# Multimodale Aufmerksamkeitssteuerung für einen mobilen Roboter

Sebastian Lang

Dipl.-Inform. Sebastian Lang
AG Angewandte Informatik
Technische Fakultät
Universität Bielefeld
E-Mail: slang@techfak.uni-bielefeld.de

Abdruck der genehmigten Dissertation zur Erlangung
des akademischen Grades Doktor-Ingenieur (Dr.-Ing.).
Der Technischen Fakultät der Universität Bielefeld
am 17. Januar 2005 vorgelegt von Sebastian Lang,
am 25. April 2005 verteidigt und genehmigt.

Gutachter:
PD Dr. Gernot A. Fink, Universität Bielefeld
Prof. Dr. Ralf Möller, Universität Bielefeld

Prüfungsausschuss:
Prof. Dr. Helge Ritter, Universität Bielefeld
PD Dr. Gernot A. Fink, Universität Bielefeld
Prof. Dr. Ralf Möller, Universität Bielefeld
Dr. Sven Wachsmuth, Universität Bielefeld

Gedruckt auf alterungsbeständigem Papier °° ISO 9706

# Multimodale Aufmerksamkeitssteuerung für einen mobilen Roboter

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

der Technischen Fakultät der Universität Bielefeld
vorgelegt von

**Sebastian Lang**

17. Januar 2005

# Inhaltsverzeichnis

# Kapitel 1

# Einleitung

Mobile Serviceroboter spielen in unserem Alltag bisher nur eine untergeordnete Rolle. Vereinzelt kann man auf Serviceroboter treffen, die zu Forschungszwecken in öffentlichen Gebäuden, wie zum Beispiel zur Führung durch Museen [Bur98, Nou99, Sch01b] oder als Einkaufsassistent in einem Baumarkt [Böh03], eingesetzt werden. Zudem werden einige erste kommerzielle Produkte angeboten, wie zum Beispiel autonome Staubsauger für den Haushalt oder ein Roboterhund als interaktives Spielzeug [Fuj98]. Während der Leistungsumfang und die Interaktionsfähigkeiten dieser Produkte noch sehr begrenzt sind, werden mit voranschreitender Entwicklung flexiblere und intelligentere Anwendungen möglich. Einsatzmöglichkeiten für Serviceroboter gibt es viele, so zum Beispiel als Assistent im Haushalt, zur Unterstützung hilfsbedürftiger Menschen, als Informationseinrichtung an öffentlichen Plätzen oder auf dem Unterhaltungssektor.

Ob solche Serviceroboter ein kommerzieller Erfolg werden und einmal einen selbstverständlichen Bestandteil unseres alltäglichen Lebens darstellen, hängt sehr davon ab, ob sie von ihren zukünftigen Nutzern akzeptiert werden. Unabhängig von der konkreten Anwendung ist der potenzielle Anwender im Allgemeinen kein Roboterexperte oder Computerspezialist. Er steht jedoch einem sehr komplexen technischen Gerät gegenüber. Damit Roboter eine hohe Akzeptanz beim Anwender erreichen, müssen sie intuitiv bedienbar sein, schnell und akkurat auf Anweisungen reagieren und klar verständliche Rückmeldungen geben.

Die Entwicklung von Robotern mit menschlichem kommunikativen Verhalten stellt einen wichtigen Bereich in der Robotikforschung dar. Es wird angestrebt, dass es dem Anwender möglich ist, so mit dem Roboter zu interagieren, wie er es von der zwischenmenschlichen Kommunikation her gewohnt ist. Nur so kann garantiert werden, dass ein Mensch jederzeit spontan auch mit ihm unbekannten Robotern problemlos umgehen kann, ohne zuvor neue Fähigkeiten erwerben zu müssen. Anstelle von Tastatur, Maus oder *Touchscreen* sind natürliche Sprache und nonverbale Signale wie Gestik, Gesichtsausdruck, Blickrichtung und Körperhaltung die Mittel für die Interaktion zwischen Mensch und Roboter. Diese muss der Roboter zum einen selber anwenden können, zum anderen muss er die Fähigkeit besitzen, die entsprechenden Signale von seinem menschlichen Interaktionspartner zu erfassen und zu interpretieren. Die grundlegende Voraussetzung dafür ist, dass der Roboter die Menschen in seiner Nähe mit Hilfe seiner Sensoren erfassen

Abbildung 1.1: Bei einem mobilen Roboter kann dem Benutzer kein fester Standort zugeordnet werden. Der Roboter muss aus dem Verhalten der Personen erkennen, wer der Benutzer ist.

und erkennen kann und im Verlauf der Zeit beobachten und verfolgen kann.

Mit seinen Sensoren nimmt ein mobiler Roboter in der Regel immer nur einen begrenzten Ausschnitt seiner Umgebung wahr. Durch aktives Verhalten ist er jedoch in der Lage, die Sensoren neu auszurichten und dadurch den erfassten Ausschnitt zu verschieben. Daher ist es nicht erforderlich, dass der Benutzer für die Interaktion einen vorgeschriebenen Standort einnimmt, wie dies bei statischen Systemen der Fall ist, zum Beispiel bei einem Informationskiosk [Reh99]. Der Anwender kann sich frei vor dem Roboter bewegen, wobei es die Aufgabe des Roboters ist, ihn im „Blick“ zu halten. Da dem Benutzer folglich keine feste Position zugeordnet werden kann, ergeben sich spezielle Anforderungen an den Roboter. Innerhalb seines Wahrnehmungsbereichs können sich mehrere Personen gleichzeitig aufhalten (siehe als Beispiel Abbildung 1.1). Jede dieser Personen stellt einen potenziellen Kommunikationspartner für ihn dar. Zunächst muss er erkennen können, ob jemand die Absicht hat, mit ihm zu interagieren. Nicht jede Person die spricht, redet zu ihm. So kann es vorkommen, dass sich zwei Menschen in der Nähe des Roboters unterhalten, oder dass jemand mit einem Mobiltelefon telefoniert. Wenn mehrere Personen anwesend sind, kann der Roboter aufgrund des beschränkten Wahrnehmungsbereichs nicht immer alle gleichzeitig beobachten. Er muss dann über eine geeignete Strategie verfügen, um schnell auf einen Interaktionswunsch reagieren zu können. Wenn er einen Kommunikationspart-

ner gefunden hat, so ergeben sich weitere Anforderungen in der folgenden Interaktionsphase: Der Roboter muss nun in der Lage sein, sich auf den Benutzer zu konzentrieren. Auch wenn er Äußerungen unbeteiligter Personen erfasst, darf er nur auf Anweisungen des Anwenders reagieren.

Sollen Roboter mit menschlichem kommunikativen Verhalten geschaffen werden, ist die Frage naheliegend, was Menschen dazu befähigt, dieselben Anforderungen zu bewältigen. Wenn wir uns zum Beispiel auf einer Party mit vielen Gästen befinden, fällt es uns nicht schwer zu erkennen, wann und von wem wir angesprochen werden. Zudem gelingt es uns dann innerhalb der Geräuschkulisse, die sich aus der Unterhaltung der zahlreichen Gäste ergibt, allein auf die Worte unseres Gesprächspartners zu achten. Das Erkennen, wann sich uns jemand zuwendet, und das Fokussieren dieser Person gelingt uns, da wir über Aufmerksamkeit verfügen. Sie ermöglicht es uns, aus dem vielfältigen Reizangebot der Umwelt einzelne, relevante Reize auszuwählen und bevorzugt zu betrachten, andere dagegen zu übergehen und zu unterdrücken [Müs00]. Sie stellt die Grundlage für ein konzentriertes, zielgerichtetes Handeln dar. Aufmerksamkeit hat einen großen Einfluss darauf, was wir bewusst wahrnehmen und bestimmt, wem wir zuhören und wohin wir schauen. Würde der Roboter über entsprechende Aufmerksamkeitsmechanismen verfügen, könnte er eine für ihn relevante Person aus einer Gruppe auswählen, um sich ihr dann zuzuwenden.

Aufmerksamkeit stellt aber nicht nur als Bestandteil des Wahrnehmungsprozesses einen wichtigen Aspekt für jedes einzelne Individuum dar. Sie hat darüber hinaus einen bedeutenden Einfluss auf das soziale Verhalten in der Kommunikation: Gesprächspartner richten ihre Aufmerksamkeit aufeinander und schauen sich gegenseitig an. Wird dies nicht getan, wird es als fehlendes Interesse und mangelnde Aufmerksamkeit interpretiert (vgl. [Arg76]). Für eine effektive Interaktion zwischen Mensch und Roboter ist es daher erforderlich, dass auch der Roboter in geeigneter Weise Aufmerksamkeit demonstriert. Damit gibt er eine intuitiv verständliche Rückmeldung an den Benutzer, die ihm zeigt, dass er vom Roboter wahrgenommen wird.

Für eine effektive und natürliche Kommunikation muss also auch ein mobiler Roboter mit einem entsprechenden Aufmerksamkeitssystem ausgestattet sein.

## 1.1 Zielsetzung

Das Ziel dieser Arbeit ist die Entwicklung eines Aufmerksamkeitssystems für einen mobilen Roboter für die Interaktion mit Menschen. Die Aufgabe des Systems ist es, die Bewegung des Roboters und die Ausrichtung der Sensoren so zu steuern, dass den im Interaktionssystem verwendeten Komponenten (Sprachverarbeitung, Gestenerkennung, ...) über die Sensoren jederzeit geeignete Daten zur Verfügung stehen. Das Aufmerksamkeitssystem soll darüber hinaus den Fokus der Aufmerksamkeit des Roboters in einer intuitiv verständlichen Weise für den menschlichen Anwender darstellen.

Das zu entwickelnde Aufmerksamkeitssystem soll den Roboter insbesondere dazu befähigen, während der Phase, in der er sich nicht in der Interaktion mit einer Person befindet, Menschen

in seiner Umgebung zu beobachten, um zielgerichtet zu erkennen, wann jemand beabsichtigt, eine Interaktion mit dem Roboter zu beginnen. In der Interaktionsphase ist es die Aufgabe des Systems, allein den Benutzer zu fokussieren und andere anwesende Personen zu ignorieren. Daneben soll es die Aktivierung der sprachverarbeitenden Komponenten im Interaktionssystem übernehmen, so dass nur die Instruktionen vom Kommunikationspartner verarbeitet werden.

Das Ziel ist schließlich die Realisierung eines robusten, lauffähigen Systems, das, integriert in das Interaktionssystem, auf einem autonomen mobilen Roboter mit Standard-Hardware einsatzfähig ist und im Betrieb eine effektive Interaktion erlaubt.

Als Beispielanwendung wird das so genannte *Home-Tour*-Szenario betrachtet. Die Vorstellung dabei ist, dass ein Benutzer einen mobilen Serviceroboter für sein Zuhause käuflich erworben hat und nun zum ersten Mal in Betrieb nimmt. Um den Roboter für zukünftige Aufgaben effektiv einsetzen zu können, muss dieser zunächst mit der für ihn neuen Umgebung vertraut gemacht werden. In einem interaktiven Prozess zeigt und beschreibt der Nutzer Gegenstände und Orte, die für spätere Interaktionen und Aufgabenstellungen relevant sind. Der Benutzer kann mit dem Roboter natürlichsprachlich kommunizieren. Um auf Objekte zu verweisen, können neben Sprache auch deiktische Gesten verwendet werden. Der Roboter erfasst benannte Objekte mit einer Kamera und speichert neben der Ansicht des Objekts Informationen wie Position, Größe und zusätzliche sprachliche Angaben des Benutzers. Um neue Objekte zu zeigen, ist es dem Benutzer darüber hinaus möglich, den Roboter durch Vorausgehen zu einer neuen Position im Haus zu führen.

## 1.2 Gliederung der Arbeit

Zu Beginn der Arbeit gibt Kapitel 2 einen Überblick über bereits existierende Ansätze, die Aufmerksamkeitsmechanismen für künstliche Systeme realisieren. Der Schwerpunkt liegt dabei auf Aufmerksamkeitssystemen, die für die Interaktion von Benutzern mit mobilen Servicerobotern entwickelt wurden. Die anschließenden Kapitel befassen sich mit dem in dieser Arbeit entwickelten Aufmerksamkeitssystem. Den Ausgangspunkt bildet Kapitel 3, das den eingesetzten Roboter beschreibt. Das Kapitel 4 erläutert dann die Konzeption des entwickelten Aufmerksamkeitssystems. Details der Realisierung werden in den drei folgenden Kapiteln beschrieben. Zunächst wird in Kapitel 5 ein Verfahren zum multimodalen Verfolgen von Objekten vorgestellt, das im Rahmen dieser Arbeit entwickelt wurde. Die konkrete Anwendung dieses Verfahrens zum Verfolgen von Personen unter Verwendung des eingesetzten Roboters ist Thema des anschließenden Kapitels 6. Die Darstellung des Aufmerksamkeitssystems erfolgt schließlich in Kapitel 7. Darin wird auch beschrieben, wie sich die Aufmerksamkeitssteuerung in das Gesamtsystem integriert, das die Interaktionsfähigkeiten des Roboters im betrachteten Szenario realisiert. In Kapitel 8 wird die Leistungsfähigkeit des vorgestellten Ansatzs anhand von Ergebnissen aus verschiedenen Benutzerexperimenten diskutiert. Die Arbeit schließt mit einer Zusammenfassung in Kapitel 9.

# Kapitel 2

# Aufmerksamkeitsmechanismen für Roboter

Viele mobile Roboter, die ihre Umgebung über Kameras oder Mikrofone erfassen, greifen bei der Verarbeitung der Sensordaten und bei der Steuerung ihres Verhaltens auf Aufmerksamkeitsmechanismen zurück. Die Gründe dafür sind vielfältig. Zum einen stehen praktische Überlegungen im Vordergrund. So werden im Hinblick auf die begrenzten Rechnerkapazitäten eines autonomen Systems Selektionsmechanismen eingesetzt, um die Menge der zu verarbeitenden Daten zu reduzieren und nur die für die jeweilige Aufgabe relevanten Daten herauszufiltern. Aufmerksamkeitssteuerungen dienen auch dazu, den Roboter und seine Sensoren in geeigneter Weise anzusteuern und auszurichten, sodass die Einschränkungen durch die begrenzten Einzugsbereiche der Sensoren überwunden werden. Zum anderen ist es das Ziel, menschliches Aufmerksamkeitsverhalten möglichst gut nachzubilden. So ist man insbesondere im Bereich der Servicerobotik bestrebt, ein intuitiv verständliches Verhalten des Roboters zu erzeugen, da dies einen wichtigen Beitrag für die Natürlichkeit der Interaktion darstellt. Darüber hinaus werden theoretische Aufmerksamkeitsmodelle auf mobilen Robotern getestet, um menschliches Verhalten zu simulieren und besser zu verstehen.

Dieses Kapitel beschreibt verschiedene Ansätze, die Aufmerksamkeitsmechanismen für mobile Roboter realisieren. Der Schwerpunkt der Auswahl liegt dabei auf Roboterapplikationen, bei denen die Interaktion mit Menschen im Zentrum steht. Um die unterschiedlichen Verfahren besser strukturieren zu können, werden zunächst vier Kriterien spezifiziert, nach denen sich Aufmerksamkeitssysteme im Bereich der Mensch-Roboter-Interaktion grob voneinander unterscheiden lassen:

**Berücksichtigte Sensoren:** Die Grundlage für die Steuerung der Aufmerksamkeit eines Roboters ist die Verarbeitung von Sensordaten. Aufmerksamkeitssysteme lassen sich danach unterscheiden, welche Arten von Sensoren sie dabei berücksichtigen. Weit verbreitet ist der Einsatz von Kameras und Mikrofonen, die das Sehen und Hören eines Roboters realisieren. Vereinzelt werden auch Sensoren verwendet, die keine Entsprechung zu den menschlichen Sinnesorganen haben, wie zum Beispiel Laser-, Sonar- oder Infrarotsensoren.

**Repräsentation der Relevanzwerte:** Der Selektionsprozess, der darüber entscheidet, worauf sich die Aufmerksamkeit eines Roboters bei einer gegebenen Aufgabe richtet, kann im Allgemeinen in drei Schritte gegliedert werden:

1. Sensordaten werden erfasst und gegebenenfalls vorverarbeitet.
2. Die (vorverarbeiteten) Daten werden bezüglich ihrer Relevanz für die Aufgabe bewertet und die Relevanzwerte in einer geeigneten Datenstruktur abgelegt.
3. Der maximale Wert wird selektiert. Er bestimmt den Fokus der Aufmerksamkeit.

Der wesentliche Aspekt, der Aufmerksamkeitssysteme voneinander unterscheidet, ist die Art der Datenstruktur, die zur Repräsentation der Relevanzwerte vorgesehen ist (Punkt 2). Man findet zwei Ansätze vor:

- Bei der ortsbasierten Repräsentation der Relevanzwerte modelliert die verwendete Datenstruktur eine gleich- oder niedrigdimensionale Abbildung des Raums, der den Roboter umgibt. Das klassische Beispiel sind so genannte Aufmerksamkeitskarten, die insbesondere bei visuellen Aufmerksamkeitssystemen Verwendung finden. Dies sind zweidimensionale Datenfelder, die in Breite und Höhe dem Kamerabild mit voller oder niedrigerer Auflösung entsprechen. Bei der ortsbasierten Aufmerksamkeit wird jedem Punkt im Raum oder jeder Richtung, die durch die Datenstruktur repräsentiert wird, ein Relevanzwert zugewiesen. Die Aufmerksamkeit richtet sich folglich auf einen Ort.
- Bei der objektbasierten Repräsentation werden zunächst die Sensordaten vorverarbeitet, indem einzelne Objekte extrahiert werden. Die Objektextraktion stellt bereits einen ersten Selektionsprozess dar, da in der Regel nur die Objekte berücksichtigt werden, die für die Anwendung relevant sind. Im Bereich der Mensch-Roboter-Interaktion stellen zum Beispiel Menschen die relevanten „Objekte“ dar. Bei der objektbeziehungsweise personenbasierten Aufmerksamkeit wird folglich jedem Objekt ein Relevanzwert zugewiesen. Die Aufmerksamkeit richtet sich also auf ein Objekt.

Mit der Unterscheidung zwischen ortsbasierten und objektbasierten Aufmerksamkeitssystemen wird dem Aspekt Rechnung getragen, dass auch die Psychologie diese beiden Arten von selektiver Aufmerksamkeit beim Menschen differenziert (vgl. [Mül02, Sch01a]).

**Bewertung der Relevanz:** Die Bewertung der Relevanz (Punkt 2, siehe oben) zu einem gegebenen Zeitpunkt hängt zunächst von den jeweils aktuell erfassten Sensordaten ab. Sie kann aber zusätzlich von vorausgegangenen Sensordaten oder dem internen Zustand des Roboters abhängen. Aufmerksamkeitssysteme werden folglich danach unterschieden, ob sie eine eindeutige, zeitunabhängige Abbildung von Sensordaten auf Relevanzwerte realisieren (statische Bewertung) oder ob andere Parameter in den Bewertungsprozess einfließen (dynamische Bewertung).

Dynamische Bewertungsverfahren sind erforderlich, um gewisse Aspekte des menschlichen Aufmerksamkeitsverhaltens zu modellieren. Ein Beispiel findet man bei der so ge-

nannten Orientierungsreaktion [Eim96]. Diese tritt bei plötzlicher Veränderung der Reizkonstellation ein, beispielsweise bei einem lauten Knall, und äußert sich in einer aktiven Aufmerksamkeitszuwendung. Regelmäßig wiederholte Darbietung des Reizes führt jedoch zur Habituierung der Orientierungsreaktion, sodass die Zuwendung nicht mehr erfolgt. Ein anderes Beispiel ist die willkürliche Selektion, von der man spricht, wenn die Intention einer Person für das Zustandekommen einer Aufmerksamkeitszuwendung verantwortlich ist [Eim96]. Sucht beispielsweise jemand nach einem Gegenstand, wird allen Gegenständen, die dem gesuchten ähneln, eine erhöhte Aufmerksamkeit zuteil.

**Personenanzahl:** Um eine personenbasierte Aufmerksamkeit für einen Roboter zu realisieren, ist ein Verfahren zur Detektion von Personen erforderlich. Aufmerksamkeitssysteme unterscheiden sich darin, wieviele Personen gleichzeitig detektiert und damit im Aufmerksamkeitszuwendungsprozess berücksichtigt werden können. Es wird zwischen einer und mehreren Personen unterschieden. Bei Aufmerksamkeitssystemen, die lediglich eine Person berücksichtigen, ist eine Bewertung der Relevanz überflüssig und die Selektion trivial. Dennoch kann man von einem Aufmerksamkeitssystem sprechen, da die Aufmerksamkeit selektiv auf Menschen gerichtet wird. Wenn ein System mehrere Personen gleichzeitig berücksichtigt, ist dagegen ein Selektionsprozess erforderlich.

Neben diesen Kriterien werden Aufmerksamkeitssysteme natürlich auch dadurch charakterisiert, wie sie die jeweils anliegenden Sensordaten nach ihrer Relevanz bewerten, das heißt welche Strategie sie verfolgen, um die Aufmerksamkeit zu steuern. Wichtig im Bereich der Mensch-Roboter-Interaktion ist auch die Frage, wie die Systeme ihren Aufmerksamkeitsfokus nach außen hin für den Beobachter darstellen.

Das wichtigste Unterscheidungskriterium ist zunächst jedoch die Repräsentation der Relevanzwerte, die mit der Entscheidung einhergeht, ob das Aufmerksamkeitssystem auf einer Objektextraktion aufsetzt. Die folgende Darstellung der Aufmerksamkeitssysteme befasst sich daher zunächst mit ortsbasierten Ansätzen in Abschnitt 2.1, gefolgt von personenbasierten Ansätzen in Abschnitt 2.2.

## 2.1 Ortsbasierte Aufmerksamkeit

Bei der ortsbasierten Aufmerksamkeit wird aus der Sicht des Roboters jedem Punkt im Raum oder jeder Richtung eine Relevanz zugeordnet. Zu diesem Zweck sind Datenstrukturen erforderlich, die eine gleich- oder niedrigdimensionale Abbildung des Raums modellieren. Bei optischen Sensoren erfolgt die Abbildung bereits bei der Messung. So stellt zum Beispiel ein Kamerabild eine zweidimensionale Abbildung des erfassten Bereichs dar. Eine Datenstruktur, die in diesem Fall geeignet ist, die Relevanzwerte aufzunehmen, ist ebenfalls ein zweidimensionales Feld. Man spricht dabei von einer so genannten Aufmerksamkeits- oder Aktivierungskarte. In dieser lässt sich direkt ablesen, welche Bereiche im Bild für den Roboter die höchste Relevanz aufweisen.

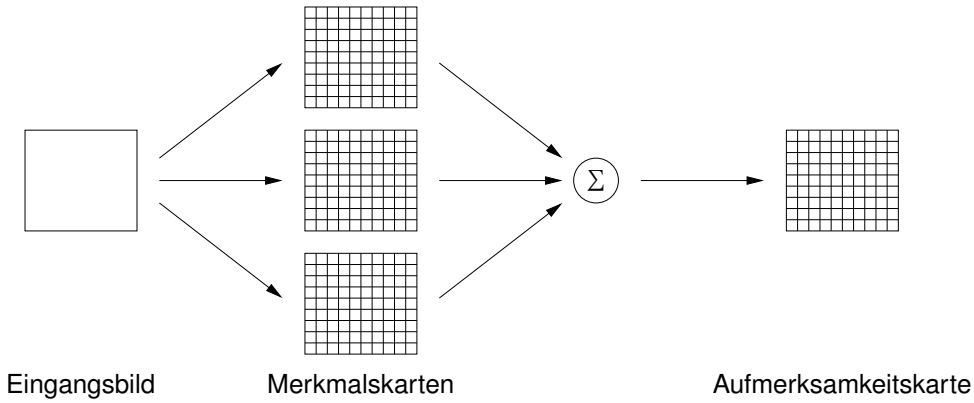



Abbildung 2.1: Prinzipieller Aufbau eines Systems zur visuellen, ortsbasierten Aufmerksamkeit: Für ein Bild werden Merkmalskarten für verschiedene Merkmale berechnet und gewichtet zur Aufmerksamkeitskarte aufsummiert.

In multimodalen Systemen, die gleichzeitig auf Sensoren mit unterschiedlichen Wahrnehmungsbereichen zurückgreifen, lässt sich eine direkte Korrespondenz zwischen Werten der Aufmerksamkeitskarte und den Sensordaten, wie sie bei rein visuellen Aufmerksamkeitssystemen existiert, nicht herstellen. Hier sind weitere Überlegungen erforderlich. Im Folgenden werden zunächst unimodale und im Anschluss multimodale Ansätze beschrieben.

### 2.1.1 Unimodale Ansätze

Der überwiegende Teil ortsbasierter, unimodaler Ansätze beschäftigt sich mit der visuellen Aufmerksamkeit. Modelle zur visuellen Aufmerksamkeit (siehe zum Beispiel [Wol94, Itt98]) ermitteln hervorstechende Bildbereiche über lokale Unterschiede bezüglich Merkmalen wie zum Beispiel Farbe, Kontrast oder Orientierung. Diese werden für das gesamte Bild berechnet und die Ergebnisse als numerische Werte in Merkmalskarten abgelegt (siehe auch Abbildung 2.1). Je größer ein numerischer Wert an einer Speicherstelle einer Merkmalskarte, desto größer ist der Unterschied bezüglich des Merkmals an entsprechender Stelle im Bild, und desto hervorstechender und auffälliger der Bildbereich. Die verschiedenen Merkmalskarten werden durch gewichtete, punktweise Summation zu einer Aufmerksamkeitskarte zusammengeführt. Die Speicherstelle der Aufmerksamkeitskarte mit dem maximalen Wert bestimmt den Bildbereich, auf den sich der Fokus der Aufmerksamkeit des Systems richtet.

