Videobasierte Handlungserkennung für die natürliche Mensch-Maschine-Interaktion

Nils Hofemann

Dipl.-Inform. Nils Hofemann
AG Angewandte Informatik
Technische Fakultät
Universität Bielefeld
email: nhofeman@techfak.uni-bielefeld.de

Abdruck der genehmigten Dissertation zur Erlangung
des akademischen Grades Doktor der Ingenieurwissenschaften (Dr.-Ing.).
Der Technischen Fakultät der Universität Bielefeld
am 13. Dezember 2006 vorgelegt von Nils Hofemann,
am 20. April 2007 verteidigt und genehmigt.

Gutachter:

Prof. Dr. Helge Ritter, Universität Bielefeld
Dr.-Ing. Jannik Fritsch, Honda Research Institute Europe

Prüfungsausschuss:

Prof. Dr.-Ing. Gerhard Sagerer, Universität Bielefeld
Prof. Dr.-Ing. Helge Ritter, Universität Bielefeld
Dr.-Ing. Jannik Fritsch, Honda Research Institute Europe
Dr.-Ing. Thomas Hermann, Universität Bielefeld

Gedruckt auf alterungsbeständigem Papier °° ISO 9706

Videobasierte Handlungserkennung für die natürliche Mensch-Maschine-Interaktion

Dissertation zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)

der Technischen Fakultät der Universität Bielefeld
vorgelegt von

Nils Hofemann

Bielefeld – Dezember 2006

Danksagung

Ich möchte an dieser Stelle die Gelegenheit nutzen und den Personen meinen Dank aussprechen, die mich bei meiner Arbeit unterstützt, gefördert und begleitet haben.

Dankbar bin ich für die gute Betreuung von Jannik Fritsch, sowie für seine kreativen Hinweise und Vorschläge. Genauso gilt mein Dank Jannik Fritsch und Helge Ritter dafür, dass sie meine Dissertation begutachten haben. Auch der Prüfungskommission möchte ich meinen Dank für ihre Arbeit aussprechen.

Meinen Kollegen und Kolleginnen in der Arbeitsgruppe *Angewandte Informatik* und im Graduiertenkolleg *Aufgaben orientierte Kommunikation* möchte ich danken für ihr Interesse, ihre Offenheit und für die gute Zusammenarbeit. Diese Kooperationen haben mir ein Arbeiten in verschiedenen interessanten Anwendungsgebieten ermöglicht. Ein besonderer Dank gilt meinen Bürokollegen und Bürokolleginnen, sowie den Kollegen und Kolleginnen aus dem COGNIRON-, Motionese- und VAMPIRE-Projekt. Dankbar bin ich auch für meine Freunde und Freundinnen, die meine Arbeit Korrektur gelesen haben. Vielen Dank für eure Mühen und eure konstruktive Kritik!

Nicht zuletzt möchte ich meinen Eltern und meiner Familie für ihre Unterstützung und Zuversicht danken. Ihr habe mich auf meinem Weg bis zur Promotion begleitet und ermutigt.

Inhaltsangabe

Die Fragestellung, unter der diese Dissertation steht, ist, wie sich das Erkennen von Gestik als sinnvoller und gewinnbringender Teil einer multimodalen Interaktion zwischen Menschen und Maschinen erreichen und nutzen lässt. Die Vision besteht darin, dass sich nicht mehr der Mensch an das System anpassen muss, sondern die üblichen und natürlichen Fähigkeiten und Modalitäten des Menschen vom Computersystem unterstützt werden.

Das in dieser Arbeit angestrebte Ziel hat sich aus der Fragestellung und der Vision entwickelt: Es soll eine robuste sowie fehlerarme Erkennung menschlicher Gesten für multimodale Systeme möglich sein. Die interaktionalen Fähigkeiten eines solchen Systems sollen so erweitert und verbessert werden.

Im Blick auf dieses Ziel beschäftigt sich die Arbeit einerseits mit den Grundlagen der Gestik in der Mensch-Maschine-Interaktion, andererseits wird aber auch die praktische Realisierung und Anwendung einer automatischen Gestenerkennung verwirklicht. Bedingung für die Gestenerkennung ist ein flexibler, modularer Aufbau, der eine gute Adaption an unterschiedliche Bedingungen in verschiedenen Systemen gewährleistet. Des Weiteren werden innovative Lösungen für die Interpretation von Handlungen in ihrem symbolischen und situativen Kontext entwickelt.

Deshalb wird der Fokus dieser Arbeit auf besonders wichtige Gesten der Interaktion gelegt. Das Merkmal dieser deiktischen und manipulativen Gesten ist, dass sie in einem Kontext mit Objekten der Umgebung auftreten und in diesem interpretierbar sind.

Die zwischenmenschliche Kommunikation und insbesondere das umfangreiche Gebiet des gestischen Repertoires des Menschen sind Gegenstand langjähriger wissenschaftlicher Untersuchungen. In dieser Arbeit werden deshalb Betrachtungen und Theorien der Kommunikationsforschung und Psychologie mit einbezogen und im Kontext der MMI diskutiert. Als junges und spannendes Gebiet ist insbesondere die Erforschung des kindlichen Lernens von Objektmanipulationen zu nennen. Theorien dieses Gebiets werden in dieser Arbeit aufgegriffen und mit automatisierten Verfahren untersucht und nachvollzogen. Die Forschungsergebnisse der Psychologie sind von grundlegender Bedeutung für das automatische Interpretieren von Gesten und Erschließen ihrer Intention. Die im Rahmen dieser Dissertation für dieses Themengebiet durchgeführten Untersuchungen konzentrieren sich auf Grundlagenforschungen in diesem für die Informatik neuen und vielversprechenden Gebiet.

Diese Arbeit leistet einen Beitrag zur Vision einer natürlichen Interaktion des Menschen mit Maschinen. Der Schwerpunkt wird auf das Erkennen und Interpretieren von Gesten und Handlungen gelegt.

Abstract

The question addressed in this thesis is how to make use of gestures in a multimodal human-machine-interaction. To interact with current computer systems the human still has to adopt himself to the interfaces of the system. The underlying vision of this work is to enable computer system to interpret the typical human forms of interaction.

Based on this question and vision the goal of this work was developed: Build a robust recognition of human gestures for a multimodal system. Thus the interactive competences of the system can be broadend and improved.

Focussing this goal this work examines the theoretic basis of the human-machine-interaction as well as the practical implementation and evaluation of an automatic recognition of gestures. The gideline for the implementation is a flexible and modular architecture. This work presents novel approaches in taking the symbolic and situational context of gesture into account (see Fritsch u. a. [2004][1] and Hofemann u. a. [2004]). Continuative work in deriving humans intent of gestures is depicted (see Li u. a. [2005b]).

This work concentrates on the interactional behaviour and gestures of humans thus mostly deictic and manipulative gestures are considered. Their context is the main aspect of these gesture, it allows reasoning about their meaning and intention.

As mentioned above the gesture recognition system is intended to be part of a multimodal system. This integration is shown by the successful use of the developed system within a mobile, social, and multimodal robot (see Haasch u. a. [2005] and Wrede u. a. [2004a]) and within an wearable assistant system (see Hanheide u. a. [2006]).

Human-human-communication and the field of gestures are in the focus of research for many years. Hence this work gives an overview on communication theory and psychological research relevant for the human-machine interaction. A young and challenging aspect is the research on the topic: "How do children learn the manipulation of objects from their parents." The theory developed by Brand u. a. [2002] is the basis for an automatic analysis of parents behaviour presented in this work (see Fritsch u. a. [2005a]).

This work is a contribution to the vision of a natural human-machine-interaction. The emphasis lies on the videobased recognition of human actions and gestures.

[1] Part of this work is published in English. Please see appendix 'E' on page 140 for references.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Seit Anbeginn menschlicher Entwicklung formt und wandelt sich Interaktion und Kommunikation zwischen Menschen. Es entstanden und entstehen Zeichen, Gesten und Sprachen, die das Austauschen von Informationen, Anweisungen und Fragen ermöglichen. Diese Entwicklungen weisen viele kulturelle Unterschiede auf, die das heute bekannte soziale Geflecht in all seinen Ausprägungen gebildet haben. Verstärkt seit der Industrierevolution im 18. Jahrhundert nehmen auch Maschinen und seit circa 40 bis 50 Jahren auch Computer eine Rolle im sozialen und kulturellen Umfeld der Menschen ein und haben dieses grundlegend verändert. Viele Entwicklungen und Annehmlichkeiten basieren auf mehr oder weniger intelligenten Maschinen oder computergestützten Systemen. Als Konsequenz dieses Prozesses gehört die Fähigkeit, Computer und Maschinen bedienen zu können, zu den üblichen aber auch geforderten Kompetenzen im Alltag und Berufsleben. Der technologische Fortschritt erfordert von uns Nutzern einen lebenslangen Prozess des Lernens, um den wandelnden Anforderungen gewachsen zu sein.

In der heutigen Interaktion zwischen Menschen und Maschinen passt sich der Mensch stark an die Interaktionsmöglichkeiten des Computers an, um diesem explizite Anweisungen und Befehle zu geben. Über Tastatur, Maus oder Spracheingaben kann der Benutzer Programme steuern und bedienen. Hierfür muss der Benutzer lernen, welche Möglichkeiten ihm ein Programm bietet und wie er diese für seine Zwecke nutzen kann. Die Auswirkung seiner Anweisungen bleiben aber meist auf das Computersystem beschränkt und haben höchstens indirekten Einfluss auf die reale Welt des Menschen. Der Computer ist in dieser Hinsicht eine Maschine in der Umwelt des Menschen, die diesem nützt, aber nicht selbst in der Umwelt situiert ist. Eine Interaktion oder Kommunikation, wie sie zwischen Menschen üblich ist, bleibt Computersystemen somit unzugänglich. Heutigen Personalcomputern ist es nur in Ansätzen möglich Gestik, Mimik oder andere Ausdrucksformen und Aktionen eines Benutzers automatisch zu erfassen und zu verstehen oder solche Ausdrücke selbst zu generieren. Nachteile, die sich aus diesem Mangel ergeben, sind beiderseitige Missverständnisse und Fehleinschätzungen. Diese führen auf der Seite des Menschen zu Vorbehalten, Misstrauen oder Ablehnung des Computersystems. Außerdem sind die Anwendungen meist auf die limitierten Möglichkeiten des Computersystems beschränkt und schöpfen

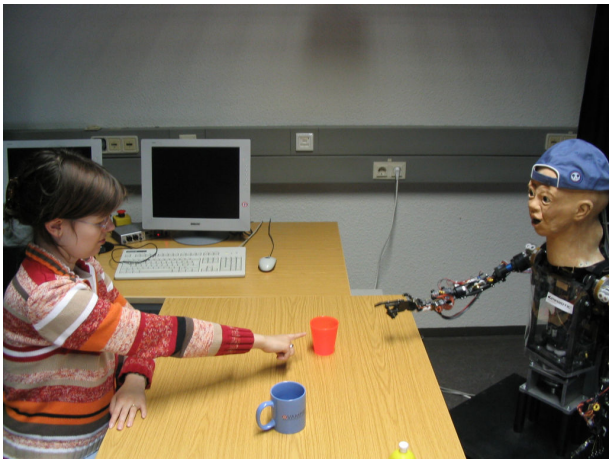


Abbildung 1.1: Eine Vision der natürliche Mensch-Maschine-Interaktion: In der gemeinsamen, situierten Interaktion zeigt eine Person auf ein Objekt und der humanoide Roboter bestätigt mit seiner Zeigegeste, dass er weiß, welches Objekt referenziert wurde.

nicht die Fähigkeiten des Menschen aus. Die Abbildung 1.1 illustriert die Vision einer multimodalen Interaktion zwischen einer Person und einem Roboter.

Computersystemen fehlen die Möglichkeiten, ihre Benutzer und ihre Umwelt wahrzunehmen und in ihr zu agieren. Direktere Interaktionen zwischen Menschen und Computern ermöglichen jedoch virtuelle Umgebungen, wie sie zum Beispiel in den Artikeln von Kehl u. Gool [2004] oder Kopp u. a. [2003] beschrieben werden. Diese virtuellen Umgebungen, bekannt unter dem Name *CAVE*[1], bezeichnen einen Raum zur Projektion einer dreidimensionalen Illusionswelt, der virtuellen Realität. Sie ermöglichen dem Benutzer, in eine computeranimierte Umgebung einzutreten und in ihr zu agieren. So kann er unter anderem auf virtuelle Objekte zeigen oder diese greifen, auch kann er in virtuellen dreidimensionalen Welten navigieren. Obwohl hier die Interaktionsmöglichkeiten direkter und multimodal sind, muss sich der Mensch dieser Umgebung mehr oder weniger stark anpassen. Denn die Bewegungen und die Position des Menschen werden für diese virtuellen Umgebungen oft mit speziellen Geräten erfasst und so für das System interpretierbar.

Entwicklung der Mensch-Maschine-Interaktion

Den Computer vom eingeschränkt nutzbaren Spezialisten zum alltäglichen Instrument und Gefährten des Menschen weiter zu entwickeltn, der adäquat mit Menschen interagieren kann, ist eine Vision der *Mensch-Maschine-Interaktion* (MMI). Ein Ziel der Forschungen auf dem Weg zu dieser Vision ist es, Systeme zu entwickeln, die sich ihrer Umgebung und ihres aktuellen Kontextes bewusst sind und dem Menschen eine natürliche Interaktion ermöglichen. Der Begriff *Mensch-Computer-Interaktion* (MCI) kann als Spezialisierung des allgemeinen Begriffs MMI aufgefasst werden. In der MCI wird der Fokus auf die direkte Interaktion und Bediehnung eines Computern gelegt, wohingegen die MMI komplexere

[1] Cave steht für *Cave Automatic Virtual Environment*, wörtlich: einer Höhle mit einer automatisierten, virtuellen Umwelt

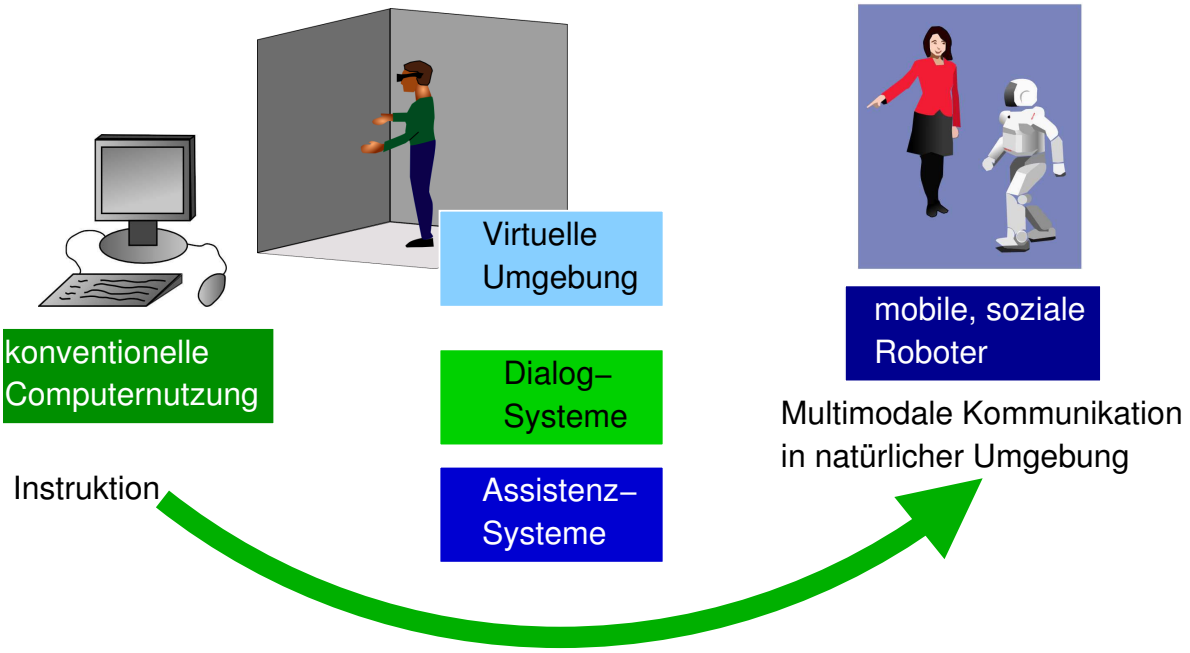


Abbildung 1.2: Der Wandel der Mensch-Maschine-Interaktion von der Instruktion zur Kommunikation.

Systeme betrachtet, die Rechnerkapazitäten beinhalten und nutzen. Mögliche Szenarien sind mobile Roboter, die als Begleiter oder Ratgeber des Menschen auftreten (siehe zum Beispiel Haasch u. a. [2004]; Kröse u. a. [2003]; Pineau u. a. [2003]) oder mobile Assistenzsysteme, die ihren Benutzer bei komplexen Handlungen oder Konstruktionen beraten und unterstützen (siehe zum Beispiel Wrede u. a. [2006b]). Mit der Entwicklung dieser mobilen Roboter beschäftigt sich das relativ junge Forschungsgebiet der *sozialen interaktiven Robotik*, über das Fong u. a. [2003] einen Überblick geben. Da sich diese Forschung mit Robotern beschäftigt, für die eine soziale Mensch-Maschine-Interaktion wichtig ist, sind die Kompetenzen unterschiedlicher Wissenschaften in einem interdisziplinären Entwicklungsprozess nötig.

Die Graphik 1.2 visualisiert die Entwicklung in der Interaktion zwischen Menschen und Maschinen. Diese Entwicklung ist der Wechsel von einer befehlsbasierten Instruktion zu einer menschenähnlichen und sozialen Kommunikation zwischen Menschen und Systemen. An Computersysteme, die diese Anforderungen erfüllen und in der natürlichen Umgebung des Menschen interagieren sollen, werden hohe Anforderungen in Bezug auf Sensorik, Perzeption und Generieren von Äußerungen und Aktionen gestellt. Das Ziel sollte hierbei sein, einem ungeübten Benutzer des Robotersystems eine menschenähnliche sowie intuitive Interaktion zu ermöglichen und eventuelle Hemmschwellen, die ihn von der Interaktion abhalten könnten, gering zu halten. Die kommunikativen Fähigkeiten des Menschen dienen hier als Vorbild und es wird versucht, diese für Robotersysteme zu erschließen. Da der

Mensch nicht nur Sprache sondern auch Mimik, Emotion und Gestik zur Kommunikation einsetzt, sind auch dieses Themen der sozialen interaktiven Robotik.

Basis der Forschung

Dieses Nachbilden einer zwischenmenschlichen Interaktion bedingt ein tiefes Verständnis der ablaufenden Prozesse und Entwicklungen. Zum Beispiel können Menschen relevante Strukturen im Ablauf einer Bewegungssequenz erkennen, sie bemerken das Ende der einen und den Anfang der folgenden Aktion. Weiterhin können Menschen die Motivation des Agierenden, die der speziellen Bewegung zugrundeliegt, identifizieren. Die Prozesse, die dies dem erwachsenen Menschen ermöglichen, sind schwer zu beobachten.

Kleinkinder liegen mit ihrern enormen Lernfähigkeiten besonders im Zentrum der Forschung der Entwicklungspsychologie. Ein besonderer von Rebecca Brand u. a. [2002] beobachteter Effekt ist, dass Eltern ihren Kleinkindern beim Verstehen und Nachmachen von Objektaktionen (z.B. bei manuellen Konstruktionen) helfen, indem sie ihre Bewegungen an die Wahrnehmungsfähigkeiten ihrer Kinder anpassen. Das ermöglicht nicht nur die Imitation, sondern auch das Ausmachen von relevanten Bewegungen und das Lenken der Konzentration. Diese Beobachtung ist in Anlehnung an den ähnlichen Effekt der *Motherese* aus der Mutter-Kind-Sprache unter dem Namen *Motionese* bekannt geworden. Ob und wie sich diese Erkenntnisse für die MMI nutzen lassen, ist eine interessante und relevante Fragestellung.

1.1 Zielsetzung der Arbeit

Die Fragestellung, unter der diese Dissertation steht, ist, wie sich das Erkennen von Gestik als sinnvoller und gewinnbringender Teil einer multimodalen Interaktion zwischen Menschen und Maschinen erreichen und nutzen lässt. Die Vision ist es, dass sich nicht mehr der Mensch an das System anpassen muss, sondern die üblichen und natürlichen Fähigkeiten und Modalitäten des Menschen vom Computersystem unterstützt werden.

Das in dieser Arbeit angestrebte Ziel hat sich aus der Fragestellung und der Vision entwickelt: Es soll eine robuste sowie fehlerarme Erkennung menschlicher Gesten für multimodale Systeme möglich sein. Die interaktionalen Fähigkeiten eines solchen Systems sollen so erweitert und verbessert werden.

Im Blick auf dieses Ziel beschäftigt sich die Arbeit einerseits mit den Grundlagen der Gestik in der Mensch-Maschine-Interaktion, andererseits wird aber auch die praktische Realisierung und Anwendung einer automatischen Gestenerkennung verwirklicht. Bedingung für die Gestenerkennung ist ein flexibler, modularer Aufbau, der eine gute Adaption an unterschiedliche Bedingungen in verschiedenen Systemen gewährleistet. Des Weiteren werden innovative Lösungen für die Interpretation von Gesten und Handlungen entwickelt.

Die angestrebte Gestenerkennung soll nicht nur als atomare und losgelöste Komponente gesehen werden, sondern zielt in der Konzeption und Realisierung auf die Integration in multimodale Systeme. Im Speziellen sollen die Vorteile einer Gestenerkennung für ein multimodales, sozial interagierendes Robotersystem und ein Assistenzsystem gezeigt werden.

Die Integration einer Komponente in Echtzeitsysteme stellt besondere Herausforderungen an die Modalität und Robustheit. Bei der Entwicklung können aber Synergien in der Arbeitsgruppe Angewandte Informatik (Universität Bielefeld) genutzt werden. Die Gruppe strebt an, komplexe, lauffähige Systeme für die multimodale Interaktion zwischen Mensch und Computer zu entwickeln. Zur wissenschaftlichen Einordnung der eigenen Entwicklungen werden alternative Verfahren zur Bewegungserkennung vorgestellt und ihre Vor- und Nachteile zu diskutiert.

Die zwischenmenschliche Kommunikation und insbesondere das umfangreiche Gebiet des gestischen Repertoires des Menschen sind Gegenstand langjähriger wissenschaftlicher Untersuchungen. In dieser Arbeit werden deshalb Betrachtungen und Theorien der Kommunikationsforschung und Psychologie mit einbezogen und im Kontext der MMI diskutiert. Als junges und spannendes Gebiet ist insbesondere die Erforschung des kindlichen Lernens von Objektmanipulationen zu nennen. Theorien dieses Gebiets werden in dieser Arbeit aufgegriffen und mit automatisierten Verfahren untersucht und nachvollzogen. Die Forschungsergebnisse der Psychologie sind von grundlegender Bedeutung für das automatische Interpretieren von Gesten und Erschließen ihrer Intention. Die im Rahmen dieser Dissertation für dieses Themengebiet durchgeführten Untersuchungen konzentrieren sich auf Grundlagenforschungen in diesem für die Informatik neuen und vielversprechenden Gebiet.

Es würde den Rahmen dieser Arbeit sprengen, wenn man die Vielfalt der menschlichen Bewegungen und Gesten betrachten würde. Deshalb wird der Fokus dieser Arbeit auf besonders wichtige Gesten der Interaktion gelegt. Das Merkmal dieser deiktischen und manipulativen Gesten ist, dass sie in einem Kontext mit Objekten der Umgebung auftreten und in diesem interpretierbar sind. Das schließt das Gebiet der Zeichensprachen aus, auch Bewegungen des ganzen Körpers, die in der Literatur oft als Gesten aufgefasst werden, sind nicht Thema der vorliegenden Arbeit.

Die Arbeit soll einen Beitrag zur Vision einer natürlichen Interaktion des Menschen mit Maschinen leisten. Der Schwerpunkt wird auf das Erkennen und Interpretieren von Gesten und Handlungen gelegt.

1.2 Gliederung

Zielvorstellung dieser Arbeit ist die multimodale Interaktion zwischen Menschen und intelligenten Robotersystemen im menschlichen Alltagsleben. So führt diese Arbeit auch mit dem Kapitel 2 zuerst in die menschliche Interaktion ein und konzentriert sich hierbei besonders auf das nonverbale Ausdrucksvermögen des Menschen durch Gesten. Eingeleitet wird der Themenkomplex der Interaktion mit einer Betrachtung der Grundlagen der Interaktion und deren Umsetzung für die Mensch-Maschine-Interaktion. Die Bedeutung der Gesten in der allgemeinen Kommunikation und speziell in der MMI wird erläutert. Abgeschlossen wird Kapitel 2 mit einer genaueren Betrachtung der menschlichen Handgesten.

Nach der Betrachtung der Interaktion und dem Herausstellen der Bedeutung von Gesten wird im Kapitel 3 erarbeitet, wie sich menschliche Gesten in die MMI integrieren lassen.

Die Grundlagen einer visuellen Handlungserkennung, die eine Detektion und Klassifikation von Bewegungen der Hand erfordert, werden erläutert. Hierzu gehören Verfahren der Musterklassifikation und Musteranalyse. Anschließend werden existierende Verfahren, die Gesten von Menschen verfolgen und typische Muster in diesen erkennen, vorgestellt.

Im Kapitel 4 wird das Verfahren zur Handlungserkennung, das im Kontext dieser Arbeit entwickelt wurde, vorgestellt. Aufbauend auf der Erläuterung der probabilistischen Grundlagen, wird der entwickelte Algorithmus dargelegt. Ein Verfahren, um die benötigten Modelle von Gesten zu erzeugen, wird im Anschluss vorgestellt. Des Weiteren wird eine Erweiterung der Gestenerkennung vorgestellt, die der Tatsache Rechnung trägt, dass viele Gesten im Kontext von Objekten ausgeführt werden.

Die Integration der Handlungserkennung in zwei unterschiedliche Systeme zur Interaktion zwischen Mensch und Computern wird im Kapitel 5 erläutert und die Erkennung von Gesten jeweils evaluiert. Das erste Evaluationsszenario ist die Interaktion mit dem mobilen Roboter BIRON[2]. Das zweite Evaluationsszenario ist das im VAMPIRE Projekt[3] entwickelte Assistenzsystem, bei dem der Benutzer seine Umwelt durch am Kopf getragene Kameras sieht.

Mit der Analyse des Lernens neuer Aktionen und den Möglichkeiten, Erkenntnisse aus der Psychologie für das maschinelle Erlernen von Gesten und Aktionen zu nutzen, beschäftigt sich Kapitel 6. Die Ergebnisse einer Studie zur automatischen Detektion von Merkmalen von Motionese werden dargestellt.

Die vorliegende Arbeit wird mit einer Zusammenfassung und Diskussion möglicher weiterer Forschungsarbeiten in Kapitel 7 abgeschlossen.

[2]BIRON: **Bi**elefeld **R**obot Compani**on** (BIRON)
[3]VAMPIRE Projekt: http://www.vampire-project.org, siehe auch Bauckhage u. a. [2005]

2. Gesten und Handlungen in der Interaktion

Das aktuelle Kapitel führt von einer Betrachtung der Grundlagen der Interaktion zwischen Menschen und Computern (2.1) zu der Kommunikation zwischen Menschen und Robotern als spezielle Ausprägungen der MMI. Des Weiteren wird die Bedeutung von Gesten und Handlungen als Teil der Kommunikation herausgearbeitet und ihre Rolle in der Mensch-Maschine-Interaktion beschrieben (2.2). Anschließend folgt eine Betrachtung der Charakteristika menschlicher Bewegungen, die bei der Ausführung von Handgesten und Handlungen auftreten (2.3).

Aus dem zwischenmenschlichen Dialog sind Gesten nicht wegzudenken. Beispielsweise zeigt ein Gesprächspartner auf Objekte, um seine Äußerung zu vervollständigen und die Aufmerksamkeit zu lenken. Auch andere Gesten tragen zum Verständnis bei und können sogar sprachliche Äußerungen ersetzen. Wird zum Beispiel eine Handlung erklärt, hört man oft den Satz „Schau mal her, ich zeig es dir gerade“. Ein Ziel der sozialen interaktiven Robotik ist es, mit Computersystemen diese Aspekte der menschlichen Kommunikation zu erfassen und zu interpretieren.

2.1 Grundlagen der Mensch-Maschine-Interaktion

Ziel des Designs einer Schnittstelle zwischen Computern und ihren Benutzern sollte es sein, eine für den Menschen und die Aufgabe adäquate Form zu finden, die dem Menschen die Interaktion erleichtert und zugänglich macht. Die ersten Schnittstellen zwischen Menschen und Maschinen bestanden aus einfachen Schaltern, Kontrollleuchten und Lochkarten. Die Entwicklung führte weiter über die Kontrolle von Computern mittels der Kommandozeile bis hin zu graphischen Benutzeroberflächen (engl. Graphical User Interface). Diese Kombination aus Tastatur, Maus und Bildschirm hat sich in den letzten Jahrzehnten als Schnittstelle für die Mensch-Computer-Interaktion etabliert. Mit dem Fortschritt der Computertechnik entstehen aber auch neue, andere Formen und Möglichkeiten der Interaktion wie

zum Beispiel virtuelle Umgebungen, konversationale Schnittstellen oder intelligente Räume. Diese neuen Anwendungsgebiete erfordern neue Konzepte für die Mensch-Maschine-Schnittstellen, die die bisherigen Schnittstellen ergänzen oder ersetzen. Herausfordernde Einsatzfelder sind unter anderem der mobile Einsatz von Computersystemen, wie zum Beispiel bei tragbaren Geräten, in Automobilen oder interaktiven Robotern.

Das Studium der Mensch-Computer-Interaktion fokussiert den Menschen und seine Fähigkeiten, um diesen Fähigkeiten in technischen Systemen Rechnung zu tragen. Ziel aktueller Forschungen ist unter anderem die Entwicklung von Schnittstellen und Interaktionsprozessen, die über die bisherige befehlsbasierte Kontrolle hinausgehen und mehr an die zwischenmenschliche Kommunikation angelehnt sind. Diese beinhaltet die Einbeziehung von Sprache, Gestik, Mimik, des Tastsinns und weiterer kommunikativer Kanäle, mithin auch einer multimodalen Prägung der MCI. Turk [2005a][1] beleuchtet in seinem Beitrag die multimodale Mensch-Computer-Interaktion, ihre geschichtlichen Wurzeln und aktuelle Entwicklungen aus der Sicht der Bildverarbeitung. Die Hauptaufgabe der Bildverarbeitung in Echtzeit für die MCI sieht er im Detektieren und Erkennen von bedeutungstragenden Kommunikationshinweisen und fasst dieses mit dem Satz „look at the user“ [2] zusammen. Außerdem stellt er die Bedeutung der Bildverarbeitung für die MCI heraus, da mit ihr unaufdringliche und nicht störende Verfahren zur Wahrnehmung menschlicher Aktivitäten möglich sind.

Ziel einer MCI sollte eine möglichst natürliche und der zwischenmenschlichen Kommunikation angenäherte Form sein. Das aktuelle Kapitel geht deswegen im Folgenden auf diese näher ein. Des Weiteren wird eine spezielle Form der Interaktion, die Kommunikation zwischen Menschen und Roboter, erarbeitet.

2.1.1 Betrachtung der Kommunikation

Die Kommunikation ist ein weit reichendes Feld, mit dessen unterschiedlichen Aspekten sich verschiedene Disziplinen beschäftigen und den Begriff Kommunikation in ihrem Umfeld nutzen. Wenn die Gegenseitigkeit in der Kommunikation herausgestellt wird, werden die Begriffe Interaktion und Kommunikation oft auch synonym verwendet. Geprägt wurde der Begriff „Kommunikation“ durch das Buch „Menschliche Kommunikation“ von Watzlawick u. a. [1971] mit dem berühmten Axiom: „Man kann nicht nicht kommunizieren.“[3] Im Rahmen dieser Arbeit wird die Kommunikation im Kontext der Interaktion zwischen Menschen und Maschinen betrachtet. Es soll die Frage aufgegriffen werden, ob und in welcher Form eine Kommunikation zwischen Menschen und Maschinen möglich ist. Aus Watzlawicks These kann man schlussfolgern, dass auch Maschinen — zum Beispiel interaktive Roboter — nicht nicht kommunizieren können. Eine Aufgabe der MMI muss es deswegen sein, die Kommunikation mit Maschinen für Menschen möglichst angenehm und natürlich zu gestalten.

Eine einheitliche Begriffsdefinition ist nicht möglich, auch ist umstritten, was als Kommunikation aufgefasst wird. Nöth [2000] stellt fest, dass Grundlage jedes kommunikativen

[1] Veröffentlicht im ersten Kapitel des Buchs von Kisacanin u. a. [2005]
[2] „auf den Benutzer schauen“
[3] Watzlawick u. a. [1971], Kap. 2.2

Prozesses jedoch die drei Konstituenten Kommunikator, Zeichen oder Botschaft und Rezipient sind (s. Graphik 2.1). Es ist aber bereits strittig, ob Kommunikator und Rezipient in diesem Sender-Nachricht-Empfänger-Modell Menschen sein müssen oder ob auch Maschinen, biologische Zellen und Ähnliches möglich sind.

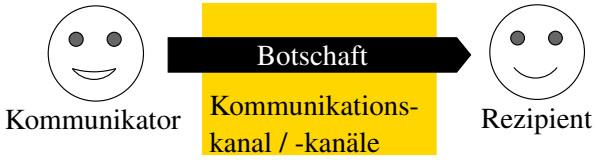


Abbildung 2.1: Konzeptionelle Darstellung eines Kommunikationsmodells.

Einige relevante Erörterungen zur Kommunikation aus dem Kapitel 4.3 des *Handbuch[s] der Semiotik* von Nöth [2000] sollen zum besseren Verständnis des Streitpunkts diskutiert werden. Die Auffassung, dass eine Kommunikation zwischen Menschen und Computern möglich ist, vertritt der Kybernetiker Georg Klaus. Er definiert Kommunikation als „den Austausch von Informationen zwischen dynamischen Systemen, die in der Lage sind, Informationen zu empfangen, zu speichern und zu verarbeiten“[4]. Diese allgemein gefasste Definition schließt sowohl die Kommunikation unter Computersystemen als auch zwischen Menschen und Computern ein. Kritisiert wird diese Auffassung, da ein Computer nicht die Funktion eines Kommunikators hat, sondern Informationen darstellt und Mitteilungen und Botschaften übermittelt. Aber das, so wird Nadin von Nöth zitiert, unterscheidet ein Computersystem nicht von einem Buch.[5] Nach dieser Auffassung kommunizieren Benutzer und Entwickler über das Medium Computer. Im Blick auf aktuelle Computerprogramme und auch die meisten Dialogsysteme, die nur einfache Kommandos erlauben, ist diese Argumentation nachzuvollziehen. Diese asymmetrische Interaktion wird schematisch in der Graphik 2.2 wiedergegeben. Festzuhalten bleibt, dass der Mensch sich an die Möglichkeiten des Computers anpassen muss. Er erteilt seine Instruktionen über die Kombination aus Tastatur und Maus und erhält die Rückmeldung über andere Kanäle, meistens über eine graphische Darstellung auf einem Monitor oder über akustische Signale.

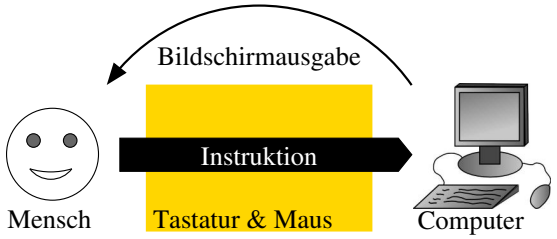


Abbildung 2.2: Übertragung einer typischen Interaktion zwischen Menschen und Computer in das Kommunikationsmodell.

[4] aus Nöth [2000] Kap. 4.3; Quelle: Klaus [1969]
[5] siehe Nöth [2000], Kap. 4.3

Natürlich ist diese Art der instruktionsbasierten Interaktion für viele Arbeiten sehr effizient und aus dem Arbeitsalltag nicht weg zu denken. Doch erstens erfordert diese Interaktion eine zum Teil aufwendige Einarbeitung des Menschen und zweitens bleibt die Interaktion mit leistungsstarken Computersystemen auf wenige Anwendungen beschränkt. Für die Disziplin der Mensch-Maschine-Interaktion stellt sich demnach die Frage, an welchen Stellen die moderne und zukünftige Interaktion von Menschen mit Maschinen über die aktuelle Vorstellung des befehlsgesteuerten Computers hinaus geht und was hierfür notwendig ist. Ein erster Schritt in diese Richtung ist das Verwenden von Kommunikationskanälen, die der zwischenmenschlichen Kommunikation entstammen, zum Beispiel der Sprache (siehe Graphik 2.3). Entsprechende Systeme haben bereits Einzug in unseren Alltag gehalten. Doch manchmal bestehen noch Probleme in ihrer Leistungsfähigkeit, da oft nur kurze sprachliche Instruktionen möglich sind. Des Weiteren fällt den menschlichen Benutzern zum Teil negativ auf, dass sie zu einer Maschine sprechen, die nur sprachliche Signale versteht, aber weder Emotionen noch sprachliche Feinheiten interpretieren kann. Dass inzwischen aber das Erkennen von Emotionen aus dem Sprachsignal möglich ist, wurde in einer Studie von Hegel u. a. [2006] gezeigt. Weitere Untersuchungen müssen zeigen, wie diese Möglichkeit der Emotionserkennung gewinnbringend für einen Dialog zwischen Mensch und Maschinen verwendet werden kann.

Abbildung 2.3: Die Sprache stellt einen unimodalen Kommunikationskanal dar, der keine oder nur wenig Anpassungen seitens des Menschen voraussetzt, um einen Dialog mit einem Computer zu führen.

Neben der weiteren Verbesserung der Dialogsysteme können die Entwicklung einer eigenen Persönlichkeit und die Möglichkeit, sich an einzelne Benutzer zu adaptieren, zu einer menschenähnlicheren Kommunikation beitragen. Ein Aspekt, der besonders in der *Mensch-Roboter-Kommunikation* (MRK) hervortritt, ist, dass der Roboter als Partner in der Umgebung des Menschen agiert. Der Roboter ist in der natürlichen Umgebung des Menschen situiert. Für dieses Szenario müssen die sensorischen und kommunikativen Fähigkeiten eines Roboters an die Bedürfnisse des Menschen und Gegebenheiten der Umwelt angepasst und erweitert werden. Ziel ist es, Roboter oder Computersysteme in einer menschenähnlichen, multimodalen Art und Weise mit Menschen kommunizieren zu lassen. Dies ist exemplarisch in dem Kommunikationsmodell in Graphik 2.4 herausgearbeitet.

2.1.2 Multimodale Kommunikation

Die Multimodalität, ein Merkmal der von der MMI angestrebten Form der Kommunikation, bedarf einer näheren Betrachtung. Eine Modalität der Wahrnehmung bezieht sich auf einen bestimmten Sinn des Menschen, zum Beispiel das Sehen oder Hören. Des Weiteren spricht man in der Kommunikationstheorie von Kanälen, über die Informationen

Abbildung 2.4: Die multimodale Kommunikation zwischen einem Menschen und einem Computersystem kommt den Bedürfnissen des Menschen entgegen, stellt aber hohe Anforderungen an das System.

übertragen werden. Dies sind unter anderem Sprache, Schrift oder Gesten. In der Kommunikationstheorie beschreibt ein kommunikativer Kanal (siehe auch Graphik 2.1) eine Interaktionstechnik, die eine Kombination der Kommunikationsmöglichkeiten des Kommunikators und Rezipienten nutzt. Am Beispiel des Kanals Sprache illustriert, steht auf Seite des Kommunikators die Technik zur Sprachgenerierung und auf der Seite des Rezipienten das Hören und Verstehen dieser Sprache. Auf die Interaktion mit einer Maschine übertragen, erfordert ein kommunikativer Kanal immer ein spezielles Gerät, das die Aktion des Menschen aufnimmt. In der heute üblichen Interaktion sind das zum Beispiel Tastatur oder Maus. Aber auch andere Aktionen des Menschen, wie gesprochene oder geschriebene Sprache oder Gesten können aufgenommen und interpretiert werden und so einen kommunikativen Kanal bilden.

Im Gegensatz zur unimodalen Kommunikation, bei der nur ein Kanal verwendet wird, spricht man von Multimodalität, wenn mehrere kommunikative Kanäle des Menschen, zum Beispiel der visuelle und auditive Kanal, benutzt werden. Meera Blattner [1992] beschreibt in ihrem Buch ausführlich die Aspekte der multimodalen Schnittstellen und der unterschiedlichen Designansätze, konzentriert sich hierbei aber auf die klassischen Schnittstellen der stationären Computer. In der Kommunikation zwischen Menschen wird meist unbewusst simultan von mehreren kommunikativen Kanälen Gebrauch gemacht. So zeigt der Sprecher zum Beispiel auf eine Person und referenziert sie sprachlich mit „du“.

Bereits an diesem Beispiel lässt sich der Nutzen der multimodalen Kommunikation erkennen, denn nur durch die gleichzeitige Interpretation von Sprache und Gestik lässt sich die Äußerung verstehen. Für ein multimodales Robotersystem bedeutet diese einfache Situation aber, dass es einen Menschen wahrnehmen, dessen verbale und nonverbale Äußerungen verstehen sowie interpretieren und auf das Verstandene reagieren sollte.

Weitere Vorteile sowie eine detaillierte Betrachtung der multimodalen Interaktion zwischen Menschen und Maschinen und ihrer Entwicklung gibt Turk [2005b]. Für die computerbasierte Bildverarbeitung sieht Turk die Herausforderungen in echtzeitfähigen Systemen, die eine robuste Erkennung in realen Umgebungen ermöglichen. Als Hauptaufgaben für die Bildverarbeitung und Mustererkennung sieht er das Finden und Erkennen von Gesichtern, Analysieren von Gesichtsausdrücken, Verfolgen von Händen und Körpern, sowie das Erkennen von Gesten und die Analyse von Aktivitäten. In diesen Feldern kann die automatische Bildverarbeitung einen Beitrag zur multimodalen MMI leisten. Folglich

erhält neben dem Erkennen von gesprochener Sprache der nonverbale Teil der Kommunikation für situierte Computersysteme eine größere Aufmerksamkeit, als es in heutigen instruktionsbasierten Systemen der Fall ist.

2.1.3 Nonverbale Kommunikation

Neben der Sprache gibt es weitere Arten der Kommunikation. Diese „nonverbale Kommunikation umfasst das Ausdruckspotential des menschlichen Körpers in Zeit und Raum“ (vergleiche Nöth [2000]). Laut Nöth lassen sich die Teilgebiete Gestik, Mimik, Blick, taktile Kommunikation, Kinesik, Proxemik und Chronemik unterscheiden. Für die MMI und insbesondere für die Kommunikation mit Robotern sind viele dieser Teilgebiete und deren Problematiken interessant. Zum Beispiel wird in der Proxemik die räumliche Nähe zwischen den Dialogpartnern untersucht. Die Chronemik beschäftigt sich mit kulturellen, soziologischen und psychologischen Aspekten der Zeit im zwischenmenschlichen Miteinander. Auch die Thematik der taktilen Kommunikation findet über Berührungssensoren an Robotern wie dem *Aibo*™ von Sony (siehe Abbildung 2.5) oder der *iCat*™(siehe van Breemen [2004]) von Philips Anwendung. Das Erkennen von Mimik und der Blickrichtung sind weitere herausfordernde Gebiete für die bildbasierte Mustererkennung. Während alle Bewegungen des menschlichen Körpers Thema der Kinesik sind, ist das umfangreiche Teilgebiet der Gestik im Kontext der vorliegenden Arbeit von besonderem Interesse und wird in Abschnitt 2.2 behandelt.

Bei der Betrachtung und Entwicklung der Mensch-Roboter-Kommunikation dient die Mensch-Mensch-Kommunikation als Vorbild. Dieses gilt für die einzelnen kommunikativen Kanäle, wie auch für die Kombination dieser zu einem kommunikativen System.

2.1.4 Mensch-Roboter-Kommunikation

Die Betrachtung der Kommunikation zwischen Menschen und sozialen, interaktiven Robotern als spezielles Gebiet der MMI erfordert zuerst eine kurze Betrachtung der Bedeutung des Wortes Roboter: Mit dem Begriff Roboter wird eine Vielzahl unterschiedlicher Maschinen bezeichnet; gemein ist ihnen, dass sie eine Aufgabe autonom ausführen. Großen wirtschaftlichen Erfolg haben die Industrieroboter, welche in der Produktion viele Arbeiten übernehmen und diese mit hoher Präzision und Wiederholungsgenauigkeit autonom ausführen können. Diese Roboter führen vorprogrammierte Bewegungen aus, sind deshalb nur sehr eingeschränkt adaptiv auf veränderte Situationen.

Neuere Entwicklungen in der Robotik stellen Serviceroboter oder Unterhaltungsroboter dar. Diese sind zum Beispiel im Haushalt einsetzbar als Staubsauger-Roboter oder dienen als Unterhaltungsobjekt, wie der Roboterhund Aibo™von Sony (siehe Abbildung 2.5). Setzt man die Entwicklung dieser Roboterarten weiter fort, gelangt man zu persönlichen, sozial interagierenden Robotern, die ihren Besitzer unterstützen, ihm helfen und sich in seinem Lebensumfeld aufhalten.

Auch wenn die kommunikativen Fähigkeiten verglichen mit den menschlichen noch sehr eingeschränkt sind, sind in den letzten Jahren vielversprechende Roboter entwickelt worden, die einen Schritt in diese Richtung darstellen. Dies sind unter anderem der bekannte gehende Roboter *Asimo*™ von Honda oder die *Partner Roboter* von Toyota, die auf der

(a) Ein Staubsager-Roboter der Firma Electrolux für den privaten Haushalt.

(b) Der Unterhaltungsroboter Aibo™der Firma Sony.

Abbildung 2.5: Zwei Beispiele für Roboter, die bereits heute in einigen Haushalten anzutreffen sind.

Expo 2005 in Japan ihr Können als Trompetenspieler zeigten. (Bilder dieser und anderer Roboter sind im Kapitel 5 auf Seite 77 abgebildet.) Für diese Vision der Robotik erlangt die Betrachtung der Interaktion zwischen Menschen und Roboter eine neue Relevanz. Das Computersystem ist mobil und in einer den menschlichen Bedürfnissen angemessenen Umgebung situiert. Eine vielschichtige und adaptive Kommunikation mit Menschen ist für diese Roboter denkbar und sinnvoll.

Kooperativ agierende Roboter, die zum Beispiel im Gespräch auf eine Zeigegeste reagieren, haben das Potential, über die aktuelle Verwendung von Robotern als nützliche Maschine hinaus zu gehen und sozial agierende Gefährten im Alltag der Menschen zu werden, diesen zu helfen und Arbeiten abnehmen zu können. Natürlich bleibt in dieser Vision die Dualität zwischen der alles beherrschenden sowie als gefährlich empfundene Maschine und dem freundlichen, hilfsbereiten Partner bestehen. Um hier den richtigen Weg einzuschlagen und auch Ressentiments vorzubeugen, muss der Mensch als Gegenüber und Benutzer des Roboters stets präsent bleiben.

Ein Mangel aktueller Computersysteme, den Turk [2005a] in dem Buch von Kisacanin u. a. [2005] herausstellt, ist, dass sich die Systeme nicht über ihre Benutzer bewusst sind. Ein System nimmt den Interaktionspartner also nicht wahr, sondern kann nur auf direkte Kommandos des Benutzers reagieren. Im Kontext der Roboter als Partner des Menschen ist dieser Mangel noch bedeutend schwerer und wird zum Beispiel von Lang [2005] und Kleinehagenbrock [2005] mit einem multimodalen Robotersystem gelöst. In diesem System, das auch später in dieser Arbeit als Integrationsplattform dient, werden die Modalitäten der Richtungserkennung von Geräuschen, dem laserbasierten Auffinden von Beinpaaren sowie der Detektion von Gesichtern fusioniert, um Interaktionspartner zu erkennen und einen sprachgestützten Dialog zu ermöglichen. Das in der Arbeitsgruppe Angewandte In-

formatik entwickelte System, der *Bielefeld Robot Companion*(*BIRON*), wird auch von Haasch u. a. [2004] sowie Fritsch u. a. [2003] eingehend vorgestellt.

In der Anwendung von Robotern als Partnern sind die üblichen Schnittstellen Maus, Tastatur und Monitor nur bedingt einsetzbar. Ziel ist es, eine Interaktion zu erlauben, die die zwischenmenschliche Interaktion als Vorbild nimmt. Eine interessante Modalität hierfür sind die Gesten, die im folgenden Unterkapitel thematisiert werden.

2.2 Gesten - ein Bestandteil der Kommunikation

Mit dem Fokus auf die Mensch-Maschine-Interaktion verschafft dieses Unterkapitel einen Überblick über das Forschungsgebiet der Gesten. Nach einer Betrachtung unterschiedlicher Definitionen und Abgrenzungen des Begriffs „Geste“ (2.2.1) wird die Bedeutung der Gesten für die MMI herausgearbeitet (2.2.2) und betrachtet, wie Gesten verstanden werden können (2.2.3). Im Anschluss werden drei Arten von Gesten näher betrachtet: Die Deiktika (2.2.4), Manipulationen (2.2.5) und Embleme (2.2.6). Grundlage der Betrachtung ist das Ziel, Gesten automatisch zu erkennen und somit in die Mensch-Maschine-Interaktion zu integrieren. Dieses automatische Erkennen von Gesten und die Integration in multimodale Systeme wird in den Kapiteln 4 und 5 behandelt.

2.2.1 Definition und Taxonomie von Gesten

Nöth [2000] definiert Gesten im engeren Sinne als das Ausdruckspotential des Menschen mittels der Arme, Hände und des Kopfes. Hiermit verortet er Gesten im Kontext der Sprache, Mimik und der nonverbalen Kommunikation. Allgemeiner können Gesten aber auch als „Körperhandlungen nichtsprachlicher Art, mit der Absicht etwas zum Ausdruck zu bringen“ [Kendon, 2004, Kap. 2] verstanden werden. Kendon prägt hierfür den Ausdruck der *deliberativen ausdrucksfähigen Bewegung* (engl. deliberate expressive movement). Der Begriff Geste wird sowohl von Noeth als auch von Kendon auf beabsichtigte Bewegungen, die der Kommunikation dienen, fokussiert.

Entsprechend der unterschiedlichen Definitionen ist auch die Klassifizierung der Gesten in der Forschung nicht eindeutig. Eine grobe Einteilung unterscheidet redebegleitende und von der Sprache autonome Gesten. Erste Arbeiten gehen auf Efron u. Veen [1972] und Ekman u. Friesen [1969] zurück. Um einen Eindruck einer möglichen Klassifikation zu vermitteln, sei hier exemplarisch die häufig verwendete Klassifikation von Ekman u. Friesen [1969] dargestellt (siehe Abbildung 2.6). Sie unterteilen die Hand- und Armbewegungen in beabsichtigte (deliberative) und unbeabsichtigte Bewegungen. Deliberative Bewegungen sind Gesten, die Ekman u. Friesen [1969] wiederum in fünf Klassen aufteilen: Die Embleme, die eine lexikalische und kulturell festgelegte Bedeutung haben, die Klassen der Affektäußerungen und der Körpermanipulationen sowie die beiden redebegleitenden Klassen Illustratoren und Regulatoren. Zu den Illustratoren zählen Ekman und Friesen auch die Zeigegesten (Deiktika). Affektäußerung sind unbewusste Bewegungen in der Kommunikation. Körpermanipulativen Bewegungen sind Änderungen der Körperhaltung oder auch ein Kratzen.

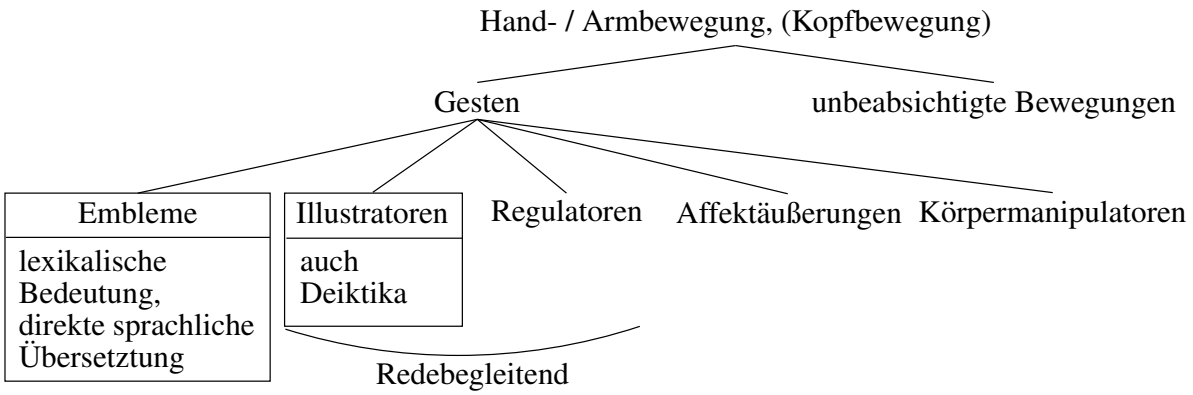


Abbildung 2.6: Die Klassifikation von Gesten nach Ekman u. Friesen [1969].

Im Gegensatz dazu versucht Kendon [1996] in seinem Artikel zur „Agenda der Gestenstudien“ gerade keine Klassifikation vorzunehmen, sondern unterteilt stattdessen die Hauptanwendungsgebiete von Gesten:

- Gesten fungieren als autonome Äußerungen (Emblem). Diese Gesten unterliegen meist einer strikten Konvention.
- Gesten sind Teil von Äußerungen in Alternation zur Sprache. Der Sprecher ersetzt einen Teil des gesprochenen Satzes durch eine Geste. Dieses können Referenzen auf Objekte durch Deiktika oder Beschreibungen durch ikonische Gesten sein.
- Gesten werden mit der Sprache kombiniert. Diese gleichzeitige Verwendung von Sprache und Gesten wird oft als Gestikulieren bezeichnet. Mit Gesten können zusätzliche aber auch redundante oder widersprüchliche Informationen dem Gesprochenen hinzugefügt werden.

Da es viele unterschiedliche Kriterien für die Unterteilung von Gesten gibt, existieren auch viele Klassifikationen und die Terminologie variiert entsprechend der Sichtweise der Autoren. Kendon vertritt deswegen die Meinung, dass kein einheitliches Klassifikationsschema entwickelt werden sollte, sondern die unterschiedlichen Klassifikationen als sinnvolle Werkzeuge in ihrem speziellen Bereich zu sehen sind (siehe Kendon [2004], Kapitel 6).

2.2.2 Gesten in der Mensch-Maschine-Interaktion

Um dieser aufgabenspezifischen Klassifikation von Gesten gerecht zu werden, empfiehlt sich eine Betrachtung der Gesten, ihrer Definition und Klassifizierung für die MMI. Exemplarisch werden hier die Ansätze von Pavlovic u. a. [1997] und Nehaniv [2005] vorgestellt und diskutiert.

Pavlovic u. a. [1997] definiert Gesten für die Interaktion zwischen Menschen und Maschinen als Aufgaben, welche mit der menschlichen Hand ausgeführt werden. Dieses umfasst

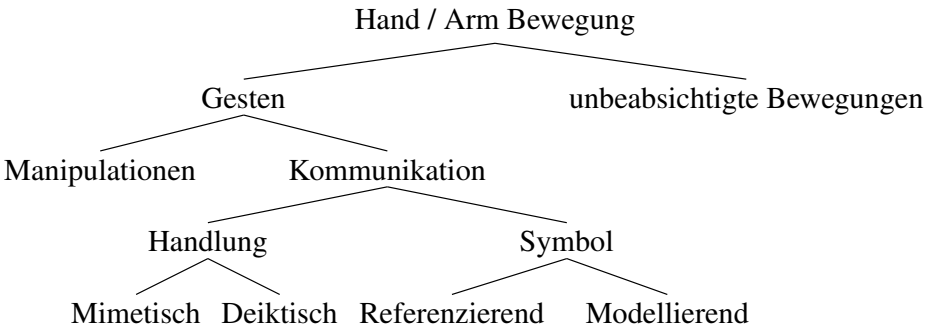


Abbildung 2.7: Taxonomie von Gesten für die HCI, nach Pavlovic u. a. [1997].

auch Manipulationen und kommunikative Akte (siehe Abbildung 2.7). Ähnlich zu Ekman unterteilt auch Pavlovic die Bewegungen in deliberative und unbeabsichtigte Gesten, jedoch werden Manipulationen bewusst als Gesten aufgefasst. Ekmans und Friesens Verständnis von Gesten entspricht eher der Untergruppe *Kommunikation* in Pavlovics Taxonomie. Hingegen werden deiktische und referenzierende Gesten in unterschiedlichen Gruppen verortet.

Einen weiteren Ansatz, das Gebiet der Gesten zumindest grob zu klassifizieren, stellt Nehaniv [2005] für die Interaktion zwischen Menschen und Robotern vor. Motivation ist für Nehaniv die Komplexität der Gesten, die im Kontext dieser situierten Interaktion auftreten, durch Klassen von Gesten für automatische Systeme handhabbar zu machen und eine Interpretation der Geste zu erlauben. Auch stellt er fest, dass eine Geste ohne Betrachtung ihres situativen Kontextes oft ambig ist. Erst die Zuordnung einer Geste zu einer Klasse stellt diese in einen Kontext und ermöglicht den Schluss auf die Intention der Geste. Zum Beispiel kann das aufgeregte Heben und Bewegen der Hände mit der Prosodie oder der Betonung der Sprache korrelieren und erhält in diesem Kontext seine Bedeutung als expressives Verhalten. Nehaniv schlägt fünf Klassen vor, erläutert aber auch, dass sich eine spezielle Geste oft nicht nur einer Klasse zuordnen lässt:

- **„Irrelevante“ / Manipulative Gesten.** Diese Gruppe fasst irrelevante Gesten, Körperbewegungen und ihre Seiteneffekte sowie Objektmanipulationen zusammen. Diese nicht kommunikativen Bewegungen verändern die Umgebung des Menschen oder seine Relation zu dieser. Beispiele sind unter anderem die Armbewegungen beim Laufen, das Spielen mit einem Kugelschreiber oder das Greifen nach einer Tasse. Trotz allem kann es, so stellt Nehaniv [2005] fest, für einen Roboter wichtig sein, die Untertypen unterscheiden zu können, um das menschliche Verhalten zu erkennen.
- **Seiteneffekt von expressivem Verhalten.** Im Gespräch sind die Bewegungen der Hände, Arme und des Gesichts Teil der allgemeinen Kommunikation. Doch wenn diese ohne spezielle symbolische, referenzierende oder interaktive Aufgabe sind, werden sie dieser Klasse zugeordnet.

- **Symbolische Gesten / Embleme.** Je nach Konvention und kulturellem Hintergrund zählen unterschiedliche Gesten mit einer definierten Bedeutung zu dieser Gruppe.
- **Interaktionale Gesten / Regulatoren.** Diese Gesten steuern die Interaktion mit einem Partner. Zum Beispiel um eine Interaktion zu initiieren, zu synchronisieren oder zu beenden.
- **Referenzierende Gesten / Deiktika.** Mit diesen Gesten werden Objekte, Personen, Richtungen oder Orte referenziert. Das kann durch gestisches Beschreiben des Objektes oder Teil des Objektes erfolgen oder durch eine Zeigebewegung mit einem Körperteil.

Vergleicht man die Taxonomien von Pavlovic u. a. [1997] und Nehaniv [2005], dann fällt auf, dass Pavlovic die Manipulationen zu den Gesten zählt, auch wenn er sie nicht dem Bereich der Kommunikation zuordnet. Diese Diskussion und Argumentation, warum Manipulationen von Objekten informationstragende kommunikative Akte sein können, wird in dem übernächsten Abschnitt (2.2.5) dargelegt.

Auch wenn Nehaniv eine gröbere und weniger scharf abgrenzende Taxonomie der Gesten für die MMI vorschlägt, lässt diese, auf die Intention einer Bewegung zielende Beschreibung, eine gute Einordnung zu.

2.2.3 Verstehen von Gesten

Was ist nötig, um Gesten zu verstehen? Warum kann zum Beispiel ein Mensch die Intention von Deiktika meist problemlos inferieren? Tomasello [2006] wirft die Frage auf, warum ein Menschenaffe begreift, dass etwas für ihn Interessantes in einem Eimer ist, wenn er sieht, wie ein Mensch versucht danach zu greifen, ohne dass der Mensch erfolgreich ist. Wird nur auf den Eimer gezeigt, erkennt der Affe die Intention hingegen nicht.

In der Psychologie wurde der Begriff *common ground* geprägt. Dieser beschreibt das notwendige Wissen beider Interaktionspartner, das ein Verstehen ermöglicht. Am Beispiel von Deiktika heißt dieses, dass der Rezipient B der Zeigegeste dem Zeigenden A ein kooperatives Verhalten unterstellt und weiß, dass das Gezeigte für ihn interessant ist. Des Weiteren muss aus der Situation hervorgehen, warum es gerade für den Rezipienten B interessant ist. Zeigt der Interaktionspartner A zum Beispiel auf einen Eimer, so hat diese Geste autonom keinen Informationsgehalt, erst im Kontext einer Frage wie „Wo ist das Tierfutter“ von B kann die Intention erschlossen werden. Auch müssen beide Partner wissen, dass der referenzierte Eimer ein typischer Aufbewahrungsort für das Tierfutter ist.

Auch Objektmanipulationen verstehen Menschen nur, weil sie Vorwissen über die manipulierten Objekte haben, also ihre typische Verwendung kennen. Wenn die Erwartung einer Manipulation oder Bewegung nicht erfüllt wird, entsteht Verwunderung, das untypische Verhalten muss erklärt werden, um wieder einen *common ground* für das Verständnis zu schaffen.

