte Abweichung von lediglich $\Delta_B \approx 1,29$ cm. Der mittlere Handstellungsfehler beträgt $\Delta_\Phi \approx 8,02^\circ$. Diese Werte liegen im Bereich der menschlichen Erkennungsleistung und sind als sehr gut zu bewerten. Sie belegen, dass die Posituredatenbank hinsichtlich ihres Umfangs sowie der Verteilung der Datenbankeinträge im Positurenraum für die Rekonstruktionsaufgabe geeignet ist.

Histogramme der einzelnen Differenzmaße sind in Anhang E.1 abgebildet. Der größte Beitrag zur Handformdifferenz ist auf den Daumen zurückzuführen, so dass hier zunächst eine feinere Quantisierung der entsprechenden Beugungswinkel sinnvoll erscheint. In der praktischen Anwendung ist die Rekonstruktionsgenauigkeit bezüglich des Daumens jedoch primär durch projektionsbedingte Mehrdeutigkeiten beschränkt. Aus diesem Grund enthält die Datenbank lediglich sieben verschiedene Daumenstellungen (vgl. Tabelle 5.1), die in der Mehrzahl der Ansichten vergleichsweise gut unterscheidbar sind.

7.1.2.2 Evaluation der Merkmalsdifferenz

In der praktischen Anwendung findet das System die Positurkandidaten über die Minimierung der Merkmalsdifferenz $\Delta_{\mathbf{F}}$, die Anforderung besteht jedoch weiterhin in möglichst geringen Abweichungen $\Delta_{\mathbf{\Phi}}$ und $\Delta_{\mathbf{B}}$. Um die Eignung der Merkmalsdifferenz – und damit auch des Merkmalsvektors $\mathbf{F}_{\text{DH}}$ – für diese Aufgabe nachzuweisen, wird für die Posituren $\mathbf{P}_{\text{in}}$ jeweils die $\Delta_{\mathbf{F}}$ minimierende Datenbankpositur $\mathbf{P}_{\text{opt},\Delta_{\mathbf{F}}}$ bestimmt:

$$\mathbf{P}_{\text{opt},\Delta_{\mathbf{F}}} = \underset{\mathbf{P} \in \mathbf{P}^{\text{db}}}{\operatorname{argmin}} \Delta_{\mathbf{F}}(\mathbf{F}_{\text{DH}}(\mathbf{P}), \mathbf{F}_{\text{DH}}(\mathbf{P}_{\text{in}})) \quad (7.2)$$

Die Suche erfolgt erneut sequenziell über die gesamte Posituredatenbank, so dass ausschließlich die Merkmalsdifferenz zum Tragen kommt. Das in Abschnitt 5.1.4 eingeführte beschleunigte Zugriffsverfahren ist in Abschnitt 7.1.2.3 separat evaluiert.

Verglichen mit dem im letzten Abschnitt gemessenen Werten steigt $E(\Delta_{\mathbf{B}})$ auf 2,22 cm und $E(\Delta_{\mathbf{\Phi}})$ auf 20,87°. Damit liegt die Abweichung der Posituren $\mathbf{P}_{\text{opt},\Delta_{\mathbf{F}}}$ etwa 1 cm bzw. 12° unterhalb des durch die Quantisierung des Positurenraums vorgegebenen Optimums.

Um die Robustheit der Merkmalsdifferenz $\Delta_{\mathbf{F}}$ nachzuweisen, wurden die Eingangsposituren $\mathbf{P}_{\text{in}}$ mit Rauschintensitäten $N \in \{2500, 5000, 7500, 10000\}$ degeneriert (s. Anhang D.3). Die Differenzmaße steigen auch bei starkem Rauschen nur geringfügig an; bei $N = 7500$ betragen sie etwa $E(\Delta_{\mathbf{B}}) \approx 2,59$ cm bzw. $E(\Delta_{\mathbf{\Phi}}) \approx 26,63^\circ$. Damit ist die Tauglichkeit der Merkmalsdifferenz für die Verarbeitung über- und untersegmentierter Eingangsbilder belegt. Dies schließt auch den Merkmalsvektor $\mathbf{F}_{\text{DH}}$ ein, wobei sich insbesondere die Verwendung der elliptischen Regionenapproximation als sinnvoll erweist. Zwar entsteht hierdurch bei der Verarbeitung ungestörter Daten ein Informationsverlust, im realistischen Fall wird jedoch das Bildrauschen wirkungsvoll kompensiert. Dieser Effekt bildet daher neben den in Abschnitt 4.3.5 beschriebenen Aspekten eine weitere Motivation für die approximative Randkurvenbeschreibung.

Auch bei Verwendung der Merkmalsdifferenz ist die Abweichung hinsichtlich der Daumenstellung aufgrund der größeren Beweglichkeit und der damit verbundenen Mehrdeutigkeiten größer als die der Finger. Die Histogramme der obigen Messungen (einschließlich $N = 0$) sind in Anhang E.2 zu finden.

7.1.2.3 Evaluation des Datenbankzugriffs

Aus Geschwindigkeitsgründen erfolgt der Datenbankzugriff in der tatsächlichen Anwendung mit Hilfe des in Abschnitt 5.1.4 beschriebenen balancierten Suchbaums. Leistung und Geschwindigkeit sind dabei von der Auswahl der zur Indexierung verwendeten Merkmale sowie vom Verzweigungsgrad des Indexbaums abhängig. Eine detaillierter Vergleich verschiedener Varianten erfolgt in Anhang E.3. Als geeignet erweist sich die Kombination der sechs normierten Radien und der normierten Flächen von Daumen- und Handrückenmarker. Die mittlere

Antwortzeit beträgt hierbei nur ca. 61 ms. Dieser Wert ist unabhängig von der Rauschintensität N . Er entspricht einer Geschwindigkeitssteigerung etwa um den Faktor 235 gegenüber der sequenziellen Suche, wobei mit $E(\Delta_{\underline{\mathbf{B}}}) \approx 2,64 \text{ cm}$ und $E(\Delta_{\underline{\Phi}}) \approx 27,66^\circ$ für $N = 7500$ annähernd dieselbe Rekonstruktionsgenauigkeit erreicht wird. Im Folgenden kommt diese Indierungsvariante bei allen Analysen sowohl für synthetische als auch für reale Eingangsdaten zur Verwendung.

7.1.2.4 Beispiele rekonstruierter Posituren

Abbildung 7.4 zeigt fünf Beispielposituren der Testmenge P_{in} und die rekonstruierten Posituren $\underline{P}_{\text{opt},\Delta_{\underline{\mathbf{P}}}}$ sowie $\underline{P}_{\text{opt},\Delta_{\underline{\mathbf{F}}}}$ bei sequenzieller bzw. indexbasierter Suche und $N = 7500$. Die entsprechenden Abweichungen sind in Tabelle 7.1 aufgelistet.

Alle Ergebnisposituren $\underline{P}_{\text{opt},\Delta_{\underline{\mathbf{P}}}}$ weisen eine hohe Übereinstimmung auf. Fehler entstehen hier ausschließlich durch „Quantisierungsrauschen“, d. h. die Entfernung zwischen der Eingangspositur und der gemäß $\Delta_{\underline{\mathbf{P}}}$ nächstgelegenen Datenbankpositur.

Erwartungsgemäß zeigen die Posituren $\underline{P}_{\text{opt},\Delta_{\underline{\mathbf{F}}}}$ geringfügig höhere Abweichungen. Hier liefert die indexbasierte Suche etwa gleich gute Ergebnisse wie die sequenzielle Suche. Bei drei der fünf Beispielposituren (a, c, d) übertrifft sie diese sogar, denn der Suchbaum findet mit-

Tabelle 7.1: Differenzmaße der in Abbildung 7.4 gezeigten Posituren.

Verfahren	Posi- tur	$\Delta_{\underline{\Phi}}$ [°]	$\Delta_{\underline{\mathbf{B}},\text{D}}$	$\Delta_{\underline{\mathbf{B}},\text{ZF}}$	$\Delta_{\underline{\mathbf{B}},\text{MF}}$ [cm]	$\Delta_{\underline{\mathbf{B}},\text{RF}}$	$\Delta_{\underline{\mathbf{B}},\text{KF}}$	$\bar{\Delta}_{\underline{\mathbf{B}}}$ [cm]	$\Delta_{\underline{\mathbf{P}}}$ [cm ²]
$\underline{P}_{\text{opt},\Delta_{\underline{\mathbf{P}}}}$	(a)	10,34	0,82	0,36	0,69	0,12	0,14	0,43	6,30
	(b)	8,20	2,76	0,86	0,00	1,02	0,38	1,00	12,70
	(c)	1,55	2,46	1,18	0,71	1,15	2,59	1,62	16,10
	(d)	3,18	3,39	0,94	1,93	0,00	1,56	1,57	20,00
	(e)	9,77	0,00	0,00	1,07	1,38	0,82	0,65	20,22
$\underline{P}_{\text{opt},\Delta_{\underline{\mathbf{F}}}}$ (sequenzi- elle Suche)	(a)	24,18	3,04	0,36	0,69	0,12	3,89	1,62	51,90
	(b)	10,21	5,08	1,72	0,00	1,02	0,38	1,64	34,84
	(c)	26,05	3,46	1,18	0,71	3,42	2,20	2,19	61,47
	(d)	5,08	3,45	1,78	9,51	0,00	1,56	3,26	171,27
	(e)	37,09	0,00	0,00	2,98	1,37	0,81	1,03	171,41
$\underline{P}_{\text{opt},\Delta_{\underline{\mathbf{F}}}}$ (index- basierte Suche)	(a)	11,91	3,04	0,36	0,69	0,12	0,14	0,87	16,48
	(b)	10,71	5,08	1,72	0,00	1,02	0,38	1,64	35,33
	(c)	7,13	3,46	1,18	0,71	2,76	2,59	2,13	30,51
	(d)	29,55	3,45	0,94	2,06	0,00	5,48	2,39	149,03
	(e)	37,27	0,00	0,00	2,98	4,17	0,81	1,56	187,50

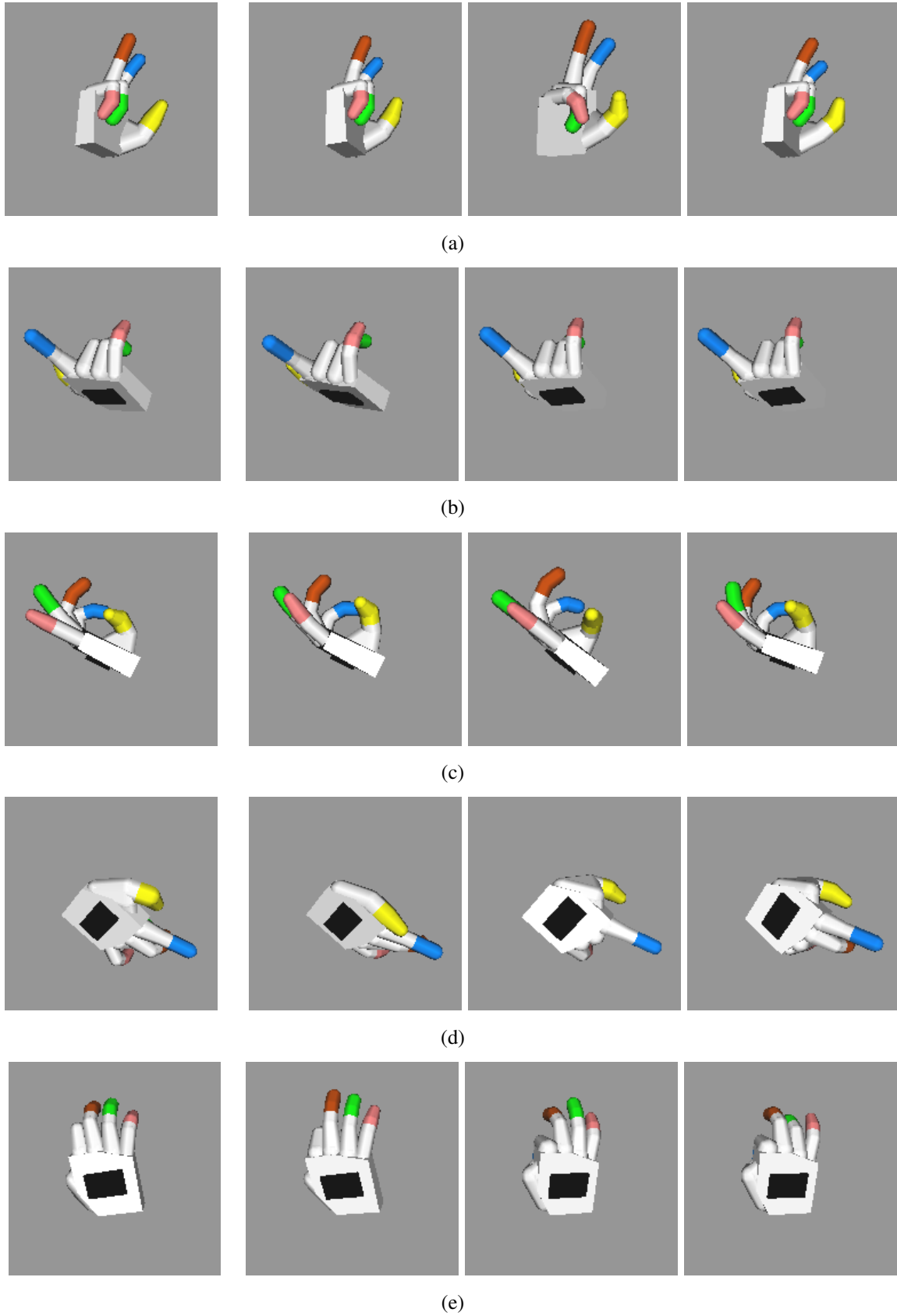


Abbildung 7.4: Testposituren aus P_{in} (1. Spalte) und Ergebnisse $\underline{P}_{opt, \Delta P}$ (2. Spalte) sowie $\underline{P}_{opt, \Delta F}$ bei sequenzieller (3. Spalte) bzw. indexbasierter Suche (4. Spalte) und $N = 7500$.

unter eine hinsichtlich $\Delta_{\underline{F}}$ schlechtere, bezüglich der Posituredifferenzmaße jedoch bessere Positur.

7.1.2.5 Variation der Markergeometrie

Form und Anordnung der Marker sind prinzipiell beliebig wählbar und sollten im Sinne maximaler Erkennungsleistung optimiert werden. Daher erfolgt in diesem Abschnitt ein Vergleich der bisher betrachteten mit drei weiteren Markergeometrien. In Abbildung 7.5 sind die vier Testfälle präsentiert.

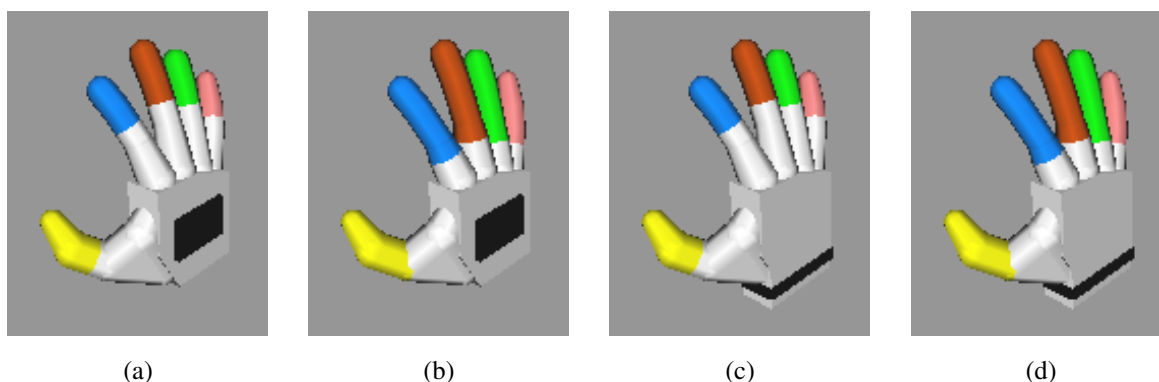


Abbildung 7.5: Verschiedene Markergeometrien. Bei der bisher betrachteten Anordnung (a) bedecken die Markierungen etwa 45% der Finger. Diese Variante wird verglichen mit einer Abdeckung von etwa 85% (b) sowie einer Anordnung, die an Stelle des Handrückenmarkers einen umlaufenden Handgelenkmarker verwendet (c). In (d) ist der Handgelenkmarker mit einer 85%igen Abdeckung kombiniert.

Tabelle 7.2 gibt die gemessenen Rekonstruktionsgenauigkeiten für die Eingangsposituren P_{in} bei $N = 7500$ an. Da die in Abbildung 7.5b gezeigte Variante die niedrigste mittlere Posituredifferenz $\Delta_{\underline{P}}$ erreicht, kommt sie in allen folgenden Analysen sowie in Realansichten zum Einsatz.

Tabelle 7.2: Rekonstruktionsgenauigkeit für die Markergeometrien in Abbildung 7.5.

Variante	Markierungen	$E(\Delta_{\underline{F}})$ [°]	$E(\Delta_{\underline{B}})$ [cm]	$E(\Delta_{\underline{P}})$ [cm ²]
(a)	Handrücken; Finger ca. 45%	27,66	2,64	124,63
(b)	Handrücken; Finger ca. 85%	22,68	2,70	103,65
(c)	Handgelenk; Finger ca. 45%	27,57	2,64	124,84
(d)	Handgelenk; Finger ca. 85%	23,37	2,77	112,55

7.1.3 Erkennungsleistung in Positurreihenfolgen

Bei der Verarbeitung von Positurreihenfolgen besteht zwischen aufeinanderfolgenden Einzelbildern eine Korrelation, die die Rekonstruktion vereinfacht. Zwei Eigenschaften des zur Generierung der Reihenfolgen verwendeten Zufallsalgorithmus (s. Anhang D.2) erschweren hingegen den Erkennungsprozess:

- Durch zufällige Richtungswechsel hinsichtlich der Änderung von Handform und -stellung (d. h. Unstetigkeit der entsprechenden Ableitungen) wird die Bewegung unregelmäßig und die Korrelation somit reduziert.
- Während in den zufällig generierten Einzelposituren P_{in} jeder Finger mit einer Wahrscheinlichkeit von etwa 13% verdeckt ist, beträgt diese Wahrscheinlichkeit in den Positurreihenfolgen lediglich ca. 5%. Da nur sichtbare Finger einen Beitrag zu Δ_B liefern, wird $E(\Delta_B)$ hierdurch erhöht.

Das Ziel der Analyse ist keine Prädiktion der Leistung in realen Posituranalysen, sondern der Nachweis der Robustheit der gesamten Verarbeitungskette. Aus diesem Grund werden auch die synthetischen Positurreihenfolgen mit Rauschen der Intensität $N = 7500$ gestört. Sie sind damit visuell stärker degeneriert als Realaufnahmen unter durchschnittlichen Bedingungen.

Insbesondere die Rekonstruktion der Handstellung profitiert von der zeitlichen Korrelation, so dass die mittlere Abweichung hier auf $16,62^\circ$ sinkt. Hinsichtlich der Handform liegt die Genauigkeit mit $E(\Delta_B) \approx 2,66$ cm auf dem für unkorrelierte Einzelposituren erreichten Niveau.

Unter Verwendung zufälliger Positurreihenfolgen lassen sich die Parameter der Rekonstruktionsstufe auf einfache Weise manuell optimieren. Die so ermittelten Werte sowie weitere Einzelheiten der Genauigkeitsmessung sind in Anhang E.4 zu finden.

7.2 Reale Posituranalysen

Die Verwendung realer Posituranalysen zur Messung der Rekonstruktionsgenauigkeit bietet den Vorteil, das tatsächliche Anwendungsszenario exakt zu reproduzieren, und gewährleistet damit eine hohe Aussagekraft. Im Unterschied zu synthetischen Posituranalysen sind die Positurparameter jedoch nicht bekannt und müssen daher manuell eingegeben werden. Zuvor erfolgt eine Anpassung der Modellgeometrie an den Benutzer, um die Erkennungsleistung zu optimieren. Diese beiden Vorgänge sind in den Abschnitten 7.2.1 und 7.2.2 erläutert.

Anschließend präsentiert Abschnitt 7.2.3 die Ergebnisse der Leistungsmessung der Handpositurerkennung. Dies umfasst die Verarbeitungsgeschwindigkeit in Hz sowie die Genauigkeit hinsichtlich der Maße Δ_Φ , Δ_B und Δ_P . Abschnitt 7.2.4 enthält eine ausführliche Diskussion anhand mehrerer Beispiele rekonstruierter Posituren sowie einen Vergleich mit existierenden Systemen.

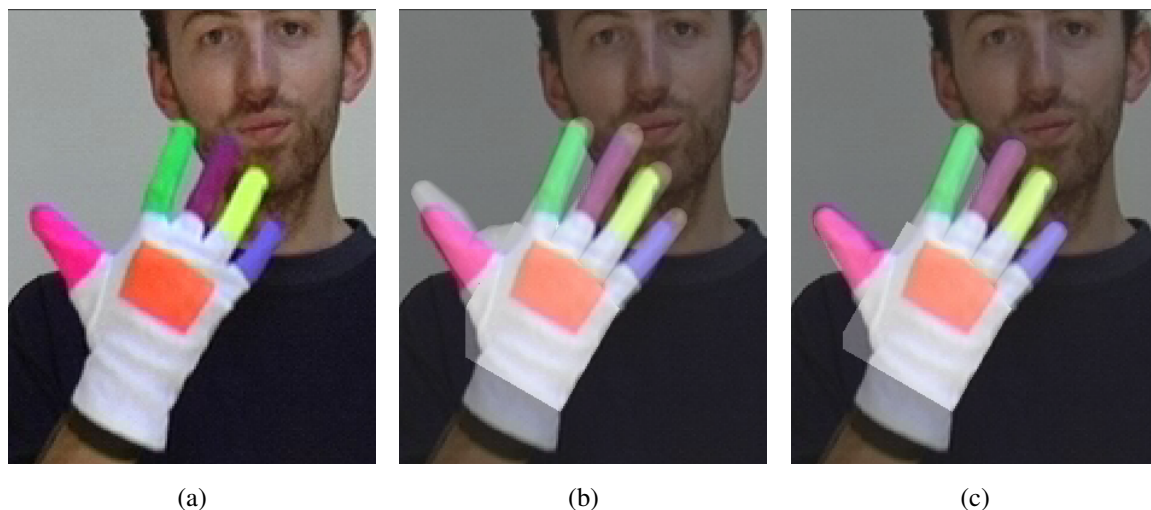


Abbildung 7.6: Anpassung der Handmodellgeometrie an den Gebärdensprecher. In die Ansicht der realen Hand (a) wird das Handmodell zunächst mit Standardabmessungen eingeblendet (b). Nach Anpassung der Modellgeometrie kommen reale und synthetische Ansicht zur Deckung (c).

7.2.1 Anpassung der Modellgeometrie

Zur Optimierung der Rekonstruktionsgenauigkeit wird die Geometrie des Handmodells im Folgenden an den Gebärdensprecher angepasst. Hierfür kommt ein interaktives Werkzeug zum Einsatz, das das Handmodell in eine Ansicht der realen Hand einblendet. Sämtliche Positurparameter sowie die Abmessungen der Modellelemente werden manuell so angepasst, dass beide Ansichten einander maximal ähnlich sind. Abbildung 7.6 zeigt hierzu ein Beispiel. Dieser Vorgang ist für verschiedene Handformen und -stellungen zu wiederholen, bis die Anpassung einzig durch Variation der Positurparameter möglich ist und die Handgeometrie konstant bleibt. Tabelle 7.3 gibt die auf diese Weise für den in Abbildung 7.6 gezeigten Gebärdensprecher bestimmten Maße der Modellelemente an (vgl. Abschnitt 3.2.1, Tabelle 3.2).

Tabelle 7.3: Anpassung der Handmodellgeometrie an den Gebärdensprecher.

Modellelement	Maße	Messwerte
Handteller	Breite×Höhe×Tiefe	8,75×9,25×3,00 cm
Daumen	PD / PP / Metakarpalknochen	1,00 / 2,25 / 2,25 cm
Zeigefinger	PD / PM / PP	1,75 / 2,50 / 3,00 cm
Mittelfinger	PD / PM / PP	2,00 / 3,00 / 3,00 cm
Ringfinger	PD / PM / PP	1,75 / 2,50 / 3,75 cm
Kleiner Finger	PD / PM / PP	1,75 / 1,75 / 3,00 cm

7.2.2 Eingabe der Positurparameter

Die quantitative Bewertung der Rekonstruktionsgenauigkeit erfordert die Kenntnis der Eingangspositur (d. h. der zu rekonstruierenden Positurparameter) in jedem Einzelbild. Die Eingabe dieser Daten geschieht – analog zu der im letzten Abschnitt beschriebenen Geometrie- anpassung – durch Einblenden des Handmodell in das Eingangsbild und manuelle Eingabe der Beugungs- und Ansichtswinkel. Abbildung 7.7 visualisiert dies anhand eines Beispiels.

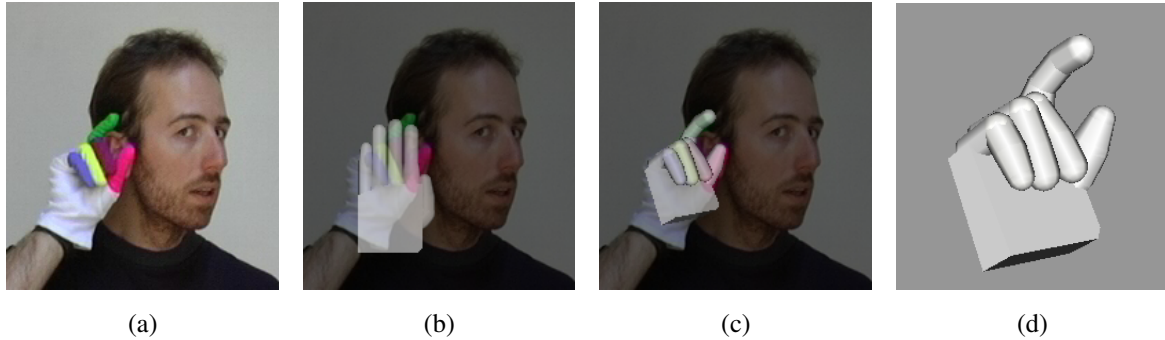


Abbildung 7.7: Eingabe der Positurparameter für ein Einzelbild der Gebärde „Hörgerät“. In das Originalbild (a) wird das Handmodell eingeblendet (b) und über die Tastatur so parametrisiert, dass es mit der realen Hand möglichst deckungsgleich ist (c). Die resultierende Positur ist in (d) vergrößert dargestellt.

Da die Handansicht hinsichtlich der Positurparameter nicht immer eindeutig ist und das Handmodell die Verformbarkeit der realen Hand nicht perfekt nachbilden kann, ergibt sich hinsichtlich der Parametrisierung ein Interpretationsspielraum. Verschiedene Bediener gelangen im Allgemeinen zu unterschiedlichen Ergebnissen, die aus der Kameraperspektive der Originalpositur entsprechen, bezüglich $\Delta_{\underline{P}}$ jedoch voneinander verschieden sind. In Abbildung 7.8 ist ein entsprechendes Beispiel zu sehen. Tabelle 7.4 enthält die Differenzen der beiden eingegebenen Posituren.

Bezüglich der Evaluation mit realen Posituransichten bedeutet dies zunächst, dass die manuell eingegebenen Positurparameter die tatsächliche Positur nicht exakt, sondern lediglich approximativ beschreiben. Die Präzision der Messung wird dadurch leicht herabgesetzt. Des Weiteren liegt die „Rekonstruktionsgenauigkeit“ eines menschlichen Beobachters nicht bei Null, sondern im Mittel bei $\bar{\Delta}_{\underline{B}} \approx 1 \text{ cm}$ und $\Delta_{\underline{\Phi}} \approx 15^\circ$ (dies wurde bereits in [AS02b, AS03b] beobachtet). Wird die Erkennungsleistung des Menschen als perfekt betrachtet, so folgt hieraus, dass das theoretische Optimum von $\Delta_{\underline{B}} = 0$ und $\Delta_{\underline{\Phi}} = 0$ aus einer monokularen Frontalperspektive weder erreichbar noch zur Gebärdenanalyse erforderlich ist.