Die Berechnung der Aufmerksamkeitskarten ist zunächst ein rein bottom-up gesteuerter Prozess, der eine unwillkürliche Aufmerksamkeitszuwendung auf einen bestimmten Bildbereich realisiert. Top-down-Einfluss, das heißt die aufgabenabhängige Intention des Systems, lässt sich über die Wahl der Gewichte bei der Summation der Merkmalskarten zur Aufmerksamkeitskarte modellieren. So können einzelne Merkmale stärker berücksichtigt werden als andere. McGuire und

Kollegen verwenden diese Strategie in einem Interaktionsszenario, in dem der Benutzer die Aufmerksamkeit des Robotersystems durch sprachliche Instruktionen beeinflussen kann [McG02]. Eine Äußerung, die zum Beispiel „ ... *der rote Würfel* ... “ beinhaltet, führt zu einer stärkeren Gewichtung der Merkmalskarte für die Farbe Rot, sodass das System vermehrt rote Objekte fixiert. Eine andere Strategie, um Top-down-Einfluss in ein visuelles Aufmerksamkeitssystem zu integrieren, wird in dem von Cave entwickelten *FeatureGate*-Modell [Cav99, Dri98] verfolgt. Dies setzt dazu direkt bei der Berechnung der Merkmalskarten an. Dort werden Bildregionen hohe Werte in den Merkmalskarten zugewiesen, die sich nicht nur stark von benachbarten Regionen abheben, sondern gleichzeitig den Merkmalen eines vorgegebenen Ziels möglichst ähnlich sind.

Die Berechnung der Aufmerksamkeitskarte für ein gegebenes Bild liefert zunächst einen Maximalwert, der den Aufmerksamkeitsfokus vorgibt. Um Folgen von Fixationspunkten (Blicksakkaden) zu erzeugen, werden die Werte im Bereich der Aufmerksamkeitskarte um den aktuellen Maximalwert für einen gewissen Zeitraum herabgesetzt, sodass ein neuer Punkt das Maximum annimmt und die nächste Aufmerksamkeitsrichtung vorgibt. Dieser Prozess wird iterativ fortgesetzt (siehe zum Beispiel [Itt98]). Der dabei realisierte Mechanismus, der die Rückkehr der Aufmerksamkeit zu einer zuvor fokussierten Stelle unterdrückt, wird als *Inhibition of Return* (Hemmung der Rückkehr) bezeichnet [Pos84]. Vijayakumar und Kollegen setzen ein entsprechendes Aufmerksamkeitssystem auf einem humanoiden Roboter ein, um Blicksakkaden durch entsprechende Ausrichtung der Kameras des Roboterkopfs zu realisieren [Vij01]. Da der *Inhibition-of-Return*-Effekt ortsgebunden ist, muss auf einem mobilen Roboter beachtet werden, dass sich die gehemmten Bereiche mit jeder Bewegung der Kameras bezüglich des Bildkoordinatensystems verschieben. Dieses wird in dem beschriebenen Verfahren durch geeignete Transformationen berücksichtigt.

Relevanzwerte von Aufmerksamkeitskarten können nicht nur auf objektunabhängigen Merkmalen, wie zum Beispiel Helligkeit, Farbe oder Orientierung berechnet, sondern auch auf Basis von Objekterkennungsergebnissen bestimmt werden. Zu diesem Zweck werden Merkmalskarten an den Stellen mit hohen Werten belegt, an denen Objekte im Bild detektiert wurden.[1] Ein Aufmerksamkeitssystem für einen Roboter, das auf diesen Ansatz zurückgreift, wird von Breazeal und Scassellati beschrieben [Bre99]. Es ist Bestandteil eines komplexen Verhaltenssystems, das den Roboter in einer Art Kind-Bezugsperson-Szenario steuert, wobei der Roboter die Rolle des Kinds übernimmt. Bei dem Roboter, genannt *Kismet*, handelt es sich um einen Roboterkopf mit beweglichen Augen, Augenlidern, Augenbrauen, Ohren und Mund. In den Augen befinden sich Kameras, mit denen die Umgebung erfasst wird. Das visuelle Aufmerksamkeitssystem basiert auf Merkmalskarten für Farbe, Bewegung und Gesichter. Die Relevanzwerte der Merkmalskarte für Gesichter entsprechen hier den Erkennungsergebnissen eines Gesichtsdetektors, der auf alle Positionen im Bild angewendet wird. Der Detektor liefert dabei keine eindeutige Entscheidung sondern bestimmt, wie wahrscheinlich das Vorhandensein eines Gesichts an der überprüften Stelle ist.

Wenn ein neuer Fokuspunkt über die Aufmerksamkeitskarte ermittelt wurde, visiert der Roboter

[1] Wenngleich eine Objektdetektion integriert ist, handelt es sich weiterhin um eine ortsbasierte Aufmerksamkeit, da sich die Aufmerksamkeit weiterhin auf einen Ort und nicht explizit auf ein Objekt richtet.

die entsprechende Richtung mit seinen Augen an. Er stellt dadurch gleichzeitig den Fokus seiner Aufmerksamkeit gegenüber dem Interaktionspartner dar. Um den Blick nach einer Sakkade eine gewisse Zeit in einer Richtung zu halten, ihn dann wieder zu lösen und auf einen neuen Ort auszurichten, werden die drei Merkmalskarten um eine zusätzliche Karte erweitert. Diese realisiert den Effekt der Habituation. Die Werte in der Karte sind zeitunabhängig durch eine zweidimensionale Gaußverteilung festgelegt, deren Maximalwert sich im Zentrum der Karte befindet. Der Einfluss der Karte wird über das zugehörige Gewicht bei der Summation der Merkmalskarten bestimmt. Dieses hat direkt nach einer Sakkade den Wert eins, sodass der zentrale Bereich hohe Aufmerksamkeit erhält und der Blick gehalten wird. Der Wert nimmt mit der Zeit linear ab und wird sogar negativ, sodass die Region in Blickrichtung keine hohen Relevanzwerte mehr annimmt und der Fokus der Aufmerksamkeit auf andere Orte gerichtet werden kann.

Die Gewichte der anderen Merkmalskarten werden durch den Zustand des Verhaltenssystems bestimmt. Dieser ergibt sich aus den Bedürfnissen des Roboters nach sozialen Stimuli (Menschen) oder nicht-sozialen Stimuli (Spielzeuge). Die Bedürfnisse des Roboters können sowohl in einem ausgeglichenen Zustand als auch über- oder unterstimuliert sein. Der Zustand der Bedürfnisse bestimmt die Auswahl von Verhaltensweisen des Roboters. Wenn *Kismet* zum Beispiel den Kontakt zu Menschen sucht, werden die Gewichte der Merkmalskarten für Bewegung und Gesichter angepasst, da diese Merkmale charakteristisch für Menschen sind. Je nachdem ob das Bedürfnis nach sozialen Stimuli über- oder unterstimuliert ist, werden die Gewichte erniedrigt beziehungsweise erhöht, sodass der Roboter entweder von Personen zurückweicht oder danach sucht. Da der Interaktionspartner in der Regel in entsprechender Weise reagiert, das heißt ebenfalls zurückweicht, wenn der Roboter zurückweicht oder näher kommt, wenn der Roboter nach ihm sucht, hat das rückgekoppelte System eine regulierende Wirkung auf den Verlauf der Interaktion.

Das von Breazeal und Scassellati entwickelte Aufmerksamkeitssystem zeichnet sich insbesondere durch seine ausgeklügelte Methode zur dynamischen Bewertung der anliegenden Sensordaten aus und führt damit zu einem dynamischen, lebendig wirkenden Verhalten des Roboters. Mit diesem System kann jedoch keine Daueraufmerksamkeit realisiert werden, wie sie für einen Serviceroboter in der Interaktionsphase benötigt wird. In dieser muss der Roboter sich auf den Interaktionspartner fixieren können, selbst wenn die perzeptuellen Stimuli aus Richtung des Benutzers zwischenzeitig gering sind, weil dieser sich zum Beispiel kurzfristig abgewendet hat und sein Gesicht nicht detektiert werden kann. Da rein visuelle, ortsbasierte Aufmerksamkeitssysteme immer auf den höchsten Reiz im Blickfeld reagieren, würden sie in dieser Situation einen Aufmerksamkeitswechsel veranlassen. Visuelle Aufmerksamkeitssysteme eignen sich daher insbesondere dazu, ein exploratives Verhalten für Roboter zu realisieren (vgl. zum Beispiel [Gon99]).

Aufmerksamkeitskarten für visuelle Aufmerksamkeitssysteme sind von ihrer Struktur her eng an das Bilddatenformat gekoppelt. Dennoch werden sie auch für Daten von anderen Sensoren eingesetzt. Frintrop und Kollegen [Fri03a] zum Beispiel berechnen Aufmerksamkeitskarten auf den Daten eines auf einem Roboter montierten beweglichen Laser-Entfernungsmessers. In einem Schwenk tastet der Laser den Bereich vor dem Roboter ab und generiert somit ein 2D-Tiefenbild. Neben Abstandswerten liefert der Sensor zusätzlich die Stärke der Reflexion des Laserlichts. Abstands- und Remissionswerte werden jeweils als Intensitäts- beziehungsweise Farbwerte in ein Farbbild kodiert, welches als Eingangsbild für das visuelle Aufmerksamkeitssystem von Itti und

Kollegen [Itt98] dient. Durch die kodierte Tiefeninformation erhalten auch die Bereiche in der Aufmerksamkeitskarte einen hohen Wert, bei denen starke lokale Änderungen in den Abstandswerten vorliegen, zum Beispiel an Objektkanten.

Für Serviceroboter in Interaktionsszenarien mit Menschen ist aber insbesondere eine Kombination aus auditiven und visuellen Sensordaten von Interesse, da sie die Grundlage für eine Aufmerksamkeitssteuerung darstellt, die sowohl auf Sprache als auch auf nonverbale Signale von Benutzern reagiert. Ortsbasierte Aufmerksamkeitssysteme, die neben Kameras auch auf Mikrofone zurückgreifen, werden im folgenden Abschnitt vorgestellt.

### 2.1.2 Multimodale Ansätze

Aufmerksamkeitskarten in der Form, wie sie bei visuellen Aufmerksamkeitssystemen Verwendung finden, eignen sich nicht für Sensoren, deren Daten (oder die daraus extrahierten Ergebnisse) sich nicht direkt in Bildkoordinaten transformieren lassen. In Interaktionsszenarien werden zum Beispiel häufig Mikrofonfelder (engl. *microphone arrays*) zur Lokalisation des Sprechers eingesetzt, mit denen sich nur der Azimut[2] der Geräuschquelle bestimmen lässt. Hierfür ließen sich eindimensionale Merkmals- und Aufmerksamkeitskarten einsetzen. Komplizierter wird der Fall, wenn ein Aufmerksamkeitssystem mit mehreren Sensoren verschiedenen Typs realisiert wird, die heterogene Daten liefern und unterschiedliche Anforderungen an die Struktur der Aufmerksamkeitskarten stellen. Zwei Ansätze, die dieses Problem durch eine geeignete Wahl der Datenstruktur lösen, werden im Folgenden beschrieben.

Déniz und Kollegen entwickeln einen Roboterkopf, der ihnen zur Erforschung von sozialen Aspekten der Mensch-Roboter-Interaktion dient. Ein Bestandteil des Systems ist ein multimodales Aufmerksamkeitssystem, das den Roboter befähigt, auf audiovisuelle Stimuli von Interaktionspartnern zu reagieren [Dén03]. Mit einem Mikrofonpaar wird die Richtung des Sprechers innerhalb des 180°-Bereichs vor dem Roboter bestimmt. Gleichzeitig werden in dem Bild einer statischen omnidirektionalen Kamera Menschen über ein Verfahren zur Hintergrundsubtraktion detektiert. Von dem Kamerabild wird dabei nur der Ausschnitt berücksichtigt, der sich mit dem Einzugsbereich des Mikrofonpaars deckt. Für detektierte Regionen wird die Richtung innerhalb des 180°-Bereichs ermittelt. Sowohl auditive als auch visuelle Aufmerksamkeitsaktivierungen beschränken sich damit auf eine Richtung in der horizontalen Ebene. Folglich können für beide Modalitäten gleich große eindimensionale Merkmalskarten eingesetzt werden, die mit dem Winkel $\rho$ parametrisiert sind. Die Merkmalskarten enthalten hohe Werte für Winkel, die die Richtungen von lokalisierten Geräuschquellen beziehungsweise segmentierten Regionen angeben. Die Aufmerksamkeitskarte $A(\cdot)$ ergibt sich, wie bei den oben beschriebenen Systemen zur visuellen Aufmerksamkeit, nach dem Prinzip der punktweisen, gewichteten Summation von Videomerkmalen $F(\cdot)$ und Audiomerkmalen $G(\cdot)$ wie folgt:

$$A(\rho) = v\,F(\rho) + s\,G(\rho) + k\,C(\rho) + T(\rho)$$

[2] bei Polarkoordinaten der Winkel zwischen der positiven Abszissenachse und dem Radiusvektor

Dabei sind $v$ und $s$ die Gewichte für die Merkmalskarten. Der Term $k\,C(\cdot)$ steuert eine zusätzliche Aktivierung für den Bereich des aktuellen Fokus der Aufmerksamkeit bei, sodass primär Personen in Blickrichtung selektiert werden. Die Funktion $T(\cdot)$ wurde von Déniz und Kollegen vorgeschlagen, um Top-down-Einflüsse zu integrieren. Sie wird in ihrer Implementierung jedoch noch nicht genutzt. Das System ist in der Lage, bei Anwesenheit mehrerer Personen den Fokus der Aufmerksamkeit auf den aktuellen Sprecher zu richten.

Auf einer komplexeren und mächtigeren Datenstruktur setzt das von Hambuchen in ihrer Dissertation entwickelte Aufmerksamkeitssystem auf [Ham04]. Die Datenstruktur, die als *Sensory Ego-Sphere* (*SES*) bezeichnet wird, dient primär als Kurzzeitgedächtnis für mobile Roboter [Pet01]. Die *SES* ist eine sphärische Schale, deren Zentrum der Ursprung des Roboterkoordinatensystems ist. Auf der *SES* können so genannte Ereignisse, das heißt Wahrnehmungen von Objekten mit den Robotersensoren, abgelegt werden. Die Position auf der Schale spezifiziert dabei die Richtung aus Sicht des Roboters. Die *SES* ist mit einer endlichen Anzahl von Speicherstellen versehen, die optimal auf der Sphäre verteilt sind. Sie bilden ein regelmäßiges Polyeder und sind netzartig miteinander verbunden, wodurch Nachbarschaftsbeziehungen definiert werden. Ein Ereignis wird entsprechend der räumlichen Position an die Speicherstelle mit ähnlichster Richtung abgelegt.

In dem Aufmerksamkeitssystem dient die Struktur der *SES* zum Aufbau eines so genannten Aufmerksamkeitsnetzwerks, das auch als eine Art Aufmerksamkeitskarte betrachtet werden kann. Mit jedem neu eintreffenden Ereignis werden an der zugehörigen Speicherstelle drei Aufmerksamkeitswerte abgelegt, die zusammen den Aktivierungswert bestimmen. Jedes Auftreten eines Ereignisses erhöht den Aktivierungswert durch einen Wert $I(\cdot)$. Gleichzeitig wird die Relevanz des Ereignisses für die momentane Aufgabenstellung des Roboters mit $TR(\cdot)$ bewertet. Über diesen Wert wird der Top-down-Einfluss integriert. Zusätzlich gibt es einen Wert für die Habituation $H(\cdot)$. Der Aktivierungswert $S(j, e)$ für ein Ereignis $j$ an einer Speicherstelle $e$ berechnet sich wie folgt:

$$S(j,e) = \big(I(j,e) + TR(j,e)\big)\; H(j,e)$$

Die Gesamtaktivierung an einer Speicherstelle der *SES* ist die Summe der Aktivierungswerte für alle dort abgelegten Ereignisse. Der Roboter steuert seine Aufmerksamkeit in die Richtung der Speicherstelle mit dem höchsten Aktivierungswert. Um das Aufmerksamkeitsverhalten zeitnah an den zuletzt registrierten Ereignissen zu orientieren, nimmt der Auftretens-Aktivierungswert $I(\cdot)$ für ein abgelegtes Ereignis ab dem Zeitpunkt seiner Zuordnung kontinuierlich ab. Das Ereignis wird aus dem Aufmerksamkeitsnetzwerk entfernt, wenn $I(\cdot)$ den Wert 0 annimmt. Wenn mehrere gleiche Ereignisse kurz hintereinander in ähnlicher Richtung beobachtet werden, nimmt die Habituation zu und der Wert $H(\cdot)$ entsprechend ab.

Das Verfahren ist auf dem humanoiden Roboter *ISAC* [Kaw00] getestet worden. Es handelt sich dabei um einen Robotertorso mit einem beweglichen Stereo-Kamera-Kopf und zwei Armen. Menschen werden über drei Sensoren erfasst: Infrarotsensoren registrieren Bewegung, mit den Kameras wird das Gesicht detektiert und mit Mikrofonen kann die Stimme lokalisiert werden. Da bei der Lokalisation der Stimme nur der Azimut bestimmt werden kann, werden die für das entsprechende Ereignis berechneten Aufmerksamkeitswerte an allen Speicherstellen abgelegt, die

einen ähnlichen Azimutwinkel aufweisen und einen Polarwinkel innerhalb eines vorgegebenen Intervalls haben. Neben Menschen können verschiedene Objekte detektiert werden. Das System ist in der Lage, den Fokus der Aufmerksamkeit aufgabenbezogen zwischen dem Benutzer und Objekten zu wechseln.

Ein Problem bei dem Ansatz von Hambuchen ergibt sich bei schnellen Bewegungen von Benutzern oder Objekten. Nacheinander registrierte Ereignisse werden dann über verschiedene Speicherstellen der *SES* verteilt, sodass sich die Aktivierungswerte von sich bewegenden Objekten nicht durch Summation verstärken, wie dies bei stationären Objekten der Fall ist. Um dieses Problem zu beheben, müsste ein Verfahren zum Verfolgen von Objekten integriert werden, über das die Zusammengehörigkeit von verteilten Ereignissen bestimmt werden kann. In dem Fall würde es sich dann aber um eine objektbasierte Aufmerksamkeit handeln.

Es ist ein genereller Nachteil von ortsbasierten Aufmerksamkeitssystemen, dass sie die raumzeitliche Entwicklung von Aktivierungswerten nicht berücksichtigen. In einem Interaktionsszenario, wie es in dieser Arbeit behandelt wird, in dem sich der Benutzer frei vor dem Roboter bewegen kann, spielt dieser Aspekt jedoch eine wichtige Rolle. Nur wenn die Bewegungen der Personen verfolgt werden, kann der Aufmerksamkeitsfokus des Roboters auf dem Benutzer gehalten werden, selbst wenn dieser zum Beispiel zwischenzeitig nicht spricht und das Ausbleiben von Bottom-up-Aktivierung zu einem geringen Aktivierungswert für den Benutzer führt. Aus diesem Grund setzen viele Aufmerksamkeitssysteme für Serviceroboter auf objekt- beziehungsweise personenbasierter Aufmerksamkeit auf. Dazu werden Personen detektiert und über die Zeit verfolgt. Die Aktivierungswerte werden dabei den beobachteten Personen zugeordnet, nicht aber objektunabhängigen Orten im Raum.

## 2.2 Personenbasierte Aufmerksamkeit

Wie in der Einleitung bei den Unterscheidungskriterien für Aufmerksamkeitssysteme bereits erwähnt, können personenbasierte Ansätze unterschieden werden in solche, die zu jedem Zeitpunkt nur eine Person berücksichtigen, und solche, die in den Aufmerksamkeitsprozess mehrere Personen einbeziehen. Dieses Kriterium ist von großer Bedeutung, da Ein-Personen-Verfahren in der Regel den Mehr-Personen-Fall in ihrem Konzept gar nicht berücksichtigen. Das Verhalten des Roboters ist dann teilweise nicht abzusehen. Im schlimmsten Fall führt der Ansatz zu einem handlungsunfähigen System.

Dieses Problem lässt sich anhand eines Experiments von Kopp und Gärdenfors erläutern [Kop02]. In ihrem Versuch beobachtet ein Roboter mit seiner Kamera ein Fließband, welches Objekte transportiert. Seine Aufgabe ist es, mit seinem Greifarm die Objekte vom Band zu heben. Diese Fähigkeit wird durch einen Satz einfacher Stimulus-Reaktions-Paare realisiert. Solange sich nur ein Objekt auf dem Fließband befindet, kann der Roboter die Aufgabe problemlos bewältigen. Er folgt dem jeweiligen Objekt, um es zu greifen. Sobald sich jedoch mehrere Objekte auf dem Fließband und damit im Wahrnehmungsbereich der Roboterkamera befinden, reagiert das System zufällig und unkoordiniert auf alle Objekte. Ein zielgerichtetes Greifen ist nicht mehr möglich.

Wenn mehrere Objekte oder, wie in den hier betrachteten Anwendungen, mehrere Personen gleichzeitig erfasst werden, ist ein geeigneter Selektionsmechanismus erforderlich. In dem folgenden Abschnitt werden zunächst die Aufmerksamkeitssysteme beschrieben, die lediglich eine Person berücksichtigen. Der anschließende Abschnitt behandelt den Mehr-Personen-Fall.

### 2.2.1 Ansätze für eine Person

Die meisten Serviceroboter verfügen über Kameras, weil mit ihnen reichhaltige Information über die beobachtete Szene gewonnen werden kann. Für die Interaktion mit Menschen eignen sich Kameras insbesondere dazu, nonverbale Signale, wie zum Beispiel Zeigegesten oder Blickrichtung, zu erkennen. Als Voraussetzung muss der Roboter den Benutzer detektieren und dann im Bild der Kamera halten. Dieses Vorgehen kann als Aufmerksamkeitsverhalten aufgefasst werden, da der Roboter seine Aufmerksamkeit aktiv auf eine Person ausrichtet.

#### Visuelle Aufmerksamkeitssysteme

Ein rein visuelles, personenbasiertes Aufmerksamkeitssystem wird von Ghidary und Kollegen in [Ghi02] beschrieben. Die von ihnen betrachtete Anwendung ist mit dem *Home-Tour*-Szenario vergleichbar. Dabei lernt ein Roboter in der Interaktion mit einem menschlichen Benutzer die Positionen von Objekten in einem Haus. Der Roboter ist mit einer Kamera ausgestattet. Für Spracheingaben dient ein Nahbesprechungsmikrofon, das der Benutzer bei sich tragen muss. Die Steuerung des Roboters erfolgt über Sprachkommandos und deiktische Gesten. Die Merkmale zur Detektion des Benutzers sind Bewegung und Hautfarbe [Ghi00]. Die Bewegung wird über Differenzbilder bestimmt unter der Annahme, dass der Hintergrund statisch ist, während der Benutzer sich bewegt. Die Kamera zentriert das Gesicht oder die Hände in der Bildmitte, um über die Fokuseinstellung der Kamera den Abstand des Benutzers zu ermitteln. Die Kamera wird den Bewegungen des Benutzers sakkadenartig nachgeführt. In der Bewegungsphase zwischen zwei Fixationspunkten wird die Bildverarbeitung deaktiviert, da die Lokalisation über Bewegungsdetektion eine ruhende Kamera erfordert.

Ein ähnliches Aufmerksamkeitssystem ist von Sidenbladh und Kollegen in dem Projekt „*Intelligent Service Robot*“ realisiert worden [Sid99]. Auch hier ist das Ziel die Entwicklung eines mobilen Serviceroboters, der im Haushalt eingesetzt werden kann. Der Roboter ist mit einer Kamera, einem Laser-Entfernungsmesser und Sonarsensoren ausgestattet. Die Detektion des Benutzers erfolgt jedoch ausschließlich über die Kamera. Mit dem Ziel, das Gesicht zu detektieren, werden in einem Bild alle Pixel daraufhin untersucht, ob sie Hautfarbe aufweisen. Liegt die Anzahl der positiv bewerteten Pixel über einem Schwellwert, gilt der Benutzer als erkannt. Durch die Detektion wird ein *Tracking*-Verfahren initiiert, das die Position der Person schätzt. Der Roboter richtet sowohl die Kamera als auch die gesamte Roboterbasis auf den Benutzer aus.

Die Gestaltung beider Aufmerksamkeitssysteme hängt stark von der dabei eingesetzten Sensorik ab. Da die Kameras jeweils über einen relativ begrenzten Öffnungswinkel verfügen, kann in der

Regel nicht mehr als eine Person zur selben Zeit erfasst werden. Für den Beginn einer Interaktion ist es erforderlich, dass sich der Benutzer in den Einzugsbereich der Kamera begibt. Es ist dann eine präzise Ausrichtung erforderlich, um den Benutzer im Bild zu halten. Die genaue Ausrichtung der Kamera zeigt damit gleichzeitig den Fokus der Aufmerksamkeit des Roboters an. Diese Rückmeldung wird in dem von Sidenbladh beschriebenen System durch die Bewegung der Roboterbasis unterstützt.