Der *common ground* in der zwischenmenschlichen Interaktion kann im Dialog hergestellt werden oder auf gemeinsamen Erfahrungen und auf Wissen über die involvierten Objekte

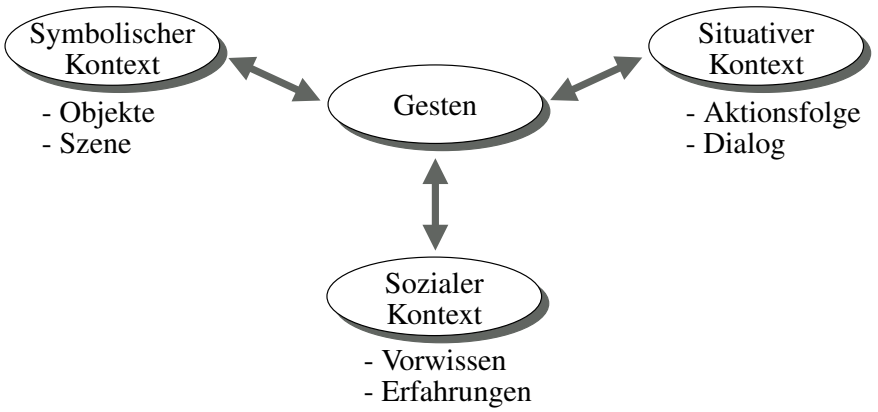


Abbildung 2.8: Gesten stehen in verschiedenen Kontexten, die das Verstehen ermöglichen. Das notwendige Wissen über diese Kontexte bildet bei einer erfolgreichen Kommunikation den *common ground* der Kommunikationspartner.

beruhen. Diese Kontexte, in denen eine Geste stehen kann, zeigt Graphik 2.8. Während das Verständnis emblematischer Gesten, z.B. ein Winken, stark vom *situativen* und *sozialen Kontext* abhängt, werden für Deiktika und manipulative Gesten Informationen über die Szene und die enthaltenen Objekte benötigt. Diese werden mit dem *symbolischen* Kontext beschrieben. Sprachliche Aspekte oder Erfahrungen können hingegen eine geringere Bedeutung haben.

Möchte man einen kooperativen Roboter entwerfen, der mit Menschen in einer sozialen und multimodalen Weise interagieren kann, müssen diese Teilgebiete, die einen *common ground* für die Interaktion bilden, modelliert werden und die Informationen aus den jeweiligen Kontexten fusioniert werden.

Ein Teil der Gestenforschung beschäftigt sich in diesem Zusammenhang intensiv mit der zeitlichen Synchronisation einzelner Phasen einer Geste mit der gesprochenen Sprache. Auf dieses umfangreiche Gebiet wird in dieser Arbeit nur am Rande eingegangen, eine genauere Darstellung liefert Kendon [2004]. In seinem Buch wird dargestellt, wann Gesten während einer Äußerung auftreten und welchen Einfluss eine Geste zum Beispiel auf Sprechpausen hat. Hingegen werden Gesten selten im symbolischen Kontext ihrer Umgebung betrachtet, dabei lässt dieser Kontext eine Eingrenzung der Bedeutung einer Geste zu und vermittelt wertvolle Informationen für das Verständnis einer Geste.

Im Kontext der vorliegenden Arbeit werden nun einige Teilgebiete der Gesten näher betrachtet: Die Zeigegesten oder auch Deiktika (2.2.4), Manipulationen von Objekten (2.2.5) und die Embleme, Gesten mit einer direkten sprachlichen Übersetzung (2.2.6).

2.2.4 Deiktika in der Kommunikation

Eine Zeigegeste steht immer im Kontext mit einem Objekt, einem Ort, einer Richtung oder einer Person und ist eine redebegleitende Geste, die eine sprachliche Äußerung ersetzt oder komplementiert.

Deiktika werden nicht nur mit der Hand ausgeführt, sondern auch mit dem Kopf, den Lippen oder dem Kinn, einem Ellbogen oder Fuß. Selbst wenn die Hand zum Zeigen benutzt

wird, existieren noch viele unterschiedliche Ausführungsvarianten der Geste: Zeigen mit dem Zeigefinger, mit der nach oben oder seitlich ausgerichteten Handfläche oder mit dem Daumen.

Clark [2003] unterteilt die Deiktika in verschiedene Klassen. Je nach verwendetem Körperteil und der Art der Ausführung ergibt sich eine andere Methode, wie die Aufmerksamkeit des Adressaten gelenkt werden kann. Eine Übersicht dieser Methoden gibt die Tabelle 2.1. Ausführlicher werden tiefergehende Bedeutungen spezieller Zeigegesten und kulturelle Unterschiede von Kendon [2004] beschrieben und anhand von Untersuchungen erläutert.

Körperteil	Index	Beispiel
Finger	Zeigen auf O	„**Das** ist das Buch.“
Arm	O präsentieren	„**All das** gehört dir.“
Kopf	nicken auf O	„Sie stand **dort**.“
Finger	tippen auf O	„**Das** ist das Buch.“
Fuß	tippen auf O	[auf eine Teppichprobe] „Ich mag **diese**.“
Torso	drehen zu P	„Lass **uns** reden.“
Gesicht	zuwenden zu P	[aufschauend] „Kann ich **Ihnen** helfen?“
Augen	P anblicken	„Ich möchte, dass **du** [Person A] mitkommst.“

Tabelle 2.1: Einige Methoden die Aufmerksamkeit mit Zeigegesten zu lenken. Hierbei steht „O“ für ein Objekt und „P“ für eine Person. (Nach Clark [2003])

So unterschiedlich die Ausführungen einer Zeigegeste auch sein können, weisen sie doch ein gewisses charakteristisches Bewegungsmuster auf. Im Kapitel 11 stellt Kendon [2004] fest, dass der Bewegungspfad — die Dynamik — so gestaltet ist, dass zumindest der letzte Teil der Bewegung linear verläuft. Des Weiteren verharrt der Körperteil, der die Geste ausführt, meist für einen kurzen Moment, wenn er seine größte Entfernung erreicht hat. Auch fällt bei der Bewegung auf, dass sie zielgerichtet verläuft, falls nicht auf ein bewegtes Objekt gezeigt wird.

Betrachten wir aus der Aufzählung möglicher Zeigemethoden einmal nur das objektreferenzierende Zeigen mit der Hand oder dem Finger, lässt sich feststellen, dass die Richtung dieser Zeigegeste sich nicht aus dem Finger, dem Unterarm oder der Blickrichtung alleine ergibt, sondern meistens eine Kombination dieser Zeigevektoren ist.

Deiktika in der MMI

Entsprechend den Variationen in der Interpretation und Ausführung von Deiktika existieren auch unterschiedliche Ansätze, Deiktika automatisch zu erkennen. Einige unterschiedliche Sichtweisen auf Zeigegesten werden hier exemplarisch veranschaulicht. Auf die technische und algorithmische Umsetzung wird an dieser Stelle verzichtet, sondern es wird die Art der zu erkennenden Gesten herausgestellt. Auf die algorithmische Umsetzung geht das Kapitel 3 ein.

Heidemann u. a. [2004] betrachten in ihrem Kontext Zeigegesten, die mit der Hand ausgeführt werden. Ihre Erkennung basiert hierfür auf einer ansichtsbasierten Erkennung und

Abbildung 2.9: Das Bild zeigt den Aufbau zur Erkennung der Zeigegeste und Richtung aus der Handstellung. (Bild aus Heidemann u. a. [2004])

Abbildung 2.10: Beispiel für eine Zeigegeste auf ein Objekt, bei der der Arm ausgestreckt wird

Klassifikation der Handstellung, der Ausrichtung und der Bewegung der Hand. In dem gewählten Aufbau (siehe Abbildung 2.9) zeigt der Benutzer auf Objekte, die vor ihm auf einem Tisch liegen. Die Hand wird von oben mit einer Videokamera aufgenommen. Relevant ist diese Art der Zeigegeste für Objekte, die einen geringen Abstand zum zeigenden Menschen haben und somit auch ohne ein Ausstrecken des Arms referenziert werden können. Zwei weitere Verfahren zum Erkennen der Handstellung und deren Anwendung werden im Buch von Kisacanin u. a. [2005] beschrieben.

Mehrere Forschergruppen beschäftigen sich mit dem Erkennen von Deiktika, bei denen der zeigende Arm ausgestreckt ist: Die Stellung der Hand findet in diesen Arbeiten wenig bis keine Beachtung. Diskutiert wird hingegen, wie sich aus einer solchen Zeigegeste die Richtung und somit das referenzierte Objekt ermitteln lässt, ein Beispiel zeigt Abbildung 2.10. Ein Erfolg versprechender Ansatz scheint die Sichtlinie vom Kopf oder den Augen über die Fingerspitze zu sein, wie es von den Gruppen um Kehl u. Gool [2004], Lee u. a. [2001] und Nickel u. Stiefelhagen [2003a] propagiert wird. Zwei weitere Alternativen, den Suchraum für ein referenziertes Objekt einzugrenzen, die Nickel u. Stiefelhagen [2004b] vorstellen, sind die Richtung des Unterarmes oder die Blickrichtung. In den Arbeiten von Howell u. Buxton [1998] und McKenna u. Gong [1998], die auf der gleichen Datenbasis arbeiten, wird eine Zeigegeste benutzt, um die Ausrichtung einer Kamera zu steuern. Eine hohe Genauigkeit der Zeigerichtung ist in diesem Szenario von geringerer Bedeutung. Des Weiteren kann auch die Richtung der Bewegung der Hand ausgewertet werden [Hofemann u. a., 2004].

2.2.5 Objektmanipulation als Form der Kommunikation

Laut der Definition von Ekman u. Friesen [1969] sind manipulative Akte keine Gesten, da mit einer solchen Handlung nicht kommuniziert wird. Pavlovic u. a. [1997] hingegen nehmen manipulative Akte mit in ihre Taxonomie von Gesten auf. Sie verorteten sie aber auf gleicher Höhe mit der Kommunikation, vertreten demzufolge auch die Meinung, dass Manipulationen nicht Teil der Kommunikation sind (siehe auch Abbildung 2.7). Einen

anderen Ansatz entwickelt Clark [2003] in seinem Kapitel zu „Pointing and Placing“[6]: Er betrachtet die Bedeutung von Deiktika und Objektplatzierungen in der zwischenmenschlichen Kommunikation und postuliert, dass die Kommunikation mit den Objekten der materiellen Umgebungen verankert ist. Außerdem stellt Clark heraus, dass es zumindest zwei Methoden gibt, um die Aufmerksamkeit des Gesprächspartners auf ein Objekt zu lenken: Zum einen mit den bereits behandelten Deiktika, zum anderen aber auch dadurch, dass Objekte bewusst platziert werden.

Bei dem Platzieren von Objekten betrachtet Clark drei Dimensionen: Welches Objekt wird platziert, wo wird es hingelegt und durch welche Aktion wird das Objekt platziert? Platziert eine Person ein Objekt, übermittelt sie mit diesem konkreten Objekt zusammen mit dem Ort, wo das Objekt hingelegt wurde, eine bestimmte Intention. Somit kann eine manipulative Geste ebenso wie Deiktika Bestandteil einer Kommunikation sein; mit ihr ist eine weitere Modalität gegeben. Clark erläutert das kommunikative Platzieren an einer Verkaufszene: Der Kunde legt seine Einkäufe auf den Tresen und stellt sich selbst vor diesen. Der Verkäufer weiß daher, wer der nächste Kunde ist und welche Gegenstände er auf die Rechnung setzen muss.

Aus dieser Betrachtung von Clark kann weiterhin abgeleitet werden, dass in manchen Interaktionssituationen nicht nur Platzierungen und Zeigegesten verwendet werden, sondern auch allgemeinere Objekthandlungen Teil der Kommunikation sind. Das ist der Fall, wenn die Manipulation nicht allein einem Selbstzweck dient, sondern damit dem Gesprächspartner eine Information übermittelt wird. Ein typisches Beispiel ist die Instruktion, bei der wichtige Handgriffe nicht nur verbal erklärt werden, sondern auch gezeigt werden. Demonstrieren Eltern ihren Kindern in der präverbalen Phase einen Konstruktionsvorgang, kommt den Bewegungen der Eltern eine besondere Bedeutung zu. Wichtige Bestandteile und Ziele können durch die Bewegung hervorgehoben werden. Mit dem spannenden Thema der Bewegungsadaption in der Kommunikation zwischen Eltern und Kind haben sich Brand u. a. [2002] beschäftigt; dies wird in Kapitel 6 aufgegriffen. Doch auch in allgemeinen Gesprächssituationen findet eine Beobachtung des Gesprächspartners statt und es wird, wenn auch oft unbewusst, mit Manipulationen kommuniziert.

In der Kommunikation ist das Beobachten und Verstehen der Gesten und Manipulationen des Gegenübers ein wesentlicher Bestandteil neben der Sprache. Über diese Handlungen wird eine Verbindung der Sprache mit der materiellen Welt hergestellt. Möchte man den Dialog zwischen Mensch und Maschine derart gestalten, dass die Maschine in der Umgebung des Menschen situiert ist, so ist es wichtig, auch Manipulationen und die bewegten Objekte zu erkennen.

2.2.6 Embleme

Embleme, die eine direkte sprachliche Übersetzung oder lexikalische Bedeutung haben, sollen in dieser Arbeit nur am Rande betrachtet werden. Grund hierfür ist, dass sie nur selten in einem Kontext zu Objekten stehen, wie es aber Deiktika und Manipulation tun. Ein Grund, trotzdem einfache befehlsartige Gesten wie das „Winken“ zu betrachten, liegt in dem Ziel einer intuitiven und natürlichen Interaktion zwischen Menschen und Robotern. Die Erfahrung zeigt, dass unvoreingenommene Benutzer einem sozialen, interaktiven

[6]Kap. 10 in dem Buch „Pointing: where language, culture, and cognition meet“ von Kita [2003]

Roboter zuwinken, auch wenn sie über die Fähigkeiten des Roboters wenig wissen. Das „Winken“ erregt, das hat die eigene Erfahrung den Benutzer gelehrt, die Aufmerksamkeit des Interaktionspartners, entspricht folglich einem „Hallo“. Eine weitere Anweisung, die auf ihre Interpretierbarkeit in einem Szenario mit Mensch und Roboter untersucht werden kann, ist die an den Roboter gerichtete „Stopp“-Geste. Also das nach vorne Führen der Hand, mit der offenen Handfläche zum Interaktionspartner.

Zusammenfassung

Aus dieser Betrachtung von Kommunikation und Gestik folgt, dass nonverbale Modalitäten und insbesondere Gesten ein wesentlicher Bestandteil der zwischenmenschlichen Kommunikation sind. Aufgrund ihres kommunikativen Charakters in der Interaktion werden im Folgenden die Manipulationen von Objekten in die Gruppe der Gesten subsumiert. Gesten sind für die MMI ein interessantes Gebiet, da sie Ambiguitäten auflösen und Informationen tragen können. Besonderes Augenmerk wird auf Gesten gelegt, die nicht nur im Kontext zur Sprache und Situation stehen, sondern auch im Kontext von Objekten der realen Welt. Diese Bindung von Zeigegesten und Manipulationen ist für soziale, interaktive Roboter relevant, da sie in der Umgebung ihres Interaktionspartners situiert sind.

Nach dieser spezifischen Betrachtung von Gesten, ihrer Verwendung und Bedeutung in der Kommunikation, soll nun auf die menschlichen Bewegungen bei der Ausführung von Gesten eingegangen werden. Ziel ist es, Strukturen und Regelmäßigkeiten zu finden, die dazu beitragen können, Gesten für die Interaktion zwischen Menschen und technischen Systemen zu erschließen.

2.3 Bewegungscharakteristik von Gesten

> Gesten werden in der zwischenmenschlichen Kommunikation intensiv genutzt und vom Interaktionspartner weitgehend verstanden. Natürlich existieren kulturelle Unterschiede, die sich insbesondere auf die Interpretation von Emblemen auswirken, doch die Intention von Deiktika und auch Manipulationen lässt sich meist aus der Interaktion erschließen. Da das dem Menschen möglich ist, lässt sich vermuten, dass die menschlichen Arm- und Handbewegungen und insbesondere die Gesten gewissen Charakteristika unterliegen.

Ein Teil dieser bedeutungstragenden Merkmale einer Geste wurde mit dem Kontext, in dem Gesten auftreten, erklärt, aber auch die Bewegung der Hand, sowie der strukturelle Aufbau und Ablauf einer Geste sind zum Erkennen wichtig. Des Weiteren ist es für Menschen leicht möglich, zwischen einer vorbereitenden Bewegung und der eigentlichen Geste zu unterscheiden. Die Fähigkeit trägt dazu bei, dass die beobachteten Bewegungen des Interaktionspartners schnell strukturiert und interpretiert werden können. Eine Beschreibung einer solchen hierarchischen Struktur von Gesten und Bewegungen wird in diesem Kapitel herausgearbeitet. An den Beispielen der Deiktika und manipulativen Gesten wird diese Struktur, sowie die Verbindung der Gesten zu ihrem Kontext beschrieben. Doch zuerst ist eine Beschreibung von Handbewegungen nötig und Charakteristika dieser müssen genannt werden.

2.3.1 Die Bewegung einer Geste

Bisher wurde die Geste als bedeutungstragende Bewegung betrachtet, die Teil der Kommunikation ist. Des Weiteren wurde die Bedeutung von Gesten in der Interaktion zwischen Robotern und Menschen herausgestellt sowie der Kontext, in dem Gesten benutzt werden, wurde beschrieben. Die Gesten, die in dieser Arbeit betrachtet werden, bestehen aus Bewegungen der Hand. Für die Interaktion zwischen Menschen und Computern und im Kontext der automatischen Erkennung von Gesten definiert Pavlovic eine Geste wie folgt:

Definition: Es sei $\mathbf{h}(t) \epsilon S$ ein Vektor, der die Haltung der Hand und/oder des Armes und ihre räumliche Position in einer Umgebung zum Zeitpunkt t im Parameterraum S beschreibt. Eine Handgeste wird durch ihre Trajektorie Z im Parameterraum S über ein bestimmtes Zeitintervall I repräsentiert. [Pavlovic u. a., 1997]

Diese Definition beschreibt allgemein die technische Repräsentation einer Arm- bzw. Handbewegung. Erst die Wahl des Intervalls lässt es zu, die Trajektorie als Geste zu verstehen.

Einen detaillierteren Blick in den Bewegungs- und Geschwindigkeitsverlauf von Handbewegungen des Menschen geben Sejnowski [1998] sowie Harris u. Wolpert [1998]. Sie untersuchen die Handbewegung, die als Trajektorie dargestellt wird. Die Bewegung des Armes und der Hand ist ein Kompromiss zwischen Zielgenauigkeit und Geschwindigkeit. Einerseits soll eine Greif- oder Zeigebewegung möglichst zielgenau sein, andererseits soll das Ziel möglichst schnell erreicht werden. Die Schwierigkeit besteht darin, dass eine starke Ansteuerung der Muskeln zu stärkerem Zittern und somit einer größerer Ungenauigkeit führt. Dieses Zittern überlagert die eigentliche Bewegung und kann in der Beobachtung der Bewegung als Rauschen aufgefasst werden.

Da zum Beispiel das Greifen kleiner Objekte eine höhere Präzision erfordert, hat das Objekt einen Einfluss auf die Geste. Gesten und ihre Bewegungen stehen folglich im symbolischen Kontext der involvierten Objekte. Die Beobachtungen von Harris u. Wolpert [1998], dass die Geschwindigkeit im zeitlichen Verlauf eine Glockenform aufweist, untermauern die These von Kendon, dass Bewegungen einer Geste meist symmetrisch sind. Des Weiteren hat dieser Verlauf den Vorteil, dass keine abrupten Richtungs- und Geschwindigkeitsänderungen vorgenommen werden, es entsteht eine weiche, glatte Bewegung. Die Graphiken 2.11 und 2.12 skizzieren eine Bewegung der Hand bei einer Zeigegeste beziehungsweise einem Winken in ihrem Geschwindigkeitsprofil. Die Bewegung der Hand bei der Ausführung einer Geste unterliegt sowohl den Bewegungsgewohnheiten des Menschen als auch dem Kontext, in dem die Geste ausgeführt wird.

2.3.2 Struktur der Bewegungen einer Geste

Einzelne Handbewegungen sind immer eingebunden in eine Bewegungshistorie und in den situativen sowie räumlichen Kontext. Das gilt auch für Gesten, die meist aus verschiedenen Phasen bestehen. Erst aus der Berücksichtigung der Folge von Bewegungen sowie ihrer Kontexte lassen sich häufig auftretende Ambiguitäten lösen und die Geste als Ganzes interpretieren. Kendon[7] beschreibt den Aufbau und die Charakteristika einer Geste, um diese von allgemeinen Bewegungen zu unterscheiden:

[7] siehe Kap. 7 im Buch „Gesture“ von Kendon [1996]

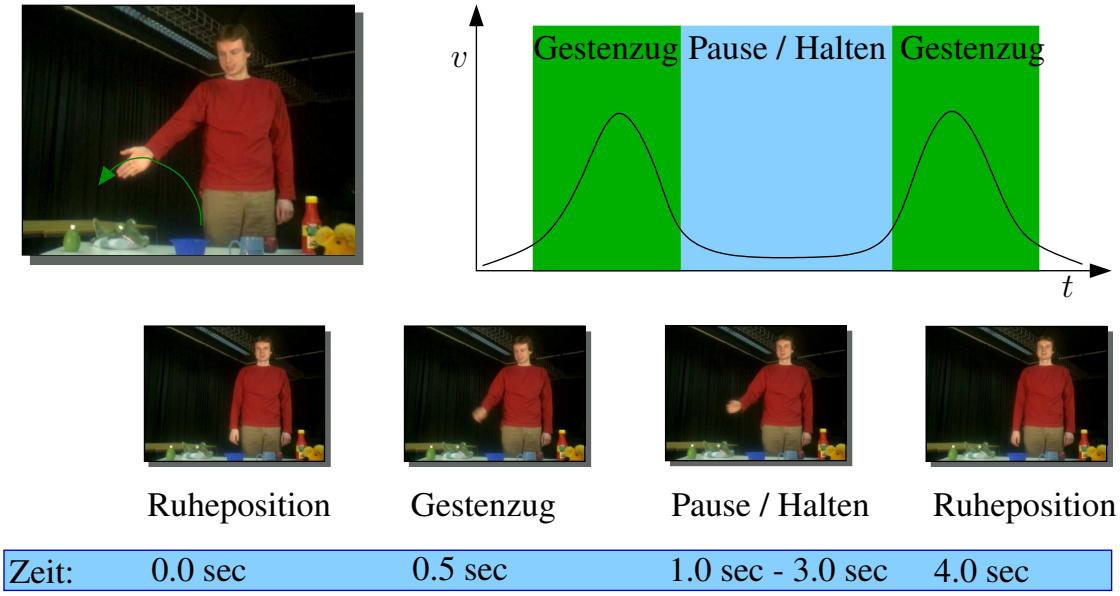


Abbildung 2.11: Eine Gesteneinheit in ihrem zeitlich Verlauf am Beispiel einer Zeigegeste.

- Der Körperteil, der eine Geste ausführt, kehrt zu seiner Ausgangsposition, der Ruheposition, zurück.
- Eine Geste hat immer einen Scheitelpunkt: das bedeutungstragende Zentrum. Es wird auch als *Gestenzug* (engl. stroke) bezeichnet.
- Gesten tendieren dazu, einen klaren Anfang und ein klares Ende zu haben. Veränderungen der Haltung geschehen dagegen meist fließend.
- Die Ausführung der Bewegung einer Geste erscheint symmetrisch. Für Menschen ist es schwer auszumachen, ob zum Beispiel eine isolierte Filmsequenz mit einer Geste vorwärts oder rückwärts präsentiert wird.

Diese Anhaltspunkte von Kendon ermöglicht die Unterteilung von Gesten in verschiedene Phasen, es lassen sich dadurch Intervalle für die Bewegungen einer Geste definieren. Kendon entwickelt eine strukturelle Taxonomie für Gesten: Die gesamte Geste bezeichnet er als *gesture unit* (*Gesteneinheit*). Sie umfasst alle Bewegungen, angefangen bei dem Verlassen der Ruheposition des Armes oder der Hand bis zur erneuten Entspannung in der Ruheposition. Die *gesture unit* kann eine oder mehrere *gesture phrases* (*Gestenphrasen*) enthalten. Eine solche Phrase wird für jede beobachtbare gestische Aktion identifiziert. Eine *gesture phrase* enthält somit auf jeden Fall eine *stroke phrase* (*Gestenzug*), also die Phase der gestischen Aktion und eine vorbereitende Bewegung, die *preparation*. Auch Pausen (*hold*) können ein Teil der *gesture phrase* sein, die nach der Vorbereitung beziehungsweise der gestischen Aktion eingefügt werden.

Als Beispiel ist in der Graphik 2.11 eine Zeigegeste mit einigen Bildern während der Ausführung und den Phasen dargestellt. Deutlich wird, dass die Hand wieder an ihren

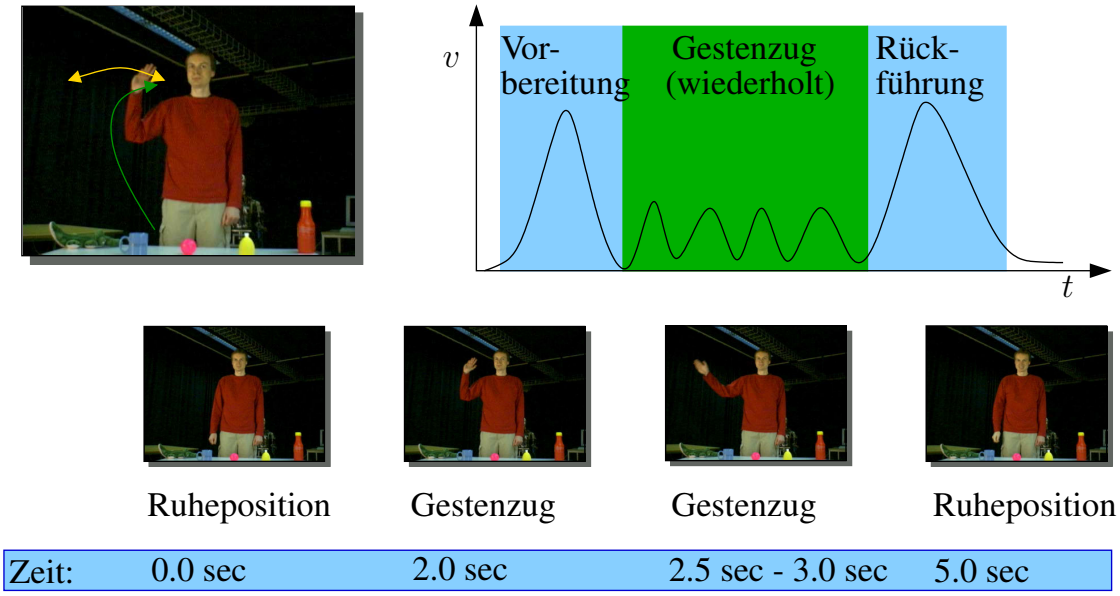


Abbildung 2.12: Die Aktivität „Winken“ mit ihren Bewegungen.

Ausgangspunkt zurückgeführt wird, aber auch eine Pause im Moment der größten Entfernung gemacht wird. An dem Diagramm des Geschwindigkeitsverlaufes und der Trajektorie wird exemplarisch die Symmetrie von Gesten sichtbar.

Die Terminologie von Kendon enthält bereits eine hierarchische Struktur, die einzelne Bewegungen zu komplexeren Sequenzen kombiniert. Bobick u. Ivanov [1998] bauen eine dreischichtige Hierarchie für allgemeine Bewegungen des Menschen auf, bestehend aus *Bewegungen* (engl. movement), *Aktivitäten* (engl. activity) und den *Aktionen* (engl. action). Jede Kategorie repräsentiert einen unterschiedlichen Grad der Erkennungskomplexität: Eine Bewegung weist nur geringfügige Variationen in der Ausführung auf und unterliegt normalerweise nur linearen Skalierungen. So zum Beispiel, wenn die Bewegung mit unterschiedlichen Geschwindigkeiten ausgeführt wurde. Diese Bewegung entspricht dem Gestenzug oder der vorbereitenden Bewegung aus Kendons Betrachtung von Gesten. Eine Aktivität beschreibt eine Sequenz von Bewegungen, kann aber komplexere, zeitliche Variationen enthalten. Die Abbildung 2.12 zeigt die Aktivität „Winken“, die aus den einzelnen Bewegungen „Heben“, „Schwenken“ und „Senken“ der Hand besteht. In der mehr auf die Semantik fokussierten Terminologie von Kendon besteht diese Aktivität aus einer Vorbereitung, dem Gestenzug und der Rückführung in die Ruheposition. Sowohl Bewegungen als auch Aktivitäten stehen nach Bobicks und Ivanovs Definition in keiner Verbindung zu Elementen, die nicht zum Körper des Menschen gehören, der die Bewegung oder Aktivität ausführt.

Interessant für die Betrachtung von Deiktika und Manipulationen sind die Aktionen, da eine Aktion definiert ist als eine Aktivität in Verbindung mit einer symbolischen Information, das kann zum Beispiel ein referenziertes Objekt sein. Eine Zeigegeste „Zeigen auf Objekt X“ kann entsprechend mit diesem Bewegungsschemata beschrieben werden: In die-

sem Beispiel sind das „Hinführen“ und „Entfernen“ der Hand Bewegungen. Die Sequenz aus diesen und einer eingefügten Pause ist eine Aktivität. Die Kombination aus dieser Aktivität und dem referenzierten Objekt ergibt eine Aktion.

Bobick u. Ivanov [1998] haben mit ihrer Hierarchie eine allgemeine Beschreibung von Bewegungen entwickelt, die auf alle Aktionen des Menschen angewandt werden kann. Das umfasst Gesten aber auch Aktionen wie Gehen und Laufen.

Für die Deiktika und Objektmanipulationen wird in den folgenden Abschnitten die hierarchische Struktur von Bobick u. Ivanov [1998] erläutert und die Einbeziehung des symbolischen Kontextes aufgezeigt. Eine Ähnlichkeit von manipulativen und deiktischen Gesten findet sich in dem „Greifen“, „Berühren“ und „Zeigen“ eines Objektes, da diese Bewegungen von dem menschlichen Bewegungsablauf und der Objektposition bestimmt werden.

2.3.3 Zeigegesten

Aus Fitts [1954] Untersuchung zur Zeigegenauigkeit am Bildschirm und weitergehenden Untersuchungen zum Beispiel von Accot u. Zhai [2003] kann abgeleitet werden, dass beim Zeigen mit der Hand auf Objekte die Größe des Objektes die Bewegung beeinflusst. Denn für ein kleines Objekt muss eine größere Genauigkeit der Handpositionierung erreicht werden. Dass kann nur über eine verringerte Geschwindigkeit am Ende der Bewegung erreicht werden. Dieser Gedankengang lässt sich auf das Greifen von und Zeigen auf Objekte in realen Umgebungen verallgemeinern: Das Objekt, in dessen Kontext die Bewegung ausgeführt wird, hat einen Einfluss auf die Art und Ausführung der Zeigegeste. Um diese These zu untermauern, sind weitere Forschungen nötig. Einen weiteren Hinweis hierauf geben Ergebnisse aus Kendons Forschung. Er hat beobachtet, dass je nach referenziertem Objekt unterschiedliche Ausführungen von Zeigegesten auftreten und dass auch verschiedene Handstellungen verwendet werden. Beispielsweise hat Kendon beobachtet, dass auf einzelne Individuen eines Objektes mit dem Zeigefinger gezeigt wird. Soll hingegen eine Gruppe referenziert werden, wird die offene Hand benutzt.

Die drei Phasen einer deiktischen Geste, bei der der Arm ausgestreckt wird und mit der offenen Hand oder dem Zeigefinger ein Objekt referenziert wird, zeigt auch die Abbildung 2.11. Eine solche Zeigegeste besteht folglich aus zwei Bewegungen und einer Pause, die ihre gestische Bedeutung aus dem situativen Kontext erhält. Dieses sind das Ausstrecken des Armes, also die vorausgehende Bewegung und eventuelle verbale Äußerungen des Akteurs. Da eine Zeigegeste im situativen Kontext des referenzierten Objektes steht, ist es erstrebenswert, die Bewegung der Hand mit dem Objekt zu verbinden. Wenn aber eine sich bewegende Hand beobachtet wird, so kann diese sich an unterschiedlichen Objekten entlang bewegen, doch nur ein spezielles Objekt wird referenziert. Die Bindung des Objektes an die Zeigegeste muss folglich in der räumlichen Relation zum Ende der Geste erfolgen. Dieses Prinzip wird schematisch in der Graphik 2.13 gezeigt. Den Suchbereich für Objekte dynamisch und abhängig von dem zeitlichen Verlauf der Geste zu gestalten, hat den Vorteil, dass zu Zeitpunkten, an denen die Hand beschleunigt oder schnell bewegt wird, kein Bereich definiert werden muss. Auch kann der Abstand des erwarteten Objektes zur Hand sowie die Richtung zu diesem definiert werden und somit die Suche nach dem referenzierten Objekt erleichtern. Dieser Ansatz wird später in dieser Dissertation aufgegriffen und das entwickelte System, das die Bewegungs- und Objektinformation fusioniert, wird vorgestellt und evaluiert.

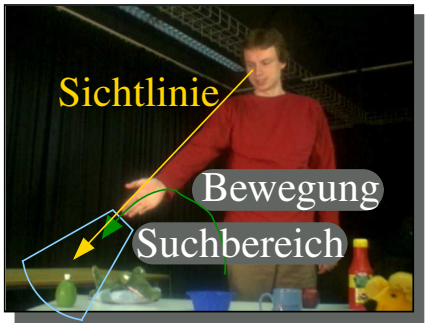


Abbildung 2.13: Schematische Darstellung des Suchbereichs. In diesem kann ein Objekt, das mit einer Zeigegeste referenziert wird, erwartet werden.

Zur Bestimmung der Zeigerichtung und damit der Ausrichtung des Suchbereichs gibt es unterschiedliche Ansätze. Die Richtung kann einerseits aus der Bewegungsrichtung der Hand geschlossen werden. Es kann aber auch die Blickrichtung oder die Sichtlinie als gedachte Linie zwischen den Augen und der Fingerspitze genommen werden. Des Weiteren können die Ausrichtungen des Unterarms oder des Zeigefingers Aufschluss über die Zeigerichtung geben.

Eine interessante Untersuchung zur Auflösung der Richtung von Deiktika wird von Kranstedt u. a. [2006] vorgestellt. In einer Experimentalumgebung wird die Position der Hand des Menschen im Raum optisch über Infrarotmarkierungen und ein aufwendiges Mehrkamerasystem erfasst, sowie die Stellung der Hand und der Finger über einen Datenhandschuh ermittelt. Während der Untersuchung zeigt der Mensch auf Objekte, die gleichmäßig auf einem Tisch liegen. Aus diesen Daten werden unterschiedliche Hypothesen der Zeigerichtung ermittelt und aus dem Abstand zu den anderen Objekten ein Kegel der Zeigerichtung berechnet (engl. pointing cone). Kranstedt u. a. [2006] unterscheiden zwei Arten von Zeigegesten, die auch in unterschiedlichen Öffnungswinkeln der Kegel für die Zeigerichtung resultieren. Zuerst wird eine Zeigegeste auf ein Objekt angenommen, entsprechend wird das referenzierte Objekte in einem kleinen Kegel (Öffnungswinkel von 12°) gesucht. Wenn in diesem Bereich kein Objekt vorhanden ist, wird ein Zeigen auf eine Region vermutet und der Winkel des Kegels wird auf 25° vergrößert. In den Studien konnten diese empirischen Werte bestätigt werden.

Das Konzept des Kegels für die Zeigerichtung zielt in die gleiche Richtung wie die in dieser Arbeit vorgestellte Definition eines Suchbereichs, der relativ zur Handbewegung ausgeprägt ist (siehe auch Hofemann u. a. [2004] und Fritsch u. a. [2004]).

2.3.4 Objektmanipulationen

Objektmanipulationen bestehen meist aus einer Greif- oder Berührbewegung der Hand gefolgt von einer Manipulation oder einem Bewegen des Objektes. Abgeschlossen wird die Sequenz von einem Ablegen und Loslassen. Teile dieser Bewegung sind stark von dem Objekt und seinen Manipulationsmöglichkeiten abhängig. Einflussfaktoren der Bewegung sind die Möglichkeiten, wie ein Objekt erfasst, bewegt oder bedient wird. Diese Bewegungen sind sehr variabel und werden von dem Objektkontext so wie der Situation in

ihrer Bewegungsform, -dauer und -richtung bestimmt. Wenn zum Beispiel ein Objekt von Position A zu Position B bewegt wird, ist die Richtung und Länge von diesem situativen Kontext abhängig, andererseits unterliegt auch diese Bewegung dem Bestreben, eine möglichst glatte Bewegung auszuführen.

Beobachtet man die Bewegung einer Hand, die ein Objekt berührt oder greift, stellt sich die Frage, wie das Objekt in der Szene gefunden werden kann. Ein Problem bei dieser Aufgabe ist das Verdecken von Objekten durch die Hand des agierenden Menschen. Menschliche Interaktionspartner können die referenzierten Objekte benennen, selbst wenn sie im Moment des Greifens durch die Hand verdeckt sind. Andererseits wissen sie auch, dass nicht jedes von der Hand oder dem Arm aus ihrer Sicht verdeckte Objekt gegriffen wurde. Doch aus der Bewegung der Hand kann geschlossen werden, wann die Hand langsamer wird um zu greifen oder das Objekt zu berühren. Es besteht also eine Erwartung der Objektposition relativ zur sich bewegenden Hand. In diesem Suchbereich kann nach potentiellen Objekten gesucht werden.

Für das automatische Erkennen von Gesten stellt die Einbeziehung des symbolischen Kontextes ein Problem dar. Denn eine Szenenanalyse, die auf einer Objekterkennung basiert, liefert nur die Objekte, aber keine Referenz auf das manipulierte Objekt. Verschwindet ein Objekt aus der Repräsentation der Szene, kann das mehrere Ursachen haben: Das Objekt kann verdeckt oder gegriffen worden sein oder die Objekterkennung kann das Objekt nicht mehr sicher detektieren. Soll aber ein Auflösen der Objktereferenz möglich sein, muss die Bewegung der Hand mit einbezogen werden, denn das gegriffene Objekt muss sich kurz vor dem Ende der Bewegung in der Bewegungsrichtung der Hand befunden haben. Wird dieses Kontextwissen benutzt, kann eine sichere Hypothese für das manipulierte Objekt aufgestellt werden. Ist außerdem bekannt, wie ein Objekt gegriffen oder manipuliert werden kann, kann das Aufschluss über die Bewegung und den Suchbereich geben. Zum Beispiel kann ein Schalter oder Hebel nur aus einer Richtung bedient werden, das kann zur Eingrenzung der eventuell manipulierten Objekte genutzt werden.

Des Weiteren sind die Manipulationen in den situativen Kontext eingebettet. Die vorherigen Aktionen können eine Auswirkung auf die Objekte der Szene haben. Beispielsweise werden Objekte in der Szene bewegt oder ihr Zustand wurde durch eine Handlung verändert. Diese Veränderungen haben wiederum Auswirkungen auf die nun möglichen und zu erwartenden Handlungen. Ein einfaches Beispiel, welches die Komplexität dieser Betrachtung aufzeigt, sei im Folgenden kurz skizziert: Eine Person ergreift eine Flasche, diese ist nun fast komplett durch seine Hand verdeckt. Die Flasche haltend führt der Mensch nun seine Hand in Richtung einer Tasse. Diese kann natürlich nicht mit dieser Hand ergriffen werden, da die Person bereits die Flasche hält. Es wird etwas aus der Flasche in die Tasse gegossen, so dass die Tasse nun gefüllt ist. Anschließend wird die Flasche wieder abgestellt. Diese einfache Sequenz besteht bereits aus vier bis fünf Bewegungen, drei Aktivitäten und enthält viele Stellen, an denen ein symbolischer oder situativer Kontext zum Verständnis notwendig ist.

Schlussfolgerungen

Kommunikation und Interaktion ist ein gewöhnlicher Vorgang im alltäglichen Miteinander unter Menschen. In der Absicht auch Computer für die situierte Interaktion mit Menschen

zu benutzen, müssen Computersysteme und Roboter mit multimodalen Interaktionsmöglichkeiten ausgestattet werden. Ein Teil hiervon sind die Gesten, die unter Verwendung optischer Methoden als eine intuitive, nicht störende Modalität eingebunden werden können. Nach der Betrachtung unterschiedlicher Klassifizierungen von Gesten wurde der *common ground* und der Kontext, in dem eine Geste steht, als Grundlage für das Verstehen von Gesten vorgestellt. Das bedeutet für das automatische Erkennen von Gesten, dass sowohl sprachliche Informationen als auch Wissen über die Objekte der Szene mit einfließen.

Zusammengefasst beginnt eine Geste meistens in einer Ruheposition, gefolgt von einer hinführenden Bewegung und dem Gestenzug. Anschließend tritt oft eine Pause ein, danach wird die Hand in die Ruheposition zurückgeführt. Eine hierarchische Betrachtung allgemeiner menschlicher Handbewegungen ist die Unterteilung in Bewegung, Aktion und Aktivität (siehe Bobick u. Ivanov [1998]). Zur Erinnerung sei gesagt, dass die Unterteilung an dem Kontextwissen festgemacht werden kann, das für die Interpretation notwendig ist.

Gesten sind ein Teilgebiet der Interaktion und Kommunikation. Hierbei werden Manipulationen bewusst zu den Gesten gefasst, da mit ihnen in der Interaktion Informationen übertragen werden können. Ihr Aufbau ist anders gestaltet als bei Zeigegesten, gemein ist ihnen aber, dass sie in einem Kontext mit den Objekten der Szene stehen und für die Interpretation ein *common ground* nötig ist.

3. Das Erkennen von Handgesten

Ein Ziel dieser Arbeit ist die videobasierte Erkennung von Gesten. Die Bedingungen und Herausforderungen dieser Aufgabe werden beschrieben (3.1) und danach wird auf die Grundlagen, das Themengebiet des Erkennens von Bewegungen (3.2), eingegangen. Anschießend werden einige spezielle Verfahren für das Erkennen von Gesten und Aktionen (3.3) vorgestellt und diskutiert.

Dariu Gavrila [1999] sieht das Erkennen von Menschen und ihrer Aktivitäten als eine Schlüsselfähigkeit für Maschinen, damit diese intelligent und effizient in der Umgebung von Menschen interagieren können. Diesem ambitionierten Ziel näher zu kommen, ist unter anderem Thema dieser Arbeit. In diesem Zusammenhang muss nun allgemein auf das computerbasierte Erkennen von Bewegungen und insbesondere auf das Erkennen von Gesten eingegangen werden. Das vorliegende Kapitel gibt einen Übersicht über unterschiedliche Verfahren der Bildverarbeitung, Musteranalyse und -klassifikation, die allgemein für die Bewegungserkennung eingesetzt werden können.

3.1 Grundlagen der videobasierten Gestenerkennung

Bereits in Kapitel 2 wurde zum einen erläutert, dass sich der Mensch bisher stark an die begrenzten Interaktionsmöglichkeiten von Computern und Robotern anpassen muss, zum anderen wurde die große Bedeutung von Gesten und Handlungen für die zwischenmenschliche Interaktion herausgearbeitet. Das Themengebiet der automatischen Gestenerkennung und daraus resultierende Anforderungen an die Systementwicklung werden nun geschildert.

Wie Clark [2003] beschreibt, ist die Kommunikation mit der materiellen Welt verankert. Diese Verbindung der Kommunikation mit den Gegenständen und Orten besteht nicht nur aus Zeigegesten sondern auch aus Objektmanipulationen. Neben der Sprache ist die

Verankerung auch für die Mensch-Maschine-Kommunikation essentiell, da mit ihr der zum Verständnis notwendige *common ground* erstellt werden kann. Erst mit Hilfe der visuellen Erkennung und Interpretation seiner Umgebung und seiner Kommunikationspartner kann zum Beispiel ein Roboter seine Umgebung wahrnehmen und darauf entsprechend reagieren.

Das Einsatzgebiet der automatischen Erkennung von Gesten und Aktionen bedingt spezielle technische Anforderungen, denen die verwendeten Algorithmen und Systeme gewachsen sein müssen. Insbesondere bei mobilen Systemen, wie sie Roboter für die MRK oder auch tragbare Assistenzsysteme darstellen, ist die Rechenleistung begrenzt. Aber eine flüssige Kommunikation erfordert schnelle Reaktionen des Systems, damit das Ziel einer möglichst großen Ähnlichkeit zur Interaktion zwischen Menschen erreicht werden kann. Diese konträren Anforderungen bei gleichzeitig hoher Erkennungsleistung müssen bei der Entwicklung eines Systems und bei der Wahl der Algorithmen bedacht werden.

Eine weitere Herausforderung für das automatische Erkennen von Handbewegungen eines Menschen besteht darin, dass diese Bewegungen keine klaren Start- und Endpunkte aufweisen und einer hohen Variabilität in der Ausführung unterliegen. So werden Bewegungen mit dem gleichen Ziel von einzelnen Personen unterschiedlich ausgeführt, aber auch der aktuelle räumliche und situative Kontext beeinflusst die Ausführung der Bewegung. Einige Systeme zur Gestenerkennung beschränken sich deswegen auf statische Hand- oder Körperkonfigurationen. Doch wie Pavlovic u. a. [1997] richtig feststellen, sind Handbewegungen dynamische Aktionen und sollten auch als solche interpretiert werden. Des Weiteren sollte ein Verfahren zur Aktionserkennung auch die von Clark [2003] postulierte Verankerung der Kommunikation mit der Umgebung schon während der Erkennung beachten. Hierfür ist auch die Integration der Gestenerkennung in ein Gesamtsystem mit mehreren Modalitäten nötig, so dass auch ein *common ground* für das Interpretieren der Gesten verwendet werden kann. Die Integration in ein komplexes System wiederum impliziert Anforderungen an die Implementierung, weshalb ein modularer Aufbau der Software von Vorteil ist.

Sollen Bewegungen automatisch erkannt werden, ist eine Vielzahl an Beispielsbewegungen unerlässlich, mit denen die Beobachtungen verglichen werden können. Das Erstellen eines solchen Trainingssets ist oft sehr aufwendig, da unterschiedliche Benutzer die Handlungen häufig wiederholen müssen. Diese Trainingsdaten müssen zusätzlich meist manuell segmentiert und annotiert werden. Ein Vorteil ist es, wenn das System auch mit kleinen Beispielmengen auskommt und diese wenigen Daten eine generalisierte Erkennungsleistung ermöglichen. Unter diesen Bedingungen kann ein Erkennungssystem schnell angepasst und in alternative Systeme integriert werden.

In der Literatur findet man verschiedene Ansätze, die versuchen, diesen Ansprüchen gerecht zu werden. Diese Systeme haben jedoch im Vergleich zu den Erkennungsleistungen des Menschen nur geringe Möglichkeiten und sind meist auf eine Domäne spezialisiert oder restringiert. Menschen ist es beispielsweise bereits möglich, aus nur wenigen bewegten Punkten die zugrundeliegende Bewegung zu erkennen. Dieses wurde in der Psychologie mit den so genannten *Johanssons' moving lights displays* gezeigt (siehe Johansson [1973]). Diese Möglichkeiten für ein Echtzeitsystem, zum Beispiel einem mobilen, interaktiven Ro-

boter, nutzbar zu machen, ist das Ziel der automatischen Aktionserkennung. Im Folgenden wird auf die Komponenten dieses Erkennungsprozesses eingegangen.

3.2 Analyse und Klassifikation von Bewegungen

Die videobasierte Gesten- und Handlungserkennung, die für diese Arbeit entwickelt wurde, ist ein Teilgebiet der videobasierten Erkennung von Bewegungen des Menschen. In diesem Unterkapitel werden Überlegungen und Methodiken zu dieser allgemeinen Betrachtung von Bewegungen diskutiert. Im Anschluss wird der Prozess des Erkennens von Bewegungen und Algorithmen, die in diesen Prozess eingebunden werden können, erläutert.

Erste Arbeiten in der Psychologie zur Analyse menschlicher Bewegungen gehen auf Johansson [1973] zurück: Allein die Bewegung von Lichtpunkten, die an den Gelenken eines Menschen fixiert sind, kann ein Mensch interpretieren. Führt der Mensch, der die Markierungen trägt, hingegen keine Bewegung aus, so lässt sich sein Körper aus diesen Punkten nicht erkennen. Auch wenn die Lichtpunkte andere Bewegungen ausführen, wird selten eine strukturierte Bewegung der Punktgruppe erkannt. Da es dem Menschen möglich ist, einzelne Lichtpunkte aus ihrer Bewegung in einen sinnvollen Zusammenhang zu setzten und somit die zugrundeliegenden Bewegungen und Aktionen zu erkennen, ist es von Interesse, diese Fähigkeit mit einem technischen System nachzubilden. Die Anwendungsgebiete für ein solches automatisches Erkennen und Analysieren menschlicher Bewegungen umfassen einen großen Bereich von medizinischen Bewegungsanalysen über die Sportmedizin bis hin zu Überwachungsaufgaben und die multimodale Mensch-Maschine-Interaktion. Insbesondere für die multimodale Interaktion zwischen Menschen und Maschinen eröffnen sich neue Möglichkeiten, die dazu beitragen können, den zwischenmenschlichen Interaktionsformen näher zu kommen.

Es wurden bereits verschiedene Ansätze und Verfahren entwickelt, die sich mit dem Problem der Bewegungserkennung eines Menschen aus Bildsequenzen beschäftigen. Der Überblick über Methoden und notwendige Komponenten, den dieses Kapitel gibt, ist inspiriert von den Arbeiten von Cédras u. Shah [1995], Gavrila [1999] sowie Pavlovic u. a. [1997]. Sie geben einen allgemeinen Überblick zur bewegungsbasierten Erkennung. Pavlovic und Kollegen fokussieren ihre Betrachtung auf die Erkennung und Interpretation von Gesten. Bobick [1997] bringt mit dem Begriff der Wahrnehmung einen weiteren Aspekt mit ein. Er untersucht den Einfluss von Wissen und Kontext für die Erkennung von Bewegungen auf unterschiedlichen Abstraktionsebenen. Mit dieser Arbeit verbindet Bobick den Bereich der Musterklassifikation mit den Kognitionswissenschaften. Der Prozess der Erkennung menschlicher Bewegungen und Aktionen wird nun betrachtet und es werden unterschiedliche Ansätze und Techniken für die jeweiligen Gebiete vorgestellt.

Aus der Betrachtung von Johanssons Arbeiten zur Bewegungswahrnehmung des Menschen ist der Ansatz entstanden, dass Aktionen erkannt werden können, ohne die Struktur des Körpers aus der Bildsequenz zu rekonstruieren. Alleine wenige Punkte des Körpers, die über die Zeit in der Bildsequenz verfolgt werden, reichen aus, um eine Bewegung oder Aktion, die ein Mensch ausführt, zu erkennen. Im Gegensatz zu diesen rein ansichtsbasierten Verfahren wird in einer anderen Forschungsdisziplin versucht aus einer oder

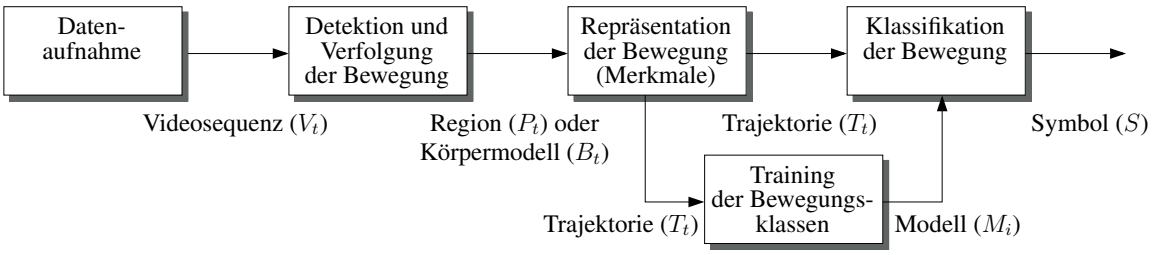


Abbildung 3.1: Schematischer Aufbau eines System zur Bewegungserkennung.

mehreren Bildsequenzen die Konfiguration des menschlichen Körpers mit einem Modell, das diesen beschreibt, zu rekonstruieren. Basierend auf einem solchen Modell kann dann eine Bewegung erkannt werden. Die Verfahren unterscheiden sich nicht unbedingt in den Bildverarbeitungsmethoden, die verwendet werden, jedoch in dem, was aus dem Bild rekonstruiert wird und wie viel Wissen über den beobachteten Körper einfließt. Dies können zum Beispiel nur Regionen oder Kanten sein, aber auch eine detaillierte Rekonstruktion der dreidimensionalen Körperkonfiguration.

Nach dem ersten Schritt der Detektion und der Verfolgung markanter Körperteile oder des Körpers liegt entweder eine Sequenz von Punkten im Bild vor, also eine Trajektorie eines Körperteils, oder es existiert die räumliche und zeitliche Dynamik der Körperkonfiguration. Den nachfolgenden Prozess kann man nach Cédras u. Shah [1995] in zwei Schritte unterteilen: Erstens muss eine adäquate Repräsentation für die gesuchten Bewegungen gefunden werden, zweitens muss eine unbekannte Beobachtung mit den bekannten Modellen verglichen und bewertet werden. Für letztere Aufgabe werden häufig typische Musterklassifikationsalgorithmen verwendet.

Die Teile des in Abbildung 3.1 dargestellten Prozesses, von der Datenaufnahme über die Detektion und Verfolgung, zur Repräsentation und der Klassifikation, werden in den nächsten Absätzen separat behandelt. Die Teile sind jedoch untereinander verzahnt, so dass eine Technik der Detektion und Verfolgung die Wahl der Repräsentation beeinflussen kann.

3.2.1 Datenaufnahme

Die Aufnahme der Daten erfolgt normalerweise über Videokameras, so dass als Ausgangsmaterial eine Bildsequenz vorliegt. Andere Verfahren, wie zum Beispiel optische Markierungen oder Datenhandschuhe, ermöglichen eine abstrahierte Aufnahme der Bewegung und deren Repräsentation in den drei Raumdimensionen. Jedoch sind die technischen Voraussetzungen höher und der beobachtete Mensch muss sich speziell an das System anpassen. Auch Stereo- oder Mehrkamera-Systeme werden für die Gestenerkennung eingesetzt, sind aber aufgrund ihrer baulichen und technischen Anforderungen in der Mensch-Roboter-Kommunikation nur bedingt zu verwenden.

Faktoren, die Einfluss auf den Verarbeitungsprozess und das Laufzeitverhalten nehmen, sind Bildgröße und Aufnahmefrequenz sowie die Abbildungsparameter der Kamera. Hierbei ist auch darauf zu achten, ob das Kamerasystem kalibriert werden muss, dass heißt, ob

die Abbildungsparameter der Kamera für die aktuelle Konfiguration ermittelt werden müssen. Ist das für die Aufnahme und weitere Verarbeitung nicht nötig, erspart dieses nicht nur eine aufwendige Kalibration, sondern erlaubt auch, das System generischer anzuwenden. Für Echtzeitsysteme muss bei der Aufnahme ein Kompromiss zwischen Detailgenauigkeit sowie zeitlicher Auflösung und der Berechenbarkeit gefunden werden. Auch Abschattung und Beleuchtung sind Faktoren, die die nachfolgende Erkennung beeinflussen. Natürlich sollten die Algorithmen auf einem breiten Spektrum von Szenarien anwendbar sein.

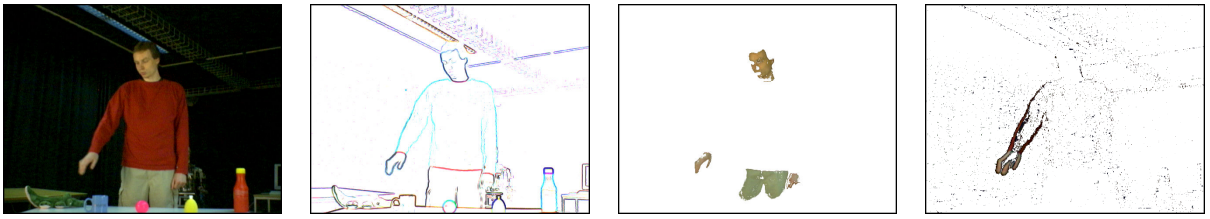
3.2.2 Detektion und Verfolgung von Bewegungen

Grundlage jeder Bewegungsklassifikation, wie sie auch im Rahmen dieser Dissertation entwickelt wurde, ist das Detektieren von Objekten oder Objektteilen und das Verfolgen dieser im zeitlichen Verlauf. Die Verfolgung hat zwei Vorteile: Zum einen ermöglicht das Wiederfinden eines Objektes in den einzelnen Bildern einer Sequenz die Rekonstruktion der Bewegung, zum anderen wird durch geschickte Algorithmen zur Verfolgung der Suchraum im Bild eingeschränkt und so der Rechenaufwand, verglichen mit einer Suche im kompletten Suchraum, reduziert. Das verfolgte Objekt, zum Beispiel die hautfarbene Region der Hand in der Bildsequenz, muss nur im Umfeld der letzten Position gesucht werden und nicht im gesamten Bild. Außerdem löst das gleichzeitige Verfolgen mehrerer Objekte das Problem der eindeutigen Referenzierung der einzelnen Objekte über die Zeit.

Die Methoden, die zum Detektieren und Verfolgen eingesetzt werden, hängen stark von dem Anwendungsgebiet und der gewünschten Rekonstruktion ab. Anhand der Rekonstruktion und dem hierfür notwendigen Wissen über das verfolgte Objekt können zwei Gruppen von Methoden zur Detektion und Verfolgung unterschieden werden: Die rein ansichtsbasierten und die modellbasierten Verfahren. Welches Wissen ist zum Beispiel über den menschlichen Körper oder eine Hand nötig, reicht die typische Hautfarbe als Merkmal aus oder müssen die Größen und Stellungen der Gliedmaße bekannt sein? Diese Unterteilung basiert auf den Beschreibungen von Pavlovic u. a. [1997] und Gavrila [1999]. Pavlovic unterscheidet nur ansichtsbasierte und modellbasierte Verfahren, Gavrila unterteilt die modellbasierten Verfahren weiter in zweidimensionale und dreidimensionale Ansätze.

Für die rein **ansichtsbasierten Verfahren** gilt die Annahme, Menschen oder relevante Körperteile des Menschen in einer Bildsequenz finden und verfolgen zu können, ohne den gesamten Körper mit allen Körperteilen zu finden. In manchen Anwendungen reicht es aus, Gesichter oder Hände mit einem Mustervergleich zu finden, der Rest des Körpers ist dafür jedoch nicht relevant. Andere Anwendungen hingegen erfordern nur das Erkennen ob ein Mensch im Bild ist; die einzelnen Körperteile müssen hingegen nicht unterschieden werden. Demnach ist es das Ziel, charakteristische Bildmerkmale eines Körperteils oder des Körpers im Bild auszumachen und diese als Repräsentanten zu detektieren und zu verfolgen. Ein entscheidender Vorteil dieser Verfahren ist zum einen, dass gängige Verfahren der Bildverarbeitung verwendet werden können, zum anderen wird von dem komplexen Körper des Menschen abstrahiert und die Analyse beruht auf einer kompakten und technisch handhabbaren Repräsentation.

Merkmale, die von Algorithmen der Bildverarbeitung ausgenutzt werden können, sind zum Beispiel Intensitätsschwellwerte oder Farbhistogramme. Mit diesen Verfahren können bestimmte Regionen, wie zum Beispiel die Hände, in einer Bildsequenz segmentiert



(a) Originalbild

(b) Detektion der Kanten mit einem Sobel-Filter

(c) Hautfarbene Regionen im Bild

(d) Differenzbild aus zwei aufeinander folgenden Bildern

Abbildung 3.2: Einige Beispiele der Merkmalsextraktion aus Bildern.

und verfolgt werden. Die Segmentierung des Bilds in hautfarbene Regionen, siehe Abbildung 3.2(c), und das Verfolgen dieser wird zum Beispiel von Fritsch [2003] beschrieben. Dieses Verfahren nutzt aus, dass die Hautfarbe in einem eingeschränkten Bereich in dem helligkeitsnormierten RG-Farbraum liegt. Aber auch Verfahren zur Kanten- und Konturdetektion und für Differenzbilder bringen relevante Merkmale, siehe Abbildung 3.2(b) und 3.2(d). Aus der Differenz zweier oder mehrerer aufeinander folgender Bilder einer Sequenz kann die Bewegung im Bild berechnet werden, der optische Fluss. Dieses Verfahren wird unter anderem von McKenna u. Morrison [2004] für den Verlauf hautfarbener Regionen aufgegriffen. Auch das von Rowley u. Rehg [1997] vorgestellte Verfahren basiert auf dieser Technik. Einige Beispiele für diese gängigen Bildanalyseverfahren werden in der Graphik 3.2 gegeben.

Bei der Verfolgung von vorher segmentierten Bildregionen können die von Kalman [1960] entwickelten Kalman-Filter oder andere Verfahren eingesetzt werden, die anhand der letzten Position eine Schätzung der nächsten Position liefern. Zum Beispiel existiert ein probabilistisches Verfahren zur Verfolgung flexibler und verformbarer Objekte von Isard u. Blake [1996]. Bei den ansichtsbasierten Verfahren wird nur die bisherige Bewegung und die Annahme über die Gleichmäßigkeit der Bewegung für das Verfolgen genutzt. Modellbasierte Verfahren bringen in diesen Prozess mehr Randbedingungen und Modellwissen mit ein.

Bei Differenzbildanalysen ist oft keine Verfolgung der bewegten Region nötig, da bereits die Ergebnisse, die Differenzbilder, als Merkmale der Bewegung dienen und diese direkt zur Bewegungsklassifikation genutzt werden können.

Cédras u. Shah [1995] stellen die Bedeutung von relativen Bewegungen zwischen sich bewegenden Punkten heraus, da dies auch in der Wahrnehmung des Menschen eine zentrale Rolle spielt. Eine Konsequenz aus dieser Betrachtung ist, nicht nur einen Teil des menschlichen Körpers zu verfolgen und für die Erkennung zu nutzen, sondern Verfahren einzusetzen, die den ganzen Körper in einer Bildsequenz verfolgen. In diesen **modellbasierten Verfahren** wird Vorwissen über den menschlichen Körper genutzt, um diesen oder Teile des Körpers in einer Bildsequenz zu detektieren und zu verfolgen. Die Art der verwendeten Modelle reicht von einfachen, zweidimensionalen Strichmodellen oder Silhouetten bis zu detaillierten dreidimensionalen Modellen, die die Ausmaße des Körpers mit Zylindern oder Polygonen annähern. Der Detailreichtum des Modells geht einher mit einer

höheren Anzahl zu bestimmender Parameter, so dass mit der Wahl des Körpermodells ein Kompromiss zwischen Berechenbarkeit und Genauigkeit gefunden werden muss.