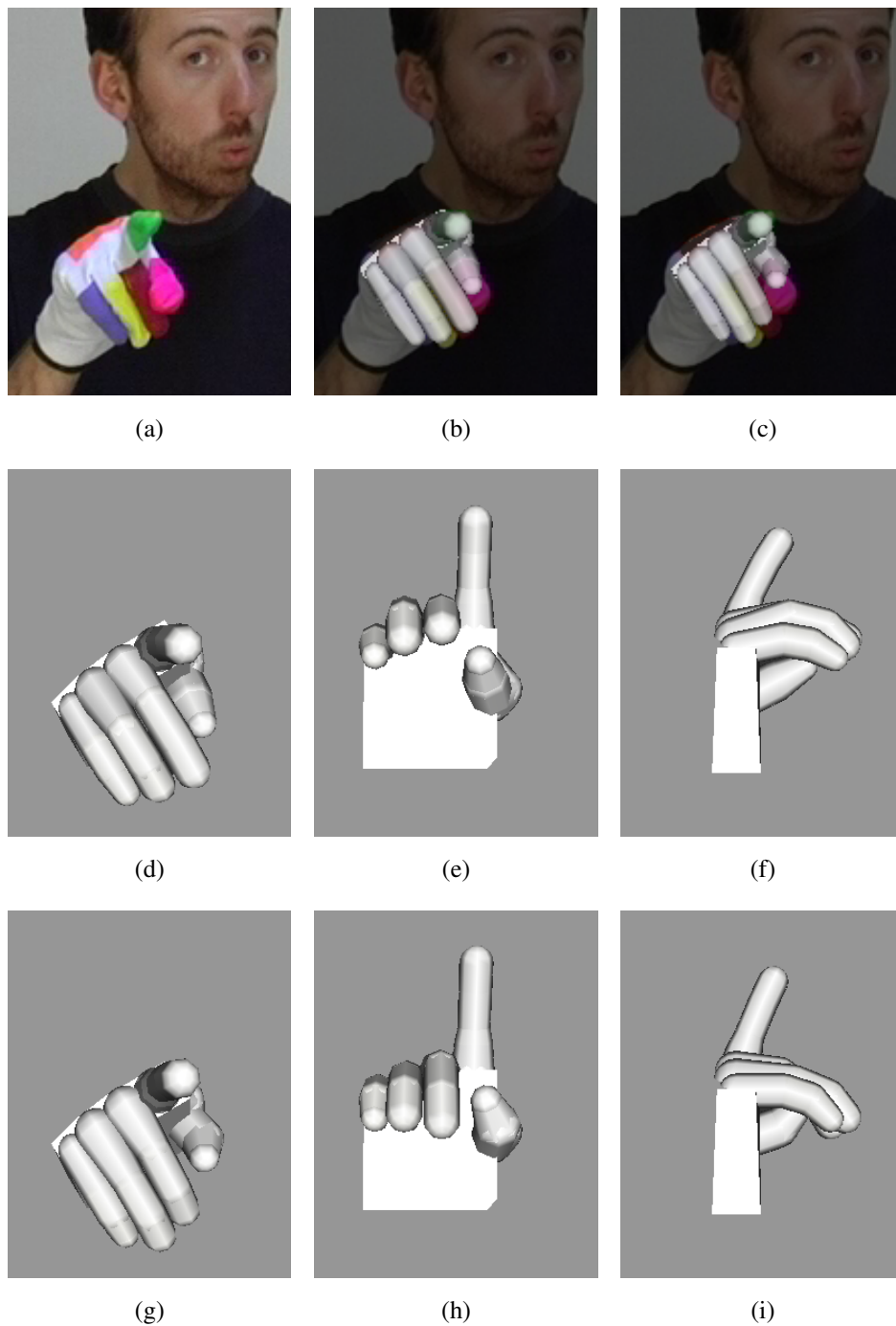


Abbildung 7.8: Varianz bei der manuellen Eingabe der Positurparameter. Für das Originalbild (a) haben zwei verschiedene Bediener jeweils unterschiedliche Modellparametrisierungen (b, c) eingegeben, die beide gut mit dem Eingangsbild übereinstimmen. In den beiden unteren Reihen sind die entsprechenden Posituren aus der tatsächlichen (d, g), der Palmar- (e, h) und einer Seitenansicht (f, i) gezeigt, wobei in den beiden letztgenannten Perspektiven die Verschiedenheit der Handform erkennbar ist.

Tabelle 7.4: Differenzmaße der Posituren in Abbildung 7.8, die von verschiedenen Bedienern für dasselbe Eingangsbild eingegebenen wurden. Die Abweichung der Handstellung ist vergleichsweise gering, während die Unterschiede der Handformen überdurchschnittlich ausgeprägt sind.

$\Delta_{\underline{\Phi}}$ [°]	$\Delta_{\underline{\mathbf{B}},\mathbf{D}}$	$\Delta_{\underline{\mathbf{B}},\mathbf{ZF}}$	$\Delta_{\underline{\mathbf{B}},\mathbf{MF}}$ [cm]	$\Delta_{\underline{\mathbf{B}},\mathbf{RF}}$	$\Delta_{\underline{\mathbf{B}},\mathbf{KF}}$	$\bar{\Delta}_{\underline{\mathbf{B}}}$ [cm]	$\Delta_{\underline{\mathbf{P}}}$ [cm ²]
1,98	0,49	0,42	3,19	2,03	1,27	1,48	16,33

7.2.3 Erkennungsleistung in Gebärden

Zur Evaluation der Rekonstruktionsgenauigkeit mit realen Posituranalysen dient ein Korpus von 20 DGS-Gebärden, für die die entsprechenden Ground-Truth-Daten wie im letzten Abschnitt beschrieben manuell erzeugt wurden. Das Testkorpus deckt zahlreiche verschiedene

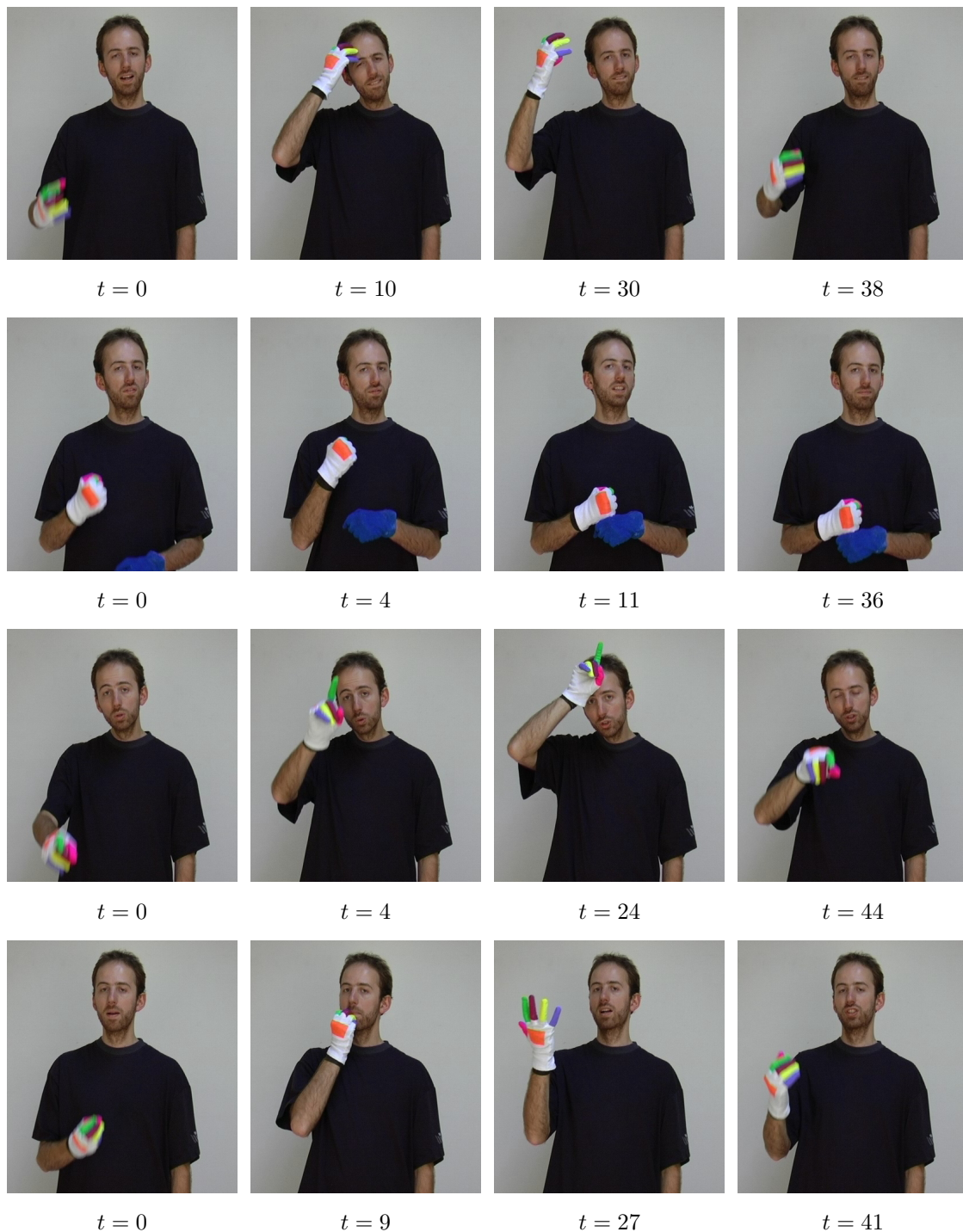


Abbildung 7.9: Die Gebärden „Hobby“, „fertig“, „deutsch“ und „praktisch“ (zeilenweise von oben nach unten) aus dem Testkorpus. Die Ausschnittaflösung beträgt 400×440 Pixel.

Handformen und -stellungen ab. In Abbildung 7.9 sind beispielhaft vier der Gebärden gezeigt. Die Bildauflösung beträgt 720×576 , wovon etwa 380×380 Pixel auf den Gebärdenraum und ca. 100×100 Pixel auf die Hand entfallen. Die Quantisierung der RGB-Farbwerte erfolgt mit jeweils 8 Bit pro Kanal.

Im Mittel beträgt die Gebärdendauer 2,79 Sekunden, wovon 1,50 Sekunden auf die tatsächliche Ausführungsphase entfallen. Vorbereitung und Rückführung benötigen jeweils ca. 0,65 Sekunden und gehen nicht in die Evaluation ein, da sich die Hand hier entweder nicht vollständig innerhalb des Bildbereichs befindet oder keine bedeutungsrelevanten Posituren annimmt, d. h. keine Informationen übermittelt. Somit stehen zur Genauigkeitsmessung insgesamt 750 Posituren zur Verfügung.

Die Parameter der Positurerkennungsalgorithmen entsprechen den unter Verwendung synthetischer Positurrensequenzen optimierten Werten aus Tabelle E.3. Für die Gesamtmenge aller Testposituren sind in Abbildung 7.10 Histogramme der Differenzmaße Δ_{Φ} , $\Delta_{\mathbf{B}}$ und $\Delta_{\mathbf{P}}$ visualisiert sowie deren Mittelwerte angegeben. Eine Aufschlüsselung nach den einzelnen Gebärdens findet sich in Tabelle 7.5.

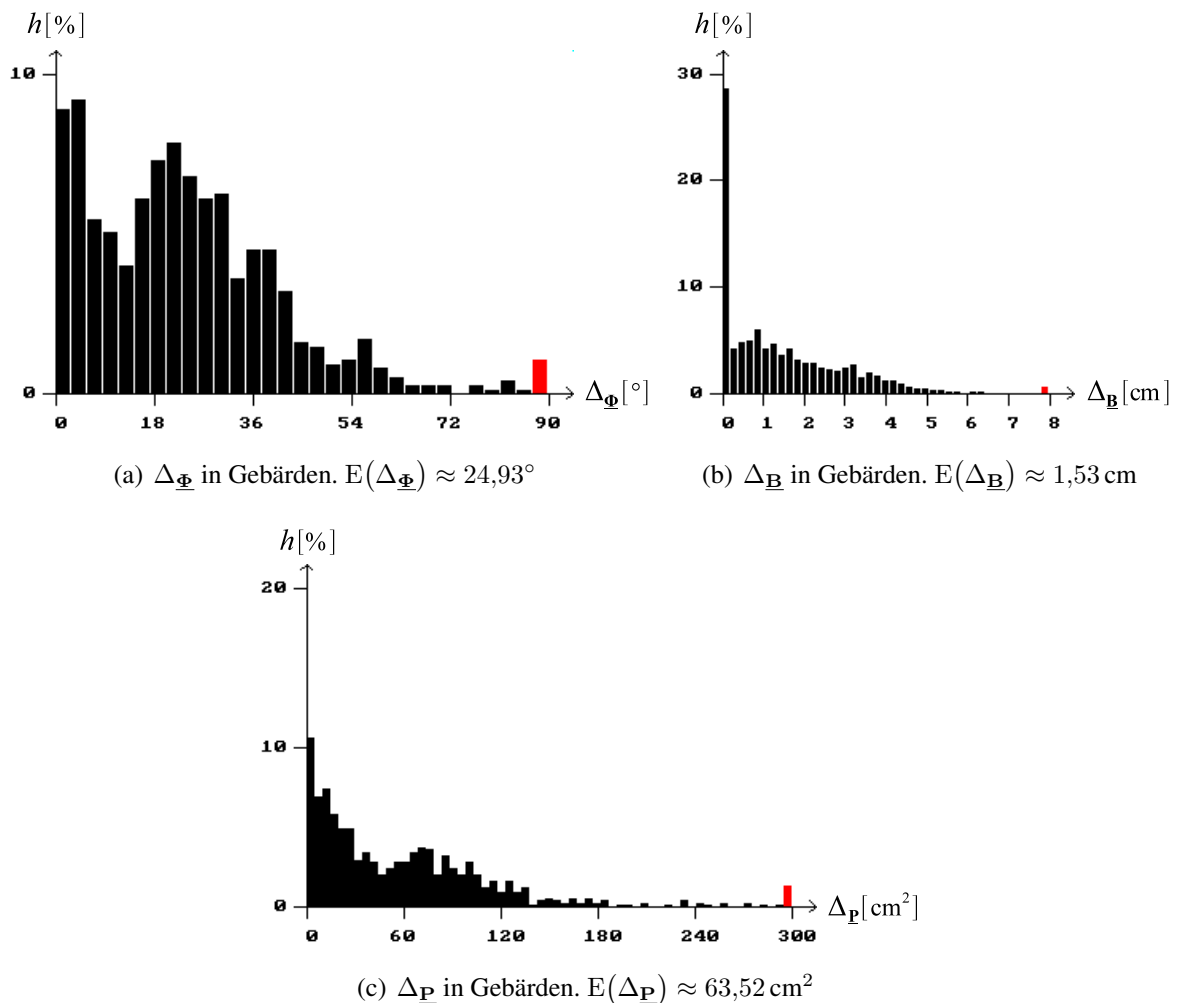


Abbildung 7.10: Rekonstruktionsgenauigkeit in einem Testkorpus von 20 Gebärdens.

Für die Handstellungsabweichung ergibt sich über alle Posituren ein Mittelwert von $E(\Delta_{\Phi}) \approx 24,93^\circ$. Hierbei liegt die Rotationsachse, um die die erkannte gegenüber der tatsächlichen Handstellung verdreht ist, vorwiegend in oder nahe der Bildebene, d. h. die Positur ist beispielsweise leicht nach hinten gekippt. Verdrehungen innerhalb der Bildebene sind hingegen wesentlich schwächer ausgeprägt. Ursache dieser ungleichmäßigen Verteilung ist die zur Aufnahme verwendete Frontalperspektive, in der sich Bewegungen entlang der z -Achse auf den Bildinhalt deutlich weniger auswirken als innerhalb der xy -Ebene.

Bei der Rekonstruktion der Handform erreicht das System eine mittleren Abweichung von $E(\Delta_{\underline{B}}) \approx 1,53$ cm. Anschaulich betrachtet sind die Fingerspitzen der rekonstruierten Positur somit eine knappe Fingerbreite von denen der Referenzpositur entfernt. Im Unterschied zur

Tabelle 7.5: Rekonstruktionsgenauigkeit in einem Testkorpus von 20 Gebärden. Angegeben sind die jeweils pro Gebärde berechneten Erwartungswerte, wobei der Operator $E(\cdot)$ in der Titelzeile aus Platzgründen nicht angegeben ist. Die Durchschnittswerte in der letzten Zeile unterscheiden sich geringfügig von denen in Abbildung 7.10, da sie nicht über alle Posituren, sondern über alle Gebärden gemittelt sind.

Gebärde	Δ_{Φ} [°]	$\Delta_{\underline{B},D}$	$\Delta_{\underline{B},ZF}$	$\Delta_{\underline{B},MF}$ [cm]	$\Delta_{\underline{B},RF}$	$\Delta_{\underline{B},KF}$	$\bar{\Delta}_{\underline{B}}$ [cm]	$\Delta_{\underline{P}}$ [cm ²]
„aber“	36,86	0,86	1,25	3,00	2,61	3,71	2,28	89,26
„Bildtelefon“	20,26	0,88	0,41	0,56	0,13	0,17	0,43	21,31
„bis“	58,44	1,22	0,73	2,26	2,73	2,25	1,84	174,29
„Brust“	9,30	1,17	1,05	0,74	0,58	0,78	0,86	22,45
„danken“	17,45	1,78	1,05	2,46	2,94	2,55	2,16	50,66
„deutsch“	33,33	1,10	1,04	2,64	2,11	1,72	1,72	64,97
„du“	13,15	0,22	0,93	0,90	1,18	0,89	0,82	12,71
„Durst“	14,72	0,90	1,70	0,96	1,82	0,80	1,23	24,52
„essen“	6,01	0,00	2,06	0,85	0,77	1,32	1,00	12,64
„Excel“	19,23	1,77	1,83	4,58	1,70	1,34	2,24	97,73
„fertig“	39,26	1,61	0,20	0,00	0,00	0,00	0,36	65,20
„gut“	25,53	1,15	0,74	1,52	1,58	2,90	1,58	52,81
„Hals“	16,57	1,03	2,44	0,19	0,00	0,00	0,73	28,57
„Hobby“	22,98	1,04	1,72	1,82	2,45	2,19	1,84	53,91
„Hörgerät“	39,07	2,42	2,62	2,27	1,69	1,76	2,15	98,75
„immer“	23,61	0,28	0,94	0,63	0,50	0,36	0,54	29,08
„Impfung“	22,40	1,24	2,42	3,77	3,75	2,74	2,78	75,40
„Jahr“	39,39	1,63	1,80	0,00	0,00	0,66	0,82	80,93
„praktisch“	7,67	0,29	0,67	0,69	0,44	0,95	0,61	14,51
„Punkt“	35,33	3,88	1,69	1,72	1,33	1,39	2,00	75,98
Durchschnitt	25,03	1,22	1,36	1,58	1,42	1,42	1,40	52,28

Handstellung ist dieser Fehler näherungsweise gleichmäßig im Raum verteilt. Eine getrennte Messung der Abweichung in und senkrecht zur Bildebene ergibt $E(\Delta_{\underline{B},xy}) \approx 1,16$ bzw. $E(\Delta_{\underline{B},z}) \approx 0,84$. In der monokularen Frontalansicht ist meist lediglich die erste Komponente erkennbar, während die Verschiebung der Fingerspitzen entlang der z -Achse nicht auffällt.

Tabelle 7.6 gibt die mittlere Rechenzeit der einzelnen Verarbeitungsstufen an. Primär fällt die Markersegmentierung ins Gewicht, die etwa zwei Drittel der Laufzeit ausmacht. Ausschlaggebend hierfür ist die Komplexität des Mean-Shift-Algorithmus in Verbindung mit der vergleichsweise große Bilddatenmenge von ca. 83 MB (29 Millionen Pixel) pro Gebärde. Die Positurerkennung selbst ist hingegen mit 10,8 Hz wesentlich schneller. Auf die Gesichtsdetektion und -segmentierung entfallen lediglich ca. 5% der Rechenzeit.

Tabelle 7.6: Geschwindigkeit der Verarbeitungsschritte zur Positurrekonstruktion.

	Gesichts- detektion	Marker- segmentierung	Positur- erkennung	Gesamt
Dauer pro Einzelbild [ms]	18,16	229,30	92,57	340,00
entspricht Bildrate [Hz]	55,06	4,36	10,80	2,94

7.2.4 Diskussion der Ergebnisse

Mit der im letzten Abschnitt gemessenen Erkennungsleistung ist das eingangs definierte Ziel, die manuellen Ausdrucksmittel der Gebärdensprache vollständig und ohne Einschränkung zu rekonstruieren, erreicht. Die Genauigkeit des Systems erlaubt eine sichere Prüfung der Anwendergebärde, und die verwendeten Algorithmen sind robust gegenüber verrauschten Eingangsdaten. Das Verfahren wird durch Verdeckungen einzelner Marker nicht negativ beeinflusst und funktioniert für alle anatomisch möglichen Handformen und -stellungen.

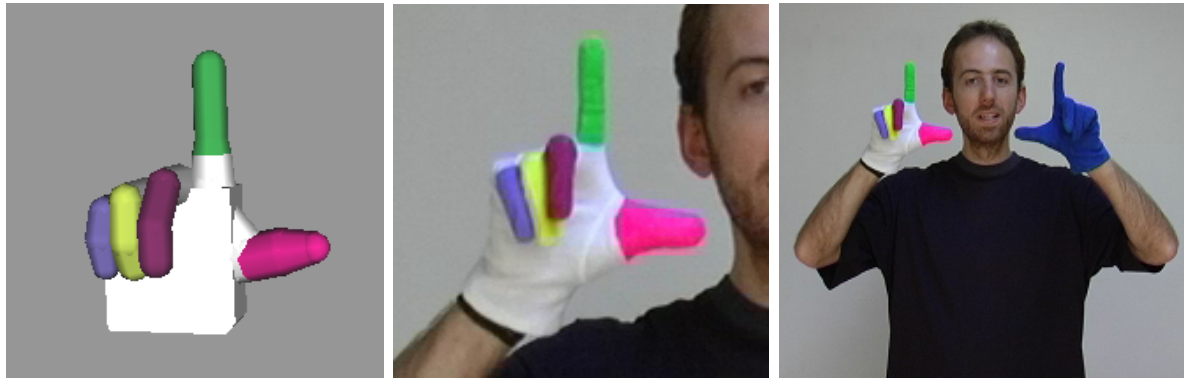
Die Evaluation mit synthetischen Testdaten hat sich als geeignet erwiesen, die einzelnen Verarbeitungsstufen zu testen und die Systemkonfiguration zu optimieren. Alle so ermittelten Parameter konnten für die Verarbeitung von Realansichten unverändert übernommen werden.

Hinsichtlich der Rechenzeit besteht noch Potenzial zur Verbesserung. Entsprechende Strategien sind in Abschnitt 8.2 skizziert.

Zur Veranschaulichung der Evaluationsergebnisse sind im Folgenden Beispiele für sehr gut sowie für durchschnittlich gut rekonstruierte Posituren gezeigt. Anschließend erfolgt eine Analyse vereinzelt auftretender Erkennungsfehler.

7.2.4.1 Beispiele sehr gut rekonstruierter Posituren

Abbildung 7.11 zeigt Beispielposituren aus den Gebärden „Bildtelefon“ und „Brust“, die mit sehr geringem Fehler rekonstruiert wurden. Die Abweichung der Fingerspitzenpositionen $\Delta_{\underline{B}}$ liegt im Bereich weniger Millimeter und ist daher visuell kaum wahrzunehmen.



„Bildtelefon“

$$\Delta_{\Phi} = 20,61^{\circ} \quad \bar{\Delta}_{\underline{B}} = 0,24 \text{ cm}$$

$$\Delta_{\underline{B},D} = 0,43 \text{ cm} \quad \Delta_{\underline{B},ZF} = 0,16 \text{ cm} \quad \Delta_{\underline{B},MF} = 0,20 \text{ cm} \quad \Delta_{\underline{B},RF} = 0,23 \text{ cm} \quad \Delta_{\underline{B},KF} = 0,16 \text{ cm}$$



„Brust“

$$\Delta_{\Phi} = 3,79^{\circ} \quad \bar{\Delta}_{\underline{B}} = 0,42 \text{ cm}$$

$$\Delta_{\underline{B},D} = 0,42 \text{ cm} \quad \Delta_{\underline{B},ZF} = 0,97 \text{ cm} \quad \Delta_{\underline{B},MF} = 0,37 \text{ cm} \quad \Delta_{\underline{B},RF} = 0,18 \text{ cm} \quad \Delta_{\underline{B},KF} = 0,16 \text{ cm}$$

Abbildung 7.11: Erkennungsbeispiele mit sehr niedrigem Rekonstruktionsfehler. Links ist jeweils das Erkennungsergebnis zu sehen, in der Mitte die Eingangspositur und rechts der vollständige Gebärdenraum. Unter den Abbildungen ist die Abweichung der rekonstruierten von der Referenzpositur angegeben.

7.2.4.2 Beispiele durchschnittlich gut rekonstruierter Posituren

Beispiele durchschnittlich gut rekonstruierter Posituren der Gebärden „du“ und „Durst“ sind in den Abbildungen 7.12 bzw. 7.13 zu sehen. Die Fehler liegen hier im Bereich des für das Testkorpus berechneten Mittelwerts und sind in den Bildausschnitten nicht ohne Weiteres auszumachen. Daher enthalten beide Abbildungen zusätzliche Ansichten der erkannten und der Referenzpositur, die mit Hilfe des Handmodells aus einer Perspektive gerendert wurden, die die Abweichungen besser erkennen lässt.

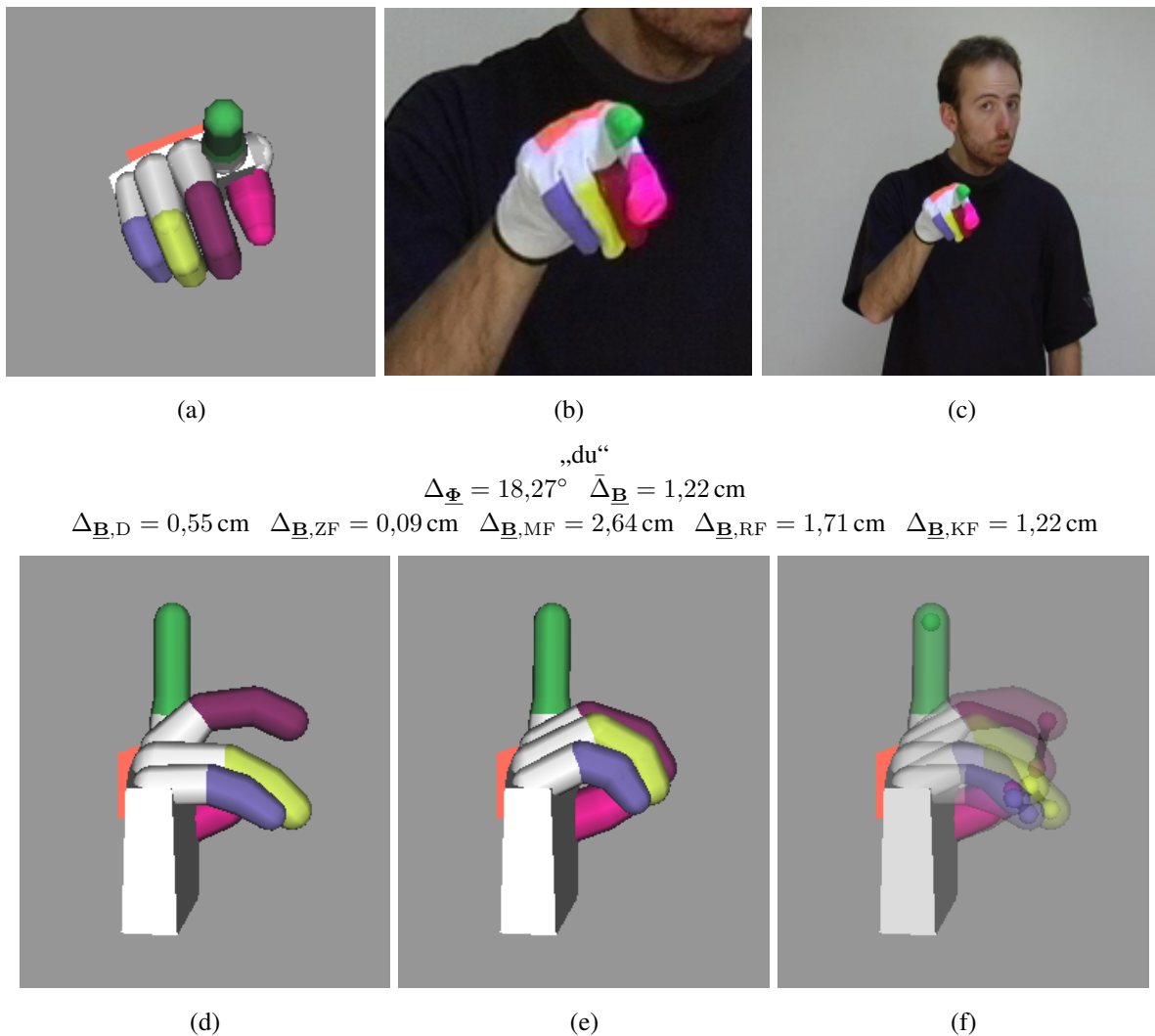


Abbildung 7.12: Beispielpositur aus der Gebärde „du“ mit durchschnittlichem Rekonstruktionsfehler. Die obere Reihe zeigt analog zu Abbildung 7.11 die erkannte (a) und die Referenzpositur (b) sowie den Gebärdenraum (c). Darunter sind beide Posituren unter Verwendung des Handmodells aus einer Perspektive betrachtet, die die Abweichung hinsichtlich der Handform deutlich macht (d bzw. e). In (f) ist die Handformdifferenz durch Überlagerung der Posituren visualisiert.