Um bei rein visuellen, personenbasierten Aufmerksamkeitssystemen einen größeren Einzugsbereich zu erlangen, sind prinzipiell zwei Strategien denkbar. Entweder wird über die Hardware ein größeres Blickfeld erreicht (Weitwinkelkamera, omnidirektionale Kamera) oder das Aufmerksamkeitssystem steuert die Kamera mit eingeengtem Blickfeld so, dass die für den Roboter relevanten Daten laufend aktualisiert werden. Ein entsprechendes System wird von Brill und Kollegen beschrieben [Bri98]. Hauptbestandteil ist ein so genannter perzeptueller Speicher. Dieser hält die für den mobilen autonomen Roboter relevanten Objekte samt deren Position vor. Es handelt sich dabei um einen Kurzzeitspeicher, der sich auf die lokale Umgebung bezieht. Die gespeicherten Elemente müssen regelmäßig validiert werden. Sie sind dazu mit einem Zuverlässigkeitswert versehen, der von einem Anfangswert zurück auf Null läuft, sobald das zugehörige Objekt außerhalb des Wahrnehmungsbereichs der Sensorik gelangt. Wenn der Zuverlässigkeitswert eines Elements Null erreicht, wird das System alarmiert, das entsprechende Objekt wieder in den Sichtbereich zu bringen. Der Anfangswert der Zuverlässigkeit hängt von der Aufgabe und von den Objekten selbst ab. Sich schnell bewegende Objekte haben beispielsweise gegenüber sich langsam bewegenden Objekten geringere Anfangswerte, da ihre Position häufiger überprüft werden muss. Wasson und Kollegen haben das von Brill beschriebene Aufmerksamkeitssystem auf einem mobilen Roboter eingesetzt [Was99]. Dieser kann damit erfolgreich zwei Personen beobachten, die so weit auseinander stehen, dass sie mit der Kamera nicht gleichzeitig erfasst werden können.

Im Verfahren von Brill wird der Fokus der Aufmerksamkeit des Roboters im Wesentlichen durch die eingeschränkte Sensorik bestimmt. Es eignet sich daher weniger im Anwendungsbereich der Mensch-Roboter-Kommunikation. Hier sollte die Aufmerksamkeit vorwiegend aufgabenbezogen und unabhängig von der Sensorik gesteuert werden. Um ein größeres visuelles Blickfeld zur Verfügung zu haben, werden daher in der Regel Weitwinkelkameras oder omnidirektionale Kameras eingesetzt. Diese werden häufig mit üblichen Kameras kombiniert, die einen kleinen Einzugsbereich aufweisen. Auf diese Weise wird, in Analogie zum menschlichen Sehsystem, peripheres und foveales Sehen imitiert. Insbesondere humanoide Roboter verfügen meistens über entsprechende Kamerakombinationen.

Ein Beispiel ist der Roboter *Cog*. Er wird dazu eingesetzt, um Modelle der Verhaltensforschung zur sozialen Entwicklung bei Kindern zu testen und zu evaluieren. Im Zuge dieses Projekts ist von Scassellati [Sca01] ein Verfahren entwickelt worden, das einen speziellen Aspekt der Aufmerksamkeit realisiert: den gemeinsamen Fokus der Aufmerksamkeit zweier Interaktionspartner (engl. *joint attention*). *Cog* erkennt den Fokus des Interaktionspartners, indem er zunächst im peripheren Bild nach dem Gesicht sucht, dann durch Kamerabewegung das Gesicht in den fovealen, hochauflösenden Bildbereich bringt, anschließend aus der Stellung des Kopfs und der Pupillen die Blickrichtung ermittelt, um schließlich, dem Blick der Person folgend, seine Auf-

merksamkeit auszurichten. Eine ähnliche Vorgehensweise findet man auch bei anderen Arbeiten, in denen die gemeinsame Aufmerksamkeit für humanoide Roboter implementiert wird (vgl. zum Beispiel [Koz01]).

Ein großer Wahrnehmungsbereich bei Kameras bringt den Vorteil, dass sich ein Benutzer nicht in einen kleinen vorgeschriebenen Bereich begeben muss, um vom Roboter erfasst zu werden. Weitwinkelkameras werden aus genau diesem Grund auf dem Roboter *Flo* eingesetzt, der zur Unterstützung hilfsbedürftiger Menschen entwickelt wurde. Da die potenziellen Benutzer in der Regel gehbehindert sind, können sie sich nicht zum Roboter begeben, sondern dieser muss jederzeit den Benutzer lokalisieren können, um auf Wunsch seine Dienste anzubieten. *Flo* dient zum einen als Gehhilfe, leistet zum anderen Gedächtnisunterstützung, gibt aktuelle Informationen und stellt ein Medium zur Kommunikation mit dem Pflegepersonal oder Verwandten dar. In [Roy00] beschreiben Roy und Kollegen, wie der Roboter seine Aufmerksamkeit auf den Benutzer richtet. Er ist mit einem aktiven Stereokamerakopf ausgestattet, deren Kameras über einen Öffnungswinkel von jeweils 100 Grad verfügen. Des Weiteren verfügt er über ein Mikrofon für Spracheingaben. Zur Detektion des Benutzers wird zunächst ein neuronales Verfahren zur Gesichtsdetektion ausgeführt. Dieses benötigt für die Verarbeitung eines einzelnen Bilds etwa vier Sekunden. Nach erfolgreicher Detektion eines Gesichts wird ein schnelles, farbbasiertes *Tracking* angestoßen, um die Person zu verfolgen. Die Kameras fokussieren den Benutzer. Gleichzeitig wird das Mikrofon auf den Benutzer ausgerichtet, um eine niedrigere Wortfehlerrate bei der Spracherkennung zu erreichen. Sprachverarbeitung wird nur für die Äußerungen im akustischen Signal durchgeführt, die mit dem Schlüsselwort „*Flo*“ beginnen. Auf diese Weise wird auch noch eine auditive Aufmerksamkeitssteuerung realisiert, die jedoch zu einer wenig natürlichen Interaktion führt.

#### Berücksichtigung akustischer Stimuli

Die Reaktion auf akustische Stimuli bietet enorme Vorteile für Roboter in der Interaktion mit Menschen. Da omnidirektionale Mikrofone den gesamten Bereich um den Roboter erfassen, muss ein Benutzer sich für eine Interaktion nicht mehr in einen bestimmten Bereich begeben. Er kann einfach über Geräusche auf sich aufmerksam machen. Die Interaktion wird damit für den Anwender erheblich erleichtert, da er sich keine Gedanken darüber machen muss, welche Position er einzunehmen hat.

Auf akustische Reize reagiert der von Park und Kollegen entwickelte Roboter *DO-U-MI* [Par01]. Dieser dient, genauso wie der Roboter *Flo*, zur Unterstützung hilfsbedürftiger Menschen. *DO-U-MI* soll in Pflegeheimen eingesetzt werden und dient als intelligente Gehhilfe. Zusätzlich bietet ein eingebauter Rechner die Möglichkeit, Musik zu hören, Videofilme zu schauen oder E-Mail zu nutzen. Der Roboter ist mit zwei Kameras (Weitwinkel und Tele) und Stereomikrofonen ausgestattet. Der Ansatz für die Aufmerksamkeitssteuerung von *DO-U-MI* ist zunächst recht ähnlich zu dem von *Flo*: Die Kameras werden genutzt, um über ein Gesichtsdetektionsverfahren, basierend auf Hautfarbensegmentierung und *Template-Matching*, den Benutzer zu lokalisieren. Zusätzlich kann der Anwender jedoch den Roboter durch Klatschen auf sich aufmerksam machen. Der Roboter lokalisiert die Schallquelle über die Phasendifferenz der beiden Mikrofonsignale

und richtet daraufhin die Kameras in die entsprechende Richtung, sodass die Gesichtsdetektion den Benutzer finden kann.

Ein weiteres Beispiel für ein Aufmerksamkeitssystem, das akustische Reize berücksichtigt, findet man bei dem von Cheng und Kollegen entwickelten humanoiden Roboter *JACK* [Che01]. Dieser ist in der Lage, menschliche Bewegungen zu imitieren. Der Roboter reagiert auf audiovisuelle Stimuli. Im Bild der Kamera werden der Kopf und die Arme einer Person über Hautfarbe detektiert. Die Position und Bewegung von Kopf und Armen der beobachteten Person werden auf die entsprechenden Bestandteile des humanoiden Roboters übertragen, der so die Bewegungen der Person imitiert. Auditive Stimuli werden über zwei Mikrofone registriert. Für beide Kanäle werden dazu Energiespektren berechnet. Der Roboter richtet sich so aus, dass die Differenz der Spektren minimiert wird. Er dreht sich folglich in Richtung von Geräuschquellen. Wenn der Roboter den Menschen verloren hat und in eine falsche Richtung schaut, ist es somit möglich, über Geräusche die Aufmerksamkeit des Roboters zu erlangen.

Mit dem Zuwenden der Aufmerksamkeit in die Richtung, aus der ein auffälliger akustischer Reiz ausgesendet wurde, imitieren die beiden Roboter *DO-U-MI* und *JACK* die Orientierungsreaktion. Allerdings reagieren beide Systeme immer in derselben Weise, das heißt sobald eine Geräuschquelle lokalisiert wurde, richten sie ihre Kameras in die entsprechende Richtung. Dieser Vorgang geschieht unabhängig vom zeitlichen Verlauf der akustischen Stimuli. Es besteht damit die Gefahr, dass der Aufmerksamkeitsfokus auf einer stetigen Geräuschquelle haften bleibt, die für den Roboter eigentlich irrelevant ist. Beim Menschen wird bei permanenter Darbietung desselben Reizes die Orientierungsreaktion jedoch habituiert, sodass die Zuwendung nicht mehr erfolgt.

Ein Ansatz, der den Effekt der Habituation realisiert, stammt von Stoytchev und Arkin. Sie beschreiben in [Sto01b] einen Roboter, dessen Steuerung über eine *Behavior*-basierte, hybride Architektur erfolgt. Der Benutzer kann die Aufmerksamkeit des Roboters erlangen, indem er Geräusche macht, zum Beispiel in die Hände klatscht. Der Roboter kann die Richtung von Geräuschquellen mit einem binauralen Mikrofonsystem auf wenige Grad genau bestimmen. Ein Bestandteil der von Stoytchev vorgeschlagenen hybriden Roboterarchitektur ist ein Motivations-Subsystem. Dieses überwacht den internen Zustand des Roboters und beeinflusst die Auswahl der aktiven *Behavior*. Wesentliche Bestandteile des Motivations-Subsystems sind so genannte Motivations-Variablen, die jeweils durch eine reelle Zahl zwischen 0 und 1 ihren Aktivierungszustand angeben. Traditionelle *Behavior*-basierte Architekturen definieren wahrnehmungsgesteuerte Auslöser (engl. *perceptual trigger*), die das System von einem Zustand in einen anderen wechseln lassen. Die Auslösebedingung hängt von äußeren Stimuli ab, die der Roboter durch seine Sensorik detektiert. In der vorgestellten Architektur sind sowohl die perzeptuellen Stimuli als auch die Motivations-Variablen für einen Zustandswechsel verantwortlich. *Behavior* können sogar allein durch Motivations-Variablen aktiviert werden. *Behavior* haben sowohl lesenden als auch zum Teil schreibenden Zugriff auf die Motivations-Variablen. Damit kann sich die Arbeitsweise eines *Behavior* in Abhängigkeit des Motivations-Zustands ändern. Andererseits kann ein *Behavior* schreibenderweise Einfluss auf den Motivations-Zustand nehmen.

Mit diesem Ansatz realisieren Stoytchev und Arkin den Effekt der Habituation. Dieser wirkt sich so aus, dass zum Beispiel ein anfänglich auftretendes Geräusch hohe Beachtung findet, aber,

wenn das Geräusch für eine lange Zeit anhält, der Roboter diesen Stimulus ignoriert und sich relevanteren Stimuli zuwenden kann. Des Weiteren können in der vorgeschlagenen Architektur zyklische Veränderungen (zum Beispiel der Tagesrhythmus) und stetige Veränderungen (zum Beispiel zunehmender Hunger) des internen Zustands realisiert werden. In einem Experiment mit einem mobilen Roboter sind vier Motivations-Variablen definiert: Neugierde, Frustration, Heimweh und Ärger. Neugierde modelliert das Interesse an externen Ereignissen, hier insbesondere Geräuschen. Frustration behandelt das Problem, eine Aufgabe nicht bewältigen zu können. Heimweh veranlasst den Roboter, nach dem Beenden einer Aufgabe zu seiner Ladestation zurückzukehren. Ärger ist mit *Behavior* verknüpft, die die Unzufriedenheit mit internen oder externen Ereignissen ausdrücken, zum Beispiel durch eine sprachliche Äußerung. Während die Aktivierungswerte für Neugierde und Heimweh stetig zunehmen, werden sie für Frustration und Ärger stetig kleiner. Der Effekt der Habituation ist in der konkreten Implementierung so realisiert, dass jedes Mal, wenn der Roboter ein Geräusch wahrnimmt und sich diesem Ereignis zuwendet, der Aktivierungswert für Neugierde verkleinert wird. Wenn ein Geräusch also ständig wiederholt wird, reagiert der Roboter darauf nicht mehr. Wenn ein Benutzer dennoch weiter Geräusche produziert, auf die der Roboter wegen fehlender Neugier nicht mehr reagiert, steigt der Aktivierungswert von Ärger, der, sobald er einen Schwellwert überschreitet dafür sorgt, dass die Person vom Roboter durch Sprachausgabe gewarnt wird, ihn nicht weiter zu belästigen.

Für einen Roboter sind Geräusche oder Sprechaktivität nicht immer ein verlässlicher Hinweis darauf, dass eine Person die Absicht hat, mit dem Roboter zu kommunizieren. Wenn jemand zum Beispiel in der Nähe eines Roboters telefoniert, sollte der Roboter die entsprechende Person ignorieren. Um entscheiden zu können, welches akustische Signal für den Roboter relevant ist, bietet es sich an, Spracherkennung in das Aufmerksamkeitssystem zu integrieren.

In der einfachsten Variante wird die Aufmerksamkeit allein durch ein Schlüsselwort aktiviert. Ein Beispiel dafür ist der mobile Büroroboter *Jijo-2*, der von Asoh und Kollegen in [Aso01] beschrieben wird. Sein Einsatzgebiet umfasst Aufgaben, wie zum Beispiel Besucher herumzuführen, Nachrichten zu überbringen oder Treffen zu organisieren. Die Sensorik des Roboters umfasst eine bewegliche Kamera und ein Mikrofonfeld, das aus acht omnidirektionalen Einzelmikrofonen besteht. Immer dann, wenn der Roboter sich in einer Bereitschaftsphase befindet, das heißt keine spezielle Aufgabe ausführt, reagiert er auf das Schlüsselwort „*hello*“. Mit seinem Mikrofonfeld ortet er die Sprachquelle und fokussiert mit der beweglichen Kamera die entsprechende Richtung im Raum. Daraufhin werden im Bild hautfarbene Bereiche lokalisiert. Es wird angenommen, dass die größte hautfarbene Region dem Gesicht des Benutzers entspricht. Die Kamera wird im Weiteren so angesteuert, dass sie die Region in der Mitte des Bilds hält. Zusätzlich wird *Beamforming* [Joh93] eingesetzt, um im akustischen Signal des Mikrofonfelds die Richtung hervorzuheben, in der der Interaktionspartner lokalisiert wurde. Damit werden Umgebungsgeräusche reduziert. Der Ansatz von Asoh und Kollegen realisiert damit eine audiovisuelle Zuwendung der Aufmerksamkeit des Roboters.

Keines der in diesem Abschnitt beschriebenen Verfahren berücksichtigt den Fall, dass sich mehrere Personen in der Nähe des Roboters aufhalten und die Aufmerksamkeitssteuerung eine Selektion vornehmen muss. Die Gründe dafür sind verschieden: Entweder liefert die eingesetzte Sensorik nicht die Daten, mit denen mehrere Personenhypothesen aufgebaut werden können

(Ghidary, Sidenbladh, Stoytchev, Asoh) oder das Forschungsvorhaben liegt nicht im Bereich der Servicerobotik (Scassellati, Cheng) oder das Anwendungsszenario geht nur von einem anwesenden Benutzer aus, wie zum Beispiel bei den Robotern zur Unterstützung hilfsbedürftiger Menschen (Roy, Park). In dem in dieser Arbeit betrachteten *Home-Tour*-Szenario können sich jedoch potenziell mehrere Personen in der Nähe des Roboters aufhalten, von denen jeweils eine für den Fokus der Aufmerksamkeit selektiert werden muss. Aufmerksamkeitssysteme, die zunächst eine Menge von Personenhypothesen aufbauen, um dann eine einzelne auszuwählen, werden im folgenden Abschnitt vorgestellt.

### 2.2.2 Ansätze für mehrere Personen

Als Voraussetzung dafür, dass in Aufmerksamkeitssystemen mehrere Personen gleichzeitig berücksichtigt werden können, muss ein Roboter aus den ihm zur Verfügung stehenden Sensordaten mehrere Personenhypothesen parallel aufbauen können. Zu diesem Zweck verwenden die im Folgenden beschriebenen Ansätze sowohl Sensoren mit großen Einzugsbereichen (omnidirektionale Kameras, Laser-Entfernungsmesser, Infrarotsensorfelder) als auch geeignete *Tracking*-Verfahren, die in der Lage sind, Personenhypothesen auch bei kurzfristigem Ausbleiben von neuer Information aufrecht zu erhalten.

Aufmerksamkeitssysteme unterscheiden sich insbesondere bezüglich der Kriterien, nach denen der Fokus der Aufmerksamkeit auf einzelne Personen gerichtet wird. Die Frage, welche Relevanz einer beobachteten Person zugeordnet wird, hängt sowohl von der betrachteten Anwendung als auch von den erfassten Merkmalen einer Person ab. In einfachen Varianten ist die Bewertung der Relevanz zeitunabhängig und nur von den jeweils aktuell anliegenden Sensordaten abhängig. Aufwändigere Ansätze berücksichtigen den zeitlichen Verlauf der Merkmale einzelner Personen und können damit Effekte wie zum Beispiel die Habituation realisieren. Zunächst werden die einfacheren Ansätze beschrieben.

#### Aufmerksamkeitssysteme mit zeitunabhängiger Relevanzbewertung

Eine besonders einfache Vorgehensweise findet man bei dem Ansatz von Feyrer und Zell [Fey00]. Sie beschreiben ein Verfahren, das es einem mobilen Roboter erlaubt, einer vorausgehenden Person hinterherzufahren. Menschen werden aus einer Kombination von Bild- und Laserdaten detektiert. Jeder detektierten Person wird in einem probabilistischen Verfahren ein Konfidenzwert zugeordnet. Wenn mehrere Personen anwesend sind, selektiert der Roboter einmalig die Person mit dem höchsten Konfidenzwert. Diese recht einfache Lösung ist für komplexere Anwendungen jedoch ungeeignet.

Ein häufig verfolgtes Ziel bei der Gestaltung von Aufmerksamkeitsmechanismen ist es, die Person zu selektieren, die offensichtlich die Absicht hat, mit dem Roboter zu interagieren. In mehreren Ansätzen wird die recht einfache Annahme gemacht, dass der Benutzer sich näher am Roboter aufhält als alle anderen Personen. Der Abstand einer Person ist somit das Merkmal, nach dem sich die Relevanz entscheidet. Genau diese Vorgehensweise haben Doi und Kollegen

für die Steuerung des Roboters *BUGNOID* gewählt [Doi01]. Der Roboter ist mit einer omnidirektionalen Kamera und einem Stereokameraaufbau ausgestattet. In einem Personendetektions-Modus werden zunächst im Bild der omnidirektionalen Kamera über optischen Fluss alle sich bewegende Personen lokalisiert. Dazu ist es erforderlich, dass sich der Roboter nicht bewegt. Der Abstand jeder Person wird aus der Größe der entsprechenden Abbildung im Kamerabild geschätzt. Wenn mehrere Personen detektiert werden, richtet der Roboter seine Aufmerksamkeit auf diejenige mit dem geringsten Abstand. In dem folgenden Personen-*Tracking*-Modus wird Farbinformation über Haut und Kleidung genutzt, um der selektierten Person hinterherzufahren. Sobald die Person stehen bleibt, erfolgt der Wechsel zum Gesichtsdetektions-Modus. Ein *Template*-basiertes Verfahren wird verwendet, um das Gesicht des Benutzers im Stereokamerabild zu lokalisieren und die Kopfstellung zu ermitteln. Kopfbewegung kann eingesetzt werden, um mit dem Roboter zu interagieren. Über Kopfschütteln ist es zum Beispiel möglich, die Interaktionsphase zu beenden. Der Roboter wechselt dann wieder in den Personendetektions-Modus. Mit dem Verfahren wird auf einfacher Ebene zwischen einer Bereitschaftsphase (Personendetektions-Modus) und einer Interaktionsphase (Personen-*Tracking*-Modus, Gesichtsdetektions-Modus) unterschieden.

Ein ähnlich einfaches Selektionskriterium wird von Kröse und Kollegen in [Krö03] beschrieben. Das Aufmerksamkeitssystem ist auf dem Roboter *Lino* realisiert, der in einem so genannten „intelligenten Raum“ eingesetzt wird. Der Raum ist mit Sensoren ausgestattet, mit denen Informationen über die anwesenden Personen gesammelt werden, die dem Roboter zur Verfügung stehen. *Lino* stellt somit die personifizierte Schnittstelle zur intelligenten Umgebung dar. Der Roboter kann natürlichsprachlich mit Benutzern interagieren. Auf der mobilen Plattform des Benutzers ist ein mechanischer Kopf befestigt, der über Stellung von Mund und Augenbrauen Emotionen ausdrückt und über Ausrichtung des Kopfs seinen Aufmerksamkeitsfokus darstellt. Im Kopf des Roboters befinden sich drei jeweils senkrecht zueinander stehende Mikrofonpaare mit einem Mikrofonabstand von 25 cm. Mit diesen können die Positionen der Sprecher im dreidimensionalen Raum bestimmt werden. Der Roboter unterscheidet zufällige Geräusche von menschlicher Stimme, indem die Tonhöhe im Signal berücksichtigt wird. *Lino* wendet seine Aufmerksamkeit dem momentanen Sprecher zu. Wenn mehrere Personen gleichzeitig sprechen, selektiert er diejenige mit der lautesten Stimme. Da das Selektionsverfahren inkonsistente Werte liefert, wenn mehrere Personen gleich laut sprechen, wird die Aufmerksamkeitszuwendung in diesem Fall deaktiviert.

Während die Ansätze von Doi und Kröse von einer sicheren Detektion von Personen ausgehen, basieren einige Verfahren auch auf unsicheren Personenhypothesen, die vor einer Interaktion durch zusätzliche Information bestätigt werden müssen. In diesem Fall dient der Selektionsmechanismus dazu, geeignete Hypothesen auszuwählen, auf die sich die Aufmerksamkeit des Roboters richtet. Die Aufmerksamkeitszuwendung versetzt das System in die Lage, an die für die Bestätigung fehlenden Daten zu gelangen. Diese Vorgehensweise wird beispielsweise von Blanco und Kollegen in [Bla03] beschrieben. Der von ihnen verwendete Roboter mit dem Namen *Albert* ist mit einem Laser-Entfernungsmesser und einer beweglichen Kamera ausgestattet. Zunächst werden die Beine von Menschen in den Laser-Messdaten ermittelt. Es wird dazu eine Hintergrundsubtraktionsmethode eingesetzt, die es erfordert, dass der Roboter sich nicht bewegt.

Die in den Laserdaten gefundenen Personenhypothesen werden nach ihrem Abstand zum Roboter sortiert. Dann wird die Kamera in Richtung der Hypothese mit geringstem Abstand bewegt, um im Bild ein Gesicht zur Bestätigung der Hypothese zu finden. Gelingt dies nicht, wendet sich der Roboter der nächsten Hypothese zu. Das Verfahren steuert somit die Aufmerksamkeit des Roboters auf die am nächsten stehende Person.

Ähnlich gelagert ist der Ansatz von Topp und Kollegen [Top04]. Der von ihnen verwendete mobile Roboter verfügt ebenfalls über eine bewegliche Kamera und einen Laser-Entfernungsmesser. Personenhypothesen werden zum einen über Hautfarbensegmentierung im Bild der Kamera und zum anderen über Bewegungs- und Forminformation in den Laserdaten aufgebaut. Für jeden hautfarbenen Bereich wird der Abstand einer zugehörigen Person aus der Fläche der entsprechenden Bildregion geschätzt. Somit sind für Hypothesen beider Modalitäten Abstands- und Richtungswert bekannt. Diese werden miteinander verglichen und gegebenfalls einander zugeordnet. Nur Hypothesen, die sowohl in den Bild- als auch in den Laserdaten Unterstützung finden, werden weiter untersucht. Zur Verifikation selektiert der Roboter zunächst die Hypothese mit geringstem Abstand zum Roboter, orientiert sich in die entsprechende Richtung und bittet durch Ansprechen („*Kann ich etwas für Sie tun?*“) um eine Bestätigung. Wird durch das sprachverstehende System eine Bestätigung festgestellt, wird die Hypothese als der Kommunikationspartner markiert, woraufhin der Roboter in den Interaktionsmodus wechselt. In diesem hält er seine Aufmerksamkeit auf die Person gerichtet. Wurde eine Ablehnung registriert oder innerhalb einer gewissen Zeitspanne keine Rückmeldung erfasst, wird die Hypothese als irrelevant markiert und die nächste Hypothese untersucht. Die hier verwendete Strategie, dass der Roboter den Benutzer für den Interaktionsaufbau anspricht, eignet sich nicht für das *Home-Tour*-Szenario, da dort davon ausgegangen wird, dass der Benutzer die Initiative für einen Kommunikationsaufbau ergreift.