Zum Beispiel zeigen Sidenbladh u. a. [2000], wie ein gehender Mensch mit einem dreidimensionalen Körpermodell in einer Videosequenz einer einzelnen Kamera verfolgt werden kann. Ein ähnliches Verfahren von Schmidt u. a. [2006] wird für die Experimente, die in dieser Arbeit vorgestellt werden, verwendet. Mit dem Verfahren ist es möglich, den menschlichen Körper und seine Gliedmaße zu verfolgen, in den Daten können Gesten erkannt werden. Auch in der Arbeit von Lange u. a. [2003] wird ein dreidimensionales Körpermodell mit Winkelbereichen für die einzelnen Gelenke verwendet. Die Gliedmaße werden in diesem Ansatz über Stöcke repräsentiert (engl. stick-figure). In einem stochastischen Suchprozess werden aus unterschiedlichen Körperkonfigurationen produzierte Bilder mit den segmentierten Bildern aus einer Videosequenz verglichen und bewertet. Zusätzlich zu der Bewertung einer Konfiguration werden auch noch Bedeutungskarten (engl. relvance maps) eingesetzt, die die Auswirkung einzelner Gelenke auf bestimmte Bildbereiche fokussieren. Die vielversprechenden Ergebnisse wurden bisher leider nur in Tests mit synthetischen Bildern erreicht.

Ein weiterer Faktor, der die Wahl des Körpermodells bestimmt, ist das Anwendungsgebiet. Für die Erkennung von Handstellungen muss das Modell die einzelnen Fingerglieder modellieren. Um hingegen einen gehenden Menschen zu detektieren, reicht oft ein zweidimensionales Modell aus. Die optische Verfolgung einer Hand in einer monokularen Bildsequenz mit einem dreidimensionalen Handmodell wird zum Beispiel von Sudderth u. a. [2004] oder auch von Stenger u. a. [2001] vorgestellt. In dem System von Stenger sind auch Mehrkamerasysteme möglich.

Natürlich werden bei modellbasierten Verfahren ähnliche Bildmerkmale wie bei den ansichtsbasierten Verfahren ausgewertet, doch über die Randbedingungen des Modells wird der Suchraum eingeschränkt, in dem zum Beispiel die Hand eines Menschen erwartet wird. Kantenmerkmale können helfen, einen Arm oder andere Gliedmaße zu finden. Die gefundenen Bildmerkmale müssen dann den Teilen des Modells zugeordnet werden. Bei dreidimensionalen Modellen können Selbstverdeckungen, Selbstkollisionen, Einschränkungen der Gelenkwinkel und Kinematiken in den Prozess der Zustandsbestimmung mit einbezogen werden. Diese Verfahren generieren somit detaillierte Modelle der verfolgten Körper aus Monokamerabildern oder auch Tiefenkamera- und Mehrkamerabildern.

Beschäftigen sich die in diesem Abschnitt vorgestellten Methoden mit der Detektion und der Verfolgung des menschlichen Körpers oder eines Körperteils im Bild, so wenden wir uns nun der Modellierung von Bewegungen des Körpers zu.

3.2.3 Repräsentation der Bewegungen

Eine verbreitete Methode, die Körperbewegungen eines Menschen zu repräsentieren und für die weitere automatische Verarbeitung verfügbar zu machen, ist es, die Körperbewegung als Bewegung repräsentativer Punkte darzustellen. Die zeitliche und räumliche Dynamik eines Punktes lässt sich als Trajektorie darstellen, siehe hierzu auch die Definition einer Geste (siehe Abschnitt 2.3.1, Seite 23) von Pavlovic.

Die trajektorienbasierten Merkmale bieten Möglichkeiten, die über die einfache Analyse des zweidimensionalen Bilds hinausgehen. Zeitliche Aspekte von Bewegungen werden gut

repräsentiert und auch Relationen zwischen einzelnen Objekten können repräsentiert werden. Trajektorien können aus den Daten einer Regionenverfolgung berechnet werden sowie zusätzliche Informationen enthalten, wie zum Beispiel die Rotation eines Objektes, welche zum Beispiel der Objektverfolger von Gräßl u. a. [2003] ermitteln kann. Aber auch die dynamischen Parameter eines Körpermodells können mit Trajektorien abgebildet werden. Beide Verfahren werden in der Evaluation und Integration in Kapitel 5 verwendet.

Die Ausgangspunkte für die trajektorienorientierte Repräsentation von Bewegungen können grob in drei Arten unterteilt werden, die die Vielfalt und Möglichkeiten diese Daten zu verwenden vorzeichnen:

- Eine Sequenz einer zweidimensionalen Bildposition: $P_t(x_{img}, y_{img})$.
- Eine Sequenz einer dreidimensionalen Raumposition: $P_t(x, y, z)$.
- Eine Sequenz einer hochdimensionalen Körperkonfiguration: B_t.

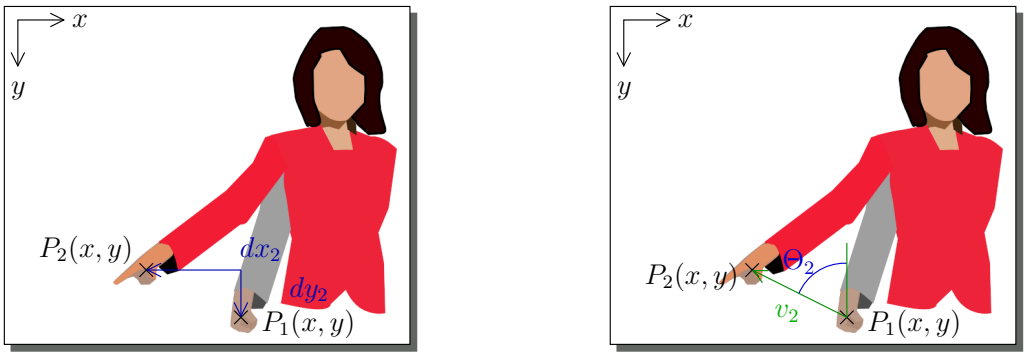
Aus diesen Daten können unterschiedliche Typen von Trajektorien $Z_{type}(t)$ für die verfolgten Punkte in Bildsequenzen berechnet werden, die die Bewegungen auf unterschiedliche Weise über die Zeit t repräsentieren. In Anlehnung an die Definition von Pavlovic u. a. [1997] (siehe Definition 2.3.1) sei hier eine Trajektorie $Z_{type}(t)$ als die Folge von Vektoren $\mathbf{s}_t$ über das Zeitintervall $t = 0$ bis $t = i$ gegeben. Wobei ein Vektor $\mathbf{s}_t$ die Stellung der Hand oder eines Objektes in einem Merkmalsraum S beschreibt. Ziel der Wahl einer günstigen Repräsentation ist es, dass ähnliche Bewegungen der Hand auf ähnlichen Trajektorien im Merkmalsraum abgebildet werden. Im Blick auf die Klassifikation von Bewegungen muss entsprechend bei der Wahl der Repräsentation beachtet werden, welche Klassen von Bewegungen erkannt werden sollen.

Beschreibungen und Diskussionen unterschiedlicher Repräsentationen geben Campbell u. a. [1996] sowie Cédras u. Shah [1995]. Nickel u. Stiefelhagen [2003a] diskutieren die Darstellungsformen von Campbell u.a. im Hinblick auf Zeigegesten. Rao u. a. [2002] konzentrieren ihre Betrachtung auf die vom Blickwinkel unabhängige Repräsentation von Aktionen in zwei Dimensionen.

In dem Artikel von Campbell u. a. [1996] werden sechs unterschiedliche Repräsentationen genannt, die auf den Raumpositionen einer Hand basieren. Diese und einige weitere Methoden der Repräsentation, die außerdem auf Bildkoordinaten, Körperkonfigurationen oder einer Objektverfolgung basieren, werden jetzt vorgestellt, sowie ihre Vor- und Nachteile diskutiert.

- **Repräsentationen basierend auf Bildkoordinaten**

  Wählt man die Sequenz von Bildkoordinaten $P_t(x_{img}, y_{img})$ als Ausgangpunkt der Bewegungsrepräsentation und -erkennung, stellt sich die Frage, wie Handbewegungen des Menschen einerseits ansichtsunabhängig und andererseits generalisierend abgebildet werden können. Die naheliegende Wahl der Bildpunkte $Z_{Pos}(t) = (x_{img}, y_{img})$ hat die Nachteile, dass die Trajektorien sowohl ansichtsabhängig, als auch ortsabhängig sind. Dass heißt, wird eine Bewegung aus unterschiedlichen Sichten beobachtetet,



(a) Repräsentation der Handbewegung als Geschwindigkeiten in den kartesichen Koordinaten des Bilds.

(b) Repräsentation der Handbewegung in der Bildebene durch die Richtung und die absolute Geschwindigkeit.

Abbildung 3.3: Repräsentationen der Handbewegung.

liegen auch unterschiedliche Trajektorien Z_{Pos} vor. Gleiches wird verursacht, wenn die Geste an einer anderen Position in der Szene ausgeführt wird. Die Repräsentation der Bewegung durch den zeitlichen Geschwindigkeitsverlauf Z_{Vel} (siehe Darstellung 3.3(a)) lässt zumindest eine ortsunabhängige Repräsentation zu:

$$Z_{Vel}(t) = \left(\frac{dx}{dt}, \frac{dy}{dt}\right). \tag{3.1}$$

Betrachtet man hingegen Zeigegesten, die in unterschiedliche Richtungen ausgeführt werden können, so fällt auf, dass diese Ausführungen auch zu unterschiedlichen Trajektorien im Merkmalsraum führen. Wie in Graphik 3.3(b) verdeutlicht wird, kann das Problem überwunden werden, indem die Bewegungsgeschwindigkeit v im Bild und die Krümmung der Bewegung $\frac{d\theta}{dt}$ verwendet werden:

$$Z_{Vel-Ang}(t) = \left(\frac{\sqrt{(dx)^2 + (dy)^2}}{dt}, \frac{d\theta}{dt}\right). \tag{3.2}$$

- **Repräsentationen basierend auf Raumkoordinaten**

  Eine Erweiterung der Ausgangsbasis auf die drei Raumdimensionen geht mit dem Vorteil einher, dass die meisten Verfolgungsverfahren, die Tiefeninformationen beinhalten, metrische Koordinaten in einem in der Kamera gelegenen Koordinatensystem nutzen. Diese Darstellung ist deutlich unabhängiger von der Ansicht. So schlägt sich zum Beispiel der Abstand der beobachteten Bewegung zur Kamera nicht in einer Skalierung der Bewegung nieder. Ansonsten lassen sich die bereits erläuterten Repräsentationen in Raumkoordinaten $Z_{3DPos}(t)$ sowie die Raumgeschwindigkeiten $Z_{3DVel}(t)$ entsprechend der zweidimensionalen Ansätze erweitern:

$$Z_{3DPos}(t) = (x, y, z) \tag{3.3}$$

$$Z_{3DVel}(t) = \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}\right) \tag{3.4}$$

In diesen Repräsentationen bleiben aber auch die meisten erwähnten Nachteile bestehen. Des Weiteren kann die Geschwindigkeit v im Raum und die Bewegungsrichtung (θ, η) beziehungsweise die Krümmung $(\frac{d\theta}{dt}, \frac{d\eta}{dt})$ der Bewegung aus den Raumpositionen P_t, berechnet werden und für die Darstellung der Handbewegungen genutzt werden:

$$Z_{3D\_Vel\_Ang}(t) = \left(v, \frac{d\theta}{dt}, \frac{d\eta}{dt}\right) \text{ mit } v = \frac{\sqrt{(dx)^2 + (dy)^2 + (dz)^2}}{dt}. \tag{3.5}$$

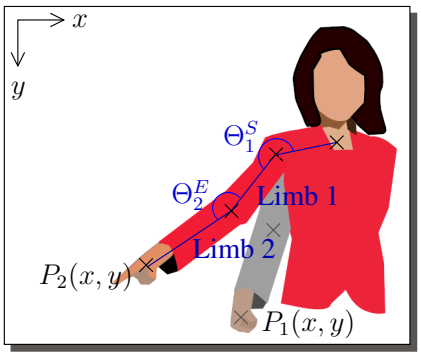
Diese Technik ist weitgehend generisch im Bezug auf den Ausführungsort und die Ausführungsrichtung der Bewegung, des Weiteren auch in weiten Teilen unabhängig von der Kameraposition und ihrer Ausrichtung. Trotzdem bleiben Mehrdeutigkeiten bestehen, wenn nur die Bewegung der Hand des Menschen betrachtet wird. Die Hinbewegung der Hand beim Zeigen kann zum Beispiel nicht von einer Rückführung der Hand unterschieden werden. Aus der Bewegung allein lassen sich diese Bewegungen nicht unterscheiden. Um dieses zu können, müssen Bewegungen im symbolischen Kontext der Objekte gesehen werden oder es muss eine Relation zum Körper des Menschen bestehen.

- **Repräsentationen basierend auf Körpermodellen**

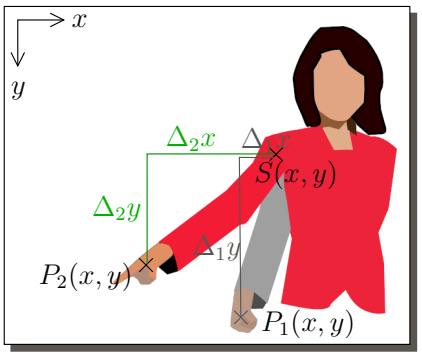
  Wird ein Verfahren zur Verfolgung des Körpers einer Person in einer Bildsequenz benutzt, welchem ein Körpermodell zugrundeliegt, existiert für jeden Zeitpunkt t die Körperkonfiguration B_t in diesem Modell. In einem solchen hochdimensionalen Merkmalsvektor lassen sich unterschiedliche Unterräume definieren, in denen die beobachteten Bewegungstypen möglichst diskriminativ abgebildet sein sollen.

  Um die Rechenkomplexität handhabbar zu halten, ist es aber auch sinnvoll, die Dimensionalität des Merkmalsraumes gering zu halten. Ein Ansatz, um Arm- und Handbewegungen zu repräsentieren, ist die Verwendung der Gelenkwinkel für die Bildung der Trajektorie. Das Vorgehen zeigt exemplarisch die Darstellung 3.4(a). In einer dreidimensionalen Körperkonfiguration sind die Winkel θ mehrdimensional, da zum Beispiel das Schultergelenk (θ_t^S) ein Kugelgelenk ist, oder für das Ellbogengelenk das Beugen und Rotieren des Unterarmes modelliert werden muss. Jedoch werden hierbei auch gleiche Bewegungen der Hand unterschieden, die von unterschiedlichen Gelenkwinkelfolgen verursacht werden.

  Ein weiterer Vorteil, der aus der Nutzung eines Körpermodells gezogen werden kann, ist, dass die Handbewegung nicht mehr losgelöst von dem Körper des Menschen betrachtet werden muss. Unter Hinzunahme dieser Informationen kann leichter unterschieden werden, ob der Mensch seine Hand zum oder vom Körper weg führt. Dieses ist, wie oben erläutert, in einer auf zwei beziehungsweise drei Dimensionen basierenden Repräsentation nicht möglich. Die Hand kann nun im Kontext zum Körper des Menschen beobachtet und ihre Bewegung in diesem räumlichen Kontext interpretiert werden.



(a) Schematische Repräsentation der Handbewegung in einem Körpermodell.



(b) Repräsentation der Handbewegung relativ zur Schulter.

Abbildung 3.4: Repräsentationen der Handbewegung.

Die Graphik 3.4(a) illustriert exemplarisch an einer vereinfachten zweidimensionalen Darstellung, wie die Handbewegung relativ zur Schulter in Gelenkwinkeln Θ_t interpretiert werden kann. Interessanter wird dieser Ansatz, wenn die Hand in räumlicher Relation zu einem Körperpunkt gesehen wird. Der Schulterpunkt (S in Darstellung 3.4(b)) bietet sich für die Interpretation von Gesten besonders an, da die Hand im entspannten Zustand sich typischerweise unterhalb dieses Punktes (P_1 in Darstellung 3.4(b)) befindet und sich mit Beginn der Geste von diesem entfernt. Zentrale Punkte, wie zum Beispiel der Kopf oder Rumpf als Referenzpunkt, haben den Nachteil, dass gleiche Gesten, zu unterschiedlichen Seiten ausgeführt, auch stark unterschiedliche Repräsentationen im Merkmalsraum haben. Am Beispiel einer mit der rechten Hand ausgeführten Zeigegeste zur rechten beziehungsweise linken Seite lässt sich erkennen, dass der Verlauf des Abstands zum Referenzpunkt S für die rechte Seite nur zunimmt, hingegen für die linke Seite erst abnimmt und dann zunimmt.

Die Verwendung eines Körpermodells ermöglicht eine Repräsentation, die unabhängig von der Translation, also der Ausführungsposition im Raum ist. Das lässt sich zum Beispiel mit den Veränderungen des Abstands zur Schulter über die Zeit erreichen:

$$Z_{RelPos}(t) = \left(\frac{dx_\Delta}{dt}, \frac{dy_\Delta}{dt}, \frac{dz_\Delta}{dt} \right) \tag{3.6}$$

$$\text{mit } n_\Delta = S_t(n) - P_t(n) \text{ für } n \in \{x, y, z\}\,. \tag{3.7}$$

Soll hingegen eine Repräsentation auch bezüglich der Ausführungsrichtung möglichst invariant sein, bieten sich zylindrische beziehungsweise sphärische Koordinatensysteme mit ihrem Ursprung in dem Referenzpunkt an. In einem zylindrischen Koordinatensystem lässt sich die Bewegung der Hand, wie folgt, repräsentieren:

$$Z_{DeltaCyl}(t) = \left(\frac{dr}{dt}, \frac{d\Theta}{dt}, \frac{dz}{dt} \right). \tag{3.8}$$

Wobei r der Abstand der Hand zum Referenzpunkt R in der $x-y$ Ebene, also der Tiefe und Breite ist, der Winkel Θ den Winkel der Hand zum Referenzpunkt in dieser Ebene beschreibt und z die Höhenkomponente ist. In einem sphärischen Koordinatensystem wird der Punkt der Hand hingegen durch die Richtungswinkel Θ_1 und Θ_2 in der $x-y$ respektive $x-z$ Ebene und dem Raumabstand r beschrieben. Auch für diese Koordinatensysteme sind wiederum Variationen möglich. Denn es kann nicht nur die zeitliche Ableitung einer Komponente (z.B. $\frac{\delta r}{\delta t}$) als Merkmal gewählt werden, sondern auch die Komponente selbst (r) oder ihr Absolutwert ($|r|$).

Die einzelnen Repräsentationen von Bewegungen haben neben den bereits genannten prinzipiellen noch weitere Vor- und Nachteile: Wählt man höhere Momente, wie die Geschwindigkeit oder Beschleunigung, kann man eine von Translation und Rotation unabhängige Darstellung der Bewegung erreichen. Das Verfahren ist also generischer und nicht nur für einen festen und kalibrierten Aufbau verwendbar. Doch das mehr oder weniger starke Rauschen, das im Signal enthalten ist, macht sich bei diesen höheren Momenten stärker bemerkbar. Der Abstand zwischen Signal und Rauschen wird dementsprechend geringer und die Erkennung schwieriger.

Dagegen sind Bewegungen im Bild, dargestellt als x und y Position, sehr speziell und eine andere Ausführungsposition im Bild, oder eine leicht unterschiedliche Richtung der Bewegung, können nicht mehr mit einem Modell repräsentiert werden. Die somit vielen, nur leicht differierenden Bewegungsmodelle, erschweren die Erkennung einer Bewegung.

3.2.4 Methoden zur Klassifikation von Bewegungen

Der nächste Schritt befasst sich mit der Erkennung typischer Bewegungen aus einer Trajektorie, die als Sequenz von Merkmalen aufgefasst werden kann. Die Repräsentation sollte den Anforderungen der Musterklassifikation genügen, also diskriminative Klassen im Merkmalsraum für verschiedene Bewegungen aufweisen. Für jede dieser Klassen wird zumindest ein Beispiel oder ein Trainingsdatensatz benötigt. Für das Bilden eines Modells oder auch Templates aus mehreren Beispielen kommen übliche Verfahren wie Mittelung, K-means, Hidden Markov Modelle (HMM) oder Neuronale Netze zum Einsatz. Alternativ kann direkt ein Template als Modell verwendet werden. Allgemein wird für die Erkennung der Abstand des unbekannten Merkmalsvektors $\mathbf{s}$ zu den bekannten Klassen $\mathbf{K}$ ermittelt. Falls der Merkmalsvektor im dem Gebiet einer Klasse k_i liegt, wird diese Klasse als erkannt ausgegeben. Pavlovic u. a. [1997] nennen die üblichen Musterklassifikationsverfahren HMM, Dynamic Time Warping (DTW), Neuronale Netze (NN) oder Particle Filtering (PF), die sich als Methoden zur Bewegungsklassifikation anbieten. Auf diese gängigen Klassifikationsverfahren möchte ich nun kurz eingehen und ihre Charakteristika sowie Vor- und Nachteile ansprechen. Weiterführende Beschreibungen der einzelnen Verfahren finden sich in den angegebenen Werken.

- **Hidden Markov Modelle**

  Die aus dem Bereich der kontinuierlichen Spracherkennung bekannte Technik der *Hidden Markov Modelle* (HMM) wird von Starner u. Pentland [1995] erfolgreich zur

Erkennung der amerikanischen Zeichensprache (*American Sign Language*) verwendet und in den Bereich der Gesten- und Aktionserkennung eingeführt. Zur leichteren Segmentierung der Hände benutzen Starner und Pentland einen strukturierten Hintergrund und farbige Handschuhe und können damit eine hohe Erkennungsrate in Echtzeit erreichen. In einem ähnlichen Aufbau setzen Campbell u. a. [1996] HMM zur Klassifikation von T'ai Chi Bewegungen unter Verwendung eines Stereokamera-Systems ein. Mit Stereodaten arbeiten auch Nickel u. Stiefelhagen [2003a], um Zeigegesten mit HMM zu erkennen. Des Weiteren wird die Technik der HMM von McKenna u. Morrison [2004] in der trajektorienbasierten Erkennung von Gesten eingesetzt. Einen anderen Ansatz verfolgen Dreuw u. a. [2005], um die Hand- und Fingerstellung zu erkennen: Nicht eine Trajektorie, sondern die Bilder werden benutzt, um in einem Zustand der HMM eine Distanz zu vorher trainierten Handstellungen zu berechnen. Dies ist die Emission des Zustands.

Die Theorie der HMM wird unter anderem von Niemann [1983] im Kapitel 4.2 vorgestellt. Bobick [1997] beschreibt die Anwendung der HMM für die kontinuierliche Gestenerkennung. An dieser Stelle sei nur das grundlegende Prinzip der HMM für die Gestenerkennung skizziert:

Das Prinzip der HMM ist das eines endlichen stochastischen Automaten. Die Handposition oder eine ähnliche Repräsentation als Signal, lässt sich als Zustandsfolge einer endlichen Menge von Zustandssymbolen annehmen. Ist der aktuelle Zustand nur von dem vorhergehenden Zustand abhängig, spricht man von einer Markov-Kette erster Ordnung. Die Übergänge zwischen den Zuständen sind über die Matrix der Zustandsübergangswahrscheinlichkeiten definiert. Die Wahrscheinlichkeiten müssen mit früheren Beobachtungsfolgen trainiert werden. Die Zustände eines HMM sind nicht zu beobachten, doch generiert das Modell eine Beobachtung, wenn es sich in einem Zustand befindet. Die Ausgabe dieser Beobachtungen wird über die Ausgabewahrscheinlichkeiten definiert.

Die HMM erlauben, unterschiedliche Bewegungssequenzen zu modellieren. Wird nun eine Sequenz in ein vorher trainiertes HMM gegeben, so lassen sich die Beobachtungen als Wahrscheinlichkeiten für die eingegebene Sequenz interpretieren. Für das Erkennen von Gesten werden meistens *links-rechts* HMM verwendet, bei denen zeitlich rückwärts gerichtete Zustandsübergänge nicht möglich sind.

Die HMM ermöglichen zeitlich sequenzielle Daten sowohl zu segmentieren als auch Modelle zu erkennen, jedoch wird eine große Trainingsmenge zum Ermitteln der internen Wahrscheinlichkeiten der HMM benötigt. Ebenso ist es kompliziert, die zeitliche Dauer von Gesten und insbesondere Bewegungspausen zu modellieren.

- **Dynamic Time Warping**

  Ein weiteres Verfahren, das robust bezüglich der zeitlichen Skalierungen des Signals beziehungsweise der Trajektorie ist, ist das *Dynamic Time Warping* (DTW). Das DTW nutzt die Methode der Dynamischen Programmierung, um das Signal optimal an ein Modell zu alignieren. Hierfür wird das Signal in kleinen Segmenten zeitlich gestaucht oder gestreckt, so dass es über alle Dimensionen und über seinen gesamten zeitlichen Verlauf möglichst dem Modell entspricht. Existieren mehrere Modelle,

kann ermittelt werden, wie groß der Aufwand für die Verzerrung und wie gut die Übereinstimmung zwischen einem Modell und der Trajektorie ist. Die Dynamische Programmierung wird unter anderem im Kapitel 1.6.8 von Niemann [1983] beschrieben.

Li u. Greenspan [2005b] benutzen das DTW auf Sequenzen von Körperkonturen, um Armbewegungen, unter anderem Zeigegesten oder Winkbewegungen, zu erkennen. Da die beobachtete Person vor einem strukturierten Hintergrund steht, sind die Bewegungen leicht zu detektieren und gut zu erkennen. Mit dem DTW ist es möglich, unterschiedlich schnelle beziehungsweise weite Ausführungen einer Geste mit einem Modell zu klassifizieren. Doch werden auch hier für die Bildung der Modelle viele Trainingssequenzen benötigt und es wird erlaubt, dass innerhalb der Ausführung einer Geste die Zeitskalierung differiert. Erkenntnisse der Bewegungsanalyse von Körding u. Wolpert [2004] legen aber nahe, dass die Skalierung während der Ausführung nahezu konstant bleibt. Des Weiteren ist das Problem der Segmentierung nicht gelöst.

- **Neuronale Netze**

  Mit den *Künstlichen Neuronalen Netzen* (engl. Artifical Neural Network) existiert eine weitere Gruppe von Verfahren, die zur Klassifikation von Körperstellungen oder Bewegungen verwendet werden können. Über die Anwendung Neuronaler Netze geben zum Beispiel Bishop [1995] oder Niemann [1983][1] eine Übersicht. Neuronale Netze zeichnen sich für die Klassifikation dadurch aus, dass mit ihnen fast beliebige Funktionen approximiert werden können. Inspiriert durch die biologischen Neuronen werden Netze künstlicher Neuronen gebildet, die gewichtete Eingaben von anderen Neuronen aufsummieren und über nichtlinearen Funktionen ihre Ausgabe berechnen.

  Neuronale Netze sind lernfähig, adaptiv und ermöglichen die parallele Verarbeitung von Informationen. Eine typische Netztopografie ist die schichtweise Anordnung der künstlichen Neuronen. Bei diesem *Mehrschicht-Perzeptron* (MLP) (engl. multilayer perceptron) sind die Neuronen einer Schicht nur mit der nächsten Schicht verbunden. Das Training eines Netzes besteht darin, über viele Beispieldaten die internen Gewichte des Netzes zu belegen.

  Zur Erkennung von Bewegungen der Amerikanischen Zeichensprache verwenden Yang u. Ahuja [2000] ein *Time Delay Neural Network* (TDNN). Als Merkmale werden die zweidimensionalen Bewegungstrajektorien der Hände verwendet. Richarz u. a. [2006] benutzen mehrere MLP für die Erkennung der Zeigerichtung aus der Stellung des menschlichen Armes. Auch Heidemann u. a. [2004] verwenden in ihrem System zur Erkennung von Zeigegesten einen Neuronalen Klassifikator.

- **Particle Filter oder Condensation**

  Eine weitere Gruppe von Klassifikationsmethoden bilden die Sequenziellen Monte Carlo Methoden (siehe auch Doucet u. a. [2001]), die ein probabilistisches Verfolgen und Erkennen ermöglichen. Stellvertretend sei hier der *Condensation*-Algorithmus

[1] siehe Niemann [1983], Kapitel 4.5

von Isard u. Blake [1998b] genannt. Das Verfahren kann als Generalisierung der HMM gesehen werden, da eine diskrete Zustandsmenge mit probabilistischen Übergängen zwischen den Zuständen erlaubt ist.

Die Variante von Black u. Jepson [1998] — das *Condensation Trajectory Recognition* (CTR) — ermöglicht den probabilistischen Abgleich von ganzen Trajektorien mit den die Gesten repräsentierenden Modellen für einzelne Zustände. In diesem Aspekt ähnelt das CTR dem DTW, aber die parametrisierte Deformationen der Modelle ist in einen wahrscheinlichkeitstheoretischen Rahmen eingebettet. Ein weiterer Vorteil des *Particle Filter* (PF) liegt in der automatischen Konzentration der Rechenressourcen auf relevante Teile und der automatischen Segmentierung. Das Verfahren ist unabhängig von der Trainingsmethode, die zur Bildung der Modelle verwendet wird. Es können auch physikalische Gesetzmäßigkeiten als Modelle verwendet werden.

Doucet u. a. [2001] geben einen detaillierten Einblick in viele unterschiedliche *Particle Filter* Verfahren, deren mathematische Hintergründe und Anwendungen. Psarrou u. a. [2002] verwenden den CTR von Black und Jepson in einem System, in dem HMM und Erwartungsmaximierung zum Training eingesetzt werden. Die trainierten Modelle nutzend werden Bewegungsabläufen von Personen in einem Büro mit dem CTR erkannt. Auf den CTR und die konkrete Implementierung für die vorliegende Arbeit wird später in Kapitel 4 ausführlich eingegangen. Es werden sowohl die technischen Aspekte, als auch ihre Umsetzung und Auswirkungen auf die Erkennung von Handgesten diskutiert.

Nach dieser allgemeinen Sicht auf den Prozess der Bewegungserkennung und verschiedene eingesetzte Methoden, beschäftigt sich das nächste Unterkapitel mit speziellen Systemen zur Erkennung von Bewegungen der Hand. Der Fokus wird hier auf Gesten und Objektmanipulationen gelegt.

3.3 Bewegungserkennung in der Mensch-Maschine-Interaktion

Nachdem in dem letzten Unterkapitel der Prozess der Gesten und Handlungserkennung erläutert und typische Methoden, die in diesem angewandt werden, beschrieben wurden, wird in diesem Unterkapitel die Konzentration auf Systeme zur Gestenerkennung gelegt. Auch hier spannt sich ein weites Forschungsfeld auf, das sowohl der Vielfalt an Gesten und ihrer Interpretationsmöglichkeiten Rechnung trägt, als auch in den diversen technischen und algorithmischen Möglichkeiten begründet liegt.

Einen guten Überblick über unterschiedliche Verfahren zur Gestenerkennung und der Extraktion zweidimensionaler Bewegungstrajektorien geben Yang u. a. [2002] in ihrem Artikel. Bei vielen der aufgeführten Arbeiten werden für die Extraktion der Merkmale jedoch Datenhandschuhe oder optische Markierungen verwendet. Diese Verfahren leisten einen wertvollen Beitrag zum tieferen Verständnis menschlicher Gesten und ihrer automatischen Erkennung. In erster Linie ermöglicht dieses Vorgehen eine deutlich einfachere und

sichere Akquise der Trajektorien sowie weniger stark verrauschte Daten. Es wird somit auch die zeitliche Entwicklung der Gestenerkennung aufgezeigt. Denn mit dem fortlaufenden Leistungszuwachs der Computertechnik werden komplexere Systeme möglich, die auch aufwendige Detektions- und Verfolgungsmethoden einschließen. Diese Entwicklung spiegelt sich auch in den Versuchsaufbauten der videobasierten Systeme wieder. Können Farbmarkierungen in einem strukturierten Aufbau noch über einfache Farbfilter verfolgt werden, erfordert es rechenintensive Algorithmen, eine menschliche Hand in einer komplexen Szene zu detektieren und zu verfolgen.

Der nun folgende Abschnitt wird durch die Gestenart und Anwendung der Gestenerkennung strukturiert und es werden für die jeweiligen Gebiete einige Beispiele genannt. Es ergeben sich drei Teilgebiete der Gestenerkennung, wovon die letzten beiden im Rahmen dieser Arbeit von Bedeutung sind:

- Erkennen der Handstellung und Bewegung der Finger.
- Erkennen der Hand- und Armbewegung im Raum.
- Erkennen der Hand- und Armbewegung im Raum in Relation zu den manipulierten Objekten.

In den vorgestellten Systemen werden auch wiederum Verfahren zur Detektion und Verfolgung von Händen und Armen des Menschen verwendet, die zum Teil bereits angesprochen wurden. Doch soll in diesem Abschnitt der Fokus auf den Systemgedanken und das Erkennen von Gesten gelegt werden.

3.3.1 Handstellung

Bei zahlreichen Gesten wird keine Arm- und Handbewegung ausgeführt, sondern die bedeutungstragenden Bewegungen werden mit den Fingern ausgeführt oder die Finger nehmen eine spezielle Stellung ein. Beispiele sind die Embleme des nach oben ausgestreckten Daumens oder die zum Schwur erhobene Hand. In Gehörlosensprachen kommt der Stellung der Finger eine ganz besondere Bedeutung zu. Solche Embleme, aber auch die Zeichen der Gehörlosensprachen, haben eine direkte lexikalische Bedeutung, die oft kulturell bestimmt ist.

Folglich beschäftigt sich ein Bereich der Gestenerkennung mit der Rekonstruktion der Finger- und Handstellung aus einer Bildsequenz oder aus Einzelbildern. Oft werden aber nur einfache Handmodelle verwendet, die nur grobe Hand- und Fingerstellungen rekonstruieren. Ausgereifte und komplexe modellbasierte Verfahren zum Verfolgen und Erkennen der Handstellung haben dagegen ein größeres Potential, siehe zum Beispiel die in 3.2.2 vorgestellten Verfahren von Sudderth u. a. [2004] oder Stenger u. a. [2001]. Die mit diesen Verfahren ermittelten Hand- und Fingerstellungen bieten detaillierte Daten für das Erkennen von Gesten, deren Bedeutung in der Handstellung liegt. An dieser Stelle liegt noch ein großes Entwicklungspotential, denn eine sichere und robuste Detektion und Verfolgung der einzelnen Finger einer Hand ermöglicht erst eine Klassifikation der Finger- und Handbewegungen in bedeutungstragende Gesten.

Ein aufwendiges Mehrkamerasystem zur hautfarbenbasierten Verfolgung von Gesichtern und Händen stellen Hongo u. a. [2000] vor. Die Distanz der gefundenen Regionen zum Kamerasystem kann über eine Stereokamera ermittelt werden. Für diese hautfarbenen Regionen werden außerdem vier gerichtete Merkmale berechnet; die größennormierten und geglätteten Merkmale zeigen Kanten in horizontaler, vertikaler beziehungsweise diagonaler Richtung in der Region auf. In einer hierarchischen linearen Diskriminanz-Analyse werden aus diesen Merkmalen zuerst Gesichter und Hände unterschieden, anschließend drei Klassen von Handsilhouetten. In diesem ansichtsabhängigen Ansatz werden folglich nur statische Handstellungen erkannt, die Bewegung wird hier nicht als Merkmal für Gesten verwendet. Aus den Versuchsbildern geht auch hervor, dass eine einfach strukturierte Szene vorliegt, die das Segmentieren von hautfarbenen Regionen begünstigt.

In der Arbeit von Bretzner u. a. [2002] wird gezeigt, wie ein Fernsehgerät über Handstellungen gesteuert werden kann. Der Testaufbau ermöglicht eine Steuerung in Echtzeit, indem die zur Kamera gerichtete Hand über Farbmerkmale gefunden und als hierarchische Struktur von Kreisen und Ellipsen repräsentiert wird. Zur Verfolgung und zum Erkennen der Stellung wird ein Particle Filtering Verfahren angewandt.

New u. a. [2003] stellen ein System vor, das eine auf einem Tisch bewegte Hand von oben aufnimmt, segmentiert und die Stellung sowie Position als Steuerbefehl für ein Computersystem interpretiert. Die Segmentierung nutzt die Merkmale Farbe, Helligkeit und Sättigung. Daraufhin wird die größte geschlossene Kontur ermittelt und als Hand des Benutzers angenommen. Über feste Schwellwerte für die Fingerbreite und den Abstand der Finger untereinander können die Finger separiert und gezählt werden. Neben diesen einfachen Gesten wird die Position der Hand in festgelegten Regionen des Bilds als Befehl gewertet. Das vorgestellte System arbeitet mit 15Hz in einem strukturierten Szenario und auch die erkannten Gesten haben eine geringe Komplexität, so ist beispielsweise die Bewegung nicht relevant.

Ein System, das eine deutlich detaillierte Erkennung und Rekonstruktion der menschlichen Hand ermöglicht, wird von Nölker u. Ritter [2002] vorgestellt. In einem zweistufigen Prozess, der künstliche Neuronale Netze verwendet, werden zuerst global die Finger und anschließend genauer die Positionen der Fingerspitzen einer Hand detektiert. Aus diesen Positionsinformationen wird, wiederum mit einem Neuronalen Netz, eine Konfiguration eines Handmodells rekonstruiert. Das Handmodell hat 20 Freiheitsgrade und Randbedingungen für die einzelnen Gelenkwinkel. Das System setzt voraus, dass der Benutzer seine Hand durch eine kleine Öffnung in einen Kasten mit gleichmäßiger Beleuchtung steckt. Unter diesen Bedingungen ist es möglich, die Stellung der Hand kontinuierlich zu verfolgen. Weitere Details und Anwendungsbeispiele dieser Mensch-Maschine-Schnittstelle, zum Beispiel zur Steuerung von Parametern der Sonifikation, werden von Nölker [2000] gegeben. Embleme und ähnliche Gesten zu erkennen, wurde hingegen nicht versucht, auch wenn das Handmodell dieses sicherlich ermöglicht.

Das Ziel der Arbeit von Bax u. a. [2003] ist, Zeigegesten und deren Richtung aus der Kontur der Hand zu erkennen und so Objektreferenzen aufzulösen. Die Hand sowie die Objekte, die auf einem Tisch liegen, werden mit einer Kamera von oben aufgenommen. Die Schätzung der Richtung einer Zeigegeste wird in diesem ansichtsbasierten Verfahren mit einem neuronalen Klassifikator bestimmt. Dazu werden hautfarbene Regionen des Bilds

in einem dreistufigen Prozess verarbeitet. Zuerst erfolgt eine Vektorquantisierung auf das Ausgangsbild, die Dimension der optischen Merkmale wird dann über eine lokale Hauptkomponentenanalyse verringert und mit einem Neuronalen Netz klassifiziert. In diesem Prozess wird klassifiziert, ob die Region eine Hand ist und wenn dies zutrifft, in welche Richtung die Hand zeigt. Das Verfahren hat den Nachteil, dass sehr viele Trainingsbilder für alle möglichen Zeigerichtungen vorliegen müssen. Das Problem gehen Heidemann u. a. [2004] an, indem sie eine graphische Schnittstelle zum einfachen Kennzeichnen der Daten entwickeln.

3.3.2 Armbewegung

Betrachten die bisher vorgestellten Verfahren nur die Hand eines Menschen und ihre Bewegung, wird nun die Betrachtung auf den ganzen Menschen und seine Arm- und Handbewegung ausgedehnt. Das hat einerseits zur Folge, dass größere und andere Bewegungen betrachtet werden, andererseits aber auch, dass Bewegungen einzelner Fingerglieder meist nicht mehr erfasst werden können.

Konturbasiertes Erkennen von Zeigegesten

In den Arbeiten von Kehl u. Gool [2004] sowie Li u. Greenspan [2005b] wird die Kontur eines Menschen aus einer Bildsequenz ermittelt und für die Detektion von Zeigegesten verwendet. Kehl u. Gool [2004] verwenden in ihrem Aufbau zur Erkennung von Zeigegesten in der virtuellen, dreidimensionalen Umgebung einer CAVE mehrere Kameras und ermitteln in den jeweiligen Bildern die Extrempunkte der Silhouette des Menschen, also die Punkte mit dem größten Abstand zum Massezentrum der Silhouette. Unter den Bedingungen, dass der Mensch aufrecht steht und von sich weg zeigt, können diese Punkte als Hände, Füße beziehungsweise als Kopf interpretiert und für die Detektion von Zeigegesten verwendet werden. Als Zeigerichtung wird die Blickrichtung verwendet, die aus der Verbindungslinie vom Kopf zur zeigenden Hand ermittelt wird. Die Richtung des Unterarmes als Zeigerichtung zu interpretieren, wie sie auch Nickel u. Stiefelhagen [2003b] benutzt haben, wurde als ein weniger natürliches Zeigen empfunden. In der strukturierten Umgebung einer CAVE erreichen sie eine hohe Präzision der Detektion von Deiktika. Die Gestenerkennung basiert hierfür auf einem kalibrierten Aufbau mit bis zu neun Kameras.

Ähnlich zu dieser Arbeit basiert auch die Erkennung von Gesten von Li u. Greenspan [2005a,b] auf der Kontur eines Menschen vor einem strukturierten, einfarbigen Hintergrund. Aus der zeitlichen Veränderung der Kontur werden Bewegungssignaturen erstellt. In einem zweiphasigen Prozess aus DTW und korrelativer Information werden vorher gelernte Signaturen mit dem Eingangssignal verglichen. In einer komplexeren Umgebung oder einem nicht kalibrierten Aufbau ist eine Detektion der Kontur des Menschen nur schwer möglich, auch verhindert der Einsatz auf einem mobilen Roboter die Nutzung von Multi-Kamerasystemen.

Zeigegesten im Raum

Neben den eben vorgestellten Verfahren, eine Zeigegeste im Raum zu erkennen, besteht auch die Möglichkeit, dies auf Grund von Tiefenbildern zu erreichen, wie Nickel u. Stiefelhagen [2004a, 2003a] in ihren Arbeiten zeigen. Aus dem Tiefenbild einer Stereokamera

und der Segmentierung von hautfarbenen Regionen werden Häufungen von hautfarbenen Pixeln im Raum detektiert und verfolgt. Des Weiteren wird ein Neuronales Netz verwendet, um die Orientierung des Kopfes zu ermitteln. Die Bewegung einer Hand wird in einem zylindrischen Koordinatensystem, dessen Basis in der Kopfposition des Benutzers liegt, dargestellt. Zeigegesten werden mit einem HMM erkannt, als Eingangssignal liegt der Abstand r der Hand zum Körper, die absolute Richtungsgeschwindigkeit $\Delta\Theta$ sowie die Veränderung der Höhe Δz vor.

Das Ziel, Zeigegesten aus den Bildern einer monokularen Kamera zu detektieren und diese in Anweisungen für den mobilen Roboter *HOROS* umzusetzen, wird in dem Ansatz von Richarz u. a. [2006] verfolgt. Eine Person, die vor dem Roboter steht, wird durch eine Detektion des Kopfes und der Schultern der Person erkannt, hierfür wird die *Boosting*-Methode von Viola u. Jones [2001] verwendet. Die Bildregion mit der Person wird skaliert und helligkeitsnormiert. Aus dem Resultat wird ein Merkmalsvektor mit Gaborfiltern berechnet. Die Zeigegesten, die immer in Richtung des Bodens ausgeführt werden, werden anschließend mit einer Kaskade mehrerer Mehrschicht-Perzeptren klassifiziert. Diese liefern als Ergebnisse die Richtung und Distanz relativ zum Roboter. Der Artikel von Richarz und seinen Kollegen stellt somit ein Evaluationssystem vor, das zeigt, dass es möglich ist, die Richtung einer Zeigegeste aus einfachen monokularen Kameras zu detektieren, doch wird die Bewegung der Zeigegesten nicht in die Betrachtung mit einbezogen. Somit ist es schwierig auszumachen, wann die Geste zu Ende ist und die Richtung an die Robotersteuerung übermittelt werden soll.

3.3.3 Handlungen im Kontext

Obwohl die Gestenerkennung in vielen Forschungsvorhaben untersucht wird, wird selten versucht, den symbolischen Kontext, in dem eine Geste steht, in den Erkennungsprozess einzubeziehen. Einer der ersten Ansätze, die Bewegung der Hände sowie die Objekte in der Szene zu berücksichtigen, geht auf Kuniyoshi u. a. [1994] zurück. In dem Ansatz wird die Bewegung der Hand verfolgt, ebenso existiert ein Modell der Umgebung. Mit einem hierarchischen Automaten, der die Aktionserkennung verwirklicht, werden Handbewegung und Objektrepräsentation verbunden.

Ayers u. Shah [1998] stellen ihren Ansatz zur Aktionserkennung am Beispiel einer Büroumgebung dar. Das Gesicht und — oder der Nacken einer Person wird mit einem einfachen Hautfarbenmodell im Kamerabild einer statischen Kamera verfolgt. Um die Objekte im Raum werden Bildregionen definiert. Intensitätsänderung in der Region eines Objektes deuten eine Interaktion mit dem Objekt an. Über einen endlichen Zustandsautomaten werden Helligkeitsänderungen in der Nähe eines Objektes und die gefundene Person zur Aktionserkennung verbunden. Ähnlich zu Kuniyoshis Vorgehen wird die Bewegung nicht explizit modelliert.

Ein System, in dem sensorische Trajektoriendaten und symbolische Objektdaten verbunden werden, stellen Moore u. Essa [2002] vor. Es wird ein Tisch von oben beobachtet, auf dem Karten liegen, die von Händen gegriffen und bewegt werden. Jedes Bildmerkmal und jede Bewegung wird über ein Symbol referenziert. Um bedeutungstragende Sequenzen dieser Ereignisse zu erkennen, wird eine kontextfreie Grammatik verwendet.

Nur das Verfahren von Moore u. Essa [2002] bezieht die Bewegung in die Erkennung ein, während in den Arbeiten von Kuniyoshi u. a. [1994] sowie Ayers u. Shah [1998] nur die Position der Hand oder des Körpers Einfluss auf die Aktionserkennung hat. Moore und seine Kollegen jedoch verwenden die Trajektorie als zusätzlichen Hinweis für die Objekterkennung. In dem in dieser Arbeit vorgestellten Verfahren werden umgekehrt die symbolischen Objektinformationen in eine trajektorienbasierte Gestenerkennung einbezogen.

Schlussfolgerungen

Für das automatische Erkennen von menschlichen Bewegungen in Kamerabildern sind Abfolgen unterschiedlicher Methoden der Bildverarbeitung und der Musterklassifikation nötig. Zur Einordnung der in dieser Dissertation präsentierten Gestenerkennung wurde in diesem Kapitel ein Überblick über dieses Themengebiet gegeben. Der Überblick wurde mit Verfahren zur Detektion und Verfolgung von Personen oder Körperteilen in Videosequenzen begonnen. Im Anschluss wurden unterschiedliche Ansätze vorgestellt, die die trajektorienbasierte Repräsentation menschlicher Bewegungen ermöglichen. Die Stärken und Schwächen der verschiedenen Repräsentation wurde herausgearbeitet. Des Weiteren wurden typische Musterklassifikationsmethoden kurz vorgestellt, die in der Bewegungserkennung Anwendung finden. Abgeschlossen wurde das Kapitel mit einem Blick auf bestehende Systeme zur Handlungserkennung.

Viele aktuelle Bestrebungen in dem Forschungsgebiet der Bewegungserkennung zielen auf automatisierte Überwachungs- und Sicherheitstechnik. Im Gegensatz dazu wurden in diesem Kapitel Repräsentationen von Bewegungen erarbeitet, die auf eine Verwendung in der MMI und MRK ausgerichtet sind. Obwohl die Verfahren der Musterklassifikation und -analyse in beiden Anwendungsgebieten eingesetzt werden können, ist die Intention der Entwicklung eine andere. Der Fokus dieser Dissertation liegt auf der Unterstützung des Menschen und der Interaktion mit sozial agierenden Robotern. Das Entwickeln und Anwenden von Überwachungsverfahren mag das Sicherheitsgefühl in der Bevölkerung stärken, kann aber auch die Sensibilität für die tieferliegenden sozialen Herausforderungen behindern.

Während in diesem Kapitel die einzelnen Schritte der Gesten- und Aktionserkennung vorgestellt und kurz erläutert wurden, wird in den folgenden zwei Kapiteln zuerst der für diese Arbeit verwendete und weiterentwickelte Algorithmus zur Handlungserkennung diskutiert. Anschließend wird der Einsatz desselben in unterschiedlichen Systemen gezeigt und evaluiert.

4. Probabilistische Gestenerkennung

In diesem Kapitel werden die Grundlagen (4.1) und Erläuterungen (4.2) des für diese Arbeit entwickelten Verfahrens zur Handlungserkennung behandelt. Des Weiteren wird eine Methode zum Training der Gestenmodelle (4.3) vorgestellt. Der darauf folgende Abschnitt (4.4) beschäftigt sich mit der Erweiterung der Handlungserkennung um eine Integration symbolischer Objektinformationen. Anwendungen des Verfahrens in komplexen MMI Systemen werden im folgenden Kapitel 5 behandelt.

Motivation für probabilistische Verfahren

Für viele Probleme des Alltags ist präzises Wissen über einen komplexen Prozess notwendig, doch können oft nur einige Prozesswerte gemessen werden und die zugrundeliegenden Werte müssen geschätzt werden. Das ist zum Beispiel bei der Wetterbeobachtung der Fall, aber auch bei der Beobachtung und Interpretation menschlicher Bewegungen. Das Problem der Wettervorhersage liegt in den relativ wenigen und ungenauen Messwerten, trotzdem wird recht erfolgreich die zugrundeliegende Dynamik erkannt. Um Daten realistischer Probleme zu analysieren, werden oft probabilistische Verfahren verwendet, dass heißt es werden anhand von statistischen oder physikalischen Modellen Schätzungen aufgestellt, die mit den Beobachtungen verglichen werden. Unter Berücksichtigung der Beobachtungen kann die Schätzung verbessert werden. Das Wissen über das beobachtete System zu einem bestimmten Zeitpunkt ist nicht mehr ein exakter Wert, sondern ein Bereich, der der Schätzung entspricht. Mit der Funktion der Wahrscheinlichkeitsdichte, der *probability density function* (pdf), kann das Wissen abgebildet werden.

Möchte man die Bewegungen und Handlungen eines Menschen in einer Bildsequenz erkennen und interpretieren, liegen auch mehr oder weniger verrauschte Daten vor. Aus diesen soll aber eine belastbare Vermutung über die ausgeführte Bewegung erstellt werden. Für die Erkennung von Handlungen, das umfasst sowohl Zeigegesten als auch Aktionen der Hand und von einer Hand bewegte Objekte, wird in dieser Arbeit deswegen ein probabilistisches Verfahren verwendet: das robuste *Condensation Trajectory Recognition* (CTR) von Black u. Jepson [1998]. Diese Methode ist eine Erweiterung des *Conditional Density*

Propagation (***Condensation***) von Isard u. Blake [1998a] und gehört somit in die Klasse der sequenziellen Monte Carlo Methoden. Der klassische Condensation-Algorithmus von Isard u. Blake [1996] wurde zum Verfolgen von Konturen verwendet. Später erweiterten Black u. Jepson dieses Verfahren um eine automatische Modellumschaltung und ermöglichten so eine Erkennung unterschiedlicher Bewegungsmodelle. Anzumerken ist, dass die Aufgaben des Verfolgens und des Erkennens von Bewegungen sich nicht stark von einander unterscheiden. Denn beim Verfolgen wird unter Verwendung eines Bewegungsmodells versucht, das bewegte Objekt zu verfolgen. Misslingt dies, ist das Modell für die beobachtete Bewegung unangemessen. Verwendet man aber mehrere unterschiedliche Modelle, lassen sich mehrere Bewegungsarten verfolgen. Gleichzeitig lässt sich sagen, dass das Modell mit der aktuell höchsten Wahrscheinlichkeit die Bewegung am besten beschreibt, die Bewegung kann folglich klassifiziert werden.

Die auf dem Condensation-Algorithmus basierende Handlungserkennung arbeitet auf mehrdimensionalen, zeitlichen Merkmalsvektoren, den Trajektorien. Eine Trajektorie mit I Dimensionen wird wie in Gleichung 4.1 dargestellt. Für einen Abschnitt vom Zeitschritt 1 bis t wird die Notation in Gleichung 4.2 verwendet. Diese Trajektorien repräsentieren Bewegungen in Merkmalsräumen, wie sie im Abschnitt 3.2.3 vorgestellt wurden.

$$Z(t) = \left\{z^1(t), \ldots, z^I(t)\right\} \text{ der Dimensionalität } I \tag{4.1}$$

$$\begin{aligned} Z(1{:}t) &= \{Z(1), Z(2), \ldots, Z(t)\} \\ z^I(1{:}t) &= \left\{z^I(1), z^I(2), \ldots, z^I(t)\right\} \end{aligned} \tag{4.2}$$

Zum Detektieren und Verfolgen von Händen, beziehungsweise von Objekten, werden in dieser Arbeit unterschiedliche Verfahren eingesetzt, die bei den Beschreibungen der Systemintegration der Handlungserkennung in Kapitel 5 erläutert werden. Eine Trennung des Verfolgens vom Erkennen erlaubt eine stärkere Modularisierung, die ein Austauschen einzelner Komponenten erlaubt. Außerdem ermöglicht es die Trennung, spezialisierte Verfahren zu verwenden.

Doch auch die verwendeten Methoden zur Verfolgung nutzen zum Teil die probabilistischen Prinzipien, die im Folgenden erläutert werden. Mit der Körperverfolgung von Schmidt u. a. [2006] kommt zum Beispiel ein Condensation-Verfahren zum Einsatz und das Verfolgen hautfarbener Region von Fritsch [2003] wird mit einem Kalman-Filter realisiert.

4.1 Sequenzielle Monte Carlo Methoden

Da der in dieser Arbeit für die Handlungserkennung verwendete Condensation-Algorithmus zur Klasse der *Sequenziellen Monte Carlo* (SMC) Filter gehört, ist es erforderlich, die diesen Filtern zugrundeliegenden Gedanken und Theorien etwas genauer zu betrachten. Eine ausführliche Beschreibung der SMC-Methoden, die über die kurze Einführung an dieser Stelle hinausgeht, geben Doucet u. a. [2001]. In dem Buch werden zahlreiche Anwendungsfälle erläutert, aber auch Beweise für die Konvergenz der Methoden geführt. Die folgenden Erläuterungen und Darstellungen basieren zum Teil auf der einführenden Beschreibung Doucets (siehe Doucet u. a. [2001], Kap. 1) und dem Artikel zu Partikel-Filtern von Arulampalam u. a. [2002].

Am Beispiel der Gestenerkennung ist es das Ziel, aus der beobachteten Bewegung einer Person auf die von ihr ausgeführte Geste oder ihre Körperkonfiguration zu schließen. Die Frage ist folglich, wie aus der Beobachtung, die zum Beispiel in Form der Trajektorie Z der Hand vorliegt, auf den nicht beobachtbaren Zustand des Körpers geschlossen werden kann.

4.1.1 Rekursive bayesche Filter

Die SMC-Methoden und somit auch der CTR sind probabilistische Simulationsverfahren, welche die Idee der rekursiven bayeschen Filter aufgreifen, um Lösungen für Probleme der oben beschriebenen Kategorie zu finden.

In vielen Anwendungen werden die Messwerte sequenziell beobachtet und die Inferenzen sollen zur Laufzeit gemacht werden. Aus diesem Grund kommt sequenziellen oder rekursiven Verfahren eine besondere Bedeutung zu, da in diesen Verfahren das Wissen über den Zustand des Systems im letzten Zeitschritt $t-1$ die Grundlage für die Schätzung des aktuellen Zustands bildet. Die Idee ist, dass die Parameter, die zu einer Sequenz von Beobachtungen geführt haben, bestimmt werden können. Der Zustand eines solchen stochastischen Prozesses ist für jeden Zeitschritt mit der Zufallsvariable q_t beschrieben, die Entwicklung dieses Zustands über die Zeit beschreibt die Systemdynamik Ω:

$$\Omega_{t-1} = \{q_a, q_2, ..., q_{t-1}\}\,. \tag{4.3}$$

Die Systemdynamik beschreibt abstrakt das Verhalten des Systems, eine konkrete Beobachtung ist nicht mit inbegriffen. Eine Methode, Beobachtungssequenzen, denen ein Zufallsprozess zugrundeliegt, zu modellieren, sind die Markov-Prozesse. Basierend auf den vorherigen Zuständen Ω_{t-1}, die die Systemdynamik beschreiben, ist es möglich, eine Wahrscheinlichkeitsdichtefunktion für den aktuellen Zustand q_t aufzustellen. Ein Markov Prozess erster Ordnung liegt vor, wenn die Funktion der Wahrscheinlichkeitsdichte — die pdf — des aktuellen Zustands q_t nur von seinem direkten Vorgänger q_{t-1} abhängt:

$$p(q_t|q_{t-1}) = p(q_t|\Omega_{t-1}). \tag{4.4}$$

Entsprechend werden bei einem Markov Prozess k-ter Ordnung die Zustände der letzten k Zeitschritte mit einbezogen. Angenommen, die benötigte pdf $p(q_{t-1}|Z(1{:}t\text{-}1))$ des Zustands q_{t-1} unter der Bedingung der Beobachtung bis zum vorherigen Zeitschritt liegt vor, so kann die pdf des aktuellen Zustands nun mit dieser und der Systemdynamik über die Chapman-Kolmogorov Gleichung vorhergesagt werden. Es ergibt sich die a priori Wahrscheinlichkeitsdichte:

$$p(q_t|Z(1{:}t\text{-}1)) = \int p(q_t|q_{t-1})\; p(q_{t-1}|Z(1{:}t\text{-}1))\; dq_{t-1}. \tag{4.5}$$

Das Integral setzt sich aus der Wahrscheinlichkeit des Zustands q_t unter der Bedingung des vorherigen Zustands q_{t-1} sowie der Wahrscheinlichkeit für den vorherigen Zustand unter der Bedingung der Beobachtung $Z(1{:}t\text{-}1)$ zusammen. Für ein erfolgreiches Verfolgen

oder Erkennen muss jedoch die aktuelle Beobachtung beachtet werden. Hierfür wird die a priori Wahrscheinlichkeitsdichte $p(q_t|Z(1\text{:t-}1))$ unter Verwendung der Beobachtung $Z(t)$ zur a posteriori Verteilung $p(q_t|Z(1\text{:t}))$ aktualisiert. Die a posteriori pdf kann über den Satz von Bayes aus der a priori pdf und der Beobachtung erschlossen werden:

$$\begin{aligned} p(q_t|Z(1:t)) &= p(q_t|Z(t), Z(1\text{:t-}1)) \\ &= \frac{p(Z(t)|q_t, Z(1\text{:t-}1))\ p(q_t|Z(1\text{:t-}1))}{p(Z(t)|Z(1\text{:t-}1))} \\ &= c\ p(Z(t)|q_t, Z(1\text{:t-}1))\ p(q_t|Z(1\text{:t-}1)) \\ &= c\ p(Z(t)|q_t)\ p(q_t|Z(1\text{:t-}1)). \end{aligned} \tag{4.6}$$

Hierbei ist

$$c = \frac{1}{p(Z(t)|Z(1\text{:t-}1))} = \int p(Z(t)|q_t)\ p(q_t|Z(1\text{:t-}1))\ dq_t \tag{4.7}$$

eine Normalisierungskonstante, die von der Ähnlichkeitsfunktion (engl. likelihood function) $p(Z(t)|q_t)$ abhängt. In der Graphik 4.1 wird dieser Prozess des Vorhersagens der a priori pdf durch die Systemdynamik und Aktualisierung der Schätzung mit den aktuellen Messdaten zur a posteriori pdf visualisiert. Das rote z in der Graphik zeigt in diesem Beispiel exemplarisch den Wert der aktuellen Messung an. Es wird deutlich, wie sich die Funktion der Wahrscheinlichkeitsdichte an diesen anpasst, aber auch andere Hypothesen für den Systemzustand behält. Dadurch dass die a posteriori pdf die Grundlage der nächsten Vorhersage ist, wird die Rekursion des bayeschen Filters geschlossen.

Liegen Beobachtungen in einem Modell eines linearen gaußschen Zustandsraums vor, kann die a posteriori Verteilung mit dem Filter von Kalman [1960] geschätzt werden. Ist aber die wirkliche Wahrscheinlichkeitsdichte nicht normalverteilt, weil sie zum Beispiel bimodal oder stark verzerrt ist, kann diese nicht über eine Normalverteilung beschrieben werden. Für nur teilweise beobachtbare Daten in einem endlichen Zustandsraum, die eine Markovkette bilden, kann die Technik der HMM verwendet werden.

Die SMC-Methoden sind simulationsbasierte Methoden, die einen überzeugenden Weg zur Schätzung der a posteriori pdf ermöglichen, auch wenn nicht normalverteilte und nichtlineare Beobachtungen vorliegen. In den letzten Jahren sind diese Methoden unter unterschiedlichen Namen bekannt geworden, etwa *Bootstrap Filter*, *Condensation*, *Particle Filter*, *Monte Carlo Filter*, *Interacting Particle Approximation* oder *Survival of the fittest*.