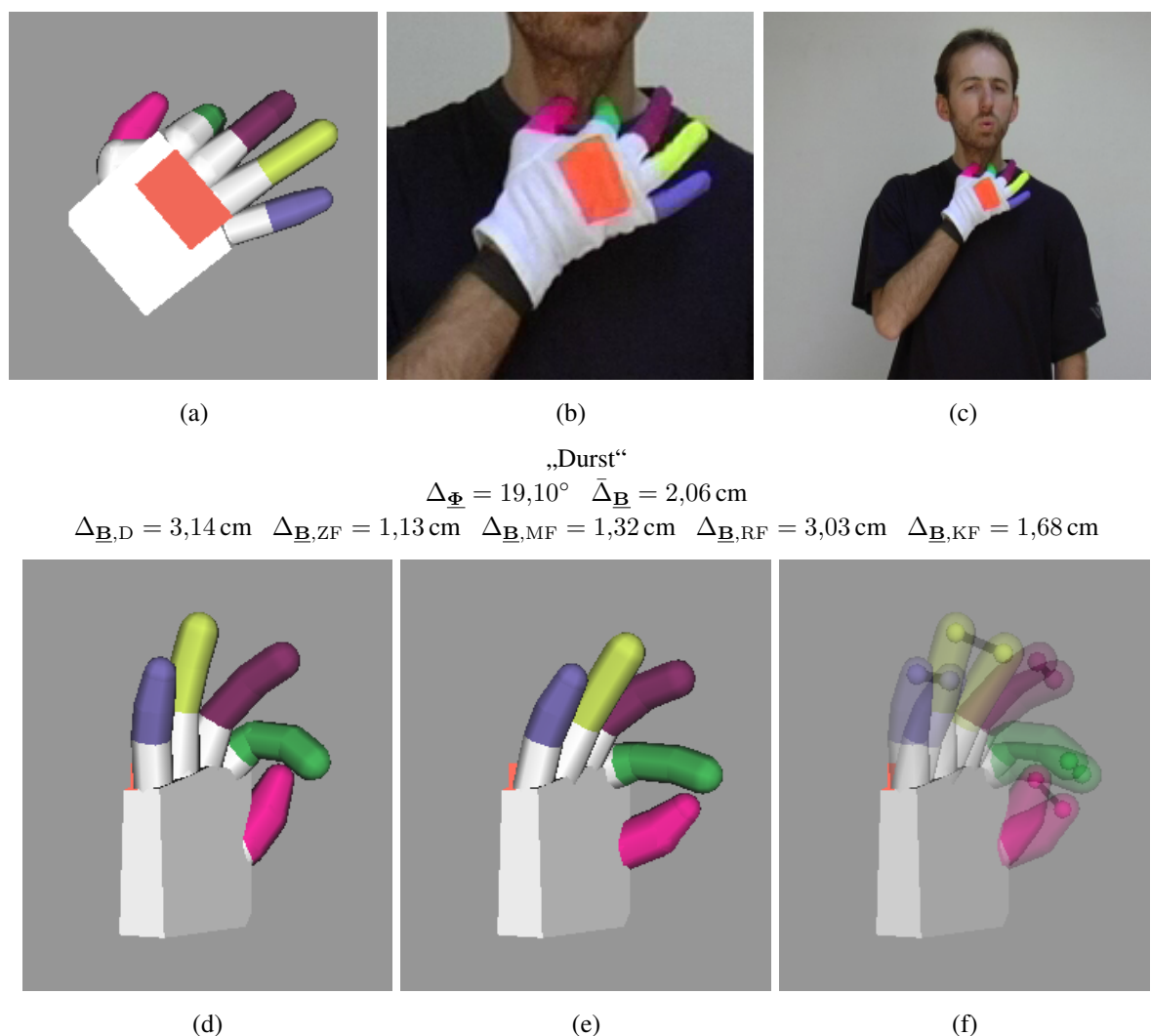


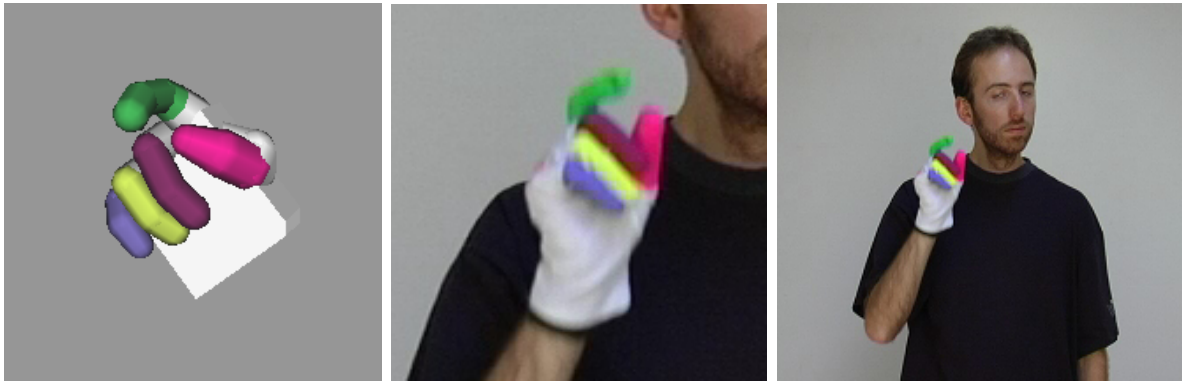
Abbildung 7.13: Beispielpositur aus der Gebärde „Durst“ mit durchschnittlichem Rekonstruktionsfehler (analog zu Abbildung 7.12).

7.2.4.3 Analyse fehlerhafter Rekonstruktionen

Für die überwiegende Mehrzahl der Posituren bewegt sich der Rekonstruktionsfehler im Bereich der oben genannten Mittelwerte. Vereinzelt treten jedoch deutliche Abweichungen auf, die in drei verschiedene Kategorien unterteilbar sind und nachfolgend erläutert werden.

Nicht auflösbare Mehrdeutigkeiten

Einzelne Gebärdenausführungen enthalten Mehrdeutigkeiten, die in der monokularen Frontalansicht auch unter Berücksichtigung des gesamten Gebärdenverlaufs nicht auflösbar sind. Im Testkorpus tritt dieser Fall bei der Gebärde „Hörgerät“ auf. Abbildung 7.14 zeigt ein Eingangsbild und das entsprechende Erkennungsergebnis.



„Hörgerät“

$$\Delta_{\Phi} = 87,88^{\circ} \quad \bar{\Delta}_{\mathbf{B}} = 2,04 \text{ cm}$$

$$\Delta_{\mathbf{B},D} = 4,40 \text{ cm} \quad \Delta_{\mathbf{B},ZF} = 1,76 \text{ cm} \quad \Delta_{\mathbf{B},MF} = 1,73 \text{ cm} \quad \Delta_{\mathbf{B},RF} = 1,67 \text{ cm} \quad \Delta_{\mathbf{B},KF} = 0,62 \text{ cm}$$

Abbildung 7.14: Fehlerhafte Rekonstruktion der Handstellung durch nicht auflösbare Mehrdeutigkeiten.

In Abbildung 7.15 sind die Kandidatenpositur, deren Merkmale sowie die Merkmale der Eingangspositur dargestellt. Zum Zeitpunkt der Kandidatenselektion hat die Positurenglättung (vgl. Abschnitt 5.3.2) noch nicht stattgefunden, so dass sich die Kandidatenpositur von dem in Abbildung 7.14 gezeigten Erkennungsergebnis leicht unterscheidet.

Trotz der stark abweichenden Handstellung ist die Merkmalsdifferenz zwischen der Kandidaten- und der Eingangspositur relativ gering. Da gleichzeitig sowohl für vorangehende als auch für nachfolgende Einzelbilder ähnliche Positurkandidaten existieren, die insgesamt eine flüssige Bewegung ergeben, wird die gezeigte Positur in der Mehrdeutigkeitsauflösung als die wahrscheinlichste bestimmt. Dies ist zwar in sich plausibel, weicht jedoch deutlich von der Referenzpositur ab.

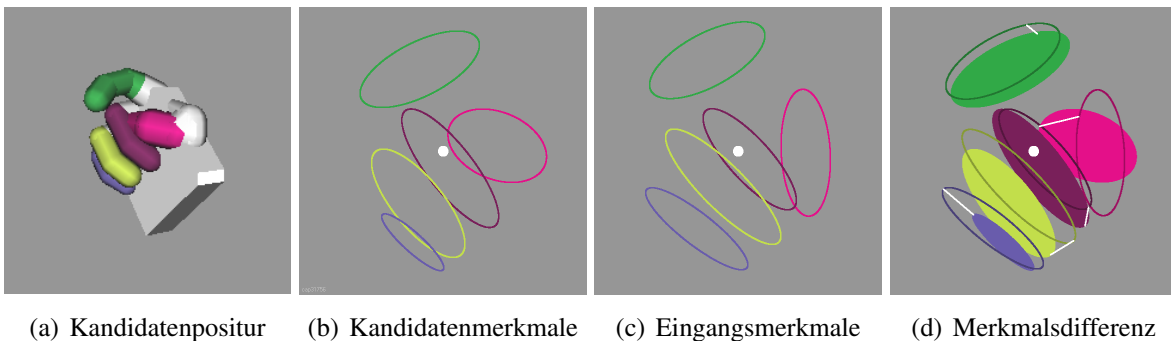


Abbildung 7.15: Mehrdeutige Merkmale in der Gebärde „Hörgerät“ (vgl. Abbildung 7.14). Durch die Projektion in die Bildebene ergibt sich für die Kandidatenpositur (a) ein Merkmalsvektor (b), der dem der Eingangspositur (c; vgl. Abbildung 7.14 Mitte)) sehr ähnlich ist, obwohl sich beide Posituren in der Handstellung stark unterscheiden.



„Hals“
 $\Delta_{\Phi} = 31,80^{\circ} \quad \bar{\Delta}_{\mathbf{B}} = 1,74 \text{ cm}$
 $\Delta_{\mathbf{B},D} = 0,46 \text{ cm} \quad \Delta_{\mathbf{B},ZF} = 2,22 \text{ cm} \quad \Delta_{\mathbf{B},MF} = 6,02 \text{ cm} \quad \Delta_{\mathbf{B},RF} = 0,00 \text{ cm} \quad \Delta_{\mathbf{B},KF} = 0,00 \text{ cm}$

Abbildung 7.16: Fehlerhafte Rekonstruktion der Handpositur durch Segmentierungsfehler.

Segmentierungsfehler

Obwohl das System robust gegenüber ungenauer Segmentierung ist, können sich wiederholt auftretende Fehler negativ auf die Rekonstruktion auswirken. Einzelne falsch erkannte Positionen werden im Allgemeinen durch die Positurrenglättung verworfen und aus Vorgängern und Nachfolgern korrekt interpoliert. Sind jedoch mehrere aufeinanderfolgende Merkmalsvektoren gestört, ist diese Korrektur aufgrund des begrenzten Stützstellenabstands der Spline-Interpolation nicht mehr möglich.

Ein Beispiel hierfür ist die Gebärde „Hals“. Da die Farbe des Daumenmarkers, abgeschattet durch den Kopf und gemischt mit der Farbe des Bartes des Gebärdenden, etwa der des Mittelfingermarkers entspricht, detektiert der Mean-Shift-Algorithmus beide Marker, obwohl tatsächlich nur der Daumenmarker sichtbar ist. Dies geschieht in mehreren aufeinanderfolgenden Einzelbildern. Abbildung 7.16 zeigt eine der infolge dieses Segmentierungsfehlers falsch



Abbildung 7.17: Fehlerhafte Segmentierung der Markerregionen in der Gebärde „Hals“ (vgl. Abbildung 7.16). Der Mean-Shift-Algorithmus detektiert irrtümlich den Mittelfingermarker am Rand des Daumenmarkers (Pfeil).

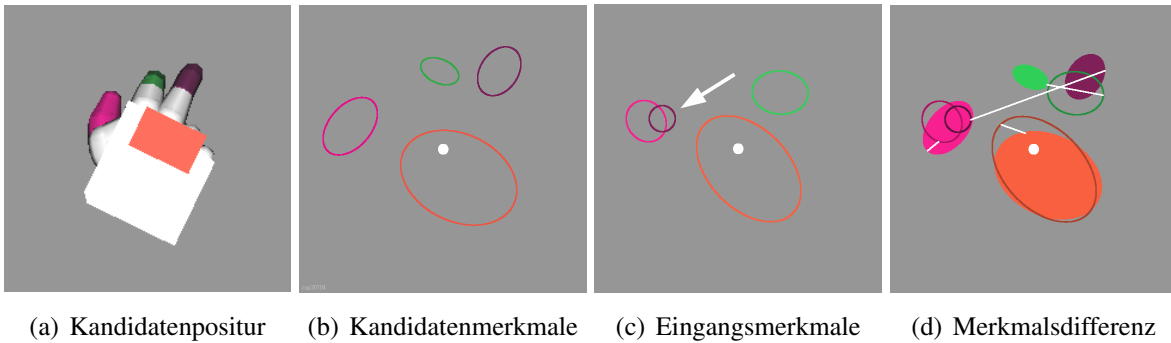


Abbildung 7.18: Kandidatenpositur (a) und -merkmale (b) in der Gebärde „Hals“ (vgl. Abbildung 7.16). Aufgrund des irrtümlich detektierten Mittelfingermarkers (c; Pfeil) ist der Mittelfinger der Kandidatenpositur ausgestreckt und sichtbar. Infolge dieses Fehlers weicht auch der im Rahmen des Merkmalsvergleichs (d) berechnete Rotationswinkel ϕ_{rot} (vgl. Abschnitt 5.2.1) um etwa 20° vom Optimalwert ab.

rekonstruierten Posituren.

Nicht modellierbare Posituren

Durch extreme Beugung oder Streckung der Gelenke in Fingern, Daumen und Handteller kann die menschliche Hand Posituren annehmen, die das Handmodell nicht nachzubilden vermag. Beispielsweise wird in der Gebärde „Impfung“ die rechte Hand gegen den linken Arm gedrückt. Diese Form kann das Modell nicht annehmen. Stattdessen ergeben sich für Mittel-, Ring- und kleinen Finger Gelenkstellungen, die zwar eine relativ niedrige Merkmalsdifferenz bewirken, jedoch keine bezüglich der Posituredifferenz optimale Parametrisierung



„Impfung“
 $\Delta_{\Phi} = 23,32^\circ \quad \bar{\Delta}_{\mathbf{B}} = 2,88 \text{ cm}$
 $\Delta_{\mathbf{B},D} = 0,73 \text{ cm} \quad \Delta_{\mathbf{B},ZF} = 0,14 \text{ cm} \quad \Delta_{\mathbf{B},MF} = 3,97 \text{ cm} \quad \Delta_{\mathbf{B},RF} = 6,30 \text{ cm} \quad \Delta_{\mathbf{B},KF} = 3,24 \text{ cm}$

Abbildung 7.19: Ungenaue Rekonstruktion bei nicht modellierbaren Posituren. Die Hand wird gegen den Arm gedrückt, so dass eine Form entsteht, die das Handmodell nicht annehmen kann.

des Modells darstellen. Ein entsprechendes Beispiel ist in Abbildung 7.19 zu sehen.

7.2.4.4 Vergleich mit existierenden Systemen

Ein direkter Vergleich der Erkennungsgenauigkeit mit den in Abschnitt 2.3 vorgestellten Arbeiten ist nicht möglich, da keines dieser Systeme die manuellen Ausdrucksmittel Form, Stellung und Position vollständig und uneingeschränkt erfasst. Durch die größtenteils stark restriktiven Randbedingungen unterscheidet sich daher bereits die Aufgabenstellung signifikant. Unabhängig davon finden sich quantitative Messungen nur in wenigen Veröffentlichungen. Diese werden im Folgenden kurz aufgeführt und der vorliegenden Arbeit – sofern möglich – gegenübergestellt.

Rosales et al. [RASS01] sowie **Shimada et al. [SKSK00, SKS01, ISS04]** geben keine Definition des verwendeten Fehlermaßes an. In der von **Athitsos et al. [AS03b, AS03a]** präsentierten Arbeit werden lediglich 26 diskrete Handformen unterschieden und das Klassifikationsergebnis entsprechend binär als richtig oder falsch bewertet. Die Genauigkeit bezüglich der Handstellung ist nicht separat evaluiert. In diesen drei Fällen stehen daher keine für einen Vergleich geeigneten Zahlenwerte zur Verfügung.

Das von **Stenger et al. [STTC06]** berechnete Differenzmaß erfasst lediglich Abweichungen in der Bildebene, nicht jedoch senkrecht hierzu. Aufgrund der monokularen Ansicht ist allerdings gerade bei letzteren der Hauptanteil des tatsächlichen Fehlers zu vermuten. Darüber hinaus erschwert die Angabe in Pixel eine anschauliche Interpretation der Werte.

Grundlage der Genauigkeitsmessung sind vier Realbildsequenzen, in denen ein mittlerer Fehler zwischen 6,6 und 9,3 Pixel ermittelt wird. Aus der Bildauflösung sowie dem relativen Anteil der Hand am Bild lässt sich grob eine Abweichung in der Größenordnung von 2 cm in der Bildebene schätzen. Für die gesamte Handpositur, d. h. Form und Stellung, stehen nur maximal drei Freiheitsgrade zur Verfügung, so dass ausschließlich die Beugung eines einzelnen Fingers oder die Änderung der Handstellung bei konstanter Handform erfasst werden kann.

Vittrup et al. [VSM02] messen die Differenz bzgl. Handform, -stellung und -position anhand der Entfernung der rekonstruierten sowie der tatsächlichen Gelenkpositionen im Raum. Zwar sind die Maße des Handmodells nicht spezifiziert, ein Mittelwert von 3,78 cm indiziert jedoch erhebliche Fehler und wird auch von den Autoren als nicht akzeptabel bewertet. Die Evaluation basiert auf lediglich sechs synthetischen Positurfolgen mit je 250 Einzelbildern, die nicht durch Rauschen degeneriert wurden.

Die von **Lathuilière et al. [LH99, LH00]** durchgeführte Evaluation verwendet nur eine unverrauschte synthetische Testsequenz mit 25 Einzelbildern. Die einzeln gemessenen Gelenkstellungsabweichungen betragen im Mittel ca. 5° , was übertragen auf die in Tabelle 7.3 angegebene Handmodellgeometrie etwa 1,5 cm entspricht. Allerdings stützt sich der auf inverser Kinematik beruhende Ansatz auf mehrere vereinfachende Annahmen hinsichtlich der Abhängigkeit der Gelenkstellungen, die die Anzahl der Freiheitsgrade des Handmodells stark reduzieren. Ein Nachweis der Gültigkeit dieser Annahmen in realen Posituransichten fehlt je-

doch, so dass die Übertragbarkeit des Ergebnisses auf die praktische Anwendung fragwürdig erscheint.

Der Ansatz erfordert die permanente Sichtbarkeit dreier Handrückenmarker und schränkt die Handstellung somit stark ein. Für die Anwendung im Bereich der Gebärdensprache ist er prinzipiell ungeeignet, da Palmar- und Seitenansichten nicht verarbeitet werden können. Die Genauigkeit der Handstellungsrekonstruktion ist nicht evaluiert und wird von den Autoren als Schwachpunkt des Ansatzes genannt.

Fillbrandt et al. [Fil03, FAK03] definieren einen skalaren Fingerbeugungsgrad mit dem Wertebereich $[0, 1]$ (vollständige Flexion bzw. vollständige Extension) und nennen eine visuell geschätzte mittlere Abweichung von 0,1. Diese Reduzierung auf einen Freiheitsgrad stellt besonders für den Daumen eine erhebliche Vereinfachung dar. Auch Abduktion und Adduktion lassen sich hiermit kaum erfassen.

Durch die Addition der einzelnen Beugungswinkel der Gelenke mit zum Ende der kinematischen Kette – d. h. zum DIP- bzw. IP-Gelenk – hin abnehmender Gewichtung und lineare Abbildung auf das Intervall $[0, 1]$ lässt sich auch für das in Kapitel 3 beschriebene Handmodell ein solcher skalarer Beugungsgrad β für jeden Finger sowie den Daumen berechnen. Eine entsprechende Bewertung der Rekonstruktionsgenauigkeit ist in Abbildung 7.20 zu sehen. Die gemessene mittlere Abweichung von $\Delta_\beta \approx 0,080$ liegt knapp, der Median von 0,045 deutlich unter 0,1.

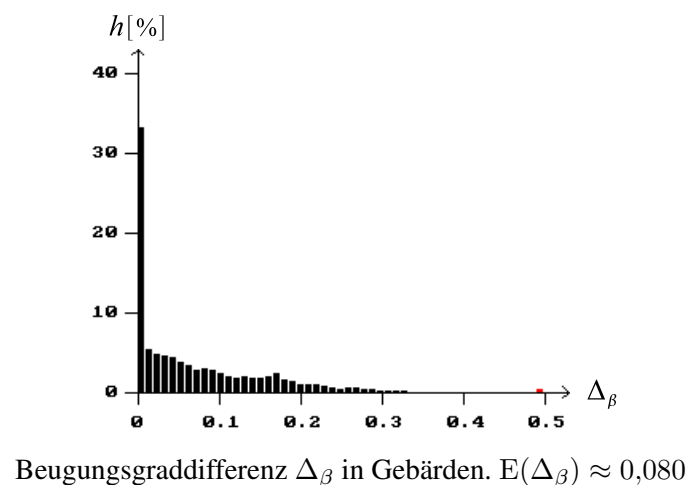


Abbildung 7.20: Angabe der in Abbildung 7.10c visualisierten Handformabweichung in Form einer Beugungsgraddifferenz Δ_β .

Somit erreichen Fillbrandt et al. zwar annähernd dieselbe Genauigkeit ohne Verwendung optischer Markierungen, allerdings sind nur 15 verschiedene Handformen sowie die Übergänge zwischen diesen erkennbar. Das Verfahren erfordert eine Initialisierung durch Einnehmen einer vorgegebenen Positur und erlaubt keine Behandlung von Mehrdeutigkeiten oder Erkennungsfehlern, etwa durch multiple Hypothesen.

Um die angegebene Abweichung zu erreichen, muss das System ferner vom Benutzer manuell trainiert werden. Dieser Vorgang entspricht prinzipiell der in Abschnitt 7.2.2 beschriebenen

Eingabe der Referenzdaten, so dass in Anbetracht der 60 verwendeten Ansichtenmodelle eine Gesamtdauer von einigen hundert Stunden anzunehmen ist.

Mit $10\text{--}20^\circ$ ist der visuell geschätzte mittlere Handstellungsfehler vergleichsweise niedrig, jedoch sind nur Ansichten aus der Nähe des „Äquators“ der Ansichtskugel ($\phi_{\text{lon}} = 0^\circ$) zugelassen und die Varianz in den Testdaten somit entsprechend gering. Ansichten aus Richtung der Finger, die in der Gebärdensprache häufig auftreten, werden nicht unterstützt. Für diese sind die verwendeten Kontur- und Texturmerkmale im Gegensatz zu optischen Markern wenig diskriminant.

Zusammenfassend ergibt sich für das in der vorliegenden Arbeit entwickelte Verfahren zur Positurerkennung die höchste Leistung. Gleichzeitig erfasst das System als einziges alle manuellen Ausdrucksmittel der Gebärdensprache ohne jede Einschränkung. Die Evaluation der Rekonstruktionsgenauigkeit basiert auf einer erheblich größeren Datenmenge und ist aufgrund der systematischen Durchführung sowie der Verwendung anschaulich nachvollziehbarer Differenzmaße wesentlich aussagekräftiger.

Kapitel 8

Zusammenfassung und Ausblick

Die Muttersprache gehörloser Menschen ist die Gebärdensprache, eine vollwertige, visuelle Sprache zur Übermittlung beliebiger konkreter sowie abstrakter Informationen. Schriftsprache, die auf der Codierung von Phonemen durch Buchstaben basiert, ist Gehörlosen nur eingeschränkt zugänglich. Zur Förderung der sozialen Integration besteht daher der Wunsch, Hörenden eine grundlegende Gebärdensprachkompetenz zu vermitteln, so dass einfache Gespräche im Alltag auch ohne Dolmetscher möglich sind.

Es existieren zahlreiche multimediale Lernprogramme zum Selbststudium der DGS. Durch Verfahren zur videobasierten Gebärdenspracherkennung lässt sich ein interaktiver Lernprozess realisieren, indem die Gebärdenausführung des Anwenders durch eine Kamera erfasst und vom Rechner kontrolliert sowie ggf. verbessert wird. Damit ist neben einer quantitativen Leistungsbeurteilung auch eine automatische Anpassung der Lerninhalte möglich.

Zentrale Komponente eines solchen Tutorsystems ist die Erfassung der manuellen Ausdrucksmittel der Gebärdensprache: Handform, -stellung, -position und -bewegung. Die vollständige und uneingeschränkte Rekonstruktion dieser Parameter eröffnet weitere Anwendungen im Bereich der virtuellen Realität und der Robotik.

8.1 Zusammenfassung

Gegenstand der vorliegenden Arbeit ist die Entwicklung eines interaktiven Gebärdensprach-tutors, der – ähnlich einem menschlichen Lehrer – die Gebärdenausführung des Lernenden prüft und bei Fehlern oder Abweichungen entsprechende Korrekturvorschläge anbietet.

Um eine zuverlässige Erkennung von Handform und -stellung zu ermöglichen, trägt der Gebärdende farbig markierte Stoffhandschuhe, die die Bewegung der Hand nicht einschränken. Ausgehend von einer monokularen Frontalansicht in unkontrollierter Umgebung erfolgt zunächst eine Modellierung des statischen Bildhintergrunds, um diesen von der weiteren Verarbeitung auszuschließen und dadurch Robustheit und Geschwindigkeit zu erhöhen. Eine anschließende Mean-Shift-Segmentierung liefert eine Menge von Regionen, die unter Verwen-

dung von Domänenwissen als Störobjekte oder Marker klassifiziert werden. Dabei verfolgt das System mit Hilfe des Viterbi-Algorithmus parallel mehrere Hypothesen und trifft eine endgültige Entscheidung erst nach Ende der Gebärde und Auswertung aller verfügbaren Informationen.

Die zweidimensionalen geometrischen Merkmale der resultierenden Markerregionen bilden die Grundlage für die Rekonstruktion der dreidimensionalen Ausdrucksmittel Handform und Handstellung, die unter dem Begriff der Handpositur zusammengefasst werden. Dieser Schritt bildet den Schwerpunkt der Arbeit. Auf Basis einer mathematisch präzisen Problembeschreibung entsteht ein neuartiges Konzept, das die prinzipielle Mehrdeutigkeit der zweidimensionalen Eingangsdaten explizit behandelt und durch Ausnutzen der zeitlichen Korrelation im Gebärdenverlauf aufzulösen vermag.