Ein weiterer Ansatz, in dem unsichere Hypothesen im Aufmerksamkeitsprozess verifiziert werden, stammt von Böhme und Kollegen [Böh03]. Sie entwickeln den Roboter *Perses*, der als Assistent in einem Baumarkt fungiert. Kunden können Warenartikel oder Kaufhausbereiche angeben, zu denen der Roboter sie hinführt. Der Roboter verfügt über eine omnidirektionale Kamera, einen Stereokamerakopf und zwei Mikrofone. Die visuelle Personendetektion erfolgt über Bewegungsinformation im Bild der omnidirektionalen Kamera. Dazu werden statistische Modelle für den Vordergrund und den Hintergrund verwendet, die einzelne Bildbereiche als rechteckige Kästen modellieren. Die erkannten Vordergrundbereiche stellen die Personenhypothesen dar. Aus der zeitlichen Veränderung der Position und Größe der zugehörigen rechteckigen Bereiche werden für jede Hypothese der Abstand, die Bewegungsrichtung und der relative Winkel zur Geradeausrichtung des Roboters geschätzt. Anhand dieser Merkmale erfolgt bei Vorhandensein mehrerer Hypothesen die Selektion einer Person. Der Roboter wählt bevorzugt Personen aus, die sich auf den Roboter zubewegen und sich relativ nah zum Roboter befinden. Wenn mehrere Personen dieselben Kriterien erfüllen, selektiert er diejenige, die für das Zuwenden eine geringere Drehbewegung erfordert.

Die beiden Mikrofone des Roboters erlauben zusätzlich eine akustische Lokalisation. Erkannt werden Händeklatschen oder Zurufen eines Kommandos. Das eingesetzte Verfahren ermittelt die Richtung der Geräuschquelle über die interaurale Zeitdifferenz und die spektralen Beschaf-

fenheiten der aufgezeichneten Tonsignale. Wenn ein auditiver Stimulus registriert wurde und dieselbe Richtung wie eine visuelle Personenhypothese aufweist, wendet der Roboter sich dieser Richtung zu. Wenn auditive und visuelle Stimuli verschiedene Richtungen vorgeben, wird diejenige gewählt, die für die Zuwendung des Roboters eine geringere Drehbewegung erfordert.

Nachdem der Roboter eine Person selektiert hat, orientiert er den Stereokamerakopf in die entsprechende Richtung, um die Hypothese zu verifizieren. Hierzu wird eine Kombination von Hautfarbensegmentierung, Detektion der Kopf-Schulter-Kontur und Gesichtsdetektion eingesetzt. Verifizierte Hypothesen werden fortan über einen Partikelfilteransatz verfolgt, der Ergebnisse von Hautfarbensegmentierung und Abstandsmessungen von Sonarsensoren integriert. Das heißt, der Roboter fixiert seine Aufmerksamkeit auf die detektierte Person für die Zeit der Interaktion.

Wenngleich das Selektionskriterium in dem Ansatz von Böhme und Kollegen ausgefeilter ausfällt als in den zuvor beschriebenen Ansätzen, kann damit nicht garantiert werden, dass sich der Roboter nur auf Personen fixiert, die tatsächlich mit ihm interagieren wollen. Zu diesem Zweck muss auch analysiert werden, was die beobachtete Person sagt. Eine entsprechende, wenn auch einfache Variante, wird von Hashimoto und Kollegen in [Has02] beschrieben. Sie wird auf dem humanoiden Roboter *Hadaly-2* eingesetzt, dessen Kopf mit Kameras und Mikrofonen ausgestattet ist. Der Roboter kann Personen in seiner Umgebung audiovisuell erfassen. Im Bild der Kamera werden Menschen über Bewegung detektiert, und mit den Mikrofonen erfolgt die Lokalisation von Sprache. Die Datenverarbeitung findet auf speziell entwickelter Hardware statt. Die Integration von Bild- und Toninformation erlaubt es dem Roboter zu entscheiden, welche der beobachteten Personen spricht. Er richtet seine Aufmerksamkeit immer auf den Sprecher und verfolgt daraufhin dessen Bewegung. Wenn der Sprecher das Schlüsselwort „*hello*“ nennt, wechselt der Roboter in einen Interaktionsmodus, in der er seine Aufmerksamkeit auf der entsprechenden Person hält. Eine detaillierte Beschreibung des Aufmerksamkeitssystems fehlt jedoch.

#### Aufmerksamkeitssysteme mit zeitabhängiger Relevanzbewertung

Aufwändigere Aufmerksamkeitssysteme beziehen bei der Bewertung der beobachteten Personen auch zeitlich zurückliegende Situationen mit ein. Einen solchen Ansatz stellen Sekmen und Kollegen in [Sek02] für den humanoiden Roboter *ISAC* (siehe [Kaw00]) vor. Der Roboter kann anwesende Personen über drei verschiedene Sensortypen lokalisieren: ein Infrarotsensorfeld, bestehend aus fünf halbkreisförmig angeordneten Sensoren, erlaubt es, warme, sich bewegende Objekte (Menschen) zu detektieren, zwei Mikrofone dienen zur Lokalisation von Sprache, und ein Stereokamerakopf wird genutzt, um Gesichter zu finden. Menschen werden primär über ein farbbasiertes Gesichts-*Tracking*-Verfahren verfolgt. Die Kamera fokussiert während der Interaktion den Kopf des Benutzers. Ist der Kopf im Bild zentriert, wird eine *Template*-basierte Gesichtsdetektion durchgeführt. Der Einzugsbereich der Kamera fällt relativ klein aus, sodass zu jedem Zeitpunkt in der Regel immer nur eine Person erfasst werden kann. Andere gleichzeitig anwesende Personen können jedoch die Aufmerksamkeit des Roboters über Bewegung oder Sprache auf sich ziehen. Wird eine Person über die Infrarotsensoren oder die Mikrofone lokalisiert, dreht

der Roboter die Kamera in die entsprechende Richtung und beginnt, die neu erkannte Person über das Gesicht zu verfolgen. Damit der Roboter nicht „hyperaktiv“ erscheint, weil er auf jedes neue Detektionsereignis mit einer Aufmerksamkeitszuwendung reagiert, schlagen Sekmen und Kollegen folgenden Aufmerksamkeitsmechanismus vor. Für jede der drei Modalitäten wird je eine Funktion $A_i(\cdot)$ eingeführt, die eine Art Aufmerksamkeitspotential modelliert. Immer, wenn ein neues Detektionsereignis innerhalb einer Modalität $i$ vorliegt, nimmt die zugehörige Funktion $A_i(\cdot)$ einen modalitätsabhängigen, positiven Initialwert ein. Von diesem Zeitpunkt an nimmt der Funktionswert exponentiell ab und nähert sich asymptotisch dem Wert Null, bis wieder ein neues Detektionsereignis vorliegt. Das Gesamtpotential $A_{Track}(t)$ zum Zeitpunkt $t$ berechnet sich aus den Einzelfunktionen wie folgt:

$$A_{Track}(t) = A_{Sprache}(t) + A_{Bewegung}(t) - A_{Gesicht}(t)$$

Der Aufmerksamkeitsmechanismus sieht vor, dass eine Aufmerksamkeitszuwendung nur dann stattfindet, wenn das Gesamtpotential von einem negativen zu einem positiven Wert wechselt. Dies kann nur dann auftreten, wenn ein Sprach- oder Bewegungsereignis detektiert wurde und daher $A_{Sprache}(\cdot)$ beziehungsweise $A_{Bewegung}(\cdot)$ einen hohen Wert aufweisen, und gleichzeitig die letzte Aufmerksamkeitszuwendung eine längere Zeitspanne zurückliegt, womit $A_{Gesicht}(\cdot)$ einen kleinen Wert aufweist. Die Aufmerksamkeit wechselt in diesem Fall in Richtung des auslösenden Ereignisses. Durch geeignete Parametrisierung (Initialwerte und Abklingfaktoren) kann die Störbarkeit des Systems durch eine sich nicht im Fokus der Aufmerksamkeit befindliche Person variiert werden. Damit können Effekte eines „höflichen“ gegenüber einem „unhöflichen“ Roboter untersucht werden (vgl. [Rog00]).

Ein weiteres zeitabhängiges Aufmerksamkeitssystem wird von Okuno und Kollegen in [Oku03] beschrieben. Das System kommt auf dem Roboter *SIG* zum Einsatz. Es handelt sich dabei um einen humanoiden Robotertorso, der sich durch Rotation um die vertikale Achse auf Benutzer ausrichten kann. Mit Kameras kann das Gesicht von Menschen lokalisiert und identifiziert werden. Über ein Mikrofonpaar wird die Stimme geortet und die Tonlage bestimmt. Jede Erkennung wird als ein auditives beziehungsweise visuelles Ereignis bezeichnet. Ereignisse innerhalb einer Modalität werden zu so genannten Strömen formiert. Ein visuelles Ereignis wird dem nächsten Strom innerhalb von $\pm 10^\circ$ zugeordnet, wenn die Gesichts-ID übereinstimmt. Entsprechend ist das Vorgehen bei auditiven Ereignissen, wobei die Stimmlage von Ereignis und Strom zueinander passen müssen. Ereignisse, die nicht zugeordnet werden können, bilden einen neuen Strom. Ein auditiver und ein visueller Strom werden zu einem assoziierten Strom verknüpft, wenn die Richtungsdifferenz der Ströme für eine gewisse Zeitspanne $10^\circ$ unterschreitet. Wächst die Richtungsdifferenz auf über $30^\circ$, werden sie wieder deassoziiert. Ein Strom löst sich auf, wenn ihm für eine halbe Sekunde kein Ereignis zugeordnet wurde. Mit der beschriebenen Technik werden die Personen, die von den Sensoren des Roboters erfasst werden, über die Zeit verfolgt.

Das Aufmerksamkeitssystem von *SIG* setzt auf den detektierten Strömen auf. Jeder Strom wird dazu mit einem Relevanzwert belegt. Die Aufmerksamkeit des Roboters richtet sich immer auf den Strom mit dem höchsten Wert. Der Relevanzwert wird einem Strom bei seinem Aufbau zugewiesen und hängt vom Status des Stroms (auditiv, visuell oder assoziiert) ab. Der Wert nimmt mit der Zeit kontinuierlich ab, wodurch der Effekt der Habituation realisiert wird. Unterschiedliche

Verhaltensweisen in der Aufmerksamkeit werden durch geeignete Wahl der initial zugewiesenen Relevanzwerte realisiert. Ein aufgabenorientiertes Verhalten erreichen Okuno und Kollegen für den Roboter *SIG*, indem assoziierte Ströme einen Wert von 2 erhalten, während unimodale Ströme mit einem Wert von 1 belegt werden. Der Roboter kann so seine Aufmerksamkeit auf einem Benutzer während der Interaktion halten und neu entstehende unimodale Ströme ignorieren. Für ein soziales Verhalten werden allen Strömen identische Relevanzwerte zugeordnet, sodass der Roboter seine Aufmerksamkeit immer auf neu auftauchende Ströme richtet. In einer Gruppe von Menschen wendet er sich immer dem Sprecher zu und stellt somit einen neugierigen Zuhörer dar. Das System weist mit seinen zwei realisierten Verhaltensweisen gewisse Ähnlichkeiten zu dem in dieser Arbeit vorgestellten Ansatz auf, der zwischen einer Bereitschaftsphase, in der der Roboter sich auf Sprecher ausrichtet, um neue Benutzer zu erkennen, und einer Interaktionsphase, in der der Roboter seine Aufmerksamkeit durchgehend auf dem Benutzer hält, unterscheidet. Okuno und Kollegen geben jedoch keine Strategie an, die einen Wechsel von Verhaltensweisen festlegt.

Der Ansatz von Matsusaka und Kollegen ist ebenfalls als zeitabhängiges Aufmerksamkeitssystem anzusehen. Sie beschreiben in [Mat03] den mobilen humanoiden Roboter *ROBITA*, der in der Lage ist, an einem Gespräch mit zwei Personen zu einem gemeinsamen Thema teilzunehmen. Der Roboter folgt dem Konversationsfluss, weiß also wer spricht und zu wem gesprochen wird und richtet seine Aufmerksamkeit in Abhängigkeit vom Gesprächsverlauf aus. *ROBITA* besteht aus einem Torso-ähnlichen Aufbau mit zwei Armen und einem Kopf auf einer mobilen Plattform. Um seine Gesprächspartner wahrzunehmen, verfügt der Roboter in seinem Kopf über zwei Mikrofone und eine Stereokamera. Er lokalisiert die Sprachquelle mit den Mikrofonen. Mit Hilfe der Kameras werden die Blickrichtung (bei einer Genauigkeit von 30°) und die Identität der Gesprächspartner ermittelt.

Die Aufmerksamkeitssteuerung von *ROBITA* dient im Wesentlichen zwei Zwecken: Durch Ausrichtung der Kamera auf den aktuellen Sprecher ist das System in der Lage, die Kopfstellung des Sprechers zu bestimmen, um daraus abzuleiten, zu wem gesprochen wird. Daneben wird die Ausrichtung des Roboterkopfs und des Rumpfs dazu verwendet, den Gesprächsteilnehmern zu zeigen, auf wen der Roboter seine Aufmerksamkeit richtet.

Die Hauptaufgabe des Systems ist es, die Rolle des Roboters und die der anderen Personen im zeitlichen Verlauf des Gesprächs zu ermitteln. Gesprächsteilnehmer können in einer Mehrpersonenkonversation eine von drei Rollen einnehmen: Sprecher, primärer Empfänger oder sekundärer Empfänger beziehungsweise Beobachter. Der Sprecher adressiert den primären Empfänger. Beide schauen sich in der Regel gegenseitig an. Alle anderen Personen sind Beobachter und schauen zum Sprecher. Der Roboter richtet seinen Kopf und seinen Rumpf entsprechend seiner Rolle aus, das heißt als Sprecher wendet er sich dem primären Empfänger zu, und als primärer Empfänger oder Beobachter schaut er zum Sprecher. Wenn der Roboter in der Rolle des Sprechers ist und ein Beobachter dazwischen spricht, wendet der Roboter den Kopf zu dem entsprechenden Beobachter und bittet ihn zu warten. Der Rumpf wird währenddessen auf den primären Empfänger ausgerichtet gehalten, um die eigentliche Absicht des Roboters, mit dieser Person zu sprechen, anzuzeigen.

Tabelle 2.1: Die beschriebenen Aufmerksamkeitssysteme im Überblick. In der Spalte der Sensoren sind diejenigen aufgelistet, die im Aufmerksamkeitssystem eingesetzt werden. Die Buchstaben bedeuten: M – Mikrofon, K – Kamera, I – Infrarotsensor, L – Laser-Entfernungsmesser

| Name | | Sensoren | Bewertung der Relevanz | Personen |
|---|---|---|---|---|
| Breazeal [Bre99] | (*Kismet*) | K | Bedürfnisse, Habituation | eine |
| Déniz [Dén03] | | M, K | multimodal | mehrere |
| Hambuchen [Ham04] | (*ISAC*) | M, K, I | multimodal, Aufgabe, Habit. | mehrere |
| Ghidary [Ghi02] | | K | – | eine |
| Sidenbladh [Sid99] | | K | – | eine |
| Scassellati [Sca01] | (*Cog*) | K | – | eine |
| Roy [Roy00] | (*Flo*) | K | – | eine |
| Park [Par01] | (*DO-U-MI*) | M, K | – | eine |
| Cheng [Che01] | (*JACK*) | M, K | – | eine |
| Stoytchev [Sto01b] | | M | Habituation | eine |
| Asoh [Aso01] | (*Jijo-2*) | M, K | Schlüsselwort | eine |
| Doi [Doi01] | (*BUGNOID*) | K | Abstand | mehrere |
| Kröse [Krö03] | (*Lino*) | M | Lautstärke | mehrere |
| Blanco [Bla03] | (*Albert*) | K, L | Abstand | mehrere |
| Topp [Top04] | | K, L | Abstand, interakt. Bestätigung | mehrere |
| Böhme [Böh03] | (*Perses*) | M, K | Gehrichtung, relativer Winkel | mehrere |
| Hashimoto [Has02] | (*Hadaly-2*) | M, K | Schlüsselwort | mehrere |
| Sekmen [Sek02] | (*ISAC*) | M, K, I | multimodal, Habituation | mehrere |
| Okuno [Oku03] | (*SIG*) | M, K | multimodal, Zustand, Habit. | mehrere |
| Matsusaka [Mat03] | (*ROBITA*) | M, K | multimodal, Konversation | zwei |

Obwohl das System eine ausgefeilte Aufmerksamkeitssteuerung für eine Kommunikationssituation mit mehreren Gesprächsteilnehmern aufweist, bleibt das Verfahren auf ein spezielles Szenario beschränkt. Der beschriebene Ansatz beschränkt sich auf ein statisches Szenario mit genau zwei Gesprächsteilnehmern. Um anhand der Kopfstellung des Sprechers den Adressaten zu ermitteln, muss der Roboter nur unterscheiden, ob der Sprecher den Roboter anschaut oder aber die andere Person. Wenn weitere Personen an dem Gespräch teilnehmen würden, müsste die Kopfstellung mit einer sehr hohen Genauigkeit festgestellt werden, um den primären Empfänger bestimmen zu können. Zudem ist das System nicht autonom, da es auf externe Rechenkapazität angewiesen ist.

## 2.3 Zusammenfassung

Tabelle 2.1 verschafft einen Überblick über die in diesem Kapitel beschriebenen Aufmerksamkeitssysteme. Die Auflistung beschränkt sich auf die Systeme, die primär für Anwendungen im

Bereich Mensch-Roboter-Interaktion konzipiert wurden. Die Ansätze sind in drei Kategorien unterteilt: ortsbasierte Ansätze, die mit Hilfe von Aufmerksamkeitskarten realisiert sind (erster Block), personenbasierte Ansätze, die nur eine Person berücksichtigen (zweiter Block), und personenbasierte Ansätze, die im Selektionsprozess zwischen mehreren Personen unterscheiden (dritter Block).

Obwohl es das gemeinsame Ziel aller Systeme ist, dass ein Roboter für die Interaktion seine Aufmerksamkeit auf einen Interaktionspartner richtet, fallen die Konzepte recht unterschiedlich aus. Dies liegt vor allem daran, dass viele Aufmerksamkeitssysteme im Hinblick auf eine konkrete Anwendung und für einen speziellen Roboter entwickelt wurden. Relativ einfache Relevanzbewertungsmechanismen finden sich häufig dort, wo die eingeschränkte Sensorik oder die eingesetzten Musterklassifikationstechniken nur die Berechnung von wenig aussagekräftigen oder wenig zuverlässigen Merkmalen zulassen. Ausgefeiltere Mechanismen greifen meistens auf multimodale Information zurück. Dabei berechnet sich die Relevanz in der Regel als gewichtete Summe von unimodalen Merkmalen. In den meisten dieser Ansätze wird auch der Effekt der Habituation modelliert.

# Kapitel 3

# Der mobile Roboter *BIRON*

In diesem Kapitel wird zunächst der verwendete Roboter *BIRON* vorgestellt, für den das in dieser Arbeit beschriebene Aufmerksamkeitssystem entwickelt wurde. Besonderes Augenmerk richtet sich dabei auf die vorhandene Sensorik, da sie bei der Gestaltung des Aufmerksamkeitssystems eine wichtige Rolle spielt.

## 3.1 Hardware

Abbildung 3.1 zeigt den in dieser Arbeit verwendeten mobilen Roboter. Er wird in der Arbeitsgruppe „Angewandte Informatik“ der Technischen Fakultät an der Universität Bielefeld zur Forschung im Bereich der Mensch-Roboter-Interaktion eingesetzt. Der Roboter trägt den Namen *BIRON* (*Bielefeld Robot Companion*). Es handelt sich bei dem Roboter um das Modell *PeopleBot* der Firma *ActivMedia Robotics*, das um einige Komponenten erweitert wurde. Der *PeopleBot* besteht aus einer mobilen Basisplattform[1] mit einem turmartigen Aufbau, dessen obere Plattform eine Höhe von 105 cm erreicht. Auf diese wurde bei *BIRON* ein Aluminiumgestell montiert, das als Halterung für einen Flachbildschirm und eine Kamera dient. *BIRON* erreicht mit diesem Aufbau eine Gesamthöhe von 145 cm. Durch die hohe und schlanke Bauweise eignet er sich insbesondere für die Interaktion mit stehenden Menschen.

Die Mobilität des Roboters wird über zwei seitlich angebrachte Antriebsräder und eine kleine Laufrolle, die sich hinten an der Basis befindet, erreicht. Der Roboter kann eine maximale Translationsgeschwindigkeit von 0,8 m/s und eine maximale Rotationsgeschwindigkeit von 130°/s erreichen. Dies genügt, um den Roboter zum Beispiel vorausgehenden Menschen folgen zu lassen oder ihn in der Interaktion schnell genug auf den Benutzer auszurichten. Die Räder des Roboters erlauben einen Einsatz auf ebenem Untergrund, wobei kleine Unebenheiten wie Türschwellen und Teppichkanten überwunden werden können.

Standardmäßig ist das Modell *PeopleBot* mit Anstoß- und Sonarsensoren ausgestattet:

[1] entspricht dem Modell *Pioneer 2-DX* mit stärkeren Antriebsmotoren

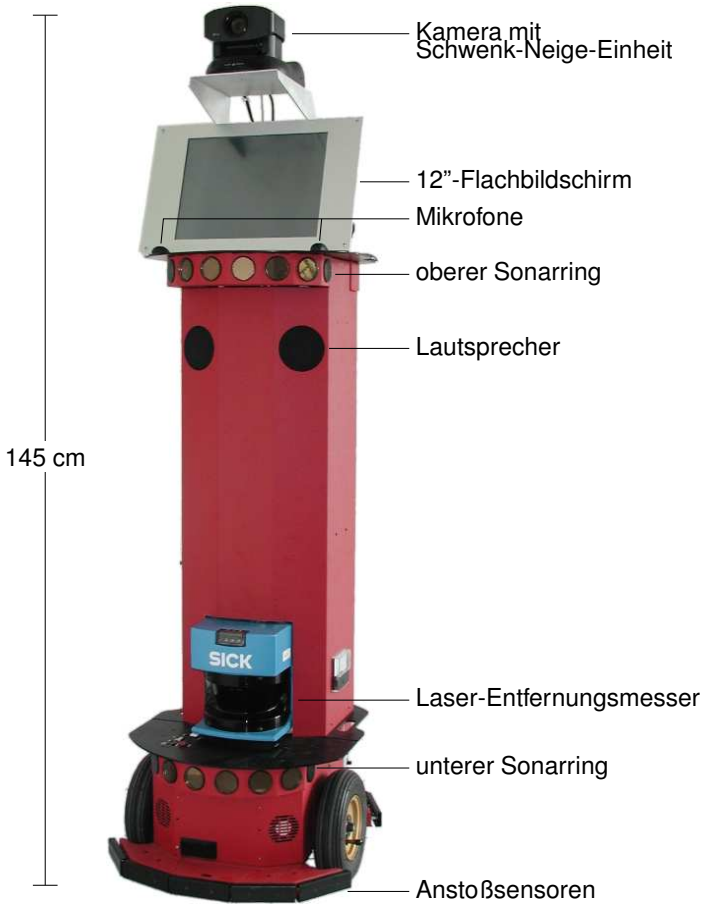



Abbildung 3.1: *BIRON*

**Anstoßsensoren:** Die Anstoßsensoren befinden sich an der unteren Kante der Roboterbasis. Sie reagieren auf Krafteinwirkung und dienen als Sicherheitsvorrichtung, um bei Kontakt mit Hindernissen den Antrieb des Roboters abzuschalten.

**Sonarsensoren:** Die Sonarsensoren befinden sich in zwei horizontalen Achtergruppen auf Höhe von etwa 20 cm und 100 cm an der Vorderseite des Roboters. Sonarsensoren bestimmen über Ultraschall den Abstand zu Hindernissen. Der Abstand $d$ errechnet sich aus der gemessenen Zeitspanne $\Delta t$ zwischen Aussenden eines gerichteten Schallimpulses und Empfang der Reflexion nach $d = c_s \, \Delta t/2$, wobei $c_s$ die Schallgeschwindigkeit ist. Jeder der beiden Sonarringe hat aufgrund der kleinen Anzahl von Einzelsensoren nur eine geringe Winkelauflösung. Die Messgenauigkeit beträgt ungefähr 10 cm im Bereich bis 5 m. Sonarsensoren werden vorwiegend zur Kollisionsvermeidung eingesetzt. Da die Gefahr von Zusammenstößen mit Hindernissen in dem bei *BIRON* betrachteten Interaktionsszenario eine untergeordnete Rolle spielt, werden die Sonarsensoren nicht eingesetzt. Da die Sensoren, obwohl Messungen auf Ultraschall basieren, Klickgeräusche erzeugen, und damit die akustische Datenverarbeitung erschweren, sind die Sonarringe auf *BIRON* vollständig

deaktiviert.