4.1.2 Das Sequential Importance Sampling

Die Idee der SMC-Methoden ist, dass die a posteriori Wahrscheinlichkeitsverteilung nicht exakt berechnet werden muss, sondern über eine Simulation approximiert wird. Das *Sequential Importance Sampling* (SIS), auch bekannt unter der Bezeichnung *Particle Filter*,

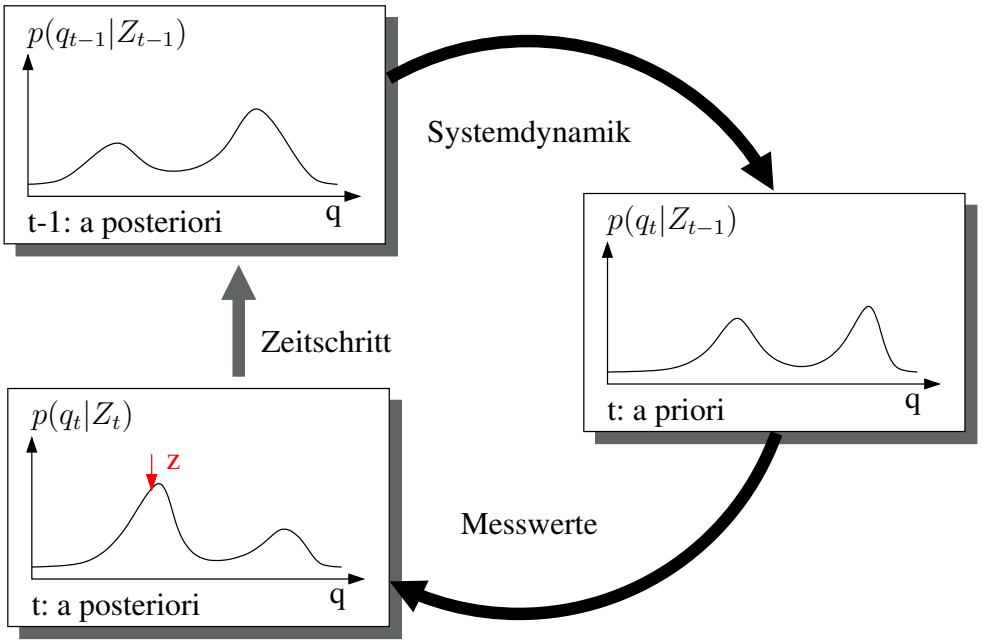


Abbildung 4.1: Schematische Darstellung der Schritte im rekursiven bayeschen Filter. Aus der Wahrscheinlichkeitsdichte des letzten Zeitschrittes $t-1$ wird mit der Systemdynamik die a priori Dichte des aktuellen Zustands ermittelt. Diese Schätzung wird mit den akutellen Messwerten, angedeutet mit z, angepasst und ist Grundlage für den nächsten Zeitschritt.

realisiert einen rekursiven bayeschen Filter mit der Idee der SMC-Simulation. Die pdf wird über eine große Anzahl von N gewichteten Stichproben, den *Sample* oder den *Partikeln* (engl. Particle), approximiert. Jedes dieser zufällig verteilten Partikel $\mathbf{s}^{(n)}$ ist ein Vektor im m-dimensionalen Merkmalsraum und repräsentiert mit seinem Gewicht $\pi^{(n)}$ eine Stützstelle der Wahrscheinlichkeitsdichte:

$$\mathbf{s}^{(n)} = (\mathbf{x}) \text{ mit } (\mathbf{x}) = (\mathbf{x_1}, \ldots, \mathbf{x_m}). \tag{4.8}$$

Wie eine Wahrscheinlichkeitsverteilung für einen eindimensionalen Merkmalsraum approximiert werden kann, zeigt die Abbildung 4.2. Die Menge N der Partikel mit ihren zugehörigen Gewichten bildet das *Partikelset*:

$$S_t = \left\{ (\mathbf{s}_t^{(1)}, \pi_t^{(1)}), \ldots, (\mathbf{s}_t^{(N)}, \pi_t^{(N)}) \right\}. \tag{4.9}$$

Die Zustandsfolge bis zum aktuellen Zeitpunkt ist wie gehabt $\Omega_t = \{q_1, \ldots, q_t\}$ (s. Gleichung 4.3) und die Sequenz der Partikel ist $\mathrm{SP} = \{s_1^{(i)}, \ldots, s_t^{(i)}\}$. Für den Fall das $N \to \infty$ kann nun die a posteriori pdf zum Zeitpunkt t approximiert werden als

$$p(\Omega_t|Z(1:t)) \approx \sum_{i=1}^{N} \pi_t^{(i)} \delta(\Omega_t - \mathrm{SP}_t^{(i)}). \tag{4.10}$$

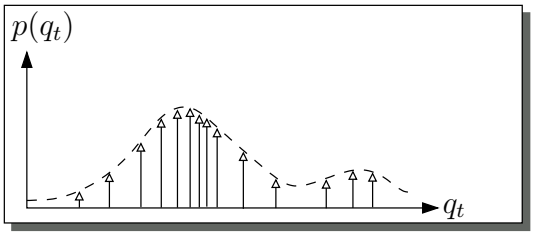


Abbildung 4.2: Approximation der pdf mit einer Menge von zufällig verteilten Partikeln $\mathbf{s}^{(n)}$. Die Höhe der Pfeile entspricht dem Gewicht $\pi^{(n)}$.

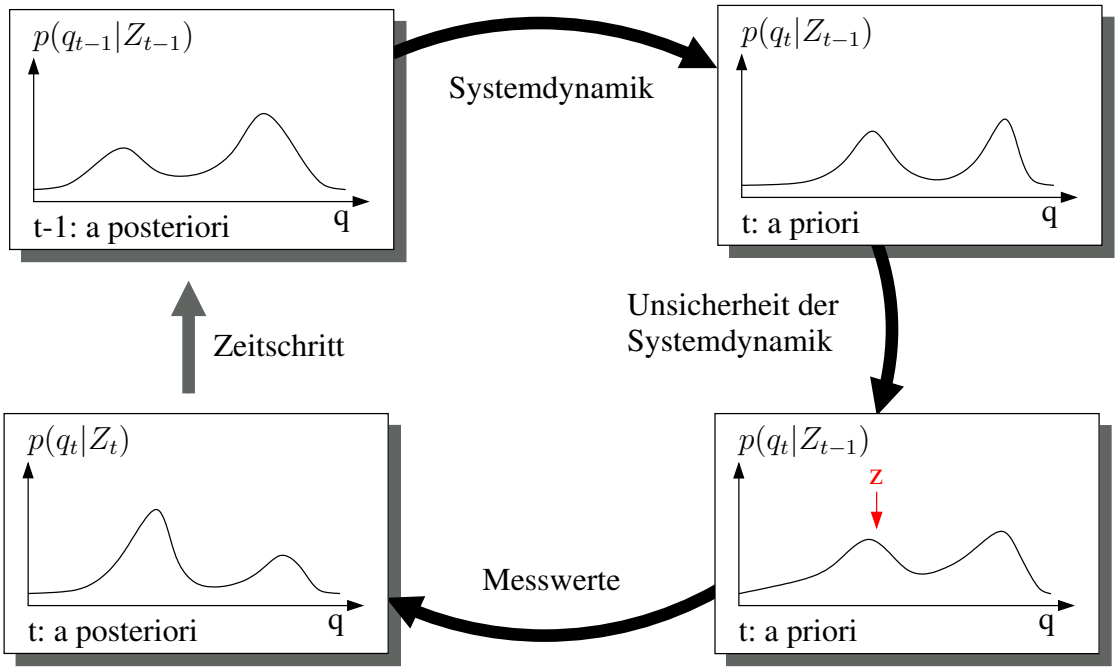


Abbildung 4.3: Schematische Darstellung der Rekursion beim *Sequential Importance Sampling*.

Es entsteht die in Abbildung 4.3 dargestellte Schleife zur Propagierung der N gewichteten Partikel. In jedem Schritt wird die Schätzung vorhergesagt und anhand der Messwerte aktualisiert. Hierbei fließen die Messwerte über die Berechnung der Gewichte $\pi_t^{(n)}$ in die Approximation ein. Im Vergleich zum bayeschen Filter wird außerdem die a priori Verteilung etwas verrauscht, um Unsicherheiten in der Systemdynamik zu modellieren. Interessant ist hierbei, dass parallel mehrere Hypothesen verfolgt werden können, also keine Normalverteilung vorliegen muss.

Die Wahl der Gewichte basiert auf dem Prinzip der gewichteten Stichprobe, dem *Importance Sampling*, das unter anderem von Doucet [2000] vorgestellt wird. Es gelte die Annahme, dass es eine Wahrscheinlichkeitsdichte $p(s) \propto r(s)$ gibt, von der es schwierig ist, Stichproben (Partikel) zu nehmen, aber wohingegen $r(y)$ ermittelt werden kann. Wenn außerdem die Partikel $s^{(i)} \sim q(y), i = 1, \ldots, N$ einfach aus der gewichteten Dichte $q(\cdot)$ (engl. importance density) erzeugt werden können, dann ist die gewichtete Approximation der Dichte $p(\cdot)$ gegeben als:

$$p(s) \approx \sum_{i=1}^{N} \pi^{(i)} \delta(s - s^{(i)}). \tag{4.11}$$

Wobei

$$\pi^{(i)} \propto \frac{r(s^{(i)})}{q(s^{(i)})} \tag{4.12}$$

das normalisierte Gewicht des n-ten Partikels ist und den Übergang von der gewichteten Dichte zu der Wahrscheinlichkeitsdichte darstellt.

Nutzt man dieses *Importance Sampling* und sind die Partikel $\mathrm{SP}_t^{(i)}$ aus einer gewichteten Dichte ermittelt, können die Gewichte aus Gleichung 4.10 berechnet werden:

$$\pi_t^{(i)} \approx \frac{p(\mathrm{SP}_t^{(i)}|Z(1{:}\mathrm{t})}{q(\mathrm{SP}_t^{(i)}|Z(1{:}\mathrm{t})}. \tag{4.13}$$

Wenn in jedem der sequenziellen Schritte nur eine Erwartung der Wahrscheinlichkeitsdichte $p(q_t|Z(1:t)$ benötigt wird und die Beobachtung als Markov-Kette erster Ordnung modelliert werden kann, haben Arulampalam u. a. [2002] gezeigt, dass die Gewichte sequenziell berechnet werden können:

$$\pi_t^{(i)} \approx \pi_{t-1}^{(i)} \frac{p(Z(t)|\mathbf{s}_t^{(i)})\ p(\mathbf{s}_t^{(i)}|\mathbf{s}_{t-1}^{(i)})}{q(\mathbf{s}_t^{(i)}|\mathbf{s}_{t-1}^{(i)}, Z(t)}. \tag{4.14}$$

Auch die a posteriori Dichte $p(q_t|Z(1{:}\mathrm{t})$ kann nun für jeden Zeitschritt t approximiert werden:

$$p(q_t|Z(1{:}\mathrm{t})) \approx \sum_{i=1}^{N} \pi^{(i)} \delta(q_t - \mathbf{s}_t^{(i)}). \tag{4.15}$$

Ein Problem des SIS liegt darin, dass das Verfahren mit der Anzahl der Rekursion scheitert, weil die Verteilung zunehmend aus Partikeln mit vernachlässigbarem Gewicht besteht und nur wenige Partikel ein hohes Gewicht haben. Nach wenigen Zeitschritten wird also mit fast allen Partikeln eine Hypothese verfolgt, andere Hypothesen werden vernachlässigt. Da die Varianz der Gewichte nur zunehmen kann, kann dieses degenerative Verhalten auch nicht vermieden werden (vergleiche Arulampalam u. a. [2002]). Lösungen dieses Problems bestehen einerseits in einer guten Wahl der gewichteten Dichte oder in einem Wiederholen des Nehmens einer Stichprobe, dem *Resampling*.

4.1.3 Sampling Importance Resampling Filter

Eine Form des Particle Filters, die ein *Resampling* verwendet, ist der *Sampling Importance Resampling* (SIR) Filter, der auch unter der Bezeichnung *Conditonal Density Propagation* (Condensation) von Isard u. Blake [1996] bekannt ist. Für die gewichtete Dichte wird die a priori Wahrscheinlichkeitsdichte gewählt:

$$t(\mathbf{s}_{t-1}^{(i)}, Z(t)) = p(\mathbf{s}_t^{(i)}|\mathbf{s}_{t-1}^{(i)}). \tag{4.16}$$

Infolge dieser Annahme kann auch die Aktualisierung der Gewichte aus der Gleichung 4.14 vereinfacht werden:

$$\pi_t^{(i)} \approx \pi_{t-1}^{(i)}\ p(Z(t)|\mathbf{s}_t^{(i)}). \tag{4.17}$$

Da das Resampling in jedem Zeitschritt ausgeführt wird, ist $\pi_{t-1}^{(i)} = 1/N\ \forall i$ und somit

$$\pi_t^{(i)} \approx p(Z(t)|\mathbf{s}_t^{(i)}). \tag{4.18}$$

Die gewichtete Dichte ist unabhängig von der Beobachtung. Die Partikel werden folglich ohne Beachtung dieser nur durch die Systemdynamik im Merkmalsraum verteilt. Auch werden durch das *Resampling* in jedem Schritt Partikel mit einem hohen Gewicht bevorzugt ausgewählt und können so im resultierenden Partikelsatz mehrfach fast identisch vorkommen. Da aber bei typischen Bildverarbeitungsproblemen die Beobachtung verrauscht ist, ist ein mögliches degeneratives Verhalten des Partikelsatzes unwahrscheinlich.

Isard u. Blake [1998a] haben den Condensation-Algorithmus so erweitert, dass automatisch zwischen mehreren Bewegungsmodellen umgeschaltet werden kann und somit nicht nur ein Verfolgen sondern auch eine Klassifikation möglich ist. Hierfür wurden die Partikel um eine Modellmarkierung μ erweitert, die bestimmt, zu welchem Bewegungsmodell ein Partikel gehört (siehe auch Gleichung 4.8):

$$\mathbf{s}^{(i)} = (\mathbf{x}, \mu) \text{ mit } (\mathbf{x}) = (x_1, \ldots, x_m),\ \mu \in \{1 \ldots l\}\,. \tag{4.19}$$

In Abhängigkeit des Modells μ zu dem ein Partikel gehört, werden nun die Partikel vom Zeitschritt t nach $t+1$ propagiert. Die Systemdynamik ist jetzt von dem entsprechenden Modell abhängig. Summiert man nun in jedem Zeitschritt für jedes Modell die Gewichte der zugehörigen Partikel auf, entsteht die Modellwahrscheinlichkeit:

$$P_t(\mu_i) = \sum_{n=1}^{N} \begin{cases} \pi_t^{(n)} & , \text{ falls } \mu_i \in \mathbf{s}_t^{(n)} \\ 0 & , \text{ sonst} \end{cases}\,. \tag{4.20}$$

An dieser lässt sich in jedem Zeitschritt ablesen, welches Modell das größte aufsummierte Gewicht hat und somit aktuell am besten zu der Beobachtung passt.

Zusammenfassung

Die Sequenziellen Monte Carlo Methoden lassen sich für viele Probleme anwenden. Es ist einerseits möglich, einen Verfolgungsalgorithmus mit ihnen zu implementieren, bei dem der Merkmalsraum, in dem gesucht werden muss, auf relevante Bereiche reduziert wird. Andererseits können die Methoden aber auch zum Erkennen von Mustern in einem Merkmalsraum benutzt werden, wie an dem im nächsten Abschnitt behandelten *Condensation Trajectory Recognition* Algorithmus gezeigt wird.

4.2 Der CTR-Algorithmus

Aufbauend auf den Erklärungen des letzten Abschnittes wird im Folgenden die Implementierung des *Condensation Trajectory Recognition* (CTR) vorgestellt, die auf der Arbeit von Black u. Jepson [1998] zur Klassifikation von Handgesten basiert.

Der CTR-Algorithmus ist eine SMC-Methode und kann so als Verallgemeinerung der HMM gesehen werden, die eine diskrete Menge von Zuständen mit Transitionen zwischen den Zuständen erlaubt. Das Verfahren des CTR bezieht jedoch in die Erkennung eines individuellen Zustands die probabilistische Übereinstimmungen einer ganzen zeitlichen Trajektorie mit ein. In diesem Aspekt ähnelt der CTR dem DTW, aber der Vergleich zwischen skalierten Modellen und den Daten ist in ein probabilistisches System eingebettet. Weitere Vorteile liegen darin, dass über die Menge der Partikel die Rechenkapazität automatisch auf die interessanten Teile konzentriert wird und dass gleichzeitig ein Erkennen und automatisches Segmentieren ermöglicht wird.

In dieser Arbeit wird das Verfahren erweitert, so dass sowohl mehrdimensionale Merkmalsvektoren wie auch der Kontext, in dem eine Handlung ausgeführt wird, integriert werden können. Black und Jepson verwenden hingegen einen zweidimensionalen Merkmalsvektor, der die Geschwindigkeit der Hand in x und y Richtung beschreibt. Zuerst wird aber das Grundkonzept des CTR erläutert, hierbei wird von der Herkunft und Art der Trajektorie abstrahiert.

Jedes Modell $\mathbf{m}$ einer Handlung besteht aus einer n-dimensionalen Trajektorie, welche die Bewegung der Hand oder des Objektes beschreibt (s. Gleichung 4.21). Diese Modelle werden in jedem Zeitschritt t, in jeder Modelldimension in einem Zeitfenster w mit dem Merkmalsvektor Z_a(t-w:t) der Messwerte verglichen. Dieser parametrisierte Vergleich ist mit dem Parametervektor $\mathbf{s}$ beschrieben (s. Gleichung 4.22). Der Vektor $\mathbf{s}_t$ beschreibt für den Zeitpunkt t zu welchem Modell μ der Partikel gehört. Des Weiteren wird die zeitliche Position in der Modelltrajektorie mit ϕ angegeben. Ein Partikel wird zum Zeitpunkt t_0 initialisiert und daraufhin in jedem Zeitschritt propagiert. Um Varianzen in der Ausführung einer Geste zu erlauben, ist mit dem Parameter α eine Skalierung der Amplitude des Modells möglich und der Parameter ρ erlaubt eine Zeitskalierung, dargestellt in Graphik 4.4 und beschrieben in Gleichung 4.21.

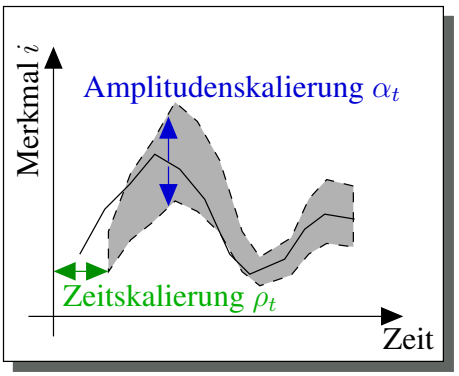
$$\mathbf{m}^{(\mu)} = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T\}, \quad \mathbf{x}_t = (f_1, f_2, \ldots, f_n) \tag{4.21}$$

$$\mathbf{s}_t = (\mu_t, \phi_t, \alpha_t, \rho_t) \tag{4.22}$$

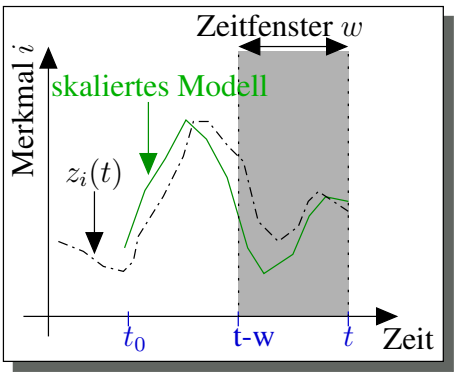
Das Ziel des Condensation-Algorithmus ist es, den optimalen Parametervektor $\mathbf{s}_t$ zu bestimmen, so dass ein skaliertes Modell $\mathbf{m}^{(\mu)}$ die größte Übereinstimmung mit der Beobachtung der letzten Zeitschritte Z_a(t-w:t) hat. Dieses wird nach dem Prinzip des SIR-Filters (siehe Abschnitt 4.1.3) mit der Propagierung von N gewichteten Partikeln $\mathbf{s}_t^{(i)}$ erreicht, die in einem Partikelsatz S_t zusammengefasst sind:

$$S_t = \left\{ (\mathbf{s}_t^{(1)}, \pi_t^{(1)}), \ldots, (\mathbf{s}_t^{(N)}, \pi_t^{(N)}) \right\}. \tag{4.23}$$

Ein Partikelsatz S_t repräsentiert die a posteriori Verteilung $p(\mathbf{s}_t|Z_a(t))$ zum Zeitpunkt t. Die Gewichte $\pi_t^{(n)}$ der Partikel $\mathbf{s}_t^{(n)}$ sind die normalisierten Wahrscheinlichkeiten $p(Z_a(t)|\mathbf{s}_t^n)$. Diese werden durch den Vergleich jeder skalierten Komponente der Modelltrajektorien in den letzten w Zeitschritten mit der Beobachtung berechnet. Das Partikel beschreibt dabei



(a) Auswirkung der Skalierung eines Modells mit den Parametern α und ρ in einer Merkmalsdimension.



(b) Vergleich eines skalierten Modells mit den Messwerten in einer Merkmalsdimension.

Abbildung 4.4: Das Bewegungsmodell μ, auf das im Partikel $\mathbf{s}_t^{(i)}$ verwiesen wird, wird zuerst skaliert (a) und anschließend in jeder Merkmalsdimension mit der Beobachtung $Z_a(t)$ verglichen. Die Güte dieses Vergleiches spiegelt sich im Gewicht $\pi_t^{(i)}$ wieder.

das mit α und ρ skalierte Modell $\mathbf{m}$ in der Dimension i an der Position ϕ. Für die Differenz zwischen Modell und Beobachtung wird eine Gauß-Verteilung mit der Varianz σ_i^2 für jeden Punkt der Trajektorie angenommen:

$$p(z_i(\text{t-w:t})|\mathbf{s}_t^{(n)}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp \frac{\sum_{j=t-w}^{t} \left(z_i(j) - \alpha m_{(\phi-\rho j),i}^{(\mu)} \right)^2}{2\sigma_i^2 (w-1)}. \tag{4.24}$$

Aus diesen bedingten Wahrscheinlichkeiten $p(z_i(\text{1:t})|\mathbf{s}_t^{(n)})$ lassen sich die normalisierten Wahrscheinlichkeiten der Gewichte berechnen:

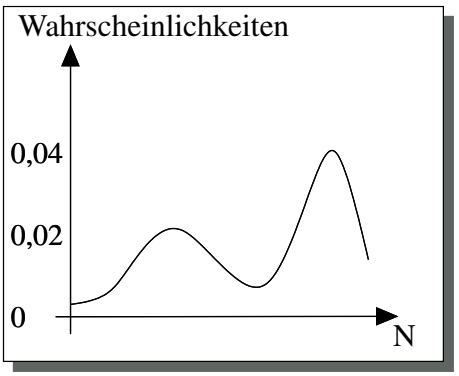
$$\pi_t^{(n)} = \frac{p(z_i(\text{1:t})|\mathbf{s}_t^{(n)})}{\sum_{j=1}^{N} p(z_i(\text{1:t})|\mathbf{s}_t^{(j)})} \quad \text{mit} \tag{4.25}$$

$$p(Z_a(t)|\mathbf{s}_t^{(n)}) = \prod_{i=1}^{I} p(z(t)_i|\mathbf{s}_t^{(n)}). \tag{4.26}$$

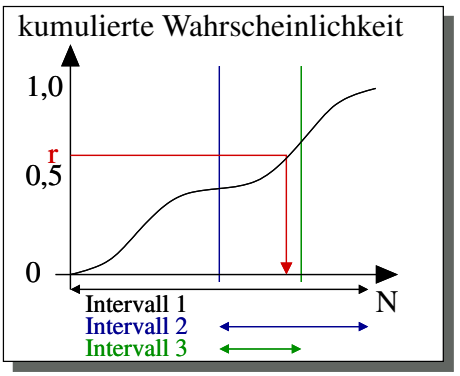
Das Propagieren der gewichteten Partikel besteht aus drei Schritten, die Ergebnisse des letzten Zeitschrittes werden hierbei verwendet.

Selektion: Selektieren von N Partikeln $\mathbf{s}_{t-1}^{(n)}$ entsprechend ihres Gewichtes $\pi_{t-1}^{(n)}$ aus dem Partikelsatz des Zeitschrittes $t-1$. Hierbei werden Partikel mit einem hohen Gewicht bevorzugt, also öfter ausgewählt. Das verwendete Verfahren wird später erläutert.

Prädiktion: Die Parameter jedes Partikels $\mathbf{s}_t^{(n)}$ werden vorhergesagt. Hierzu wird Gaußsches Rauschen zu α_{t-1} und ρ_{t-1} addiert, sowie die Position ϕ_{t-1} in jedem Zeitschritt



(a) Exemplarische Verteilung der Wahrscheinlichkeiten $\pi_t^{(n)}$ der einzelnen Partikel $\mathbf{s}_t^{(n)}$.



(b) Exemplarische Verteilung der aufsummierten Wahrscheinlichkeiten und das Selektionsverfahren mit Intervallschachtelung.

Abbildung 4.5: Das Verfahren zur Auswahl der Partikel über die kumulierten Wahrscheinlichkeiten.

um ρ_t angehoben. Sobald ϕ_t größer als die maximale Modelllänge $\phi_{\max}$ ist, wird ein neuer Partikel $\mathbf{s}_t^{(n)}$ initialisiert.

Aktualisierung: Die Gewichte $\pi_t^{(n)}$ werden auf der Basis von $p(Z_a(1{:}t)|\mathbf{s}_t^{(n)})$ neu berechnet.

Für die Klassifikation werden die gewichteten Partikel genutzt. Die Wahrscheinlichkeit, dass ein Modell μ_i zum Zeitpunkt t abgeschlossen ist, wird mit der Endwahrscheinlichkeit $P_{\text{end}}(\mu_i)$ angegeben. Sie berechnet sich aus der Summe aller Gewichte $\pi_t^{(n)}$ eines speziellen Modells mit einem Wert von $\phi_t > 0.9\phi_{\max}$:

$$P_{\text{end},t}(\mu_i) = \sum_{n=1}^{N} \begin{cases} \pi_t^{(n)} & \text{, falls } \mu_i \in \mathbf{s}_t^{(n)} \wedge (\phi > 0.9\phi_{\max}) \\ 0 & \text{, sonst} \end{cases} \tag{4.27}$$

Selektionsmethode

Die Selektion eines Partikels basiert auf den aufsummierten Wahrscheinlichkeiten $\pi_t^{(i)}$ der Partikel. Es wird über eine Intervallschachtelung das Partikel ausgewählt, dessen aufsummierte Wahrscheinlichkeit am dichtesten an einer Zufallszahl zwischen 0 und 1 liegt. Da Partikel mit einem hohen Gewicht mehr zu der aufsummierten Wahrscheinlichkeit beitragen, werden diese häufiger ausgewählt als solche mit niedrigerem Gewicht. Das Prinzip des Verfahrens, das in der Effizienzklasse $O(\log N)$ liegt, ist in der Graphik 4.5 visualisiert.

4.2.1 Modellierung von Bewegungssequenzen

Black u. Jepson [1998] verwenden in ihrem Algorithmus eine einfache hierarchische Struktur aus Kind- und Elternmodellen um erwartete Sequenzen von Bewegungen zu modellieren. Neben dem Kindmodell μ wird auch das Elternmodell ν im Vektor des Partikels gespeichert:

$$\mathbf{s}_t^{(i)} = (\nu_t, \mu_t, \phi_t, \alpha_t, \rho_t). \tag{4.28}$$

Ist ein Partikel in seinem Endzustand angekommen, wird entsprechend der Matrix mit Übergangswahrscheinlichkeiten $A^{(\nu)}_{\mu^i \to \mu^j}$ des Elternmodells ν ein neues Kindmodell ausgewählt. Existiert kein Übergang in ein weiteres Kindmodell, so wird die Endwahrscheinlichkeit des Modells $P_{\text{end}}(\mu^i)$ zu der Endwahrscheinlichkeit des Elternmodells aufsummiert:

$$P_{Parent-End}(\nu^i) = P_{Parent-End}(\nu^i) + p_{\text{end}}(\mu^i). \tag{4.29}$$

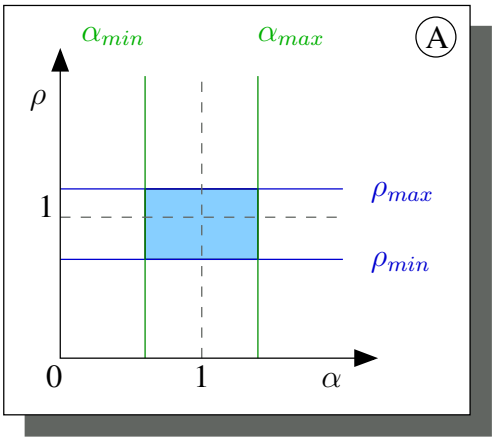
Initialisiert werden nur Kindmodelle μ, die aus der Matrix der Übergangswahrscheinlichkeiten als Startmodell hervorgehen. Diese Methode ermöglicht es, eine erwartete Sequenz von Basisaktionen, das sind die Kindmodelle, zu modellieren und auch alternative Modelle und Auslassungen im Ablauf der Modellsequenz zu berücksichtigen. Die Bewegungen einer Geste, dieses sind der Gestenzug sowie die Vorbereitung und Rückführung (siehe 2.3.2, Seite 23), können über Kindmodelle modelliert werden; die Geste wird dann durch ein Elternmodell repräsentiert. Dennoch ist dieser Ansatz anfällig für Fehlerkennungen, denn eine nicht oder fälschlich erkannte Bewegung kann die Menge der erwarteten Bewegungen verändern und somit die korrekte Erkennung folgender Bewegungen verhindern. Um dieses Problem zu lösen, müssen zum einen mehr Kontextinformationen eingebunden werden, andererseits wäre auch auf dieser Ebene ein probabilistisches Verfahren wünschenswert, das mit unterschiedlichen Hypothesen für Aktivitäten arbeiten kann und die erkannten Bewegungen probabilistisch einbezieht.

4.2.2 Initialisieren eines Partikels

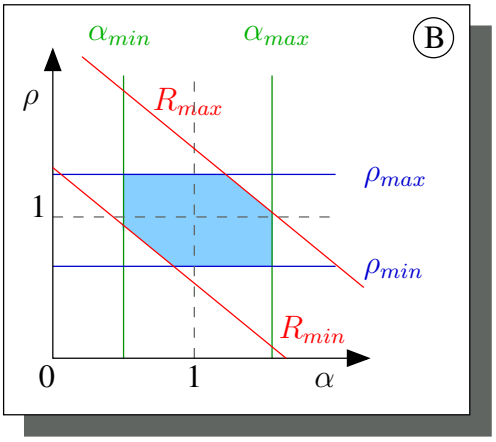
Neben dem Start des Algorithmus muss noch in drei weiteren Fällen ein neues Partikel für den CTR initialisiert werden. Sobald ein Partikel den Endzustand des zugehörigen Kindmodells erreicht, werden die Parameter dieses Partikels im nächsten Schritt neu belegt. Wenn das Kindmodell μ zu einem Elternmodell gehört, wird ein möglicher Nachfolger entsprechend der Übergangswahrscheinlichkeit des Elternmodells ν gewählt. Existiert kein gültiger Nachfolger, wird der Partikel zufällig neu belegt, indem zuerst ein neues Elternmodell ν und anschließend ein mögliches Kindmodell μ gewählt wird. Werden nur Kindmodelle verwendet, kann ein Kindmodell μ zufällig selektiert werden. Des Weiteren wird die Position ϕ zurückgesetzt und die Skalierungsparameter α und ρ werden in ihren Grenzen initialisiert.

Damit der Algorithmus sich schnell auf veränderte Beobachtungen einstellen kann, wird in jedem Schritt ein bestimmter Prozentsatz N_{neu} des Partikelsets neu initialisiert. Der letzte Fall tritt sehr selten ein, denn ist die Bewertung eines Partikels zu schlecht, wird dieses auch neu initialisiert. Doch die Selektion der Partikel verhindert recht zuverlässig, dass solche Partikel in das neue Partikelset übernommen werden.

Nach Black u. Jepson [1998] werden die Parameter α und ρ unabhängig von einander mit einer gleichverteilten Wahrscheinlichkeit in den Grenzen $\alpha_{min} \dots \alpha_{max}$ beziehungsweise $\rho_{min} \dots \rho_{max}$, wie in Graphik 4.6(a) dargestellt, zufällig gewählt. Der Wert von α skaliert hierbei die Amplitude der Modellwerte, ρ hingegen die zeitliche Ausdehnung. Betrachten



(a) Initialisierung der α und ρ Werte nach Black u. Jepson [1998].



(b) Zusätzliche Einschränkung der möglichen Skalierung durch den minimalen und maximalen Aktionsradius.

Abbildung 4.6: Wertebereiche der Skalierungsparameter.

wir einen Geschwindigkeitswert v der mit $\alpha > 1$ skaliert wird, so wird die Trajektorie T im Raum länger: Die Bewegung reicht weiter. Ebenso kann die Bewegung auch durch $\rho > 1$ länger werden, da sie in diesem Fall bei ursprünglicher Geschwindigkeit länger anhält.

Dieses Abhängigkeitsverhältnis ist für den allgemeinen Fall einer beliebigen Trajektorie zutreffend. Betrachtet man jedoch Zeigegesten oder Handlungen, ist eine Abhängigkeit zwischen den Parametern geben. Die Analyse dieser gegenseitigen Abhängigkeit hilft, den Suchraum für die optimale Kombination aus α und ρ zu präzisieren.

Der mögliche Aktionsradius R einer Bewegung im Raum erlaubt eine Relation zwischen α und ρ zu formulieren (siehe Gleichung 4.30). Visualisiert wird der neue Skalierungsbereich in Abbildung 4.6(b).

$$R_{min} < (\alpha + \rho) < R_{max} \tag{4.30}$$

Zum anderen kann eine modellierte typische Bewegung aber nur in einem begrenzten Rahmen verkürzt oder verlängert werden. Das Gleiche gilt auch für die Geschwindigkeit. Deswegen bleiben die Grenzen für α und ρ bestehen.

$$\rho_{min} < \rho < \rho_{max} \tag{4.31}$$
$$\alpha_{min} < \alpha < \alpha_{max} \tag{4.32}$$

Aber dadurch dass Parameterpaarungen, die sehr kurze oder sehr lange Bewegungen bedingen würden, ausgeschlossen sind, können die minimalen und maximalen Werte für α und ρ großzügiger gewählt werden. In diesem hier beschriebenen Ansatz werden die Skalierungsparameter nicht mehr unabhängig in ihren Grenzen gewählt und propagiert, sondern

die Grenzen, in denen ein Parameter zufällig gesetzt wird, werden von dem jeweils anderen Parameter beeinflusst.

4.3 Generieren neuer Modelle

> Eine besondere Bedeutung im beschriebenen CTR-Verfahren haben die Modelle μ, da mit ihnen die Gesten und Aktionen beschrieben sind, die erkannt werden können. Ein Modell soll für die Klasse der beschriebenen Bewegungen möglichst generisch sein, aber andererseits auch diskriminativ zu den anderen möglichen Klassen.

Diese Anforderung für die Modelle zur Gestenerkennung liegt in der gewünschten hohen Erkennungsrate bei gleichzeitig möglichst geringer Fehlerrate begründet. Weitere Anforderungen, die sich aus dem Einsatz des implementierten CTR-Verfahrens ergeben, sind, dass oft nur eine kleine annotierte Trainingsstichprobe verfügbar ist und dass Modelle für neue Szenarien schnell und einfach gebildet werden können müssen.

Die einfachste Methode basiert auf dem Prinzip des „one-shot-learning“, hierfür wird von jeder gewünschten Aktion ein repräsentatives Modell aus einer Sequenz gebildet. Der klare Vorteil ist, dass der Aufwand für die Erzeugung von Trainingsdaten sehr gering ist, doch fordert das Vorgehen eine robuste und auch tolerante Gestenerkennung.

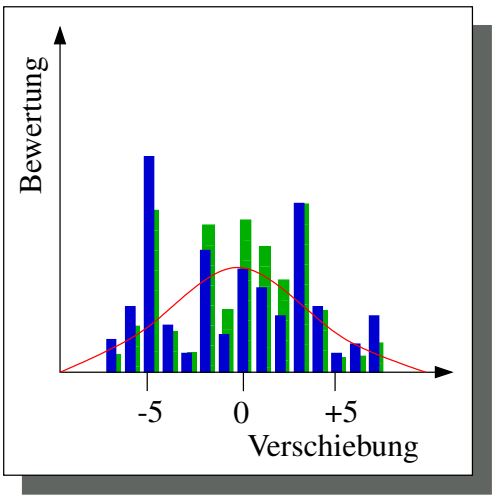
Da die Modelle möglichst generisch sein sollen, sowie außerdem die verwendeten Modell-Varianzen eine fundierte statistische Basis verlangen, wird im nächsten Abschnitt nun ein im Zusammenhang mit dieser Dissertation neu entwickeltes Verfahren vorgestellt. Die Problematik, die in den von Hand annotierten oder auch automatisch erzeugten Sequenzen liegt, ist, dass der charakteristische Teil der Aktion nicht an einer festen Stelle in den Sequenzen anfängt. In manchen Sequenzen ist zum Beispiel zum Beginn der Sequenz ein unbedeutender oder sogar unpassender Bewegungsteil enthalten. Die charakteristischen Teile sind also nicht aligniert und somit ist eine einfache Mittelwertbildung über alle Sequenzen nicht möglich.

Für das notwendige mehrdimensionale Alignement sind in dem Bereich des Sequenzalignements in der Bioinformatik Verfahren entwickelt worden. Beschrieben sind die Methoden unter anderem von Merkl u. Waak [2003]; der weit verbreitete Algorithmus CLUSTAL W wurde von Thompson u. a. [1994] vorgestellt. In Anlehnung an diese für die Bioinformatik entwickelten Verfahren wird nun ein Ansatz für das Alignieren von multidimensionalen Trajektorien entwickelt.

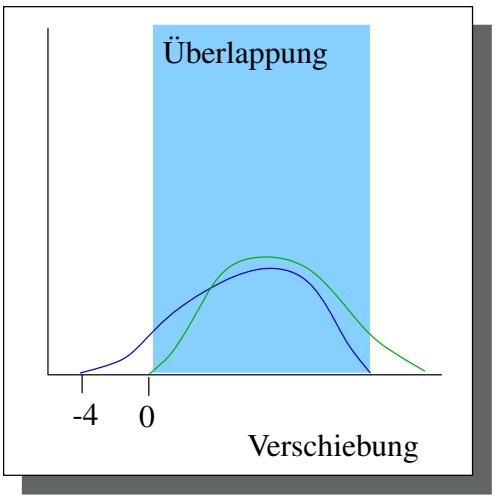
Aus den Messwerten Z_a einer Repräsentation a werden die I annotierten Sequenzen B_i^k (siehe Gleichung 4.33) einer Klasse $k \in K$ eingelesen und jede dieser Sequenzen wird in jeder Dimension unter Beibehaltung der Nulldurchgänge normiert:

$$B_i^k(t_1 : t_e) = \text{norm}\left(Z_a(t^i_{Beginn} : t^i_{Ende})\right). \tag{4.33}$$

Paarweise wird nun eine Kreuzkorrelation dieser normierten Sequenzen B_i^k durchgeführt. Da die optimalen Verschiebungen für die einzelnen Dimensionen unterschiedlich sein kön-



(a) Bewertung (Blau) und gewichtete Bewertung (Grün) der Verschiebung zwischen zwei Modellen.



(b) Verschiebung zweier Modelle zueinander.

Abbildung 4.7: Paarweise wird die beste Verschiebung zwischen zwei Sequenzen ermittelt. Damit der Bereich der Überlappung möglichst groß ist, werden geringe Verschiebungen positiv gewichtet.

nen, wird die Gesamtbewertung $\mathbf{j}$ aus der Summation der Bewertungen $\mathbf{j}_m$ der einzelnen Merkmalsdimensionen ermittelt:

$$\mathbf{j} = \sum_{m=1}^{M} \mathbf{j}_m. \tag{4.34}$$

Die Abbildung 4.7(b) zeigt exemplarisch für eine Dimension die Verschiebung von zwei Sequenzen zueinander. Damit Verschiebungen, die eine möglichst große Überlappung der Sequenzen bewirken, bevorzugt werden, wird die Bewertung mit einer Normalverteilung ($\mathcal{N}$) gewichtet (siehe Graphik 4.7(a)).

Aus dieser gewichteten Bewertung $\mathbf{j}_g$ kann nun die optimale Verschiebung d_{opt} der einzelnen Kreuzkorrelationen ermittelt werden, indem die Verschiebung gewählt wird, für die die Bewertung am höchsten ist:

$$\begin{aligned} d_{\text{opt}} &= \max(\mathbf{j}_g) \\ \text{mit } \mathbf{j}_g &= \mathcal{N}\mathbf{j}. \end{aligned} \tag{4.35}$$

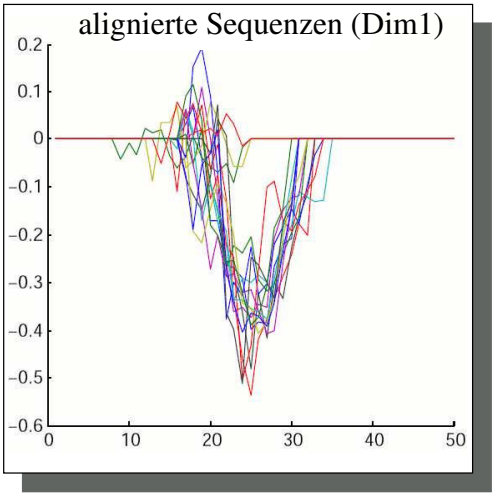
Die Verschiebung und die entsprechende Bewertung für das Sequenzpaar B_i^k, B_j^k werden in die $I \times I$ große Verschiebungsmatrix D^k und Bewertungsmatrix J^k eingetragen:

$$\begin{aligned} D^k(i,j) &= -D^k(j,i) = d_{opt}(B_i^k, B_j^k) \\ J^k(i,j) &= J^k(j,i) = j_{opt}(B_i^k, B_j^k). \end{aligned} \tag{4.36}$$

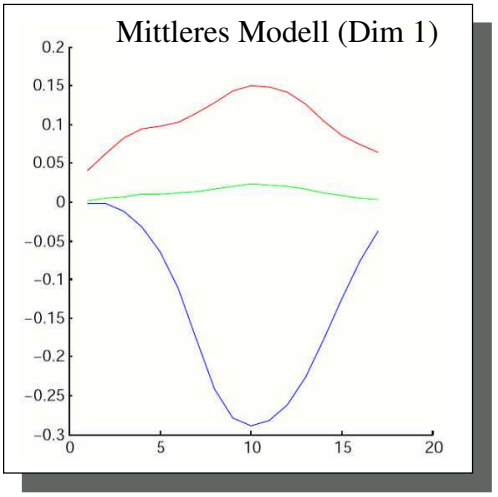
In einem direkten und schnellen Verfahren wird die Sequenz mit der höchsten aufsummierten Bewertung

$$j_{sum}(i) = \sum_{j=0}^{I} j_{opt}(J^k(i,j)) \tag{4.37}$$

ausgewählt und die anderen Sequenzen gemäß ihrer Verschiebung $d_{opt}(B_o^k, B_j^k)$ zur Sequenz B_o^k aligniert. Hierbei werden Sequenzen, deren Verschiebung zur Zielsequenz größer als ihre halbe Sequenzlänge ist $d > t_1 - t_e$ oder deren Korrelationsbewertung niedriger als das arithmetische Mittel aller ist, ausgelassen. Die Sequenzen, die diese Bedingungen nicht einhalten, würden das resultierende Modell stören. Für diese ist es sinnvoller, ein neues Modell zu bauen. Exemplarisch zeigt die Abbildung 4.8 eine Dimension einer Gruppe alignierter Sequenzen und das daraus resultierende Modell.



(a) Eine Gruppe alignierter Sequenzen. Die Sequenzen sind eine Dimension von Zeigegesten, die in einem zylindrischen Koordinatensystem repräsentiert werden.



(b) Das durch Mittelung entstandene Modell. Die blaue Kurve zeigt das mittlere Modell. Die grüne steht für die Varianz und die rote für die Standardabweichung.

Abbildung 4.8: Relativ zu der Sequenz mit der am besten bewerteten Verschiebung werden die anderen Sequenzen verschoben. Aus dieser Überlagerung wird das Bewegungsmodell gebildet. Weitere Beispiele finden sich im Anhang: Abbildung 8.2 bis 8.4.

Da nur der charakteristische Bereich der Sequenzen modelliert werden soll, der in den meisten Sequenzen S_k enthalten ist, werden die Randbereiche, die nur Werte von weniger als einem Viertel der Sequenzen enthalten, abgeschnitten. Von dem verbleibendem Kernbereich wird aus dem arithmetischen Mittel jeder Dimension ein Modell gebildet.

Bei der Verwendung in Experimenten zeigt dieses neu entwickelte globale Alignement seine Funktionalität zur Bildung generischer Modelle.

4.4 Erkennung von Gesten in ihrem Kontext

Die trajektorienbasierte Gestenerkennung, wie sie bisher vorgestellt wurde, verwendet die dynamischen Charakteristika der Handlung. Für verschiedene Anwendungen

wie zum Beispiel Zeichensprache oder andere symbolische Gesten ist dieser Ansatz ausreichend, da die Geste losgelöst aus ihrem räumlichen und symbolischen Kontext interpretiert werden kann. Werden aber Objektmanipulation und Zeigegesten betrachtet, so stehen diese immer im Kontext zu realen Objekten im Raum und zur aktuellen Situation.

Zur Gestenerkennung kann, wie gezeigt, die Bewegung der Hand verwendet werden. Andere Ansätze verwenden nicht die Trajektorie und somit die dynamischen Eigenschaften, sondern beschränken sich auf die räumliche Anordnung der Hand zu den umgebenen Objekten. Die Vorteile beider Ansätze können aber in einem probabilistischen Verfahren, aufbauend auf dem CTR, vereint werden, wie wir in den Veröffentlichungen von Fritsch u. a. [2004] und Hofemann u. a. [2004] gezeigt haben. Das Innovative dieses Ansatzes ist, dass der Kontext, in dem eine Handlung ausgeführt wird, während der Verarbeitung der Trajektorie in den Erkennungsprozess eingebunden wird. Es werden die sensorischen Daten der Bewegung mit den symbolischen Beschreibungen der Szene verbunden. So kann zum Beispiel nicht nur eine Zeigegeste erkannt werden, sondern es wird auch das referenzierte Objekt der Geste zugeordnet. Für eine Geste kann so nicht nur der Bewegungsverlauf definiert werden, sondern auch welches Objekt wann und wo relativ zur Bewegung erwartet wird.

Die symbolische Szenenrepräsentation Θ enthält für jedes Objekt O_l der Szene den Objekttypen und eine eindeutige Objektreferenz, sowie die Position des Objektes:

$$\Theta_t = \{O_1, \ldots, O_l, \ldots, O_L\} \tag{4.38}$$

$$O_l = (\text{type}_l, \text{ID}_l, \text{Pos}_l) \qquad \forall\, l \in \{1, ..., L\}. \tag{4.39}$$

Der Kontext, in dem eine Handlung ausgeführt wird, wird unterschieden in den *situativen* und den *räumlichen* Kontext, die im Folgenden erläutert werden:

- **Situativer Kontext**

  Der situative Kontext umfasst die Vorbedingungen, die für die Ausführung einer Geste erfüllt sein müssen, und den Effekt, den diese für die Szene hat. Für die Handlungserkennung ist der Zustand der Hand, zum Beispiel „Die Hand hält ein Objekt“, von besonderer Bedeutung. Dieser Zustand der Hand wird mit dem *global hand state* (GHS) beschrieben:

  $$\text{GHS}(t) = \{\emptyset | O_l\}. \tag{4.40}$$

  Die Vorbedingung und der Effekt eines Modells arbeiten auf diesem GHS. Zum Beispiel muss für eine Aktion „Greife Tasse“ die Hand frei sein ($GHS(t) = \{\emptyset\}$). Wird diese Aktion aber erkannt, wird der GHS mit dem Symbol O_1 der gegriffenen Tasse belegt ($GHS(t) = \{O_1\}$).

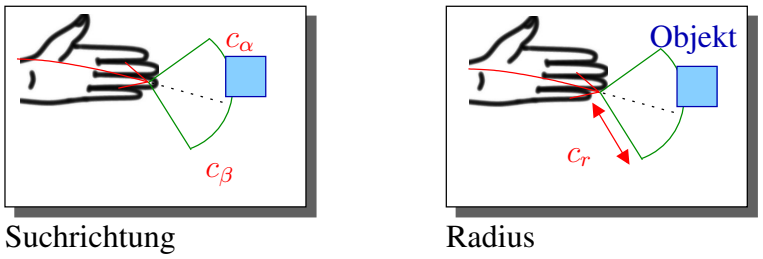


Abbildung 4.9: Die *Context Area* wird über die Suchrichtung und den Radius für ein Bewegungsmodell definiert. Für eine Bewegung kann so definiert werden, wann und wo ein Objekt im Kontext der Bewegung erwartet wird.

- **Räumlicher Kontext**

  Für die Handlungserkennung ist es nötig, ein manipuliertes oder referenziertes Objekt mit der Trajektorie der Hand zu verknüpfen. Ayers u. Shah [1998] beschreiben, dass eine Person, die dicht an einem statischen und vorher bekannten Objekt im Raum ist, einen Zustandsübergang in einem *state model* auslöst. Die Bewegung der Personen beachten sie jedoch nicht. Eine Näherelation, die beschreibt, wann die Hand dicht an einem Objekt ist, reicht aber in komplexen Szenen alleine nicht aus. Denn die verfolgte Hand kann bei einer „Greifen"-Handlung ein Objekt im geringen Abstand passieren, es aber nicht greifen. So können Ambiguitäten entstehen, die nur schwer aufgelöst werden können, da das Objekt zum Beispiel von der Hand oder dem Arm verdeckt wird. Somit kann auch nicht aus dem Verschwinden eines Objektes geschlossen werden, dass es gegriffen wurde.

  Eine Betrachtung der Relation einer Trajektorie zu jedem Objekt der Szene würde eine unnötige hohe Komplexität verursachen. Deswegen wird als neue Herangehensweise ein für die Handlung relevanter Bereich definiert, die *Context Area*. In diesem Bereich des Raumes, relativ zur aktuellen Position der Hand, werden potentiell für die Handlung relevante Objekte erwartet. Ist ein Objekt in dem Bereich vorhanden, wird dieses Auftreten dynamisch in den Erkennungsprozess der Geste einbezogen. Die *Context Area* wird als Kreissegment mit dem Radius c_r und einem Winkelbereich von c_α bis c_β definiert (siehe Abbildung 4.9). Für Objekte, deren Handhabung keine spezielle Richtung vorschreibt, wird die Orientierung c_{orient} der *Context Area* relativ zur Bewegungsrichtung der Hand interpretiert. Ergibt sich jedoch aus der Handhabung des Objektes eine typische absolute Raumrichtung, kann die *Context Area* absolut interpretiert werden.

  Neben der Definition der geometrischen Parameter des symbolischen Kontextes ist es auch nötig, eine neue semantische Beschreibung zu definieren. Mit dem Kontexttypen c_{type} wird der Typ des erwarteten Objektes beschrieben. Des Weiteren kann mit der Kontext Bedeutung c_{imp} spezifiziert werden, ob für die Handlung ein Objekt in der *Context Area* vorhanden sein muss (*necessary*), dieses optional (*optional*) oder nicht notwendig ist (*irrelevant*). Die Aufschlüsselung ist notwendig, da bei der Manipulation eines Objektes dieses zeitweise verdeckt sein kann, aber trotzdem die Erkennung der Handlung möglich sein soll. Der räumliche Kontext C_T setzt sich also aus einem individuellen Kontext $\mathbf{c}_t$ für jeden Zeitschritt t eines Handlungsmodells zusammen:

$$\mathbf{c}_t = (c_{imp}, c_{orient}, c_\alpha, c_\beta, c_r, c_{type}) \quad (4.41)$$
$$C_T = (\mathbf{c}_1, \ldots, \mathbf{c}_t, \ldots, \mathbf{c}_T). \quad (4.42)$$

Der Objekttyp c_{type} kann eine Klasse von Objektem beschreiben, die eine ähnliche Eigenschaft haben, wie zum Beispiel die Klasse der Objekte, die zum Trinken benutzt werden können (Tassen, Flaschen, etc.).

Einbeziehung der Objektinformationen

In den Modellen (siehe Gleichung 4.21) wird zusätzlich die Vorbedingung (*Precondition*) und der Effekt (*Effect*) einer Handlung spezifiziert. Desgleichen enthält das Modell nun auch den zeitlich dynamischen Kontext C_T und repräsentiert somit nicht mehr eine Bewegung sondern eine Aktion (siehe auch Abschnitt 2.3.2):

$$\begin{aligned} &X_T = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\} \\ &\mathbf{m}^{(\mu)}_{\text{action}} = (\text{Precondition}, X_T, C_T, \text{Effect}). \end{aligned} \quad (4.43)$$

Das Aktionsmodell enthält neben den Daten für den CTR auch statische Informationen über den symbolischen Kontext der Aktion. In einer ähnlichen Weise, wie im Vektor des Partikels $\mathbf{s}_t$ die Parameter (ϕ, α, ρ) die Trajektorie des Modells X_T mit der beobachteten Bewegung verbinden, muss auch der räumliche Kontext C_T mit den beobachteten symbolischen Daten verbunden werden. Zum Beispiel muss das generische Modell des „Tasse greifens“ mit einer konkreten Instanz eines Objektes in der Szene verbunden werden. Um den Typ c_{type} des Kontextes $\mathbf{x}_t$ mit einem spezifischen Objekt O_l der Szene Θ_t zu assoziieren, wird nun eine symbolische Objektreferenz ψ_t eingeführt, die den Bezeichner ID_t des Objektes speichert. Diese Referenz wird dem Partikelvektor $\mathbf{s}_t$ hinzugefügt:

$$\mathbf{s}_t = (\nu_t, \mu_t, \phi_t, \alpha_t, \rho_t, \psi_t) \text{ mit } \psi_t = (\text{ID}_t, \text{HS}_t). \quad (4.44)$$

Darüber hinaus enthält die symbolische Objektreferenz ψ_t auch einen für das Partikel spezifischen Handstatus (HS), der ähnlich definiert ist wie der GHS (siehe Gleichung 4.40). Initialisiert wird der Bezeichner ψ_t, wenn das erste Mal ein Objekt in dem Objektkontext C_T des Modells erwartet wird und ein passendes Objekt der Szene in dem definierten Suchbereich liegt. Durch diese Bindung eines Partikels und seines Modells an ein spezifisches Objekt kann sichergestellt werden, dass die Verbindung zwischen dem Partikel und dem Objekt auch in den nächsten Zeitschritten bestehen bleibt und kontrolliert werden kann.

Um den erweiterten Partikelvektor für die Aktionserkennung nutzen zu können, müssen die Schritte Selektion, Prädiktion und Aktualisierung des CTR angepasst werden. Zum einen wird der **situative Kontext** für die Selektion und Initialisierung eines Partikels verwendet, um nur solche Partikel auszuwählen oder zu selektieren, deren Vorbedingung mit der aktuellen Situation, also dem GHS vereinbar ist. Ebenso wird der GHS entsprechend dem Effekt einer erkannten Aktion verändert. Zum anderen wird der Suchbereich

beim Aktualisieren verwendet, indem das Gewicht $\pi_t^{(i)}$ eines Partikels $s_t^{(i)}$ verändert wird, je nachdem ob der Objektkontext des Modells durch ein Objekt der Szene erfüllt ist.

Zu diesem Zweck wird die Berechnung des Partikelgewichtes aus Gleichung 4.25 um einen multiplikativen *Kontextfaktor* P_{symb} erweitert, der repräsentiert, wie gut die beobachtete Szene Θ_t dem erwarteten Kontext entspricht:

$$\pi_t^{(i)} \propto p(\mathbf{z}_t|\mathbf{s}_t^{(i)})\ P_{symb}(\Theta_t, \mathbf{s}_t^{(i)}). \tag{4.45}$$

Ein einfacher, aber wirksamer Ansatz ist es, einen konstanten Faktor P_{symb} zu wählen. Der ist aber abhängig davon, ob ein erwartetes Objekt im Suchbereich vorhanden ist oder nicht. Ist eine Geste erkannt (siehe Gleichung 4.27), werden die symbolischen Daten aller zur Endwahrscheinlichkeit $P_{\text{end},t}(\mu)$ beitragenden Partikel untersucht, um eine objektspezifische Endwahrscheinlichkeit $P_{Obj,t}$ zu erhalten. Für jedes Objekt O_l werden die Gewichte der Partikel, die zu dem erkannten Modell μ_i gehören und das jeweilige Objekt enthalten, aufsummiert:

$$P_{Obj,t}(O_l, \mu_i) = \sum_{n=1}^{N} \begin{cases} \pi_t^{(n)} & \text{if } O_l \in \psi_t^{(n)} \wedge \mu_i \in \mathbf{s}_t^{(n)} \wedge\ \phi > 0.9\phi_{\max} \\ 0 & \text{sonst.} \end{cases} \tag{4.46}$$

Das Objekt mit der höchsten Wahrscheinlichkeit $P_{Obj,t}(O_l, \mu_i)$ wird als das manipulierte ausgewählt. Somit ist es möglich, manipulierte Objekte sowohl anhand ihres Typs und des Typs der Geste als auch der räumlichen und zeitlichen Relation zwischen Objekt und Geste auszuwählen.

4.4.1 Ergebnisse

An der Erkennung von Zeigegesten im Kontext von Objekten wird die Integration des Objektkontextes in die Gestenerkennung evaluiert. Die Ergebnisse sind von Hofemann u. a. [2004] veröffentlicht worden. In dem Versuch führen fünf Personen Zeigegesten auf Objekte aus. Es liegen 14 Videosequenzen mit insgesamt 84 Zeigegesten vor. Die beobachtete Person steht circa 2m vor der Kamera, so dass der Oberkörper und die bewegte Hand im Sichtbereich der Kamera sind. Mit der rechten Hand zeigt die Versuchperson auf sieben Objekte, jeweils drei auf der rechten und linken Seite, sowie ein Objekt vor ihr. Da in dem Experiment nur die Erkennung von Zeigegesten evaluiert werden soll, wird eine perfekte Objekterkennung angenommen. Die Bilder liegen mit einer Frequenz von 15 Bildern pro Sekunde bei einer Auflösung von 320x240 Pixel vor. Unter diesen Bedingungen ist eine Echtzeiterkennung auf einem Bürocomputer möglich. Die Hände werden auf Grund ihrer Hautfarbe mit der von Fritsch [2003] in seiner Arbeit im Kapitel 4.1 vorgestellten Methode verfolgt. Die Bewegung wird als Geschwindigkeit der Hand im Bild und ihrer Richtungsänderung repräsentiert $Z_{Vel-Ang}$ (siehe Gleichung 3.2, Seite 39 und Abbildung 3.3(b), Seite 39). Ein exemplarisches Bewegungsmodell mit einem Objektkontext und ein Beispielsbild aus dem Versuch sind im Anhang in Tabelle 8.1 und Abbildung 8.1 zu sehen.

In der Evaluation, deren Ergebnisse in Tabelle 4.1 dargestellt sind, werden unterschiedliche Parametrisierungen der Gestenerkennung getestet. Die Parameter des CTR Algorithmus

sind auf N=1000 Partikel gesetzt mit den Skalierungsfaktoren α und ρ zwischen 0.65 und 1.35 mit einer Standardabweichung von $\sigma = 0.15$.

Die Erkennungsleistung spiegelt sich in den korrekt erkannten Gesten (k), den nicht erkannten Gesten (Löschung l), den falschen Erkennungen (Vertauschung v) und den Einfügungen (e) von Gesten wieder. Aus diesen Werten lassen sich zwei Kennziffern berechnen. Das ist zum einen die Erkennungsrate

$$ER = \frac{k}{n} \tag{4.47}$$

und zum anderen die Fehlerrate

$$FR = \frac{l + v + e}{n}. \tag{4.48}$$

	Kontext								
	keine	Distanz	gerichtet	gewichtet					
$P_{missing}$	-	1.0	1.0	0.8	0.6	0.4	0.2	0.1	0.0
Korrekt (k=84)	83	69	74	72	75	77	76	78	82
Einfügungen (e)	81	9	5	5	5	5	6	5	18
Löschungen (l)	1	10	10	12	9	7	6	6	2
Vertauschung (v)	0	5	0	0	0	0	0	0	0
Fehlerrate in %	97.6	28.6	17.8	20.2	16.7	14.3	14.3	13,3	23.8
Erkennungsrate in %	98.8	82.2	88.1	85.7	89.3	91.7	90.4	92.8	97.6

Tabelle 4.1: Erkennung von Zeigegesten im Kontext von Objekten.

Die erste Spalte der Tabelle 4.1 zeigt die Ergebnisse mit dem unveränderten CTR von Black u. Jepson [1998]. Ohne das Einbinden des symbolischen Kontextes ist hier keine Unterscheidung zwischen einem Annähern und Entfernen der Hand von einem Objekt möglich. Jede Bewegung, die dem Bewegungsmodell eines Zeigens entspricht, wird als „Zeigen“ interpretiert. Entsprechend ist bei diesem Ansatz auch eine hohe Anzahl von Einfügungen aufgetreten. Dieses wirkt sich negativ auf die Fehlerrate aus. Des Weiteren steht bei einer erkannten Geste keine Information über das referenzierte Objekt zur Verfügung.

Nutzt man ein einfaches Distanzmaß (Spalte *Distanz* in 4.1), um eine Bindung der Hand an Objekte zu etablieren, werden hauptsächlich Zeigegesten erkannt. Aber es liegen immer noch viele Einfügungen vor und oft wird auch das falsche Objekt mit der Geste verbunden, das zu den beobachteten Ersetzungen führt. Mit dem gerichteten Kontextbereich (Spalte *gerichtet*) konnte sowohl eine geringere Fehlerrate als auch eine höhere Erkennung erreicht werden. Mit Verwendung der gewichteten Einbeziehung des Objektkontextes (Spalte *gewichtet*) konnte die Erkennungsleistung noch weiter gesteigert werden.

Wird jedoch der Kontextfaktor $P_{missing} = 0$ gesetzt, ist wohl die Erkennungsrate höher, aber auch die Fehlerrate steigt wieder an. Dieser Effekt liegt darin begründet, dass Partikel, deren Objektkontext nicht erfüllt ist, gelöscht werden. Indirekt werden Partikel, deren

Trajektorien nicht gut zu den Messwerten $Z(t)$ passen, die aber an ein Objekt gebunden wurden, propagiert. In diesem Fall wird der Fokus auf die Näherelation der Hand zu den Objekten der Umgebung gelegt, die Bewegung als Merkmal wird hingegen weniger beachtet.

Gerade dieses Ergebnis zeigt auf, wie wichtig die Fusion aus symbolischen und sensorischen Daten zur Erkennung von menschlichen Bewegungen im Kontext von Objekten ist. Erst ein Verfahren, dass beide Faktoren gleichzeitig beachtet, ermöglicht das robuste Erkennen von Aktionen, deren Bedeutungen im situativen und symbolischen Kontext verständlich werden.