Kernstück dieses Ansatzes ist ein 3D-Handmodell, das über 20 Freiheitsgrade verfügt und die funktionale Anatomie der menschlichen Hand einschließlich der Abhängigkeiten zwischen verschiedenen Gelenken akkurat nachbildet. Das Handmodell weist dieselben Markierungen wie der vom Benutzer getragene Stoffhandschuh auf, so dass derselbe Merkmalsvektor berechnet werden kann. Die Erkennungsaufgabe besteht damit in der Bestimmung der Gelenkstellungen und des Ansichtswinkels des Modells derart, dass beide Merkmalsvektoren einander möglichst ähnlich sind. Hierfür wurde ein auf der Hausdorff-Distanz beruhendes aussagekräftiges Differenzmaß entwickelt, das die Verschiedenheit zweier Merkmalsvektoren skalar quantifiziert.

Vor der Nutzung generiert das System mit Hilfe des Handmodells eine Datenbank, die etwa 100 Millionen verschiedene Handposituren sowie die entsprechenden Merkmalsvektoren umfasst und alle anatomisch möglichen Posituren abdeckt. Die Suche in dieser Datenbank erfolgt für jedes der T Einzelbilder einer Gebärde zweistufig: Unter Verwendung eines balancierten Indexbaums wird zunächst eine Vorselektion von Positürkandidaten erstellt und anschließend auf Grundlage der Merkmalsdifferenz verfeinert. Pro Einzelbild resultieren so $k > 1$ Posituren, aus denen sich k^T mögliche Gebärdenverläufe ergeben. Dieser Hypothesenraum bildet die Mehrdeutigkeit der Eingangsdaten ab.

Zur Bestimmung des tatsächlichen Gebärdenverlaufs wird der Hypothesenraum nach einer Positurenfolge durchsucht, die sowohl eine möglichst niedrige kumulative Merkmalsdifferenz aufweist, als auch eine flüssige und damit plausible Bewegung darstellt. Die Auswertung des letztgenannten Kriteriums basiert auf einem weiteren Differenzmaß, das zwei Handposituren mittels eines anschaulichen geometrischen Verfahrens vergleicht und gut mit der Bewertung eines menschlichen Beobachters korreliert. Durch erneuten Einsatz des Viterbi-Algorithmus reduziert sich die Komplexität dieser Suche auf $O(Tk^2)$.

Da die Datenbank eine Diskretisierung des Positurenraums darstellt, ist die so ermittelte Positurenfolge zunächst noch mit Quantisierungsrauschen überlagert. Eine Spline-Interpolation der Handmodellparameter kompensiert diesen Effekt und optimiert gleichzeitig das Erkennungsergebnis, indem als Stützstellen Zeitpunkte mit minimaler Merkmalsdifferenz selektiert werden. Isolierte Rekonstruktionsfehler in einzelnen Eingangsbildern wirken sich daher kaum

auf das Endergebnis aus. Des Weiteren ist der Wertebereich der Ausgabe nicht auf die endliche Menge der Datenbankposituren beschränkt, sondern umfasst den gesamten Positorenraum.

Um die Gebärde des Lernenden mit der Referenzausführung zu vergleichen, findet zunächst eine zeitliche Anpassung mittels Dynamic Time Warping (DTW) statt. Anschließend prüft das Tutorsystem die manuellen Ausdrucksmittel Handposition, Bewegungsgeschwindigkeit, Handstellung und Handform. Dieser Prozess berücksichtigt die verschiedenen zeitlichen Gebärdenphasen durch entsprechende Gewichtung der jeweils relevanten Komponenten. Die Bewertung stützt sich damit auf linguistische Kriterien und ist deshalb für den Anwender nachvollziehbar.

Im Falle einer Abweichung erfolgt eine textuelle Rückmeldung sowie eine visuelle Korrektur der Positur. Hierzu dient erneut das Handmodell, das sich von der fehlerhaften in die Referenzpositur bewegt und dadurch die Verschiedenheit verdeutlicht. Dieses anschauliche Feedback ist mit dem eines menschlichen Lehrers vergleichbar und daher intuitiv verständlich.

Die Leistung der Handpositurerkennung wurde in einer umfassenden Evaluation quantitativ gemessen. Durch die Verwendung des Handmodells zur automatischen Generierung umfangreicher Testdaten ist eine systematische Analyse einzelner Systemkomponenten, Designparameter und Algorithmen möglich, die in der Literatur bisher nicht beschrieben ist. So ergibt sich etwa durch simuliertes Bildrauschen der Nachweis der Robustheit, und die Geometrie der optischen Marker lässt sich bereits vor der Erstellung der Gebärdenaufnahmen optimieren. Als weiteres Novum wurden die Positurparameter für 20 Gebärden manuell eingegeben und das Ergebnis der automatischen Positurerkennung mit den so erhaltenen Ground-Truth-Daten verglichen. Der mittlere Rekonstruktionsfehler bezüglich der Position der Daumen- bzw. Fingerspitzen im Raum beträgt lediglich 1,53 cm, die Abweichung hinsichtlich der Handstellung $24,93^\circ$. Damit liegt die erreichte Genauigkeit nur wenig unter der eines menschlichen Beobachters und erlaubt eine sichere Prüfung der Anwendergebärde.

Im Vergleich mit weiteren Systemen zur Handpositurrekonstruktion, die zu Beginn dieser Arbeit ausführlich besprochen und eingeordnet werden, ist das vorgestellte Verfahren für die Anwendung in einem Gebärdensprachtutor am besten geeignet. Es erlaubt als einziges eine Erkennung aus beliebigem Ansichtswinkel, schränkt Form und Stellung der Hand in keiner Weise ein und extrahiert in vollem Umfang alle manuellen Ausdrucksmittel der Gebärdensprache. Die durch die monokulare Ansicht bedingte Mehrdeutigkeit der Eingangsdaten ist im Systemkonzept methodisch berücksichtigt. Das parallele Verfolgen multipler Hypothesen sowohl auf Merkmals- als auch auf Positurebene ermöglicht in jedem Einzelbild die Auswertung sämtlicher vorheriger sowie nachfolgender Eingangsdaten und stellt dadurch Robustheit und Fehlertoleranz der Rekonstruktionsstufe sicher.

Der Einsatz videobasierter Verfahren zur Handpositurerkennung im Kontext eines interaktiven Gebärdensprachtutors stellt eine innovative Anwendung dar, die bislang in keiner Veröffentlichung diskutiert ist. Die Rekonstruktion aller manuellen Ausdrucksmittel ohne Beschränkung hinsichtlich Handform und -stellung führt zu hohen Anforderungen, die das hier entwickelte Verfahren vollständig erfüllt.

8.2 Ausblick

Im Folgenden werden verschiedene Ansatzpunkte zur Weiterentwicklung des Gebärdensprachtutors sowie speziell der Handpositurerkennung kurz diskutiert. Die Reihenfolge der Aufzählung orientiert sich dabei an der Verarbeitungskette des Systems.

Die Schwierigkeit der Positurrekonstruktion ist in erster Linie durch die vergleichsweise geringe Aussagekraft der zweidimensionalen Ansicht hinsichtlich der dreidimensionalen Handkonfiguration bedingt. Eine Steigerung der Erkennungsgenauigkeit ist daher von einer Erweiterung auf zwei oder mehr Kameras zu erwarten (vgl. Abschnitt 2.5.5). Insbesondere die Kombination von Frontal- und Seitenansicht reduziert die Mehrdeutigkeit der Eingangsdaten erheblich und wäre diesbezüglich der stereoskopischen Perspektive des Menschen überlegen, schränkt allerdings den Bewegungsraum der Hand auf die Schnittmenge der beiden Kamerasisichtfelder ein.

Das Handmodell berücksichtigt derzeit nicht die Verformbarkeit des Handtellers. Durch Hinzufügen eines oder zweier Freiheitsgrade ist in Einzelfällen eine bessere Übereinstimmung mit der Eingangspositur erreichbar. Um die Größe der Posituredatenbank nicht unangemessen zu erhöhen, sollte für die hinzugefügten Gelenke neben anatomischen Abhängigkeiten jedoch auch die Relevanz der Gelenkstellung für die Interpretation der Handform berücksichtigt und zur Datenreduzierung ausgenutzt werden.

Beim Einsatz der Positurrekonstruktion für fehlerfreie Gebärden sind verschiedene linguistisch motivierte Optimierungen denkbar. Durch eine Beschränkung auf die in der jeweiligen Gebärdensprache verwendeten Handposituren – einschließlich der Übergänge zwischen diesen – lassen sich große Teile des Positurenraums ausblenden und folglich die gesamte Erkennungsaufgabe deutlich vereinfachen. Anwendung fände ein solches Erkennungssystem etwa bei der Erfassung von Gebärdeninhalten zur Reproduktion durch einen Avatar.

Zur Erhöhung der Verarbeitungsgeschwindigkeit ist die Parallelisierung einzelner Verarbeitungsstufen nahe liegend. Besonders eignen sich hierfür die Segmentierung der Eingangsbilder, die gegenwärtig den Hauptanteil der Rechenzeit beansprucht, sowie die Abfrage der Posituredatenbank. Beide Schritte operieren jeweils auf einzelnen Bildern bzw. Merkmalsvektoren und lassen sich daher problemlos auf mehrere Rechner verteilen.

Mittels eines Datenhandschuhs, der äußerlich dem optisch markierten Stoffhandschuh entspricht, ist die Erzeugung von Ground-Truth-Daten deutlich schneller und mit höherer Präzision möglich als über die Methode der manuellen Positureingabe. Bei der Evaluation der Rekonstruktionsgenauigkeit in Realansichten ist dies von Vorteil, da mit geringem Aufwand vergleichsweise große Mengen von Testdaten generierbar sind und sich die Aussagekraft der Messung entsprechend erhöht.

Neben der Genauigkeit der Positurrekonstruktion sind auch Gestaltung und Effektivität des interaktiven Feedbacks für eine Evaluation von Interesse. Hierfür wären entsprechend umfangreiche Untersuchungen an Versuchspersonen durchzuführen und der Lernfortschritt mit einer Kontrollgruppe zu vergleichen, die mit nicht interaktiven Medien arbeitet.

Schließlich ergeben sich für die Handpositurerkennung zahlreiche Einsatzmöglichkeiten außerhalb der Gebärdensprache. Die berührungslose visuelle Messung von Gelenkstellungen und Ansichtswinkel erlaubt die Realisierung intuitiver Benutzungsschnittstellen beispielsweise in VR- und CAD-Systemen oder zur Steuerung von Robotern. Das entwickelte Verfahren ist aufgrund seiner Flexibilität hinsichtlich Markergeometrie und Kameraperspektive sowie der vollständigen Abdeckung des Positurenraums für diese Anwendungen sehr gut geeignet.

Anhang A

Algorithmus zur multimodalen Hintergrundmodellierung

Die Schätzung der Verteilungsparameter $\underline{\mu}$ und Σ sowie der Gewichte ω und der Anzahl der Komponenten K in (4.26) kann mit Hilfe des EM-Algorithmus (*Expectation Maximization*) geschehen, der für eine gegebene Menge von Farbwerten $\mathcal{C} = \{\underline{C}_1, \underline{C}_2, \dots\}$ die Wahrscheinlichkeit $\prod_i P(\underline{C}_i)$ maximiert. In Verbindung mit einer noch zu leistenden Unterteilung Vorder- und Hintergrundkomponenten ist damit eine Hintergrundsubtraktion möglich, indem für eine gegebene Farbe $\underline{C}$ die wahrscheinlichste Verteilungskomponente ermittelt und im Hintergrundfall das Pixel ausgeblendet wird.

Für eine praktische Anwendung erweist sich die hohe Laufzeitkomplexität des exakten EM-Algorithmus als nachteilig. Daher kommt eine in [SG99] vorgestellte und in [KB03] erweiterte schnelle Approximation zum Einsatz, die ähnlich einem Kalman-Filter beobachtete Farbwerte nicht explizit, sondern in Form eines Systemzustands speichert. Pro Pixel werden lediglich die Verteilungsparameter $\underline{\mu}_k$, Σ_k und ω_k sowie K verwaltet und mit jedem Bild aktualisiert. Abbildung A.1 zeigt ein Ablaufdiagramm dieses Algorithmus. Anschließend sind die grundlegenden Annahmen sowie die Funktionsweise beschrieben.

Annahmen Zunächst erfolgt hinsichtlich der Kovarianzmatrix Σ dieselbe Vereinfachung wie im unimodalen Fall (vgl. (4.16)). Allerdings wird Σ nicht, wie in (4.22), für alle Pixel gleich angenommen, sondern für jede einzelne Modellkomponente separat berechnet.

Die K Gaußkomponenten einer Verteilung sind absteigend nach dem als *fitness* bezeichneten Quotienten ω_k/σ_k sortiert, d. h.

$$\frac{\omega_k}{\sigma_k} \geq \frac{\omega_{k+1}}{\sigma_{k+1}} \quad (\text{A.1})$$

Hintergrundkomponenten weisen im Vergleich zum Vordergrund eine geringere Varianz sowie ein höheres Gewicht auf und belegen in dieser Sortierung daher die vorderen Positionen.

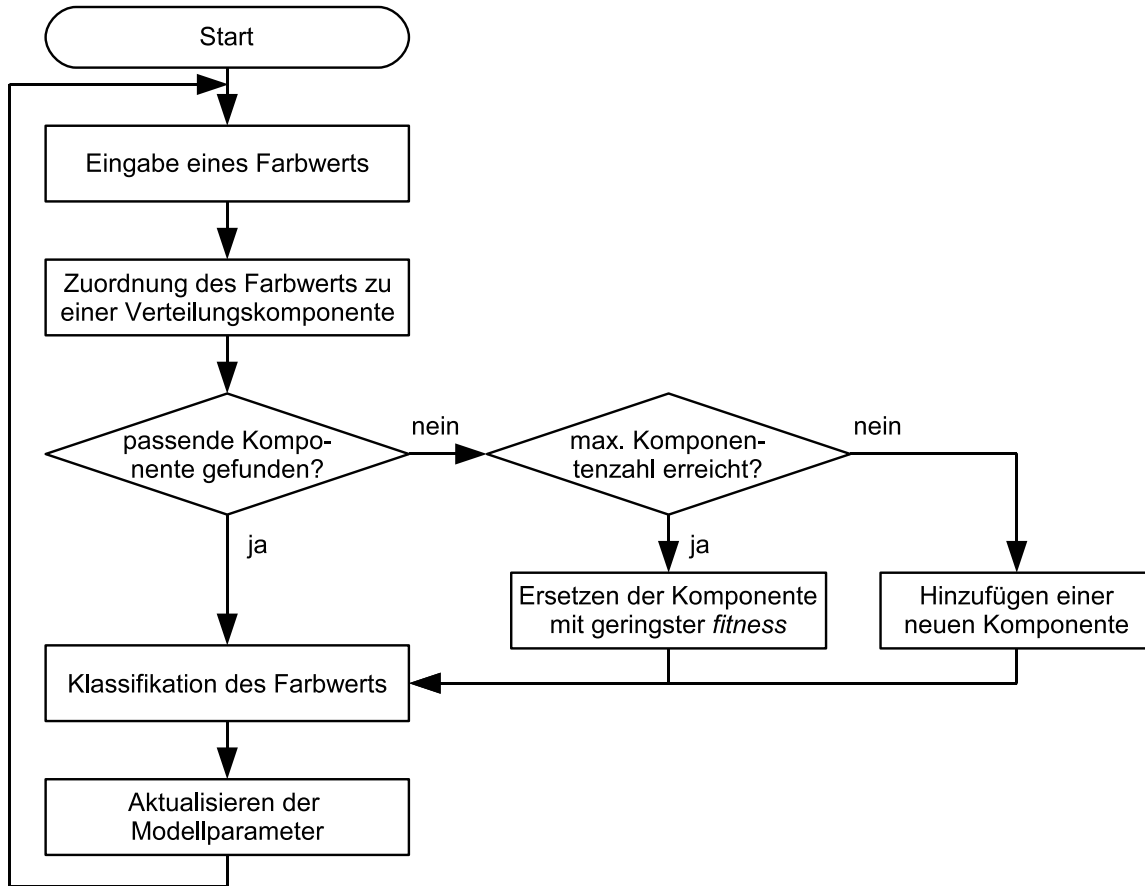


Abbildung A.1: Algorithmus zur Schätzung der Verteilungsparameter eines multimodalen Hintergrundmodells.

Zuordnung eines Farbwerts zu einer Verteilungskomponente Mit jedem Einzelbild erhält der Algorithmus für jedes Pixel einen neuen Farbwert $\underline{C}$, der in das Modell eingeht. Zunächst wird geprüft, ob dieser Wert einer der bestehenden Verteilungskomponenten zuzuordnen ist. Dazu durchsucht der Algorithmus die Komponenten in der oben beschriebenen Reihenfolge absteigender Fitness-Werte und selektiert die erste Komponente, die die Bedingung

$$\left\| \underline{C} - \underline{\mu}_k \right\| < 2,5\sigma_k \quad (\text{A.2})$$

erfüllt. Trifft (A.2) auf keine der Komponenten zu (etwa zu Beginn der Bildsequenz, wenn noch keine Komponenten vorhanden sind), so wird eine neue Komponente mit $\underline{\mu} = \underline{C}$, einem geringen Anfangsgewicht ω und einer großen Varianz σ hinzugefügt. Falls die maximale Anzahl von Komponenten erreicht ist, ersetzt die neue Komponente diejenige mit dem geringsten Fitness-Wert.

Im Folgenden gibt k' den Index der zugeordneten bzw. neu erzeugten Komponente in der gemäß (A.1) sortierten Liste an.

Klassifikation Die Klassifikation des Farbwerts beruht auf der im letzten Schritt zugeordneten bzw. erzeugten Verteilungskomponente. Der Farbwert gilt als Hintergrund, falls

$$\sum_{k=1}^{k'-1} \omega_k < \Theta \quad (\text{A.3})$$

mit dem Schwellwert $\Theta \in (0, 1)$ als globalem Parameter. Das Hintergrundmodell besteht somit aus einer variablen Anzahl von Verteilungskomponenten, deren Gewichtssumme sich in der Nähe des Schwellwerts befindet.

Update der Modellparameter Falls der betrachtete Farbwert einer bestehenden Verteilungskomponente zugeordnet wurde, erfolgt eine Anpassung von Varianz, Mittelwert und Gewicht, um die neue Beobachtung zu berücksichtigen und die Komponente ggf. an variable Bildinhalte anzupassen. Die neuen Parameter ergeben sich nach [KB03] zu

$$\omega_k^{(t+1)} = (1 - \alpha) \omega_k^{(t)} + \begin{cases} \alpha & \text{für } k = k' \\ 0 & \text{sonst} \end{cases} \quad (\text{A.4})$$

$$\underline{\mu}_{k'}^{(t+1)} = (1 - \rho) \underline{\mu}_{k'}^{(t)} + \rho \underline{\mathbf{C}} \quad (\text{A.5})$$

$$\left(\sigma_{k'}^{(t+1)} \right)^2 = (1 - \rho) \left(\sigma_{k'}^{(t)} \right)^2 + \rho \left(\underline{\mathbf{C}} - \underline{\mu}_{k'}^{(t)} \right) \left(\underline{\mathbf{C}} - \underline{\mu}_{k'}^{(t)} \right)^T \quad (\text{A.6})$$

Dabei sind die Lernraten α und ρ abhängig vom Zeitindex t :

$$\alpha = \max \left(\frac{1}{t+1}, \alpha_0 \right) \quad (\text{A.7})$$

$$\rho = \max \left(\frac{1}{m_{k'}+1}, \rho_0 \right) \quad (\text{A.8})$$

Der Wert $m_{k'}$ gibt die Anzahl der Farbwerte an, die der Komponente k' bereits zugeordnet wurden. Varianzen und Mittelwerte der übrigen Komponenten bleiben unverändert.

Beide Lernraten steuern den Anteil des neuen Farbwerts am aktualisierten Modell. Die Abhängigkeit von der Zeit bzw. der Anzahl bisher beobachteter Farbwerte fördern eine schnelle Adaption zu Beginn der Bildsequenz und für neu erstellte Komponenten. Im stationären Zustand erreichen α und ρ die unteren Grenzwerte α_0 bzw. ρ_0 , die in [KB03] gleich gewählt sind.

Dynamisches Verhalten Da die Summation in (A.3) bei Index 1 beginnt, folgt mit (A.1), dass diejenigen Komponenten als Hintergrund angenommen werden, die ein hohes Gewicht und niedrige Varianz aufweisen. Durch das Update dieser Parameter in (A.4) und (A.6) passt sich die Klassifikation dynamisch an den Bildinhalt an. Eine bisher nicht beobachtete Farbe

erzeugt zunächst eine neue Verteilungskomponente mit niedrigem Fitness-Wert und wird als Vordergrund betrachtet. Tritt derselbe Farbwert jedoch häufiger auf, so steigt ω , während σ sinkt; entsprechend (A.1) erhält die Komponente einen niedrigeren Index und wird schließlich als Hintergrund klassifiziert.

Schattenerkennung Die Erkennung von Schatten erfolgt wie im unimodalen Fall (vgl. Abschnitt 4.2.1). Während der Schwellwert Θ_c konstant und global gesetzt wird, berechnet sich Θ_b aus der Standardabweichung der betrachteten Verteilungskomponente zu $\Theta_b = 2,5\sigma$. Die Bedingung (A.2) ändert sich somit zu

$$\Delta_c < \Theta_c \quad \wedge \quad \Delta_b < 2,5\sigma_k \quad (\text{A.9})$$

mit Δ_c und Δ_b entsprechend (4.24) bzw. (4.25).

Anhang B

Viterbi-Algorithmus

Eine rekursive Implementierung des Viterbi-Algorithmus ist in [RJ86] beschrieben. Abbildung B.1 zeigt eine nichtrekursive Umsetzung, die eine Beschleunigung der Berechnung durch vorzeitiges Abbrechen wenig aussichtsreicher Pfade erlaubt.

Viterbi-Algorithmus

1. Deklariere t als Schleifenindex, i als Hilfsindex und $W(t)$ als Menge von Pfaden zu einer Hypothese des Zeitindex t .
2. Initialisiere $t = 0$, $W(0) = \{\langle 0 \rangle, \langle 1 \rangle, \dots, \langle n(0) - 1 \rangle\}$ und $W(t) = \emptyset$ für $t = 1, 2, \dots, T - 1$.
3. Finde zu jeder Hypothese $H_i(t + 1)$ mit $i = 0, 1, \dots, n(t + 1) - 1$ den günstigsten Pfad $\widehat{W}_i(t + 1)$ gemäß

$$W'(t) = \operatorname{argmax}_{W(t) \in W(t)} (P(W(t)) \cdot P_{\text{dyn}, w(t)i}(t + 1) \cdot P_{\text{stat}, i}(t + 1))$$

$$\widehat{W}_i(t + 1) = \langle w'(0), w'(1), \dots, w'(t), i \rangle$$

Füge $\widehat{W}_i(t + 1)$ zu $W(t + 1)$ hinzu.

4. Erhöhe t um 1.
5. Wenn $t < T - 1$, gehe zu Schritt 3.
6. Der Pfad

$$\widehat{W}(T - 1) = \operatorname{argmax}_{W(T-1) \in W(T-1)} P(W(T - 1))$$

ist die wahrscheinlichste Interpretation des beobachteten Vorgangs.

Abbildung B.1: Nichtrekursive Implementierung des Viterbi-Algorithmus.

Der Viterbi-Algorithmus wird in der vorliegenden Arbeit mehrfach eingesetzt:

- Im Rahmen der Merkmalsextraktion zur Zuordnung der segmentierten Regionen zu den

einzelnen Markern (vgl. Abschnitt 4.5)

- Zur Mehrdeutigkeitsauflösung nach der Berechnung der k Positurkandidaten pro Einzelbild (vgl. Abschnitt 5.3.1)
- Zur dynamischen Zeitanpassung beim Vergleich von Anwender- und Referenzgebärde (hierbei ist die Zustandsabfolge innerhalb eines Pfades durch Randbedingungen eingeschränkt; vgl. Abschnitt 6.2.1)

Anhang C

Berechnung der Wahrscheinlichkeiten für Markerzuordnungshypothesen

Im Folgenden sind die zur Berechnung der verschiedenen statischen und dynamischen Zuordnungswahrscheinlichkeiten verwendeten Parameter der Sigmoid-Bewertungsfunktion (4.66) aufgelistet (s. Abschnitt 4.5.3). Die Werte $x_{0,9}$ und $x_{0,1}$ haben die anschauliche Entsprechung „sehr plausibel“ bzw. „sehr unwahrscheinlich“ und wurden empirisch ermittelt, die Konstante b_G steht für die die Gesichtsbreite (vgl. Abschnitt 4.4). Tabelle C.1 enthält die Angaben für die in Abbildung 3.4a/b gezeigte Markergeometrie, in der die Markierungen etwa 45% der Finger bedecken. Für die in Abbildung 7.5b visualisierte Markergeometrie, die eine Abdeckung von etwa 85% verwendet, sind einige Werte gemäß Tabelle C.2 zu ändern.

Tabelle C.1: Parameter zur Hypothesenbewertung bei Verwendung der in Abbildung 3.4a/b gezeigten Markergeometrie, in der die Finger zu ca. 45% bedeckt sind.

Kriterium	Marker	Komponente	$x_{0,9}$	$x_{0,1}$
Flächeninhalt a	Finger, Daumen	$P_{\text{stat},a}$	$0,08b_G^2$	$0,16b_G^2$
	Handrücken	$P_{\text{stat},a}$	$0,18b_G^2$	$0,36b_G^2$
	nichtdominante Hand	$P_{\text{stat},a}$	$1,5b_G^2$	$2b_G^2$
Flächenänderung $ \Delta a $	Finger, Daumen	$P_{\text{dyn},a}$	$0,04b_G^2$	$0,07b_G^2$
	Handrücken	$P_{\text{dyn},a}$	$0,075b_G^2$	$0,13b_G^2$
	nichtdominante Hand	$P_{\text{dyn},a}$	$0,3b_G^2$	$0,7b_G^2$
Entfernung von $\underline{s}_{\text{DH}}$	Finger, Daumen, Handrücken	$P_{\text{stat},r}$	b_G	$1,8b_G$
Diff. zu Hintergrund	alle	$\gamma - P_{\text{stat},\mu}^a$	15^b	25^b
Bewegung	alle	$P_{\text{dyn},s}$	$0,3b_G$	$0,8b_G$

^aDa die Abweichung vom Hintergrundfarbwert nicht negativ, sondern positiv mit der Plausibilität der Hypothese korreliert, wird die Bewertungsfunktion für dieses Kriterium gespiegelt, so dass sie streng monoton steigt.

^bEuklidische Distanz im RGB-Farbraum bei $R, G, B \in \{0, \dots, 255\}$.

Tabelle C.2: Geänderte Parameter zur Hypothesenbewertung bei Verwendung der in Abbildung 7.5b gezeigten Markergeometrie, in der die Finger zu ca. 85% bedeckt sind.

Kriterium	Marker	Komponente	$x_{0,9}$	$x_{0,1}$
Flächeninhalt a	Finger, Daumen	$P_{\text{stat},a}$	$0,13b_G^2$	$0,26b_G^2$
Flächenänderung $ \Delta a $	Finger, Daumen	$P_{\text{dyn},a}$	$0,05b_G^2$	$0,08b_G^2$

Anhang D

Generierung synthetischer Testdaten

Zur Evaluation der Positurerkennung in Kapitel 7 dienen synthetische Posituransichten, die zufällig generiert werden und somit eine Überanpassung der Rekonstruktionsalgorithmen vermeiden. Abschnitt D.1 erläutert zunächst die Erzeugung zufälliger Einzelposituren. Hierauf basiert die in Abschnitt D.2 beschriebene Generierung zufälliger Positurrensequenzen.

Eine wichtiger Aspekt der synthetischen Testdaten ist die Modellierung des in der praktischen Anwendung auftretenden Bildrauschens, das Über- und Untersegmentierung zur Folge haben kann und dadurch die Merkmalsextraktion stört. Abschnitt behandelt einen Algorithmus zur Simulation dieses Rauschprozesses in synthetischen Posituransichten.

D.1 Erzeugen zufälliger Posituren

Durch die Verwendung unabhängiger gleichverteilter Zufallszahlen $x \in [0, 1]$ für die Beugungsgrade $b_{D,i}$ bis $b_{KF,i}$ (vgl. (3.1)) sind sämtliche möglichen Handformen $\underline{\mathbf{B}}$ abgedeckt. Da jedoch eine vollständige Flexion oder Extension der Finger, die in der Realität vergleichsweise häufig ist, hierbei mathematisch eine Wahrscheinlichkeit von Null hat, wird zunächst jeder Finger mit einer Wahrscheinlichkeit von P_{Flex} bzw. P_{Ext} vollständig gebeugt bzw. gestreckt; die Verwendung zufälliger Werte erfolgt mit der Komplementärwahrscheinlichkeit von $1 - P_{\text{Flex}} - P_{\text{Ext}}$. Abduktion und Adduktion sowie die Gelenkstellungen des Daumens bleiben hiervon unbeeinflusst und werden immer zufällig bestimmt. Aufgrund der in Abschnitt 3.2.3.1 beschriebenen Interpretation der Beugungsgrade ist die Stellung jedes einzelnen Fingergelenks somit im anatomisch möglichen Bereich gleichmäßig verteilt.