*BIRON* verfügt über weitere Sensoren, die nachgerüstet wurden:

**Laser-Entfernungsmesser:** Direkt auf der unteren Basisplattform ist ein auf Lasertechnik basierender Entfernungsmesser (Modell *LMS 200* der Firma *SICK*) montiert. Er erfasst Objekte auf einer Höhe von 30 cm innerhalb eines 180° Öffnungswinkels im Bereich vor dem Roboter. Laser-Entfernungsmesser werden auf vielen mobilen Forschungsrobotern zur Navigation und Lokalisation eingesetzt. In dieser Arbeit dient der Sensor zur Detektion von Beinen.

Der Laser-Sensor misst Entfernungen zu Objekten nach dem so genannten *Time-of-Flight*-Prinzip, das heißt es wird aus der Zeit $\Delta t$, die zwischen Emission eines Laserimpulses und dem Empfang der Reflexion des Impulses an einem Hindernis vergeht, der Abstand als $d = c_l \, \Delta t/2$ ermittelt, wobei $c_l$ die Geschwindigkeit von Licht ist. Ein rotierender Spiegel lenkt die Laserimpulse um und fächert sie innerhalb der horizontalen Ebene auf. Das eingesetzte Sensormodell ist so konfiguriert, dass innerhalb eines 180° Öffnungswinkels mit einer Winkelauflösung von 0,5° gemessen wird. Jede Messung resultiert folglich in einem Datensatz bestehend aus 361 Abstandswerten. Ein Beispiel zeigt Abbildung 6.1 auf Seite 65. Die Reichweite des Laser-Entfernungsmessers ist 30 m. Der statistische Fehler beträgt $\pm$15 mm in dem für die geplante Anwendung relevanten Bereich bis 8 m. Obwohl Messungen mit 25 Hz durchgeführt werden, wird die Rate aufgrund der Übertragung der Daten über eine serielle Schnittstelle auf 4,7 Hz begrenzt.

**Mikrofone:** Auf der oberen Plattform sind zwei Mikrofone (Modell *C 400 BL* der Firma *AKG*) angebracht. Es handelt sich um Grenzflächenmikrofone, deren Frequenzbereich zur Aufnahme von Sprache optimiert ist. Sie befinden sich vor dem Flachbildschirm und haben einen Abstand von 28,1 cm zueinander. Die Mikrofone werden auf *BIRON* für die Lokalisation von Sprechern und für die Sprachverarbeitung eingesetzt.

**Kamera:** Auf dem Aluminiumgestell, das als Befestigung für den Flachbildschirm dient, ist auf einer Höhe von 135 cm eine Farbkamera (Modell *EVI-D31* der Firma *SONY*) montiert. Das Kameramodell verfügt über eine Schwenk-Neige-Einheit, die es erlaubt, die Kamera in der Horizontalen seitlich um $\pm$100° zu schwenken und in der Vertikalen nach oben und unten um $\pm$25° zu neigen. Die maximale Geschwindigkeit liegt in der Horizontalen (Vertikalen) bei 80°/s (50°/s). Dies erlaubt ein relativ zügiges Ausrichten auf einen gewünschten Zielpunkt. Die Kamera verfügt über ein Zoomobjektiv mit einer Brennweite von 5,4 mm bis 64,8 mm. Im Weitwinkelbereich beträgt der horizontale (vertikale) Bildwinkel 48,8° (37,6°) und im Telebereich 4,3° (3,3°). In dieser Arbeit wird die Kamera ausschließlich im Weitwinkelmodus betrieben, um einen möglichst großen Einzugsbereich zu erzielen. Die Kamera verfügt über die Standard PAL Auflösung. Um schnellere Raten bei der Verarbeitung einzelner Bilder zu erzielen, wird im Rahmen dieser Arbeit auf skalierten Bildern mit einer Größe von 256$\times$192 Pixeln gearbeitet.

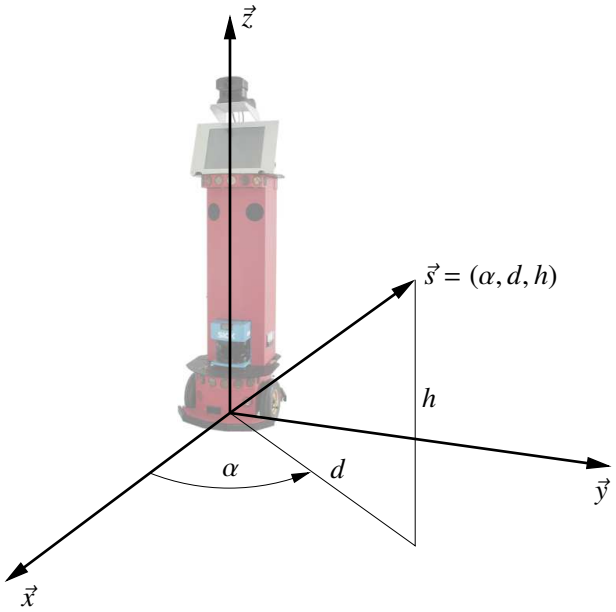



Abbildung 3.2: Lokales Koordinatensystem von *BIRON*. $\vec{x}$ ist die Geradeausrichtung und entspricht 0°. Punkte werden aus Robotersicht in Zylinderkoordinaten angegeben.

Das Aluminiumgestell, auf dem die Kamera montiert ist, dient auch als Befestigung für einen berührungsempfindlichen 12”-Flachbildschirm. Durch seine leicht geneigte Anbringung auf einer Höhe von ungefähr 120 cm eignet er sich gut für interaktive Eingaben durch den Benutzer. In dieser Arbeit wird der Bildschirm jedoch nur zur Darstellung eines animierten Gesichts eingesetzt. Des Weiteren sind in der vorderen Front des Roboterturms auf einer Höhe von 90 cm Stereolautsprecher eingebaut, die für Sprachausgaben genutzt werden.

Das Modell *PeopleBot* ist standardmäßig mit einem *Intel-Pentium-III*-System mit 850 MHz Prozessortakt und 256 MB Hauptspeicher ausgestattet. Der Rechner verfügt über eine Soundkarte, an welche die beiden Mikrofone angeschlossen sind. Der Roboter *BIRON* wurde um einen zusätzlichen Rechner mit *Intel-Pentium-III*-Prozessor erweitert. Dieser wird mit einer Taktrate von 500 MHz betrieben und ist ebenfalls mit 256 MB Hauptspeicher bestückt. Er ist mit einer handelsüblichen TV-Karte ausgestattet, die zur Aufnahme von Bildern der Kamera dient. Der Rechner wird des Weiteren zur Ansteuerung der Schwenk-Neige-Einheit der Kamera über eine serielle Schnittstelle eingesetzt. Die beiden Rechner von *BIRON* sind über 100-Mbit-*Ethernet* miteinander verbunden. Zudem besteht die Möglichkeit über Funk-*Ethernet* von einem externen Rechner auf das Robotersystem zuzugreifen.

Die Stromversorgung erfolgt über drei Hochleistungs-Bleiakkus, die bei vollem Betrieb eine maximale Einsatzdauer von ungefähr 30 Minuten erlauben. Alternativ lässt sich ein externes Netzteil anschließen, wodurch die Einsatzzeit prinzipiell unbegrenzt ist.

## 3.2 Lokales Koordinatensystem

Zur einheitlichen Repräsentation der Positionen von Objekten wird für *BIRON* ein lokales Koordinatensystem definiert (siehe Abbildung 3.2). Der Ursprung des Koordinatensystems befindet sich auf der zentralen, vertikalen Achse des Roboters auf Bodenhöhe. Die $\vec{x}$-Achse des rechtwinkligen Koordinatensystems ist nach vorne, in Fahrtrichtung orientiert. Die $\vec{y}$-Achse verläuft parallel zur Achse der Antriebsräder und zeigt aus Sicht des Roboters nach links. Die $\vec{z}$-Achse zeigt nach oben. Die Position eines Objekts wird in Zylinderkoordinaten durch Winkel $\alpha$, Abstand $d$ und Höhe $h$ beschrieben.

# Kapitel 4

# Konzeption des Aufmerksamkeitssystems

In diesem Kapitel wird die Konzeption des Aufmerksamkeitssystems vorgestellt. Sie verschafft einen Überblick über den in dieser Arbeit entwickelten Ansatz. Die Beschreibung technischer Details von einzelnen Komponenten erfolgt im Anschluss an dieses Kapitel.

Das Aufmerksamkeitssystem soll den Roboter befähigen, seine Wahrnehmung in der Interaktion mit Benutzern zu optimieren, sodass den im Interaktionssystem verwendeten Komponenten (Sprachverarbeitung, Gestenerkennung, ...) über die Sensoren jederzeit geeignete Daten zur Verfügung stehen. Die Realisierung des Gesamtsystems erfolgt auf dem mobilen Roboter *BIRON*. Als Beispiel für ein Interaktionsszenario wird das *Home-Tour*-Szenario gewählt.

Das hier vorgestellte System realisiert ausschließlich eine auf Personen gerichtete Aufmerksamkeit. Selbstverständlich ist es notwendig, den Fokus der Aufmerksamkeit auch auf Objekte zu richten, die vom Benutzer in der Interaktion sprachlich oder durch deiktische Gesten referenziert werden. Diese Anforderung wird zwar berücksichtigt, sodass sie durch das Gesamtsystem geleistet werden kann. Eine allgemeine Aufmerksamkeitssteuerung, die gleichzeitig Menschen und Objekte einbezieht, liegt jedoch außerhalb des in dieser Arbeit gesteckten Ziels.

Die grundsätzliche Notwendigkeit einer personenbezogenen Aufmerksamkeitssteuerung ergibt sich im Wesentlichen aus zwei Gründen. Zum einen können die Sensoren aufgrund technisch bedingter Einschränkungen der jeweiligen Einzugsbereiche nicht immer alle notwendigen Informationen aufnehmen und müssen daher in geeigneter Weise ausgerichtet werden. Zum anderen liefert die Darstellung des Aufmerksamkeitsfokus einen wichtigen Beitrag für eine intuitive Interaktion mit dem Benutzer.

Für die Gestaltung der Aufmerksamkeitssteuerung ist es erforderlich, die Anforderungen an das System genauer zu beleuchten. Unabhängig von dem konkreten Einsatzgebiet lassen sich für einen Serviceroboter, der seine Dienste Menschen zur Verfügung stellt, zwei Phasen voneinander unterscheiden. Einerseits gibt es die Interaktionsphase mit einem Benutzer.[1] Da die Dauer

[1] Hier werden nur eins-zu-eins-Situationen betrachtet. Eine Mehrpersonenkonversation, wie sie zum Beispiel mit dem Roboter *Robita* [Mat01] realisiert ist, wird nicht berücksichtigt.

einer Interaktion endlich ist, gibt es andererseits für den Roboter immer eine Zeitspanne ohne Kommunikationspartner. In dieser Bereitschaftsphase ist er bestrebt, seine Dienste anderen Menschen anzubieten und wieder in die Interaktionsphase zu wechseln.

## 4.1 Bereitschaftsphase

Werfen wir einen genaueren Blick auf die beiden Phasen und beginnen mit der Bereitschaftsphase. Es gibt verschiedene Verhaltensstrategien für den Roboter, um wieder in eine Interaktion zu gelangen. Ausgeschlossen werden unnatürliche, technisch anmutende Lösungen, wie zum Beispiel die in [Bur98] beschriebene, bei der der Benutzer zum Beginnen einer Interaktion einen Knopf am Roboter zu drücken hat. Eine geeignete Möglichkeit ist, dass der Roboter selber aktiv wird, indem er Personen in seiner Nähe anspricht. Diese Strategie wird zum Beispiel von Topp und Kollegen auf ihrem Roboter angewandt [Top04]. Diese Vorgehensweise kann jedoch störend wirken, wenn der Roboter Personen kontaktiert, die sich gerade im Gespräch miteinander befinden, oder wenn er Menschen mehrfach anspricht, die keine Hilfe wollen. Eine andere Möglichkeit ist, dass der Roboter sich passiv verhält und darauf wartet, von einem Benutzer angesprochen zu werden. In [Aso01] wird zum Beispiel die Nennung eines Schlüsselworts vom Benutzer gefordert, damit der Roboter den Beginn der Interaktionsphase erkennt. Nachteilig dabei ist, dass der Nutzer dieses Schlüsselwort kennen muss.

In dieser Arbeit wird ein flexiblerer Mechanismus eingesetzt, der auf natürlichen Verhaltensmustern basiert. Dass wir angesprochen werden erkennen wir Menschen daran, dass jemand in unserer Nähe spricht und seine Äußerung an uns richtet. Dabei ist der Inhalt der Äußerung gar nicht von so großer Bedeutung: Wir fühlen uns auch dann angesprochen, wenn es sich dabei um eine fremde Sprache handelt. Wichtig ist also zu erkennen, an wen die Äußerung gerichtet ist. In der Regel schaut ein Sprecher seinen Adressaten an. Die Blickrichtung ist daher das zweite wichtige Merkmal neben der Sprechaktivität. Diese Merkmale werden vom Roboter genutzt, um einen Interaktionsbeginn zu erkennen. Der Sprecher wird dabei mit den zwei Mikrofonen lokalisiert. Die Blickrichtung wird aus dem Bild der Kamera erkannt. Genauer gesagt wird lediglich das Gesicht detektiert und vereinfachend angenommen, dass die Ausrichtung des Kopfs und die Blickrichtung übereinstimmen. Menschen erfassen selbstverständlich weitere Merkmale, die ihnen das Erkennen des Interaktionsbeginns erleichtern, wie zum Beispiel die Mimik des Sprechers oder der situative Kontext. Da diese Information für ein technisches System nur schwer zu gewinnen sind, wird auf eine Umsetzung in dieser Arbeit verzichtet.

Allein aus der Kombination von Sprechaktivität und Blickrichtung zum Roboter lässt sich ein Interaktionsbeginn jedoch nicht sicher ableiten. So kann zum Beispiel eine Person mit einer anderen über den Roboter sprechen und dabei zwischenzeitlich den Roboter anschauen. Beide Merkmale sind gegeben, ohne dass der Sprecher mit dem Roboter in Kontakt treten möchte. Um den Anmeldeprozess robuster zu gestalten, wird daher auch der Inhalt der Äußerung berücksichtigt. Es muss sich um eine Äußerung handeln, die sich in dem betrachteten *Home-Tour*-Szenario als Beginn einer Interaktion eignet, wie zum Beispiel *„Hallo BIRON!“* oder *„Folge mir bitte!“*.

Alle Merkmale müssen für eine Person zeitgleich vorliegen und vom System erkannt werden. In dem dynamischen Szenario mit einem mobilen Roboter besteht jedoch das Problem, dass den potenziellen Kommunikationspartnern kein fester Standort zugeschrieben werden kann. Da die Kamera von *BIRON* in der Regel nicht mehr als eine Person erfassen kann, liegt das Merkmal „Blickrichtung zum Roboter“ nur für die Person vor, auf die die Kamera gerichtet ist. Außerdem darf nicht jede sprachliche Äußerung interpretiert werden, da das von der Sprachverarbeitung verwendete Lexikon im Wesentlichen auf das *Home-Tour*-Szenario begrenzt ist, und daher die Gefahr besteht, dass nicht aufgabenbezogene Äußerungen falsch interpretiert werden.

Die von der Aufmerksamkeitssteuerung in der Bereitschaftsphase zu durchlaufenden Teilschritte sehen unter diesen Anforderungen wie folgt aus:

1. Es wird eine Liste der Menschen erstellt, die sich in der Nähe des Roboters aufhalten, da diese die Menge der potenziellen Kommunikationspartner darstellen. Dieser Schritt wird über ein geeignetes Verfahren zum Verfolgen (engl. *Tracking*) von Personen realisiert. Um möglichst viele Personen in der Nähe des Roboters zu erfassen, greift das Verfahren auf den Laser-Entfernungsmesser zurück, der mit seinem Einzugsbereich von 180° den gesamten Halbraum vor dem Roboter abdeckt. In den Laserdaten werden Menschen über ihre Beine lokalisiert. Da Menschen den Wahrnehmungsbereich des Roboters betreten und wieder verlassen, muss das Verfahren in der Lage sein, eine variable Anzahl von Personen zu verfolgen.

2. Die Relevanz der beobachteten Personen wird aus den Merkmalen „Sprechaktivität“ und „Blick zum Roboter“ bestimmt. Diese Information wird ebenfalls aus dem *Tracking*-Verfahren gewonnen, das heißt die Laserdaten werden mit audio-visuellen Sensordaten integriert. Es handelt sich folglich um ein multimodales Verfahren zum Verfolgen von Personen. Der multimodale Ansatz bietet den Vorteil, robuster gegenüber dem Ausbleiben von Information einzelner Modalitäten zu sein. Dieser Überlegung folgend wird das *Tracking*-Verfahren um eine zusätzliche Informationsquelle erweitert, sodass neben den Beinen, der Stimme und dem Gesicht auch der Oberkörper im Bild der Kamera über die Farbe der Kleidung lokalisiert wird.

3. Der Fokus der Aufmerksamkeit wird auf die Person mit höchster Relevanz gerichtet, welche in der Regel der aktuelle Sprecher ist. Das Ausrichten geschieht durch das Bewegen der Kamera. Somit gelangt der Sprecher ins Kamerabild, sodass auch die Blickrichtung überprüft werden kann. Die Sprachverarbeitung wird immer dann aktiviert, wenn eine Person gleichzeitig spricht und zum Roboter schaut, da angenommen wird, dass dann die Äußerung an ihn adressiert ist. Gleichzeitig richtet der Roboter durch Rotation der Basis seine Mikrofone auf den Sprecher, um eine bessere Erkennungsrate der Sprachverarbeitung zu ermöglichen. Darüber hinaus wird die Positionsinformation aus der Sprecherlokalisation genutzt, um akustische Signale aus Richtung des Sprechers zu verstärken.

Die Aufmerksamkeitssteuerung der Bereitschaftsphase ist für den Beobachter an der Bewegung der Kamera und der Rotation der Basis zu erkennen. Die Kamera, die immer den Kopf des

Sprechers fokussiert, signalisiert diesem damit, dass der Roboter zuhört und bereit ist. Obwohl der Roboter nicht aus eigenem Antrieb einen Interaktionsaufbau initiiert, nimmt er wegen der durch die Aufmerksamkeitssteuerung realisierten Fähigkeiten aktiv an diesem Prozess teil.

## 4.2 Interaktionsphase

Wenn für eine Person alle der drei erforderlichen Merkmale erkannt wurden, so ist ein neuer Benutzer gefunden, und es folgt der Wechsel zur Interaktionsphase. Prinzipiell werden weiterhin dieselben Teilschritte durchlaufen, wie sie für die Bereitschaftsphase beschrieben wurden. Allerdings wird die Relevanz nicht mehr anhand der dort betrachteten Merkmale bestimmt, sondern der Benutzer hat automatisch die höchste Relevanz. Das heißt der Roboter fokussiert seine Aufmerksamkeit durch Ausrichten von Kamera und Roboterbasis allein auf den Kommunikationspartner, während er andere anwesende Personen ignoriert. Die Aktivierung der Sprachverarbeitung verhält sich analog zur Bereitschaftsphase. Die Aktivierung erfolgt, wenn der Benutzer spricht und gleichzeitig zum Roboter schaut und bleibt so lange eingeschaltet, bis die Äußerung beendet ist, also eine Pause in der Sprechaktivität des Kommunikationspartners vorliegt. Gegen eine dauerhafte Aktivierung spricht der unnötige Verbrauch von Rechnerressourcen und die Tatsache, dass selbst in der Interaktionsphase die Möglichkeit besteht, dass der Benutzer Äußerungen nicht an den Roboter sondern an Dritte richtet. Für die auditive Aufmerksamkeit des Roboters ist es zwingend erforderlich, den Kopf des Benutzers im Einzugsbereich der Kamera zu halten, um jederzeit das Merkmal „Blick zum Roboter“ erkennen zu können. Die Kamera wird primär auf das Gesicht des Benutzers zentriert, um so Bereitschaft und Aufmerksamkeit zu signalisieren. Da das *Home-Tour*-Szenario jedoch mehr Interaktionsmöglichkeiten als das reine Gespräch bietet, sind abweichende Verhaltensweisen erforderlich, die insbesondere die Ausrichtung der Kamera betreffen.

Im *Home-Tour*-Szenario besteht für den Benutzer die Möglichkeit, in der Interaktion auf Objekte zu verweisen, für die der Roboter Informationen in seiner Wissensbasis aufnehmen soll. Dazu ist es erlaubt, deiktische Gesten einzusetzen. Um diese zu erkennen, müssen sich die Hände des Benutzers im Blickfeld der Kamera befinden. Bei einem für eine Interaktion üblichen Abstand des Benutzers zum Roboter werden die Hände in der Regel jedoch nicht immer erfasst, wenn die Kamera auf den Kopf des Benutzers zentriert ist. Die Kamera muss daher zu einem geeigneten Zeitpunkt nach unten geneigt werden. Dieser ergibt sich in der Regel aus dem Verlauf des Dialogs.

Um an visuelle Information über ein vom Benutzer referenziertes Objekt zu gelangen, ist es notwendig, die Kamera neu auszurichten. Dazu wird die visuelle Aufmerksamkeit des Roboters vom Benutzer gelöst und dann auf das entsprechende Objekt verschoben. Das in dieser Arbeit entwickelte personenbasierte Aufmerksamkeitssystem übergibt dazu die Ansteuerung der Kamera an eine eigens für diese Aufgabe vorgesehene Komponente, die über Fähigkeiten zur Detektion und Analyse von Objekten verfügt. Nachdem ein Objekt analysiert wurde, wechselt die Ansteuerung der Kamera wieder zurück an die personenbasierte Aufmerksamkeitssteuerung, die daraufhin mit der Kamera wieder das Gesicht des Benutzers fokussiert.

Im *Home-Tour*-Szenario besteht für den Benutzer weiterhin die Möglichkeit, durch Vorausgehen den Roboter zu einem anderen Ort zu führen. Da die Aufmerksamkeitssteuerung für die Ansteuerung der Roboterbasis zuständig ist, übernimmt sie auch beim Hinterherfahren die Kontrolle über Richtung und Geschwindigkeit. Da der vorausgehende Benutzer in der Regel in Gehrichtung schaut und nicht spricht, steht dem Personen-*Tracking* in dieser Situation lediglich Positionsinformation von den Beinen durch den Laser-Entfernungsmesser und vom Oberkörper durch die Kamera zur Verfügung. Da es für die farbbasierte Lokalisation des Oberkörpers vorteilhaft ist, wenn sich ein möglichst großer Teil der Kleidung im Bild der Kamera befindet, neigt die Aufmerksamkeitssteuerung die Kamera wieder so weit nach unten, dass sich der Kopf des Benutzers am oberen Bildrand befindet.

## 4.3 Resümee

Das Aufmerksamkeitsverhalten des Roboters ist bei der Bereitschaftsphase und der Interaktionsphase grundlegend verschieden. In der Bereitschaftsphase kann sich die Aufmerksamkeit auf jede Person richten, die sich im Einzugsbereich der Sensoren befindet und damit im *Tracking*-Prozess berücksichtigt werden kann. Die Aufmerksamkeit wird durch externe Reize gesteuert, wobei die Sprechaktivität eine bedeutende Rolle spielt. In dieser Phase lässt sich die Aufmerksamkeit sehr leicht durch Geräusche beeinflussen. Der Roboter verhält sich rein reaktiv. Er verfügt über keinen Mechanismus, um einzelne Personen von seiner Aufmerksamkeit vollständig auszuschließen. Durch sein Verhalten stellt sich der Roboter als neugieriger Zuhörer dar. In der Bereitschaftsphase realisiert die Aufmerksamkeitssteuerung eine reizbasierte, reaktive beziehungsweise bottom-up gesteuerte Aufmerksamkeit.

Im Gegensatz dazu spielen in der Interaktionsphase die Reize wie Sprechaktivität und Blickrichtung keine Rolle für das sichtbare Verhalten des Roboters. Die Aufmerksamkeit richtet sich allein auf den Benutzer, während andere Menschen, die ebenfalls von den Sensoren erfasst werden, vollständig ignoriert werden. Unterschiede im Aufmerksamkeitsfokus der Kamera ergeben sich aus der Dialogsituation, in der die Kamera entweder direkt auf das Gesicht gerichtet ist oder versucht, auch die Hände oder den Oberkörper besser ins Bild zu bekommen. Der Benutzer kann durch geeignete Instruktionen das Verhalten des Roboters steuern. In der Interaktionsphase realisiert die Aufmerksamkeitssteuerung eine selektive, willkürliche beziehungsweise top-down gesteuerte Aufmerksamkeit.