Resumé und offene Fragen

Der im Rahmen dieser Dissertation weiterentwickelte CTR erweist sich als gutes Verfahren für die sequenzielle Gestenerkennung. Mit dem Algorithmus wird gleichzeitig eine Segmentierung und Klassifizierung der Bewegungsdaten vorgenommen. Die Daten müssen als Trajektorien vorliegen, die Bewegungen repräsentieren. Der entwickelte Algorithmus zur Gestenerkennung kann mit unterschiedlichen Repräsentationen und unterschiedlichen Dimensionalitäten der Trajektorien arbeiten. Ein weiterer Vorteil des vorgestellten Verfahrens ist, dass verhältnismäßig wenig Trainingsdaten vorliegen müssen, um gute Modelle zu generieren. Bereits aus einstelligen Beispielssätzen für jede Geste lassen sich gute Modelle bilden. Ein Erkennen von Gesten in Echtzeit auf normalen Bürocomputern ist gut möglich. Auch wenn viele Partikel ($n = 1000 \ldots 5000$) zu einer besseren Erkennungsleistung führen, lässt sich die Anforderung an die Rechenleistung über die Anzahl der verwendeten Partikel an die gegebenen Bedingungen anpassen. Ein Merkmal des Algorithmus ist, dass die Rechenleistung auf relevante Bereiche des Merkmalsraums konzentriert wird. Sobald die Beobachtung dem Modell einer Geste ähnelt, werden mehr Partikel für das Erkennen dieser Geste verwendet. Doch werden gleichzeit andere Hypothesen mit genügend anderen Partikeln verfolgt, so dass der Algorithmus innerhalb weniger Zeitschritte auf eine veränderte Beobachtung eingestellt werden kann.

Mit dem Objektkontext wurde im letzten Abschnitt eine zuverlässige Methode gezeigt, wie symbolische und sensorische Merkmale für das Erkennen von Gesten vereinigt werden können. Diese Kombination erlaubt nicht nur ein Klassifizieren der Bewegung einer Geste, sondern auch das Erkennen von einfachen Aktivitäten, die eine Verbindung der Bewegung zu den Objekten einschließt.

Die Anwendung der Gestenerkennung für die Interaktion zwischen Menschen und Computern wirft aber einige Fragen auf:

- Auch mit der Verwendung des Objektkontextes kann nicht erkannt werden, ob der Mensch ein Objekt berührt oder greift. Wie kann allgemein die Bedeutung einer Aktion erschlossen werden?

- Kann das zeitliche und räumliche Erkennen einer Bewegung und das Wissen über den Objektkontext dazu beitragen, unbekannte Objekte zu detektieren?

- In komplexen Szenen ist es schwierig, die menschliche Hand im Bild zu finden und zu verfolgen. Gibt es alternative Verfahren, die eine sichere und detailliertere Verfolgung der Hand oder des menschlichen Körpers erlauben? Können auch die bewegten Objekte für die Aktionserkennung verwendet werden?
- Können aus den vielen Handbewegungen eines Menschen neue Gesten und Aktionen gelernt werden? Kann aus der Bewegung erkannt werden, wann eine neue Handlung gezeigt wird?

Der Objektkontext ist ein Beitrag zum Erkennen der Bedeutung oder der Intention einer Handbewegung. Meistens muss aber ein größerer situativer Kontext beachtet werden und Vorwissen über typischen Aktivitäten bestehen, um die Intention des beobachteten Menschen erkennen zu können. Dass der Ansatz mit Kind- und Elternmodellen hierfür nicht ausreichend und zu statisch ist, wurde bereits erläutert. Aber auch die starre Anwendung des Objektkontextes kann zu ähnlichen Fehlinterpretationen führen. Die Lösung dieses Problems kann in einem probabilistischen Überbau für die Aktivitätserkennung bestehen. Ein solches System muss mit stochastischen Prozessen die Informationen über erkannte Bewegungen und Objekte verbinden und dynamische Hypothesen der ausgeführten Aktivität erstellen. Dieser Gedankengang führt zu der Kooperation mit dem Kollegen Zhe Li, der auch in der Gruppe Angewandte Informatik arbeitet. Aufbauend auf den hier beschriebenen Ansätzen wurde in dieser Kooperation ein hierarchisches Modell zur Erkennung von manipulativen Gesten entwickelt und vorstellt (siehe Li u. a. [2005b]). In diesem System, das die Theorie der Hierarchischen HMM verwendet, werden Bewegungen im Kontext von Objekten erkannt, aber parallel auf abstrakteren Ebenen Hypothesen für die aktuelle Aktivität verfolgt. Das Greifen einer Zuckerdose ist eine Bewegung, die sowohl Teil der Vorbereitung eines Tees oder Kaffees sein kann. Diese Hypothesen müssen folglich beide verfolgt werden, bis weitere Informationen die eine Aktivität bestätigen und die andere unwahrscheinlich erscheinen lassen.

Eine Antwort auf die zweite Frage wird im nächsten Kapitel (5.2) mit einem System zur Auflösung von Objektreferenzen und dem Lernen von Ansichten der referenzierten Objekte gegeben. Anhand eines Systems wird gezeigt, wie die Gestenerkennung ein Beitrag zur multimodalen Interaktion sein kann, denn in dem Modul zur Objektaufmerksamkeit vereinigt Haasch u. a. [2005] sowohl sprachliche als auch deiktische Informationen.

Der Objektkontext einer Handlung lässt sich auch anders als bisher erläutert interpretieren, denn es kann nicht nur die Hand, die ein Objekt bewegt, zur Aktionserkennung verfolgt werden, sondern auch die bewegten Objekte selbst. Werden die bewegten Objekte verfolgt, erübrigt sich auch die Frage, ob und wann welches Objekt manipuliert wird. Dieser Ansatz, der auch im nächsten Kapitel (5.3) vorgestellt wird, zeigt somit eine Alternative zur Verfolgung der menschlichen Hände auf. Eine weitere Alternative für die Verfolgung der Hände wird mit einer Körperverfolgung gegeben. Wird ein solches Verfahren genutzt, steht die Bewegung der Hand außerdem im Kontext des menschlichen Körpers. Mehrdeutigkeiten, die sich bei der Betrachtung der Handbewegung ergeben, werden durch die Körperverfolgung gelöst.

Die letzte Frage betrifft das Erlernen neuer Bewegungen und das Erkennen, wann eine Bewegung erlernt werden soll. Um Antworten auf diese Frage zu finden, wird im Kapi-

tel 6 ein Blick auf die frühkindliche Entwicklung geworfen. Es wird überlegt, warum es Kleinkindern möglich ist, relevante Bewegungen zu erkennen und zu lernen.

5. Multimodale Systeme

Im vorherigen Kapitel wurde ein probabilistisches Verfahren zur Gesten- und Handlungserkennung vorgestellt. Das Verfahren erlaubt es, computergestützt Bewegungen und Aktionen, die ein Mensch mit einer Hand ausführt, zu klassifizieren und den genauen Zeitpunkt des Abschlusses einer Geste zu detektieren. Es wurde auch gezeigt, dass Gesten und Manipulationen nicht losgelöst auftreten, sondern in den situativen und symbolischen Kontext der Szene eingebunden sind. Nur unter Beachtung dss Kontextes können Gesten und ihre Bedeutung sicher erkannt werden. Des Weiteren ist auch aus Kapitel 2 bekannt, dass Gesten meistens von Menschen in einer multimodalen Interaktion verwendet werden.

Diese Beobachtungen sind die Motivation dafür, Gesten als weitere Modalität in die Interaktion zwischen Menschen und Computern zu integrieren. Die Gestenerkennung wird zum einen in das System des mobilen, multimodalen Roboter BIRON eingebunden (5.2). Zum anderen findet sie Verwendung im Rahmen eines virtuellen Assistenzsystems (5.3). Das System des Roboters BIRON und weitere Projekte, die sich mit sozial interagierenden Robotern beschäftigen, werden vorher (5.1) vorgestellt.

Bei der Integration einer Komponente, wie es die Gestenerkennung ist, in ein komplexes System stellen sich unterschiedliche Herausforderungen. Einige davon liegen in der Natur der Systemintegration und Kommunikation zwischen unabhängigen und nicht synchron entwickelten Modulen, die aber zur Laufzeit in einem integrierten System arbeiten und Daten austauschen müssen. Für die Systeme konnte auf Arbeiten der Arbeitsgruppe Angewandte Informatik an der Universität Bielefeld zurückgegriffen werden, die den Bau komplexer Systeme ermöglichen und unterstützen. Auf der Seite der Bildverarbeitung und Mustererkennung ist das die komfortable Plattform *IceWing*, die von Lömker u. a. [2006] entwickelt wurde. Für den Austausch, das Speichern und asynchrone Wiedergeben von Daten über verschiedene Computer hinweg wird bei den Entwicklungen der in diesem Kapitel (5.2 und 5.3) vorgestellten Systeme das *XML enabled Communication Framework* (XCF) von Wrede u. a. [2006b] verwendet.

5.1 Interaktion mit einem Roboter

In diesem Unterkapitel wird ein Überblick über Robotersysteme gegeben, die eine multimodale Interaktion verwirklichen. Detaillierter werden der mobile Roboter BIRON und seine Komponenten zur multimodalen MRK beschrieben.

Das Ziel einer menschenähnlichen Interaktion zwischen Menschen und sozial agierenden Robotern schließt das Erkennen der Aktionen des Menschen ein. Es ist aber nur ein Teilbeitrag zu der angestrebten multimodalen Interaktion. Das Erkennen von Gesten muss deswegen in ein System zur multimodalen Interaktion eingebettet sein. Erst Robotersysteme, in denen unterschiedliche Komponenten zusammenwirken, können eine multimodale Interaktion erlauben.

Ein Anwendungsszenario für diese Roboter ist der persönliche Serviceroboter, der in einer Wohnung situiert ist und agiert. Der Roboter ist ein Begleiter des Menschen, kennt sich in der Wohnung aus und kann im Gespräch instruiert werden. Der Roboter muss also Personen erkennen und Dialoge führen können, sowie in der Wohnung navigieren können. Der *common ground* des Dialogs sollte nicht nur auf die Sprache beschränkt bleiben, sondern mit der Umgebung und deren Objekten verbunden sein. Deswegen ist es sinnvoll, wenn es dem Roboter möglich ist, Referenzen auf Gegenstände aufzulösen und ihre verbale Beschreibung zu lernen. Hier kann das Erkennen von Zeigegesten das Finden referenzierter Gegenstände ermöglichen und einen Beitrag zur Multimodalität leisten. Aber nicht nur Deiktika, auch andere Gesten tragen zur menschenähnlichen Interaktion bei.

Die Komponenten der multimodalen Interaktion müssen auf einem Roboter gleichzeitig, kooperativ und in Echtzeit laufen, damit eine dem Benutzer angemessene Interaktion möglich ist. Einem solchen kooperativen und situierten Roboter kann sein neuer Besitzer nach dem Erwerb seine Wohnung zeigen und erläutern. Das so gewonnene Wissen dient dem Roboter später als Grundlage für die Interaktion mit Personen in der Wohnung. Es sind somit auch die Anwendungsfälle denkbar, dass der Roboter neue Gäste in einer Wohngemeinschaft begrüßt, als Makler agiert oder körperlich beeinträchtigten Menschen hilft. Weiterhin könnte der Roboter auch Handwerker in die Wohnung lassen und ihnen zeigen, was repariert werden muss. Für viele dieser Anwendungsfälle ist es unter anderem wünschenswert, dass der Roboter mit entsprechenden Aktoren Aufgaben verrichtet und Dinge für den Besitzer holen oder wegbringen kann. Aber auch die Rolle als Mittler, der spezialisierten Robotern Anweisungen geben kann, ist möglich.

5.1.1 Multimodale Roboter

Die meisten bisher entwickelten Roboter, die eine menschenähnliche, multimodale Interaktion anstreben, sind Forschungsprojekte. Doch gibt es zum Beispiel mit dem *Wakamaru* der japanischen Firma Mitsubishi auch kommerzielle Produkte (siehe Abbildung 5.1(a)). Auch der laufenden Roboter *Asimo* von Honda und der *Partner Robot* von Toyota werden industriell entwickelt und produziert (siehe Abbildungen 5.1(b) und 5.1(c)). Sie können eine Plattform für Module stellen, die eine natürliche Interaktion zwischen Menschen und Robotern erlauben. Entwicklungen der Universitäten und Forschungseinrichtungen konzentrieren sich hingegen häufiger auf die einzelnen Modalitäten und Interaktionsformen.

(a) Wakamaru[1]

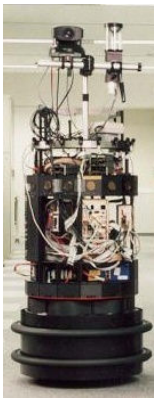
(b) Asimo[2]

(c) Partner Robot[3]

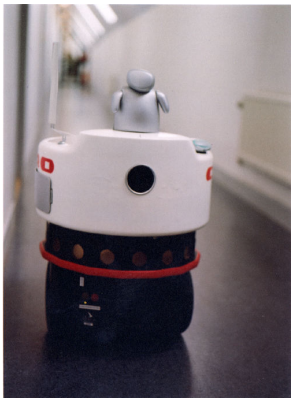
(d) PeopleBot[4]

(e) Lino[5]

(f) BIRON



(g) Jijo 2[6]



(h) CERO[7]

Abbildung 5.1: Es werden verschiedene mobile Roboter gezeigt, mit denen die Möglichkeiten der Interaktion zwischen Menschen und Robotern erforscht werden.

Hierbei sind die Roboter teilweise Eigenkonstruktion, zum Teil wird aber auch auf Roboterplattformen wie dem *Pioneer PeopleBot* der Firma ActiveMedia aufgebaut (siehe Abbildung 5.1(d), 5.1(f) und 5.2(d)). Einen guten Überblick über unterschiedliche Roboter und ihre Anwendungen geben Fong u. a. [2003].

Für die sozial interagierenden Roboter gibt es unterschiedliche Anwendungsszenarien. Neben dem Einsatz im Haushalt (zum Beispiel *Wakamaru, Lino, BIRON*) oder im Büro (zum Beispiel *Jijo-2*) ist auch die Anwendung in der Unterstützung älterer und pflegebedürftiger Menschen ein Thema, wie mit den Roboter *CERO* und *Care-O-Bot II* gezeigt wird (siehe Abbildungen 5.1 und 5.2).

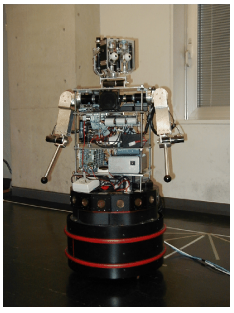
[1]Wakamaru von Mitsubishi™; Bildquelle: http://www.mhi.co.jp/kobe/wakamaru/
[2]Asimo von Honda™; Bildquelle: http://www.dld-conference.com/2005/12/
[3]Partner Robot[3] von Toyota™; Bildquelle: http://www.toyota.co.jp/en/special/robot/
[4]PeopleBot[4] von ActiveMedia™; Bildquelle: http://www.activrobots.com/
[5]Lino, aus Kröse u. a. [2003]
[6]Jijo 2, siehe Matsui u. a. [1998]; Bildquelle: http://staff.aist.go.jp/h.asoh/jijo2/
[7]CERO, aus Hüttenrauch u. Eklundh [2002]

(a) Care-O-Bot II[8]

(b) Armar[9]

(c) HERMES[10]

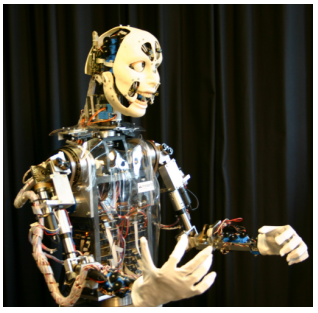
(d) Horos[11]



(e) ROBITA[12]

(f) ROBOVIE[13]

(g) iCat[14]



(h) Barthoc[15]

Abbildung 5.2: Zu sehen sind weiterer mobile und stationäre Roboter, mit denen die Möglichkeiten der Interaktion zwischen Menschen und Robotern erforscht werden.

Eine Übersicht über einige Robotersysteme gibt die Tabelle 8.2 im Anhang. Hier werden tabellarisch die Artikel, in denen die Roboter beschrieben werden, und einige Fähigkeiten der Roboter aufgelistet. In allen Anwendungsgebieten finden sich aber ähnliche Anforderungen an die interaktiven Kapazitäten und sonstigen Fähigkeiten des Roboters.

5.1.2 Modalitäten und Fähigkeiten von Robotern

Die Modalitäten und Fähigkeiten von Robotern, deren Ziel eine soziale Interaktion mit Menschen ist, werden nun beschrieben und an Beispielen einiger Implementationen konkretisiert.

[8] Care-O-Bot; Bildquelle: http://www.care-o-bot.de; siehe Graf u. a. [2004]
[9] Armar, aus Asfour u. a. [2001]
[10] HERMES, aus Bischoff u. Graefe [2002]
[11] Horos, siehe Richarz u. a. [2006]
[12] Robita; Bildquelle: http://tosa.mri.co.jp/sounddb/robot; siehe Matsusaka u. a. [2003]
[13] ROBOVIE, aus Ishiguro u. a. [2001]
[14] iCat, siehe van Breemen [2004]
[15] Barthoc, siehe Hackel u. a. [2005]

Personenaufmerksamkeit

Die erste Aufgabe, die ein sozial interagierender Roboter bewältigen muss, ist das Auffinden von möglichen Interaktionspartern. Spricht ein Mensch einen Roboter an, ist der Anspruch an den Roboter, dass er sich seinem Gesprächspartner zuwenden muss. Für die natürliche Interaktion ist diese Fähigkeiten wichtig, da dem Menschen so auf sehr intuitive Art vermittelt werden kann, dass der Roboter ihn als Gesprächspartner erkannt hat. Technisch kann das mit einer Gesichtserkennung oder der Lokalisation von Geräuschquellen gelöst werden. Mit der Fähigkeit potentielle Gesprächspartner zu erkennen und diese zu verfolgen, sind die meisten Roboter ausgestattet. Matsui u. a. [1998] stellen mit dem Roboter *Jijo-2* zum Beispiel einen Roboter vor, der mit einem Mikrofonfeld die Richtung ausmachen kann, aus der ein „Hallo“ gesprochen wird. Der Roboter richtet sich daraufhin in die Richtung aus und versucht eine hautfarbene Region zu finden. Diese Region beinhaltet vermutlich das Gesicht und wird im Bild der dreh- und schwenkbaren Kamera zentriert, daraufhin wird versucht, ein Gesicht im Kamerabild zu finden. Da das Verfahren Trainingsbilder verwendet, kann über das ähnlichste Trainingsbild auch eine Hypothese für die erkannte Person aufgestellt werden.

Spracherkennung und Dialog

Ein weiterer Schritt zur menschenähnlichen Interaktion ist das Führen eines Dialoges. Bei den meisten Systemen ist es möglich, einen natürlichsprachlichen Dialog mit dem Roboter zu führen. Die Fähigkeiten und die Natürlichkeit dieser Dialogsysteme zu vergleichen fällt hingegen schwer. Oft sind die Systeme auf einfache Befehle oder ein begrenztes Lexikon beschränkt. Probleme bestehen auch in der Erkennung der Sprache, denn Nahbesprechungsmikrofone sind unerwünscht, da sie eine Vorbereitung für die Interaktion bedingen. Des Weiteren können die Eigengeräusche oder Geräusche der Umgebung die Erkennungsleistung mindern. Die Spracherkennung wird meistens mit kommerziellen Programmen verwirklicht, für die Modellierung der Dialoge werden oft endliche Automaten verwendet.

Agieren und Manipulieren

Da mobile Systeme in der Umgebung des Menschen situiert sind und in dieser agieren, ist nicht nur eine Lokalisation der Gesprächspartner sinnvoll, sondern auch die Selbstlokalisation des Roboters im Raum. Bespiele für mobile Roboter sind *Amila, Armar, BIRON, Care-O-Bot, CERO, HERMES, Horos, Jijo 2, Lino, ROBITA* und *Robovie* (siehe Abbildungen 5.1 und 5.2). Damit ein Roboter zum Beispiel auf Anweisung in einen Raum fahren oder ein Objekt holen kann, muss der Roboter ein Weltbild von seiner Umgebung haben, in der er sich befindet. Für die Aufgaben werden oft Laserscanner verwendet, die ein Tiefenbild in einer bestimmten Höhe über dem Boden aufnehmen und somit eine Hypothese für die Wände und Objekte in der Umgebung erlauben. In einer so erstellten Karte kann der Roboter seine Position bestimmen und zu bestimmten Positionen navigieren.

Aber auch Kamerasysteme können verwendet werden, um bildbasiert die Räume einer Wohnung zu unterscheiden. Spexard u. a. [2006] stellen zum Beispiel ein System vor, das es dem Roboter BIRON ermöglicht, Fragen nach seiner Position zu beantworten. Die Räume werden mit einer omnidirektionalen Kamera aufgenommen und anschließend klassifiziert.

Für viele mobile Robotersysteme, die im Haushalt oder in einem Pflegeszenario eingesetzt werden sollen, existiert die Anforderung, dass der Roboter Dinge holen und transportieren können soll. Hierfür benötigt der Roboter Aktuatoren, die es ermöglichen Objekte zu greifen und zu transportieren. Das Detektieren und Greifen von Objekten ist eine komplexe Aufgabe, die in vielen Forschungsprojekten nicht aufgegriffen wird. Folglich werden an vielen Robotern keine Aktuatoren montiert, die zudem auch die Konstruktion des Roboters und die Energieversorgung schwieriger gestalten. Unter den Forschungsentwicklungen haben die Roboter *Armar, Care-O-Bot* und *HERMES* Aktuatoren (siehe Abbildung 5.2). Die Arme der Roboter *Robita* und *Robovie* haben keine Endeffektoren, die ein Greifen ermöglichen. Sie werden zum Generieren von Gesten und Armbewegungen verwendet.

Persönlichkeit

Ein weiterer Themenbereich, der für die menschenähnliche Interaktion notwendig ist, ist das personenspezifische und emotionale Verhalten der Roboter. Einige Roboter können bereits Personen identifizieren. Hierfür wird meistens eine videobasierte Gesichtserkennung verwendet. Der Roboter Wakamaru (Abbildung 5.1(a)) wird zum Beispiel damit beworben, dass er bis zu zehn Personen erkennen kann und hiervon zwei als seine Besitzer registriert werden können. Bei dem Roboter *Lino* (Abbildung 5.1(e)) haben Kröse u. a. [2003] einen alternativen Weg der Identifizierung gewählt. Mit generalisierten HMM werden die Sprachsignale klassifiziert und so unterschiedlichen Benutzern zugeordnet. Der Roboter fungiert als personifizierter Charakter für ein intelligentes Haus, das über ihn gesteuert werden kann. Kröse u. a. [2003] versuchen deswegen einen Roboter zu entwickeln, der sowohl ein intelligentes Verhalten zeigt, als auch eine situierte natürliche Interaktion erlaubt. Eine Besonderheit des Roboters Lino ist auch, dass er ein mechanisches Gesicht hat, das emotionale Mimiken produzieren kann. Mit der Entwicklung der *iCat* hat van Breemen [2004] diese Idee weiter verfolgt. Die iCat ist aber kein mobiler Roboter, sondern ein kleiner statischer, der als Schnittstelle zwischen Menschen und Haushaltsgeräten konzipiert ist (siehe Abbildung 5.2(g)).

Ein anderer Weg der Emotionsproduktion wird mit dem humanoiden Oberkörper der *Barthoc* Roboter verfolgt (siehe Abbildung 5.2(h)). Die von Hackel u. a. [2005] vorgestellte Plattform, beinhaltet ein menschenähnliches Gesicht, mit dem sich Mimiken produzieren lassen. In einer Studie von Hegel u. a. [2006] wird die Mimikproduktion verwendet, des Weiteren zeigt die Studie auch, dass das Erkennen menschlicher Emotionen aus dem Sprachsignal möglich ist.

Gestenerkennung für mobile Roboter

Eine multimodale Interaktion unter Einbeziehung einer Gestenerkennung wird für wenige Roboter beschrieben. Auf dem Roboter Armar arbeitet die Gestenerkennung von Nickel u. Stiefelhagen [2003a]. Richarz u. a. [2006] stellen eine Zeigeerkennung für den Roboter Horos vor.

Eine einfache Aktionserkennung, die auf eine Anwendung spezialisiert ist, wurde auch für den *Care-O-bot II* entwickelt. Der in Abbildung 5.2(a) dargestellte Roboter kann Menschen mit seinem Arm eine Visitenkarte anbieten. Mit einem Bewegungssensor wird erkannt, ob die vor dem Roboter stehende Person die Karte vermutlich ergriffen hat.

Hypothesen für Personen in der Umgebung des Roboters werden aus den Tiefendaten eines Laserscanners gewonnen. Der Roboter wurde für die optische Detektion und Manipulation von Objekten im Haushalt entwickelt. Über eine graphische Bedieneinheit oder Sprachbefehle kann der Roboter zum Beispiel angewiesen werden, Objekte zu holen oder aufzuräumen. Die Vision der Anwendung im Pflegebereich wird deutlich, da der Roboter auch als Gehhilfe und Gehilfe dienen kann. Der *Care-O-Bot II* wird unter anderem von Schraft u. a. [2004] beschreiben.

Die Systemarchitektur des Roboters *Armar* (siehe Abbildung 5.2(b)) und das Generieren von Bewegungen wird von Asfour u. a. [2001] beschrieben. Interessanter für die Interaktion zwischen Menschen und Robotern ist die Integration der Ergebnisse der Gestenerkennung von Nickel u. Stiefelhagen [2003a] in einen Dialog. Die von Holzapfel u. a. [2004] entwickelte Fusion von sprachlichen und gestischen Informationen benutzt die Sprache als Hauptmodalität, die Zeigegeste wird zum Disambiguieren der Objektreferenzen eingesetzt. Wie bereits in 3.3.2 beschrieben, setzen Nickel und Stiefelhagen eine Stereokamera ein und ermitteln die Position der zeigenden Hand und die Zeigerichtung im Raum. Über eine regelbasierte Grammatik werden Gesten, die eine Referenz auf ein Objekt geben, mit den sprachlichen Äußerungen vereint.

Mit der Entwicklung des Roboters *Horos* wird ein Einkaufsassistent angestrebt. Der Roboter soll folglich kostengünstig, aber auch einfach zu bedienen sein. Der auf dem Pioneer PeopleBot aufbauende Roboter Horos und hat eine Spracherkennung und Sprachgenerierung. Für den Roboter Horos haben Richarz u. a. [2006] eine einfache Methode der Zeigegestendetektion und Richtungsauflösung entwickelt. Für das Verfahren ist bereits eine kostengünstige Kamera ausreichend. Die Autoren erläutern, dass die Richtung einer auf den Boden gerichteten Geste mit ihrem Verfahren genauer bestimmt werden kann, als es Menschen unter denselben Umständen möglich ist.

Ein Mensch, der vor dem Roboter Horos steht, wird über einen Schulterdetektor detektiert. Die Bildregion, die den Oberkörper und die Arme des Menschen umfasst, wird skaliert. Mit einem Gabor-Filter werden hieraus Merkmale extrahiert. In einer Kaskade von MLP Klassifikatoren werden aus diesen die Distanz und der Winkel des referenzierten Punktes auf dem Boden relativ zum Menschen ermittelt.

Leider arbeitet das Verfahren auf den Einzelbildern der Videosequenz, die die ruhige Hand beim Zeigen enthalten. Über einen sprachlichen Befehl wird die Erkennung der Zeigerichtung angestoßen. Im eigentlichen Sinn wird folglich nicht erkannt, ob und wann eine Zeigegeste erfolgt, sondern die Richtung wird ermittelt, wenn eine Zeigegeste sprachlich angedeutet wurde. Die Autoren hoffen, die Genauigkeit zu verbessern, wenn Bewegungsinformationen hinzugenommen werden.

Auch die von Ghidary u. a. [2002] vorgestellte Methode, mit der Gesten und Objektreferenzen erkannt werden können, detektiert nur statische Stellungen der Hand. Mit einem sprachlichen Befehl wird ein Prozess angestoßen, in dem zuerst hautfarbene Regionen im Bild der Kamera gesucht werden. Mit einem ansichtsbasierten Verfahren werden anschließend in den größeren dieser Regionen Gesichter und Hände gesucht. Dass nur Hände im Bild gefunden werden können, die mit der offenen Handfläche zur Kamera zeigen, ist eine Einschränkung der Interaktion. Da die Autoren eine dreidimensionale Karte der

Umgebung des Roboters aufbauen wollen, müssen sie die Entfernung des Benutzers ermitteln, der auf ein Objekt zeigt. Hierfür verwenden sie den Autofokus einer der dreh- und schwenkbaren Kamera. Die Objekte werden als rechtwinklige Bereiche repräsentiert, eine Detektion von Objektgrenzen oder eine Objekterkennung haben die Autoren angedacht, aber noch nicht implementiert. Die Methode der Tiefenmessung ist ein weiterer Nachteil, da die Hand für die Dauer der Tiefenmessung nicht bewegt werden sollte.

Das Erkennen von Gesten zielt in dieser Arbeit und in den bisher vorgestellten Robotersystemen auf die Perzeption von Gesten. Einen interessanten Ansatz zur Generierung von Gesten wird mit dem Robotersystem *HERMES* vorgestellt (siehe Abbildung 5.2(c)). Der Roboter besitzt zwei Arme mit je sechs Freiheitsgraden, ist mit einem Stereovideosystem ausgestattet und hat taktile Sensoren. Die Kommunikation mit dem System kann über gesprochene und geschriebene Sprache erfolgen. Bischoff u. Graefe [2004] unterstreichen, dass die Dialogführung und das Sprachverstehen über einen situativen Kontext verbessert wird. Der Kontext wird aus vorherigen Äußerungen gewonnen und erlaubt die Antworten, die der Roboter vom Menschen erwarten kann, einzugrenzen. Dieser Roboter kann bisher keine Gesten von Menschen wahrnehmen, nutzt selber aber seine Arme um sprachbegleitend Gesten auszuführen. Unter anderem kann der Roboter Hermes winken, um die Aufmerksamkeit von Menschen auf sich zu lenken.

Eine natürliche Interaktion zwischen Robotern und Menschen erfordert viele der bisher aufgezählten Fähigkeiten, die aber in einem multimodalen System zusammengefasst sein sollten. Das System kann dann eine menschenähnlichere und für den Interaktionspartner angenehmere Kommunikation erlauben als die einzelnen Modalitäten.

Benutzerstudien

Wie Menschen auf Roboter reagieren, was sie von diesen erwarten und wie die Fähigkeiten der Roboter kommuniziert werden können, ist auch ein wichtiges Thema der MRK. Leider mangelt es in den Studien aber an Robotern, die bereits eine robuste multimodale Interaktion erlauben.

Der Gebrauch von mobilen und sozialen Robotern im persönlichen Umfeld ist nicht üblich und erfordert eine Erforschung der Ansprüche des Menschen an diese Art von Robotern. Entsprechend wird in einigen Forschungsprojekten die Interaktion zwischen Menschen und Robotern beobachtet und versucht, Erkenntnisse für ein angemessenes Verhalten und Design zukünftiger Roboter zu gewinnen. Der Roboter *CERO*, (siehe Abbildung 5.1(h)) den Hüttenrauch u. Eklundh [2002] vorstellen, ist für einfache Hol- und Tragefunktionen entwickelt worden. Der Roboter hat keinen Aktuator, sondern nur eine Ablagefläche. Gesteuert wird er über einen PDA[5] oder im Dialog mit sprachlichen Anweisungen. Zielrichtung der Entwicklung ist hier, genauso wie bei dem *Care-O-Bot II*, die Unterstützung körperlich beeinträchtigter Benutzer.

5.1.3 Die Integrationsplattform BIRON

Die Übersicht über die verschiedenen Robotersysteme zeigt auf, wie umfangreich das Gebiet der Interaktionsmodalitäten eines sozial interagierenden Roboters sein muss. Es er-

[5]PDA: Ein Persönlicher Digitaler Assistent ist ein kleiner, etwa handgroßer Computer.

geben sich viele Herausforderungen, von der Detektion von Menschen über die Dialogsteuerung und Navigation bis zur Gesten- und Objekterkennung. Das Zusammenspiel vieler spezialisierter Module muss in einem Robotersystem koordiniert und getestet werden. Die Evaluation schließt hierbei nicht nur den Test einzelner Komponenten oder des Gesamtsystems ein, sondern auch die Benutzerfreundlichkeit des Roboters. Für den Roboter BIRON (siehe Abbildung 5.3) werden Module für unterschiedliche Modalitäten und Anforderungen entwickelt, integriert und evaluiert.

Der **Bi**elefeld **R**obot Compani**on** (BIRON) wurde in der Arbeitsgruppe für Angewandte Informatik (Universität Bielefeld) entwickelt. Als Basis dient ein Modul des *Pioneer PeopleBot* der Firma ActiveMedia. In den Artikel von Haasch u. a. [2004], Fritsch u. a. [2005b] und Wrede u. a. [2004a] wird der Roboter detailliert beschrieben. An dieser Stelle werden die Hauptkomponenten und die Leistungsfähigkeit des Roboters vorgestellt. Das Zusammenspiel der Komponenten erlaubt eine situierte, multimodale Interaktion mit dem Roboter. Die Entwicklungsarbeiten, die im Rahmen dieser Dissertation ausgeführt wurden, ermöglichen auch die gestische Interaktion mit dem Roboter.

Abbildung 5.3: Der mobile Roboter BIRON .

Die Softwarearchitektur des Roboters ist schematisch in Graphik 5.4 aufgezeichnet. Ausführlicher ist diese Drei-Ebenen-Architektur von Kleinehagenbrock [2005] beschrieben und wird hier kurz vorgestellt, um eine Einordnung der für die Gestenerkennung nötigen Module zu erlauben.

Die in der deliberativen Ebene angeordnete Dialogsteuerung von Li u. a. [2005a] erlaubt es, Instruktionen des Menschen zu verstehen und Interaktionsprobleme wie zum Beispiel Mehrdeutigkeiten aufzulösen. Hierfür stellt der Roboter Nachfragen und zeigt so ein kooperatives Verhalten. Eingaben erhält die Dialogsteuerung von der Sprachverstehenskomponente, die ebenfalls in der deliberativen Ebene angeordnet ist und auf Informationen unterer Ebenen, unter anderem des Szenemodells, zurückgreifen kann (siehe Hüwel u. Wrede [2006]). Zum Beispiel kann das Objekt, das mit einer Zeigegeste referenziert wird, in die Aussage „Die Tasse dort“ eingebunden werden. Eine Anforderung an das Sprachverstehen ist der erfolgreiche Umgang mit unvollständigen, sowie grammatikalisch inkorrekten Sätzen der Alltagssprache. Der *Planer*, als weitere Komponente in der Deliberativen Ebene verortet, wird bisher nur für Navigationsaufgaben verwendet, kann aber um weitere, komplexere Fähigkeiten erweitert werden.

Die Ablaufsteuerung steht im engen Zusammenspiel mit dem Planer und erlaubt ein Zerlegen von komplexen Plänen in Einzelschritte. Ebenso wie der Planer befindet sich diese Komponente noch in der Entwicklung und erlangt eine stärkere Bedeutung, wenn der

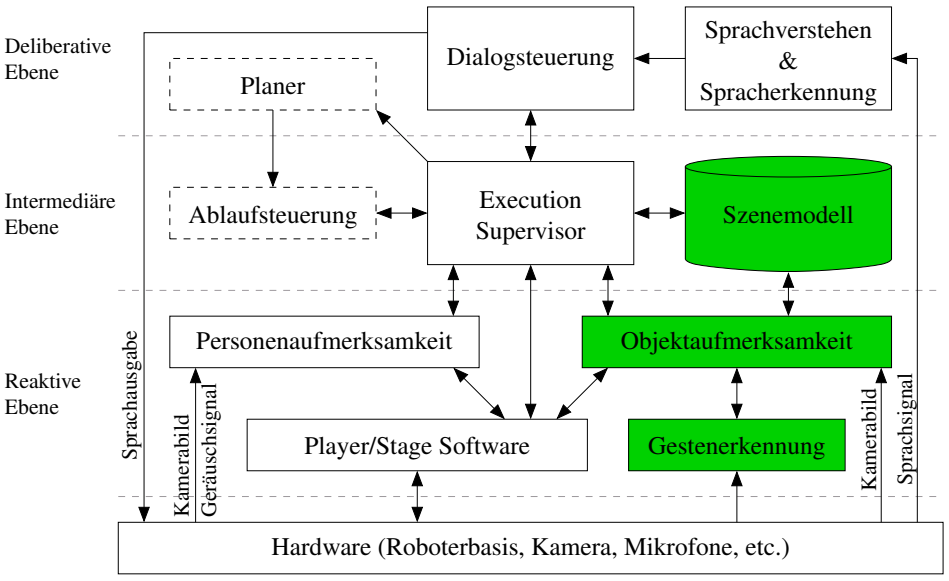


Abbildung 5.4: Schematische Systemarchitektur des Roboters BIRON. Grün hinterlegt sind die Module, die im Kontext der Gestenerkennung relevant sind. (Graphik entnommen und angepasst aus Wrede u. a. [2004a].)

Roboter autonomer agiert. Die Zerlegung der Pläne ist notwendig, da der *Execution Supervisor* jeweils nur ein Kommando verarbeiten kann. Dieser Komponente kommt aber eine zentrale Rolle zu, da mit ihr der Zustand des Gesamtsystems überwacht und gesteuert wird. Der *Execution Supervisor* ist mit einem endlichen Automaten umgesetzt, der unterschiedliche Zustände der Interaktion abbildet und entsprechend der multimodalen Eingaben Anweisungen an die einzelnen Module gibt. Es gibt zum Beispiel den Zustand „Beobachte“, in dem Personen im Umkreis des Roboters gesucht werden oder auch den Zustand „Erwarte Instruktion“, der die Aufmerksamkeit des Roboters auf Zeigegesten lenkt. Als weitere Komponente ist das Szenemodell in der intermediären Ebene angeordnet. Aufgabe dieses Moduls ist es, Informationen über Objekte, ihre Attribute und Lage, die für die Interaktion wichtig sein könnten, zu speichern und eine Schnittstelle für Anfragen bereit zu stellen.

Neben dem Modul der Objektaufmerksamkeit und der Gestenerkennung befindet sind in der reaktiven Ebene auch das Modul zur Ansteuerung der Roboterhardware und der Sensorik, der *Player/Stage Software* von Gerkey u. a. [2003]. Diese Software stellt einfache Schnittstellen für die Ansteuerung wie zum Beispiel Fahranweisungen zur Verfügung. Das von Lang [2005] entwickelte Modul zur Personenaufmerksamkeit vereinigt effizient Daten von drei unterschiedlichen Sensoren, um Personen im Umfeld des Roboters räumlich und zeitlich zu verfolgen. Ein Laserscanner wird verwendet, um in dem Tiefenbild Beinpaare von Personen zu finden. Auch wird damit die Entfernung und Ausrichtung des Menschen zum Roboter gemessen. Die Bilder der Kamera, die oben auf dem Roboter horizontal und vertikal schwenkbar montiert ist (siehe Abbildung 5.3), werden für eine Gesichtserkennung verwendet, die auf dem Verfahren von Viola u. Jones [2001] beruht. Ein erkanntes Gesicht erlaubt Schlüsse auf seine Entfernung sowie seine Ausrichtung. Auch die Größe der Person kann aus der Ausrichtung der Kamera sowie der Entfernung und Position des Kopfes

im Kamerabild berechnet werde. Des Weiteren werden zwei Mikrofone genutzt, um Geräuschquellen in der Umgebung vor dem Roboter zu lokalisieren. Diese unterschiedlichen Modalitäten werden fusioniert und ergeben Hypothesen für mögliche Interaktionspartner des Roboters und erlauben deren Verfolgung.

Der *Execution Supervisor* kann die Kontrolle von der Personenaufmerksamkeit zur Objektaufmerksamkeit verlagern, um deiktische Objektreferenzen des Benutzers zu erkennen. Die Kombination der Gesten und sprachlicher Instruktionen erlaubt es dem System zur *Objektaufmerksamkeit* (*Object Attention System* (OAS)), die bewegliche Kamera des Roboters anzusteuern, um Detailaufnahmen eines gestisch referenzierten Objektes zu gewinnen. Die so gewonnenen Informationen werden im Szenemodell für die spätere Interaktion abgelegt. Das in Kooperation des Autors mit Haasch u. a. [2005] entwickelte System wird im nächsten Unterkapitel (5.2) im Zusammenhang mit der Gestenerkennung für den Roboter behandelt.

Kommunikationssoftware

An dieser Stelle soll auch auf die genutzten Methoden der Systemkommunikation eingegangen werden. Alle Datenströme in dem System des Roboters BIRON sind XML-kodierte Nachrichten, die über das von Wrede u. a. [2006b] entwickelte *XML enabled Communication Framework* (XCF) ausgetauscht werden. Diese Technik erlaubt einen direkten Transport von XML Strukturen, inklusive einer auf XML-Schemata basierenden Validierung. Das XCF ist somit eine einfach zu verwendende Middleware für eine verteilte Architektur, das synchrone sowie asynchrone Funktionsaufrufe bereitstellt und einen Publisher-/Subscriber-Mechanismus implementiert. Auch können in den XML-Strukturen binäre Daten, wie zum Beispiel Bilder, referenziert werden. Diese Software erlaubt es, die Module des Roboters und somit die Rechenlast auf mehrere Rechner zu verteilen. In der aktuellen Konfiguration des Roboters werden die Module auf vier Rechner verteilt, das schließt einen portablen Rechner für die Sprachverarbeitung und Dialogsteuerung sowie einen für die Gestenerkennung und Objektaufmerksamkeit ein. Zwei weitere Computer sind fest in den Pioneer PeopleBot der Firma ActiveMedia eingebaut. Auf dem einen PC (Pentium III, 850MHz) läuft die Motoransteuerung, die Verarbeitung der Sensordaten des Roboters und die Geräuschverarbeitung. Der andere PC (Pentium III, 500MHz) wird für Bildverarbeitung und Assoziation von Daten verwendet. Weitere technische Details der Plattform werden von Wrede u. a. [2004a] beschrieben.

Des Weiteren lässt sich das XCF-System mit der Bildverarbeitungsplattform *IceWing* kombinieren, die Lömker u. a. [2006] entwickelt haben, so dass auch die rechenintensiven Bildverarbeitungsmodule auf unterschiedliche Computer verteilen werden können. Dieses System aus XCF und IceWing bietet neben der Lesbarkeit der Daten einen weiteren Vorteil, da die Ergebnisse eines einzelnen Moduls gespeichert werden können und andere Module diese asynchron als Eingabe nutzen können. Der Aufbau lässt ein separates Testen und Evaluieren einzelner Module zu und erleichtert die Fehlerverfolgung.

Interaktion mit BIRON

Um einen Einblick in die Möglichkeiten der Interaktion zu geben, die mit dem Roboter BIRON möglich sind, wird nun der Ablauf einer möglichen Interaktion skizziert:

- Sind mehrere Personen anwesend, wird die für den Roboter interessanteste mit der dreh- und schwenkbaren Kamera fokussiert.
- Ein Benutzer kann die Interaktion mit dem Roboter initialisieren, indem er ihn zum Beispiel mit „Hello BIRON“ grüßt.
- Von nun an fokussiert der Roboter seine Aufmerksamkeit auf diese Person und kann nicht von anderen Personen abgelenkt werden, die in seiner Umgebung sprechen.
- Die Person kann den Roboter anweisen, ihm zu folgen, um dem Roboter neue Räume oder Objekte zu zeigen.
- Folgt der Roboter der Person, so wird über die Robotersteuerung versucht, eine konstante Distanz einzuhalten.
- Ist das für den Roboter BIRON nicht möglich, weil die Person sich zu schnell bewegt, teilt der Roboter das sprachlich mit.
- Der Roboter kann jederzeit vom Benutzer angehalten werden (zum Beispiel mit „BIRON, please stop“)
- Sagt der Benutzer, dass er dem Roboter ein neues Objekt beibringen möchte, teilt der Roboter mit, wenn er bereit ist, eine Geste des Benutzers zu beobachten.
- Verlässt der Benutzer den Roboter oder verabschiedet er sich, nimmt der Roboter an, dass die Interaktion abgeschlossen ist und schaut sich nach weiteren potentiellen Interaktionspartnern um.

Der vorletzte Interaktionsfall ist eng mit der Gestenerkennung, die Thema dieser Dissertation ist, verbunden und wird näher betrachtet: Eine Äußerung wie „This is my favorite cup“ kann mit einer erkannten Geste verknüpft werden. Der Roboter versucht in der Region, auf die der Interaktionspartner seine Aufmerksamkeit gelenkt hat, ein Objekt zu finden. Hierfür wird die bewegliche Kamera des Roboters auf die Region ausgerichtet. Werden zusätzliche Informationen, zum Beispiel die Farbe der Tasse benötigt, erfragt der Roboter diese. Im Anschluss an diese Interaktion schwenkt die Kamera des Roboters wieder auf das Gesicht des Benutzers. Der Roboter hat eine Ansicht des Objektes gelernt und die aktuelle Position vermerkt. Eine weitere Interaktionsform ist das Zeigen neuer Räume, die der Benutzer benennt und die daraufhin vom Roboter wiedererkannt werden können.

5.2 Gestenerkennung für einen mobilen Roboter

Die im Kapitel 4 vorgestellte Gestenerkennung wird in zwei unterschiedlichen Konfigurationen für die multimodale Interaktion mit dem Roboter BIRON benutzt, um eine Objektaufmerksamkeit des Roboters zu ermöglichen. Ziel ist es, dem Roboter in einem multimodalen Dialog Objekte zeigen zu können.

In diesem Kapitel werden zwei leicht unterschiedliche Systeme zur Gestenerkennung und Auflösung von Objektreferenzen für den interaktiven, multimodalen Roboter BIRON vorgestellt. Das erste System arbeitet mit zweidimensionalen Bewegungsdaten, das zweite verwendet die dreidimensionalen Daten einer Körperverfolgung.

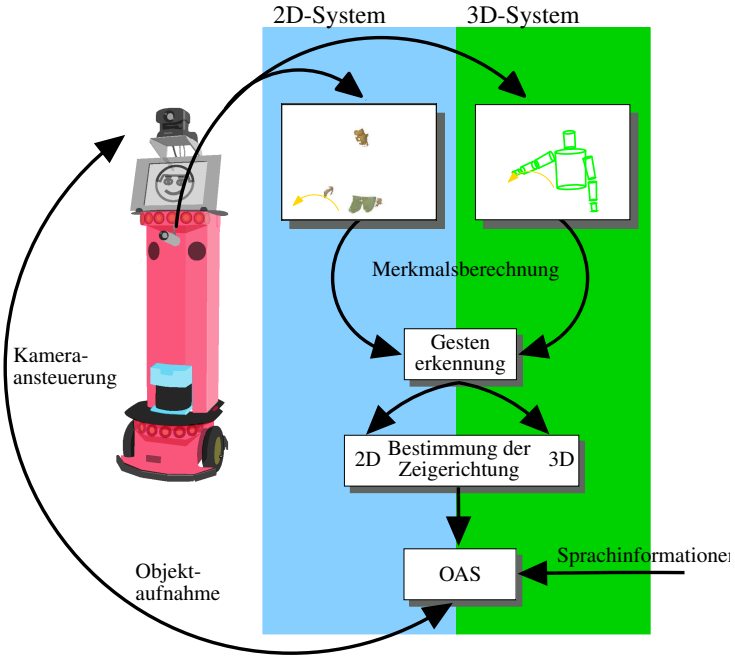


Abbildung 5.5: Die Graphik zeigt den schematischer Verarbeitungsprozess der Gestenerkennung und Auflösung von Objektreferenzen für den Roboter BIRON.

Der Roboter kann sprachliche und visuelle Informationen von seinem Interaktionspartner aufnehmen und verarbeiten. Das gestische Lenken der Aufmerksamkeit des Roboters erlaubt das Reden über die Gegenstände seiner Umwelt. Der Mensch kann dann fragen, wo seine Tasse steht und den Roboter bitten, ihm diese zu bringen. Wurde dem Roboter vorher die Tasse gezeigt und gesagt, wem sie gehört, hat er nun in seiner internen Repräsentation alle notwendigen Informationen zur Lösung dieser Aufgabe. In der Abbildung 5.5 werden die Prozessabläufe der beiden Systeme, die zur Gestenerkennung und Objektauflösung entwickelten wurden, in einer Graphik schematisch aufgezeichnet.

Im ersten System, das in Echtzeit bei der Demonstration des Roboters BIRON läuft, werden die Hände des Interaktionspartners als hautfarbene Regionen im Bild detektiert und verfolgt. Diese Trajektorien werden ausgewertet und Zeigegesten inklusive der Zeigerichtung werden erkannt. In dieser Konfiguration ist die Gestenerkennung komplett in das System des Roboters integriert. Deutet eine sprachliche Äußerung auf eine Geste hin, kann diese durch Informationen, die über eine deiktische Geste gegeben wurden, vervollständigt werden. Die Kollegin Sonja Hüwel zeigt in ihren Arbeiten, wie mit der Kombination aus Sprache und Gestik Ellipsen, also unvollständige Sätze, ergänzt werden können (siehe Hüwel u. a. [2006] und Hüwel u. Wrede [2006]). Ein Beispiel ist der Satz „Das ist meine Tasse“; die Referenz ist ohne eine Zeigegeste meist nicht aufzulösen.

Die Genauigkeit der Auflösung der Zeigerichtung aus der Handbewegung im Bild ist begrenzt. Um die Objektreferenzen genauer zu bestimmen, wird auch die dreidimensionale Körperverfolgung von Schmidt u. a. [2006] verwendet, die unter anderem die Position der Hand im Raum verfolgen kann. Für die Gestenerkennung bedeuten diese unterschiedlichen Systemkonfigurationen nur, dass unterschiedliche Modelle trainiert werden müssen und dass die Daten von unterschiedlichen Modulen geliefert werden. Für das Objektauf-

merksamkeitssystem (OAS) bedingen die unterschiedlichen Systeme auch unterschiedliche Berechnungen der Richtung der Zeigegeste und der Ermittlung der *Aufmerksamkeitsregion* (*Region-Of-Interest* (ROI)).

Unter Verwendung der Körperverfolgung kann die Hand im Kontext des Körpers des Menschen gesehen werden, folglich ist es möglich, die Zeigerichtung nicht nur aus der Bewegung der Hand zu schließen, sondern auch aus der Blickrichtung. Diese wird aus der Verbindung zwischen Kopf und Fingerspitze im Raum gewonnen und in das Koordinatensystem des Roboters transformiert. Das auf der Körperverfolgung aufbauende System ist zurzeit noch nicht in die Demonstration des Roboters eingebaut, da die Körperverfolgung zum einen eine manuelle Initialisierung benötigt. Zum anderen läuft das Verfolgen des Körpers noch nicht in Echtzeit, wenn ein robustes und genaues Ergebnis sicher gestellt sein soll.

Das in Kapitel 4.4 vorgestellte Einbeziehen der symbolischen Objektinformationen kommt in diesem System nicht zum Einsatz, da auf bisher unbekannte Objekte gezeigt wird und das Auflösen der Objektreferenzen somit zur Objektaufmerksamkeit verlagert wird. Des Weiteren steht auch kein zuverlässiger und für das gewählte Szenario trainierter Objekterkenner zur Verfügung. Das auf der Verfolgung hautfarbener Regionen basierende System ist in Kooperation mit Haasch u. a. [2005] entwickelt und veröffentlicht worden, seine Einbindung in das System des Roboters BIRON wird in dem Artikel von Wrede u. a. [2004a] dargestellt.

Systemaufbau

Bevor in den folgenden Abschnitten die Einzelmodule zum Erkennen von Zeigegesten behandelt werden, wird in diesem Abschnitt ihr Zusammenspiel erläutert. Die Unterschiede in dem auf der Hautfarbendetektion (siehe Graphik 5.6) aufbauendem System und dem System, dass die Körperverfolgung (siehe Graphik 5.7) verwendet, werden hervorgehoben.

Für das Erkennen von Zeigegesten wird eine zusätzliche Kamera (Apple iSight) benötigt, die am Torso des Roboters angebracht ist. Die Kamera und ihre Position wurde gewählt, weil die bewegliche Kamera oben auf dem Roboter einerseits nur einen kleinen Bildausschnitt aufnimmt und andererseits von der Personenaufmerksamkeit bewegt wird. So kann nicht sichergestellt werden, dass die Arme des Menschen, der sich vor dem Roboter befindet, während einer Aktion im Bild sichtbar sind. Für die Objektaufmerksamkeit wird hingegen die obere Kamera benutzt, weil diese auf ein referenziertes Objekt ausgerichtet werden kann und eine größere Ansicht des Objektes für die Analyse genommen werden kann. Die Objekterkennung basiert auf Farbkarten und kann das referenzierte Objekt aus der Region, auf die gezeigt wurde, extrahieren.

Die agierende Hand des Benutzers kann über hautfarbene Regionen im Bild detektiert und verfolgt werden. Die Bildkoordinaten dieser Hand werden in Form einer Trajektorie zur Verfügung gestellt und in die Geschwindigkeit und Richtungsänderung in der Bildebene umgerechnet (siehe Abbildung 5.6). Wird eine Geste erkannt, ist zu dem aktuellen Zeitpunkt der Ort der Hand im Bild und die Richtung der Geste bekannt. Die Richtung ist die gemittelte Bewegungsrichtung aus den letzten drei Zeitschritten. Die Region für die Aufmerksamkeit, die *Region-Of-Interest*, die Axel Haasch mit dem OAS berechnet und in der das Objekt gesucht wird, liegt im Bild in der Bewegungsrichtung vor der zeigenden Hand.

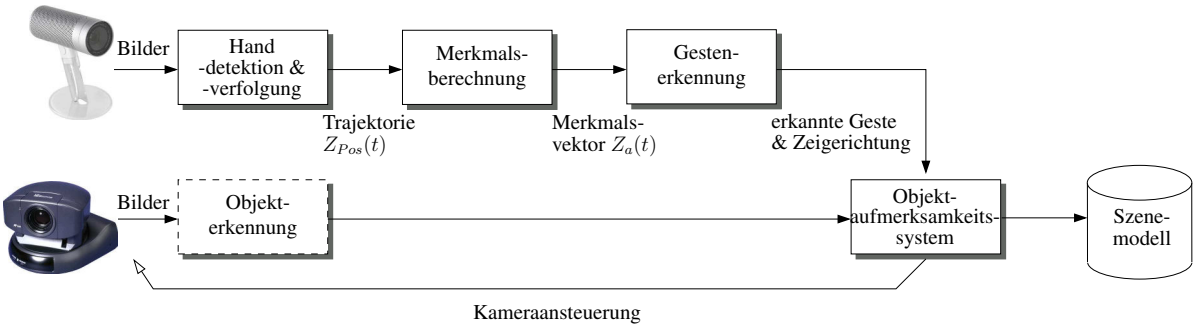


Abbildung 5.6: Architekturskizze und Prozesskette der Gestenerkennung und Objektaufmerksamkeit. Die Gestenerkennung arbeitet mit zweidimensionalen Trajektorien, die das Modul zur Handdetektion und -verfolgung ausgibt. Die Richtung der Zeigegeste wird aus der Handbewegung bestimmt.

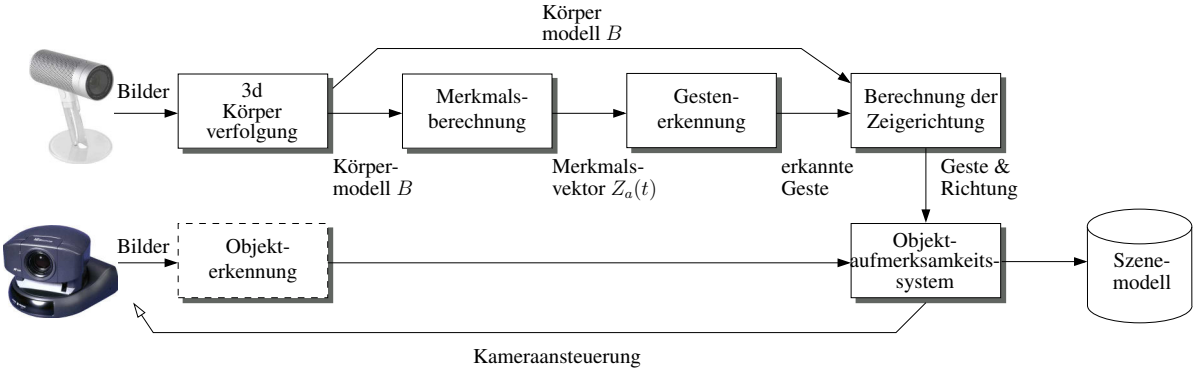


Abbildung 5.7: Architekturskizze und Prozesskette der Gestenerkennung und Objektaufmerksamkeit. Der Körper der Versuchsperson wird mit einer Körperverfolgung erkannt und über die Zeit verfolgt. Aus dieser werden die Merkmale für die Gestenerkennung gewonnen. Die Richtung einer Zeigegeste wird zum Zeitpunkt einer Erkennung aus der Körperkonfiguration ermittelt.

Die Unterscheidung zwischen den sehr ähnlichen Bewegungen des Zeigens und Zurückführens der Hand wird gelöst, indem eine Körperregion definiert ist, in die der Mensch seine Hand zurückführt, auf die er aber nicht zeigt. Der Roboter und somit auch die Kamera werden von der Personenaufmerksamkeit auf den Interaktionspartner ausgerichtet. So ist sichergestellt, dass die Körperregion in der Mitte des Kamerabilds liegt.

Alternativ kann mit der 3D-Körperverfolgung von Schmidt u. a. [2006] die Stellung und Position eines Menschen im Bild der Kamera erkannt und der Körper über die Zeit verfolgt werden (siehe Abbildung 5.7). Bei der Verwendung der Körperverfolgung ist natürlich exakt das Wissen, ob sich die Hand zum Körper bewegt oder sich entfernt, verfügbar und wird entsprechend in die Bewegungsrepräsentation aufgenommen. Das Ergebnis liegt in einer XML-Struktur vor, die die Körperkonfiguration des Menschen beschreibt. Für die weitere Verwendung wird die Position der rechten Hand und der rechten Schulter extrahiert. Es folgt die Berechnung von Merkmalen der Handbewegung, diese werden für die Gestenerkennung verwendet. Für den weiteren Erkennungsprozess wird die erkannte Ges-

te sowie ein kurzer zeitlicher Verlauf der Raumposition der Hand und des Kopfes in einer XML-Struktur ausgegeben. Aus diesen Daten wird in einer Nachverarbeitung die erwartete Raumrichtung der Zeigegeste berechnet. Genauso wie im oben beschriebenen Fall steuert das OAS die bewegliche Kamera in diese Richtung und versucht das referenzierte Objekt zu finden.

In dem OAS werden außerdem sprachliche Details zum Objekt, wie zum Beispiel die Farbe oder Größe, hinzugezogen. Das gefundene Objekt wird daraufhin mit der akquirierten visuellen Ansicht und symbolischen Daten, wie zum Beispiel der Farbe oder Größe, in das Szenemodell eingefügt.

Nach diesem Überblick wird auf die genannten Komponenten genauer eingegangen. Bestandteil der vorliegenden Dissertation sind die Merkmalsberechnung und Vorverarbeitung der Trajektoriendaten sowie die Gestenerkennung. Für die zweidimensionale Repräsentation schließt das die Ermittlung der Zeigerichtung mit ein, für den dreidimensionalen Fall hat Axel Haasch nach gemeinsamer Konzeption die Implementierung übernommen. Das Bestimmen der referenzierten Region, in der das Objekt vermutet wird, ist Thema der Dissertation von Axel. Schwerpunkt der Evaluation wird in dieser Arbeit das Erkennen von Gesten und die zeitliche Genauigkeit dieser Erkennung sein.

5.2.1 Verfolgen einer Person

Im Folgenden werden die zwei unterschiedlichen Methoden vorgestellt, mit denen eine Trajektorie der Hand aus einer Bildsequenz gewonnen werden kann. Obwohl beide Verfahren nur eine einfache Farbkamera verwenden, ist es doch mit dem zweiten Verfahren möglich, Tiefeninformationen zu extrahieren, wie im nächsten Abschnitt erläutert wird.

Detektion und Verfolgen der Hände

Die Detektion und Verfolgung einer menschlichen Hand aufgrund ihrer speziellen Hautfarbe wurde von Fritsch [2003] entwickelt und detailliert beschrieben. Das Verfahren beruht auf einer adaptiven Segmentierung der hautfarbenen Pixel in dem helligkeitsnormierten RG-Farbraum. Über zeitliche Schwellwerte und Schwellwerte für die minimale Größe der Regionen werden über die Zeit stabile Regionen detektiert und Regionen, die sich bewegen, werden mit einem Kalman-Filter verfolgt. Das Ergebnis ist eine Trajektorie der Hand in der Bildebene der Bildsequenz.

Das Verfahren zum Verfolgen einer Person

Das von Schmidt u. a. [2006] entwickelte probabilistische Verfahren zur Verfolgung eines Menschen ermöglicht mit nur einer Kamera ein Verfolgen des Körpers in drei Dimensionen. Hierfür werden Bildmerkmale und ein dreidimensionales Modell des menschlichen Körpers verwendet (siehe Abbildung 5.8). Die Parameter der Gelenkwinkel des Körpermodells bestimmen dessen Konfiguration. Für ein so konfiguriertes Modell werden für alle Gliedmaße Bildmerkmale berechnet, die eine Bewertung einer Konfiguration ergeben.