Zur Festlegung der Ansichtswinkel ϕ_{lat} und ϕ_{lon} ist eine Gleichverteilung ungeeignet, da sie eine Häufung von Ansichtspunkten in Polnähe zur Folge hätte. Beispielsweise entfielen auf das Intervall $\phi_{\text{lat}} \in [-45^\circ, +45^\circ]$ die Hälfte der Punkte, die entsprechende Kugelzone ist hingegen größer als die beiden Kugelkappen mit $|\phi_{\text{lat}}| > 45^\circ$. Aus diesem Grund wird stattdessen ein zufälliger Vektor $\underline{\mathbf{x}} = [x \ y \ z]^T$ mit $0 < \|\underline{\mathbf{x}}\| \leq 1$ und einer innerhalb dieses Volumens

konstanten Verteilungsdichte generiert.¹ Der Schnittpunkt der Ansichtskugel und der Halbgeraden $a\mathbf{x}$ mit $a > 0$ bestimmt den Ansichtspunkt, d. h. ϕ_{lat} und ϕ_{lon} , während ϕ_{rot} erneut eine davon unabhängige gleichverteilte Zufallszahl ist.²

Ist die resultierende Handform oder -stellung anatomisch unzulässig (beispielsweise aufgrund maximaler Extension des Ringfingers bei gleichzeitiger maximaler Flexion des Mittelfingers; vgl. Abschnitt 3.2.4), wird der Vorgang wiederholt. Im Unterschied zu einer Modifikation der Positur – etwa durch Flexion oder Extension der betroffenen Finger – entstehen auf diese Weise keine Häufungspunkte an den Grenzen des zulässigen Wertebereichs. In Abbildung 7.1 (s. Seite 152) sind einige auf diese Weise erzeugte Posituren gezeigt. Tabelle D.1 in Abschnitt D.2 enthält die für die Wahrscheinlichkeiten P_{Flex} und P_{Ext} verwendeten Werte.

D.2 Erzeugen zufälliger Positurrensequenzen

Eine zufällige Positurrensequenz der Länge T stellt eine flüssige Bewegung der Hand von einer Ausgangspositur $\mathbf{P}(0)$ in eine Endpositur $\mathbf{P}(T - 1)$ dar. Das direkte Ausführen dieser Bewegung durch schrittweises Angleichen der Positurparameter resultiert jedoch in einem überwiegend linearen Verlauf der Beugungsgrade sowie der Ansichtspunkte auf der Oberfläche der Ansichtskugel. Durch die Interpolation der Positurparameter wäre die Rekonstruktion einer solchen Positurrenfolge vergleichsweise einfach.

In der Praxis treten hingegen innerhalb einer Gebärde im Allgemeinen Änderungen der Bewegungsrichtung und -geschwindigkeit auf. Aus diesem Grund wird eine zufällige Positurrensequenz aus n Segmenten $i \in \{0, 1, \dots, n - 1\}$ mit variabler Länge T_i zusammengesetzt, die jeweils eine Bewegung von einer Positur $\mathbf{P}_i$ in eine Positur $\mathbf{P}_{i+1}$ repräsentieren. Die Gesamtsequenz ist somit eine Bewegung von $\mathbf{P}(0) = \mathbf{P}_0$ nach $\mathbf{P}(T - 1) = \mathbf{P}_n$.

An den Übergangsposituren $\mathbf{P}_1$ bis $\mathbf{P}_{n-1}$ ist der Verlauf der Beugungsgrade $\mathbf{B}(t)$ stetig, jedoch nicht notwendigerweise stetig differenzierbar. Drehachse und -winkel zweier aufeinanderfolgender Handstellungen $\mathbf{\Phi}(t)$ können sich hier beliebig ändern, sind jedoch innerhalb der einzelnen Segmente konstant.

Die Generierung eines Segments i der Länge T_i beginnt mit der Festlegung einer Startpositur $\mathbf{P}_i = \langle \mathbf{B}_i, \mathbf{\Phi}_i \rangle$. Für das erste Segment ist dies eine gemäß Abschnitt D.1 erzeugte Zufallspositur, für die folgenden Segmente wird die Bewegung ausgehend von der letzten Positur des vorherigen Segments $\mathbf{P}_{i-1}$ fortgeführt. Um Gebärdenphasen zu repräsentieren, in denen Handform und -stellung konstant bleiben, werden anschließend mit einer Wahrscheinlichkeit von P_{const} alle Posituren des Segments gleich der Startpositur gesetzt. Andernfalls erfolgt die zufällige Bestimmung einer Handstellung $\mathbf{\Phi}_{i+1}$, in die sich die Hand ausgehend von $\mathbf{\Phi}_i$ dreht. Die Geschwindigkeit dieser Drehung ist konstant und ergibt sich aus der Segmentlänge T_i ,

¹Die gewünschte Verteilungsdichte ergibt sich durch Gleichverteilung der drei Koordinaten $x, y, z \in [-1, 1]$ und Verwerfen der Punkte $x = y = z = 0$ sowie $x^2 + y^2 + z^2 > 1$.

²Der Einfluss der Rotation ist aufgrund der in Abschnitt 5.2.1 beschriebenen Optimierung (vgl. (5.4)) bis auf Quantisierungseffekte vernachlässigbar und wird daher primär aus Gründen der Vollständigkeit betrachtet.

wird jedoch durch eine Obergrenze $\hat{\Delta}_{\Phi}$ beschränkt, so dass $\Delta_{\Phi}(\Phi(t), \Phi(t+1)) \leq \hat{\Delta}_{\Phi}$ gilt. Die Drehachse liegt so, dass keine anatomisch unmöglichen Stellungen auftreten.³

Anschließend ist für jeden einzelnen Beugungsgrad b eine zufällige Schrittweite Δb zu bestimmen, die innerhalb eines Intervalls $[-\hat{\Delta}_b, +\hat{\Delta}_b]$ gleichverteilt ist. Ausgehend vom Startwert $b(t_0)$ ergibt sich der Verlauf des Beugungsgrades für $t > t_0$ damit zu

$$b(t) = \begin{cases} 0 & \text{falls } b(t-1) + \Delta b < 0 \\ 1 & \text{falls } b(t-1) + \Delta b > 1 \\ b(t-1) + \Delta b & \text{sonst} \end{cases} \quad (\text{D.1})$$

Durch die unabhängige Änderung der Beugungsgrade können die Handformparameter dabei auf Bereiche des Positurrenraums treffen, die anatomisch nicht zulässig sind. In diesem Fall wird $\underline{B}(t)$ entlang der diesen Bereich begrenzenden Hyperfläche so bewegt, dass für alle Beugungsgrade gilt:

$$\begin{aligned} 0 \leq \frac{b(t) - b(t-1)}{\Delta b} \leq 1 & \quad \text{falls } \Delta b \neq 0 \\ b(t) = b(t-1) & \quad \text{sonst} \end{aligned} \quad (\text{D.2})$$

Dieser Vorgang ist vom Verlauf der Ansichtswinkel unabhängig.

Zur Erzeugung einer zufälligen Positurrensequenz ist zunächst die gewünschte Gesamtlänge T vorzugeben. Die Anzahl der Segmente n wird auf Grundlage einer diskreten Wahrscheinlichkeitsverteilung $P_{\text{Seg}}(n)$ bestimmt. Hieraus resultieren $n - 1$ Übergangspunkte t_i , die im Intervall $[1, T - 2]$ gleichverteilt sind, wobei jedoch $t_{i+1} > t_i$ gilt, um Segmente der Länge

Tabelle D.1: Parameter der zur Generierung zufälliger Posituren sowie Positurrensequenzen eingesetzten Algorithmen. In der rechten Spalte sind die im Rahmen der Evaluation verwendeten numerischen Werte angegeben.

Parameter	Beschreibung	Wert
P_{Flex}	Wahrscheinlichkeit maximaler Flexion eines Fingers	0,1
P_{Ext}	Wahrscheinlichkeit maximaler Extension eines Fingers	0,1
T	Gesamtlänge der Zufallssequenz	50
P_{const}	Wahrscheinlichkeit eines statischen Sequenzabschnitts	0,2
$\hat{\Delta}_{\Phi}$	maximale Ansichtswinkeldifferenz aufeinanderfolgender Posituren	20°
$\hat{\Delta}_b$	maximale Beugungsgraddifferenz aufeinanderfolgender Posituren	0,1
$P_{\text{Seg}}(n)$	Wahrscheinlichkeit der Teilung in n Segmente mit $n \in \{1, \dots, 4\}$	0,25
	mit $n \notin \{1, \dots, 4\}$	0

³Eine geeignete Drehachse existiert immer, da die zu vermeidenden Ansichtspunkte ausschließlich auf der durch $\phi_{\text{lat}} > 65^\circ$ beschriebenen Kugelkappe liegen.

Null zu vermeiden. Abbildung 7.2 (s. Seite 153) zeigt exemplarisch eine Zufallssequenz, die aus drei Segmenten besteht. Die Übergänge erfolgen bei $t = 15$ sowie bei $t = 24$. Tabelle D.1 fasst die Parameter des beschriebenen Verfahrens zur zufälligen Positur- sowie Positurrensequenzgenerierung zusammen und gibt die zur Evaluation verwendeten Werte an.

D.3 Simulation des Bildrauschens

Während die Segmentierung der synthetischen Handansicht grundsätzlich fehlerfrei arbeitet, stimmen die vom Mean-Shift-Algorithmus in Realansichten segmentierten Randkurven infolge des Bildrauschens im Allgemeinen nicht mit den tatsächlichen Regionengrenzen überein. Ist die berechnete Region größer als die tatsächliche, wird dies als Untersegmentierung bezeichnet, im umgekehrte Fall als Übersegmentierung. In Abbildung D.1 ist diese Abweichung anhand eines Beispiels visualisiert.

Ziel des künstlichen Bildrauschens ist es, die synthetischen Ansichten so zu stören, dass ebenfalls „Unter-“ und „Übersegmentierungen“ auftreten. Dies lässt sich erreichen, indem mehreren Pixeln die Farbe eines benachbarten oder nahe gelegenen Pixels zugewiesen wird. Auf diese Weise entstehen ähnliche Zerklüftungen wie in Abbildung D.1b gezeigt. Die Segmentierung dieser modifizierten Regionen erfolgt noch immer fehlerfrei, im Vergleich zum Originalbild entspricht der Effekt jedoch einer Unter- bzw. Übersegmentierung.

Das künstliche Rauschen ist als iterativer Prozess implementiert, der das Bild in jedem Durchlauf weiter verschlechtert. Zunächst erfolgt die zufällige Auswahl einer quadratischen Region der Seitenlänge s . Aus einem konzentrischen Quadrat der Seitenlänge $s + 2\epsilon$ wird ebenfalls

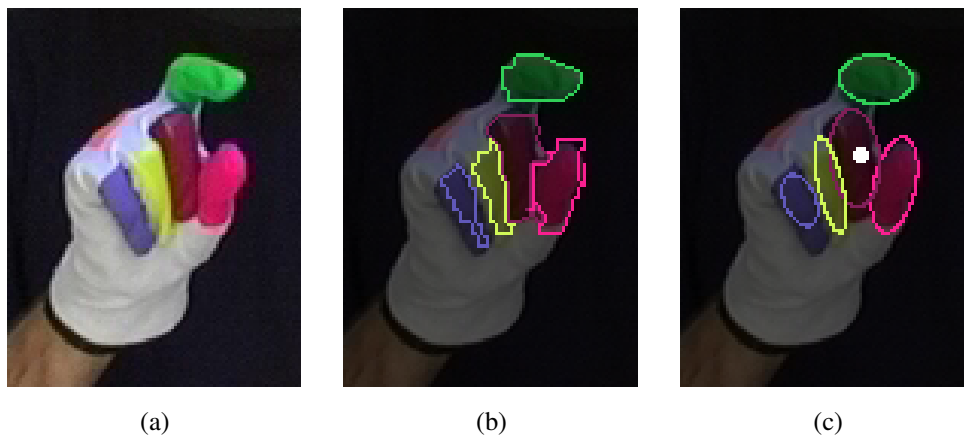


Abbildung D.1: Die vom Mean-Shift-Algorithmus im Eingangsbild (a) segmentierten Regionen (b) weichen von den tatsächlichen Randkurven ab. Im gezeigten Beispiel ist besonders die Übersegmentierung des kleinen Fingers zu erkennen. Durch die elliptische Regionenapproximation (c) wird zwar die Zerklüftung der Randkurven eliminiert, die Übersegmentierung ist jedoch noch immer sichtbar. Der weiße Punkt in der Bildmitte kennzeichnet den gemeinsamen Markerschwerpunkt s_{DH} .

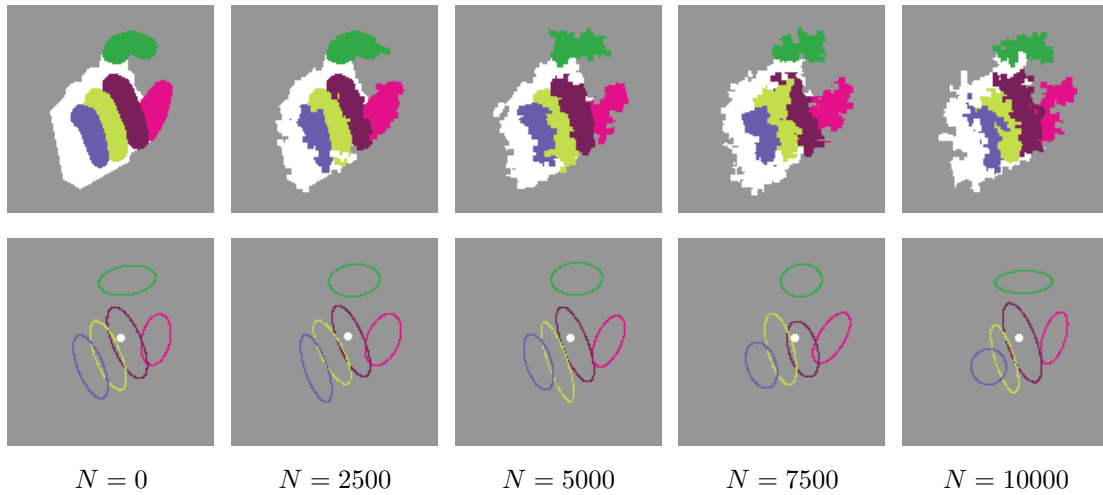


Abbildung D.2: Originalansicht (links) sowie durch simuliertes Bildrauschen gestörte Ansichten derselben Positur für verschiedene Werte N . Die Ausschnittaauflösung beträgt 130×130 Pixel, die Parameter s und ϵ des Algorithmus sind konstant mit $s = 3$ und $\epsilon = 1$. In der oberen Zeile sind die unbeleuchteten Posituransichten (vgl. Abschnitt 5.1.2) gezeigt, jeweils darunter die entsprechenden elliptischen Randkurvenapproximationen.

zufällig ein Pixel bestimmt, dessen Farbe der Algorithmus allen Pixeln des ersten Quadrats zuweist. Die Anzahl der Wiederholungen N dieses Vorgangs sowie die beiden Parameter s und ϵ bestimmen die Stärke des Rauschens. Für die Evaluation der Rekonstruktionsleistung wurden $s = 3$ und $\epsilon = 1$ konstant gewählt und die Intensität über N gesteuert. Abbildung D.2 zeigt eine Abbildung D.1a entsprechende synthetische Posituransicht bei verschiedenen Rauschintensitäten sowie die daraus berechneten elliptischen Randkurvenapproximationen.

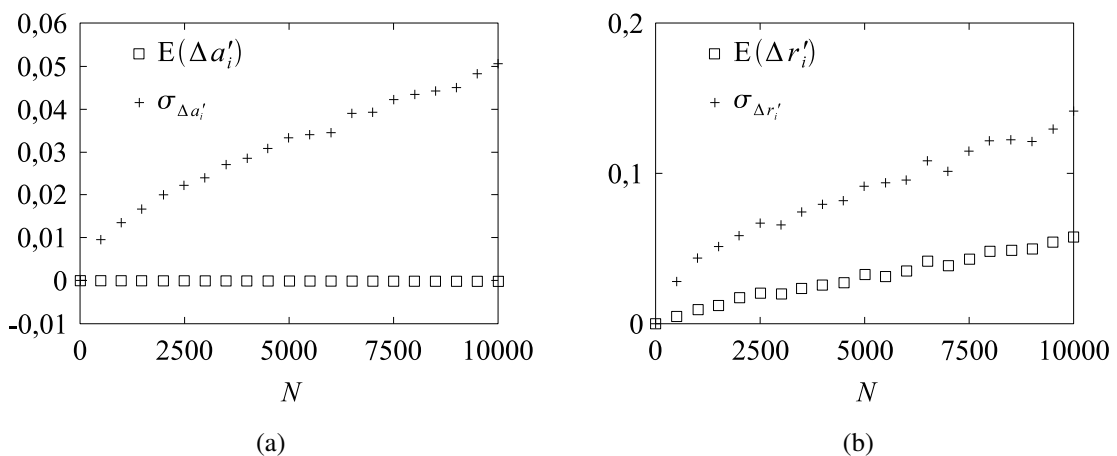


Abbildung D.3: Mittelwert und Standardabweichung des Rauschprozesses im Merkmalsraum für die Komponenten a'_i (a) und r'_i (b) des Merkmalsvektors $\underline{F}_{DH}$ bei verschiedenen Rauschintensitäten N . Berücksichtigt sind nur Marker, die im unverrauschten Bild sichtbar sind, d. h. für die $a'_i \neq 0$ gilt. Für den Radius (b) muss zusätzlich $a'_i \neq 0$ erfüllt sein, da dieses Merkmal andernfalls in der verrauschten Ansicht undefiniert ist.

Der im Vergleich zur Störung der 2D-Ansicht geringe Einfluss des Rauschens auf die elliptische Randkurvenapproximation belegt die Robustheit des Merkmalsvektors $\underline{F}_{DH}$, die auch bei der Verarbeitung realer Ansichten zum Tragen kommt. Um die Auswirkung im Merkmalsraum zu quantifizieren, wird der Rauschprozess als Störsignal Δf aufgefasst, das einem Merkmal f additiv überlagert ist:

$$\tilde{f} = f + \Delta f \quad (D.3)$$

Durch Vergleich von gestörten und ungestörten Merkmalsvektoren jeweils derselben Positur lassen sich Mittelwert $E(\Delta f)$ sowie Standardabweichung $\sigma_{\Delta f}$ des Signals Δf in Abhängigkeit von der Iterationsanzahl N bestimmen. Abbildung D.3 zeigt diese beiden Maße für die Merkmale a'_i und r'_i aller Marker. Jeder Stützpunkt basiert auf der Auswertung 1000 zufälliger Posituren. Marker, die in der Originalansicht verdeckt sind, sind in der gestörten Ansicht ebenfalls nicht zu sehen, d. h. $a'_i = \tilde{a}'_i = \Delta a_i = 0$. Diese Werte werden von der Messung ausgeschlossen. Im Fall des Radius r'_i gilt dies auch für Marker, die infolge des Rauschens vollständig verschwinden, da hier keine Abweichung quantifizierbar ist.

Der positive Erwartungswert $E(\Delta r'_i)$ entsteht dadurch, dass eine zufällige Verschiebung des Markerschwerpunkts $\underline{s}'_i$ aus geometrischen Gründen häufiger eine Entfernung vom gemeinsamen Schwerpunkt bewirkt als eine Annäherung. Dies ist in Abbildung D.4 veranschaulicht.

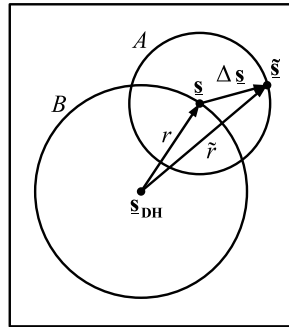


Abbildung D.4: Der Rauschprozess bewirkt eine Verschiebung des Markerschwerpunkts $\underline{s}$ um $\Delta \underline{s}$. Für einen gegebenen Abstand $\|\Delta \underline{s}\|$ liegt der resultierende Schwerpunkt $\tilde{\underline{s}}$ auf einem Kreis A um $\underline{s}$. Nimmt man die Richtung der Verschiebung $\Delta \underline{s}$ als gleichverteilt an, so entspricht das Verhältnis der Segmente des Kreises A innerhalb und außerhalb des Kreises B um $\underline{s}_{DH}$ mit dem Radius r dem Verhältnis der Wahrscheinlichkeiten $P(\tilde{r} < r)$ und $P(\tilde{r} > r)$. Aus $P(\tilde{r} < r) < P(\tilde{r} > r)$ folgt der in Abbildung D.3b erkennbare Anstieg des gestörten gegenüber dem ursprünglichen Radius.

Anhang E

Analyse und Optimierung der Rekonstruktionsleistung

Auf Basis der in Abschnitt 7.1 beschriebenen Evaluation mit synthetischen Posituransichten lassen sich Histogramme der Rekonstruktionsfehler berechnen, die aussagekräftiger als die dort angegebenen Mittelwerte sind. Diese Histogramme werden im Folgenden präsentiert. Messwerte außerhalb des dargestellten Wertebereichs sind in der Visualisierung jeweils der letzten Klasse des Histogramms zugeordnet (der Erwartungswert bleibt hiervon unbeeinflusst).

Eng mit der Leistungsanalyse verknüpft ist die Bestimmung geeigneter Werte für die verschiedenen Systemparameter. Ergebnisse dieser empirischen Parameteroptimierung sind jeweils bei der entsprechenden Messung angegeben.

Die nachfolgend verwendete Testpositurenmenge P_{in} (vgl. Abschnitt 7.1.2) ist für alle Auswertungen konstant, so dass die Ergebnisse miteinander vergleichbar sind. Messungen der Verarbeitungsgeschwindigkeit beziehen sich auf einen Standard-PC mit 1,6 GHz Taktfrequenz (AMD-Sempron-CPU), der die Datenbank vollständig im Arbeitsspeicher hält.

E.1 Optimale Positur bezüglich $\Delta_{\underline{P}}$

Die Abweichung der bezüglich $\Delta_{\underline{P}}$ optimalen Positur $\underline{P}_{\text{opt},\Delta_{\underline{P}}}$ von der zufälligen Testpositur quantifiziert die mit der verwendeten Posituredatenbank maximal zu erwartende Rekonstruktionsleistung. In den Abbildungen E.1 und E.2 sind die entsprechenden Fehlerhistogramme zu sehen. Die Häufung bei Null für die Differenzmaße $\Delta_{\underline{B},D}$ bis $\Delta_{\underline{B},KF}$ entsteht durch Finger, die sowohl in $\underline{P}_{\text{in}}$ als auch in $\underline{P}_{\text{opt},\Delta_{\underline{P}}}$ verdeckt sind und damit keine Abweichung aufweisen.

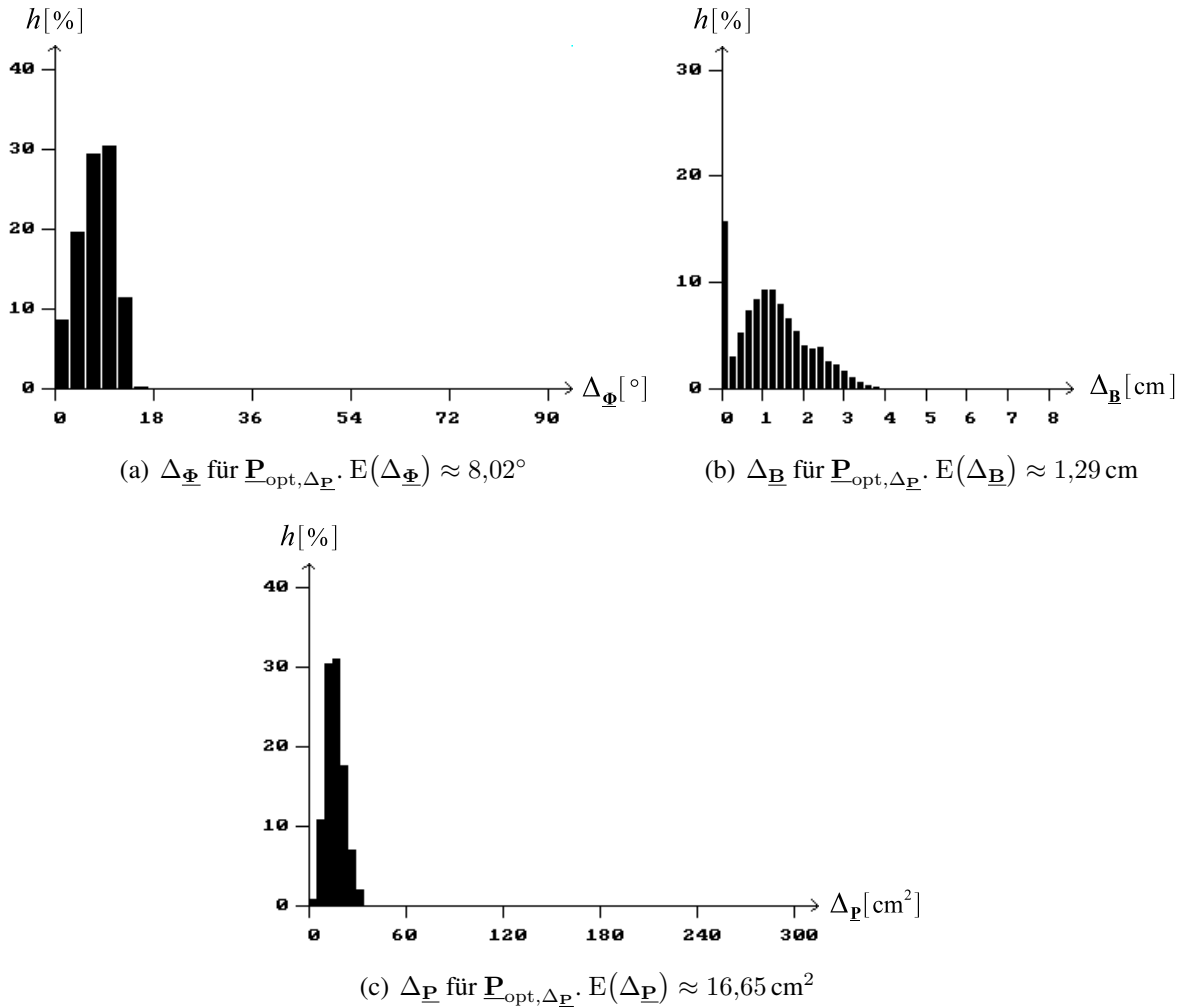
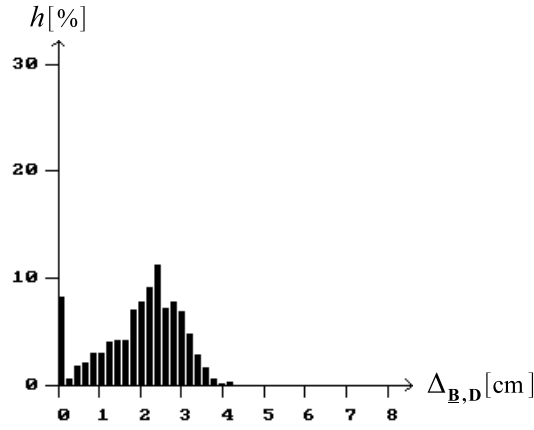
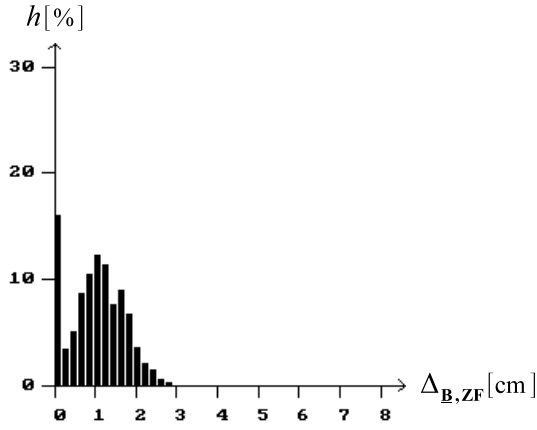


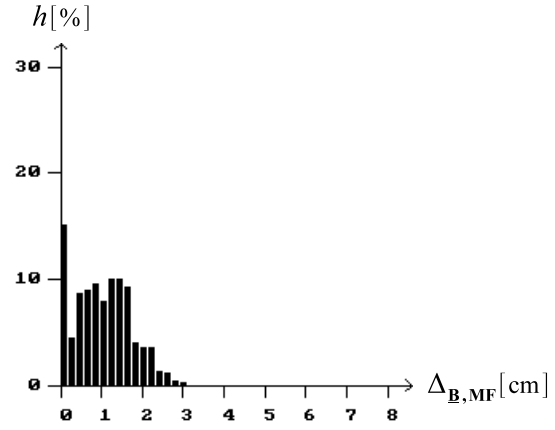
Abbildung E.1: Histogramme und Erwartungswerte der Differenzmaße Δ_{Φ} , $\Delta_{\underline{B}}$ und $\Delta_{\underline{P}}$ für 1000 zufällige Eingangsposituren $\underline{P}_{\text{in}}$ und die jeweils bezüglich $\Delta_{\underline{P}}$ ähnlichste Datenbankpositur $\underline{P}_{\text{opt},\Delta_{\underline{P}}} \in \mathcal{P}^{\text{db}}$.