Neben den nach außen hin sichtbaren Verhaltensweisen, die sich durch Bewegung von Kamera und Roboterbasis zeigen, spielt die auditive Aufmerksamkeit eine wichtige Rolle. Sie unterscheidet sich nicht zwischen den beiden Phasen. Die Sprachverarbeitung wird aktiviert, wenn der beobachtete Sprecher zu Beginn seiner Äußerung zum Roboter schaut und endet, wenn eine Sprechpause erfolgt. Zugleich werden die mit den beiden Mikrofonen aufgezeichneten Daten so miteinander kombiniert, dass akustische Signale aus Richtung des Sprechers verstärkt werden. Der Roboter verfügt somit über die Fähigkeit zur selektiven auditiven Aufmerksamkeit.

Für das gesamte Aufmerksamkeitssystem stellt die Fähigkeit des Roboters, Menschen in seiner

Nähe wahrzunehmen, eine grundlegende Voraussetzung dar. Diese Fähigkeit wird durch ein Verfahren zum Verfolgen von Personen realisiert. Dabei werden durch das Aufmerksamkeitssystem besondere Anforderungen an das Verfahren gestellt: Es deckt einen möglichst großen räumlichen Bereich ab, die Anzahl der zu verfolgenden Personen ist variabel, für jedes verfolgte Individuum können Sprechaktivität und Blickrichtung ermittelt werden, und es ist darüber hinaus möglichst robust gegenüber variierenden äußeren Bedingungen. In Rahmen dieser Arbeit wurde eigens ein neues *Tracking*-Konzept entwickelt, um damit ein geeignetes Verfahren zum multimodalen Verfolgen von Personen von einer mobilen Plattform realisieren zu können.

Um die erforderlichen Interaktionsfähigkeiten des Roboters im *Home-Tour*-Szenario zu realisieren, sind neben dem *Tracking*-Verfahren eine Vielzahl weiterer Komponenten erforderlich, wie zum Beispiel Dialogsteuerung, Gestenerkennung oder eine Wissensbasis. Die Komplexität der Anwendung macht es erforderlich, den Datenaustausch und die zeitlichen Abläufe der einzelnen Komponenten zu koordinieren. Zu diesem Zweck ist für den Roboter *BIRON* eine Softwarearchitektur entwickelt worden, in die die Aufmerksamkeitssteuerung und das multimodale Personen-*Tracking* als Bestandteile eingebettet sind.

In den folgenden Kapiteln werden die Details des realisierten Aufmerksamkeitssystems beschrieben, beginnend mit dem Ansatz zum multimodalen Verfolgen von Personen.

# Kapitel 5

# Multimodales *Anchoring*

Das Verfolgen von Personen von einem mobilen Roboter aus ist eine anspruchsvolle Aufgabe. Zum einen handelt es sich um einen hochdynamischen Prozess, da die relativen Positionsveränderungen der Personen nicht nur aus den Bewegungen der Personen, sondern gleichzeitig aus der Eigenbewegung des Roboters resultieren. Zum anderen führt die Mobilität des Roboters zu variierenden äußeren Gegebenheiten, was bei der Verarbeitung von Sensordaten berücksichtigt werden muss. Beim Wechsel des Standorts des Roboters können sich die Beleuchtungsverhältnisse und die akustischen Eigenschaften des jeweiligen Raums drastisch ändern und dadurch die Personenerkennung erschweren. Ein generelles Problem beim Verfolgen von Objekten stellen Verdeckungen dar, während derer das jeweilige Objekt nicht von den Sensoren erfasst werden kann. In dem betrachteten Szenario können Personen sich gegenseitig verdecken oder durch andere Gegenstände teilweise oder ganz verdeckt werden. Eine weitere Schwierigkeit ergibt sich dadurch, dass die Wahrnehmungsbereiche von Sensoren im Allgemeinen eingeschränkt sind und deshalb in der Regel nicht immer alle Personen gleichzeitig erfasst werden können.

Die Größe des Wahrnehmungsbereichs ist abhängig von der Art und Bauweise eines Sensors. Ein Laser-Entfernungsmesser erfasst zum Beispiel nur Objekte innerhalb einer Ebene, während ein Mikrofon Geräuschquellen an beliebiger Position wahrnehmen kann. Von der Art des Sensors hängt es auch ab, wie sich extreme äußere Bedingungen auf die Qualität der Personenerkennung auswirken. Während zum Beispiel eine geringe Beleuchtung die Erkennung von Personen mit einer Kamera erschweren kann, behindern laute Umgebungsgeräusche die Sprecherlokalisation mit Mikrofonen. Die Einschränkungen und Schwächen sind also sensorspezifisch und können durch gleichzeitige Berücksichtigung mehrerer verschiedenartiger Sensoren kompensiert werden. Um das Verfahren zum Verfolgen von Personen von einem mobilen Roboter aus robust zu gestalten, bietet sich ein multimodaler Ansatz an, bei dem die Personen gleichzeitig durch mehrere Sensoren verschiedener Art erfasst werden. Der in dieser Arbeit verwendete Roboter *BIRON* (siehe Abschnitt 3.1) verfügt über eine Farbkamera, mit der Gesichter und Oberkörper erkannt werden, zwei Mikrofone, mit denen Geräusche (Stimmen) lokalisiert werden und einen Laser-Entfernungsmesser, der Beine erfasst.

Die Sensoren des Roboters führen fortlaufend Messungen durch. Das Verfolgen ist der Prozess,

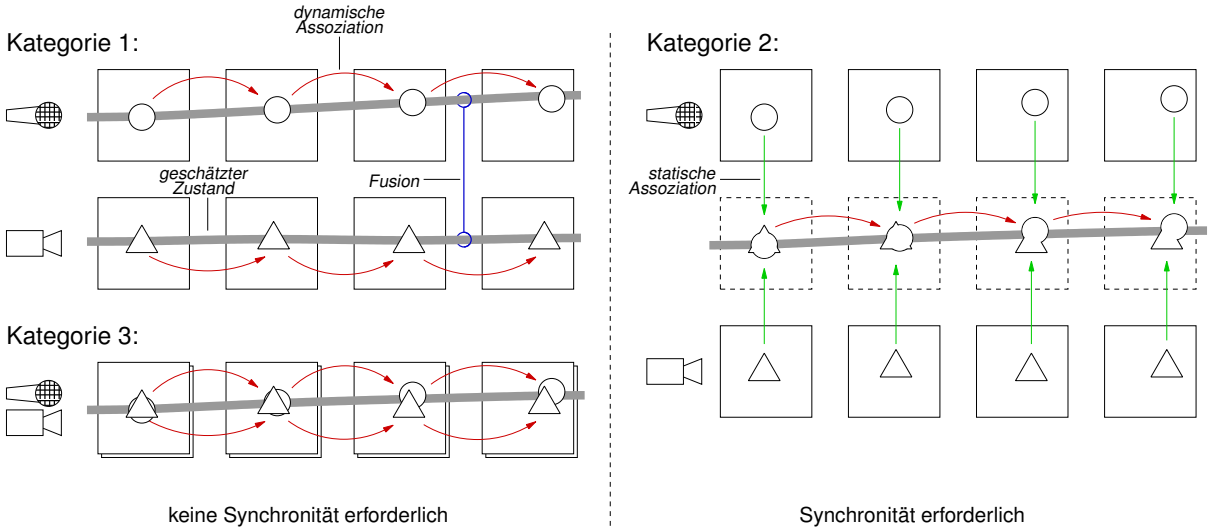



Abbildung 5.1: Multisensor-*Tracking*-Verfahren lassen sich in drei Kategorien aufteilen. Da in Kategorie 2 eine statische Assoziation von Messdaten erfolgt, ist, im Gegensatz zu den Kategorien 1 und 3, eine Synchronisierung der Sensormessungen erforderlich.

bei dem die Messdaten einer Person verarbeitet werden, um den aktuellen Zustand der Person zu schätzen. Der Zustand umfasst im Allgemeinen Bewegungsinformation (Position, Geschwindigkeit, ...), aktuelle Merkmale (spricht, schaut zum Roboter, ...) und Eigenschaften, die sich nicht oder nur sehr langsam ändern (Identität, Farbe der Kleidung, ...). Zum Verfolgen müssen die über den Lauf der Zeit anfallenden Messdaten von einer Person einander zugeordnet werden. Es handelt sich um das Problem der Datenassoziation. In einem Multisensorsystem müssen einerseits die zeitgleichen Messdaten der verschiedenen Sensoren einander zugeordnet werden (statische Assoziation), andererseits müssen zeitlich aufeinander folgende Daten über alle Modalitäten hinweg assoziiert werden (dynamische Assoziation). Auf den assoziierten Daten erfolgt sodann die Schätzung des aktuellen Zustands der Person.

Um die verschiedenen Lösungsstrategien für das Verfolgen mit einem Multisensorsystem zu unterteilen, haben Bar-Shalom und Li in [Bar95], Seite 432ff, drei Kategorien vorgeschlagen (siehe auch Abbildung 5.1):[1]

**Kategorie 1:** Für jeden Sensor gibt es einen eigenen Verfolgungsprozess. Das heißt, die zeitlich aufeinander folgenden Messdaten werden separat, innerhalb der einzelnen Modalitäten einander zugeordnet (dynamische Assoziation). Der aktuelle Zustand der Person wird auf Basis der assoziierten Daten für jeden Sensor einzeln geschätzt, zum Beispiel die Position des Kopfs mit einer Kamera und die Richtung der Stimme mit Mikrofonen. Erst dann

[1] Die hier genannten Kategorien entsprechen den von Bar-Shalom und Li vorgeschlagenen Konfigurationen II bis IV.

werden die Datenströme der verschiedenen Sensoren fusioniert, das heißt einander zugeordnet. Auf den fusionierten Daten erfolgt keine erneute Schätzung des Systemzustands. In diesem Ansatz müssen die Sensoren nicht synchronisiert sein.

**Kategorie 2:** Es erfolgt zunächst eine statische Assoziation der Messdaten der verschiedenen Sensoren, wobei vorausgesetzt wird, dass die Sensoren synchronisiert sind. Es entstehen zu jedem Zeitpunkt „Supermessungen“. Sie dienen als Grundlage für den Verfolgungsprozess. Das heißt, aus den dynamisch assoziierten Supermessungen wird der Zustand der Person geschätzt.

**Kategorie 3:** Die dynamische Assoziation der Messdaten geschieht für jeden Sensor separat. Die Schätzung des Zustands der Person erfolgt sodann unter gleichzeitiger Berücksichtigung der Messdaten aller Sensoren. Während beim Ansatz der Kategorie 1 mehrere Verfolgungsprozesse parallel und voneinander unabhängig ablaufen, hat bei der Vorgehensweise der Kategorie 3 die Schätzung des Gesamtzustands Einfluss auf den *Tracking*-Ablauf beziehungsweise auf die dynamische Assoziation der Messdaten für die einzelnen Sensoren. In diesem Ansatz müssen die Sensoren wiederum nicht synchronisiert sein.

Ein Beispiel für ein *Tracking*-Verfahren der Kategorie 1 ist der Ansatz von Nakadai und Kollegen [Nak01], der auf dem Roboter *SIG* eingesetzt wird, um ein Aufmerksamkeitssystem zu realisieren. Personen werden dabei über Stimme und Gesichter verfolgt. Zeitlich aufeinander folgende Messungen, die aus ähnlichen Richtungen kommen, bilden einen sogenannten Strom. Der Aufbau von Strömen geschieht für die beiden Modalitäten getrennt. Auditive und visuelle Ströme werden dann fusioniert, wenn die Winkeldifferenz innerhalb einer einsekündigen Zeitspanne 10° unterschreitet. Die Trennung erfolgt, wenn die Winkeldifferenz über drei Sekunden mehr als 30° beträgt. Da die Fusion der Daten von verschiedenen Sensoren erst nach dem separaten *Tracking* innerhalb der einzelnen Modalitäten erfolgt, wird bei Verfahren der Kategorie 1 nicht der Vorteil genutzt, dass sich die multimodalen Messungen gegenseitig ergänzen und unterstützen und dadurch die Datenassoziation erleichtert wird.

Im Gegensatz dazu werden bei Verfahren der Kategorie 2 die multimodalen Informationen zunächst kombiniert, bevor die Schätzung des Zustands des beobachteten Objekts erfolgt. Ein Beispiel ist die Methode von Feyrer und Zell [Fey00], die dazu eingesetzt wird, einen mobilen Roboter einer vorausgehenden Person folgen zu lassen. Mit einer Kamera werden Gesichter erkannt, ein Laser-Entfernungsmesser dient zur Lokalisation von Beinen. Die Fusion der Daten wird über eine Maximumsuche in einem Potenzialfeld realisiert, das aus der Überlagerung von zwei zweidimensionalen Gaußfunktionen resultiert. Jede der beiden Funktionen repräsentiert die Position an der eine Person durch den Laser oder die Kamera detektiert wurde. Das *Tracking* auf den fusionierten Daten erfolgt dann über einen Kalman-Filter.

Zur Kategorie 2 sind auch die Ansätze zu zählen, die Partikelfilter zum multimodalen Verfolgen einsetzen. In der Regel wird ein videobasiertes *Tracking* durchgeführt, bei dem die Messungen anderer Sensoren die Gewichtung der Partikel beeinflussen. Sowohl Gatica-Perez und Kollegen [GP04] als auch Vermaak und Kollegen [Ver01] verfolgen Personen über die Kopf-Schulter-

Kontur. Ergebnisse von akustischer Sprecherlokalisation werden in Bildkoordinaten transformiert und dienen zur Gewichtung der Partikel. Wilhelm und Kollegen [Wil02] verfolgen eine ähnliche Strategie. Hier werden jedoch Sonardaten anstelle von akustischen Daten fusioniert. Einen weiteren probabilistischen Ansatz schlagen Beal und Kollegen [Bea03] vor. Sie verwenden ein grafisches Modell, das auditive und visuelle Daten kombiniert.

Charakteristisch für Verfahren der Kategorie 2 ist, dass die Datengewinnung über die verschiedenen Sensoren synchron sein muss. Dies kann zum Nachteil werden, wenn sich die Zeiten, die zur Verarbeitung der Sensordaten benötigt werden, zwischen den Modalitäten stark unterscheiden, da das langsamste System die Geschwindigkeit des Gesamtverfahrens bestimmt.

Das in dieser Arbeit zum Einsatz kommende *Tracking*-Verfahren realisiert ein multimodales Verfolgen von Personen von einem mobilen Roboter aus. Daten von drei verschiedenen Sensoren (Kamera, Mikrofone und Laser-Entfernungsmesser) werden asynchron verarbeitet. Das Verfahren ist der Kategorie 3 zuzuordnen.

Das verwendete *Tracking*-Verfahren baut auf einem Ansatz von Coradeschi und Saffiotti auf, der als *Anchoring* bezeichnet wird [Cor00, Cor01b, Cor03]. *Anchoring* ist ein unimodales *Tracking*-Verfahren, das heißt zur Beobachtung der zu verfolgenden Objekte ist lediglich ein einzelner Sensor vorgesehen. Um das vorgeschlagene Konzept auch bei Verwendung mehrerer Sensoren nutzen zu können, wurde in Kooperation mit Marcus Kleinehagenbrock eine Erweiterung des Verfahrens entwickelt, die entsprechend als multimodales *Anchoring* bezeichnet wird [Fri03b]. Bevor in den Abschnitten 5.2 und 5.3 die für multimodales *Anchoring* erforderlichen Erweiterungen dargestellt werden, beschreibt der folgende Abschnitt zunächst das ursprüngliche *Anchoring*.

## 5.1 *Anchoring*

*Anchoring* ist ein domänenunabhängiges *Tracking*-Konzept, das ein ganzheitliches Rahmenwerk für die verschiedenen Aspekte der Objektverfolgung definiert. Es bietet die Möglichkeit, beim Verfolgen von Objekten Einschränkungen auf bestimmte Objektklassen oder auch ein einzelnes Individuum vorzunehmen. Daneben berücksichtigt es sowohl die Initialisierung der *Tracking*-Prozedur als auch das Problem verfolgte Objekte wiederzufinden, die für eine längere Zeit nicht beobachtet werden konnten.

*Anchoring* ist für autonome Systeme entwickelt worden, die ihr Handeln auf Basis von symbolischem Schließen planen. Objekte, die für das System von Interesse sind und verfolgt werden sollen, werden intern durch Symbole repräsentiert. Die Schnittstelle zur Außenwelt stellt der verwendete Sensor dar, mit dem das System die Objekte erfassen kann. Während ein Symbol über den gesamten *Tracking*-Prozess erhalten bleibt, liefert der Sensor fortlaufend neue Messungen des entsprechenden Objekts. *Anchoring* befasst sich folglich mit dem Problem, Verknüpfungen zwischen Objektreferenzen auf symbolischer Ebene und sensorischer Ebene aufzubauen und über die Zeit aufrecht zu erhalten. Die Verknüpfungen müssen dynamisch sein, da ein Symbol

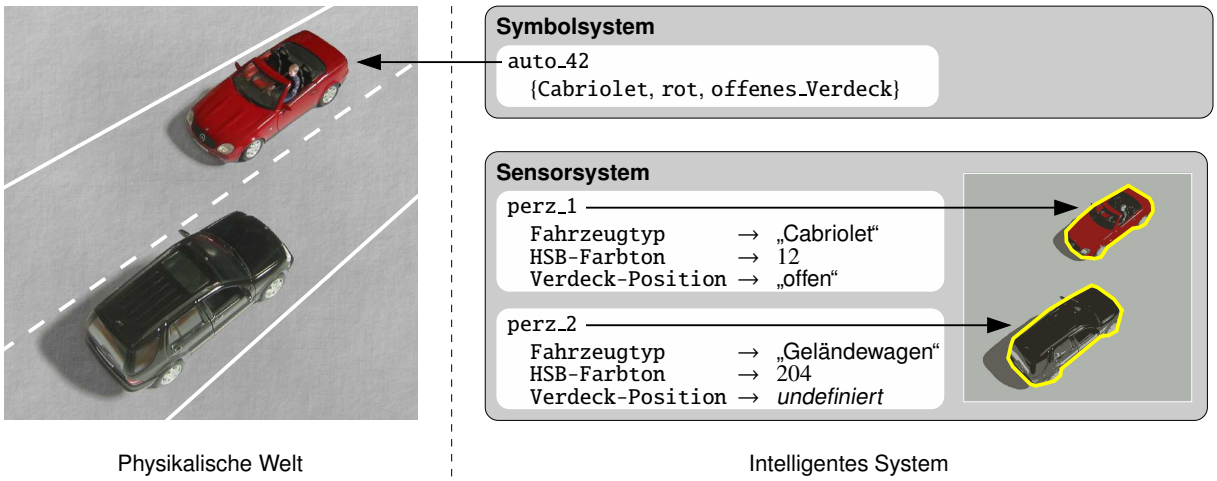



Abbildung 5.2: Das zu verfolgende rote Cabriolet wird im Symbolsystem durch das Symbol `auto_42` bezeichnet. Im Sensorsystem wurden zwei Perzepte generiert, von denen `perz_1` von dem gesuchten Fahrzeug stammt.

jedes Mal an neue Messwerte geknüpft werden muss, wenn eine neue Beobachtung des Objekts vorliegt.

### 5.1.1 Das *Anchoring*-Problem

In diesem Unterabschnitt wird das *Anchoring*-Problem, das den korrekten Aufbau von Verknüpfungen zwischen Objektreferenzen der symbolischen und der sensorischen Ebene beinhaltet, formal beschrieben. Dazu müssen zunächst die grundlegenden Bestandteile von *Anchoring* definiert werden. Die behandelten Definitionen werden anhand eines Beispiels veranschaulicht. Das Beispiel wird an den entsprechenden Stellen immer wieder aufgegriffen.

**Beispiel:** In einem fiktiven Szenario beobachtet ein unbemannter Überwachungshubschrauber eine Straße, auf der verschiedene Fahrzeuge unterwegs sind. Abbildung 5.2 zeigt auf der linken Seite den Blickwinkel des Systems. Es ist das Ziel der Anwendung, das rote Cabriolet zu verfolgen. Um das gesuchte Auto zu bezeichnen, wird im zugehörigen *Anchoring*-Prozess das Symbol `auto_42` verwendet. Da das Symbol keine Objekteigenschaften vorgibt, könnte es für beide der beobachteten Fahrzeuge verwendet werden. Um den *Tracking*-Prozess jedoch auf das Cabriolet zu lenken, wird zusätzlich zu dem Symbol eine Auswahl von so genannten Prädikaten angegeben, die das Auto so beschreibt, dass es sich von anderen unterscheiden lässt. In diesem Beispiel werden dazu die Prädikate `Cabriolet`, `offenes_Verdeck` und `rot` verwendet. Diese Menge der Prädikate stellt die symbolische Beschreibung des gesuchten Fahrzeugs dar. ◇

Formal lassen sich die im Beispiel dargestellten Bestandteile von *Anchoring* wie folgt beschreiben.[2] Auf Ebene der abstrakten Repräsentation der zu verfolgenden Objekte wird im *Anchoring* ein Symbolsystem $\Sigma = (X, P)$ eingeführt. Es setzt sich aus einer Menge von Symbolen $X = \{x_1, x_2, \dots\}$ und einer Menge von Prädikaten $P = \{p_1, p_2, \dots\}$ zusammen. Die Symbole aus $X$ dienen als Bezeichner für physikalische Objekte. Jedes Symbol kann für ein beliebiges Objekt stehen, welches vom Sensorsystem erkannt werden kann. Es sagt nichts über die Eigenschaften eines Objekts aus. Um ein durch ein Symbol referenziertes Objekt zu spezifizieren, werden Prädikate aus $P$ verwendet. Jedes Prädikat gibt eine bestimmte Eigenschaft eines Objekts an. Als symbolische Beschreibung wird jede Teilmenge von Prädikaten $\sigma \in 2^P$ bezeichnet. Sie listet die für die perzeptuelle Erkennung eines Objekts relevanten Prädikate auf.

**Beispiel:** Das System beobachtet das Szenario mit einer Farbkamera. Über geeignete Techniken der Bildverarbeitung sei das System in der Lage, die Fahrzeuge zu detektieren und segmentieren. Die segmentierten Bildbereiche heißen im *Anchoring* Perzepte. In der gegebenen Situation hat das Sensorsystem folglich zwei Perzepte generiert, die mit `perz_1` und `perz_2` bezeichnet werden. Um die detektierten Fahrzeuge voneinander unterscheiden zu können, ist es notwendig, Merkmale zu bestimmen. Dazu werden aus den zugehörigen segmentierten Bildbereichen Werte für bestimmte, vorgegebene Attribute extrahiert. In diesem Beispiel werden die Attribute `Fahrzeugtyp`, `Verdeck-Position` und `HSB-Farbton`[3] verwendet. Die für die Perzepte berechneten Attributwerte sind in der Abbildung 5.2 angegeben. Für Perzept `perz_2` war dabei für das Attribut `Verdeck-Position` keine Angabe möglich, da das entsprechende Fahrzeug über kein mechanisches Verdeck verfügt. ◇

Die zugehörigen Definitionen sind wie folgt. Auf Ebene der sensorischen Erfassung von Objekten wird das Sensorsystem $\Xi = (\Pi, \Phi)$ eingeführt. Es besteht aus einer Menge von Perzepten $\Pi = \{\pi_1, \pi_2, \dots\}$ und einer Menge von Attributen $\Phi = \{\phi_1, \phi_2, \dots\}$. Ein Perzept ist eine strukturierte Ansammlung von Messwerten, die alle von demselben physikalischen Objekt stammen. Ein Perzept ist zum Beispiel das Ergebnis eines Segmentierungsprozesses oder eines Objekterkenners. Die Menge $\Pi$ umfasst alle theoretisch möglichen Perzepte, von denen in jedem Arbeitsschritt des Sensorsystems jedoch immer nur eine sehr kleine Teilmenge aus den Messdaten extrahiert und verwendet wird. Perzepte selbst beschreiben keine Objekteigenschaften, jedoch können aus Perzepten Eigenschaften ermittelt werden. Welche Objekteigenschaften betrachtet werden, wird durch die Menge der Attribute $\Phi$ angegeben. Jedes Attribut $\phi_i$ ist eine messbare Eigenschaft eines Perzepts mit Werten aus einer Menge $D(\phi_i)$. Als perzeptuelle Signatur wird jede partielle Abbildung $\gamma : \Phi \to D(\Phi)$ von Attributen auf Attributwerte bezeichnet. Sie enthält die Werte der gemessenen Attribute eines Perzeptes. Nicht immer können alle Attributwerte bestimmt werden. Auf diesen Attributen bleibt die Abbildung undefiniert. Die Menge der Attribute, auf denen die perzeptuelle Signatur definiert ist, wird mit $\mathrm{feat}(\gamma)$ bezeichnet.

**Beispiel:** Das zu verfolgende Cabriolet wird auf symbolischer Ebene durch ein Symbol repräsentiert. Auf sensorischer Ebene wurden zwei Perzepte generiert, von denen eines vom gesuchten Fahrzeug stammt. Es

[2] Die Notation orientiert sich an der in [Cor01b].

[3] HSB bezeichnet einen Farbraum, in dem Farben durch drei Werte für Farbton (engl. *hue*), Sättigung (engl. *saturation*) und Helligkeit (engl. *brightness*) spezifiziert werden. Der Farbton wird hier in Grad angegeben.