Mit dem Condensation-Algorithmus (siehe 4.1.3) werden diese Konfigurationen probabilistisch propagiert und so die Bewegungen des beobachteten Körpers verfolgt. Als Bildmerkmale werden Kanten- und Eckenmerkmale, die mittlere Farbe in eines Gliedmaßes

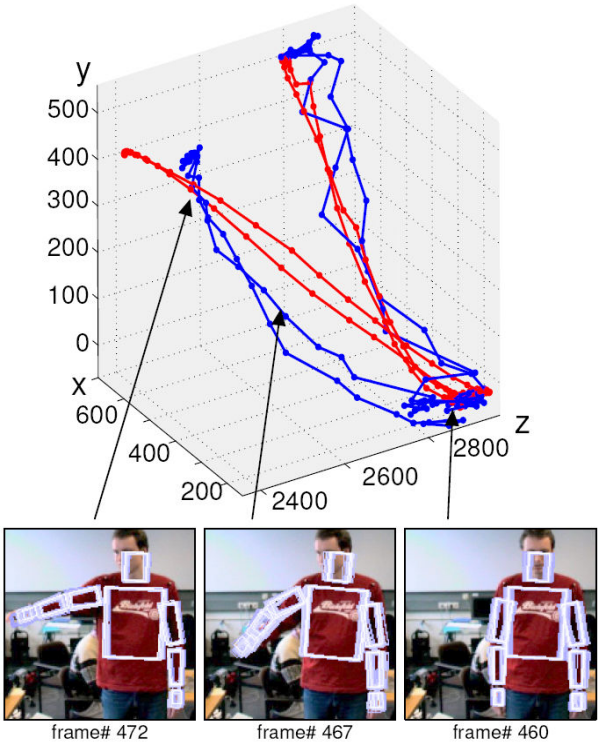


Abbildung 5.8: Visualisierung der Ergebnisse der Körperverfolgung im metrischen Raum. Abgetragen ist die erkannte Bewegung der Hand (Blau), sowie die mit einem aktiven Verfolgungsverfahren ermittelten Werte (Rot). Im unteren Teil sind exemplarisch einige Bilder mit der besten Körperkonfiguration abgebildet. (Graphik entnommen aus Schmidt u. a. [2006].)

sowie die Hautfarbe für die Hände und den Kopf verwendet. Die Anzahl der benötigten Partikel kann gering gehalten werden, da der *Mean-Shift* Algorithmus benutzt wird, um die Partikel in Gebiete hoher Gewichte zu verschieben (siehe Schmidt u. a. [2006]). Mit dieser Technik ist eine Verfolgung in Echtzeit möglich, allerdings wird bisher ein zuverlässiges und weniger verrauschtes Resultat noch nicht ganz in Echtzeit erreicht.

Ein entscheidender Vorteil dieses Verfahrens für das Erkennen von Zeigegesten besteht in der zusätzlichen Tiefeninformation, die so auch mit einem einfachen Kameraaufbau auf einem mobilen Roboter möglich ist. Des Weiteren ist es möglich, nicht nur die Bewegung der Hand im Bild zu betrachten, sondern ihre Bewegung relativ zum Körper. Der Ablauf der Verfolgung eines menschlichen Körpers ist mit einer Bildsequenz in Graphik 5.8 dargestellt, zu sehen sind sowohl die besten Körperkonfigurationen als Überlagerung auf den Bildern der Sequenz, also auch die Bewegung der Hand im Raum. Die Referenzdaten wurden mit einem Infrarotsystem, das aktive Markierungen verwendet, aufgenommen.

Die Positionen der Gliedmaßen liegen in einem metrischen Koordinatensystem vor, das seinen Ursprung in der Kamera hat. Die 3D-Körperverfolgung ist ebenso wie die für diese Arbeit entwickelten Module als Plug-in für das Bildverarbeitungsprogramm IceWing entwickelt und gibt die beste Körperkonfiguration eines Zeitschrittes als XML-Struktur aus. Ein Beispiel der verwendeten XML-Struktur, in der die relevanten Teile hervorgehoben sind, zeigt die Abbildung 8.5 im Anhang.

5.2.2 Verwendung des CTR

Die Bewegungen des Menschen sollen von der Gestenerkennung mit möglichst wenigen generischen Modellen klassifiziert werden. Hierzu ist eine Abbildung der Bewegungen in geeigneten Repräsentationen notwendig, wie die in Unterkapitel 3.2 vorgestellten. Auch die Modelle müssen für den gewählten Aufbau und den Merkmalsraum gebildet werden, bevor das Erkennen von Zeigegesten möglich ist. Hierfür wird das in 4.3 vorgestellte Verfahren verwendet.

Berechnen der Merkmale aus der Trajektorie der Hand

Für die Repräsentation der Bewegung in der Bildebene wird, wie schon für die Experimente in Abschnitt 4.4, die Berechnung der Geschwindigkeit und Richtungsänderung zur Trajektorie $Z_{Vel-Ang}(t)$ gewählt (siehe Gleichung 3.2).

Berechnen der Merkmale aus dem Körpermodell

Für die Merkmale, in denen die Bewegungen repräsentiert werden, wird ein schulterzentriertes Zylinder-Koordinatensystem verwendet. Vorteile sind, dass Bewegungen, die vom Körper wegführen, rotationsinvariant abgebildet werden, denn die Richtung, in die eine Zeigegeste ausgeführt wird, ist für die Erkennung und Ermittlung des Zeitpunktes der Geste nicht relevant. Hierfür muss nur der Bewegungsverlauf betrachtet werden; die Richtung und Körperkonfiguration ist für das spätere Berechnen der Zeigerichtung und das Auflösen des referenzierten Objektes nötig. Ein Ziel der Evaluation ist es deswegen, zu überprüfen, ob wenige oder nur ein Modell für das Zeigen auf Objekte, die auf einem Tisch vor der Versuchsperson liegen, genügen.

Die Merkmale, die für das Erkennen benutzt werden, sind der Abstand r der Hand zum Körperpunkt in der horizontalen Ebene sowie der Winkel δ in dieser. Die Vertikalachse z ist ein weiteres Merkmal. Die Merkmale können sowohl absolut verwendet werden, als auch in ihren ersten Ableitungen, also der Geschwindigkeiten beziehungsweise der Winkelgeschwindigkeit. Wie bereits in Abschnitt 3.2.3 erläutert, haben die Ableitungen den Vorteil, dass sie invariant im Bezug auf die Raumposition sind, doch auch den Nachteil, dass diese höheren Momente ein schlechteres Signal-Rausch-Verhältnis aufweisen. Das zugrundeliegende Koordinatensystem dieser genannten Merkmale ist in Graphik 5.9 dargestellt.

Um die Daten der Körperverfolgung für die Gestenerkennung einsetzen zu können, müssen sie, neben der Merkmalsberechnung, geglättet werden. Da die Daten sequenziell vorliegen und außerdem die Extrempunkte möglichst erhalten bleiben sollen, wird der Savitzky-Golay Filter angewandt (siehe Press u. a. [1992], Kap. 14.8). Dieser Filter lässt sich kausal anwenden und sein Glättungsverhalten über die Fenstergröße und den Grad des Polynoms einstellen. Da der Savitzky-Golay Filter aber die Werte mit einer Schwingung überlagern kann, kann auch ein Mittelwert in einem Fenster der Größe drei verwendet werden. Dieser Filter ist nicht kausal, verursacht folglich eine Verzögerung der Messwerte. Dieser systematische Fehler kann aber vernachlässigt werden, wenn die gleiche Glättungsmethode für die Merkmalsberechnung der Daten und des Trainings benutzt wird.

Ein Beispiel der geglätteten Merkmale ist in Abbildung 5.10 aufgetragen. Der zeitliche Verlauf der Raumposition im euklidischen Koordinatensystem der Hand ist im Kasten

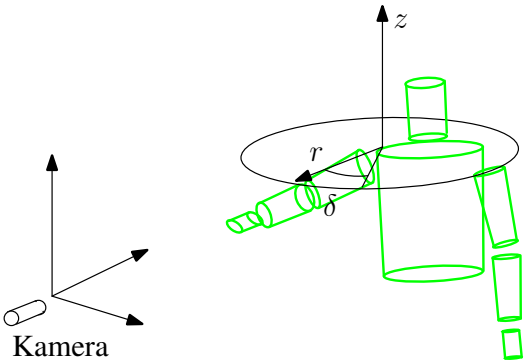


Abbildung 5.9: Koordinatensystem für die Merkmalsberechnung. Die Bewegung der Hand wird in einem zylindrischen Koordinatensystem analysiert. Die Schulter des Körpermodells ist der Ursprung des Systems.

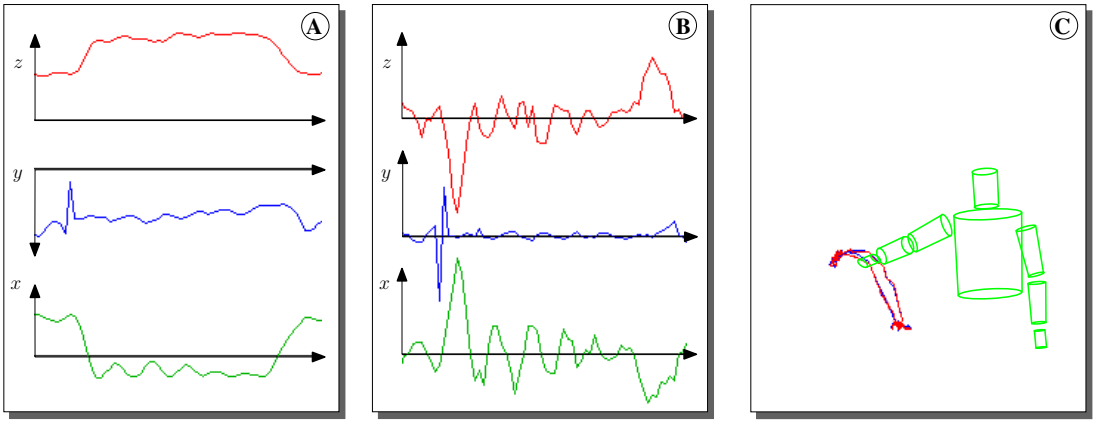


Abbildung 5.10: Die geglättete Bewegung der Hand im kartesischen Raum. Im Graphen (A) werden die absoluten Werte abgebildet, während in (B) die Geschwindigkeiten in den jeweiligen Raumdimensionen zu sehen sind. Die Trajektorie der Handbewegung in der Bildebene zeigt (C).

(A) dargestellt, die erste Ableitung dieser Kurve wird in (B) gezeigt. Zu beobachten ist hier, dass die Tiefenwerte (x) stärker verrauscht sind als die Werte in der Höhe und Breite, da diese der Bildebene entsprechen. In (C) ist der Verlauf der Handbewegung in der Bildebene wiedergegeben. Die Merkmale werden in der in Abbildung 5.11 aufgezeigten XML-Struktur abgelegt.

Gestenerkennung

Für die Verwendung der Gestenerkennung in diesem Szenario sind nur wenige Anpassungen des im Kapitel 4 vorgestellten Verfahrens notwendig. Natürlich müssen die Modelle der Gesten generiert werden und über die Parameter des Programms muss angegeben werden, welche Merkmale aus der XML-Struktur, die das Modul der Merkmalsberechnung erzeugt, genutzt werden. Die Gestenerkennung fügt erkannte Gesten dem bestehenden XML-Fragment der Merkmalsextraktion hinzu (siehe Graphik 5.12). Diese werden als Eingangsdaten für das OAS benutzt.

```
<DATA>
  <FEATURE Framerate="40" ID="" ImgNum="127" Reset="0">
    <VELOCITY A="0.097" X="-0.095" Y="-0.001" Z="0.019"/>
    <DIRECTION DDelta="-0.148" Delta="1.404"/>
    <CYLINDRICAL DDistance="-0.001" DEta="0.083" DHeight="-0.010"
                 Distance="0.321" Eta="-1.476" Height="0.181" />
    <RAW>
      <STEP T="0">
        <RIGHTHANDPOS X="2.428" Y="-0.016" Z="-0.395"/>
        <RIGHTFINGERTIP X="2.328" Y="-0.014" Z="-0.365"/>
        <HEADPOS X="2.461" Y="-0.339" Z="0.335"/>
      </STEP>
    </RAW>
  </FEATURE>
</DATA>
```

Abbildung 5.11: XML-Struktur mit den berechneten Merkmalen und den Originaldaten. (Siehe auch 8.6)

```
<DATA>
  ...
  <GESTURE>
    <PROGRESS>0.949996</PROGRESS>
    <CHILD>Point</CHILD>
  </GESTURE>
</Data>
```

Abbildung 5.12: XML-Struktur einer erkannten Geste. (Siehe auch 8.6)

In Abbildung 5.13 ist exemplarisch der Verlauf der Wahrscheinlichkeiten während der Erkennung einer Zeigegeste aufgezeichnet. Es ist der zeitliche Verlauf der Endwahrscheinlichkeiten für die unterschiedlichen Gestenmodelle dargestellt. Einige Bilder der zugehörigen Bildsequenz sind zusätzlich enthalten, um den zeitlichen Ablauf der Geste zu verdeutlichen.

5.2.3 Berechnung der Zeigerichtung

In Zusammenarbeit mit Joachim Schmidt und Axel Haasch wurde die Berechnung der Zeigerichtung entworfen, die Implementation wurde von Axel Haasch vorgenommen. Ziel der Berechnung ist es, die dreh- und schwenkbare Kamera des Roboters möglichst genau auf die Position im Raum, auf die gezeigt wurde, auszurichten. Der nun skizzierte Vorgang wird durch das Erkennen einer Geste initialisiert.

Berechnung für den zweidimensionalen Fall

Damit die Berechnung der *Region-Of-Interest* nicht abhängig von der Bildgröße der Kamera ist, werden die Koordinaten der Hand von der Gestenerkennung auf das Intervall

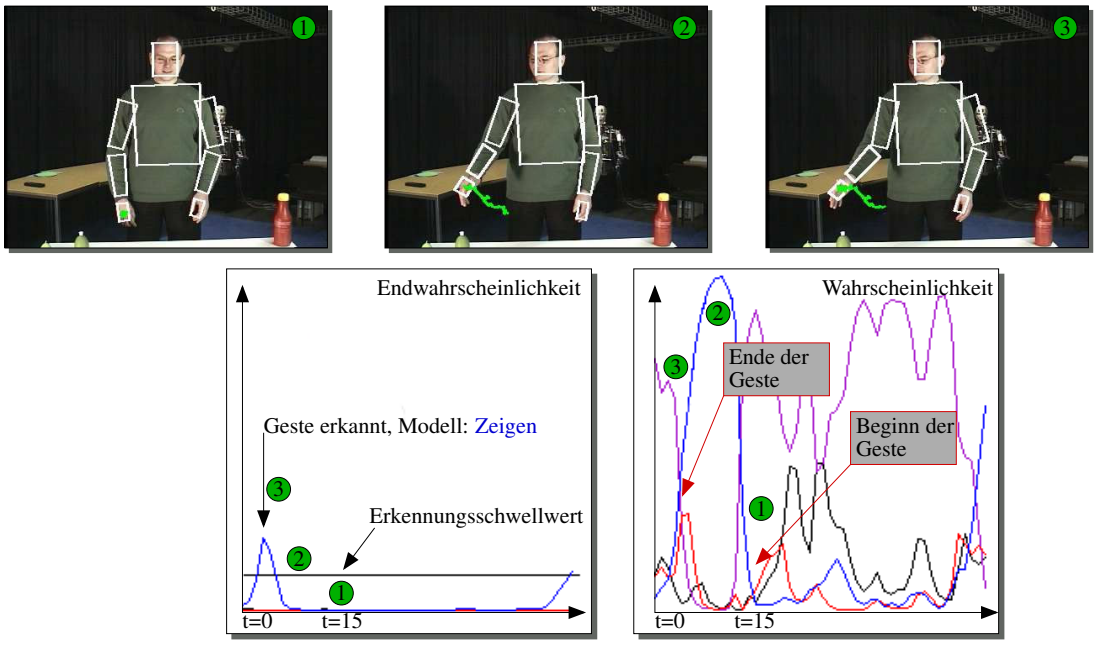


Abbildung 5.13: Darstellung der Ergebnisse der Gestenerkennung mit den zugehörigen Bildern aus der Sequenz. Abgebildet wird rechts die Wahrscheinlichkeit, die sich auf mehrere Modelle verteilt. Links sind die Endwahrscheinlichkeiten, die den einzelnen Modellen zugeordnet sind, aufgetragen. Sichtbar wird wie die Kurve für die Zeigengeste im rechten Graphen mit Beginn der Geste ansteigt und mit ihrem Ende wieder abfällt. Der linke Graph zeigt zu diesem Zeitpunkt eine Spitze in der Kurve des Modells „Zeigen“. Da diese Spitze über dem Schwellwert der Erkennung liegt, wird die Geste als *erkannt* ausgegeben.

[0..1] normiert. Aus dieser Position im Bild und der Bewegungsrichtung der Hand wird ein Punkt gewählt, der 10% der Bilddimensionen vor der Handposition im Bild liegt. Diese Position ist die vermutete Position des referenzierten Objektes im Bild. Die Hypothese der Entfernung zwischen der Kamera und dem referenzierten Objekt basiert auf der Entfernung des Menschen zum Roboter. Diese wird mit dem Laser gemessen und wird von der Komponente zur Personenaufmerksamkeit bereitgestellt.

Wird die Zeigegeste im Bild nach unten ausgeführt, besteht die Annahme, dass auf ein Objekt gezeigt wird, das vor dem Benutzer auf einem Tisch liegt. Da der Mensch in diesem Fall seine Hand nach vorne streckt, wird die Hypothese für die Entfernung entsprechend verringert. Die ROI ist ein Rechteck, mit dem berechneten Punkt als Mittelpunkt.

Berechnung für den dreidimensionalen Fall

Die Positionen der Fingerspitze (RIGHTFINGERTIP) der rechten Hand werden aus dem XML-Fragment der Gestenerkennung extrahiert. Diese liegt für drei Zeitschritte vor; die vermutete Fingerposition wird aus diesen drei Werten gemittelt. Die Kopfposition der Person, die auch in den XML-Daten enthalten ist, sowie die Position der Finger, liegen in einem kartesischen Koordinatensystem vor. Dessen Ursprung liegt in der Kamera für die Körperverfolgung.

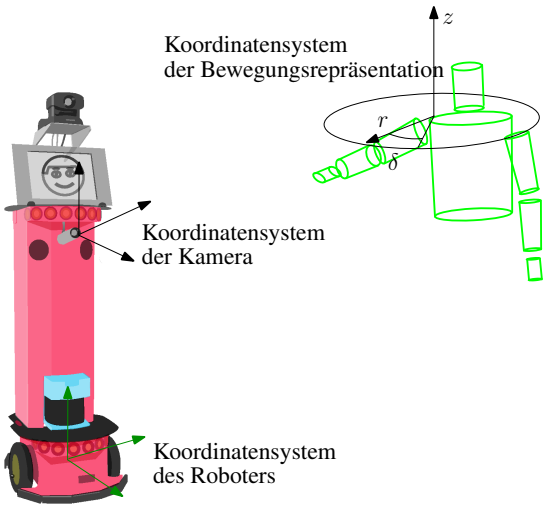


Abbildung 5.14: Schematische Darstellung der unterschiedlichen Koordinatensysteme für das Auflösen von Objektreferenzen. Die Körperkonfiguration der beobachteten Person liegt in dem Koordinatensystem der Kamera vor. Die Handbewegung wird umgerechnet in das zylindrische Koordinatensystem, das in der Schulterposition der Person liegt. Damit die schwenk- und drehbare Kamera auf ein referenziertes Objekt ausgerichtet werden kann, wird die Handposition in das Koordinatensystem des Roboters umgerechnet.

Unter Berücksichtigung des vertikalen und horizontalen Versatzes sowie der vertikalen Neigung dieser Kamera werden die Raumpositionen der Fingerspitzen und des Kopfes in das zylindrische Koordinatensystem des Roboters umgerechnet (siehe Abbildung 5.14). Der Vektor, der von der Kopfposition zur Position der Fingerspitzen im Raum führt, wird entsprechend der verbalen Spezifikation der Objektgröße verlängert. Die Größe wirkt sich auch auf den Zoomfaktor aus, mit dem die dreh- und schwenkbar Kamera das referenzierte Objekt aufnimmt. Die Kamera kann nun auf die errechnete Position des Objektes ausgerichtet werden.

5.2.4 Die Objektaufmerksamkeit

Das von Haasch u. a. [2005] entwickelte Aufmerksamkeitssystem für Objekte koordiniert sprachliche und gestische Informationen sowie Objektmerkmale und die Ansteuerung der Roboterhardware (vergleiche auch Abbildung 5.7). Die Informationsfusion ermöglicht es, Objekte in der Umgebung des Roboters zu lernen und für die spätere Interaktion verfügbar zu machen. Ähnlich wie das System zur Personenaufmerksamkeit ist es somit ein essentieller Beitrag, um einen situierten, sich seiner Umwelt bewussten Roboter zu konstruieren.

Das OAS wird aufgrund von sprachlichen Äußerungen, die Objekte involvieren, aktiviert. Dieses kann zum Beispiel die Äußerungen „Darf ich dir etwas zeigen“ oder „Das ist Brittas Tasse“ sein. Wird zudem eine Geste erkannt, so kann die *Region-Of-Interest* über die bereits erläuterte Berechnung der Zeigerichtung bestimmt werden.

Mit dem OAS wird versucht, eine möglichst genaue Repräsentation des referenzierten Objektes zu erstellen. In Abhängigkeit von der Objektgröße wird eine Gaußglocke über die

Region-of-Interest gelegt, um nahe stehende ähnliche Objekte ausblenden zu können. Für diese Region werden Farbkarten generiert. Sie können mit sprachlich gegebenen Farbinformationen verglichen werden. Ist das Objekt mehrfarbig, so werden die farbigen Regionen des Objektes in einem Merkmalsgraph repräsentiert. Des Weiteren werden Merkmale für eine nachfolgende Objekterkennung aus der Objektansicht extrahiert und im Szenemodell abgelegt.

Das Szenemodell benutzt das Konzept des *Aktiven Speichers* (*active memory*) von Wrede u. a. [2004b]. In dem Modell können sowohl Objektinformationen und Objektansichten für die Objekterkennung gespeichert werden, als auch sprachliche Informationen abgelegt werden. Veränderungen in der Umgebung des Roboters sind typisch für das Szenario. Den Umgang mit dieser Dynamik erlaubt der *Aktive Speicher* durch Vergessenskonzepte und das Aktualisieren von Informationen. Zum Beispiel ist die Information, wo eine Tasse vor drei Monaten stand, irrelevant, doch die Information, wem die Tasse gehört, ist auch über längere Zeit gültig.

5.2.5 Evaluation des Systems

Methode der Annotation und Auswertung

Für die Evaluation wurde neben dem Modul zur Merkmalsberechnung und Gestenerkennung noch ein Annotationsmodul entwickelt, das auf dem Bildverarbeitungswerkzeug *IceWing* aufsetzt. Dieses Annotationswerkzeug ermöglicht es, manuell die Merkmalsdaten um eine Annotation zu erweitern. Es werden die Namen der zu erkennenden Modelle zur Verfügung gestellt und sie können ausgewählt werden, wenn eine entsprechende Bewegung abgeschlossen wird. Diese zusätzlich in das XML-Fragment eingetragenen Daten (siehe Abbildung 5.15) werden im Anschluss an die Gestenerkennung genutzt, um die Erkennung zu evaluieren.

```
<DATA>
  ...
  <ANNOTATION Still="0" Untracked="1">Point</ANNOTATION>
</Data>
```

Abbildung 5.15: XML-Struktur mit einer Geste und Annotation. In diesem Fall ist die annotierte Geste „Point“. Des Weiteren kann in der Struktur vermerkt werden, ob die Hand im Moment nicht bewegt wird oder die Verfolgung ein Ergebnis liefert. (Siehe auch 8.6)

Bei der Evaluation der Gestenerkennung muss beachtet werden, ob eine Geste richtig erkannt wurde und wann diese erkannt wurde. Die Anzahl n der vorhandenen Gesten basiert auf der Annotation. Die Erkennungsleistung spiegelt sich in den korrekt erkannten Gesten k, den nicht erkannten Gesten (l Löschung), den falschen Erkennungen (v Vertauschung) und den Einfügungen e von Gesten wieder. Aus diesen Werten lassen sich die Fehlerrate FR und die Erkennungsrate ER berechnen (siehe Gleichung 4.48 und 4.47).

Sowohl in der Annotation als auch in der automatischen Erkennung wird das Ende einer Geste als Zeitbereich angegeben. Aufgrund dieser Zeitbereiche muss bei der Evaluation

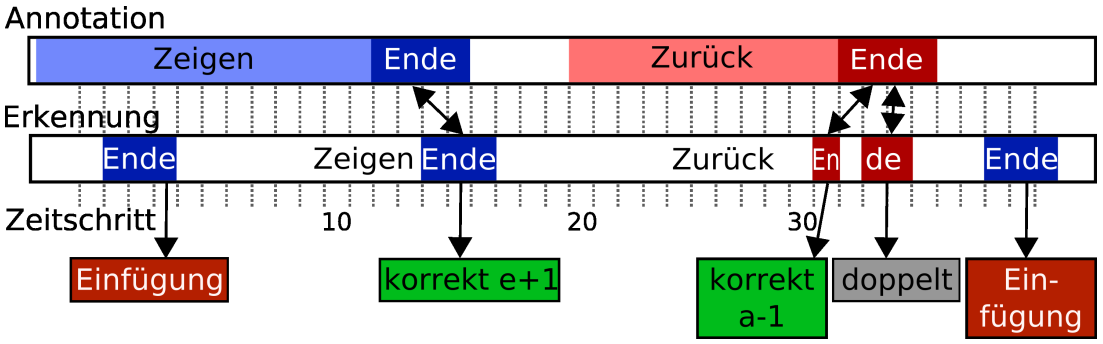


Abbildung 5.16: Eine schematische Darstellung der Auswertung von erkannten und nicht erkannten Gesten. Für zwei exemplarische Gesten ist oben die Annotation aufgetragen, darunter sind Beispiele für Erkennungsergebnisse dargestellt. An diesen ist die entsprechende Auswertung notiert.

die Entscheidung getroffen werden, ob eine Geste erkannt wurde oder ob die Geste fehlt. Wurde eine Geste nicht erkannt, kann das auch darin begründet liegen, dass sie zu spät erkannt wurde. Dieser Fall führt direkt zu zwei Fehlern, denn die Geste wurde erstens nicht erkannt und zweitens wurde sie eine kurze Zeit fälschlich erkannt, also eingefügt.

In der Natur des probabilistischen CTR liegt es, dass es zu einer Geste auch mehrere richtige Erkennungen geben kann. Denn die Endwahrscheinlichkeit wird für jeden Zeitschritt berechnet. Nun kann es aber vorkommen, dass die Endwahrscheinlichkeit in einem Zeitschritt für ein Modell über dem Schwellwert der Erkennung liegt, im nächsten nicht, aber im darauf folgenden Zeitschritt wieder. In diesem Fall wird die zweite erkannte Geste als doppelte Erkennung d gezählt.

In der für diese Dissertation ausgeführten Evaluation wird angenommen, dass eine Geste, die in einem eine Sekunde großen Fenster erkannt wird, als korrekt gewertet wird. Bei einer Videosequenz, die mit 15 Bildern pro Sekunde aufgenommen wurde, heißt das, dass eine korrekt erkannte Geste frühstens sieben Zeitschritte vor der Annotation und spätesten sieben nach ihr auftreten darf.

Die zeitliche Genauigkeit der Erkennung wird aus dem Versatz der Erkennung berechnet. Dazu wird für jedes korrekte oder doppelte Erkennungsergebnis $i \in I_c$ der Versatz des Beginns b_i beziehungsweise des Endes e_i zur jeweiligen Annotation ausgewertet. In der Abbildung 5.16 endet das korrekte „Zeigen“ einen Zeitschritt nach der Annotation, da das die zweite erkannte Geste ist, ist $e_2 = +1$. Entsprechend wird $a_3 = -1$ für die „Zurück“-Bewegung gesetzt. Doppelte Erkennungen werden hierbei auch mit aufgenommen. In einer formalisierten Betrachtung ist $\mathbf{A} = (A_1, \ldots, A_J)$ die Menge der Annotationen einer Videosequenz und $\mathbf{E} = (E_1, \ldots, E_I)$ die Menge der Erkennungen. Eine Annotation A_j enthält einen Anfang (b), einen Endbereich ($e1$ bis $e2$) und eine Beschreibung (g) der vorliegenden Geste:

$$A_j = \left\{ A_j^b; A_j^{e1}; A_j^{e2}; A_j^g \right\}. \tag{5.1}$$

Entsprechend enthält die Erkennung auch einen Endbereich und die Beschreibung des erkannten Gestenmodells:

$$E_j = \left\{ E_j^{e1}; E_j^{e2}; E_j^{g} \right\}. \tag{5.2}$$

Nach der zeitlichen Zuordnung einer korrekten oder doppelten Erkennung zu einer Annotation ($E_i^g = A_j^g$) kann der zeitliche Versatz ermittelt werden:

$$\text{wenn } E_i^{e1} > A_j^{e1} \text{ dann } a_i = |A_j^{e1} - E_i^{e1}| \tag{5.3}$$
$$\text{wenn } E_i^{e2} < A_j^{e2} \text{ dann } e_i = |A_j^{e2} - E_i^{e2}|. \tag{5.4}$$

Für fehlerhafte Erkennungen E_i werden $e_i = 0$ und $a_i = 0$ gesetzt. Die zeitliche Präzision der korrekten und doppelten Erkennungen in einer Sequenz wird über die jeweilige empirische Standardabweichung (S_{Beginn} und S_{Ende}) des Versatzes am Beginn b_i und Ende e_i der erkannten Gesten angegeben:

$$S_{Beginn} = \sqrt{\frac{1}{I_c - 1}\left(\sum_{i \in I_c} (b_i - \overline{b_i})^2\right)} \quad \text{mit } \overline{b_i} = \frac{1}{I_c}\sum_{i \in I_c} b_i \tag{5.5}$$

$$S_{Ende} = \sqrt{\frac{1}{I_c - 1}\left(\sum_{i \in I_c} (e_i - \overline{e_i})^2\right)} \quad \text{mit } \overline{e_i} = \frac{1}{I_c}\sum_{i \in I_c} e_i. \tag{5.6}$$

Zur Bewertung der Gestenerkennung werden diese Maßstäbe für das nun beschriebene Experiment verwendet.

Versuchsaufbau und Durchführung

Das Experiment, das in Kooperation mit Joachim Schmidt und Axel Haasch durchgeführt wurde, sollte zeigen, ob die Auflösung von Objektreferenzen in der Interaktion zwischen einem Menschen und dem Roboter BIRON erfolgreich ist. Dafür wurde das vorgestellte System aus Körperverfolgung, Gestenerkennung und Objektaufmerksamkeit verwendet.

In den Versuchdaten führen fünf Personen (drei Männer, zwei Frauen) drei unterschiedliche Aktionen aus. In erster Linie sollen Zeigegesten auf Objekte, die vor der Versuchperson auf einem Tisch liegen, untersucht werden. Das Zeigen besteht aus den zwei Bewegungen „Zeigen“ und „Rückführen“; zwischen diesen Bewegungen liegt eine Pause. Als weitere Geste wird die Aktion „Winken“ aufgenommen. Das Lenken der Aufmerksamkeit durch Winken ist für die Interaktion mit Robotern in diesem Szenario interessant und bei anderen Untersuchungen mit dem Roboter BIRON häufig von den Versuchspersonen ausgeführt worden. Die Geste gliedert sich in ein Heben des Armes, einer mehrmals wiederholten Winkelbewegung und der Rückführung in die Ruheposition. Die beiden Gesten „Winken“ und „Zeigen“ und ihre Bestandteile wurden auch in 2.3.2 beschrieben und ihr zeitlicher Ablauf visualisiert. Des Weiteren führen die Personen noch eine Aktion aus, bei der sie die Objekte präsentieren. Hierbei führen sie die Hand zuerst zu einem Objekt auf der rechten

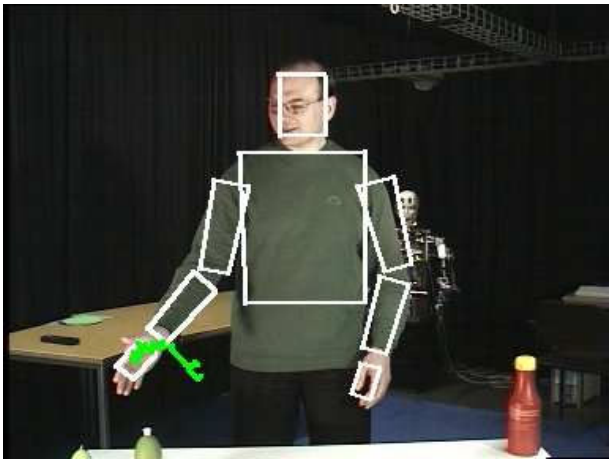


Abbildung 5.17: Ein Bild während des Experiments zur Evaluation der Gestenerkennung. Die Versuchsperson steht vor einem Tisch, auf dem einige Objekte liegen. Das Bild zeigt die Szene aus der Position des Roboters, der gegenüber der Person steht. Die weißen Kästen repräsentieren das Köpermodell der Körperverfolgung. In Grün ist die Trajektorie der rechten Hand dargestellt.

beziehungsweise linken Tischseite und lassen dann ihre Hand über alle Objekte hinweg bis zur gegenüberliegenden Tischseite schweifen. Anschließend wird die Hand wieder in die Ruheposition zurückgeführt. Diese Geste kann zum Beispiel verwendet werden, um dem Roboter zu sagen: „Diese Gegenstände gehören für mein Frühstück auf den Tisch."

In dem Experiment zeigen die Personen auf fünf Objekte, die vor ihnen auf dem Tisch liegen. In einem Datensatz zeigt eine Person jeweils zweimal auf jedes Objekt, führt zweimal die Aktion „Winken" aus und präsentiert einmal die Objekte von links nach rechts und einmal von rechts nach links. Damit werden 14 Aktionen ausgeführt, die insgesamt minimal 32 Bewegungen enthalten. Die Anzahl der Bewegungen variiert, da die Bewegung „Winken" als ein Schwenken des Armes in Höhe des Kopfes nach außen und zurück definiert ist. Wie häufig die Versuchsperson das tat, war ihr freigestellt. Den Aufbau des Versuchs zeigt die Abbildung 5.17. Die Versuchspersonen haben jeweils zweimal den genannten Ablauf an Gesten absolviert.

Auswertung der Annotation

Aus den Annotationen der Sequenzen können nicht nur die Modelle zur Erkennung gebildet werden, sondern auch statistische Aussagen über die Ausführungsdauer der Gesten gewonnen werden. Die Ergebnisse dieser Analyse werden in Tabelle 5.1 dargestellt. Demzufolge werden die meisten Gesten innerhalb von 12 bis 22 Zeitschritten ausgeführt. Bei der gewählten Bildfrequenz von 15 Bildern pro Sekunde entspricht das circa einer Dauer von 0,7 bis 1,5 Sekunden für eine Geste. Der Mittelwert über alle Personen liegt mit 16,79 Bildern knapp über einer Sekunde. Über alle Gesten ergibt sich dabei eine durchschnittliche empirische Standardabweichung in der Größenordnung einer drittel Sekunde (5,05 Zeitschritte).

	Person	I	II	III	IV	V	⌀
Anzahl		146	138	149	178	136	
mittlere Dauer	frames	16,32	17,95	19,31	13,6	21,3	17,7
mittlere Dauer	msec	1088	1197	1287	906	1422	1180
Standardabw.	frames	3,78	4,86	4,31	7,25	5,91	5,22

Tabelle 5.1: Die Ausführungsdauer der Gesten. Für jede Versuchsperson ist die mittlere Ausführungsdauer einer Geste zum einen in der Anzahl der Bilder (frames) und in Millisekunden (msec) aufgetragen. Zum anderen wird die empirische Standardabweichung von diesem Mittelwert angegeben. In der letzten Spalte sind die Durchschnittswerte über alle Personen angegeben.

Datenvorverarbeitung

Zuerst wird aus den Videosequenzen die Körperkonfiguration der Person mit der Personenverfolgung extrahiert und abgespeichert. Bei diesem Prozess kann es vorkommen, dass nicht die gesamte Videosequenz erfolgreich verarbeitet werden kann, sondern das Verfolgen abbricht. Von den insgesamt fast 26 Minuten konnten mit dem Verfahren von Schmidt u. a. [2006] 22,5 Minuten verfolgt werden und standen der Merkmalsextraktion zur Verfügung.

Zum Training werden aus drei der vier Videosequenzen (A-D) alle Bewegungssequenzen mit der gleichen Annotierung gewählt und aus diesen für jede Gestenart ein Modell gebildet. Mit dem in 4.3 vorgestellten Verfahren können aus den Vidoesequenzen einer Person leider keine Modelle gebildet werden. Aufgrund von Rauschen in den Daten der Körperverfolgung und einer variablen Ausführung der Gesten schlägt das automatische Bilden von Modellen in vielen Fällen bei der Versuchsperson V fehl. Es wird keine genügend große Überlappung zwischen verschiedenen Ausführungen einer Bewegung gefunden. Das heißt, dass keine genügend große Ähnlichkeit in den Bewegungssequenzen mit gleicher Annotation gefunden werden kann. Man kann darüber spekulieren, warum die Bewegungen dieser Versuchsperson so unterschiedlich sind. Sei es, dass versucht wurde, die Bewegungen zu deutlich auszuführen, oder dass die Art der Bewegungsausführung schwer zu Verfolgen ist. Die Datenbasis reduziert sich somit auf knapp 19 Minuten, die insgesamt 534 Bewegungen enthalten.

5.2.6 Ergebnisse der Gestenerkennung

An den oben beschriebenen Daten wird ein automatisierter Versuch durchgeführt, der die Leistung der Gestenerkennung testet. Für jede Person wird jede Sequenz gegen ein Training, das auf den anderen drei Sequenzen der Person basiert, durchgeführt. Damit liegen für ein Modell zwischen fünf und 20 Trainingsbewegungen vor. Hier wird ein Vorteil des CTR deutlich, da auch schon kleine Trainingsmengen genügen, generische Modelle zu erzeugen.

Als Merkmal wird die Distanzänderung und die Höhenänderung in dem zylindrischen Koordinatensystem gewählt. Aus den Erkenntnissen von Vorversuchen und dem in 4.4 beschriebenen Versuch konnten diese Merkmale als vielversprechend ausgewählt werden. Die

Richtungsänderung, die in der zweidimensionalen Verfolgung der Hand verwendet wird, muss in diesem Versuch entfallen, da das Rauschen der Körperverfolgung dieses Merkmal ungeeignet macht. Können jedoch die Ergebnisse der Körperverfolgung mit Hilfe von Bewegungsmodellen verbessert werden, kann dieses Merkmal problemlos zur Gestenerkennung hinzugenommen werden.

Als Parameter des CTR wird eine Anzahl von $n = 5000$ Partikel gewählt, die Grenzen der Zeit- und Amplitudenskalierung liegen bei $\alpha_{min} = \rho_{min} = 0.85$ sowie $\alpha_{max} = \rho_{max} = 1.15$, und die Standardabweichung wird auf $\sigma = 0.2$ gesetzt.

Für das „Zeigen“ und „Rückführen“ der Hand werden jeweils zwei Modelle gebildet, da zu erwarten ist, dass ein „Zeigen“ nach vorne sich in der Bewegung von dem „Zeigen“ auf die seitlich liegenden Objekte unterscheidet. Gewertet werden beide Bewegungen jedoch als „Zeigen“.

Ein weiteres Problem in diesem Merkmalsraum besteht darin, dass die Bewegung des Präsentierens zu keinen Wertänderungen führt. Bei dieser Bewegung ist der Arm der Versuchsperson ausgestreckt und wird über die Objekte geschwenkt. Hierbei ändert sich weder der Abstand der Hand zur Schulter, noch die Höhe der Hand. Die Winkeländerung im zylindrischen Koordinatensystem könnte für diese Bewegung das charakteristische Merkmal sein. Doch leider ist dieses durch die Art der Daten der Körperverfolgung bisher nicht verwendbar. Auf einer abstrakteren Ebene kann aber auf die Bewegung des Präsentierens geschlossen werden. Angenommen ein Objekt A liegt auf der linken Seite des Tisches und ein Objekt B auf der gegenüberliegenden rechten Seite, die Gestenerkennung stellt in dieser Situation zum Beispiel zuerst ein „Zeigen“ in Richtung Objekt A fest und anschließend die Bewegung „Rückführen“ aus der Richtung des Objektes B. Daraus kann vermutet werden, dass in der Zwischenzeit eine präsentierende Bewegung ausgeführt wurde.

In der Tabelle 5.2 sind die Erkennungs- und Fehlerergebnisse dargestellt. Da das „Winken“ einer Testperson (IV) nur aus einer Bewegung der Hand besteht, die Körperverfolgung aber nur den Unterarm verfolgt, werden diese Bewegungen von dieser Person aus der Statistik entfernt. Dieser Umstand zeigt leider die Grenzen der trajektorienbasierten Gestenerkennung auf, denn nur die Bewegungen, die die Verfahren zur Gewinnung der Trajektorien registrieren, können ausgewertet werden. Im Anhang sind die detaillierten Tabellen für die einzelnen Personen aufgelistet (siehe Tabelle 8.3 bis 8.10).

Die Tabelle 5.2 zeigt zum einen die Erkennungsrate, die mit 94.56% auf eine gute Erkennung der Gesten hinweist und zum anderen die Fehlerrate. Die Fehlerrate, als Verhältnis aller Fehler zur Anzahl vorliegender Gesten, erlaubt einen tieferen Blick in das Verhalten des Systems.

Mit 11,9% tragen die falsch positiven Erkennungen — die Einfügungen — (e=59) am stärksten zur Fehlerrate bei. Diese Fehler können in dem Robotersystem zu der falschen Annahme führen, dass eine Geste ausgeführt wurde. Eine Kombination aus Sprache und Gestik ist deswegen sinnvoll, da zwei Informationsquellen zur Bestätigung einer Geste verwendet werden können. Im Vergleich zu anderen Verfahren ist aber keine sprachliche Aktivierung notwendig, um den Zeitpunkt der Geste zu bestimmen. Denn die Gestenerkennung kann sowohl den Zeitpunkt des Endes einer Geste als auch die Art der Geste erkennen.

Gesten	n	496
Korrekt	k	469
Doppelt	d	151
Löschung	l	26
Vertauscht	v	1
Eingefügt	e	59
Versatz Beginn	$\sum b$	847
Versatz Ende	$\sum e$	1235
Fehlerrate in %	FR	17.34
Erkennungsrate in %	ER	94.56
Standardabweichung	S_{Beginn}	1,61
Standardabweichung	S_{Ende}	1,66

Tabelle 5.2: Fehlerrate bei der Erkennung von Gesten. Die Standardabweichung ist der Anzahl der Bilder (frames) angegeben, ein frame enspricht 66.7ms.

Bewegung		Zeigen	Zurück	Hoch	Winken	ohne IV	Runter
Anzahl	n	182	187	29	109	71	27
Korrekt	k	167	182	29	65	65	26
Doppelt	d	49	102	0	0	0	0
Löschung	l	15	5	0	42	5	1
Eingefügt	e	24	26	1	2	2	6
Vertauscht	v	0	0	0	2	1	0
Fehlerrate in %	FR	21,43	16,58	3.45	42.20	11.27	17.34
Erkennungsrate in %	ER	91,76	97.33	100.00	59.63	91.55	94.56

Tabelle 5.3: Fehlerrate und Erkennungrate für die einzelnen Bewegungen

Die Zahl der nicht erkannten Gesten (l=26) entspricht dem geringen Fehler von 5.2%. Eine weitere Reduzierung dieses Fehlers durch eine sensiblere Einstellung des CTR würde aber zur Steigerung der falsch positiven Erkennung (e) führen. Die relativ hohe Zahl der doppelt erkannten Gesten ist zum einen in dem probabilistischen Verfahren begründet, da für jeden Zeitschritt die Endwahrscheinlichkeit berechnet wird und über dem Schwellwert für die Erkennung liegen muss. Zum anderen werden sowohl für die „Zeigen“- als auch „Zurück“-Bewegung zwei Modelle verwendet, die manchmal beide erkannt werden. Die doppelte Erkennung ist in sofern kein Fehler, da es sich auch um eine korrekte Erkennung handelt. Für nachfolgende Verfahrensschritte ist es aber sicherlich sinnvoll, mehrere gleiche und zeitlich dicht aufeinander folgende Erkennung zu einer zu verschmelzen.

Positiv zu sehen ist auch, dass die Anzahl der Vertauschungen (v=1) sehr gering ist, da ein solcher Fehler in dem Gesamtsystem schwerer zu beheben ist. Bei der Fusion der Ergebnisse aus unterschiedlichen Modalitäten kann eine Vertauschung zu sich widersprechenden Aussagen führen, bei denen die Gewichtung der Modalitäten schwer abzuschätzen ist.

		Zeigen	Zurück	Hoch	Winken	Runter	⌀
Beginn	$\overline{b_i}$	1,31	0,83	1,86	2,03	0,62	1,33
Beginn	S_{Beginn}	1,80	1,51	1,75	1,80	1,22	1,61
Ende	$\overline{e_i}$	1,82	1,46	1,93	1,86	1,27	1,67
Ende	S_{Ende}	1,81	1,48	1,78	1,46	1,76	1,66

Tabelle 5.4: Zeitliche Genauigkeit der Erkennung für die einzelnen Gesten. Aufgetragen ist der mittlere Versatz für den Beginn $\overline{b_i}$ und das Ende $\overline{e_i}$ der Erkennung zur Annotation. Zusätzlich sind die Standardabweichungen S_{Beginn} und S_{Ende} dieses Versatzes über alle Erkennungen angegeben. Alle Werte sind in frames angegeben, ein frame entspricht 66,7ms.

Einen Einblick in die Erkennung der einzelnen Gesten bietet die Tabelle 5.3. Für das „Winken“ werden die Ergebnisse jeweils mit (Spalte „Winken“) den nicht erkannten Bewegungen der Versuchsperson IV und ohne diese (Spalte *ohne IV*) angegeben. Die Person führt nur ein Winkeln mit der Hand aus, da in dem Körpermodell das Handgelenk nicht modelliert ist, kann keine Bewegung festgestellt werden. Zu sehen ist, dass einige Gesten offensichtlich leichter und fehlerfreier zu erkennen sind als andere. Zum Beispiel ist das Hochführen der Hand (Hoch) zum Winken eine klare und weite Bewegung, die von der jeweiligen Versuchsperson immer in einer sehr ähnlichen Weise ausgeführt wird. Das resultiert in der sehr hohen Erkennungsrate von 100%, aber auch in der geringen Fehlerquote (3,5%). Hingegen treten beim Zeigen auf ein Objekt (Spalte „Zeigen“) größere Varianzen in der Bewegung auf, wenn auf unterschiedliche Objekte gezeigt wird. Doch wie bei allen anderen wird auch bei dieser Bewegung eine Erkennungsrate von über 90% erreicht. Deutlich wird auch, wie stark das bereits erwähnte Nichterkennen der „Winken“-Bewegung bei einer Versuchperson das sonst gute Erkennungsergebnis Bewegung beeinträchtigt.

Bisher wurden Aussagen darüber getroffen, ob eine Bewegung zum entsprechenden Zeitpunkt richtig erkannt wurde. Wie genau diese Erkennung in zeitlicher Hinsicht ist, zeigt die Tabelle 5.4. Über alle Gesten hinweg bleibt die Standardabweichung unterhalb von zwei Zeitschritten, das entspricht etwa 130ms. Laut den Werten der Ausführungsdauer (siehe Tabelle 5.4) tritt die automatische Erkennung folglich meistens während der letzten 10% einer Geste auf. In diesem zeitlichen Bereich verzögert der Mensch seine Handbewegung und erreicht das Ziel seiner Geste. Erfreulich ist, dass auch bei der Bewegung des Winkens die Abweichung so gering ist. Zwischen den einzelnen Schwenkbewegungen wird keine Pause gemacht. Aus der Bildsequenz ist ersichtlich, dass die Hand für nur ein Bild an dem Wendepunkt ist. In der Annotation sind bei diesen Gesten deswegen auch keine Zeitspannen über mehrere Bilder angegeben, sondern das Ende der Geste ist auf ein Bild begrenzt. Die Schwierigkeit für die Erkennung ist, dass exakt zu diesem Zeitpunkt eine Geste aufhört und die nächste beginnt. In den umgebenden Zeitschritten hat die Hand jedoch eine noch relativ hohe Geschwindigkeit beziehungsweise beschleunigt bereits.

Zusammenfassung

Die in dieser Arbeit entwickelte Gestenerkennung ist ein Teil des interaktiven Systems des Roboters BIRON. Über die für die Demonstrationen integrierte Version hinaus, die

auf dem Verfolgen hautfarbener Region aufsetzt, wurde in diesem Unterkapitel eine Evaluation der Gestenerkennung vorgestellt. Die ausgeführten „Zeige“- und „Winken“-Gesten können mit einer hohen Rate erkannt werden. Dabei ist es möglich, auch die Fehlerrate akzeptabel niedrig zu halten. Die Gestenerkennung kann dabei über die gesamte Dauer einer Interaktion laufen, der Zeitpunkte des Abschlusses einer Geste kann dabei auf weniger als zwei Zeitschritte genau ermittelt werden.

Eine manuelle Anpassung einiger Modelle könnte sicherlich die zeitliche Präzision und Erkennungsleistung weiter erhöhen, da bei den Modellen manchmal noch eine kurze Ruhephase nach erfolgter Bewegung mit modelliert wird. Die Evaluation zeigt aber, dass das Erkennen mit automatisch generierten Modellen gewährleistet ist. Dabei werden nur wenige Trainingsdaten für die Modelle benötigt.

Auch eine Erweiterung des Gestenrepertoires ist möglich, solange die Gesten in dem gewählten Merkmalsraum diskriminativ sind. Für das Unterscheiden von sehr ähnlichen Bewegungen muss hingegen der Kontext in die Betrachtung eingezogen werden. Das Thema der Genauigkeit der Zeigerichtung ist in der kooperativ ausgeführten Evaluation dem Bereich der Objektaufmerksamkeit zugeordnet. Die Ergebnisse werden in der Dissertation von Axel Haasch vorgestellt.

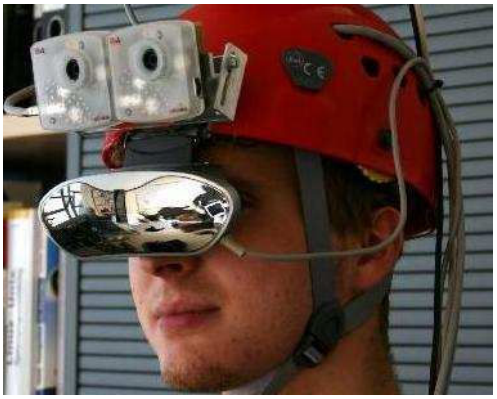
Mit der Kombination aus sprachlichen Hinweisen und einer zeitlichen und räumlichen Gestendetektion ist es möglich, Objektreferenzen aufzulösen und Informationen über Objekte zu sammeln, die dem System bisher unbekannt waren. Dieses Beispiel einer multimodalen Interaktion zwischen einem Roboter und einem Menschen zeigt die Stärken der Kombination auf, da erst die Fusion der unterschiedlichen Modalitäten eine erfolgreiche Kommunikation erlaubt. Des Weiteren wird mit der Verwendung der Körperverfolgung von Schmidt die Adaptivität der entwickelten Gestenerkennung gezeigt.

Die Gestenerkennung ist neben der Spracherkennung und Dialogführung ein wichtiger Baustein für die Interaktion zwischen Menschen und sozial agierenden Robotern. Trotzdem ist noch viel Forschung nötig, um eine sozial angemessene Interaktion zu verwirklichen. Zum einen ist eine Weiterentwicklung der technischen Möglichkeiten notwendig, zum anderen muss aber auch die Anpassung an die Bedürfnisse der Menschen weiter fortschreiten. Damit die Vision der sozial agierenden Roboter eine breite Akzeptanz findet, ist in Zukunft verstärkt ein interdisziplinäres Forschen unter Beteiligung von Psychologen, Designern und Informatikern nötig.

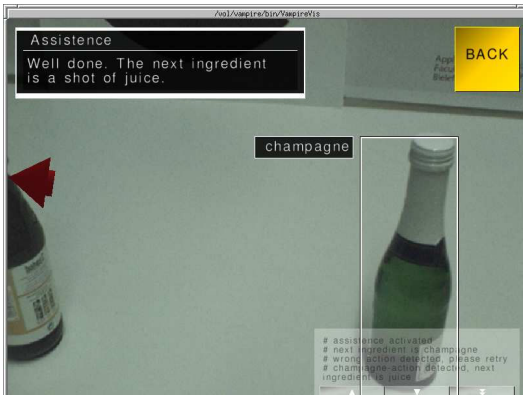
5.3 Gestenerkennung für ein Assistenzsystem

> Nicht nur in der Mensch-Roboter-Interaktion ist das Erkennen menschlicher Aktionen und Gesten wichtig. Auch um Benutzer von Assistenzsystemen instruieren und kontrollieren zu können, muss das computergestützte System Aktionen und Manipulationen erkennen können.

Ein Assistenzsystem ist ein computergestütztes System, das seinem Benutzer bei Handlungen in der realen Welt akustisch und visuell Informationen zu der erfolgreichen Ausführung einer Aufgabe gibt. Diese Systeme können im beruflichen Umfeld unter anderem



(a) Ein Benutzer trägt eine digitale Brille, in der die Bilder der Kameras angezeigt werden. Die Kameras sind oberhalb der Brille befestigt.



(b) Der Blick des Benutzers auf die Szene. Es werden Anweisungen und zusätzliche Informationen eingeblendet.

Abbildung 5.18: Das tragbare Assistenzsystem, das in dem Projekt VAMPIRE entwickelt wurde. Mit diesem System wurden die Experimente zur Aktionerkennung in dieser Arbeit durchgeführt.

bei Reparaturen oder Installationen komplexer Apparate eingesetzt werden. Bei solchen Arbeiten ist viel Erfahrung erforderlich oder die Arbeit muss ansonsten durch umständliches Nachlesen in Handbüchern unterbrochen werden. Ein häufiges Problem dabei ist, dass oft schwierige und lange Arbeitsabfolgen nötig sind oder relevante Bauteile verdeckt oder schlecht sichtbar sind.

Bei den hier betrachteten Assistenzsystemen sieht der Benutzer zum Beispiel durch eine Brille, in der zusätzliche Informationen eingeblendet werden, oder er sieht seine Umgebung in den Displays einer digitalen Brille (engl. Head-Up-Display). In letzterem Fall wird das Blickfeld des Benutzers mit zwei Kameras aufgenommen, die in der Nähe seiner Augen angebracht sind. In der Abbildung 5.18(a) ist dieser Aufbau abgebildet. Das Sichtfeld des Benutzers, angereichert mit einigen Informationen zu den sichtbaren Objekten, zeigt Abbildung 5.18(b). Für den Menschen entsteht eine vermischte Realität (engl. augmented reality).

Ziel dieser Assistenzsysteme ist die Unterstützung des Menschen bei komplexen Aufgaben, indem ihm Zusatzinformationen und Instruktionen gegeben werden. Hierfür ist eine computerbasierte Objekterkennung und Szenenanalyse nötig. Jedoch reicht die reine Perzeption der Szene mit ihren Objekten und deren räumlicher Anordnung nicht aus, da der Benutzer in die Szene eingreift. Er manipuliert Objekte in seinem Gesichtsfeld. Hierbei befolgt er die eingeblendeten Anweisungen oder nimmt eigenständig Veränderungen vor. Die Herausforderung an das System besteht darin, auch diese Manipulationen zu erkennen und richtig im Kontext der Aufgabe zu bewerten. Das System muss unterscheiden, ob dem Benutzer ein Fehler unterlaufen ist oder ob er mehrere von einander unabhängige Bearbeitungsschritte in einer alternativen Reihenfolge ausführt. Wrede u. a. [2006b] geben Ideen, wie das Problem angegangen werden kann.

Eine gemeinsame Vorarbeit zur Erkennung von Objektmanipulationen wird von Fritsch u. a. [2004] vorgestellt. In diesen Experimenten wird mit einer statischen Kamera die Bewegung der rechten Hand des Benutzers ermittelt und für die Erkennung von Aktionen benutzt. Es wird nicht nur das Zeigen auf Objekte untersucht, sondern auch das Manipulieren und Greifen dieser. Durch die Erweiterung des CTR um den Objektkontext (siehe Kapitel 4.4) tragen symbolische Informationen über die Objekte im Bild einen Kontext bei, in dem die Aktionen ausgeführt werden. Für die Erkennung von Manipulationen wird auch das Verfahren zur Detektion und Verfolgung hautfarbener Regionen im Kamerabild verwendet.

Eine Schwierigkeit, die bei Objektmanipulationen besteht, ist es, zu erkennen, welches Objekt manipuliert wird und ob ein Objekt ergriffen wurde. Um einzugrenzen, welches Objekt im Kontext mit einer Handlung steht, wird ein zeitlich dynamischer, räumlicher Objektkontext relativ zur Hand definiert (siehe auch Kapitel 4.4). Ein Problem bei diesem Ansatz besteht darin, dass die agierende Hand nicht von einem Objekt verdeckt sein darf und dass nicht immer aufgelöst werden kann, welches Objekt manipuliert wird. Zum Beispiel ist es mit diesem Ansatz nicht möglich zu erkennen, ob ein Objekt gegriffen oder nur berührt wurde.

Alternativ dazu wird in dem hier vorgestellten System ein Objekt erkannt und daraufhin im Bild verfolgt, das der Benutzer anschaut und bewegt. Der Benutzer kann das Objekt berühren oder greifen und bewegen. Da die Bewegung des manipulierten Objektes aber von der Eigenbewegung des Kopfes des Benutzers und somit der Kamera überlagert ist, muss diese Bewegungskomponente ermittelt und mit der Objektbewegung im Bild verrechnet werden. Von den beiden Kameras des Assistenzsystems wird aus Gründen der Rechenleistung nur das Bild der einen verwendet. Das Berechnen eines Tiefenbildes ist bei dem gewählten Aufbau der Kameras und dem Ziel des mobilen Einsatzes nicht möglich. Die hierfür verwendeten Algorithmen und der Architekturentwurf werden in dem nächsten Abschnitt kurz vorgestellt, bevor die Verwendung der Gestenerkennung in diesem Kontext erläutert wird. Anschließend wird auf die Evaluation des Systems und die erzielten Ergebnisse eingegangen.

5.3.1 Systemkomponenten

Die in diesem Kapitel vorgestellte Anwendung der Gestenerkennung ist eine Komponente eines integrierten kognitiven Assistenzsystems von Wrede u. a. [2006b], das für Objektmanipulationen mit einer Hand entwickelt wurde. Die konkrete Einbindung der Aktionserkennung wurde zusammen mit Hanheide u. a. [2006] entwickelt und beschrieben. Im Folgenden werden die in Graphik 5.19 aufgezeichneten Komponenten des Systems beschrieben. Die Komponenten wurden von Kollegen in der Arbeitsgruppe als Plug-ins für die Bildverarbeitungsplattform *IceWing* umgesetzt.

Objektverfolgung

Zur Erkennung von Objektmanipulationen werden in diesem Ansatz die Trajektorien der Objekte herangezogen. Um diese zu bekommen, gibt es verschiedene Ansätze, deren Vor- und Nachteile Deutsch u. a. [2005] herausgearbeitet haben. In das behandelte System, das Aktionen in Echtzeit erkennen kann, wurden die zwei im Folgenden kurz vorgestellten Verfahren integriert, die Objekte in Echtzeit verfolgen können.

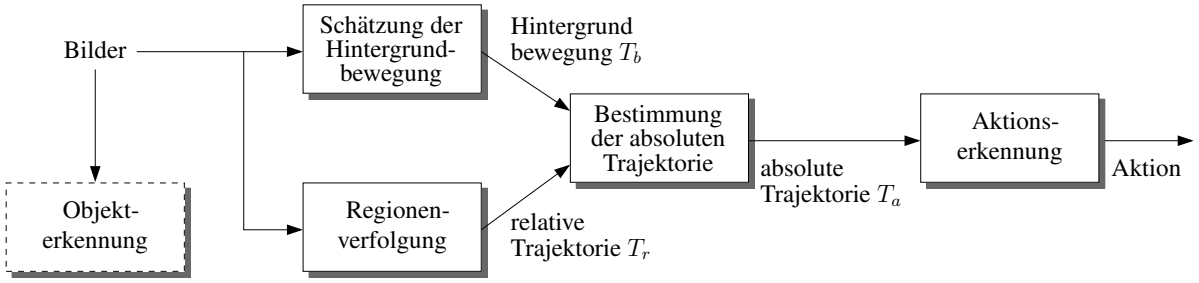


Abbildung 5.19: Skizze der Systemarchitektur für das Erkennen von Objektmanipulationen. (Graphik entnommen aus Hanheide u. a. [2006].)

- **Hyperplane Tracking:**

  Das *Hyperplane Tracking*, es wird detailliert in dem Artikel von Gräßl u. a. [2003] beschrieben, ist eine *template* basierte Technik. Die Bewegung der Objekte wird entsprechend unterschiedlicher Bewegungsmodelle verfolgt. Diese Modelle können einfache Translationen sein oder komplexere Bewegungen wie affine oder projektive Translationen. Das Verfahren nimmt an, dass zum Zeitpunkt t eine Intensitätsänderung einer Pixelgruppe I aus der verfolgten Region $\mathcal{R}$ durch eine Transformation T erklärt werden kann. Das Verfahren benötigt eine Trainingsphase und es wird die Annahme gemacht, dass die Transformation A unabhängig vom aktuellen Zeitpunkt t bestimmt werden kann:

$$\delta T_r \quad = \quad A_h \left(I(F(\mathcal{R}, T_r), t) - I(F(\mathcal{R}, T_r^*), t) \right). \tag{5.7}$$

  Das Verfahren erlaubt das schnelle Verfolgen von Objekten, die auch rotieren dürfen. Doch zeigt sich das Verfahren als relativ sensibel auf Verdeckungen, da diese starke Änderungen in der Intensität bewirken.

- **Kernel-based Tracking:**

  Als ein Verfolgungsverfahren, das robuster gegenüber Verdeckungen ist, wird das *Kernel-based Tracking* von Comaniciu u. a. [2003] verwendet. Dieses Verfahren benötigt außerdem keine zeitaufwendige Trainingsphase. Es werden Farbhistogramme als Merkmale für die Ähnlichkeit des Referenzmodells und des untersuchten Modells benutzt. Über den *Mean-Shift* Algorithmus werden die Bewegungsparameter iterativ verbessert, so das der Abstand iterativ minimiert wird. Das Verfahren hat sich als sehr robust und genau erwiesen, hat aber den Nachteil, dass es auf Translationen beschränkt ist.

Hintergrundbewegung

Eine weitere Herausforderung für die Objektverfolgung in dem vorgestellten Assistenzsystem ist das Ermitteln der Eigenbewegung des Benutzers, um die absolute Trajektorie eines Objektes zu erhalten.