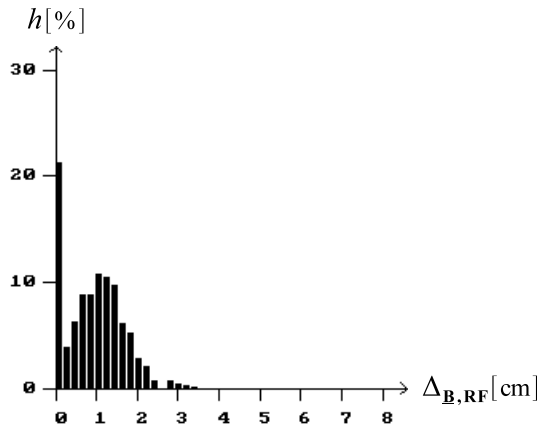
(a) $\Delta_{\underline{B},D}$ für $\underline{P}_{\text{opt},\Delta_{\underline{P}}}$. $E(\Delta_{\underline{B},D}) \approx 2,10$ cm



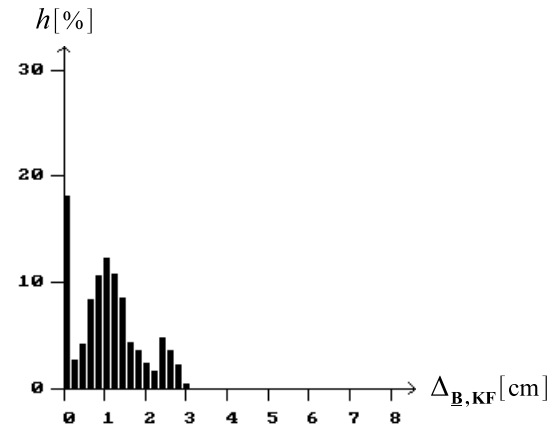
(b) $\Delta_{\underline{B},ZF}$ für $\underline{P}_{\text{opt},\Delta_{\underline{P}}}$. $E(\Delta_{\underline{B},ZF}) \approx 1,09$ cm



(c) $\Delta_{\underline{B},MF}$ für $\underline{P}_{\text{opt},\Delta_{\underline{P}}}$. $E(\Delta_{\underline{B},MF}) \approx 1,10$ cm



(d) $\Delta_{\underline{B},RF}$ für $\underline{P}_{\text{opt},\Delta_{\underline{P}}}$. $E(\Delta_{\underline{B},RF}) \approx 1,01$ cm



(e) $\Delta_{\underline{B},KF}$ für $\underline{P}_{\text{opt},\Delta_{\underline{P}}}$. $E(\Delta_{\underline{B},KF}) \approx 1,15$ cm

Abbildung E.2: Histogramme und Erwartungswerte der Differenzmaße $\Delta_{\underline{B},D}$ bis $\Delta_{\underline{B},KF}$ für 1000 zufällige Eingangsposituren P_{in} und die jeweils bezüglich $\Delta_{\underline{P}}$ ähnlichste Datenbankpositur $\underline{P}_{\text{opt},\Delta_{\underline{P}}} \in P^{\text{db}}$.

E.2 Optimale Positur bezüglich $\Delta_{\underline{F}}$ bei sequenzieller Suche

Die durch Anwendung der Merkmalsdifferenz ermittelten Posituren $\underline{P}_{\text{opt},\Delta_{\underline{F}}}$ zeigen erwartungsgemäß höhere Abweichungen als $\underline{P}_{\text{opt},\Delta_{\underline{P}}}$. Histogramme der Differenzmaße sind in den Abbildung E.3 und E.4 visualisiert. Der numerisch relativ starke Anstieg der Posituredifferenz $\Delta_{\underline{P}}$ geht auf die quadratische Natur dieses Maßes zurück. Im Mittel benötigt die sequenzielle Suche nach $\underline{P}_{\text{opt},\Delta_{\underline{F}}}$ etwa 14 Sekunden.

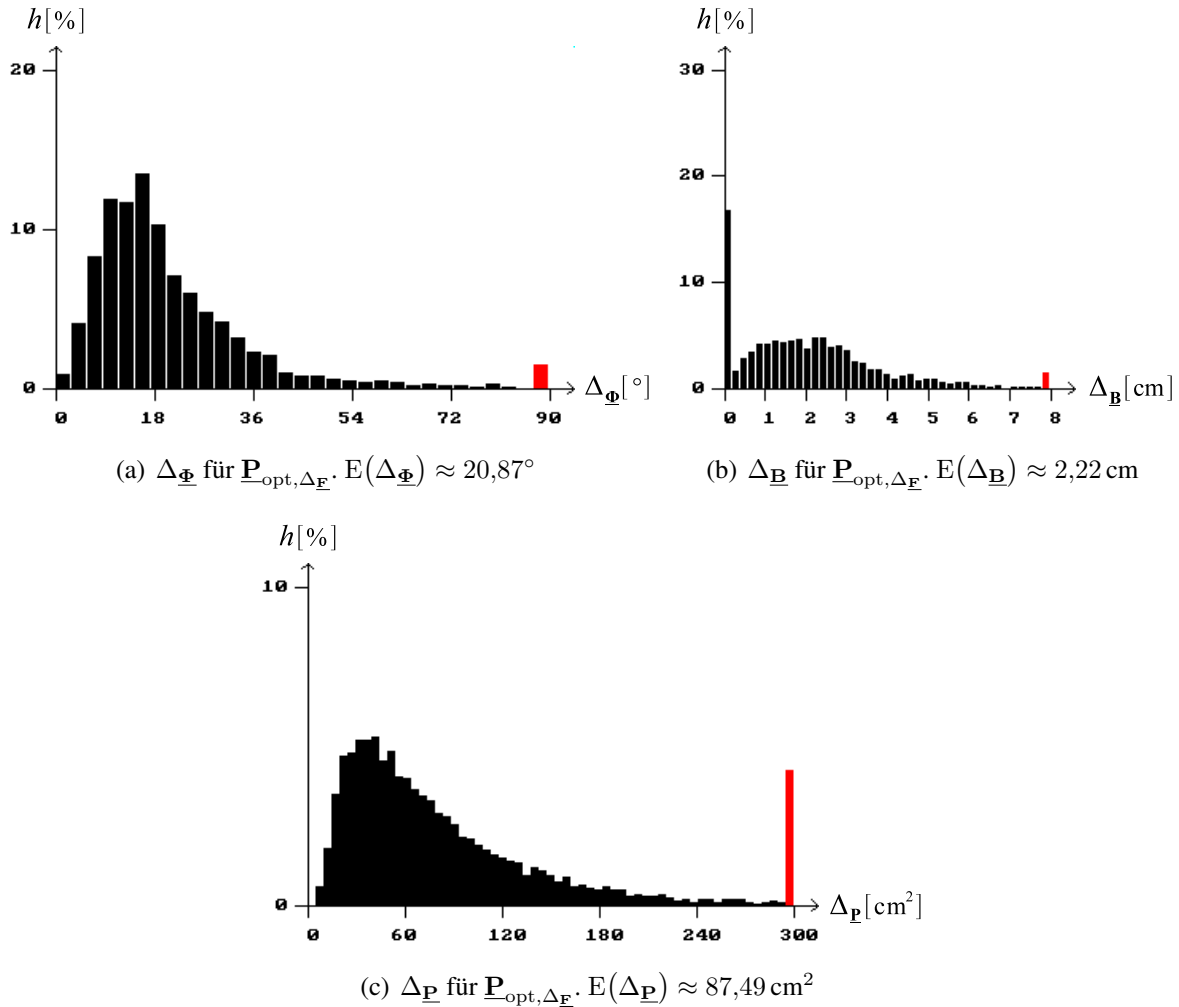
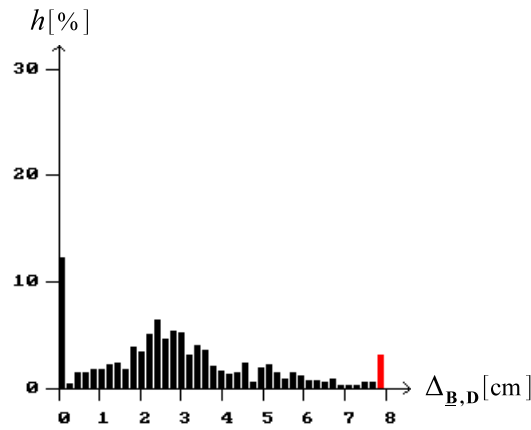
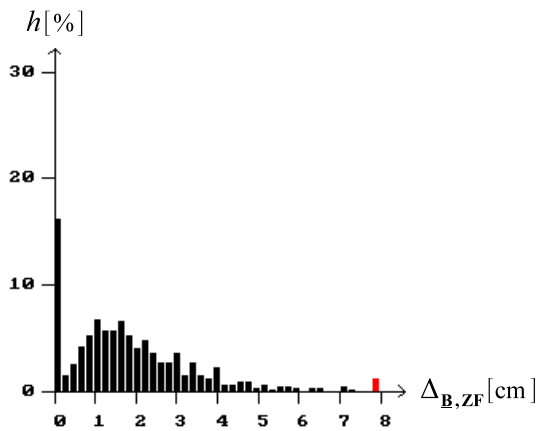


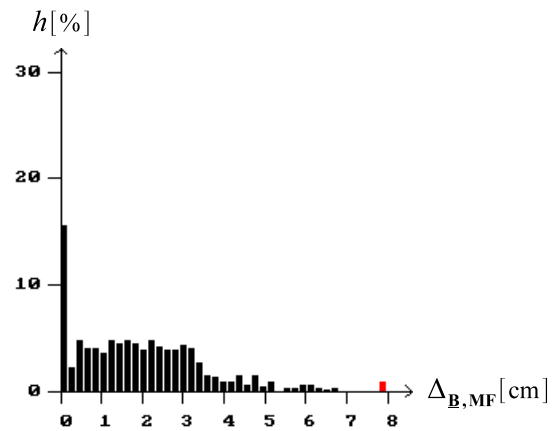
Abbildung E.3: Histogramme und Erwartungswerte der Differenzmaße $\Delta_{\underline{\Phi}}$, $\Delta_{\underline{B}}$ und $\Delta_{\underline{P}}$ für 1000 zufällige Eingangsposituren P_{in} und die jeweils bezüglich $\Delta_{\underline{F}}$ ähnlichste Datenbankpositur $\underline{P}_{\text{opt},\Delta_{\underline{F}}} \in P^{\text{db}}$.



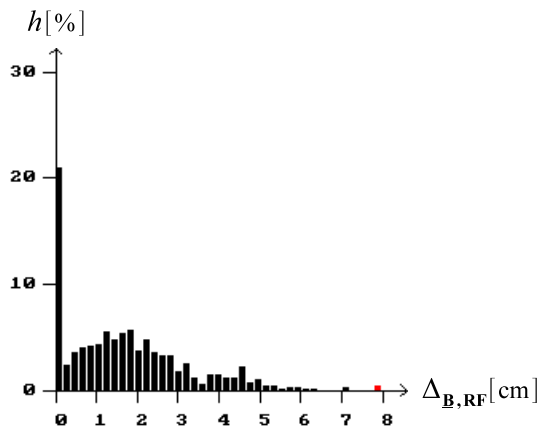
(a) $\Delta_{\underline{B},D}$ für $\underline{P}_{\text{opt},\Delta_{\underline{F}}}$. $E(\Delta_{\underline{B},D}) \approx 3,08$ cm



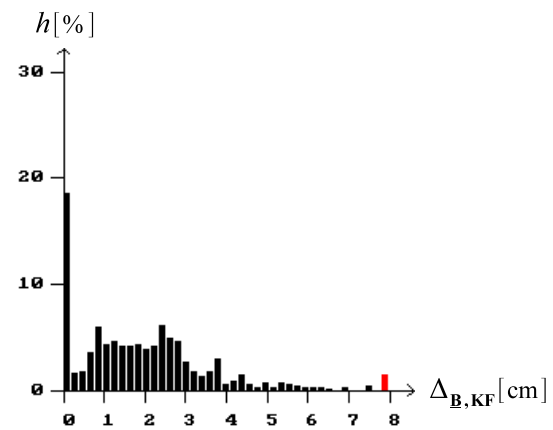
(b) $\Delta_{\underline{B},ZF}$ für $\underline{P}_{\text{opt},\Delta_{\underline{F}}}$. $E(\Delta_{\underline{B},ZF}) \approx 2,00$ cm



(c) $\Delta_{\underline{B},MF}$ für $\underline{P}_{\text{opt},\Delta_{\underline{F}}}$. $E(\Delta_{\underline{B},MF}) \approx 2,09$ cm



(d) $\Delta_{\underline{B},RF}$ für $\underline{P}_{\text{opt},\Delta_{\underline{F}}}$. $E(\Delta_{\underline{B},RF}) \approx 1,85$ cm



(e) $\Delta_{\underline{B},KF}$ für $\underline{P}_{\text{opt},\Delta_{\underline{F}}}$. $E(\Delta_{\underline{B},KF}) \approx 2,08$ cm

Abbildung E.4: Histogramme und Erwartungswerte der Differenzmaße $\Delta_{\underline{B},D}$ bis $\Delta_{\underline{B},KF}$ für 1000 zufällige Eingangsposituren P_{in} und die jeweils bezüglich $\Delta_{\underline{F}}$ ähnlichste Datenbankpositur $\underline{P}_{\text{opt},\Delta_{\underline{F}}} \in P^{\text{db}}$.

Das Verrauschen der Positurenansicht reduziert die Rekonstruktionsleistung nur geringfügig. Die Abbildungen E.5 bis E.8 zeigen die resultierenden Abweichungen und belegen die Robustheit der Merkmalsdifferenz.¹

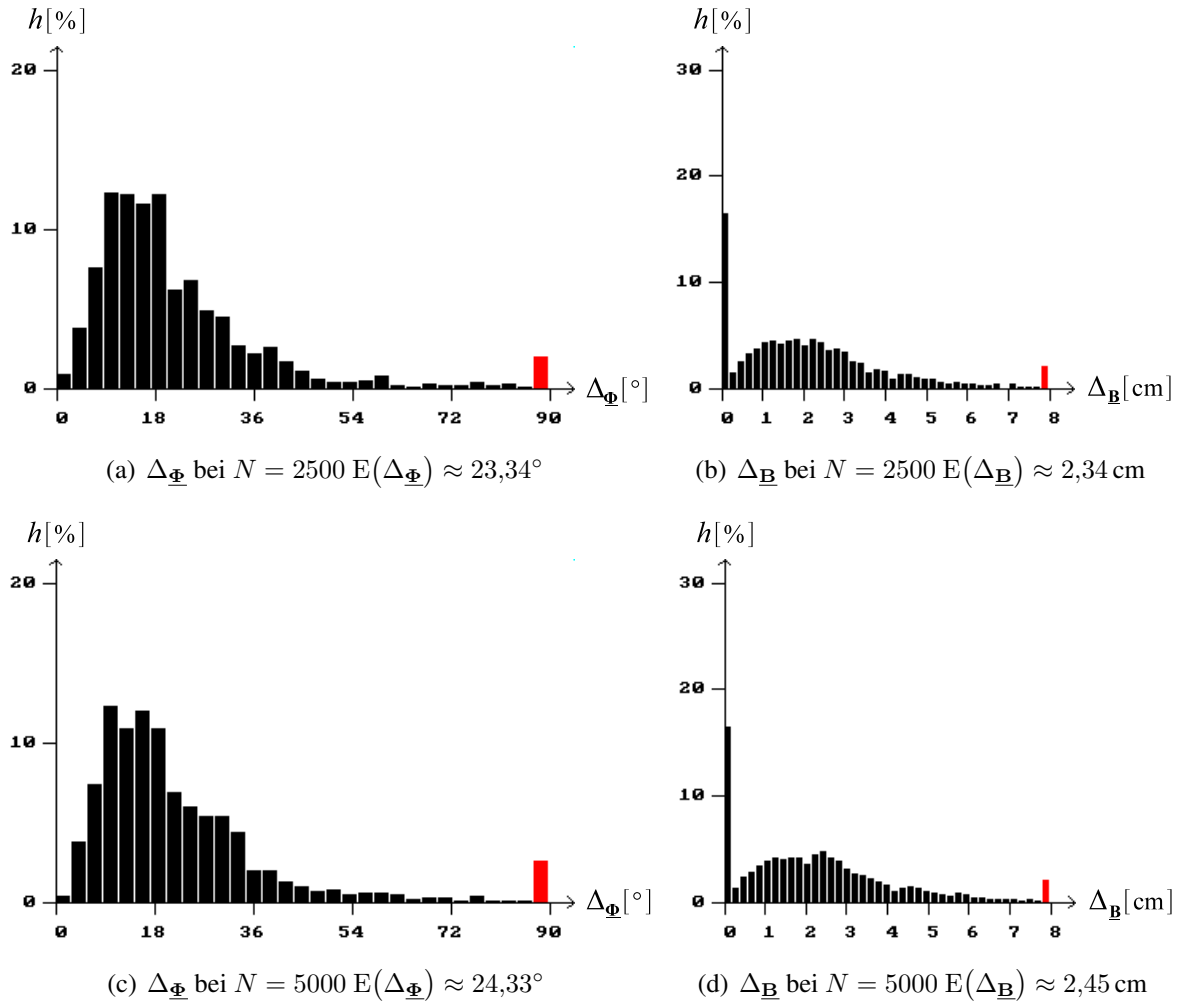
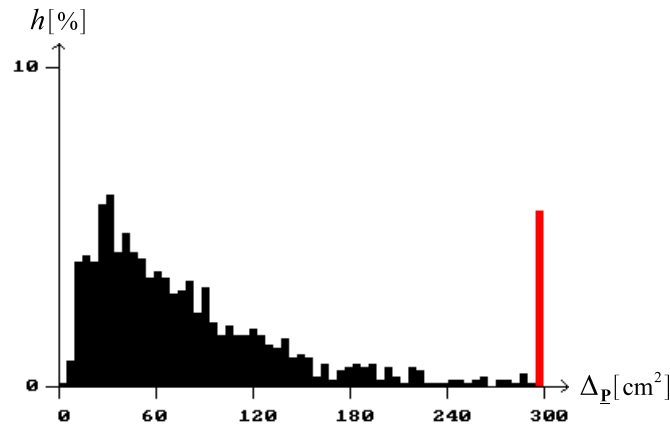
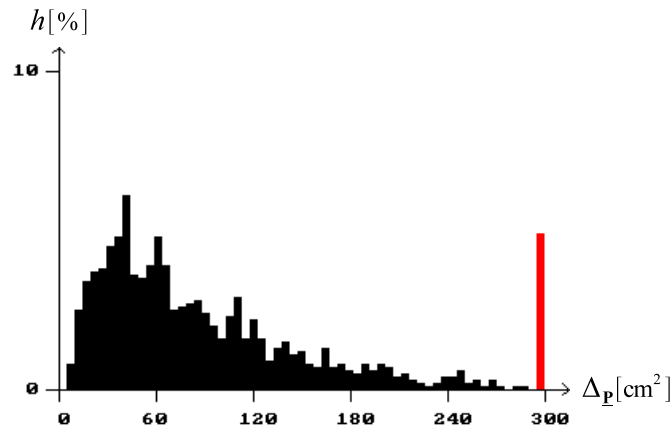


Abbildung E.5: Histogramme und Erwartungswerte der Differenzmaße $\Delta_{\underline{F}}$ und $\Delta_{\underline{B}}$ für 1000 zufällige Eingangsposituren P_{in} und die jeweils bezüglich $\Delta_{\underline{F}}$ ähnlichste Datenbankpositur $\underline{P}_{opt, \Delta_{\underline{F}}} \in P^{db}$ bei den Rauschintensitäten $N = 2500$ (a, b) und $N = 5000$ (c, d).

¹Aus Platzgründen wird auf die Aufschlüsselung nach Fingern verzichtet. Die Distanzen $\Delta_{\underline{B}, D}$ bis $\Delta_{\underline{B}, KF}$ sind annähernd identisch verteilt und werden daher durch das gemeinsame Histogramm $h(\Delta_{\underline{B}})$ hinreichend gut repräsentiert.

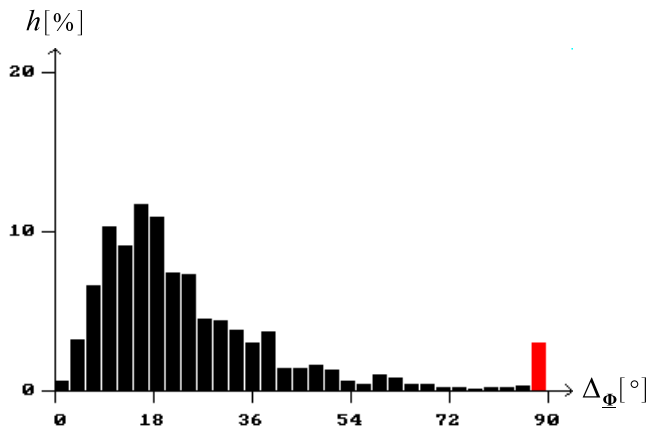


(a) $\Delta_{\underline{P}}$ bei $N = 2500$. $E(\Delta_{\underline{P}}) \approx 99,76 \text{ cm}^2$

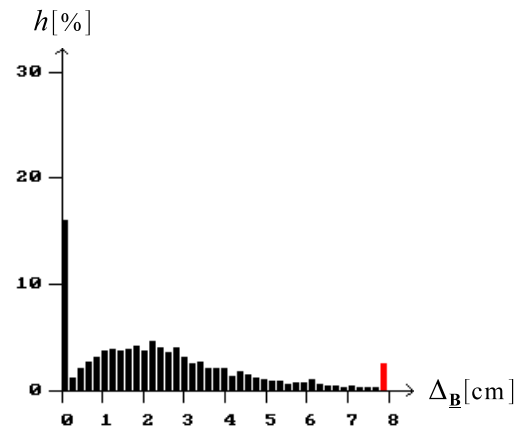


(b) $\Delta_{\underline{P}}$ bei $N = 5000$. $E(\Delta_{\underline{P}}) \approx 105,88 \text{ cm}^2$

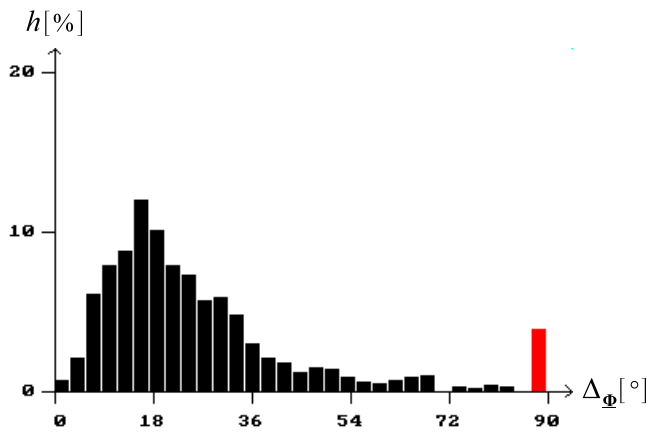
Abbildung E.6: Histogramme und Erwartungswerte der Posituredifferenz $\Delta_{\underline{P}}$ für 1000 zufällige Eingangsposituren $\underline{P}_{\text{in}}$ und die jeweils bezüglich $\Delta_{\underline{F}}$ ähnlichste Datenbankpositur $\underline{P}_{\text{opt}, \Delta_{\underline{F}}} \in \mathcal{P}^{\text{db}}$ bei den Rauschintensitäten $N = 2500$ (a) und $N = 5000$ (b).



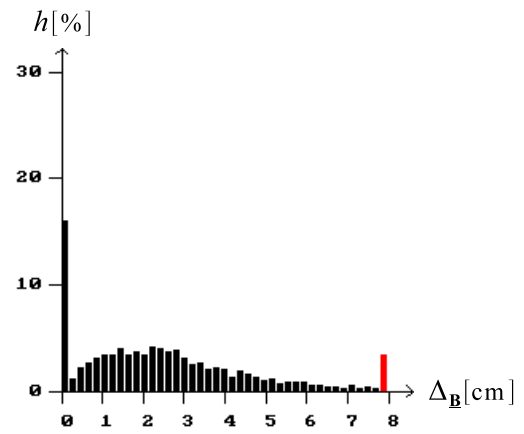
(a) Δ_{Φ} bei $N = 7500$ $E(\Delta_{\Phi}) \approx 26,63^\circ$



(b) $\Delta_{\underline{B}}$ bei $N = 7500$ $E(\Delta_{\underline{B}}) \approx 2,59$ cm

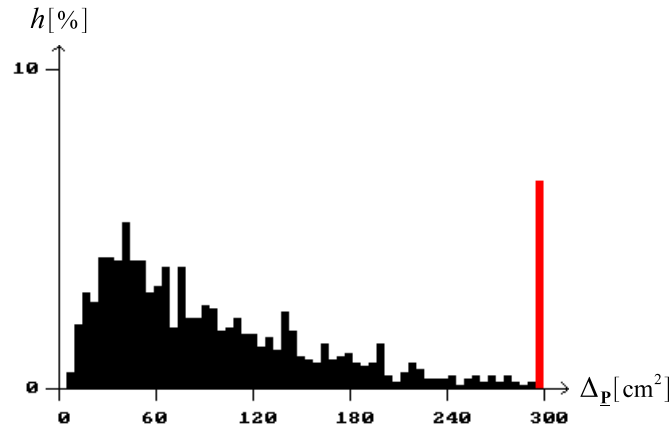


(c) Δ_{Φ} bei $N = 10000$ $E(\Delta_{\Phi}) \approx 28,66^\circ$

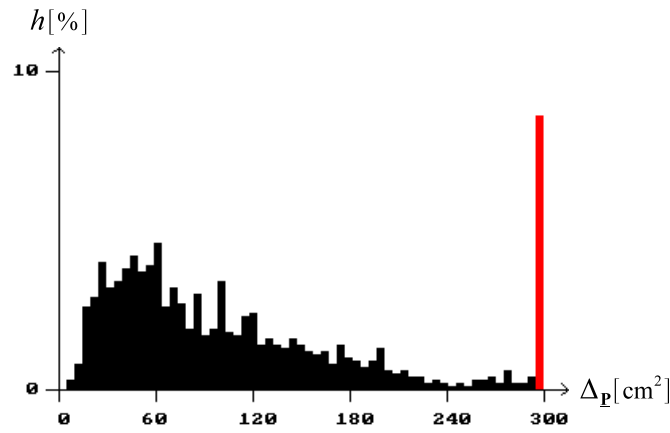


(d) $\Delta_{\underline{B}}$ bei $N = 10000$ $E(\Delta_{\underline{B}}) \approx 2,71$ cm

Abbildung E.7: Histogramme und Erwartungswerte der Differenzmaße Δ_{Φ} und $\Delta_{\underline{B}}$ für 1000 zufällige Eingangsposituren P_{in} und die jeweils bezüglich $\Delta_{\underline{F}}$ ähnlichste Datenbankpositur $\underline{P}_{\text{opt}, \Delta_{\underline{F}}} \in P^{\text{db}}$ bei den Rauschintensitäten $N = 7500$ (a, b) und $N = 1000$ (c, d).