Tabelle 5.1: Beispiel für eine Prädikat-*Grounding*-Relation.

| Prädikat | Attribut | Attributwert |
|---|---|---|
| `rot` | `HSB-Farbton` | $\{0, \ldots, 19, 340, \ldots, 359\}$ |
| `blau` | `HSB-Farbton` | $\{220, \ldots, 259\}$ |
| ⋮ | ⋮ | ⋮ |
| `Cabriolet` | `Fahrzeugtyp` | „Cabriolet“ |
| `Geländewagen` | `Fahrzeugtyp` | „Geländewagen“ |
| ⋮ | ⋮ | ⋮ |
| `offenes_Verdeck` | `Verdeck-Position` | „offen“ |
| `geschlossenes_Verdeck` | `Verdeck-Position` | „geschlossen“ |
| ⋮ | ⋮ | ⋮ |

ist nun die Aufgabe von *Anchoring*, eine Verknüpfung von dem Symbol und einem geeigneten Perzept aufzubauen. Welches der beiden Perzepte in Frage kommt, lässt sich durch Vergleich der symbolischen Beschreibung und der jeweiligen perzeptuellen Signatur entscheiden. Im vorliegenden Beispiel können das Prädikat `Cabriolet` mit dem Attribut `Fahrzeugtyp`, das Prädikat `offenes_Verdeck` mit dem Attribut `Verdeck-Position` und das Prädikat `rot` mit dem Attribut `HSB-Farbton` in Beziehung gesetzt werden. Beim Prädikat `Cabriolet` ist die Entscheidung einfach: Ein geeignetes Perzept muss für das Attribut `Fahrzeugtyp` den Wert „Cabriolet“ aufweisen. Dies trifft auf das Perzept `perz_1`, nicht aber auf Perzept `perz_2` zu. Ähnliches gilt für das Prädikat `offenes_Verdeck`. Beim Prädikat `rot` ist der Zusammenhang nicht so eindeutig, da verschiedene HSB-Farbtöne als rot empfunden werden können. Für die Anwendung wird festgelegt, dass Werte in den Bereichen von $0$ bis $19$ und $340$ bis $359$ als rot akzeptiert werden. Da auch der Wert für das Attribut `HSB-Farbton` bei Perzept `perz_1` mit $12$ in dem geforderten Intervall liegt, sind die perzeptuelle Signatur und die symbolische Beschreibung konsistent. Symbol `auto_42` kann folglich an Perzept `perz_1` geknüpft werden. ◇

Im *Anchoring* wird der Vergleich zwischen symbolischer Beschreibung und perzeptueller Signatur formal über die so genannte Prädikat-*Grounding*-Relation

$$g \subseteq P \times \Phi \times D(\Phi)$$

realisiert. Jedes Tripel $(p, \phi, d)$ aus der Relation $g$ setzt ein Prädikat $p$ mit einem Attribut $\phi$ über einen zulässigen Attributwert $d$ in Beziehung. Im *Anchoring* wird keine Aussage darüber gemacht, wie die Prädikat-*Grounding*-Relation zustande kommt. Sie kann beispielsweise heuristisch festgelegt oder gelernt sein.

**Beispiel:** Einen Ausschnitt der Prädikat-*Grounding*-Relation, die von dem Beispielsystem zum Verfolgen von Autos eingesetzt wird, ist in Tabelle 5.1 angegeben. ◇

Wenn die Konsistenz einer symbolischen Beschreibung $\sigma$ und einer perzeptuellen Signatur $\gamma$ auf Grundlage der Prädikat-*Grounding*-Relation $g$ überprüft werden soll, muss beachtet werden, dass nicht notwendigerweise für jedes Prädikat der symbolischen Beschreibung eine entsprechende

Beobachtung auf der Seite des Perzepts vorliegt, da die perzeptuelle Signatur teilweise undefiniert sein kann. Ob es zu einem Prädikat $p$ ein berechnetes Attribut $\phi$ gibt, das in der Prädikat-*Grounding*-Relation zu dem Prädikat in Beziehung gesetzt wird, bestimmt die Funktion $\mathrm{obs}(\cdot)$:

$$\mathrm{obs}(p,\gamma) \quad \Leftrightarrow \quad \exists \phi \in \mathrm{feat}(\gamma) \;\; \exists d \in D(\phi) \;\; \big((p,\phi,d) \in g\big) \tag{5.1}$$

Alle Prädikate, für die $\mathrm{obs}(\cdot)$ zutrifft, müssen zu dem entsprechenden Attributwert $\gamma(\phi)$ passen, das heißt es muss in der Prädikat-*Grounding*-Relation ein entsprechender Eintrag bestehend aus dem Prädikat $p$, dem Attribut $\phi$ und dem Attributwert $\gamma(\phi)$ vorhanden sein. Diese Bedingung wird durch die Funktion $\mathrm{cons}(\cdot)$ überprüft:

$$\mathrm{cons}(p,\gamma) \quad \Leftrightarrow \quad \exists \phi \in \mathrm{feat}(\gamma) \;\; \big((p,\phi,\gamma(\phi)) \in g\big) \tag{5.2}$$

Schließlich kann die Konsistenz einer symbolischen Beschreibung $\sigma$ und einer perzeptuellen Signatur $\gamma$ durch die Funtkion $\mathrm{match}(\cdot)$ bestimmt werden. Diese überprüft für jedes Prädikat $p$ der symbolischen Beschreibung $\sigma$, dass, wenn es zu dem Prädikat $p$ eine Beobachtung gibt (Funktion 5.1), diese Beobachtung auch konsistent ist (Funktion 5.2):

$$\mathrm{match}(\sigma,\gamma) \quad \Leftrightarrow \quad \forall p \in \sigma \;\; \big(\neg\, \mathrm{obs}(p,\gamma) \;\vee\; \mathrm{cons}(p,\gamma)\big)$$

Die bis hierher beschriebenen Bestandteile von *Anchoring* sind in der jeweiligen Anwendung konstant und von Beginn an gegeben. Da das Verfolgen von Objekten jedoch ein dynamischer Prozess ist, gibt es entsprechend auch zeitabhängige Bestandteile, die im Folgenden beschrieben werden.

Die Dynamik des *Anchoring*-Prozesses wird durch das Sensorsystem bestimmt. Fortlaufend werden dort Perzepte generiert. In einer realen Anwendung braucht jeder Arbeitszyklus des Sensorsystems eine gewisse Zeitspanne, die sich aus der Dauer für den Messvorgang, der Extraktion von Perzepten aus den Messdaten und der Bestimmung der perzeptuellen Signatur für die entsprechenden Perzepte zusammensetzt. Die dafür benötigte Zeit ist nicht notwendigerweise konstant. Sie kann zum Beispiel von der Anzahl der aus den Messwerten extrahierten Perzepte abhängig sein. Das Sensorsystem gibt einen diskreten Zeittakt vor, auf den sich die folgenden Angaben beziehen. Es sei $T = \{t_1, t_2, \dots\}$ die Menge der Zeitpunkte, zu denen das Sensorsystem Perzepte generiert.

Um das *Anchoring*-Konzept möglichst allgemein zu halten, ist es notwendig, dass sich sowohl die Zuordnung von Symbol zur symbolischen Beschreibung als auch die von Perzept zur perzeptuellen Signatur mit der Zeit ändern kann.

**Beispiel:** Es beginnt zu regnen, woraufhin der Fahrer des Cabriolets das Verdeck seines Fahrzeugs schließt. Die symbolische Beschreibung ist fortan aufgrund des Prädikats `offenes_Verdeck` nicht mehr konsistent zu der perzeptuellen Signatur des gesuchten Fahrzeugs. Um das Cabriolet weiterhin verfolgen zu können, muss entweder das Prädikat aus der symbolischen Beschreibung entfernt oder durch ein entsprechendes Prädikat (`geschlossenes_Verdeck`) ersetzt werden. ◇

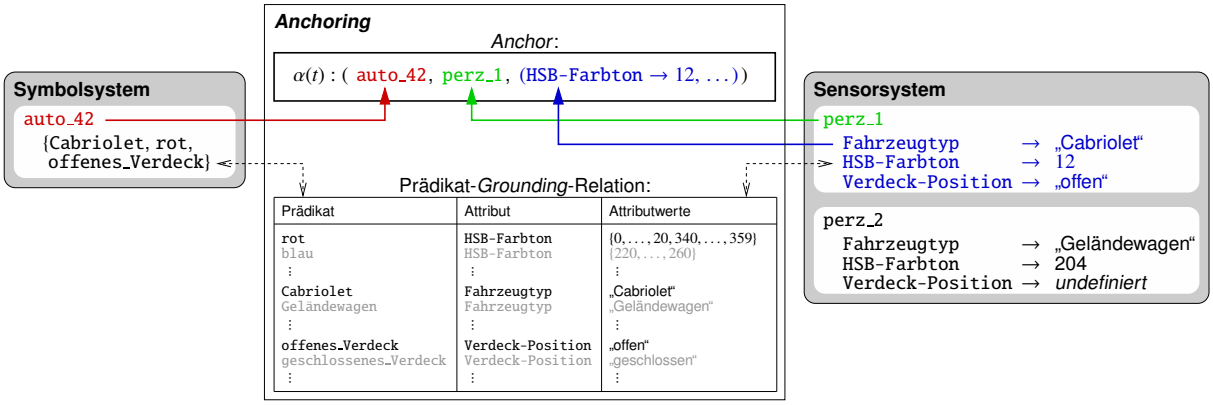



Abbildung 5.3: Beispiel für ein *Anchoring*-System

Um zu berücksichtigen, dass sich die symbolische Beschreibung im Verlauf der Zeit ändern kann, wird der Beschreibungszustand $DS_t : X \rightarrow 2^P$ definiert. Er weist jedem individuellen Symbol $x$ zum Zeitpunkt $t$ seine symbolische Beschreibung zu.

Auch auf der sensorischen Ebene kann sich die Berechnung der Attributwerte über die Zeit ändern.

**Beispiel:** Durch die neue Wetterlage haben sich die Lichtverhältnisse geändert. Bei Regen erscheinen die Farben der Autos im Kamerabild anders als bei Sonne. Dieser Einfluss kann bei der Berechnung des HSB-Farbtons berücksichtigt werden: Bei jedem neu generierten Perzept wird zunächst eine beleuchtungsabhängige Farbkorrektur des segmentierten Bildbereichs durchgeführt, bevor der HSB-Farbton berechnet wird. Die Berücksichtigung eines sich dynamisch verändernden Kontexts führt folglich dazu, dass es keine eindeutige Zuordnung von Perzept zu perzeptueller Signatur gibt. ◇

Analog zum Beschreibungszustand gibt es auf der Ebene des Sensorsystems den perzeptuellen Zustand $PS_t : \Pi \rightarrow \Gamma$, wobei $\Gamma := (\Phi \rightarrow D(\Phi))$ die Menge aller Signaturen ist. Der perzeptuelle Zustand weist jedem Perzept $\pi$ zum Zeitpunkt $t$ seine perzeptuelle Signatur $\gamma$ zu. In den meisten Anwendungen wird der Einfachheit halber die Zuordnung konstant sein, das heißt für dasselbe Perzept werden immer dieselben Attributwerte berechnet.

Mit diesen Definitionen sind die Bestandteile von *Anchoring* auf der Ebene des Symbolsystems und des Sensorsystems hinreichend spezifiziert. Der Rest des Unterabschnitts behandelt den zentralen Aspekt von *Anchoring*: die Verknüpfung zwischen Objektreferenzen auf symbolischer und sensorischer Ebene.

**Beispiel:** Da die zum Symbol `auto_42` gehörende symbolische Beschreibung nach der Prädikat-*Grounding*-Relation konsistent zu der für das Perzept `perz_1` bestimmten Signatur ist, können das Symbol und das Perzept nun miteinander verknüpft werden. Zu diesem Zweck werden sie, zusammen mit der perzeptuellen Signatur, in einer Datenstruktur abgelegt (siehe Abbildung 5.3). Diese stellt das Bindeglied zwischen der symbolischen und der sensorischen Ebene dar. ◇

Verknüpfungen zwischen symbolischer und sensorischer Ebene werden im *Anchoring* als *Anchor* bezeichnet. Ein *Anchor* $\alpha$ ist eine partielle Abbildung der Zeit auf die Menge der Tripel, die aus einem Symbol, einem Perzept und einer perzeptuellen Signatur bestehen:

$$\alpha : T \rightarrow X \times \Pi \times \Gamma \tag{5.3}$$

Zu jedem Zeitpunkt $t$ beinhaltet ein *Anchor* $\alpha(t)$ ein Symbol $x \in X$ aus dem Symbolsystem $\Sigma$, welches ein Objekt bezeichnet, ein Perzept $\pi \in \Pi$, welches durch Beobachtung des Objekts innerhalb des Sensorsystems $\Xi$ generiert wurde, und eine Signatur $\gamma \in \Gamma$, welche die beste Schätzung der Attributwerte für das entsprechende Objekt enthält. Die Bestandteile werden mit $\alpha_t^{sym}$, $\alpha_t^{per}$ und $\alpha_t^{sig}$ bezeichnet. Wenn zu einem Zeitpunkt das Objekt nicht beobachtet werden konnte, dann ist $\alpha_t^{per}$ das Nullperzept $\perp$. Die Signatur $\alpha_t^{sig}$ des *Anchor* enthält in diesem Fall aber weiterhin die bestmögliche Schätzung. Ein *Anchor* $\alpha$ wird zum Zeitpunkt $t$ als *grounded* bezeichnet, genau dann wenn $\alpha_t^{per} \in V_t$, wobei $V_t$ die Menge der Perzepte bezeichnet, die zum Zeitpunkt $t$ vom Sensorsystem generiert wurde.

Mit einem *Anchor* werden für jeden Zeitpunkt die Zuordnungen von einem Symbol zu einem Perzept (oder dem Nullperzept) ausgedrückt. Dabei besteht sinnvollerweise ein eindeutiger Bezug zwischen einem *Anchor* und einem Symbol. Das Symbol $\alpha_t^{sym}$ ist demnach in jedem *Anchor* über die Zeit konstant. Das wesentliche Ziel von *Anchoring* ist, dass das in einem *Anchor* enthaltene Symbol und Perzept dasselbe physikalische Objekt referenzieren. Diese Forderung kann innerhalb des Systems nicht formal spezifiziert werden. Informell gesagt wird ein *Anchor* $\alpha$ genau dann als bezugsmäßig korrekt bezeichnet, wenn zu jedem Zeitpunkt $t$, zu dem der *Anchor* *grounded* ist, das durch $\alpha_t^{sym}$ bezeichnete physikalische Objekt der Ursprung für das Perzept $\alpha_t^{per}$ war.

Schließlich kann das so genannte *Anchoring*-Problem formuliert werden: Es ist das Problem, bezugsmäßig korrekte *Anchor* zu finden. Darüber hinaus sollte für einen *Anchor* $\alpha$ natürlich gelten, dass die Werte in der Signatur $\alpha^{sig}$ ein gute Schätzung der Eigenschaften des entsprechenden physikalischen Objekts darstellen.

### 5.1.2 Basisfunktionen

Um die Definition von *Anchoring* für eine Anwendung nutzbar zu machen, werden Mechanismen benötigt, die einen *Anchor* für einen Zeitpunkt zum ersten Mal aufbauen und seine Definition dann auf folgende Zeitpunkte erweitern. Zu diesem Zweck existieren vier Basisfunktionen FIND, ACQUIRE[4], REACQUIRE und TRACK, die den grundlegenden Anforderungen der meisten Anwendungen gerecht werden. Die beiden erstgenannten Basisfunktionen dienen zum Aufbau von *Anchor*-Funktionen. Dabei werden top-down und bottom-up gerichtete Prozesse unterschieden. Im Top-down-Modus (FIND) ist ein Symbol samt symbolischer Beschreibung vorgegeben und wird an geeignete Perzepte gebunden. Im Bottom-up-Modus (ACQUIRE) bestimmt der Fluss

[4]Die Basisfunktion ACQUIRE wurde zuerst in [Che04] erwähnt.

Gegeben seien ein Zeitpunkt $t$ und ein Symbol $x$.

Geeignete Perzepte bestimmen:

(1) $\Psi \leftarrow \{\pi \in V_t | \, \mathrm{match}(DS_t(x), PS_t(\pi))\}$

(2) $\Psi' \leftarrow \mathrm{Select}(\Psi, t)$

Neue *Anchor* aufbauen:

(3) **for all** $\pi_i \in \Psi'$ **do**

(4) $\alpha_i \leftarrow \mathrm{NewAnchor}(x, \pi_i, PS_t(\pi_i))$

(5) **od**.

Abbildung 5.4: Funktion FIND zum top-down gerichteten Aufbau von *Anchor*-Funktionen.

der Sensordaten, welche *Anchoring*-Prozesse aufgerufen werden. Die zwei anderen Basisfunktionen dienen zur Erweiterung der Definitionsbereiche von *Anchor*-Funktionen auf neue Zeitpunkte. Dabei wird unterschieden, ob sich das durch den *Anchor* referenzierte Objekt unter konstanter Beobachtung befindet (TRACK) oder von dem Objekt für mindestens einen Zeitschritt kein Perzept generiert werden konnte (REACQUIRE). Im Folgenden werden die vier Basisfunktionen eingehend beschrieben.

**FIND:** Die Funktion FIND dient dem top-down gerichteten Aufbau von *Anchor*-Funktionen. In diesem Fall werden für ein vorgegebenes Symbol Perzepte gesucht. Durch die zugehörige symbolische Beschreibung wird dabei die Menge der geeigneten Perzepte eingegrenzt. Es hängt dabei von der Anzahl und Auswahl der Prädikate der symbolischen Beschreibung ab, ob eine Klasse von Objekten („ein Transportwagen“) oder ein bestimmtes Individuum („das rote Cabriolet“) gesucht wird. Im ersteren Fall können potenziell mehrere geeignete Perzepte identifiziert werden. Es kommt dann auf die beabsichtigte Anwendung an, ob für jedes Perzept ein eigener *Anchor* aufgebaut wird oder zunächst aus der Menge der Perzepte eine geeignete Auswahl getroffen wird. In dem Fall, dass ein bestimmtes Individuum gesucht wird, ist der Ablauf dagegen eindeutig: Steht ein geeignetes Perzept zur Verfügung, so wird der entsprechende *Anchor* aufgebaut.

In Abbildung 5.4 ist der Pseudocode von FIND angegeben. Zu einem Zeitpunkt $t$ ist ein Symbol $x$ gegeben, für das ein oder mehrere *Anchor* aufgebaut werden sollen. Die symbolische Beschreibung ist im *Anchoring*-System durch den Beschreibungszustand $DS_t$ zugreifbar. In Zeile (1) wird aus der Menge $V_t$ aller zum Zeitpunkt $t$ im Sensorsystem generierten Perzepte die Teilmenge $\Psi$ der geeigneten Perzepte extrahiert, also diejenigen bei denen die jeweilige Signatur $PS_t(\pi)$ und die symbolische Beschreibung $DS_t(x)$ des Symbols laut Prädikat-*Grounding*-Relation dieselbe Art von Objekt referenzieren. In Zeile (2) wird daraus eine Auswahl von Perzepten $\Psi'$ getroffen. Die entsprechende Unterfunktion $\mathrm{Select}(\cdot)$ ist domänenabhängig und muss für die jeweilige Anwendung eigens

Gegeben seien ein Zeitpunkt $t$ und eine Menge von Perzepten $V' \subseteq V_t$.

Für die gegebenen Perzepte neue *Anchor* aufbauen:

(1) **for all** $\pi_i \in V'$ **do**
(2) $x_i \leftarrow \text{NewSymbol}()$
(3) $\alpha_i \leftarrow \text{NewAnchor}(x_i, \pi_i, PS_t(\pi_i))$
(4) **od**.

Abbildung 5.5: Funktion ACQUIRE zum bottom-up gerichteten Aufbau von *Anchor*-Funktionen.

spezifiziert werden. Die Zeilen (3) bis (5) erzeugen für jedes der ausgewählten Perzepte $\pi_i \in \Psi'$ einen neuen *Anchor* $\alpha_i$. Die *Anchor* sind für den aktuellen Zeitpunkt *grounded*. Sie enthalten jeweils die durch den perzeptuellen Zustand gegebene Signatur $PS_t(\pi_i)$ des zugehörigen Perzepts $\pi_i$.

**ACQUIRE:** Im Bottum-up-Modus wird der Aufbau von *Anchor*-Funktionen durch die im Sensorsystem generierten Perzepte initiiert. Da in diesem Fall keine Einschränkungen durch eine symbolische Beschreibung des Symbolsystems vorliegt, wird für jedes Perzept ein *Anchor* aufgebaut. Es ist nicht immer sinnvoll, für alle zu einem Zeitpunkt $t$ im Sensorsystem neu generierten Perzepte $V_t$ *Anchor* aufzubauen: Wenn zum Beispiel ein Perzept aus der Beobachtung eines Objekts resultiert, für das zu einem früheren Zeitpunkt bereits ein *Anchor* aufgebaut wurde, dann sollte das Perzept dem bestehenden *Anchor* zugeordnet werden. Aus diesem Grund wird in der Funktion ACQUIRE nur eine Teilmenge $V' \subseteq V_t$ der generierten Perzepte betrachtet (siehe Pseudocode in Abbildung 5.5). Für jedes der Perzepte wird in Zeile (2) ein Symbol gewählt, welches bisher kein Objekt referenzierte, um damit in Zeile (3) den entsprechenden *Anchor* aufzubauen.

Bei dem bottom-up gerichteten Aufbau einer *Anchor*-Funktion ist zunächst keine symbolische Beschreibung vorhanden. Um einen *Anchor* bezugsmäßig korrekt fortzuführen, spielt im *Anchoring* die symbolische Beschreibung aber eine wichtige Rolle, da über sie Objekte eindeutig identifiziert und verfolgt werden können. Eine symbolische Beschreibung kann im Bottom-up-Modus aus der Signatur der Perzepte bestimmt werden. Die Menge der zur Signatur konsistenten Prädikate $\omega(\gamma)$ lässt sich mit Hilfe der Funktion $\text{cons}(\cdot)$ (siehe Gleichung 5.2 auf Seite 46) wie folgt bestimmen:

$$\omega(\gamma) \;=\; \{p \in P \mid \text{cons}(p, \gamma)\}$$

**REACQUIRE und TRACK:** Die Funktionen REACQUIRE und TRACK dienen dazu, die Definition einer *Anchor*-Funktion auf einen gegebenen Zeitpunkt $t$ zu erweitern. Das heißt es wird ein Perzept $\pi$ gesucht, dessen perzeptueller Zustand $PS_t(\pi)$ zum entsprechenden Zeitpunkt $t$ zu der symbolischen Beschreibung $DS_t(\pi)$ konsistent ist. In Abbildung 5.6 ist

Gegeben seien ein Zeitpunkt $t$ sowie ein Anchor $\alpha$, der zu einem früheren Zeitpunkt $t'$ *grounded* ist.

Symbol bestimmen:

(1) $x \leftarrow \alpha_{t'}^{sym}$

Signatur schätzen:

(2) $\hat{\gamma} \leftarrow \mathrm{Predict}(\alpha, t', t)$

Geeignetes Perzept ermitteln:

(3) $\Psi \leftarrow \{\pi \in V_t \mid \mathrm{Verify}(DS_t(x), PS_t(\pi), \hat{\gamma})\}$

(4) $\pi \leftarrow \mathrm{Select}(\Psi, t)$

Definition der Anchor-Funktion erweitern:

(5) **if** $\pi \neq \perp$ **then**

(6) $\gamma \leftarrow \mathrm{Update}(\hat{\gamma}, DS_t(x), PS_t(\pi))$

(7) **fi**

(8) $\alpha(t) \leftarrow (x, \pi, \gamma)$.

Abbildung 5.6: Funktion REACQUIRE zum Erweitern der Definition einer *Anchor*-Funktion auf einen neuen Zeitpunkt.

der Pseudocode für die Funktion REACQUIRE angegeben. In Zeile (2) wird zunächst aus der Signatur eines vorhergehenden Zeitpunkts die zu erwartetende Signatur $\hat{\gamma}$ durch die Unterfunktion $\mathrm{Predict}(\cdot)$ bestimmt. Diese Funktion kann beliebig komplex sein und zum Beispiel Weltwissen einbeziehen.

**Beispiel:** Beim Verfolgen des Cabriolets kann Weltwissen wie folgt genutzt werden: Das Cabriolet kann, da es durch einen Tunnel fährt, momentan nicht beobachtet werden, aber mit Wissen über den Streckenverlauf und die Geschwindigkeit kann die Position des Autos geschätzt werden. ◇

In Zeile (3) werden unter Verwendung der Unterfunktion $\mathrm{Verify}(\cdot)$ alle geeigneten Perzepte bestimmt. Diese müssen zum einen bezüglich der symbolischen Beschreibung $DS_t(x)$ konsistent sein, was üblicherweise durch die Funktion $\mathrm{match}(\cdot)$ überprüft wird. Zum anderen muss die aktuelle perzeptuelle Signatur $PS_t(\pi)$ mit der vorhergesagten Signatur $\hat{\gamma}$ konsistent sein.[5] Sollten mehrere Perzepte diese Bedingungen erfüllen, so dient die Unterfunktion $\mathrm{Select}(\cdot)$ in Zeile (4) dazu, eine geeignete Auswahl zu treffen. Konnte ein Perzept ermittelt werden, so wird in Zeile (6) die Signatur, die im *Anchor* gespeichert wird, aus dem vorhergehenden Zustand des *Anchor* und der aktuellen Signatur durch die Unterfunktion $\mathrm{Update}(\cdot)$ neu berechnet. Zuletzt wird in Zeile (8) die Definition des *Anchor* auf den aktuellen Zeitpunkt erweitert. Die Unterfunktionen $\mathrm{Predict}(\cdot)$, $\mathrm{Verify}(\cdot)$, $\mathrm{Select}(\cdot)$ und

[5] Zu diesem Zweck kann die Signatur auch über Attribute verfügen, für die keine entsprechenden Prädikate vorhanden sind. Ein Beispiel sind Attribute, die die räumliche Position des Objektes beschreiben.