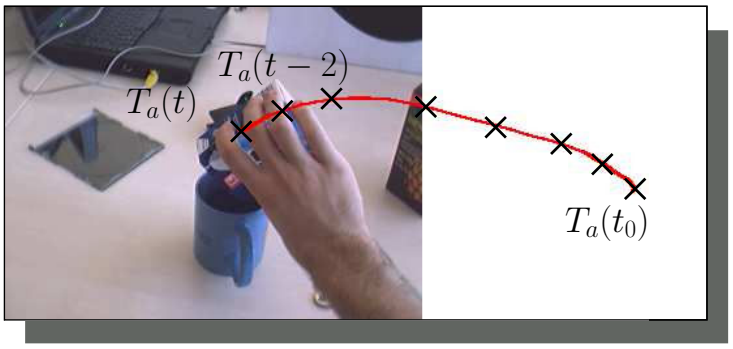


Abbildung 5.20: Die rekonstruierte Trajektorie der Objektbewegung.

Die Projektpartner Zinßer u. a. [2004] verwenden hierfür einen Ansatz von Shi u. Tomasi [1994], der den Hintergrund über ein affines Bewegungsmodell verfolgen kann, indem Stücke (patches) des Bilds verfolgt werden. Für jedes Stück wird die erwartete affine Bewegung berechnet und über eine *Least Median of Squares* (LMedS) Regression (siehe Rousseeuw [1984]) die globale Hintergrundverschiebung T_b ermittelt. Die LMedS ist robust gegenüber Ausreißern und ermöglicht die Berechnung der Bewegung des Bilds, auch wenn Abweichungen von dieser durch bewegte Objekte im Bild vorliegen.

Die Abbildung 5.20 stellt die erwartete Gesamttrajektorie einer „Gießen“ Aktion da. Die komplette Trajektorie ist nicht sichbar im aktuellen Bild $\mathcal{I}_t$, konnte aber aus der Bestimmung der Hintergrundbewegung berechnet werden.

Aktionserkennung

Um die Gestenerkennung auf die rekonstruierte, absolute Trajektorie $T_a(t)$ anzuwenden, wird aus dieser die Merkmalstrajektorie $Z_{Vel}(t)$ berechnet. Diese Repräsentation enthält die Geschwindigkeit in der Bildebene. Wird jedoch das *Hyperplane Tracking* verwendet, kann der Merkmalsvektor um die Winkeländerung der Ausrichtung des Objektes erweitert werden:

$$Z_{VelAng}(t) = \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{d\gamma}{dt} \right). \tag{5.8}$$

5.3.2 Evaluation

Die Leistung des vorgestellten Systems und die Erkennung von Manipulationen wird in dem Szenario eines Assistenzsystems für Barkellner evaluiert, das von Wrede u. a. [2006b] genauer beschrieben wird. Die Aufgabe des Systems ist es, eine Person zu instruieren, einen Cocktail anzufertigen. Das Assistenzsystem kann den Benutzer überwachen und bei Fehlern des Benutzers Hinweise gegen. Dafür müssen die Aktionen des Benutzers klassifiziert werden.

In der Evaluation werden die drei Aktionen „Gießen“, „Bewegen“ und „Schütteln“ betrachtet. Die Aktionen werden von zwei Personen mehrmals ausgeführt. Fehler anderer Module werden bei der Betrachtung der Leistung der Aktionserkennung ausgeblendet. So ist zum Beispiel eine sichere Objektdetektion und -klassifikation gewährleistet. Die Ergebnisse unter Benutzung des *Kernel-based Trackers* sind in Tabelle 5.5 dargestellt, Tabelle 5.6 zeigt die Ergebnisse mit dem *Hyperplane Tracker*.

Aktion	n	k	ER %	v	l	e	FR %
gießen links	9	9	100.0	-	-	-	0.0
gießen rechts	8	7	87.5	-	1	1	25.0
nach links bewegen	8	5	62.5	3	-	-	37.5
nach rechts bewegen	6	4	66.6	1	3	-	66.7
schütteln	9	8	87.5	-	1	-	11.1
Summe	40	33		4	5	1	
ER & FR in%			82.5	10.0	12.5	2.5	25.0

Tabelle 5.5: Ergebnisse, der in dieser Arbeit entwickelten Gestenerkennung, unter Verwendung des *Kernel-based Trackers. (FR: Fehlerrate; ER Erkennungsrate)*

Aktion	n	k	ER %	v	l	e	FR %
gießen links	9	9	100.0	-	-	1	11.1
gießen rechts	8	7	87.5	-	1	-	12.5
nach links bewegen	8	7	87.5	-	1	-	12.5
nach rechts bewegen	6	4	66.7	2	-	1	50.0
schütteln	-	-	-	-	-	-	-
Summe	31	26		2	2	2	
ER & FR in%			83.8	6.4	6.4	6.4	19.1

Tabelle 5.6: Ergebnisse, der in dieser Arbeit entwickelten Gestenerkennung, unter Verwendung des *Hyperplane Trackers. (FR: Fehlerrate; ER: Erkennungsrate)*

Auch in dieser Evaluation wird für jede Aktion die Anzahl der Aktionen n, der korrekt erkannten Gesten k sowie die Anzahl von Vertauschungen v, Einfügungen e und Löschungen l gezählt. Mit diesen Werten kann wiederum die Fehlerrate FR und Erkennungsrate ER berechnet werden.

Bedingt durch die hohe Geschwindigkeit und starke Verdeckung des Objektes durch die Hand des Benutzers, konnte das Objekt bei der Aktion „Schütteln“ nicht robust mit dem *Hyperplane Tracker* verfolgt werden (siehe Tabelle 5.6). Bei der Verwendung des *Kernel-based Trackers* ist es schwer, die Aktionen „Bewegen“ und „Gießen“ zu unterscheiden. Für diese Aktionen ist die Rotation des Objektes ein weiteres wichtiges Merkmal, das die Anzahl der nicht erkannten und vertauschen Aktionen reduziert.

Das Training jedes Modells basiert in diesem Experiment auf nur einer Bewegung einer Person. Dass trotzdem ein gutes Erkennen möglich ist, zeigt, dass die verwendete Gestenerkennung ein robustes Verfahren ist und ein generalisiertes Erkennen ermöglicht.

Zusammenfassung

In diesem Unterkapitel wurden die Resultate der Kooperation mit Hanheide u. a. [2006] vorgestellt. Es wurde ein einzigartiger Ansatz erarbeitet, mit dem es möglich ist, menschliche Aktionen aus der Sicht der ausführenden Person zu erkennen. Die Ergebnisse belegen die Tauglichkeit der Kombination aus einer lokalen Objektverfolgung und der globalen Bestimmung der Kamerabewegung für die bewegungsbasierte Aktionsklassifikation. Es wurden zwei unterschiedliche Verfahren zur Objektverfolgung getestet, die unterschiedliche

Grenzen in ihrer Anwendung haben. Der eine (*Hyperplane Tracker*) erlaubt komplexere Bewegungsmodelle, die das Unterscheiden von ähnlichen Bewegungen ermöglichen. Doch lassen sich mit dem anderen Verfahren (*Kernel-based Tracker*) auch schnelle Bewegungen, wie zum Beispiel das Schütteln eines Objektes verfolgen.

Deswegen ist es sinnvoll in Zukunft die Verfolgungsalgorithmen zu kombinieren. Die Kombination kann zum einen robustere Ergebnisse produzieren und zum anderen mehr Bewegungsmerkmale beinhalten. Eine weitere Idee ist es, sowohl die Bewegung der Hand als auch die Objektbewegungen zu analysieren, um die Intention einer Aktion zu erkennen. Mit der vorgestellten Verwendung der Gestenerkennung ist ein weiterer Schritt auf dem Weg zu intelligenten Assistenzsystemen erreicht. Aber auch auf die Überlegungen der Aktionserkennung im Kontext von sozial interagierenden Robotern hat das entwickelte System Auswirkungen. Denn auch in diesem Szenario ist ein Objekterkennen und -verfolgen sowie die Klassifikation der Objektbewegungen sinnvoll.

6. Aspekte demonstrationsgestützten Lernens

Motionese — eine Methodik des Lernens

Eine Frage, die zum Schluss des Kapitels 4 gestellt wurde, betrifft das Lernen neuer Bewegungen und ihrer Bedeutung. Das bisherige Training von Modellen setzt die aufwendige manuelle Annotation der Bewegungen voraus. Konkret lässt sich die Frage stellen, wie es möglich ist, automatisch zu erkennen, ob eine Person eine interessante Bewegungen ausführt. In einer Interaktion kann so bemerkt werden, wann dem Beobachter oder dem technischen System eine neue Handlung gezeigt wird.

Stellen wir uns die Situation vor, dass ein Besitzer seinem neuen, sozial agierenden Roboter beibringen möchte, wie er Aufgaben im Haushalt ausführen soll. Der Benutzer demonstriert einige Aufgaben und wird unbewusst Wert darauf legen, dass er die relevanten Bewegungen deutlich ausführt. Ist es möglich, diese für den Roboter interessanten Bewegungen zu detektieren? Dabei muss aus den vielen Bewegungen, die der Roboter beobachtet, herausgefiltert werden, wann sich etwas im Bewegungsmuster ändert. Wenn das möglich ist, kann ein Dialog vom Roboter ausgehen, in dem er nachfragt, was für eine Aktion das war und welche Intention mit ihr verbunden ist. Der Mensch muss für das Lernen nicht aufwendig markieren, wann die Aktion, die er dem Roboter beibringen möchte, anfängt und wann sie zu Ende ist. Die Vision ist, dass ein Roboter in einer implizit strukturierten Interaktion neue Aktionen lernen kann und genau weiß, wann er eine Person imitieren soll.

Aus Sicht der Entwicklungspsychologie haben Brand u. a. [2002] dieses Problem aufgegriffen. Sie untersuchen, wie es Kleinkindern möglich ist, die Objektmanipulationen, die sie beobachteten, zu strukturieren und zu erkennen. Die Fähigkeiten eines erwachsenen Menschen in diesem Gebiet beruhen auf vielen Erfahrungen und Beobachtungen. Selbst die Intention einer Manipulation kann aus einer flüchtigen Beobachtung erschlossen werden. Aber die Frage, wie sich Kleinkinder diese Fähigkeiten aneignen, bleibt bestehen.

Ein Bestreben der Forschung, die sich mit der Mensch-Maschine-Interaktion befasst, ist, Effekte und Lernprozesse wie den beschriebenen in technischen Systemen zu realisieren.

In der Robotik ist die Programmierung durch Demonstration ein aktives Forschungsgebiet, das die Kontrolle und Steuerung von Robotern umfasst, aber auch visuelle Aktions- und Gestenerkennung beinhaltet. Nachteil dieser Art des Programmierens von Bewegungen ist aber, dass der Demonstrator seine Bewegungen bewusst an die Möglichkeiten der Maschine anpassen muss. Inspiriert durch das aus der Biologie bekannte Phänomen *Imitation*, ist jedoch in letzter Zeit eine neue Perspektive mit neuen Fragestellungen für die Robotik entstanden. Dautenhahn u. Nehaniv [2002] fassen diese neue Herangehensweise in fünf Fragestellungen für einen Roboter, der mit Menschen interagiert, zusammen:

- Wer wird imitiert?
- Wann wird imitiert?
- Was wird imitiert?
- Wie kann imitiert werden?
- Wie kann ein erfolgreiches Imitieren evaluiert werden?

Mit technischen und mathematischen Methoden wird intensiv an der Lösung dieser Fragen gearbeitet, zum Beispiel von Billard u. Siegwart [2004]. Jedoch werden hauptsächlich die Fragen „Was wird imitiert?“ und „Wie kann imitiert werden?“ aufgegriffen. Diese Fragen sind eng mit Forschungen auf der neuronalen Basis von Imitation verbunden. Hingegen sind die ersten beiden Fragen „Wer wird imitiert?“ und „Wann wird imitiert“ nur im Kontext der sozialen Interaktion von Imitierenden und Imitierten zu beantworten. Das Problem besteht darin, aus den vielen möglichen Beobachtungen die relevanten auszufiltern. Diesem Problem stellen sich Kleinkinder erfolgreich, da es ihnen möglich ist, aus einer großen Anzahl von Reizen neues Wissen zu gewinnen. Insbesondere in der Interaktion zwischen einem Kind und seinem Erzieher sind die sozialen Aspekte relevant, da es dem Erzieher möglich ist, die Aufmerksamkeit des Kindes auf Aktionen zu lenken, die das Kind lernen soll. Kleinkinder können diese Hinweise bemerken und nutzen. Ein weiteres Phänomen, das ein erfolgreiches Lernen in der Interaktion ermöglicht, ist, dass der Demonstrator sich in die Situation des Lernenden versetzt und weiß, welches Wissen diesem fehlt.

Hinweise, die die Aufmerksamkeit lenken sollen, sind nicht nur symbolische Start- oder Stoppmarkierungen, wie zum Beispiel die gesprochenen Befehle „Guck hier“ oder „Das ist ...“. Es konnte vielmehr nachgewiesen werden, dass das Demonstrationsverhalten selbst verändert wird. Diese Beobachtung wird von der Entwicklungspsychologie getragen, in der mit wachsender Unterstützung diskutiert wird, dass Kinder von speziell für sie angepassten Eingaben lernen. Im Bereich der an Kinder gerichteten Sprache ist dieses Phänomen unter dem Namen *Motherese* bekannt geworden. Die Merkmale dieser lehrenden Sprache und einige Beispiele werden unter anderem von Grimm u. Weinert [2002] beschrieben. Nach Dominey u. Dodane [2004] wird hiermit die Aufmerksamkeit des Kindes auf relevante

Aspekte des Sprachsignals gelenkt. Die Charakteristika von *Motherese* sind gut bekannt und werden unter anderem von Breazeal [2002] in technischen Systemen eingesetzt.

Gogate u. a. [2000] untersuchen den zeitlichen Synchronismus zwischen sprachlichen Markierungen und Gesten, also multimodales Motherese. Einen anderen Effekt — *Motionese* genannt — in der Interaktion mit Kindern haben Brand u. a. [2002] bei physischen Objektmanipulationen beobachtet. Sie konnten feststellen, dass Eltern ihre Bewegungen für ihre Kinder anpassen, indem sie Teile der Handlung betonen, Pausen einfügen oder die Handlung wiederholen. Diese Erkenntnisse sind für die Forschung an Robotern in sozialen Umfeldern interessant, da die Anpassung der Bewegungen eines Betreuers oder Erziehers ein starker Hinweis auf relevante Bewegungen sein können. Eine vom Erzieher beziehungsweise Interaktionspartner unbewusste ausgeführte Modifikation seiner Bewegungen kann es möglich machen, mit einem technischen System zu erkennen „wer“ und „wann“ demonstriert.

In diesem Kapitel wird zuerst auf die Arbeiten von Brand u. a. [2002] eingegangen. Anschließend folgen eine technische Analyse der Merkmale von Motionese und die Ergebnisse einer Studie, bei der die entwickelte automatische Analyse eingesetzt wird.

6.1 Motionese: Einordnung und Definition

Brand u. a. [2002] haben in ihrer Studie beobachtet, dass die an Kleinkinder gerichteten Bewegungen von Müttern spezielle charakteristische Merkmale aufweisen, die die Bedeutung und Struktur der Bewegung betonen. In einem Experiment zeigten die Mütter ihren Kindern und einem anderen vertrauten Erwachsenen den Gebrauch neuer Objekte. Die zwei untersuchten Altersgruppen von Kindern waren im Alter von sechs bis acht beziehungsweise elf bis dreizehn Monaten. Die Aktionen der Mütter wurden für die spätere Analyse auf Video aufgenommen. In Brands psychologischen Untersuchungen wurden acht intuitive Kategorien analysiert: Die Weite der Bewegung, Wiederholungen, Nähe zum Partner, Enthusiasmus, Stärke der Interaktion, Punktierung der Bewegung, Vereinfachung der Bewegung und die Geschwindigkeit. Die Merkmale mit einer kurzen Beschreibung sind in Tabelle 6.1 aufgelistet. In einer Videoanalyse wurden die Demonstrationen der Mütter ausgewertet und in jeder Kategorie auf einer Skala (0 bis 4) bewertet.

Trotz der Alters- und Entwicklungsunterschiede zwischen den beiden Kindergruppen konnten in dieser Studie keine signifikanten Unterschiede im Verhalten der Mütter zwischen den Gruppen festgestellt werden. Doch im Vergleich der kindgerichteten und erwachsenengerichten Demonstrationen konnten für alle Merkmale bis auf die Geschwindigkeit signifikante Änderungen festgestellt werden. Diese Änderungen waren bis auf die Merkmale Punktierung und Geschwindigkeit auch signifikant.

Die Beobachtungen von Brand u. a. [2002] werden auch durch andere Untersuchungen untermauert — zum Beispiel von Gergely u. Csibra [2003], die das Schlussfolgern und das Erkennen von Intentionen untersuchten. Sie haben festgestellt, dass gerade Bewegungen das sind, was Kinder erwarten, wenn sie das Greifen nach einem Objekt beobachten. Gergely u. Csibra [2003] schlussfolgern, dass Kinder zielorientierte Bewegungen erwarten. Der offensichtliche Vorteil einer zielorientierten Ausführung einer Bewegung liegt darin, dass schon früh das Ziel auszumachen ist. Auch Woodward u. Sommerville [2000] stellen in

Merkmal	Originalname	Beschreibung
Weite der Bewegung	range of motion	Führt der Demonstrator sehr kurze, begrenzte oder ausladende Bewegungen aus.
Wiederholungen	repetitiveness	Die Bewegungen werden nicht oder sehr häufig wiederholt.
Nähe zum Benutzer	proximitry	Die Demonstration findet meist im Raum des Demonstrators oder im Raum des Beobachters statt. Hierfür wurden Bereiche auf dem Tisch markiert.
Enthusiasmus	enthusiasm	Enthusiasmus, den der Demonstrator gegenüber dem Objekt zeigt. Lachen oder Grinsen, das sich nicht auf das Objekt bezieht, wird nicht beachtet.
Interaktionalität	interactiveness	Bewertung, ob wenig oder viel Interaktion stattfindet.
Punktierung	punctation	Beurteilung, ob die Bewegungen sehr fließend und kontinuierlich ausgeführt werden oder scharf, abrupt und punktiert.
Vereinfachung	simplification	Führt der Demonstrator komplexe Kombinationen aus vielen Bewegungen aus oder dominieren kurze und einfache Bewegungseinheiten.
Geschwindigkeit	rate	Die Bewegungen werden sehr langsam oder sehr schnell ausgeführt.

Tabelle 6.1: Die Merkmale von Motionese nach Brand u. a. [2002] und die Beschreibung dieser Dimensionen.

ihrer Arbeit fest, dass ein Zielwechsel während einer Bewegung für Kleinkinder wichtiger ist als ein Wechsel des Pfades. Dieses haben sie mit der Dauer des Blickkontakts - als Ausdruck der Aufmerksamkeit des Kindes - belegt.

Um jedoch den Effekt Motionese für ein Robotersystem zugänglich und für das Imitationslernen nutzbar zu machen, müssen zuerst technisch messbare Größen gefunden werden. Grundlage hierfür können die eher qualitativen Bewertungen der menschlichen Bewerter in Brands Experimenten sein. Diese Überlegungen führten zu der Entwicklung eines bildbasierten Verfahrens zur Detektion von Motionese, das in dieser Arbeit vorgestellt wird und weitere objektive Hinweise für Brands These der Motionese gibt.

6.2 Technische Analyse von Motionese

Wenn ein Mensch aus einer Videoaufnahme bewerten kann, ob eine Objektmanipulation einem Kleinkind oder einem Erwachsenen gezeigt wird, so sollten diese in der Bildsequenz vorhandenen Informationen auch technisch und quantitativ erfasst werden können. Das umfasst die Aufnahme und Repräsentation der Bewegungen des Demonstrators sowie die Extraktion aussagekräftiger Merkmale und deren Analyse.

Aufnahme der Bewegung

In einer ersten Studie zu Motionese, deren Ergebnisse mit Fritsch u. a. [2005a] veröffentlicht wurden, wird ein bildbasierter Ansatz benutzt, der die Hände aufgrund ihrer Hautfarbe verfolgt. In einer späteren, breiter angelegten Untersuchung wird die modellbasierte

(a) Eine Mutter zeigt ihrem Kind, wie Becher ineinander gestapelt werden.

(b) Eine Mutter bei der Demonstration der Handlung.

Abbildung 6.1: Die Bilder zeigen das Szenario der Versuche zum Motionese-Effekt. Bilder entnommen aus Rohlfing u. a. [2006].

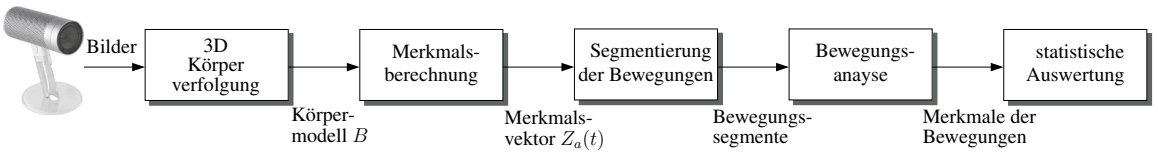


Abbildung 6.2: Konzept der Verarbeitungssequenz für die automatische Analyse von Motionese-Merkmalen in Videosequenzen.

Körperverfolgung von Schmidt u. a. [2006] eingesetzt. Dieses Verfahren, das bereits in 5.1 beschrieben wurde, ermöglicht die Extraktion der dreidimensionalen Körperkonfiguration des beobachteten Menschen aus einem Videobild.

Andere Methoden der Körperverfolgung, zum Beispiel optische Markierungen oder ein externes Skelett, sind für dieses Szenario ungünstig, da sie nicht nur eine natürliche Bewegung behindern könnten, sondern auch die Aufmerksamkeit des Kindes ablenken würden. Die Studie wurde in Kooperation mit Katharina Rohlfing, Jannik Fritsch, Tanja Jungmann und Joachim Schmidt durchgeführt. Die Abbildung 6.1 vermittelt einen Eindruck der Interaktion und der Objektmanipulation, die die Mütter und Väter ausführten. Die Ergebnisse werden von Rohlfing u. a. [2006] beschrieben. Im Rahmen dieser Dissertation wurden die Merkmalsextraktion, die Segmentierung der Bewegung und die Analyse der Daten für die Studie entwickelt und durchgeführt. Die Auswahl und Definition der Merkmale, die in der Bewegungsanalyse berechnet werden, entstanden in Diskussion mit den an der Studie beteiligten Kollegen und ist in den Erfahrungen aus den Vorversuchen begründet.

Der Verarbeitungsprozess ist schematisch in Abbildung 6.2 dargestellt. Die Personen werden im Video verfolgt, die Trajektorien der Hand können in einer zwei- bzw. dreidimensionalen Repräsentation vorliegen. Die Sequenz wird in Bewegungen segmentiert, für einzelne Bewegungen können unterschiedliche Parameter berechnet werden. Diese werden dann quantitativ auf charakteristische Merkmale untersucht. Es findet eine signalnahe Analyse

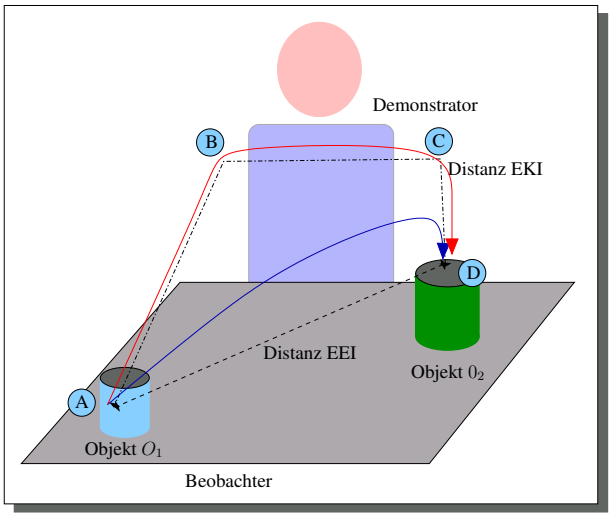


Abbildung 6.3: Schematischer Aufbau des Szenarios zur Untersuchung von Motionese. Vor dem Demonstrator stehen einige Becher, die er ineinander stapeln soll. In Rot ist exemplarisch der Bewegungspfad der EKI über die Punkte A, B, C und D dargestellt. Die rundere Bewegung der EEI ist in Blau aufgezeichnet. Des Weiteren sind die entsprechenden Abstände D_m für die Bewegungsphasen eingetragen. Zu der Aktion „Becherstapeln“ gehören auch die Bewegungen des Greifens und das Zurückführen der Hand.

der Bewegungen statt, in die keine symbolischen Informationen eingehen, sondern nur die Bewegung der Hand betrachtet wird. Der Ablauf der Verarbeitung und die Merkmale werden nun beschrieben.

Körperverfolgung und Repräsentation der Bewegung

Für das Verfolgen des Körpers des Demonstrators wird das schon in Kapitel 5.2.1 (Seite 90) beschriebene Verfahren von Schmidt u. a. [2006] verwendet. Aus der resultierenden Körperkonfiguration wird die Position der rechten Hand extrahiert und die Geschwindigkeit der Hand berechnet. Die Daten werden mit bekannten Methoden wie einem laufenden Mittelwert geglättet.

Segmentierung der Bewegungssequenzen

Damit die gleichen Aktionen aus den unterschiedlichen Demonstrationen verglichen werden können, wurden manuell die zeitlichen Bereiche einer Aktion ermittelt. Als Aktion wird die abgeschlossene Manipulation eines Objektes definiert. Ein Beispiel für Aktion ist das Greifen eines Bechers und das Setzen desselben in einen anderen. Eine Aktion besteht aber aus mehreren einzelnen, voneinander getrennten Bewegungen. Über Schwellwerte für die absolute Geschwindigkeit der Hand sowie minimale Zeiten für Pausen und Bewegungen werden die Aktionen in Bewegungs- und Pausenphasen segmentiert. Als Kriterien dienen die Geschwindigkeit v der Hand sowie die minimale Zeitdauer T_{min}^{Pause} für Pausen beziehungsweise T_{min}^{Bew} für Bewegungen. Abstrahierte Beispiele für die Aktion „Becherstapeln“ sind in der Abbildung 6.3 zu sehen.

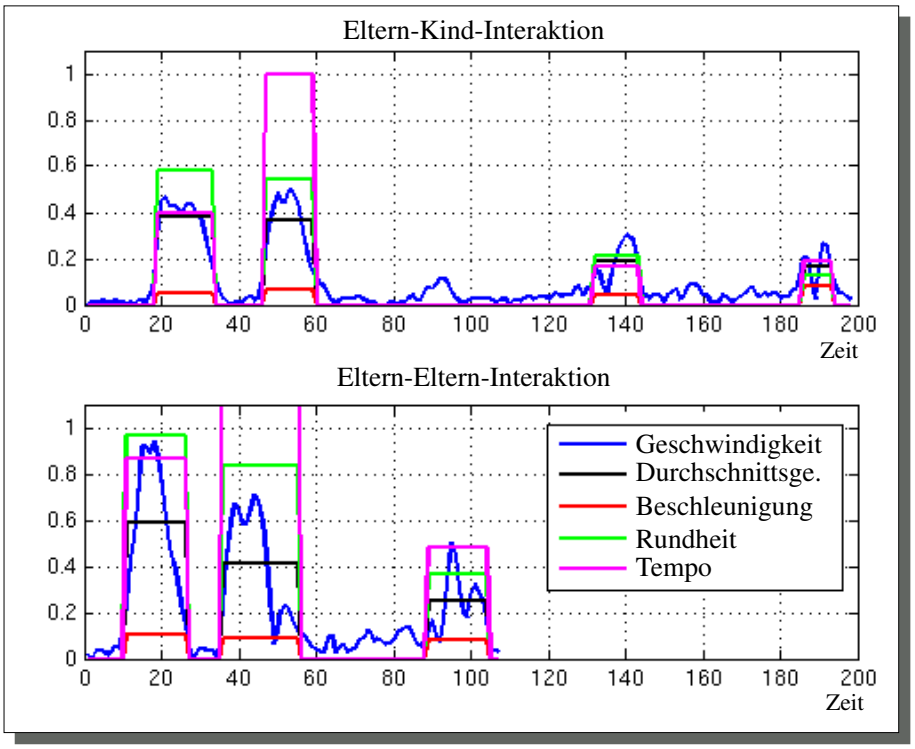


Abbildung 6.4: Merkmale der Eltern-Kind- und Eltern-Eltern-Interaktion. Aufgetragen ist in Blau der Verlauf der absoluten Geschwindigkeit der rechten Hand. Für die automatisch segmentierten Bewegungen sind die Durchschnittswerte jeder Bewegung eingetragen. Die EKI ist 200 Bilder und somit circa 13 Sekunden lang; die EEI dauert dagegen nur circa 7 Sekunden. (Graphik angepasst aus Fritsch u. a. [2005a].)

In der *Eltern-Kind-Interaktion* (EKI) wird das Objekt O_1 ergriffen und über die Punkte A, B, und C zum Zielpunkt D geführt. Dabei wird an den Punkten B und C jeweils eine Pause gemacht, in der die Aufmerksamkeit des Kindes auf das Objekt gelenkt wird. In der *Eltern-Eltern-Interaktion* (EEI) dagegen wird das Objekt direkt ohne eine Pause von A nach D geführt. In Abbildung 6.4 sind die Segmentierungen von zwei Bewegungssequenzen zu sehen. Nur für die Bewegungsphasen sind die Merkmale abgetragen.

Die Merkmale

In einer Pilotstudie wurde ein Erwachsener per Video aufgenommen, wie er seinem zehn Monate alten Kind und einem anderen Erwachsenen die Benutzung eines Spielzeugs zeigte (Stapeln von Holzblöcken auf einen Stab). Die Analyse (siehe Rohlfing u. a. [2004]) weist Unterschiede zwischen der EKI und EEI auf. Diese minimale Studie erlaubt die Vermutung, dass Brands These auch mit mathematisch analytischen Verfahren bestätigt werden kann.

Schwieriger ist es hingegen, die Kategorien aus Brands Studie auf quantitativ messbare Parameter abzubilden. Zum Beispiel ist Brands Punktierung charakterisiert durch die Beschleunigung der Hand und die Anzahl der Pausen. Die Pausen beeinflussen aber auch die

Kategorie „Vereinfachung“, da viele Pausen die Bewegung strukturieren und leichter interpretierbar machen. Andere Parameter aus der Studie von Brand sind mit einem automatischen, bildbasierten Verfahren nur schwer zu erfassen, so zum Beispiel der Enthusiasmus oder die Stärke der Interaktion.

In den Experimenten der Pilotstudie konnten mehrere Parameter identifiziert werden, die aus den Trajektorien der Hände berechnet werden können.

Für jede Bewegungsphase m wird die Pfadlänge L_m und der Abstand D_m der Endposition P_m^{Ende} von der Startposition P_m^{Start} ermittelt (siehe Abbildung 6.3). Die Dauer der Bewegungsphase t_m^{Bew} und der ihr vorausgehenden Pause t_m^{Pause} werden zusätzlich ermittelt. Aus diesen Messwerten werden für jede Bewegung folgende Merkmale berechnet, die die Bewegung charakterisieren:

- Die *durchschnittliche Geschwindigkeit* $\bar{v} = \frac{\sum v}{t_m^{Bew}}$ der Hand. Dieses Merkmal entspricht der Dimension *rate* von Brand.
- Die *durchschnittliche Beschleunigung* $\bar{a} = \frac{\sum \frac{dv}{dt}}{t_m^{Bew}}$ gibt genauer Aufschluss über die Art der Bewegungsgeschwindigkeit.
- Die *Rundheit* (engl. roundness) $r = \frac{L_m}{D_m}$ ist ein Maß für die Form der Bewegung (siehe Abbildung 6.3). Der Pfad der Bewegung wird durch den Abstand der Endposition zum Startpunkt der Bewegung geteilt. Beschreibt die Hand eine gebogene Trajektorie, ergeben sich als Resultat hohe Werte. Wird die Hand gradlinig bewegt, ist der Wert für Rundheit nahe $r = 1$. Es ist zu beobachten, dass der Mensch dazu neigt, sanft gebogene Bewegungen auszuführen. Die Hand wird nicht gradlinig bewegt, sondern in Kurven, die aus einer gleichmäßigen Veränderung der Gelenkwinkel resultieren. Ein Experiment, das diese Beobachtung belegt, hat Sejnowski [1998] durchgeführt.
- Das *Tempo* (engl. pace) $p = \frac{t_m^{Bewegung}}{t_m^{Pause}}$ ist das Verhältnis zwischen Bewegungspausen und Bewegungsphasen. Längere Pausen können ein Mittel sein, eine Bewegungssequenz zu strukturieren.

Beispiele für die berechneten Merkmale werden in Abbildung 6.4 für kindgerichtete und erwachsenengerichtete Demonstrationen abgetragen. Im Folgenden wird die durchgeführte Studie zu Motionese vorgestellt und die erzielten Ergebnisse werden diskutiert.

6.3 Experiment und Ergebnisse

Teilnehmer der detaillierten Studie zu Motionese waren 16 Elternpaare (32 Personen) und ihre Kinder im präverbalen Alter von acht bis elf Monaten. Das Durchschnittsalter der Kinder lag bei 10,2 Monaten (Standardabweichung $\sigma_{Alter} = 1,16$). Die Interaktionspartner sitzen sich an einem Tisch gegenüber, der Demonstrator wird von vorne, nahe der Position des Beobachters, aufgenommen. Abbildung 6.1(b) (Seite 117) zeigt die Sicht der Kamera auf die Szene. Die Eltern wurden gebeten, ihrem Gegenüber die Funktion der Objekte zu zeigen. In der Interaktion mit einem Kind sollten sie dem Kind zeigen und erklären, wie

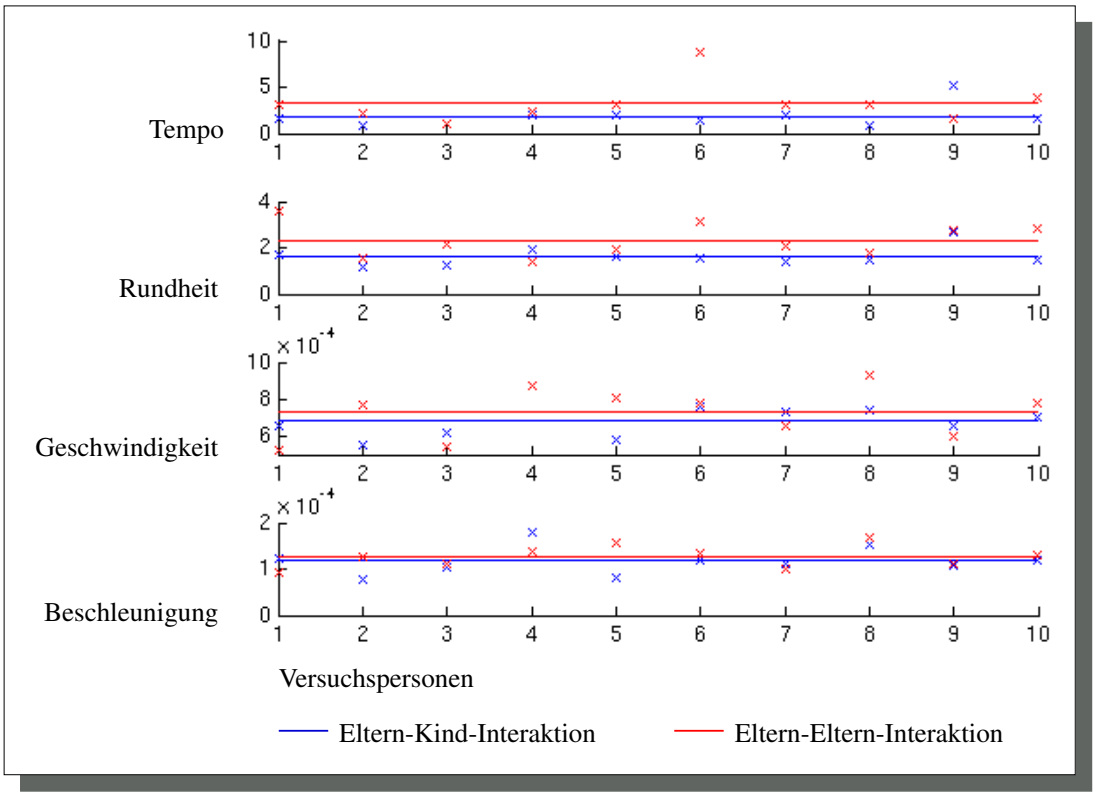
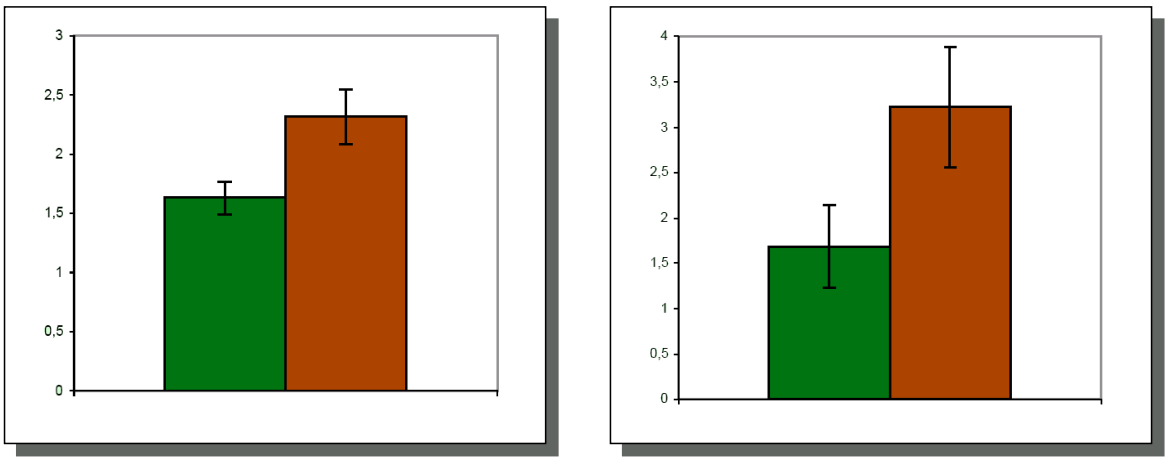


Abbildung 6.5: Mittelwerte über alle Aktionen für alle Versuchpersonen. Für die vier untersuchten Merkmale werden die Mittelwerte für jede Aktion abgetragen, des Weiteren sind die Mittelwerte der EKI (Rot) und EEI (Blau) angegeben. (Graphik angepasst aus Fritsch u. a. [2005a].)

zum Beispiel die Becher ineinander gestapelt werden. Dem Erwachsenen sollten sie zeigen, welche Funktion der Objekte sie dem Kind gezeigt haben.

Von den 32 aufgezeichneten Versuchspersonen wurden 22 wegen Variabilität im Verhalten aus der Analyse ausgeschlossen. Die Demonstration der Benutzung des Kinderspielzeugs wurde von vier Müttern und sechs Vätern ausgeführt. Der Analyseprozess ist zum Beispiel nicht möglich oder die Ergebnisse sind nicht vergleichbar, wenn ein Objekt mit beiden Händen zugleich manipuliert wird, die verfolgte Hand verdeckt ist, die Handlung nicht vergleichbar ausgeführt wird oder die Bildqualität gestört ist. Für die verbleibenden zehn Eltern-Kind-Gruppen betrug das mittlere Alter der Kinder 10,3 Monate (Standardabweichung $\sigma_{Alter} = 1,10$). Eine Gruppe bestand aus jeweils einem Kind sowie einem erwachsenen Demonstrator und Beobachter. Der Demonstrator zeigte die Objektmanipulation einmal seinem Kind und einmal seinem Ehepartner.

Die Unterschiede in der Ausführung der Demonstration in der Eltern-Kind- beziehungsweise Eltern-Eltern-Interaktion sind exemplarisch in Abbildung 6.4 dargestellt. Deutlich zu sehen ist, dass die Aktion in der EKI deutlich länger dauert als in der EEI. Doch ist im direkten Vergleich auch zu sehen, dass das Tempo, also die Abfolge der Bewegungen, höher ist und die Aktion in der EKI in mehr Bewegungen aufgeteilt wird.



(a) Rundheit (Roundness) (b) Tempo (Pace)

Abbildung 6.6: Mittelwerte und Standardabweichung der signifikanten Merkmale Tempo (Pace) und Rundheit (Roundness). In Grün die Werte der EKI und in Rot die der EEI. (Graphik angepasst aus Rohlfing u. a. [2006].)

Die Werte der Merkmale sind für alle Versuchspaare in Abbildung 6.5 abgetragen. Für jede Demonstration (EKI bzw. EEI) einer Versuchperson sind die Merkmale über alle automatisch segmentierten Bewegungsphasen einer Demonstration gemittelt. Der t-Test für abhängige Stichproben ergibt eine signifikante Abhängigkeit vom Interaktionspartner für die Rundheit r (visualisiert in 6.6(a)). Der Mittelwert in der EKI beträgt $\overline{r}^{EKI} = 1,631$ ($\sigma_r^{EKI} = 0,138$) gegenüber $\overline{r}^{EEI} = 2,315$ in der EEI ($\sigma_r^{EEI} = 0,738$; $t(9) = -2,97$, $P < 0,05$)[1]. Der t-Test und sein nonparametrisches Äquivalent, der Wilcoxon-Test, zeigen einen signifikanten Trend für das Tempo (siehe 6.6(b)). In der EKI beträgt der Mittelwert $\overline{p}^{EKI} = 1,685$ ($\sigma_p^{EKI} = 0,458$) gegenüber $\overline{p}^{EEI} = 3,224$ in der EEI ($\sigma_p^{EEI} = 0,666$; $Z = -1,89$, $P < 0,59$).

Bei den Daten aus der dreidimensionalen Körperverfolgung kann kein signifikanter Effekt für die Geschwindigkeit oder Beschleunigung festgestellt werden. Die Analyse zeigt aber eine höhere Gesamtsumme der Einzelaktionen in der EKI als in der EEI.

Im Experiment wurde nicht nur das Videosignal aufgezeichnet, sondern auch das Audiosignal. In dem Artikel von Rohlfing u. a. [2006] werden auch die Ergebnisse der Analyse dieser Daten ausgeführt. Insgesamt lässt sich feststellen, dass auch in den sprachlichen Informationen die Komplexität reduziert wurde, wie es auch die Bewegungsanalyse zeigt. Mit dem t-Test für gepaarte Stichproben kann ein signifikanter Unterschied in der Dauer der Sprache festgestellt werden. In der EKI sprechen die Demonstratoren deutlich weniger. Des Weiteren zeigen die Daten, dass die Varianz der Sprachdauer in der EKI ($\sigma^2_{EKI-Sprache} = 3,144$) deutlich geringer ist als in der EEI ($\sigma^2_{EEI-Sprache} = 9,098$). Diese Erkenntnisse über Unterschiede in der Sprache verlangen noch keine Spracherkennung und können doch Aufschluss über die Interaktionsform geben.

[1] Mittelwert der Rundheit $\overline{r} = \frac{1}{I}\sum_{i=1}^{I} r_i$ für I Versuchpersonen; die Standardabweichung von r ist σ_r. Entsprechend für das Tempo p: $\overline{p}$ und σ_p

6.4 Diskussion und Zusammenfassung

Das vorgestellte Verfahren ist eine algorithmische Herangehensweise an die Frage „Wann soll imitiert werden“, die auf den unterschiedlichen Bewegungen in der Interaktion zwischen Eltern und ihren Kindern basiert. Sowohl über die Bewegungsanalyse als auch die Sprachanalyse konnten signifikante Änderungen im Verhalten der Eltern festgestellt werden.

Die gemeinsame Auswertung der beiden Modalitäten ist der nächste Schritt in der analytischen Untersuchung von Motionese. Aber schon aus den einzelnen Modalitäten werden die Möglichkeiten deutlich, die die sozialen Aspekte bieten, um Objektmanipulationen zu lernen.

Im Gegensatz zu Brands Untersuchung konnte der Effekt Motionese nicht nur bei Müttern sondern allgemein bei Eltern nachgewiesen werden. Die Hypothesen, dass in der EKI längere Pausen zwischen Bewegungsphasen eingefügt werden und dass die Interaktionszeiten deutlich länger waren, konnten bestätigt werden. Ähnlich zu Brands Ergebnissen konnte keine statistische Signifikanz in Geschwindigkeit und Beschleunigung gezeigt werden. Das entspricht der Beobachtung, dass die Bewegungen in der EKI nicht im „Zeitlupentempo“ ausgeführt werden.

Zusammenfassend sei gesagt, dass Eltern komplexe Aktionen für ihre Kinder in Teilbewegungen zerlegen, an deren Ende das Objekt, seine Funktion oder eine Eigenschaft hervorgehoben wird. Auch sind die Bewegungen eher gradlinig auf das Bewegungsziel ausgerichtet, im Gegensatz zur EEI, in der die gebräuchlicheren geschmeidigen Bewegungen bevorzugt ausgeführt werden.

Ein sehr interessanter Versuch, der sich an das Experiment anschließen kann, ist es, die Ergebnisse auf Roboter zu übertragen. Die Vermutung ist, dass ein Mensch einen sozialen und kindlich aussehenden Roboter so behandeln würde wie ein kleines Kind. Entsprechend kann die Hypothese aufgestellt werden, dass der Demonstrator auch seine Bewegung in einer ähnlichen Weise ausführt wie in der EKI. Die Bestätigung der Hypothese stellt hohe Anforderungen an das Design des Roboters, der einen kindlichen Eindruck erwecken soll. Zum anderen muss der Roboter aber auch seine Fähigkeiten dem Benutzer glaubwürdig vermitteln, so dass der Benutzer schon unbewusst seine Demonstration anpasst. Ein Roboter für einen solchen Versuch kann die Barthoc-Plattform der Arbeitsgruppe Angewandte Informatik sein, die von Hackel u. a. [2005] entwickelt wurde (siehe auch Abbildung 1.1, Seite 2).

In den vorherigen Kapiteln dieser Dissertation wurde ein System zur Gestenerkennung entwickelt und seine Tauglichkeit in multimodalen Systemen zur Interaktion zwischen Menschen und Maschinen gezeigt. Auch wenn die vorgestellte Gestenerkennung über das reine Klassifizieren einzelner Gesten oder Bewegungen hinausgeht und die menschlichen Bewegungen in ihrem situativen und symbolischen Kontext gesehen werden, bleibt doch die Intention oft verborgen. In diesem Kapitel konnte aber ein innovativer Ansatz vorgestellt und evaluiert werden, der es ermöglicht, eine spezielle Intention der beobachteten Person auszumachen: Die Intention eines Erwachsenen seinem Kind etwas zu zeigen.

Diese Intention lässt sich nicht lokal an bestimmten Gesten festmachen, sondern verlangt einen globalen und abstrahierten Blick auf die Art und Weise der Bewegungsausführung.

Das Interessante an dem Verfahren zur Detektion von Motionese ist, dass die gleichen Bewegungstrajektorien verwendet werden, wie sie auch schon in der Gestenerkennung eingesetzt wurden. Die vorgestellten Arbeiten können nur ein erster Schritt in das spannende Gebiet des automatischen Interpretierens von menschlichen Gesten und Bewegungen sein, zeigen aber neue Möglichkeiten auf, um die Grenzen heutiger Handlungserkennung und -interpretation zu überwinden.

7. Zusammenfassung und Ausblick

Rechnergestützte Systeme treffen wir immer häufiger im Alltag an und es ist erforderlich, dass diese von vielen Anwendern ohne größere Schwierigkeiten bedient werden können. Diese Interaktion mit einem Rechner kann erfolgreich sein oder erscheint dem Benutzer fremd und unvertraut und kann so Frustrationen und Angstgefühle auslösen.

Auch wenn diese Interaktion eines Menschen mit einer Maschine nicht den Definitionsansprüchen des Begriffs *Kommunikation* genügen, subjektiv erlebt der Benutzer, wie das System mit ihm kommuniziert. Ob diese gefühlte Kommunikation dem Benutzer gefällt oder ihn stört, hängt stark ab von den Fähigkeiten des Systems und wie diese Fähigkeiten dem Benutzer verständlich gemacht werden. Man kann also feststellen, dass Computersysteme eine bidirktionale Kommunikation mit ihren Benutzern eingehen. Das ist eine Interpretation und Verallgemeinerung von Watzlawicks Axiom: „Man kann nicht nicht kommunizieren“ (siehe Watzlawick u. a. [1971]). Es sollte deswegen eine Interaktion angestrebt werden, die sich an den Modalitäten und Gegebenheiten der zwischenmenschlichen Interaktion orientiert. Beobachten wir unsere Gesprächspartner, so fällt auf, dass auch der Bereich der Gestik ein fundamentaler Bestandteil der Kommunikation ist.

Die Vision, die hinter dieser Dissertation steht, ist, dass eine Interaktion zwischen Computern und Menschen nicht besondere Fähigkeiten des Benutzers verlangt, sondern dass die Interaktions- und Kommunikationsfähigkeiten des Menschen von einem Computersystem oder Roboter unterstützt werden. In dieser Arbeit konnte gezeigt werden, wie eine Gestenerkennung als weitere Modalität in die bestehenden Möglichkeiten heutiger Computer integriert werden kann, und so einen Beitrag zur Verwirklichung der angestrebten Vision leistet. Die Funktionalität der Gestenerkennung wurde mit der Integration in multimodale Echtzeitsysteme und Forschungssysteme gezeigt.Die Kombination von Sprache, Gestik, Personenaufmerksamkeit und Objekterkennung ermöglicht, Wissen über komplexe Umgebungen rechnergestützt wahrzunehmen und zu nutzen. Das Erkennen von Deiktika erlaubt zum Beispiel das Auflösen von Referenzen und erkannte Aktionen des Menschen können so in der Spracherkennung und im Dialog verwendet werden.

Die Ideen zum Erkennen von Gesten und von neuen, relevanten Bewegungen sind inspiriert von der zwischenmenschlichen Interaktion, insbesondere von dem enormen Lernpotential

von Kleinkindern. Es wird aber in der Mensch-Maschine-Forschung nicht das Nachbauen dieser Fähigkeiten oder das komplette Nachbilden der menschlichen Intelligenz angestrebt. In der heutigen Robotikforschung muss das, wie Sebastian Thrun anführt, auch nicht das Ziel sein, denn „Heute gehts's darum, clevere Systeme zu bauen“[1].

Eine perfekt funktionierende Einzelkomponente kann ein Beitrag zur Lösung eines Problems sein, aber erst ihre Integration in lauffähige Systeme, die über Laboruntersuchungen hinausgehen, macht sie zu einer wertvollen und sinnvollen Erweiterung. Aus diesem Grund wurde in dieser Arbeit nicht nur Wert auf die Entwicklung eines Moduls zur Gestenerkennung gelegt, sondern ein Schwerpunkt der Arbeit war die Integration dieser in unterschiedliche Systeme. In unterschiedlicher Weise wurde die Gestenerkennung für die multimodale Interaktion mit einem mobilen Roboter verwendet. Die Vision, die mit dem Roboter BIRON verfolgt wird, ist die der sozial agierenden persönlichen Roboter. Diese zukünftigen Roboter sollen mit ihren Besitzern in deren Alltag interagieren können und viele unterschiedliche, alltägliche Aufgaben des Menschen übernehmen können. Aber auch für ein mobiles Assistenzsystem kann die Gestenerkennung verwendet werden und dazu beitragen, dass das System die Aktionen des Menschen wahrnehmen und interpretieren kann. Hervorzuheben ist, dass die genannten Systeme nicht als reine Laborsysteme konzipiert sind, sondern zum Beispiel mit dem Roboter BIRON bereits Experimente in normalen Wohnungen ausgeführt wurden (siehe Abbildung 7.1). In der Konfiguration für das Experiment konnte die Gestenerkennung erfolgreich genutzt werden.

Abbildung 7.1: Während des Experiments in einer Wohnung zeigt eine Person dem Roboter BIRON einen Laptop, der auf dem Tisch im Hintergrund liegt. Aufgrund der Zeigegeste kann der Roboter die referenzierte Region ermitteln und fokussieren. (Bild entnommen aus Wrede u. a. [2006a].)

Betrachtet man die dynamischen Umgebungen, in denen sozial interagierende Roboter eingesetzt werden sollen, wird ersichtlich, dass ein umfassendes Wissen über die Umwelt und das soziale Miteinander der Menschen nicht im Vorhinein implementiert werden kann. Auch das technische Wahrnehmen und Erfassen der Unzahl von Informationen stellt für

[1] aus DIE ZEIT, Nr. 29 vom 13. Juli 2006; Rubrik Wissen, Seite 33; „Blitzrechner ohne Geist“ von Christoph Drösser

solche Systeme eine Herausforderung dar. Es werden immer Unwägbarkeiten, Unsicherheiten und unvollständiges Wissen über die Umgebung als Problem für die technischen Systeme bleiben. Ein Ansatz, mit diesem Problem umzugehen, der in dieser Dissertation am Beispiel der Gestenerkennung herausgearbeitet wurde, sind statistische Verfahren, die es erlauben, trotz der genannten Einschränkungen Schlüsse zu ziehen und Erkenntnisse zu gewinnen. Die entwickelte Gestenerkennung nutzt ein probabilistisches Verfahren, das es ermöglicht, sequenzielle Bewegungsdaten zu segmentieren und zu klassifizieren. Dabei hat sich das *Condensation Trajectory Recognition*(CTR) als ein robuster Algorithmus erwiesen, der aus den vorhandenen Merkmalsdaten zuverlässige Hypothesen für Gesten ermittelt.

Die Komplexität der Umgebungen und der menschlichen Gesten schlägt sich auch darin nieder, dass nicht eine Wahrnehmungsquelle genügt. So reicht es nicht aus, nur die Bewegung des Menschen zu betrachten, nein, die Objekte und die Situation, in der eine Geste beobachtet wird, muss berücksichtigt werden. Die Erweiterung des CTR um einen Objektkontext und situative Bedingungen zeigt für dieses Problem einen neuen Weg auf. Das Konzept ist jedoch noch nicht umfassend genug, um die menschlichen Aktionen und Gesten komplett zu erfassen. Zum Beispiel müssen Vorwissen und Erwartungshaltungen in einem stochastischen Prozess in die Erkennung und Bewertung der Situation einfließen, damit Hypothesen möglich werden.

Eine Aussage, welche Informationsquelle — die Bewegungen oder die Objekte – mehr zur Erkennung beiträgt, lässt sich nicht treffen. Die Bewegung der Hand einer Person ist sehr generisch und lässt selten einen zuverlässigen Schluss auf eine ausgeführte Aktion zu. Aber auch Objekte, die in der Nähe der Hand sind oder Zustandsänderungen in der Umgebung reichen nicht aus, um die Aktion der Person zu ermitteln. Erst die Fusion der unterschiedlichen Wahrnehmungen lässt Hypothesen von Gesten und Aktionen zu. Der Beitrag der Bewegungserkennung ist eine Segmentierung sowie eine räumliche und zeitliche Fokussierung. Die Art der Bewegung kann außerdem die Art der Aktion einschränken. Wissen über die Umgebung, ihre zeitlichen Änderungen und Objekte im Kontext der Bewegungen ermöglichen es, zusammen mit der Bewegungserkennung, Erkenntnisse über die Aktionen und Intentionen der beobachteten Person zu gewinnen. Dass das Prinzip erfolgreich ist, zeigen die im Rahmen dieser Arbeit ausgeführten Experimente.

Ein rechnergestütztes System wie zum Beispiel ein mobiler, sozial interagierender Roboter kann aber nicht alle möglichen Gesten und Handlungen, die er erkennen soll, von Anfang an kennen. Um einen solchen Roboter sinnvoll nutzen zu können, sind Lernkonzepte nötig, die dem Besitzer ermöglichen, seinem Roboter neue Dinge beizubringen. Versuche, Roboter aus der unstrukturierten Vielfalt ihrer Wahrnehmungen lernen zu lassen, müssen aufgrund der Komplexität dieses Problems scheiterten. In dieser Arbeit werden deshalb Experimente durchgeführt, in denen das Verhalten, das Eltern gegenüber ihren Kindern zeigen, beoachtet und untersucht wird. Kleinkinder beweisen uns immer wieder, dass sie mit der Fülle an Eindrücken und Erfahrungen umgehen können und sich so ein Bild von ihrer Umgebung aufbauen. Sie entwickeln schnell Erwartungen von typischen und untypischen Aktionen und können die Intention von Bewegungen erahnen. Mit der Betrachtung der Motionese wurde ein Blick auf dieses Gebiet der Entwicklungspsychologie geworfen. Die These, für die in den ausgeführten Untersuchungen Belege gefunden wurden, sagt aus,

dass Eltern ihr Verhalten und ihre Bewegungen an die Erfahrungs- und Wahrnehmungswelt ihres Kleinkindes anpassen, um diesem das Lernen von Objektmanipulationen zu erleichtern. Diese Anpassungen werden von Eltern unbewusst vorgenommen, lassen sich aber mit technisch analytischen Methoden nachweisen.

Für die soziale Robotik bildet sich aus diesen Erkenntnissen eine visionäre These: Wenn es gelingt, dass Menschen im Umgang mit einem Roboter annehmen, dass dieser ihre Bewegungen wahrnehmen kann und von ihnen wie ein kleines Kind lernen kann, dann werden sie ihre Bewegungsmuster dementsprechend anpassen. Der Roboter kann daraus erkennen, ob der Mensch etwas Neuartiges vormacht oder nur Bewegungen ausführt, die für den Roboter nicht relevant sind. Dieses Konzept bietet eine implizite Möglichkeit, die Menge an Signalen für das technische System zu reduzieren und verlangt von dem Menschen nicht, einen Trainingskodex zu erlernen.

Die Arbeit zeigt Schritte zur Realisierung der Vision, dass Roboter im Alltag des Menschen die Gesten und Aktionen von Personen erkennen, verstehen und erlernen können, auf. Weitere Schritte sind das Erkennen und Interpretieren von komplexeren Gesten und Aktionsabläufen, aber auch die Generalisierung der Erkennung auf bimanuelle Handlungen. Für diese Herausforderung lassen sich unterschiedliche Wege einschlagen: Es kann signalnah versucht werden, die Bewegungen beider Hände einer Person als Ganzes aufzufassen. Des Weiteren ist auch eine symbolische Integration von Einzelerkennung denkbar. Eine weitere Möglichkeit besteht darin, bei der Handlungserkennung einer Hand die andere als Kontext aufzufassen, in der die Bewegung steht.

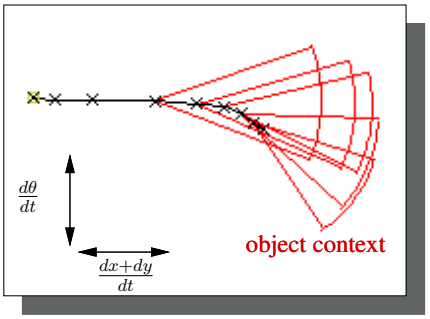
Fragen, die den wissenschaftlichen Komplex dieser Dissertation abstecken, sind, wie sich menschliche Bewegungen und Aktionen erkennen, interpretieren und verstehen lassen. Wie können die mannigfaltigen Möglichkeiten der menschlichen Interaktion und Manipulation erfasst werden? In der Arbeit konnte ein Teil dieses großen Themenbereichs betrachtet und Lösungen für einige Probleme entwickelt werden. Die entwickelte und evaluierte Gestenerkennung liefert in unterschiedlichen Szenarien hohe Erkennungsraten und mit dem analytischen Nachweis des Motionese-Effekts konnte ein Einblick in das kindliche Lernen und Erfassen von Objektmanipulationen gegeben werden.

8. Anhang

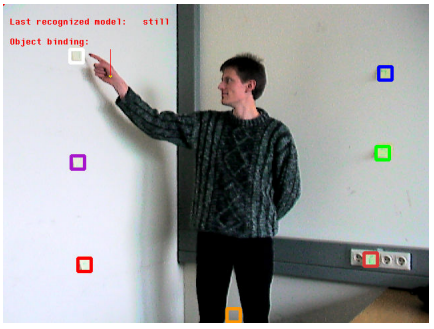
A: Bewegungsmodelle

Time	Kontext Relevanz c_{imp}	*Context Area* Orientierung c_{orient}	Start [W] c_α	Ende [W] c_β	Radius c_r	Kontext Objekttyp c_{type}
t_1	irrelevant					
t_2	notwendig	absolut	30	90	20	Tasse
t_3	notwendig	absolut	30	90	20	Tasse
t_4	notwendig	absolut	30	90	15	Tasse
t_5	möglich	absolut	30	90	15	Tasse

Tabelle 8.1: Tabellarisches Beispiel für den Objektkontext eines Aktionsmodells. Für jeden Zeitschritt wird die Größe (c_r) und Ausrichtung ($c_\alpha, c_\beta, c_{orient}$) der *Context Area* definiert. Außerdem wird die Relevanz des Objektkontextes und der erwartete Objekttype angegeben.



(a) Ein Bewegungsmodell für eine „Zeigen"-Geste mit eingetragener *Context Area*.



(b) Beispielsbild aus den durchgeführten Experimenten zur Zeigegestenerkennung.

Abbildung 8.1: Das Modell einer Zeigegeste und ein Beispiel einer Zeigegeste.