(a) $\Delta_{\underline{P}}$ bei $N = 7500$. $E(\Delta_{\underline{P}}) \approx 119,20 \text{ cm}^2$



(b) $\Delta_{\underline{P}}$ bei $N = 10000$. $E(\Delta_{\underline{P}}) \approx 132,15 \text{ cm}^2$

Abbildung E.8: Histogramme und Erwartungswerte der Posituredifferenz $\Delta_{\underline{P}}$ für 1000 zufällige Eingangsposituren $\underline{P}_{\text{in}}$ und die jeweils bezüglich $\Delta_{\underline{F}}$ ähnlichste Datenbankpositur $\underline{P}_{\text{opt}, \Delta_{\underline{F}}} \in \mathcal{P}^{\text{db}}$ bei den Rauschintensitäten $N = 7500$ (a) und $N = 10000$ (b).

E.3 Optimale Positur bezüglich $\Delta_{\underline{F}}$ bei indexbasierter Suche

In der praktischen Anwendung des Systems erfolgt der Zugriff auf die Positurrendatenbank mit Hilfe des in Abschnitt 5.1.4 erläuterten indexbasierten Verfahrens. Dabei hängen Antwortzeit und Qualität der zurückgelieferten Posituren von den zur Indexierung verwendeten Merkmalen sowie dem Verzweigungsgrad des Indexbaums ab. Nachfolgend wird die Variation und Optimierung dieser Parameter erläutert. Vereinfachend ist jeder Indexbaum auf allen Ebenen i gleich stark verzweigt, d. h. $n_i = n$ für alle i .

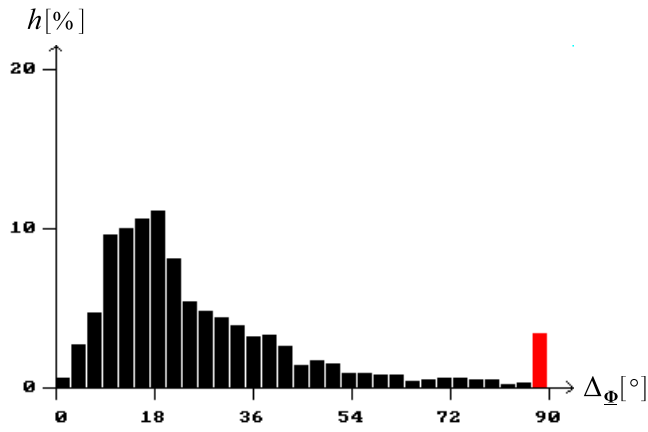
Indexmerkmale

Hinsichtlich der Auswahl der Indexmerkmale f_i werden drei verschiedene Varianten evaluiert, die in Tabelle E.1 aufgelistet sind. Der Verzweigungsgrad beträgt konstant $n = 6$. Die Höhe h des Indexbaums wird auf 9 beschränkt, da der benötigte Speicherplatz exponentiell ansteigt und bereits für $h = 10$ etwa 70% (ca. 200 MB) der Positurdaten ausmacht.

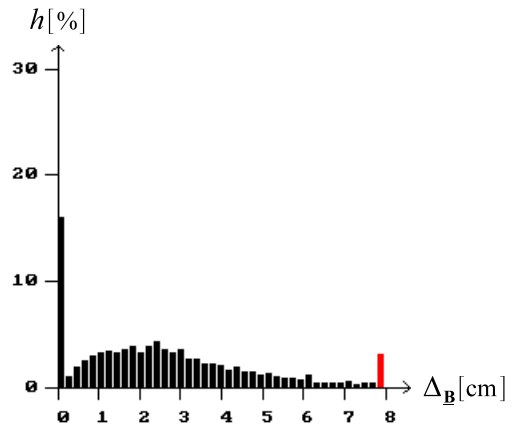
Tabelle E.1: Varianten der Datenbankindexierung mit unterschiedlichen Indexmerkmalen f_i und gleichem Verzweigungsgrad $n = 6$. Die in Variante 3 verwendeten Flächen a'_1 und a'_6 entsprechen dem Daumen- bzw. dem Handrückenmarker.

Variante	Indexmerkmale	h
1 (Fläche)	$\langle a'_1, \dots, a'_6 \rangle$	7
2 (Radius)	$\langle r'_1, \dots, r'_6 \rangle$	7
3 (Radius und Fläche)	$\langle r'_1, \dots, r'_6, a'_1, a'_6 \rangle$	9

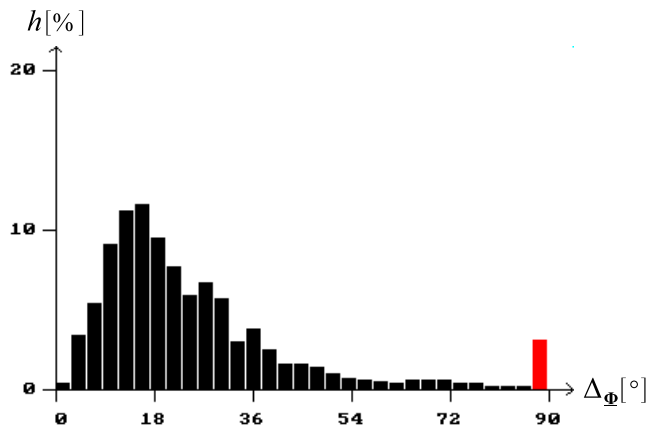
In den Abbildungen E.9 und E.10 ist die jeweils ermittelte Rekonstruktionsgenauigkeit für die Eingangsposituren P_{in} bei konstantem Rauschen mit $N = 7500$ angegeben. Der Vergleich der ersten beiden Varianten zeigt, dass r' zur Indexierung besser geeignet ist als a' . Eine weitere, jedoch geringe Verbesserung ergibt sich in Variante 3 durch die Hinzunahme von a'_1 und a'_6 . Bei Verwendung der zweiten oder der dritten Variante erreicht das Verfahren annähernd dieselbe Rekonstruktionsgenauigkeit wie die sequenzielle Suche (vgl. Abbildungen E.7a/b und E.8a), die mittlere Antwortzeit ist hingegen etwa um den Faktor 30 (Variante 2) bzw. 89 (Variante 3) kürzer und beträgt ca. 476 ms bzw. 162 ms. Dies gilt ebenfalls für $N \in \{0, 2500, 5000, 10000\}$. Damit ist die Effizienz des in Abschnitt 5.1.4 beschriebenen Zugriffsverfahrens nachgewiesen.



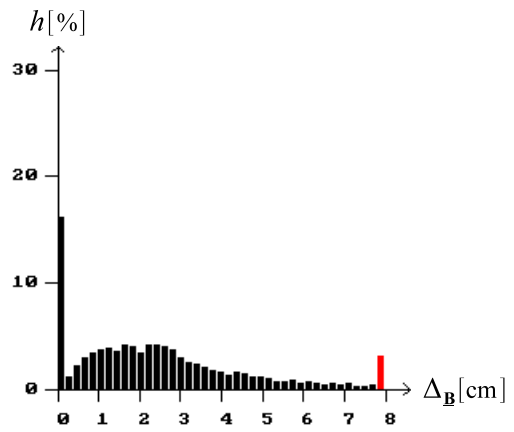
(a) $\Delta_{\underline{\Phi}}$ für Variante 1. $E(\Delta_{\underline{\Phi}}) \approx 29,08^\circ$



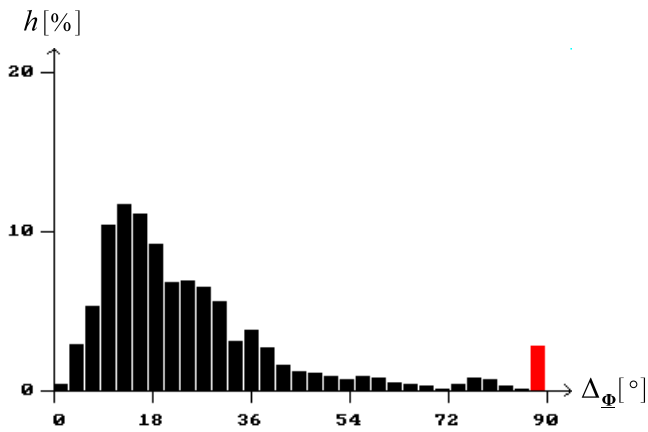
(b) $\Delta_{\underline{B}}$ für Variante 1. $E(\Delta_{\underline{B}}) \approx 2,75 \text{ cm}$



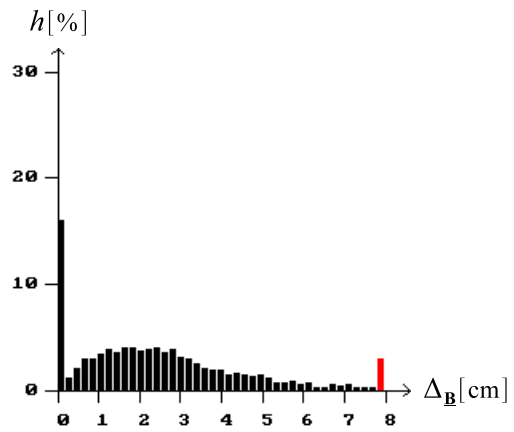
(c) $\Delta_{\underline{\Phi}}$ für Variante 2. $E(\Delta_{\underline{\Phi}}) \approx 27,25^\circ$



(d) $\Delta_{\underline{B}}$ für Variante 2. $E(\Delta_{\underline{B}}) \approx 2,62 \text{ cm}$

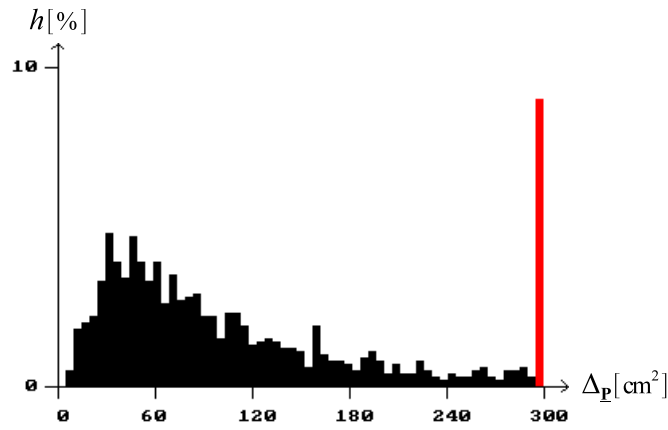


(e) $\Delta_{\underline{\Phi}}$ für Variante 3. $E(\Delta_{\underline{\Phi}}) \approx 26,88^\circ$

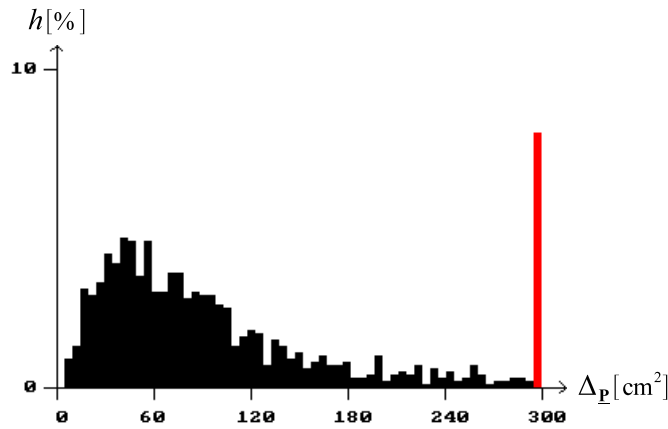


(f) $\Delta_{\underline{B}}$ für Variante 3. $E(\Delta_{\underline{B}}) \approx 2,64 \text{ cm}$

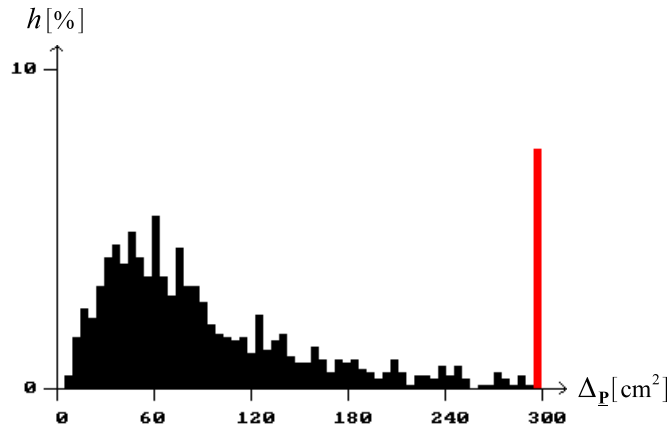
Abbildung E.9: Differenzmaße $\Delta_{\underline{\Phi}}$ und $\Delta_{\underline{B}}$ für die in Tabelle E.1 aufgeführten Varianten der Datenbankindexierung. Die Rauschintensität beträgt $N = 7500$.



(a) $\Delta_{\underline{P}}$ für Variante 1. $E(\Delta_{\underline{P}}) \approx 133,18 \text{ cm}^2$



(b) $\Delta_{\underline{P}}$ für Variante 2. $E(\Delta_{\underline{P}}) \approx 122,56 \text{ cm}^2$



(c) $\Delta_{\underline{P}}$ für Variante 3. $E(\Delta_{\underline{P}}) \approx 120,60 \text{ cm}^2$

Abbildung E.10: Posituredifferenz $\Delta_{\underline{P}}$ für die in Tabelle E.1 aufgeführten Varianten der Datenbankindexierung. Die Rauschintensität beträgt $N = 7500$.

Verzweigungsgrad

Der Verzweigungsgrad n des Baums bestimmt die Anzahl der vorselektierten Posituren $\tilde{\mathbf{P}}$, für die $\Delta_{\mathbf{F}}$ zu berechnen ist. Da einerseits die Güte der resultierenden Positurkandidaten $\hat{\mathbf{P}}$ mit $|\tilde{\mathbf{P}}|$ zunimmt, andererseits jedoch die Antwortzeit des Systems etwa proportional zu $|\tilde{\mathbf{P}}|$ ist, muss ein Kompromiss zwischen Genauigkeit und Geschwindigkeit gefunden werden.

In Tabelle E.2 sind Zugriffszeiten sowie Erwartungswerte für $\Delta_{\mathbf{P}}$, Δ_{Φ} und $\Delta_{\mathbf{B}}$ für die Indexmerkmale $\langle r'_1, \dots, r'_6, a'_1, a'_6 \rangle$ (Variante 3 in Tabelle E.1) und verschiedene Verzweigungsgrade n angegeben.² Die Ergebnisse zeigen, dass Verzweigungsgrade bis zu $n = 7$ und damit Zugriffszeiten von etwa 60 ms möglich sind, ohne die Rekonstruktionsgenauigkeit merklich zu beeinträchtigen. Daher kommt dieser Wert in der vorliegenden Arbeit global zum Einsatz.

Tabelle E.2: Rekonstruktionsgenauigkeit und Antwortzeit für unterschiedliche Verzweigungsgrade n und die Indexmerkmale $\langle r'_1, \dots, r'_6, a'_1, a'_6 \rangle$ bei $N = 7500$. Der Anstieg der Antwortzeit für $n > 8$ erklärt sich durch die Reduzierung der Indexintervallgrößen, die zu häufigerem Verzweigen der Suche führt. Die Werte für $n = 6$ entsprechen Abbildung E.9e/f bzw. Abbildung E.10c.

n	$E(\Delta_{\Phi})$ [°]	$E(\Delta_{\mathbf{B}})$ [cm]	$E(\Delta_{\mathbf{P}})$ [cm ²]	Antwortzeit [ms]
5	26,85	2,64	120,07	950
6	26,88	2,64	120,60	162
7	27,66	2,64	124,63	61
8	28,93	2,73	131,62	34
9	29,15	2,77	134,01	85

E.4 Positurrensequenzen

In Abbildung E.11 sind die mittleren Handpositurabweichungen für 1000 Zufallssequenzen angegeben, die unter Verwendung der in Tabelle D.1 aufgelisteten Parameter sowie einer Rauschintensität von $N = 7500$ generiert wurden. Die Parameter der Erkennungsstufe ergeben sich durch manuelle Optimierung und sind in Tabelle E.3 zusammengefasst. Bei der Verarbeitung realer Posituransichten kommen dieselben Werte zur Anwendung (vgl. Abschnitt 7.2.3).

²Da die Abweichungen analog zur sequenziellen Suche verteilt sind, wird auf die Darstellung der Histogramme verzichtet.

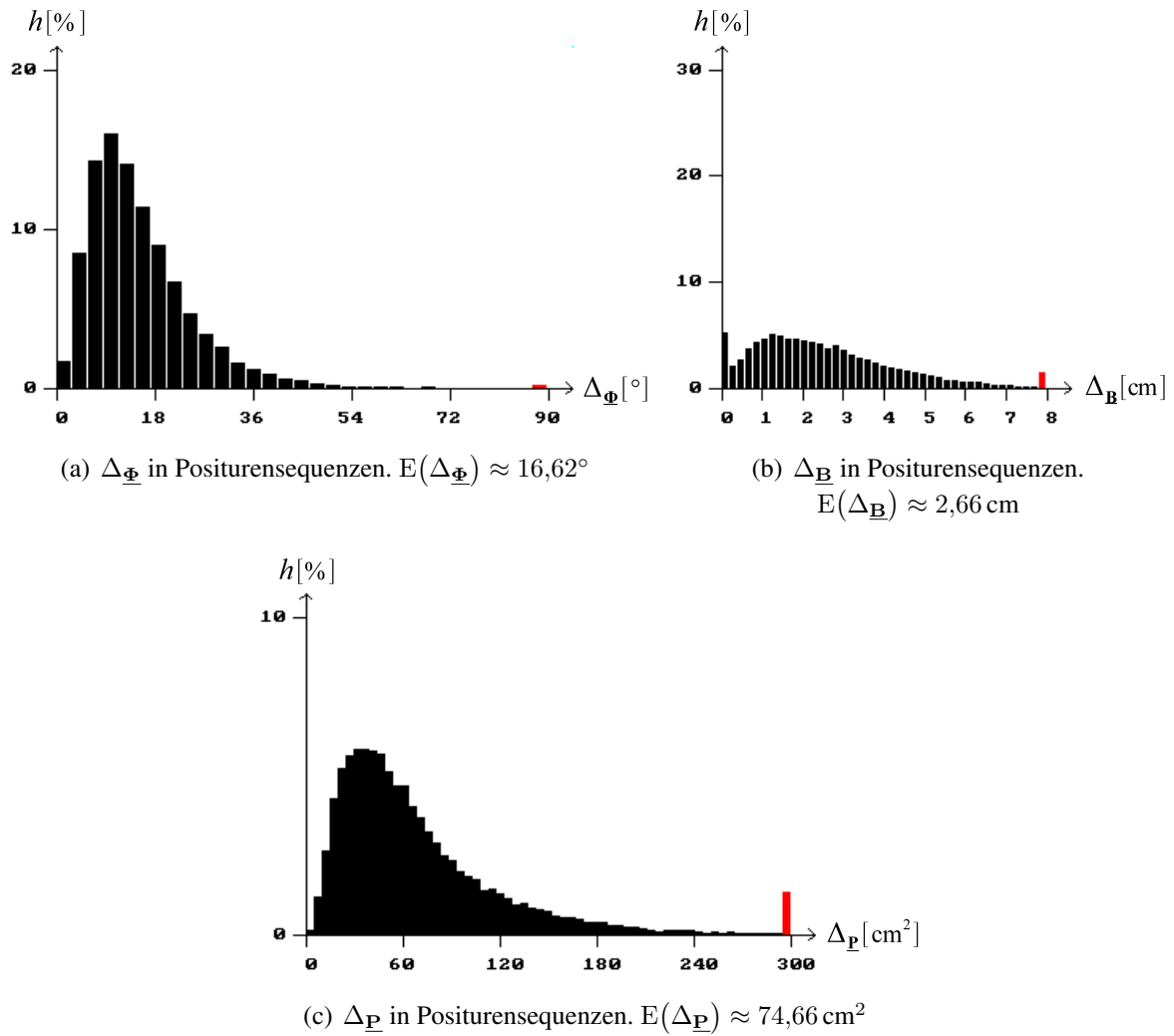


Abbildung E.11: Rekonstruktionsgenauigkeit in 1000 zufälligen Positurssequenzen. Die Rauschintensität beträgt $N = 7500$.

Tabelle E.3: Manuell optimierte Parameter der Algorithmen zur Handpositurerkennung. Die angegebenen Werte werden sowohl für synthetische als auch für reale Eingangsdaten verwendet.

Parameter	Beschreibung	Wert
k	Anzahl der Positurkandidaten pro Einzelbild (vgl. Abschnitt 5.3)	5
$\Delta_{\underline{P}}^*$	maximale Posituredifferenz ohne Einfluss auf P_{dyn} (vgl. (5.12))	30 cm^2
β	Gewichtung P_{stat} vs. P_{dyn} (vgl. Abschnitt 5.3.1.2)	0,999

Abbildungsverzeichnis

1.1	Einzelbilder der Gebärde „Olympia“	2
1.2	Mehrdeutige Handposituren	4
1.3	Anwendungsszenario für einen interaktiven Gebärdensprachtutor	5
1.4	Ausführung einer Gebärde mit farbig markierten Stoffhandschuhen	5
1.5	3D-Handmodell zur Erkennung der Handpositur	6
1.6	Schematische Darstellung des Verfahrens zur Handpositurerkennung	7
2.1	Die DGS-Gebärde „Buch“	12
2.2	Ausdrucksmittel in der Gebärdensprache	13
2.3	Die Gebärden „essen“ und „rot“	14
2.4	Die Gebärden „du“ und „heute“	14
2.5	Der Gebärdenraum	15
2.6	Die Gebärden „fragen“ und „Hunger“	15
2.7	Trajektorien der Gebärde „bookings“	16
2.8	Die Gebärdenphasen Vorbereitung, Ausführung und Rückführung	17
2.9	Die Geste und DGS-Gebärde „Uhrzeit“	18
2.10	Schematische Übersicht der Verfahren zur videobasierten Erfassung manueller Ausdrucksmittel	19
2.11	Handpositurrekonstruktion mittels inverser Kinematik	21
2.12	Handpositurrekonstruktion mittels Modellanpassung	22
2.13	Für die Hand verfügbare Bildauflösung	47
2.14	Veranschaulichung der geringen Aussagekraft der Handkontur	48
2.15	Auswirkung von Bewegungsunschärfe auf die unmarkierte Hand	49
2.16	Beispiel eines farbig markierten Stoffhandschuhs	50

2.17	Auswirkung von Bewegungsunschärfe auf die markierte Hand	50
2.18	Mehrdeutigkeiten bei der farbbasierten Markerdetektion	51
2.19	Markierung der nichtdominanten Hand	56
2.20	Verarbeitungskette zur Rekonstruktion einer Positurresequenz	58
2.21	Vergleich von Benutzer- und Referenzgebärde	58
3.1	Röntgenaufnahme einer menschlichen Hand	62
3.2	Schematische Darstellung des entwickelten Handmodells	64
3.3	Palmaransicht des Handmodells bei maximaler Abduktion der Finger	66
3.4	Modellierung einer Markergeometrie	67
3.5	Wertebereich des Beugungswinkels des DIP-Gelenks	68
3.6	Verschiedene Beugungswinkel in Palmar- und Seitenansicht	69
3.7	Verschiedene Breitengrade bei $\phi_{\text{lon}} = \phi_{\text{rot}} = 0^\circ$	69
3.8	Verschiedene Längengrade bei $\phi_{\text{lat}} = \phi_{\text{rot}} = 0^\circ$	70
3.9	Verschiedene Rotationen bei $\phi_{\text{lat}} = -36^\circ$ und $\phi_{\text{lon}} = 144^\circ$	70
3.10	Anatomisch unmögliche Gelenk- und Handstellungen	72
3.11	Handkoordinatensystem mit Ursprung im Mittelpunkt des Handtellers	72
3.12	Lage der Referenzpunkte zum Vergleich zweier Posituren	73
3.13	Mehrdeutigkeit hinsichtlich des Beugungswinkels in PIP- und DIP-Gelenken	75
3.14	Berechnung der Handformdifferenz $\Delta_{\underline{\mathbf{B}}}$	77
3.15	Berechnung der Handstellungsdifferenz $\Delta_{\underline{\Phi}}$	78
3.16	Berechnung der Posituredifferenz $\Delta_{\underline{\mathbf{P}}}$	79
3.17	Vergleich ähnlicher Posituren	80
4.1	Schematische Darstellung von Segmentierung und Merkmalsextraktion . . .	82
4.2	Hintergrundfarbmodell eines einzelnen Pixels	88
4.3	Hintergrundfarbmodell mit Schattenerkennung	89
4.4	Anwendung des unimodalen Hintergrundmodells	89
4.5	Diskrete Verteilung und kontinuierliche Approximation mittels Gaußkernel .	92
4.6	Anwendung des Mean-Shift-Segmentierungsalgorithmus auf ein Farbbild . .	93
4.7	Manuelle Maskierung der Markerregionen in einer Kalibrierungssequenz . .	94

4.8	4er- und 8er-Nachbarschaft	95
4.9	Repräsentation einer Bildregion als Liste ihrer Randpixel sowie als Polygon	96
4.10	Parametrische Beschreibung einer Ellipse	100
4.11	Bestimmung der Ellipsenradien	100
4.12	Randkurven der Markerregionen und approximierende Ellipsen	101
4.13	Scheitelpunkt und Breite der Gesichtsregion	101
4.14	Mehrdeutigkeiten bei der Detektion der farbigen Markierungen	102
4.15	Trellisdiagramm des Hypothesenraums	104
4.16	Sigmoidfunktion zur Berechnung statistischer Wahrscheinlichkeiten	108
4.17	Berechnung der 2D-Merkmale	112
4.18	Koordinatensystem zur Angabe der Handposition	114
5.1	Verteilung der Ansichtspunkte auf der Ansichtskugel	117
5.2	Ansicht der in der Posituredatenbank verwendeten Daumenstellungen	118
5.3	Synthetische Posituransicht bei aktivierter bzw. deaktivierter Beleuchtung	119
5.4	Diskretisierung der Ellipsenrandkurve mit acht Punkten	120
5.5	Eingangspositur und Modellpositur sowie entsprechende Merkmale	124
5.6	Berechnung der Kovarianzmatrix Σ für die parametrische Merkmalsdifferenz	125
5.7	Hausdorff-Distanz zweier Ellipsen A und B	126
5.8	Anwendung der Hausdorff-Distanz zum Vergleich zweier Merkmalsvektoren	127
5.9	Vergleich zweier Posituren bei einseitig fehlender Markierung	128
5.10	Merkmalsdifferenz bei einseitig fehlender Markierung in den Datenbank- merkmalen	129
5.11	Merkmalsdifferenz bei einseitig fehlender Markierung in den extrahierten Merkmalen	130
5.12	Glättung der Ansichtswinkel ϕ_{lat} , ϕ_{lon} und ϕ_{rot}	133
5.13	Glättung der Beugungsgrade durch Interpolation mit kubischen Splines	134
6.1	DTW-Matrix	138
6.2	Mögliche Schritte auf dem Pfad durch die DTW-Matrix	138
6.3	Berechnung eines Pfades durch die DTW-Matrix	140
6.4	Ablaufdiagramm des Gebärdenvergleichs	141

6.5	Die Gebärde „Bildtelefon“, aufgezeichnet mit 25 Hz	143
6.6	Die Gebärde „gut“, aufgezeichnet mit 25 Hz	144
6.7	Korrektives Feedback bei abweichender Handposition	144
6.8	Korrektives Feedback bei abweichender Ausführungsgeschwindigkeit	145
6.9	Korrektives Feedback bei abweichender Handstellung	145
6.10	Bewegung des Handmodells von der abweichenden in die Referenzstellung	146
6.11	Korrektives Feedback bei abweichender Handform	147
6.12	Bewegung des Handmodells von der abweichenden in die Referenzform	148
6.13	Feedback bei vollständig korrekter Gebärdenausführung	149
7.1	Zufällig generierte Einzelposituren	152
7.2	Zufällig generierte Positurrensequenz	153
7.3	Auswirkung des Rauschprozesses bei verschiedenen Iterationsanzahlen N	153
7.4	Testposituren aus P_{in} und Erkennungsergebnisse	157
7.5	Vergleich unterschiedlicher Markergeometrien	158
7.6	Anpassung der Handmodellgeometrie an den Gebärdensprecher	160
7.7	Eingabe der Positurparameter	161
7.8	Varianz bei der manuellen Eingabe der Positurparameter	162
7.9	Beispielgebärden aus dem Testkorpus	164
7.10	Rekonstruktionsgenauigkeit in Gebärdens	165
7.11	Erkennungsbeispiele mit sehr niedrigem Rekonstruktionsfehler	168
7.12	Erkennungsbeispiel mit durchschnittlichem Rekonstruktionsfehler	169
7.13	Erkennungsbeispiel mit durchschnittlichem Rekonstruktionsfehler	170
7.14	Rekonstruktionsfehler durch nicht auflösbare Mehrdeutigkeiten	171
7.15	Mehrdeutige Merkmale in der Gebärde „Hörgerät“	171
7.16	Rekonstruktionsfehler durch Segmentierungsfehler	172
7.17	Fehlerhafte Segmentierung der Markerregionen in der Gebärde „Hals“	172
7.18	Kandidatenpositur und -merkmale in der Gebärde „Hals“	173
7.19	Ungenauere Rekonstruktion bei nicht modellierbaren Posituren	173
7.20	Angabe der Handformabweichung in Form einer Beugungsgraddifferenz	175