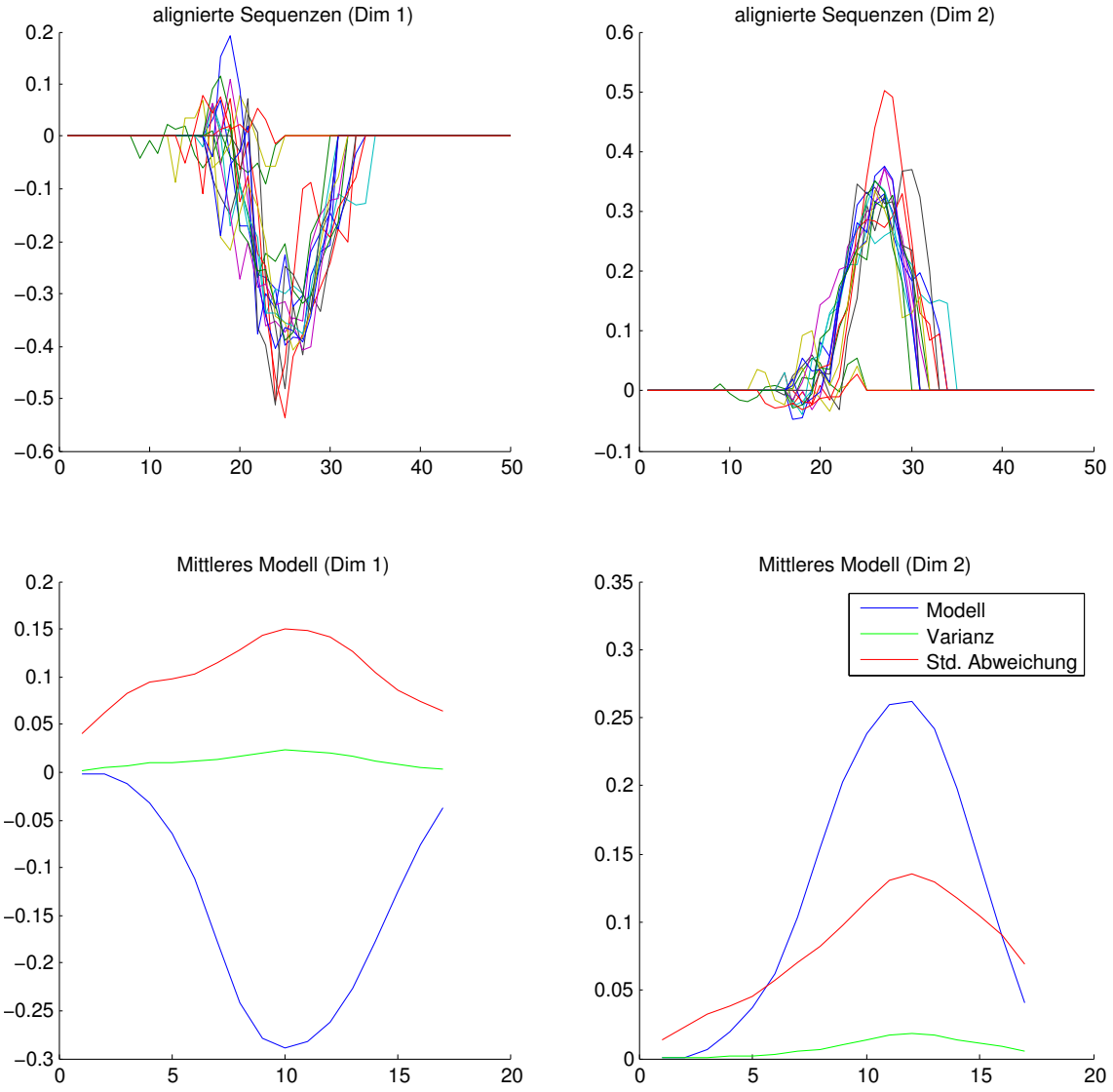


Abbildung 8.2: In der oberen Spalte sind alignierten Bewegungssequenzen einer Zeigegeste zu sehen. Die untere Spalte zeigt das gebildete Modell, mit der berechneten Varianz und Standardabweichung. Auf der linken Seite ist jeweils die Distanzänderung abgetragen und auf rechten die Höhenänderung.

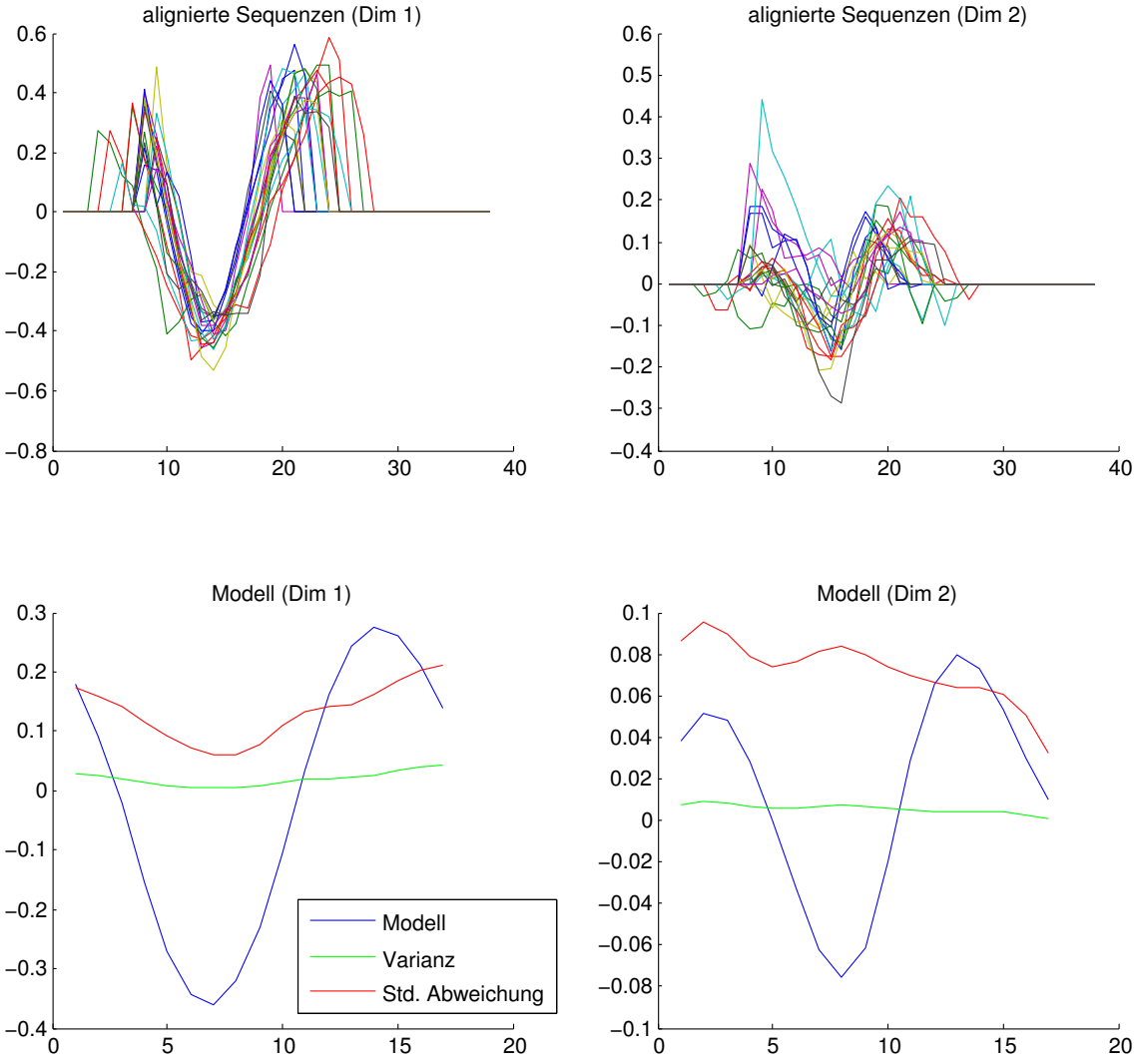
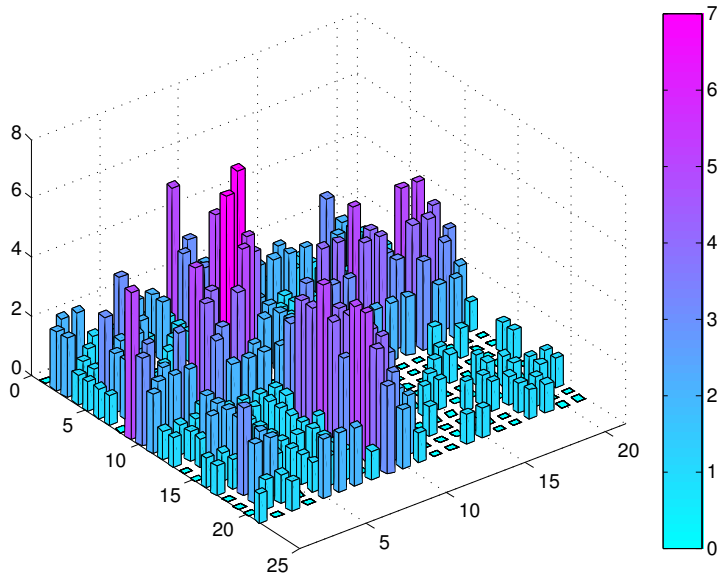
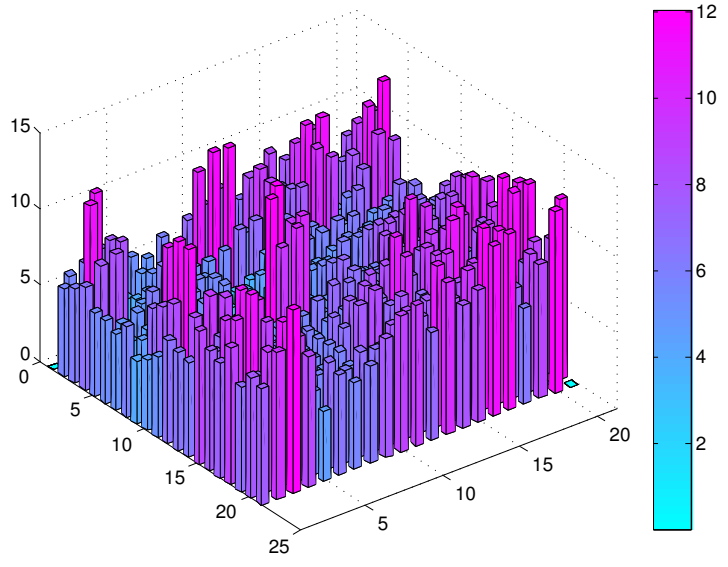


Abbildung 8.3: In der oberen Spalte sind alignierten Bewegungssequenzen einer Winkenbewegung zu sehen. Die untere Spalte zeigt das gebildete Modell, mit der berechneten Varianz und Standardabweichung. Auf der linken Seite ist jeweils die Distanzänderung abgetragen und auf rechten die Höhenänderung.



(a) Die Abbildung zeigt die aus den Koorelation zwischen den Sequenzen berechneten besten Verschiebungen.



(b) Mit den in dieser Graphik dargestellen Bewertungen der jeweiligen Verschiebungen wird die Sequenz ausgewählt, an der die anderen zur Modellbildung ausgerichtet werden.

Abbildung 8.4: Für die Modellbildung werden die besten Verschiebungen zwischen jeweils zwei Sequenzen und deren Bewertung berechnet. Zu sehen ist, dass einige Sequenzen eine geringe Verschiebung zueinander haben (siehe rechts in(a)) und diese auch eine gute Bewertung (siehe rechts in (b)) erhalten. Diese Bewegungssequenzen haben eine hohe Ähnlichkeit untereinander.

B: Mobile, interaktive Robotersysteme

System	Quelle	mobil	Dialog	Gestik	Aktuator	Emotionen, Gesicht
Armar	Asfour u. a. [2001]	j	j	j	j	-
Asimo	komerzielle Plattform	j	?	-	j	-
Barthoc	Hackel u. a. [2005]	-	j	-	j	-
BIRON	Haasch u. a. [2004]	j	j	j	-	-
Care-O-Bot	Graf u. a. [2004]	j	j	-	j	-
CERO	Hüttenrauch u. Eklundh [2002]	j	j	-	-	-
HERMES	Bischoff u. Graefe [2002]	j	j	-	j	-
Horus	Richarz u. a. [2006]	j	j	j	j	-
iCat	van Breemen [2004]	-	-	-	-	j
Jijo	Matsui u. a. [1998]	j	j	-	-	-
Leonardo	Breazeal u. a. [2004]	-	j	j	j	j
LINO	Kröse u. a. [2003]	j	j	-	-	j
Partner Robot		j	-	-	j	-
ROBITA	Matsusaka u. a. [2003]	j	j	-	j	-
Robovie	Ishiguro u. a. [2001]	j	-	-	j	-
SIG	Okuno u. a. [2002]	-	j	-	-	-
Wakamaru	komerziell	j	j	-	j	-

Tabelle 8.2: Eine Übersicht über bestehende Robotersysteme, die für die Interaktion zwischen Menschen und Robotern entwickelt werden. Zu jedem System ist — soweit möglich — eine Quelle angegeben, in der der Roboter näher beschrieben wird.

C: XML-Strukturen

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<HumanModel confidence="0" currentTime="..." depthConfidence="0">
  <FileInfo fileName="[...]/claudia_A_00300.png" numOfImage="300"/>
  <RightHandPos dirX="-0.81" dirY="-0.57" dirZ="-0.08"
        imageDirX="0.95" imageDirY="0.28" imageX="356" imageY="330"
        localXx="-0.54" localXy="0.80" localXz="-0.23" localYx="0.20"
        localYy="-0.14" localYz="-0.96" localZx="-0.81" localZy="-0.57"
        localZz="-0.08" phi="0" psi="0" theta="0" velPhi="0" velPsi="0"
        velTheta="0" velX="0.44" velY="-0.01" velZ="0.24"
        x="2.28" y="-0.11" z="-0.28"/>
   <LeftHandPos ... />
   <HeadPos .../>
   <TorsoPos .../>
   <RotationLimits ... />
   <TranslationLimits ... />
   <CurrentConfig ...>
        ...
     <MarkerPos number="2">
       <marker computeXCoord="0" isMarkerVisible="1" length="405"
           name="torso_rechts" x2d="280" x3d="136.48" x_coord="0"
           y2d="280" y3d="-137.31" y_coord="-200" z3d="2500.20"/>
       <marker ... />
     </MarkerPos>
       ...
     <MarkerPos number="1">
       <marker computeXCoord="0" isMarkerVisible="1" length="200"
           name="fingertip_right" x2d="407" x3d="-243.58" x_coord="0"
           y2d="367" y3d="-354.62" y_coord="0" z3d="2030.03"/>
     </MarkerPos>
     ...
   </CurrentConfig>
</HumanModel>
```

Abbildung 8.5: XML-Struktur der Körperverfolgung. Notiert sind nur Teile, die für die Gestenerkennung und Objektaufmerksamkeit relevant sind. Zahlenwerte wurden gerundet.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- This file is generated by iceWing's XMLFileWriter plugin.-->
<XMLSTREAM creator="XMLFileWriter">
  <CHUNK ... >...</CHUNK>
  <CHUNK fname="" frame="14" generator="CTR1" id="CTRRecog" loop="14"
         nr="14"  timestamp="71989" xmlns:xs="...">
  <MSG xmlns:xs="..." xs:type="trajectory">
    <GENERATOR>CtrXMLImport</GENERATOR>
    <TIMESTAMP>1160050246495</TIMESTAMP>
    <ID />
    <Origin Mod="GrabImg" Timestamp="">[...]claudia_A_00300</Origin>
    <Origin Mod="BodyModelTracker" Timestamp=""/>
    <Origin Mod="CtrXMLImport" Timestamp="..."/>
    <Origin Mod="GR" Timestamp="..."/></ID>
    <NAME>Trajectory</NAME>
    <DATA>
      <FEATURE Framerate="15.00" ID="" ImgNum="300" Reset="0">
        <RAW>
          <SHOULDER X="2.50" Y="0.14" Z="-0.13"/>
          <STEP T="0">
            <RIGHTHANDPOS X="2.28" Y="-0.11" Z="-0.28"/>
            <RIGHTHANDTIP X="2.13" Y="-0.23" Z="-0.41"/>
            <HEADPOS X="2.44" Y="-0.078" Z="0.034"/>
          </STEP>
          <STEP T="-2"><RIGHTHANDPOS .../>
            <RIGHTHANDTIP .../><HEADPOS .../></STEP>
          <STEP T="-1"><RIGHTHANDPOS .../>
            <RIGHTHANDTIP .../><HEADPOS .../></STEP>
        </RAW>
        <VELOCITY A="0.26" X="0.16" Y="0.14" Z="-0.16"/>
        <DIRECTION DDelta="-0.10" DGamma="-0.068"
                   Delta="0.063"  Gamma="-0.404836"/>
        <CYLINDRICAL DDistance="-0.14" DEta="-0.13" DHeight="0.16"
                     Distance="0.33" Eta="0.88" Height="0.17"/>
        <SPHERICAL .../>
      </FEATURE>
      <OBJECT ID=""/>
      <GESTURE>
        <PROGRESS>0.971813</PROGRESS>
        <SCORE>0.000000</SCORE>
        <CHILD>cm_Point125</CHILD>
      </GESTURE>
    </DATA>
    <ANNOTATION Still="0" Untracked="0">Point</ANNOTATION>
</MSG>
</XMLSTREAM>
```

Abbildung 8.6: XML-Struktur der Gestenerkennung mit den Bewegungsmerkmalen, der Erkennung und Annotation. Zahlenwerte wurden gerundet.

D: Ergebnisstabellen

Im Folgenden werden die detaillierten Ergebnisstabellen des Experiments aus Kapitel 5.2 gegeben. Für jede Person sind jeweils die Erkennungs- und Fehlerraten nach Gesten und nach Videosequenz aufgetragen:

Versuchsperson I

Bewegung		Zeigen	Zurück	Hoch	Winken	Runter	Summe
Anzahl	n	48	50	8	26	8	140
Korrekt	k	43	48	8	20	8	127
Doppelt	d	6	35	0	0	0	41
Gelöscht	l	5	2	0	5	0	12
Eingefügt	e	5	17	0	2	0	24
Vertauscht	v	0	0	0	1	0	1
Fehlerrate	FR	0,21	0,38	0	30,77	0	26,43
Erkennungsrate	ER	0,9	0,96	100	76,92	100	90,71

Tabelle 8.3: Erkennungsergebnisse für Versuchperson I. Aufgetragen sind die Fehlerraten und Erkennungsraten für die einzelnen Bewegungen.

	A	B	C	D	Summe
Anzahl	33	35	36	36	140
Korrekt	27	34	32	34	127
Doppelt	11	9	7	14	41
Gelöscht	5	1	4	2	12
Eingefügt	7	5	5	7	24
Vertauscht	1	0	0	0	1

Tabelle 8.4: Erkennungsergebnisse für Versuchperson I. Aufgetragen sind die Fehlerraten und Erkennungsraten für die einzelnen Sequenzen.

Versuchsperson II

Bewegung		Zeigen	Zurück	Hoch	Winken	Runter	Summe
Anzahl	n	46	46	7	17	5	121
Korrekt	k	43	45	7	17	4	116
Doppelt	d	20	22	0	0	0	42
Gelöscht	l	3	1	0	0	1	5
Eingefügt	e	7	4	1	0	6	18
Vertauscht	v	0	0	0	0	0	0
Fehlerrate	FR	0,22	0,11	14,29	0	140	19,01
Erkennungsrate	ER	0,93	0,98	100	100	80	95,87

Tabelle 8.5: Erkennungsergebnisse für Versuchperson II. Aufgetragen sind die Fehlerraten und Erkennungsraten für die einzelnen Bewegungen.

	A	B	C	D	Summe
Anzahl	33	36	24	28	121
Korrekt	31	35	23	27	116
Doppelt	7	10	16	9	42
Gelöscht	2	1	1	1	5
Eingefügt	14	1	1	2	18
Vertauscht	0	0	0	0	0

Tabelle 8.6: Erkennungsergebnisse für Versuchperson II. Aufgetragen sind die Fehlerraten und Erkennungsraten für die einzelnen Sequenzen.

Versuchsperson III

Bewegung		Zeigen	Zurück	Hoch	Winken	Runter	Summe
Anzahl	n	46	50	8	28	8	140
Korrekt	k	41	49	8	28	8	134
Doppelt	d	8	32	0	0	0	40
Gelöscht	l	5	1	0	0	0	6
Eingefügt	e	8	1	0	0	0	9
Vertauscht	v	0	0	0	0	0	0
Fehlerrate	FR	0,28	0,04	0	0	0	10,71
Erkennungsrate	ER	0,89	0,98	100	100	100	95,71

Tabelle 8.7: Erkennungsergebnisse für Versuchperson III. Aufgetragen sind die Fehlerraten und Erkennungsraten für die einzelnen Bewegungen.

	A	B	C	D	Summe
Anzahl	36	36	33	35	140
Korrekt	35	34	32	33	134
Doppelt	13	9	9	9	40
Gelöscht	1	2	1	2	6
Eingefügt	0	1	3	5	9
Vertauscht	0	0	0	0	0

Tabelle 8.8: Erkennungsergebnisse für Versuchperson III. Aufgetragen sind die Fehlerraten und Erkennungsraten für die einzelnen Sequenzen.

Versuchsperson IV

Bewegung		Zeigen	Zurück	Hoch	Winken	Runter	Summe	ohne Winken
Anzahl	n	42	41	6	38	6	133	95
Korrekt	k	40	40	6	0	6	92	92
Doppelt	d	15	13	0	0	0	28	28
Gelöscht	l	2	1	0	37	0	40	3
Eingefügt	e	4	4	0	0	0	8	8
Vertauscht	v	0	0	0	1	0	1	0
Fehlerrate	FR	0,14	0,12	0	100	0	36,84	11,58
Erkennungsrate	ER	0,95	0,98	100	0	100	69,17	96,84

Tabelle 8.9: Erkennungsergebnisse für Versuchperson IV. Aufgetragen sind die Fehlerraten und Erkennungsraten für die einzelnen Bewegungen.

	A	B	C	D	Summe	ohne Winken
Anzahl	36	26	28	43	133	95
Korrekt	28	18	20	26	92	92
Doppelt	3	1	10	14	28	28
Gelöscht	8	8	8	16	40	3
Eingefügt	4	2	1	1	8	8
Vertauscht	0	0	0	1	1	0

Tabelle 8.10: Erkennungsergebnisse für Versuchperson IV. Aufgetragen sind die Fehlerraten und Erkennungsraten für die einzelnen Sequenzen.

E: Schriftenverzeichnis

- Fritsch u. a.(2005a)
  Fritsch, Jannik. ; Hofemann, Nils. ; Rohlfing, Katharina: Detecting 'When to Imitate' in a Social Context with a Human Caregiver. In: *Proceedings of the IEEE International Conference on Robotics & Automations, Workshop on The Social Mechanisms of Robot Programming by Demonstration.* Barcelona, Spain : IEEE, 2005

- Fritsch u. a.(2004)
  Fritsch, Jannik ; Hofemann, Nils ; Sagerer, Gerhard: Combining Sensory and Symbolic Data for Manipulative Gesture Recognition. In: *Proceedings of the International Conference on Pattern Recognition* Bd. 3. Cambridge, UK : IEEE, 2004, S. 930–933

- Haasch u. a.(2005)
  Haasch, Axel ; Hofemann, Nils ; Fritsch, Jannik. ; Sagerer, Gerhard: A Multi-Modal Object Attention System for a Mobile Robot. In: *Proceedings of the IEEE / RSJ International Conference on Intelligent Robots and Systems* IEEE/RSJ, IEEE, 2005, 1499–1504

- Hanheide u. a.(2006)
  Hanheide, Marc ; Hofemann, Nils ; Sagerer, Gerhard: Action Recognition in a Wearable Assistance System. In: *Proceedings of the International Conference on Pattern Recognition* Bd. 2, IEEE, 2006, S. 1254–1257

- Hofemann u. a.(2004)
  Hofemann, Nils ; Fritsch, Jannik ; Sagerer, Gerhard: Recognition of Deictic Gestures with Context. In: Rasmussen, C. E. (Hrsg.) ; Bülthoff, H. H. (Hrsg.) ; Giese, M. A. (Hrsg.) ; Schölkopf, B. (Hrsg.): *Pattern Recognition, 26th DAGM Symposium, Tübingen, Germany. Proceedings* Bd. 3175. Heidelberg, Germany : Springer, 2004 (Lecture Notes in Computer Science), S. 334–341

- Li u. a.(2005b)
  Li, Zhe ; Hofemann, Nils ; Fritsch, Jannik ; Sagerer, Gerhard: Hierarchical Modeling and Recognition of Manipulative Gesture. In: *Proceedings of the IEEE International Conference on Computer Vision, Workshop on Modeling People and Human Interaction.* Beijing, China : IEEE Computer Society, 2005

- Wrede u. a.(2004a)
  Wrede, Britta. ; Haasch, Axel ; Hofemann, Nils ; Hohenner, Sascha. ; Hüwel, Sonja ; Kleinehagenbrock, Marcus. ; Lang, Sebastian ; Li, Shuyin. ; Toptsis, Ioannis. ; Fink, Gernot A. ; Fritsch, Jannik. ; Sagerer, Gerhard.: Research Issues for Designing Robot Companions: BIRON as a Case Study. In: Drews, P. (Hrsg.): *Proceedings of the International Conference on Mechatronics & Robotics* Bd. 4. Aachen, Germany : Eysoldt-Verlag, Aachen, 2004, S. 1491–1496

Literatur

[Accot u. Zhai 2003] Accot, J. ; Zhai, S.: Refining Fitts' law models for bivariate pointing. In: Cockton, Gilbert (Hrsg.) ; Korhonen, Panu (Hrsg.): *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. Fort Lauderdale, Florida, USA : ACM, 2003. – ISBN 1–58113–630–7, 193–200

[Arulampalam u. a. 2002] Arulampalam, S. ; Maskell, S. ; Gordon, N. ; Clapp, T.: A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking. In: *IEEE Transactions on Signal Processing* 50 (2002), Nr. 2, S. 174–188

[Asfour u. a. 2001] Asfour, T. ; Ude, A. ; K.Berns ; Dillmann, R.: Control of armar for the realization of anthropomorphic motion patterns. In: *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, 2001, S. 22–24

[Ayers u. Shah 1998] Ayers, D. ; Shah, M.: Monitoring Human Behavior in an Office Environment. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Santa Barbara, California, USA : IEEE Computer Society, 1998

[Bauckhage u. a. 2005] Bauckhage, Christian ; Hanheide, Marc ; Wrede, Sebastian ; Käster, Thomas ; Pfeiffer, Michael ; Sagerer, Gerhard: Vision Systems with the Human in the Loop. In: *EURASIP Journal on Applied Signal Processing* 2005 (2005), Nr. 14, 2375–2390. http://dx.doi.org/10.1155/ASP.2005.2375. – DOI 10.1155/ASP.2005.2375

[Bax u. a. 2003] Bax, I. ; Bekel, H. ; Heidemann, G.: Recognition of Gestural Object Reference with Auditory Feedback. In: *Proceedings of the International Conference on Neural Networks*. Istanbul, Turkey, 2003, 425–432

[Billard u. Siegwart 2004] *Kapitel* EDITORIAL: Robot learning from demonstration. In: Billard, A. ; Siegwart, R.: *Robotics & Autonomous Systems*. Elsevier, 2004, S. 65–67

[Bischoff u. Graefe 2002] Bischoff, R. ; Graefe, V.: Demonstrating the Humanoid Robot *HERMES* at an Exhibition: A Long-Term Dependability Test. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems; Workshop on Robots at Exhibitions*. Lausanne, Switzerland, 2002

[Bischoff u. Graefe 2004] Bischoff, R. ; Graefe, V.: HERMES - a Versatile Personal Assistant Robot. In: *Proceedings of the IEEE Special Issue on Human Interactive Robots for Psychological Enrichment*, 2004, 1759–1779

[Bishop 1995] Bishop, Christopher M.: *Neural Networks for Pattern Recognition*. Oxford University Press, 1995

[Black u. Jepson 1998] Black, M. J. ; Jepson, A. D.: A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In: Burkhardt, H. (Hrsg.) ; Neumann, B. (Hrsg.): *European Conference on Computer Vision* Bd. 1406. Freiburg, Germany : Springer-Verlag, 1998 (Lecture Notes in Computer Science), S. 909–924

[Blattner 1992] Blattner, Meera M. (Hrsg.): *Multimedia interface design*. New York, USA : ACM Press, 1992 (ACM Press frontier series). – 438 S.

[Bobick u. Ivanov 1998] Bobick, A. ; Ivanov, Y.: Action Recognition Using Probabilistic Parsing. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Santa Barbara, California, USA : IEEE Computer Society, 1998, S. 196–202

[Bobick 1997] Bobick, A. F.: Movement, activity and action: the role of knowledge in the perception of motion. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 352 (1997), Nr. 1358, 1257–1265. http://dx.doi.org/10.1098/rstb.1997.0108. – DOI 10.1098/rstb.1997.0108

[Brand u. a. 2002] Brand, R. J. ; Baldwin, D. A. ; Ashburn., L. A.: Evidence for 'motionese': modifications in mothers' infant-directed action. In: *Developmental Science* 5 (2002), S. 72–83

[Breazeal 2002] Breazeal, C.: *Designing Social Robots*. Cambridge, UK : MIT Press, 2002

[Breazeal u. a. 2004] Breazeal, C. ; Brooks, A. ; Gray, J. ; Hoffman, G. ; Kidd, C. ; Lee, H. ; Lieberman, J. ; Lockerd, A. ; Chilongo, D.: Tutelage and Collaboration for Humanoid Robots. In: *Int. J. of Humanoid Robotics* 1 (2004), Nr. 2, S. 315–348

[van Breemen 2004] Breemen, A. J. N.: Animation Engine for Believable Interactive User-Interface Robots. In: *Proceedings of the IEEE / RSJ International Conference on Intelligent Robots and Systems* IEEE/RSJ, IEEE, 2004

[Bretzner u. a. 2002] Bretzner, L. ; Laptev, I ; Lindeberg, T.: Hand Gesture Recognition using Multi-Scale Colour Features, Hierarchical Models and Particle Filtering. In: *Proceedings of the International Conference on Automatic Face and Gesture Recognition*. Washington DC, USA : IEEE Computer Society, 2002, S. 423–428

[Campbell u. a. 1996] Campbell, L. ; Becker, D. ; Azarbayejani, A. ; Bobick, A. ; Pentland, A.: Invariant Features for 3-D Gesture Recognition. In: *Proceedings of the International Conference on Automatic Face and Gesture Recognition*. Killington, Vermont, USA : IEEE Computer Society, 1996, S. 157

[Clark 2003] *Kapitel* 10. In: Clark, Herbert H.: *Pointing: where language, culture, and cognition meet: Pointing and Placing*. Mahwah : Erlbaum, 2003, S. 243–268

[Comaniciu u. a. 2003] Comaniciu, D. ; Ramesh, V. ; Meer, P.: Kernel-Based Object Tracking. In: *Pattern Analysis and Machine Intelligence* 25 (2003), S. 564–575

[Cédras u. Shah 1995] Cédras, Claudette ; Shah, Mubarak: Motion-based recognition a survey. In: *Proceedings of the International Conference on Image and Vision Computing* 13 (1995), S. 129–155

[Dautenhahn u. Nehaniv 2002] *Kapitel* 1. In: Dautenhahn, K. ; Nehaniv, C. L.: *The Agent-Based Perspective on Imitation*. MIT Press, 2002, S. 1–40

[Deutsch u. a. 2005] Deutsch, B. ; Graessl, Ch. ; Bajramovic, F. ; Denzler, J.: A Comparative Evaluation of Template and Histogram Based 2-d Tracking Algorithms. In: Kropatsch, Walter (Hrsg.) ; Sablatnig, Robert (Hrsg.) ; Hanbury, Allan (Hrsg.): *Proceedings of the 27th DAGM Symposium* Bd. 3663. Heidelberg : Springer, 2005 (Lecture Notes in Computer Science), S. 269

[Dominey u. Dodane 2004] Dominey, Peter F. ; Dodane, Christelle: Indeterminacy in language acquisition: The role of child directed speech and joint attention. In: *Journal of Neurolinguistics* 17 (2004), S. 121–145

[Doucet 2000] Doucet, A.: On Sequential Monte Carlo Sampling Methods for Bayesian Filtering. In: *Satistics and Computing* 10 (2000), S. 197–208

[Doucet u. a. 2001] Doucet, A. (Hrsg.) ; Freitas, N. de (Hrsg.) ; Gordon, N. (Hrsg.): *Sequential Monte Carlo Methods in Practice*. New York, USA : Springer, 2001 (Information Science and Statistics)

[Dreuw u. a. 2005] Dreuw, P. ; Keysers, D. ; Deselaers, T. ; Ney, H.: Gesture Recognition Using Image Comparison Methods. In: Gibet, S. (Hrsg.) ; Courty, N. (Hrsg.) ; Kamp, J.-F. (Hrsg.): *Proceedings of the International Gesture Workshop* Bd. 3881. Berlin Heidelberg : Springer, 2005 (Lecture Notes in Computer Science), S. 124–128

[Efron u. Veen 1972] Efron, D. ; Veen, S. v.: *Gesture, race and culture*. The Hague [u.a.] : Mouton, 1972 (Approaches to semiotics ; 9). – 226 S.

[Ekman u. Friesen 1969] Ekman, P. ; Friesen, W. V.: The repertoire of nonverbal behavior : categories, origins, usage and coding. In: *Semiotica* 1 (1969), S. 50–98. – Mouton de Gruyter (Berlin)

[Fitts 1954] Fitts, P. M.: The information capacity of the human motor system in controlling the amplitude of movement. In: *Journal of Experimental Psychology* 47 (1954), Nr. 6, S. 381–391. – Reprinted in Journal of Experimental Psychology: General, 121(3):262–269, 1992

[Fong u. a. 2003] Fong, T. ; Nourbakhsh, I. ; Dautenhahn, K.: A Survey of Socially Interactive Robots. In: *Robotics and Autonomous Systems* 42 (2003), S. 143–166

[Fritsch 2003] Fritsch, J. N.: *Vision-based Recognition of Gestures with Context*, Universität Bielefeld, Technische Fakultät, Dissertation, 2003. http://bieson.ub.uni-bielefeld.de/volltexte/2003/285/pdf/Fritsch2003.pdf

[Fritsch u. a. 2005a] Fritsch, J. N. ; Hofemann, N. ; Rohlfing, K.: Detecting 'When to Imitate' in a Social Context with a Human Caregiver. In: *Proceedings of the IEEE International Conference on Robotics & Automations, Workshop on The Social Mechanisms of Robot Programming by Demonstration.* Barcelona, Spain : IEEE, 2005

[Fritsch u. a. 2004] Fritsch, J. N. ; Hofemann, N. ; Sagerer, G.: Combining Sensory and Symbolic Data for Manipulative Gesture Recognition. In: *Proceedings of the International Conference on Pattern Recognition (ICPR)* Bd. 3. Cambridge, UK : IEEE, 2004, S. 930–933

[Fritsch u. a. 2005b] Fritsch, J. N. ; Kleinehagenbrock, M. ; Haasch, A. ; Wrede, S. ; Sagerer, G.: A Flexible Infrastructure for the Development of a Robot Companion with Extensible HRI-Capabilities. In: *Proceedings of the IEEE / RSJ International Conference on Intelligent Robots and Systems.* Barcelona, Spain : IEEE, April 2005, S. 3419–3425

[Fritsch u. a. 2003] Fritsch, J. N. ; Kleinehagenbrock, M. ; Lang, S. ; Plötz, T. ; Fink, G. A. ; Sagerer, G.: Multi-Modal Anchoring for Human-Robot-Interaction. In: *Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems* 43 (2003), Nr. 2–3, S. 133–147

[Gavrila 1999] Gavrila, D. M.: The Visual Analysis of Human Movement: A Survey. In: *Computer Vision and Image Understanding* 73 (1999), Nr. 1, S. 82–98

[Gergely u. Csibra 2003] Gergely, G. ; Csibra, G.: Telelogical reasoning in infancy: the naive thery of rational actions. In: *TRENDS in Cognitive Sciences* 7 (2003), Nr. 7, S. 287–291

[Gerkey u. a. 2003] Gerkey, B. P. ; Vaughan, R. T. ; Howard, A.: The Player/Stage Project: Tools for Multi-Robot and Distributed Sensor Systems. In: *Proceedings of the International Conference on Advanced Robotics.* Coimbra, Portugal, 2003, 317–323

[Ghidary u. a. 2002] Ghidary, S.S. ; Nakata, Y. ; Saito, H. ; Hattori, M. ; Takamori, T.: Multi-Modal Interaction of Human and Home Robot in the Context of Room Map Generation. In: *Autonomous Robots* 13 (2002), Nr. 2, S. 169–184

[Gogate u. a. 2000] Gogate, L. J. ; Bahrick, L. E. ; Watson, J. D.: A study of multimodal motherese: the role of temporal synchrony between verbal labels and gestures. In: *Conference on Child Development* Bd. 71, 2000 (4), S. 878–894

[Graf u. a. 2004] Graf, B. ; Hans, M. ; Schraft, R. D.: Care-O-bot II — Development of a Next Generation Robotic Home Assistant. In: *Autonomous Robots* 16 (2004), Nr. 2, 193–205. http://www.care-o-bot.de/english/index.php. – no document

[Gräßl u. a. 2003] Gräßl, C. ; Zinsser, T. ; Niemann, H.: Illumination Insensitive Template Matching with Hyperplanes. In: *Proceedings of the 25th DAGM Symposium* Bd. 2781, 2003 (LNCS), S. 273–280

[Grimm u. Weinert 2002] *Kapitel* 15: Sprachentwicklung. In: Grimm, H. ; Weinert, S.: *Entwicklungsphychologie.* Beltz PVU, 2002, S. 547–549

[Haasch u. a. 2005] Haasch, A. ; Hofemann, N. ; Fritsch, J. ; Sagerer, G.: A Multi-Modal Object Attention System for a Mobile Robot. In: *Proceedings of the IEEE / RSJ International Conference on Intelligent Robots and Systems* IEEE/RSJ, IEEE, 2005, 1499–1504

[Haasch u. a. 2004] Haasch, A. ; Hohenner, S. ; Hüwel, S. ; Kleinehagenbrock, M. ; Lang, S. ; Toptsis, I. ; Fink, G. A. ; Fritsch, J. ; Wrede, B. ; Sagerer, G.: BIRON – The Bielefeld Robot Companion. In: Prassler, E. (Hrsg.) ; Lawitzky, G. (Hrsg.) ; Fiorini, P. (Hrsg.) ; Hägele, M. (Hrsg.): *Proceedings of the International Workshop on Advances in Service Robotics*. Stuttgart, Germany : Fraunhofer IRB Verlag, 2004, S. 27–32

[Hackel u. a. 2005] Hackel, M. ; Schwope, St. ; Fritsch, J. ; Wrede, B. ; Sagerer, G.: A Humanoid Robot Platform Suitable for Studying Embodied Interaction. In: *Proceedings of the IEEE / RSJ International Conference on Intelligent Robots and Systems*. Edmonton, Alberta, Canada : IEEE, August 2005, S. 56–61

[Hanheide u. a. 2006] Hanheide, M. ; Hofemann, N. ; Sagerer, G.: Action Recognition in a Wearable Assistance System. In: *Proceedings of the International Conference on Pattern Recognition (ICPR)* Bd. 2, IEEE, 2006, S. 1254–1257

[Harris u. Wolpert 1998] Harris, C. M. ; Wolpert, D. M.: Signal-dependent noise determinesmotor planning. In: *NATURE* 394 (1998), S. 780–784

[Hegel u. a. 2006] Hegel, F. ; Spexard, T. ; Vogt, T. ; Horstmann, G. ; Wrede, B.: Playing a different imitation game: Interaction with an Empathic Android Robot. In: *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, IEEE, December 2006

[Heidemann u. a. 2004] Heidemann, G. ; Bekel, H. ; Bax, I. ; Saalbach, A.: Hand Gesture Recognition: Self-Organising Maps as a Graphical User Interface for the Partitioning of Large Training Data Sets. In: Kittler, J. (Hrsg.) ; Petrou, M. (Hrsg.) ; Nixon, M. (Hrsg.): *Proceedings of the International Conference on Pattern Recognition (ICPR)* Bd. 4. Cambridge, UK : IEEE CS-Press, 2004, S. 487–490

[Hofemann u. a. 2004] Hofemann, N. ; Fritsch, J. ; Sagerer, G.: Recognition of Deictic Gestures with Context. In: Rasmussen, C. E. (Hrsg.) ; Bülthoff, H. H. (Hrsg.) ; Giese, M. A. (Hrsg.) ; Schölkopf, B. (Hrsg.): *Proceedings of the 26th DAGM Symposium* Bd. 3175. Heidelberg, Germany : Springer, 2004 (Lecture Notes in Computer Science), S. 334–341

[Holzapfel u. a. 2004] Holzapfel, Hartwig ; Nickel, Kai ; Stiefelhagen, Rainer: Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures. In: *Proceedings of the 6th international conference on Multimodal interfaces* SIGCHI: ACM Special Interest Group on Computer-Human Interaction, ACM: Association for Computing Machinery, 2004, S. 175–182

[Hongo u. a. 2000] Hongo, H. ; Ohya, M. ; Yasumoto, M. ; Yamamoto, K.: Face and Hand Gesture Recognition for Human-Computer Interaction. In: *Proceedings of the International Conference on Pattern Recognition (ICPR)* Bd. 2, IEEE, 2000, S. 2921

[Howell u. Buxton 1998] Howell, A. J. ; Buxton, H.: Learning Gestures for Visually Mediated Interaction. In: Carter, J. N. (Hrsg.) ; S., Nixon M. (Hrsg.): *Proceedings of the British Machine Vision Conference*. Southampton, UK : British Machine Vision Association, 1998. – ISBN 1–901725–04–9

[Hüttenrauch u. Eklundh 2002] Hüttenrauch, H. ; Eklundh, K. S.: Fetch-and-carry with CERO: Observations from a long-term user study with a service robot. In: *Proceedings of the IEEE International Workshop on Robot-Human Interactive Communication (ROMAN)*, IEEE Press, 2002, 158–163

[Hüwel u. Wrede 2006] Hüwel, S. ; Wrede, B.: Situated Speech Understanding for Robust Multi-Modal Human-Robot Communication. In: *Proceedings of the International Conference on Computational Linguistics (COLING/ACL)*, ACL Press, 2006. – to appear

[Hüwel u. a. 2006] Hüwel, S. ; Wrede, B. ; Sagerer, G.: Robust speech understanding for multi-modal human-robot communication. In: *Proceedings of the IEEE International Workshop on Robot-Human Interactive Communication (ROMAN)*, IEEE Press, 2006

[Isard u. Blake 1996] Isard, M. ; Blake, A.: Contour tracking by stochastic propagation of conditional density. In: *European Conference on Computer Vision* Bd. 1. Cambridge, UK : Springer, 1996 (Lecture Notes in Computer Science), S. 343–356

[Isard u. Blake 1998a] Isard, M. ; Blake, A.: A mixed-state Condensation tracker with automatic model-switching. In: *Proceedings of the IEEE International Conference on Computer Vision*, 1998, S. 107 – 112

[Isard u. Blake 1998b] Isard, M. ; Blake, A.: Condensation – conditional density propagation for visual tracking. In: *International Journal of Computer Vision* 29 (1998), Nr. 1, S. 5–28

[Ishiguro u. a. 2001] Ishiguro, H. ; Ono, T. ; Imai, M. ; Maeda, T. ; Kanda, T. ; Nakatsu, R.: Robovie: an interactive humanoid robot. In: *Int. J. of Industrial Robot* 28 (2001), Nr. 6, S. 498–503

[Johansson 1973] Johansson, G.: Visual perception of biological motion and a model for its analysis. In: *Perception and Psychophysics* 14 (1973), Nr. 2, S. 201–211

[Kalman 1960] Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems. In: *ASME-Journal of Basic Engineering* 82 (1960), S. 35–45

[Kehl u. Gool 2004] Kehl, R. ; Gool, L. v.: Real-Time Pointing Gesture Recognition for an Immersive Environment. In: *Proceedings of the International Conference on Automatic Face and Gesture Recognition*. Seoul, Korea : IEEE Computer Society, 2004, S. 577–582

[Kendon 1996] Kendon, A.: An Agenda for Gesture Studies. In: *Semiotic Review of Books* 7 (1996), Nr. 3, S. 6–12

[Kendon 2004] Kendon, A.: *Gesture*. Cambridge [u.a.] : Cambridge Univ. Press, 2004

[Kisacanin u. a. 2005] Kisacanin, B. (Hrsg.) ; Pavlovic, V. (Hrsg.) ; Huang, T. S. (Hrsg.): *Real-Time Vision for Human-Computer Interaction*. Springer, 2005 http://www.springer.com/0-387-27697-1

[Kita 2003] Kita, S. (Hrsg.): *Pointing: where language, culture, and cognition meet*. Mahwah : Erlbaum, 2003. – Workshop on Pointing (1997; Nijmegen)

[Klaus 1969] Klaus, G. (Hrsg.): *Wörterbuch der Kybernetik*. Frankfurt am Main [u.a.] : Fischer-Bücherei, 1969 (Fischer-Handbücher)

[Kleinehagenbrock 2005] Kleinehagenbrock, M.: *Interaktive Verhaltenssteuerung für Robot Companions*, Universität Bielefeld, Technische Fakultät, Dissertation, 2005. http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:hbz:361-6849. – URN urn:nbn:de:hbz:361–6849

[Kopp u. a. 2003] Kopp, S. ; Jung, B. ; Lessmann, N. ; Wachsmuth, I.: Artikel: Max - A Multimodal Assistant in Virtual Reality Construction. In: *KI-Küstliche Intelligenz* 4 (2003), S. 11–17. – arenDTap Verlag, Bremen

[Körding u. Wolpert 2004] Körding, K.P. ; Wolpert, D.M.: Bayesian integration in sensorimotor learning. In: *NATURE* 427 (2004), Nr. 427, S. 244 – 247

[Kranstedt u. a. 2006] Kranstedt, Alfred ; Luecking, Andy ; Pfeiffer, Thies ; Rieser, Hannes ; Wachsmuth, Ipke: Deixis: How to Determine Demonstrated Objects Using a Pointing Cone. In: Gibet, S. (Hrsg.) ; Courty, N. (Hrsg.) ; Kamp, J.-F. (Hrsg.): *Proceedings of the International Gesture Workshop* Bd. 2881. Berlin Heidelberg : Springer, 2006 (Lecture Notes in Computer Science), S. 300–311

[Kröse u. a. 2003] Kröse, B. J. A. ; Porta, J. M. ; Breemen, A. J. N. ; Crucq, K. ; Nuttin, M. ; Demeester, E.: Lino, the User-Interface Robot. In: *Proceedings of the First European Symposium on Ambience Intelligence (EUSAI)* Bd. 2875. Eindhoven, The Netherlands : Springer, 2003 (Lecture Notes in Computer Science), S. 264–274

[Kuniyoshi u. a. 1994] Kuniyoshi, Y. ; Inaba, M. ; Inoue, H.: Learning by Watching: Extracting Reusable Task Knowledge from Visual Observation of Human Performance. In: *IEEE Trans. on Robotics and Automation* 10 (1994), Nr. 6, S. 799–822

[Lang 2005] Lang, S.: *Multimodale Aufmerksamkeitssteuerung für einen mobilen Roboter*, Universität Bielefeld, Technische Fakultät, Dissertation, 2005. http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:hbz:361-6792. – URN urn:nbn:de:hbz:361–6792

[Lange u. a. 2003] Lange, C. ; Hermann, T. ; Ritter, H.: Holistic Body Tracking for Gestural Interfaces. In: *Proceedings of International Workshop on Gesture-Based Communication in Human-Computer Interaction* Bd. 2915. Genova, Italy : Springer, 2003 (Lecture Notes in Artificial Intelligence)

[Lee u. a. 2001] Lee, M.-S. ; Weinshall, D. ; Cohen-Solal, E. ; Colmenarez, A. ; Lyons, D.: A Computer Vision System for On-Screen Item Selection by Finger Pointing. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Bd. 1. Hawaii, USA : IEEE Computer Society, 2001, S. 1026

[Li u. Greenspan 2005a] Li, H. ; Greenspan, M.: Continuous Time-Varying Gesture Segmentation by Dynamic TimeWarping of Compound Gesture Models. In: *Proceedings of the International Workshop on Human Activity Recognition and Modelling (HAREM)* Bd. 1. U.K., 2005, S. 35–42

[Li u. Greenspan 2005b] Li, H. ; Greenspan, M.: Multi-Scale Gesture Recognition from Time-Varying Contours. In: *Proceedings of the IEEE International Conference on Computer Vision* Bd. 1, 2005, S. 236–243

[Li u. a. 2005a] Li, S. ; Haasch, A. ; Wrede, B. ; Fritsch, J. ; Sagerer, G.: Human-style interaction with a robot for cooperative learning of scene objects. In: Lazzari, Gianni (Hrsg.) ; Pianesi, Fabio (Hrsg.) ; Crowley, James L. (Hrsg.) ; Mase, Kenji (Hrsg.) ; Oviatt, Sharon L. (Hrsg.): *Proceedings of the 6th international conference on Multimodal interfaces*, ACM, 2005. – ISBN 1–59593–028–0, S. 151–158

[Li u. a. 2005b] Li, Z. ; Hofemann, N. ; Fritsch, J. ; Sagerer, G.: Hierarchical Modeling and Recognition of Manipulative Gesture. In: *Proceedings of the IEEE International Conference on Computer Vision, Workshop on Modeling People and Human Interaction*. Beijing, China : IEEE Computer Society, 2005

[Lömker u. a. 2006] Lömker, F. ; Wrede, S. ; Hanheide, M. ; Fritsch, J.: Building Modular Vision Systems with a Graphical Plugin Environment. In: *Proceedings of the IEEE International Conference on Computer Vision Systems (ICVS)* IEEE, 2006

[Matsui u. a. 1998] Matsui, T. ; Asoh, H. ; Fry, J. ; Motomura, Y. ; Asano, F. ; Kurita, T. ; Hara, I. ; Otsu, N.: Integrated Natural Spoken Dialogue System of Jijo-2 Mobile Robot for Office Services. In: *Proceedings of the IEEE / RSJ International Conference on Intelligent Robots and Systems* Bd. 2. Victoria, Canada, 1998, S. 1278–1283

[Matsusaka u. a. 2003] Matsusaka, Y. ; Tojo, T. ; Kobayashi, T.: Conversation Robot Participating in Group Conversation. In: *IEICE Trans. on Information and System* E86-D (2003), Nr. 1, 26–36. http://tosa.mri.co.jp/sounddb/robot/indexe.htm. – no document

[McKenna u. Gong 1998] McKenna, Stephen J. ; Gong, Shaogang: Gesture recognition for visually mediated interaction using probabilistic event trajectories. In: Carter, John N. (Hrsg.) ; S., Nixon M. (Hrsg.): *Proceedings of the British Machine Vision Conference*. Southampton, UK : British Machine Vision Association, 1998

[McKenna u. Morrison 2004] McKenna, Stephen J. ; Morrison, Kenny: A comparison of skin history and trajectory-based representation schemes for the recognition of user-specified gestures. In: *Pattern Recognition* 37 (2004), Nr. 5, S. 999 – 1009

[Merkl u. Waak 2003] Merkl, R. ; Waak, S.: *Bioinformatik Interaktiv: Algorithmen und Praxis*. Wiley-VCH, 2003

[Moore u. Essa 2002] Moore, Darnell J. ; Essa, Irfan A.: Recognizing Multitasked Activities from Video Using Stochastic Context-Free Grammar. In: *18th National Conference on Artificial Intelligence*. Edmonton, Alberta, Canada : AAAI Press, 2002, S. 770–776

[Nehaniv 2005] Nehaniv, Chrystopher L.: Classifying Types of Gesture and Inferring Intent. In: *Proceedings of the Symposium on Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction* AISB: The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, 2005, S. 74 – 81

[New u. a. 2003] New, J. ; Hasanbelliu, E. ; Aguilar, M.: Facilitating User Interaction with Complex Systems via Hand Gesture Recognition. In: *Proceedings of the Southeastern ACM Conference*. Savannah, GA, USA : Association for Computing Machinery, 2003

[Nickel u. Stiefelhagen 2003a] Nickel, Kai ; Stiefelhagen, Rainer: Pointing gesture recognition based on 3D-tracking of face, hands and head orientation. In: Oviatt, Sharon L. (Hrsg.) ; Darrell, Trevor (Hrsg.) ; Maybury, Mark T. (Hrsg.) ; Wahlster, Wolfgang (Hrsg.): *Proceedings of the 6th international conference on Multimodal interfaces*, 2003, 140–146

[Nickel u. Stiefelhagen 2003b] Nickel, Kai ; Stiefelhagen, Rainer: Real-Time Recognition of 3D-Pointing Gestures for Human-Machine-Interaction. In: Michaelis, Bernd (Hrsg.) ; Krell, Gerald (Hrsg.): *Proceedings of the 25th DAGM Symposium* Bd. 2781, Springer, 2003 (Lecture Notes in Computer Science), 557–565

[Nickel u. Stiefelhagen 2004a] Nickel, Kai ; Stiefelhagen, Rainer: 3D-Tracking of Head and Hands for Pointing Gesture Recognition in a Human-Robot Interaction Scenario. In: *Proceedings of the International Conference on Automatic Face and Gesture Recognition*. Seoul, Korea : IEEE Computer Society, 2004, 565–570

[Nickel u. Stiefelhagen 2004b] Nickel, Kai ; Stiefelhagen, Rainer: Real-Time Person Tracking and Pointing Gesture Recognition for Human-Robot Interaction. In: Sebe, Nicu (Hrsg.) ; Lew, Michael S. (Hrsg.) ; Huang, Thomas S. (Hrsg.): *European Conference on Computer Vision , Workshop on HCI* Bd. 3058. Prague, Czech Republic : Springer, 2004 (Lecture Notes in Computer Science), 28–38

[Niemann 1983] Niemann, Heinrich: *Klassifikation von Mustern*. Berlin : Springer-Verlag, 1983 http://www5.informatik.uni-erlangen.de/Personen/niemann/klassifikation-von-mustern/m00links.html?language=en. – 2. revised version, online available

[Nölker u. Ritter 2002] Nölker, C. ; Ritter, H.: Visual Recognition of Continuous Hand Postures. In: *IEEE Trans. Neural Networks* 13 (2002), Nr. 4, S. 983–994

[Nölker 2000] Nölker, Claudia: *GREFIT: Ein System zur Visuellen Erkennung von Handposturen*, Technische Fakultät, Universität Bielefeld, Diss., 2000. http://bieson.ub.uni-bielefeld.de/volltexte/2003/341/

[Nöth 2000] Nöth, Winfried: *Handbuch der Semiotik*. 2., vollst. neu bearb. und erw. Aufl. Stuttgart [u.a.] : Metzler, 2000. – XII, 667 S. : Ill., graph. Darst.

[Okuno u. a. 2002] Okuno, H. G. ; Nakadai, K. ; Kitano, H.: Social Interaction of Humanoid Robot Based on Audio-Visual Tracking. In: *Proceedings of the International*

Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE), 2002, 725–734

[Pavlovic u. a. 1997] Pavlovic, Vladimir ; Sharma, Rajeev ; Huang, Thomas S.: Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 19 (1997), Nr. 7, S. 677–695

[Pineau u. a. 2003] Pineau, J. ; Montemerlo, M. ; Pollack, M. ; Roy, N. ; Thrun, S.: Towards Robotic Assistants in Nursing Homes: Challenges and Results. In: *Robotics and Autonomous Systems* 42 (2003), Nr. 3-4, S. 271–281

[Press u. a. 1992] Press, William H. ; Teukolsky, Saul A. ; Vetterling, William T. ; Flannery, Brian P.: *Numerical Recipes in C, The Art of Scientific Computing*. Second Edition. Cambridge University Press, 1992 http://library.lanl.gov/numerical/bookcpdf.html. – ISBN 0-521-43108-5

[Psarrou u. a. 2002] Psarrou, Alexandra ; Gong, Shaogang ; Walter, Michael: Recognition of human gestures and behaviour based on motion trajectories. In: *Image and Vision Computing* 20 (2002), S. 349 – 358

[Rao u. a. 2002] Rao, Cen ; Yilmaz, Alper ; Shah, Mubarak: View-Invariant Representation and Recognition of Actions. In: *Int. J. of Computer Vision archive* 50 (2002), S. 203 – 226

[Richarz u. a. 2006] Richarz, J. ; Martin, C. ; Scheidig, A. ; Gross, H.-M.: There you go! - Estimating Pointing Gestures in Monocular Images for Mobile Robot Control. In: *Proceedings of the IEEE International Workshop on Robot-Human Interactive Communication (ROMAN)*. Hatfield, UK, 2006, 546–551

[Rohlfing u. a. 2004] Rohlfing, Katharina ; Fritsch, Jannik ; Wrede, Britta: Learning to Manipulate Objects: A Quantitative Evaluation of Motionese. In: *Proceedings of the International Conference on Development and Learning (ICDL)*. La Jolla, CA, 2004, S. 27. – ISBN 0-615-12704-5

[Rohlfing u. a. 2006] Rohlfing, Katharina J. ; Fritsch, Jannik ; Wrede, Britta ; Jungmann, Tanja: How can multimodal cues from child-directed interaction reduce learning complexity in robotos? In: *Advanced Robotics* 20 (2006), Nr. 10, S. 1183–1199

[Rousseeuw 1984] Rousseeuw, P.J.: Least Median of Squares Regression. In: *Journal of the American Statistical Association* 79 (1984), S. 871–880

[Rowley u. Rehg 1997] Rowley, Henry ; Rehg, Jim: Analyzing Articulated Motion Using Expectation-Maximization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997, S. 935 – 941

[Schmidt u. a. 2006] Schmidt, J. ; Kwolek, B. ; Fritsch, J.: Kernel Particle Filter for Real-Time 3D Body Tracking in Monocular Color Images. In: *Proceedings of the International Conference on Automatic Face and Gesture Recognition*. Southampton, UK, 2006

[Schraft u. a. 2004] Schraft, R.D. ; Helms, E. ; Hans, M. ; Thiemermann, S.: Man-Machine-Interaction and Co-Operation for Mobile and Assisting Robots. In: *Proceedings of the International Symposium on Engineering of Intelligent Systems (EIS)*, 2004. – Care-o-Bot

[Sejnowski 1998] Sejnowski, Terrence J.: Making Smooth Moves. In: *NATURE* 394 (1998), S. 725 – 726

[Shi u. Tomasi 1994] Shi, Jianbo ; Tomasi, Carlo: Good Features to Track. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994

[Sidenbladh u. a. 2000] Sidenbladh, H. ; Black, M. J. ; Fleet, D.J.: Stochastic tracking of 3D human figures using 2D image motion. In: Vernon, D. (Hrsg.): *European Conference on Computer Vision* Bd. 1843. Dublin, Ireland : Springer, 2000 (Lecture Notes in Computer Science), S. 702–718

[Spexard u. a. 2006] Spexard, Thorsten ; Li, Shuyin ; Wrede, Britta ; Fritsch, Jannik ; Sagerer, Gerhard ; Booij, Olaf ; Zivkovic, Zoran ; Terwijn, Bas ; Kröse, Ben: BIRON, where are you? - Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization. In: *Proceedings of the IEEE / RSJ International Conference on Intelligent Robots and Systems*, IEEE, October 2006

[Starner u. Pentland 1995] Starner, T. ; Pentland, A.: Visual recognition of american sign language using hidden markov models. In: *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 1995

[Stenger u. a. 2001] Stenger, B. ; Mendonca, P.R.S. ; Cipolla, R.: Model-based 3D Tracking of an Articulated Hand. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001

[Sudderth u. a. 2004] Sudderth, Erik B. ; Mandel, Michael I. ; Freeman, William T. ; Willsky, Alan S.: Visual Hand Tracking Using Nonparametric Belief Propagation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Workshop* Bd. 12. Washington, DC, USA : IEEE Computer Society, 2004, S. 189

[Thompson u. a. 1994] Thompson, J.D. ; Higgins, D.G. ; Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. In: *Nucleic Acids Research* 22 (1994), S. 4673?–4680

[Tomasello 2006] *Kapitel* Why don't apes point? In: Tomasello, Michael: *Roots of Human Sociality: Culture, Cognition and Interaction*. Berg Publishers Ltd, 2006

[Turk 2005a] *Kapitel* TRV4HCI: A Historical Overview. In: [**Kisacanin u. a., 2005**], 3–13

[Turk 2005b] *Kapitel* Multimodal Human Computer Interaction. In: [**Kisacanin u. a., 2005**], 269–283

[Viola u. Jones 2001] Viola, P. ; Jones, M.: Robust Real-time Object Detection. In: *Proceedings of the IEEE International Workshop on Statistical and Computational Theories of Vision*. Vancouver, Canada : IEEE, 2001

[Watzlawick u. a. 1971] Watzlawick, Paul ; Beavin, Janet H. ; Jackson, Don D.: *Menschliche Kommunikation*. Bern, Schweiz : Huber, 1971. http://dx.doi.org/3-456-30389-0. http://dx.doi.org/3-456-30389-0

[Woodward u. Sommerville 2000] Woodward, Amanda L. ; Sommerville, J. A.: Twelve-month-old infants interpret action in context. In: *Psychological Science* 11 (2000), Nr. 1, S. 73–75

[Wrede u. a. 2004a] Wrede, B. ; Haasch, A. ; Hofemann, N. ; Hohenner, S. ; Hüwel, S. ; Kleinehagenbrock, M. ; Lang, S. ; Li, S. ; Toptsis, I. ; Fink, G. A. ; Fritsch, J. ; Sagerer, G.: Research Issues for Designing Robot Companions: BIRON as a Case Study. In: Drews, P. (Hrsg.): *Proceedings of the International Conference on Mechatronics & Robotics* Bd. 4. Aachen, Germany : Eysoldt-Verlag, Aachen, 2004, S. 1491–1496

[Wrede u. a. 2006a] Wrede, Britta ; Kleinehagenbrock, Marcus ; Fritsch, Jannik: Towards an Integrated Robotic System for Interactive Learning in a Social Context. In: *Proceedings of the IEEE / RSJ International Conference on Intelligent Robots and Systems*. Bejing, 2006

[Wrede u. a. 2004b] Wrede, S. ; Hanheide, M. ; Bauckhage, C. ; Sagerer, G.: An Active Memory as a Model for Information Fusion. In: *Proceedings of the International Conference on Information Fusion*, 2004 (1), S. 198–205

[Wrede u. a. 2006b] Wrede, Sebastian ; Hanheide, Marc ; Wachsmuth, Sven ; Sagerer, Gerhard: Integration and Coordination in a Cognitive Vision System. In: *Proceedings of the IEEE International Conference on Computer Vision Systems (ICVS)* IEEE, 2006. – accepted for publication

[Yang u. Ahuja 2000] Yang, M.-H. ; Ahuja, N.: Recognizing Hand Gesture Using Motion Trajectories. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Bd. 1, IEEE Computer Society, 2000, S. 466–472

[Yang u. a. 2002] Yang, M.-H. ; Ahuja, N. ; Tabb, M.: Extraction of 2D Motion Trajectories and its Application to Hand Gesture Recognition. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24 (2002), Nr. 8, S. 1061–1074

[Zinßer u. a. 2004] Zinsser, T. ; Gräßl, Ch. ; Niemann, H.: Efficient Feature Tracking for Long Video Sequences. In: Rasmussen, C. E. (Hrsg.) ; Bülthoff, H. H. (Hrsg.) ; Giese, M. A. (Hrsg.) ; Schölkopf, B. (Hrsg.): *Proceedings of the 26th DAGM Symposium* Bd. 3175. Heidelberg, Germany : Springer, 2004 (Lecture Notes in Computer Science), S. 326–333

Index