A.1	Schätzung der Verteilungsparameter eines multimodalen Hintergrundmodells	184
B.1	Nichtrekursive Implementierung des Viterbi-Algorithmus	187
D.1	Fehler bei der Mean-Shift-Segmentierung	194
D.2	Simulation des Bildrauschens in der synthetischen Posituransicht	195
D.3	Mittelwert und Standardabweichung des Rauschprozesses im Merkmalsraum	195
D.4	Verschiebung des Markerschwerpunkts durch den Rauschprozess	196
E.1	Histogramme der Differenzmaße Δ_{Φ} , $\Delta_{\underline{B}}$ und $\Delta_{\underline{P}}$ der bezüglich $\Delta_{\underline{P}}$ optimalen Posituren	198
E.2	Histogramme der Differenzmaße $\Delta_{\underline{B}D}$ bis $\Delta_{\underline{B}KF}$ der bezüglich $\Delta_{\underline{P}}$ optimalen Posituren	199
E.3	Histogramme der Differenzmaße Δ_{Φ} , $\Delta_{\underline{B}}$ und $\Delta_{\underline{P}}$ der bezüglich $\Delta_{\underline{F}}$ optimalen Datenbankposituren (sequenzielle Suche)	200
E.4	Histogramme der Differenzmaße $\Delta_{\underline{B}D}$ bis $\Delta_{\underline{B}KF}$ der bezüglich $\Delta_{\underline{F}}$ optimalen Datenbankposituren (sequenzielle Suche)	201
E.5	Histogramme der Differenzmaße Δ_{Φ} und $\Delta_{\underline{B}}$ der bezüglich $\Delta_{\underline{F}}$ optimalen Datenbankposituren (indexbasierte Suche)	202
E.6	Histogramme der Posituredifferenz $\Delta_{\underline{P}}$ der bezüglich $\Delta_{\underline{F}}$ optimalen Datenbankposituren (indexbasierte Suche)	203
E.7	Histogramme der Differenzmaße Δ_{Φ} und $\Delta_{\underline{B}}$ der bezüglich $\Delta_{\underline{F}}$ optimalen Datenbankposituren (indexbasierte Suche)	204
E.8	Histogramme der Posituredifferenz $\Delta_{\underline{P}}$ der bezüglich $\Delta_{\underline{F}}$ optimalen Datenbankposituren (indexbasierte Suche)	205
E.9	Differenzmaße Δ_{Φ} und $\Delta_{\underline{B}}$ für verschiedene Varianten der Datenbankindexierung	207
E.10	Posituredifferenz $\Delta_{\underline{P}}$ für verschiedene Varianten der Datenbankindexierung	208
E.11	Rekonstruktionsgenauigkeit in 1000 zufälligen Positurensequenzen	210

Tabellenverzeichnis

2.1	Wichtige Randbedingungen bei der visuellen Handpositurrekonstruktion . . .	24
2.2	Existierende Ansätze zur Handpositurerkennung	27
2.3	Übersicht der vorgestellten Systeme zur Handpositurrekonstruktion	43
3.1	Gelenke des Handmodells und deren Freiheitsgrade	65
3.2	Elemente des Handmodells, die die Handgeometrie bestimmen	65
3.3	Differenzmaße der verglichenen Posituren	77
3.4	Differenzmaße der verglichenen Posituren	78
3.5	Differenzmaße der verglichenen Posituren	79
3.6	Differenzmaße der verglichenen Posituren	80
4.1	Abhängigkeit der geometrischen Merkmale von Translation, Rotation und Skalierung	99
5.1	Beugungswinkel der Finger und des Daumens	118
5.2	Struktur eines Datenbankeintrags	121
7.1	Differenzmaße der in Abbildung 7.4 gezeigten Posituren	156
7.2	Rekonstruktionsgenauigkeit für unterschiedliche Markergeometrien	158
7.3	Anpassung der Handmodellgeometrie an den Gebärdensprecher	160
7.4	Varianz bei der manuellen Eingabe der Positurparameter	163
7.5	Rekonstruktionsgenauigkeit in Gebärden	166
7.6	Geschwindigkeit der Verarbeitungsschritte zur Positurrekonstruktion	167
C.1	Parameter zur Hypothesenbewertung	189
C.2	Geänderte Parameter zur Hypothesenbewertung	190

D.1	Parameter bei der Generierung zufälliger Posituren und Positurensequenzen .	193
E.1	Varianten der Datenbankindexierung	206
E.2	Rekonstruktionsgenauigkeit und Antwortzeit für unterschiedliche Verzweigungsgrade	209
E.3	Manuell optimierte Parameter der Algorithmen zur Handpositurerkennung . .	210

Literaturverzeichnis

- [AG97] ASSAN, M. und K. GROBEL: *Video-Based Sign Language Recognition Using Hidden Markov Models*. In: *Proceedings of the International Gesture Workshop*, 1997.
- [AKA05] ARAN, O., C. KESKIN und L. AKARUN: *Sign Language Tutoring Tool*. In: *EUSIPCO*, 2005.
- [Aky03] AKYOL, S.: *Nicht-intrusive Erkennung isolierter Gesten und Gebärden*. Doktorarbeit, Lehrstuhl für Technische Informatik, RWTH Aachen, 2003.
- [Alv04] ALVARADO MOYA, J. P.: *Segmentation of Color Images for Interactive 3D Object Retrieval*. Doktorarbeit, Lehrstuhl für Technische Informatik, RWTH Aachen, 2004.
- [AS01] ATHITSOS, V. und S. SCLAROFF: *3D Hand Pose Estimation by Finding Appearance-Based Matches in a Large Database of Training Views*. Technischer Bericht BU-CS-TR-2001-021, Computer Science Department, Boston University, 2001.
- [AS02a] ATHITSOS, V. und S. SCLAROFF: *An Appearance-Based Framework for 3D Hand Shape Classification and Camera Viewpoint Estimation*. Technischer Bericht BU-CS-TR-2001-022, Computer Science Department, Boston University, 2002.
- [AS02b] ATHITSOS, V. und S. SCLAROFF: *An Appearance-Based Framework for 3D Hand Shape Classification and Camera Viewpoint Estimation*. In: *IEEE Conference on Automatic Face and Gesture Recognition*, Mai 2002.
- [AS03a] ATHITSOS, V. und S. SCLAROFF: *Database Indexing Methods for 3D Hand Pose Estimation*. In: *Gesture Workshop*, 2003.
- [AS03b] ATHITSOS, V. und S. SCLAROFF: *Estimating 3D Hand Pose from a Cluttered Image*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Juni 2003.

- [AZ02] AKYOL, S. und J. ZIEREN: *Evaluation of ASM Head Tracker for Robustness against Occlusion*. In: ARABNIA, H. R. und Y. MUN (Herausgeber): *Proceedings of the International Conference on Imaging Science, Systems, and Technology*, Band 1, Seiten 28–34, Juni 2002.
- [Bal04] BALLÉ, J.: *Implementierung und Evaluierung eines multimodalen Hintergrundmodells als Vorstufe eines Objektverfolgungssystems*. Studienarbeit, Lehrstuhl für Technische Informatik, RWTH Aachen, 2004.
- [Bau04] BAUER, B.: *Erkennung kontinuierlicher Gebärdensprache mit Untereinheiten-Modellen*. Doktorarbeit, Lehrstuhl für Technische Informatik, RWTH Aachen, 2004.
- [Bec97] BECKER, C.: *Zur Struktur der Deutschen Gebärdensprache*. Wissenschaftlicher Verlag Trier, 1997.
- [BHE⁺02] BEBIS, G., F. HARRIS, A. EROL, B. YI, J. MARTINEZ, J. HERNANDEZ-USABIAGA und S. FRITZINGER: *Development of a Nationally Competitive Program in Computer Vision Technologies for Effective Human-Computer Interaction in Virtual Environments*. Technischer Bericht, BioVIS Technology Center of NASA Ames Research Center, 2002.
- [Bou85] BOURGAIN, J.: *On Lipschitz embeddings of finite metric spaces in Hilbert space*. Israel Journal of Mathematics, 52:46–52, 1985.
- [Boy95] BOYES BRAEM, P.: *Einführung in die Gebärdensprache und ihre Erforschung*. Signum Press, Hamburg, 1995.
- [Bro98] BROOKNER, E.: *Tracking and Kalman filtering made easy*. John Wiley & Sons, 1998.
- [BS87] BAR-SHALOM, Y.: *Tracking and Data Association*. Academic Press Professional, 1987.
- [BS91] BRONSTEIN, I. N. und K. A. SEMENDJAJEW: *Taschenbuch der Mathematik*. Harri Deutsch, 25. Auflage, 1991.
- [BTBW77] BARROW, H. G., J. M. TENENBAUM, R. C. BOLLES und H. C. WOLF: *Parametric correspondence and chamfer matching: Two new techniques for image matching*. In: *IJCAI*, Seiten 659–663, 1977.
- [CGH02] CHUA, C.-S., H. GUAN und Y.-K. HO: *Model-based 3D hand posture estimation from a single 2D image*. International Journal of Image and Vision Computing, 20(3), 2002.
- [CM02] COMANICIU, D. und P. MEER: *Mean shift: A robust approach toward feature space analysis*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Band 24, Seiten 603–619, Mai 2002.

- [CT98] CUTLER, R. und M. TURK: *View-based Interpretation of Real-time Optical Flow for Gesture Recognition*. In: *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, Seiten 416–421, Nara, Japan, 1998.
- [CTCG95] COOTES, T. F., C. J. TAYLOR, D. H. COOPER und J. GRAHAM: *Active Shape Models: Their Training and Application*. *Computer Vision and Image Understanding*, 61(1):38–59, Januar 1995.
- [Der04] DERPANIS, K. G.: *A Review of Vision-Based Hand Gestures*. Technischer Bericht, Department of Computer Science, York University, 2004.
- [Dic06] DICK, T.: *Entwicklung eines interaktiven Gebärdensprachtutors mit video-basierter Lernkontrolle*. Diplomarbeit, Lehrstuhl für Technische Informatik, RWTH Aachen, 2006.
- [DZK06] DICK, T., J. ZIEREN und K.-F. KRAISS: *Visual Hand Posture Recognition in Monocular Image Sequences*. In: FRANKE, K., K.-R. MÜLLER, B. NICKOLAY und R. SCHÄFER (Herausgeber): *Proceedings of the 28th DAGM Symposium*, Band 4174 der Reihe *Lecture Notes in Computer Science*. Springer, 2006.
- [FAK03] FILLBRANDT, H., S. AKYOL und K.-F. KRAISS: *Extraction of 3D Hand Shape and Posture from Image Sequences for Sign Language Recognition*. In: AZADA, D. (Herausgeber): *IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, 2003.
- [Fil03] FILLBRANDT, H.: *Rekonstruktion von Handformen aus monokularen Bildfolgen für die Erkennung von Gebärdensprache*. Diplomarbeit, Lehrstuhl für Technische Informatik, RWTH Aachen, 2003.
- [GCH99] GUAN, H. Y., C. S. CHUA und Y. K. HO: *Hand Posture Estimation from 2D Monocular Image*. In: *International Conference on 3-D Imaging and Modeling*, 1999.
- [GCH01] GUAN, H., C.-S. CHUA und Y.-K. HO: *3D hand pose retrieval from a single 2D image*. In: *International Conference on Image Processing*, 2001.
- [Gol97] GOLDSTEIN, E. B.: *Wahrnehmungspsychologie*. Spektrum Akademischer Verlag, 1997.
- [HH96a] HEAP, T. und D. HOGG: *3D Deformable Hand Models*. In: *Gesture Workshop*, Seiten 131–139, 1996.
- [HH96b] HEAP, T. und D. HOGG: *Towards 3D Hand Tracking using a Deformable Model*. In: *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*. IEEE Computer Society, 1996.

- [HLB⁺05] HENDERSON, V., S. LEE, H. BRASHEAR, H. HAMILTON, T. STARNER und S. HAMILTON: *Development of an American Sign Language Game for Deaf Children*. In: *Proceeding of the 2005 Conference on Interaction Design and Children*, Juni 2005.
- [HOR99a] HOLDEN, E.-J., R. OWENS und G. G. ROY: *3D Hand Tracker for Visual Sign Recognition*. Technischer Bericht 99/2, Department of Computer Science, University of Western Australia, 1999.
- [HOR99b] HOLDEN, E.-J., R. OWENS und G. G. ROY: *Adaptive Fuzzy Expert System for Sign Recognition*. Technischer Bericht 99/3, Department of Computer Science, University of Western Australia, 1999.
- [Hu62] HU, M.-K.: *Visual Pattern Recognition by Moment Invariants*. IRE Transactions on Information Theory, IT-8:179–187, 1962.
- [ISS04] IMAI, A., N. SHIMADA und Y. SHIRAI: *3-D Hand Posture Recognition by Training Contour Variation*. In: *International Conference on Automatic Face and Gesture Recognition*, 2004.
- [Jai89] JAIN, A. K.: *Fundamentals of Digital Image Processing*. Prentice-Hall, 1989.
- [Jel97] JELINEK, F.: *Statistical Methods for Speech Recognition*. MIT Press, 1997.
- [JR98] JONES, M. J. und J. M. REHG: *Statistical Color Models with Application to Skin Detection*. Technischer Bericht CRL 98/11, Compaq Cambridge Research Lab, Dezember 1998.
- [Kal60] KALMAN, R. E.: *A New Approach to Linear Filtering and Prediction Problems*. Transactions of the ASME–Journal of Basic Engineering, 82(Series D):35–45, 1960.
- [Kap84] KAPANDJI, I. A.: *Funktionelle Anatomie der Gelenke*. Enke Verlag Stuttgart, 1984.
- [KB03] KAEWTRAKULPONG, P. und R. BOWDEN: *A real time adaptive visual surveillance system for tracking low-resolution colour targets in dynamically changing scenes*. IVC, 21:913–929, 2003.
- [Ker01] KERRES, M.: *Multimediale und telemediale Lernumgebungen. Konzeption und Entwicklung*. Oldenbourg Verlag, 2001.
- [Kes02] *777 Gebärden*. DVD-ROM, Verlag Karin Kestner, 2002.
- [Lan04] *Der Englisch-Kurs*. CD-ROM, Langenscheidt Fachverlag, 2004.
- [LH99] LATHUILLIÈRE, F. und J.-Y. HERVÉ: *Visual Tracking of Hand Posture in a Robot Control Application*. In: *Proceedings of the Vision Interface Conference*, 1999.

- [LH00] LATHUILIÈRE, F. und J.-Y. HERVÉ: *Visual Tracking of Hand Posture with Occlusion Handling*. In: *Proceedings of the 15th International Conference on Pattern Recognition Volume 3*, 2000.
- [LHH⁺05] LEE, S., V. HENDERSON, H. HAMILTON, H. BRASHEAR, T. STARNER und S. HAMILTON: *A Gesture-Based American Sign Language Game for Deaf Children*. In: *Extended Abstracts of the Conference for Human-Computer Interaction*, Seiten 1589–1592. ACM Press, April 2005.
- [Lid84] LIDDELL, S.: *Think and believe: Sequentiality in American Sign Language*. *Language*, 60(2):372–399, 1984.
- [LJ89] LIDDELL, S. und R. JOHNSON: *American Sign Language: The Phonological Base*. *Sign Language Studies*, Seiten 195–277, 1989.
- [LK95] LEE, J. und T. L. KUNII: *Model-Based Analysis of Hand Posture*. *IEEE Computer Graphics and Applications*, 15(5):77–86, 1995.
- [LMSO03] LU, S., D. METAXAS, D. SAMARAS und J. OLIENSIS: *Using Multiple Cues for Hand Tracking and Model Refinement*. In: *CVPR*, 2003.
- [Low91] LOWE, D. G.: *Fitting Parameterized Three-Dimensional Models to Images*. *IEEE TPAMI*, 13(5):441–450, Mai 1991.
- [McC00] MCCANE, B.: *Snakes and Spiders*. In: *International Conference on Pattern Recognition*, 2000.
- [Met97] METAXAS, D. N.: *Physics-Based Deformable Models – Applications to Computer Vision, Graphics and Medical Imaging*. Kluwer Academic Publishers, 1997.
- [MG98] MCKENNA, S. J. und S. GONG: *Gesture Recognition for Visually Mediated Interaction Using Probabilistic Event Trajectories*. In: *Proceedings of the British Machine Vision Conference*, Band 2, Seiten 498–508, Southampton, England, 1998.
- [MR81] MYERS, C. S. und L. R. RABINER: *A comparative study of several dynamic time-warping algorithms for connected word recognition*. *The Bell System Technical Journal*, 60(7):1389–1409, September 1981.
- [MT05] METAXAS, D. und G. TSECHPENAKIS: *Dynamic Data Driven Coupling of Continuous and Discrete Methods for 3D Tracking*. In: *ICCS*, 2005.
- [Nöl00] NÖLKER, C.: *Grefit: Ein System zur visuellen Erkennung von Handposturen*. Doktorarbeit, Technische Fakultät der Universität Bielefeld, 2000.
- [OH99] OUHADDI, H. und P. HORAIN: *Hand Tracking by 3D Model Registration*. In: *Actes du Colloque Scientifique International Réalité Virtuelle et Prototypage*, 1999.

- [OR05] ONG, S. C. W. und S. RANGANATH: *Automatic Sign Language Analysis: A Survey and the Future Beyond Lexical Meaning*. IEEE TPAMI, 27(6):873–8914, Juni 2005.
- [Pal99] PALMER, S.: *Vision Science: Photons to Phenomenology*. MIT Press, 1999.
- [PKB90] POIZNER, H., E. S. KLIMA und U. BELLUGI: *Was die Hände über das Gehirn aussagen. Neuropsychologische Aspekte der Gebärdensprachforschung*. Signum (Internationale Arbeiten zur Gebärdensprache und Kommunikation Gehörloser), 12, 1990.
- [RASS01] ROSALES, R., V. ATHITSOS, L. SIGAL und S. SCLAROFF: *3D Hand Pose Reconstruction Using Specialized Mappings*. In: *International Conference on Computer Vision (ICCV)*, 2001.
- [Reh95] REHG, J. M.: *Visual Analysis of High DOF Articulated Objects with Application to Hand Tracking*. Doktorarbeit, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1995.
- [RF99] ROBINSON, T. L. und A. J. FUGLEVAND: *Independence of finger movement in normal subjects and in concert-level pianists*. Society for Neuroscience Abstracts, 25:1149, 1999.
- [Rit93] RITTER, H.: *Parameterized Self-Organizing Maps*. In: *Proceedings of the International Conference on ANN*, 1993.
- [RJ86] RABINER, L. und B.-H. JUANG: *An Introduction to Hidden Markov Models*. IEEE ASSP Magazine, 3(1):4–16, 1986.
- [RK93] REHG, J. und T. KANADE: *DigitEyes: Vision-Based Human Hand Tracking*. Technischer Bericht CMU-CS-93-220, School of Computer Science, Carnegie Mellon University, 1993.
- [RK94] REHG, J. M. und T. KANADE: *Visual Tracking of High DOF Articulated Structures: an Application to Human Hand Tracking*. In: *European Conference on Computer Vision*, 1994.
- [RK95] REHG, J. M. und T. KANADE: *Model-Based Tracking of Self-Occluding Articulated Objects*. In: *Proceedings of the 5th International Conference on Computer Vision*, 1995.
- [RKE98] RIGOLL, G., A. KOSMALA und S. EICKELER: *High Performance Real-Time Gesture Recognition Using Hidden Markov Models*. In: WACHSMUT, I. und M. FRÖHLICH (Herausgeber): *Gesture and Sign Language in Human-Computer Interaction*, Seiten 69–80. Springer, 1998.
- [Sch96] SCHÜRMANN, J.: *Pattern Classification*. John Wiley & Sons, 1996.

- [Sch05] SCHADEMANN, N.: *Entwicklung und Implementierung einer laufzeiteffizienten Suche in einer Handstellungsdatenbank zur 3D-Handstellungserkennung aus 2D-Bilddaten*. Studienarbeit, Lehrstuhl für Technische Informatik, RWTH Aachen, 2005.
- [SG99] STAUFFER, C. und W. E. L. GRIMSON: *Adaptive background mixture models for real-time tracking*. In: *Computer Vision and Pattern Recognition 1999*, Band 2, 1999.
- [SHB98] SONKA, M., V. HLAVAC und R. BOYLE: *Image Processing, Analysis and Machine Vision*. Brooks Cole, 1998.
- [SKS01] SHIMADA, N., K. KIMURA und Y. SHIRAI: *Real-Time 3-D Hand Posture Estimation Based on 2-D Appearance Retrieval Using Monocular Camera*. In: *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, 2001.
- [SKSK00] SHIMADA, N., K. KIMURA, Y. SHIRAI und Y. KUNO: *Hand Posture Estimation by Combining 2-D Appearance-Based and 3-D Model-Based Approaches*. In: *International Conference on Pattern Recognition*, 2000.
- [SMC01a] STENGER, B., P. R. S. MENDONCA und R. CIPOLLA: *Model-Based 3D Tracking of an Articulated Hand*. In: *CVPR*, Band II, Seiten 310–315, 2001.
- [SMC01b] STENGER, B., P. R. S. MENDONCA und R. CIPOLLA: *Model-Based Hand Tracking Using an Unscented Kalman Filter*. In: *BMVC*, 2001.
- [Sob93] SOBOTTA, J.: *Atlas der Anatomie des Menschen*. Urban und Schwarzenberg, 1993.
- [SS03] SHAMAIE, A. und A. SUTHERLAND: *Accurate Recognition of Large Number of Hand Gestures*. In: *Proceedings of the 2nd Iranian Conference on Machine Vision and Image Processing*, 2003.
- [ST02] SAGAWA, H. und M. TAKEUCHI: *A Teaching System of Japanese Sign Language Using Sign Language Recognition and Generation*. In: *Proceedings of the 10th ACM International Conference on Multimedia*, 2002.
- [Ste96] STEGER, C.: *On the Calculation of Arbitrary Moments of Polygons*. Technical Report FGBV-96-05, Forschungsgruppe Bildverstehen (FG BV), Informatik IX, Technische Universität München, Oktober 1996.
- [Ste01] STENGER, B.: *3D Model-Based Hand Tracking*. Technischer Bericht, University of Cambridge, Department of Engineering, 2001.
- [STTC06] STENGER, B., A. THAYANANTHAN, P. H. S. TORR und R. CIPOLLA: *Model-Based Hand Tracking Using a Hierarchical Bayesian Filter*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

- [Stu92] STURMAN, D. J.: *Whole-hand Input*. Doktorarbeit, School of Architecture and Planning, Massachusetts Institute of Technology, 1992.
- [TSS02] TANIBATA, N., N. SHIMADA und Y. SHIRAI: *Extraction of Hand Features for Recognition of Sign Language Words*. In: *Proceedings of the International Conference on Vision Interface*, Seiten 391–398, 2002.
- [TT97] THAM, K. und R. TEGNER: *Video Feedback in the Rehabilitation of Patients with Unilateral Neglect*. *Archives of Physical Medicine and Rehabilitation*, 78(4):410–413, April 1997.
- [TvdM02] TRIESCH, J. und C. VON DER MALSBERG: *Classification of Hand Postures Against Complex Backgrounds Using Elastic Graph Matching*. *International Journal of Image and Vision Computing*, 2002.
- [vASZK06] AGRIS, U. VON, D. SCHNEIDER, J. ZIEREN und K.-F. KRAISS: *Rapid Signer Adaptation for Isolated Sign Language Recognition*. In: *IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.
- [vAZC⁺07] AGRIS, U. VON, J. ZIEREN, U. CANZLER, B. BAUER und K.-F. KRAISS: *Recent Developments in Visual Sign Language Recognition*. *Journal on Universal Access in the Information Society*, 2007.
- [Vit02] VITTRUP, M.: *Pose Estimation by Applied Numerical Techniques*. Diplomarbeit, Department of Computer Science, University of Otago, 2002.
- [Vri97] VRIJ, R.: *Marketing Plan for the Sign Printing System*. Technischer Bericht 6.2, RNR, Amsterdam, 1997.
- [VSM02] VITTRUP, M., M. K. D. SØRENSEN und B. MCCANE: *Pose Estimation by Applied Numerical Techniques*. In: *Image and Vision Computing, New Zealand*, 2002.
- [WADP97] WREN, C., A. AZARBAYEJANI, T. DARRELL und A. PENTLAND: *Pfinder: Real-time tracking of the human body*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.
- [WB04] WELCH, G. und G. BISHOP: *An Introduction to the Kalman Filter*. Technischer Bericht TR 95-041, Department of Computer Science, University of North Carolina at Chapel Hill, 2004.
- [WLH01] WU, Y., J. Y. LIN und T. S. HUANG: *Capturing Natural Hand Articulation*. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2001.
- [ZC06] ZIEREN, J. und U. CANZLER: *Advanced Man-Machine Interaction*, Kapitel Non-Intrusive Acquisition of Human Action, Seiten 7–94. Springer, 2006.

- [ZCBK06] ZIEREN, J., U. CANZLER, B. BAUER und K.-F. KRAISS: *Advanced Man-Machine Interaction*, Kapitel Sign Language Recognition, Seiten 95–139. Springer, 2006.
- [ZK04] ZIEREN, J. und K.-F. KRAISS: *Non-Intrusive Sign Language Recognition for Human-Computer Interaction*. In: *Proceedings of the IFAC-HMS Symposium*, 2004. Submitted to the IFAC Symposium 2004, Atlanta, USA.
- [ZK05] ZIEREN, J. und K.-F. KRAISS: *Robust Person-Independent Visual Sign Language Recognition*. In: MARQUES, J. S., N. PÉREZ DE LA BLANCA und P. PINA (Herausgeber): *2nd Iberian Conference on Pattern Recognition and Image Analysis*, Band 3523 der Reihe *Lecture Notes in Computer Science*, 2005.
- [ZKN05] ZAHEDI, M., D. KEYSERS und H. NEY: *Appearance-Based Recognition of Words in American Sign Language*. In: MARQUES, J. S., N. PÉREZ DE LA BLANCA und P. PINA (Herausgeber): *2nd Iberian Conference on Pattern Recognition and Image Analysis*, Band 3523 der Reihe *Lecture Notes in Computer Science*, 2005.
- [ZUA02] ZIEREN, J., N. UNGER und S. AKYOL: *Hands Tracking from Frontal View for Vision-Based Gesture Recognition*. In: GOOL, L. VAN, J. HARTMANIS und J. VAN LEEUWEN (Herausgeber): *24th DAGM Symposium*, Band 2449 der Reihe *Lecture Notes in Computer Science*, Seiten 531–539, September 2002.

Lebenslauf des Verfassers

Persönliche Daten

Vor- und Zuname: Jörg Zieren
Geburtsdatum/-ort: 29.09.1974 in Neheim-Hüsten
Staatsangehörigkeit: deutsch

Schulbildung

1981 – 1985 Grundschule in Mettmann
1985 – 1994 Konrad-Heresbach-Gymnasium in Mettmann
31.05.1994 Abschluss: Allgemeine Hochschulreife

Studium

10/1994 – 11/2000 Studium der Elektrotechnik an der RWTH Aachen
18.11.2000 Abschluss: Diplom-Ingenieur

Berufliche Tätigkeit

02/2001 – 04/2006 Wissenschaftlicher Mitarbeiter am Lehrstuhl
für Technische Informatik der RWTH Aachen
seit 11/2006 IVU AG, Aachen