$\mathrm{Update}(\cdot)$ sind domänenabhängig und müssen in der jeweiligen Anwendung spezifiziert werden.

Die Funktion TRACK ist eine Spezialisierung von REACQUIRE mit analogem Aufbau. TRACK wird genau dann eingesetzt, wenn sich das durch den gegebenen *Anchor* referenzierte Objekt unter konstanter Beobachtung befindet, also der *Anchor* zum direkt vorhergehenden Zeitpunkt *grounded* war. In diesem Fall können die Unterfunktionen $\mathrm{Predict}(\cdot)$ und $\mathrm{Verify}(\cdot)$ wesentlich einfacher gehalten werden. Wenn zum Beispiel allein durch den Abgleich der vorhergesagten und der tatsächlichen Signatur sichergestellt werden kann, dass das entsprechende Perzept immer noch von demselben Objekt stammt, ist ein Abgleich der symbolischen Beschreibung mit dem perzeptuellen Zustand nicht notwendig.

### 5.1.3 Zusammenfassung und Diskussion

Mit *Anchoring* wird formal das Problem der Verknüpfung von symbolischer Repräsentation und sensorischer Wahrnehmung von Objekten beschrieben. Es ist laut Coradeschi und Saffiotti der erste Ansatz, das Problem des *Symbol-Grounding* für Objekte systematisch zu analysieren [Cor01b]. Mit *Anchoring* ist ein generelles Verfahren zum Verfolgen von Objekten gegeben. Dabei wird neben dem reinen Verfolgen auch die Problematik der Initialisierung des Verfahrens und der Wiederaufnahme des Verfolgens nach längerfristigem Ausbleiben von Sensorinformationen behandelt. Die Definitionen und Funktionen von *Anchoring* geben ein klar strukturiertes Rahmenwerk vor. Vor dem Einsatz für eine konkrete Anwendung müssen jedoch diverse Bestandteile explizit gestaltet werden. Dies betrifft insbesondere die in den Basisfunktionen zahlreich verwendeten Unterfunktionen.

Während die grundlegende Problematik des Aufbaus von Verknüpfungen zwischen symbolischer und sensorischer Ebene durch *Anchoring* hinreichend behandelt wird, bleibt eine Reihe ungeklärter Probleme bestehen. Einige der Schwierigkeiten sind auch von Coradeschi und Saffiotti in [Cor00] und [Cor03] identifiziert worden. Im Folgenden sollen einige Probleme angesprochen werden.

Im *Anchoring* wird implizit die Annahme gemacht, dass die Generierung von Perzepten und die Berechnung von Attributwerten durch das Sensorsystem fehlerfrei sind. In realen Anwendungen ist diese Annahme aufgrund von Messfehlern der Sensoren und einer im Allgemeinen fehlerbehafteten Extraktion von Perzepten aus dem Sensorsignal nicht gewährleistet. Es können dadurch nicht bezugsmäßig korrekte *Anchor* entstehen oder sogar Symbole an Perzepte geknüpft werden, die von gar keinem Objekt stammen (Falsch-Positive). Das Problem der Unsicherheit bei der Generierung von Perzepten kann zum Beispiel durch eine Methode zur Verwaltung mehrerer *Anchor*-Hypothesen aufgefangen werden.

Unsicherheit ist auch ein Aspekt bei der symbolischen Beschreibung. Prädikate, insbesondere solche, die in natürlicher Sprache verwendet werden, wie zum Beispiel `rot`, haben häufig keine präzise Definition in Bezug auf die messbaren Attribute. Die Übereinstimmung von Prädikaten und Attributen sollte daher eher durch Ähnlichkeit als durch Identität bestimmt sein. Ein

Lösungsansatz, der sich diesem Problem widmet, ist durch *Fuzzy Anchoring* [Cor01a], einer Modifikation von *Anchoring*, gegeben. In diesem Ansatz werden Attribute als linguistische Variablen modelliert und die Prädikat-*Grounding*-Relation durch eine Funktion ersetzt, die ein Maß für die Übereinstimmung zwischen Prädikat und Attribut berechnet.

Ein weiteres Problem ergibt sich im *Anchoring* durch eine implizit gegebene Restriktion bei der Berechnung von Attributwerten. Die Signatur wird immer für ein Perzept zu einem Zeitpunkt bestimmt. Damit beinhaltet sie lediglich Information über statische Eigenschaften eines Objekts. Dynamische Eigenschaften, die sich erst über die Veränderung von Objekteigenschaften definieren, beispielsweise die Geschwindigkeit eines Autos, können dadurch nicht berücksichtigt werden. Das Problem im *Anchoring* ist, dass Attributwerte bestimmt werden müssen, bevor ein Perzept einem *Anchor* (und damit implizit Perzepten vorhergehender Zeitpunkte) zugeordnet werden kann, aber eine Zuordnung zeitlich aufeinander folgender Perzepte Voraussetzung dafür ist, dynamische Objekteigenschaft zu berechnen.

### 5.1.4 Notwendige Erweiterungen

Für die in dieser Arbeit entwickelte Aufmerksamkeitssteuerung bildet das Personen-*Tracking* die wesentliche Grundlage. In diesem Abschnitt soll diskutiert werden, inwiefern *Anchoring* sich für das Ziel dieser Arbeit eignet, eine personengerichtete Aufmerksamkeitssteuerung für einen mobilen Roboter aufzubauen.

Für die Aufmerksamkeitssteuerung ist es erforderlich, die Personen in der Nähe des Roboters zu verfolgen. Da alle Personen für die Aufmerksamkeitssteuerung relevant sind, bietet sich ein bottom-up gesteuerter Aufbau von *Anchor*-Funktionen an. Die Anzahl der Personen ist variabel, da nicht feststeht, wann sich Personen dem Roboter nähern oder den Wahrnehmungsbereich der Sensoren wieder verlassen. Es müssen daher zu jedem Zeitpunkt verschiedene Basisfunktionen eingesetzt werden. Zu diesem Zweck sollte es eine zusätzliche Funktion geben, die für die Menge der verwendeten *Anchor* den Aufruf der Basisfunktionen regelt.

Bei gleichzeitiger Verfolgung mehrerer Personen ergibt sich ein Problem bei der Auswahl von Perzepten durch die Unterfunktion $\mathrm{Select}(\cdot)$. Da die Funktion $\mathrm{Select}(\cdot)$ in einem Zeitschritt für mehrere *Anchor* nacheinander aufgerufen wird, kann nicht garantiert werden, dass ein Perzept nicht fälschlicherweise mehreren *Anchor*-Funktionen zugeordnet wird. Die Zuordnung wäre in diesem Fall nicht konsistent. Es ist daher eine parallelisierte Variante notwendig, bei der die Zuordnung von Perzepten für alle *Anchor* gleichzeitig bestimmt wird.

In der geplanten Anwendung müssen die Personen mit Hilfe mehrerer unterschiedlicher Sensoren beobachtet werden, um alle für die Aufmerksamkeitssteuerung relevanten Daten zu erfassen. Das heißt aber, dass ein Symbol, das eine Person bezeichnet, gleichzeitig an mehrere Perzepte aus verschiedenen Sensorsystemen gebunden werden muss. Die Definition von *Anchoring* erlaubt aber nur die Verknüpfung eines Symbols mit genau einem Perzept. Für das multimodale Verfolgen von Personen ist es daher erforderlich, die Definition des *Anchor* zu modifizieren.

Die Aufmerksamkeitssteuerung stellt folgende Anforderungen an das Verfahren:

- Die Anzahl der zu verfolgenden Menschen ist variabel.
- Multimodale Information, die durch verschiedenartige Sensoren geliefert wird, muss integriert werden.

Das multimodale *Anchoring* mehrerer Personen geht über das von Coradeschi und Saffiotti beschriebene *Anchoring* hinaus. Im Rahmen dieser Arbeit wurde das *Tracking*-Konzept daher in beide Richtungen erweitert. Im folgenden Abschnitt werden zunächst neue Funktionen vorgestellt, mit denen *Anchoring* für eine variable Anzahl von Objekten realisiert werden kann. Das Kapitel im Anschluss beschreibt multimodales *Anchoring*, das *Anchoring* für ein Multisensorsystem realisiert.

## 5.2 *Anchoring* von mehreren Objekten

In den Basisfunktionen REACQUIRE und TRACK (siehe Seite 51) wird ein Perzept gesucht, mit dem der Definitionsbereich der betrachteten *Anchor*-Funktion auf den aktuellen Zeitpunkt erweitert werden kann. Konnten mehrere geeignete Perzepte gefunden werden, weil zum Beispiel die symbolische Beschreibung wenig einschränkend ist, dann sorgt die Unterfunktion $\mathrm{Select}(\cdot)$ für die Auswahl eines geeigneten Perzepts. Wenn im *Anchoring*-System mehrere Objekte gleichzeitig verfolgt werden, kann dabei folgendes Problem auftreten: Für jeden *Anchor*, der ein Objekt referenziert, wird eine der beiden Funktionen REACQUIRE und TRACK aufgerufen. Da die Aufrufe unabhängig voneinander erfolgen, kann es prinzipiell passieren, dass ein Perzept zwei verschiedenen *Anchor*-Funktionen zugeordnet wird. Da ein Perzept per Definition nur von einem Objekt stammen kann, läge in diesem Fall ein Widerspruch zu dem Ziel, bezugsmäßig korrekte *Anchor* zu finden, vor.

Eine Vorgehensweise, um dieses Problem zu beheben ist es, weiterhin die Basisfunktionen REACQUIRE und TRACK für jeden *Anchor* in sequenzieller Folge aufzurufen und dabei bei nachfolgenden Aufrufen nur noch die Perzepte aus $V_t$ zu betrachten, die nicht bereits einem bearbeiteten *Anchor* zugeordnet wurden. Diese Vorgehensweise hat die Nachteile, dass den später bearbeiteten *Anchor*-Funktionen tendenziell weniger Perzepte zur Auswahl stehen, und die gesamte Zuordnung global betrachtet nicht optimal ist. Um diese Nachteile zu vermeiden, muss die Auswahl geeigneter Perzepte parallelisiert werden. Zu diesem Zweck wird eine Erweiterung der Basisfunktionen REACQUIRE beziehungsweise TRACK eingeführt, die mit MULTIREACQUIRE bezeichnet wird:

**MULTIREACQUIRE:** Der Funktion MULTIREACQUIRE wird, im Gegensatz zu REACQUIRE und TRACK, nicht nur ein, sondern die Menge aller verwendeten *Anchor* übergeben. Das Ziel ist es, die Definitionsbereiche der *Anchor* durch Zuordnung von Perzepten auf den aktuellen Zeitpunkt $t$ zu erweitern. Die Auswahl geeigneter Perzepte geschieht dabei unter gleichzeitiger Berücksichtigung aller *Anchor*. Die Funktionsweise soll anhand des Pseudocodes in Abbildung 5.7 erläutert werden.

Gegeben seien ein Zeitpunkt $t$ und eine Menge von $n$ *Anchor*-Funktionen $A = \{\alpha_1, \ldots, \alpha_n\}$, die jeweils zu einem früheren Zeitpunkt *grounded* sind.

Für jeden *Anchor* die Menge der geeigneten Perzepte bestimmen:

(1) **for all** $\alpha_i \in A$ **do**
(2) $t' \leftarrow \max\{t \in T | \alpha_{it}^{per} \neq \perp\}$
(3) $x_i \leftarrow \alpha_{it'}^{sym}$
(4) $\hat{\gamma}_i \leftarrow \text{Predict}(\alpha_i, t', t)$
(5) $\Psi_i \leftarrow \{\pi \in V_t \mid \text{Verify}(DS_t(x_i), PS_t(\pi), \hat{\gamma}_i)\}$
(6) **od**

Parallele Zuordnung von Perzepten:

(7) $\{\pi_1 \ldots, \pi_n\} \leftarrow \text{Assign}(\{\Psi_1, \ldots, \Psi_n\}, \{\alpha_1, \ldots, \alpha_n\}, \{\hat{\gamma}_1, \ldots, \hat{\gamma}_n\}, t)$

Definitionen der *Anchor*-Funktionen erweitern:

(8) **for all** $i \in \{1, \ldots, n\}$ **do**
(9) **if** $\pi_i \neq \perp$ **then**
(10) $\gamma_i \leftarrow \text{Update}(\hat{\gamma}_i, DS_t(x_i), PS_t(\pi_i))$
(11) **fi**
(12) $\alpha_i(t) \leftarrow (x_i, \pi_i, \gamma_i)$
(13) **od**.

Abbildung 5.7: Funktion MULTIREACQUIRE zum Erweitern der Definitionen einer Menge von *Anchor*-Funktionen auf einen neuen Zeitpunkt.

In den Zeilen (1) bis (6) wird für jeden *Anchor* jeweils die Menge aller Perzepte bestimmt, die zu der vorhergesagten Signatur und der symbolischen Beschreibung passen. Dies geschieht in Analogie zu den ersten drei Zeilen in der Basisfunktion REACQUIRE. Um auch die TRACK-Variante zu berücksichtigen, müsste es innerhalb der Schleife eine Fallunterscheidung bezüglich *Anchor*-Funktionen geben, die zum unmittelbar vorhergehenden Zeitpunkt *grounded* oder nicht *grounded* sind. Der Übersichtlichkeit halber wurde hier darauf verzichtet.

Die wesentliche Änderung von MULTIREACQUIRE gegenüber REACQUIRE und TRACK ist durch die Unterfunktion $\text{Assign}(\cdot)$ in Zeile (7) gegeben. Diese führt, im Gegensatz zu der Unterfunktion $\text{Select}(\cdot)$, eine Auswahl von Perzepten unter gleichzeitiger Berücksichtigung aller gegebenen *Anchor* durch. Es wird garantiert, dass kein Perzept mehr als einem *Anchor* zugewiesen wird. Sei $\pi_i$ das Perzept, das dem $i$-ten *Anchor* $\alpha_i$ zugeordent wird. Es gilt:

$$\forall \pi_i, \pi_j \in \{\pi_1, \ldots, \pi_n\} \left(\pi_i \neq \pi_j \vee \pi_i = \pi_j = \perp\right)$$

In den Zeilen (8) bis (13) werden anschließend die Definitionsbereiche der *Anchor* auf

den aktuellen Zeitpunkt $t$ erweitert, wobei den *Anchor*-Funktionen die durch $\mathrm{Assign}(\cdot)$ zugewiesenen Perzepte zugeordnet werden. Die Unterfunktion $\mathrm{Assign}(\cdot)$ ist wiederum domänenabhängig.

### 5.2.1 Variable Anzahl von Objekten

Im allgemeinen Fall ist die Anzahl der in einem *Anchoring*-System betrachteten Objekte variabel. Im Top-down-Modus werden in der Regel nicht für alle Symbole gleichzeitig *Anchor* aufgebaut, da zum Beispiel nicht alle referenzierten Objekte zur selben Zeit vom Sensor erfasst werden können. Im Bottom-up-Modus bestimmt der Fluss der Sensordaten, wann neue *Anchor* aufgebaut werden. Dies kann zu verschiedenen Zeitpunkten geschehen. Zusätzlich ist es möglich *Anchor* zu verwerfen, die für die entsprechende Anwendung nicht mehr relevant sind, da sie zum Beispiel Objekte referenzieren, die nicht mehr im Erfassungsbereich des Sensors liegen. Bei einer variablen Anzahl von Objekten müssen zu einem Zeitpunkt potenziell verschiedene Basisfunktionen (FIND, ACQUIRE, MULTIREACQUIRE) eingesetzt werden. Es ist daher notwendig, den Aufruf der verschiedenen Basisfunktionen zu koordinieren. Dies geschieht unter Verwendung der Funktion MULTIANCHORING:

**MULTIANCHORING:** Die hier beschriebene Funktion MULTIANCHORING widmet sich dem bottom-up gesteuerten *Anchoring*-Prozess für eine variable Anzahl von Objekten. Der Funktion wird die Menge $A$ aller aktuell verwendeten *Anchor* übergeben. Ziel ist es, die Definitionsbereiche der *Anchor* auf den aktuellen Zeitpunkt $t$ durch Zuordnung von Perzepten zu erweitern. Gleichzeitig werden nicht mehr benötigte *Anchor* aus der Menge entfernt und neu aufgebaute *Anchor* hinzugefügt. Der zugehörige Pseudocode ist in Abbildung 5.8 angegeben.

In den Zeilen (1) bis (5) werden zunächst die für die Anwendung nicht mehr relevanten *Anchor* aus der Menge $A$ entfernt. Die Entscheidung, ob ein *Anchor* nicht weiter betrachtet wird, realisiert die domänenabhängige Unterfunktion $\mathrm{IsIrrelevant}(\cdot)$. In Zeile (6) werden die Menge der verbliebenen *Anchor* an die Basisfunktion MULTIREACQUIRE übergeben. Bei diesem Aufruf wird nicht notwendigerweise jedes Perzept aus $V_t$ einem *Anchor* zugeordnet. Die Menge der verbleibenden Perzepte $V'$ wird in Zeile (7) bestimmt. Für diese werden in Zeile (8) durch die bottom-up arbeitetende Basisfunktion ACQUIRE neue *Anchor* aufgebaut. Die Top-down-Variante von MULTIANCHORING lässt sich in analoger Weise unter Einsatz der Basisfunktion FIND realisieren.

## 5.3 Multimodales *Anchoring*

Im *Anchoring* werden Symbole und Sensordaten, die dieselben physikalischen Objekte referenzieren, mit Hilfe von *Anchor*-Funktionen miteinander in Beziehung gesetzt. Ein *Anchor* verknüpft dabei laut Definition auf Seite 48 ein Symbol mit genau einem Perzept und einer zugehörigen Signatur. Viele Objekte lassen sich jedoch aufgrund ihrer komplexen Natur nicht durch

Gegeben seien ein Zeitpunkt $t$ und eine Menge von $n$ *Anchor*-Funktionen $A = \{\alpha_1, \ldots, \alpha_n\}$, die jeweils zu einem früheren Zeitpunkt *grounded* sind.

Nicht mehr relevante *Anchor* entfernen:

(1) **for all** $\alpha_i \in A$ **do**
(2) **if** $\text{IsIrrelevant}(\alpha_i, t)$ **then**
(3) $A \leftarrow A \setminus \{\alpha_i\}$
(4) **fi**
(5) **od**

MULTIREACQUIRE (Algorithmus 5.7, Seite 55) aufrufen:

(6) $A \leftarrow \text{MULTIREACQUIRE}(A, t)$

ACQUIRE (Algorithmus 5.5, Seite 50) für nicht zugeordnete Perzepte aufrufen:

(7) $V' \leftarrow \{\pi \in V_t | \forall \alpha \in A(\pi \neq \alpha_{t_j}^{per})\}$
(8) $A \leftarrow A \cup \text{ACQUIRE}(\Pi', t)$.

Abbildung 5.8: Funktion MULTIANCHORING zur Koordination der Aufrufe der Basisfunktionen im bottom-up gesteuerten *Anchoring*-Prozess.

einen Sensor alleine vollständig erfassen beziehungsweise nicht durch ein einzelnes Perzept in ihrer Gesamtheit beschreiben. Dies ist zum Beispiel der Fall bei der Problemstellung in dieser Arbeit: Ein Serviceroboter soll auf Wunsch in Interaktion mit einem Menschen treten. Zu diesem Zweck beobachtet er die Personen in seiner Nähe. Um zu entscheiden, ob ein Mensch ihn anspricht, also ob die Person gleichzeitig spricht und ihn anschaut, verwendet der Roboter Mikrofone und eine Kamera. In dieser Anwendung werden für Menschen Perzepte zweier Modalitäten durch zwei verschiedene Sensorsysteme generiert.

Im allgemeinen Fall werden Objekte durch mehrere, potenziell verschiedenartige Sensoren erfasst. Jedes Sensorsystem ist für die Beobachtung eines oder mehrerer Bestandteile eines Objekts zuständig und generiert entsprechende Perzepte. Es stehen mehrere Perzepte einem Symbol gegenüber, welches das Gesamtobjekt beschreibt. In der ursprünglichen Definition deckt *Anchoring* diesen Fall nicht ab. Daher wurde im Rahmen dieser Arbeit das so genannte multimodale *Anchoring* [Fri03b] entwickelt, das eine entsprechende Erweiterung von *Anchoring* darstellt. In diesem Abschnitt wird zunächst kurz der modulare Aufbau des multimodalen *Anchoring*-Systems vorgestellt und im Anschluss die sich ergebenden Fragestellungen behandelt.

### 5.3.1 Modularer Ansatz

Im multimodalen *Anchoring* werden anstelle eines gesamten Objekts nur einige seiner Bestandteile unter Verwendung verschiedenartiger Sensoren erfasst. Für jeden Bestandteil werden Perzepte generiert, die dem Symbol, welches das Gesamtobjekt bezeichnet, zugeordnet werden müs-

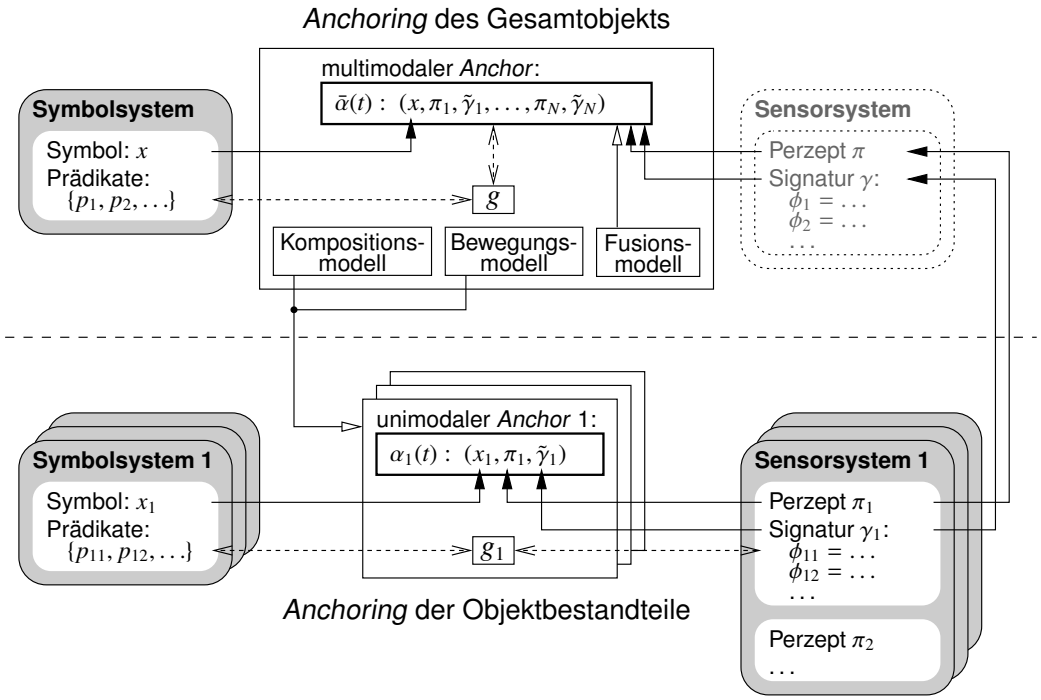



Abbildung 5.9: Schematischer Aufbau vom multimodalen *Anchoring*.

sen. Da ein *Anchor* nach der ursprünglichen Definition 5.3 ein Symbol mit nur einem Perzept verknüpft, muss die Definition für multimodales *Anchoring* erweitert werden. Demnach ist ein multimodaler *Anchor* eine partielle Abbildung

$$\bar{\alpha} : T \rightarrow X \times \big(\Pi_1 \times \Gamma_1\big) \times \ldots \times \big(\Pi_N \times \Gamma_N\big),$$

wobei $N$ die Anzahl der Sensorsysteme ist. Zu jedem Zeitpunkt $t$ beinhaltet ein *Anchor* $\bar{\alpha}(t)$ ein Symbol, $N$ Perzepte und $N$ Signaturen.

Anstatt multimodales *Anchoring* allein durch einen einzigen komplexen *Anchoring*-Prozess zu beschreiben, der ein Symbolsystem und mehrere Sensorsysteme umfassen würde und eine aufwändige Anpassung der Basisfunktionen erforderte, bietet sich vielmehr ein modularer, hierarchischer Ansatz an: Für jedes der beobachteten $N$ Objektbestandteile wird ein normaler, eigenständiger *Anchoring*-Prozess eingesetzt. Die $N$ *Anchoring*-Prozesse laufen parallel und unabhängig voneinander. Jeder *Anchoring*-Prozess wird durch die individuellen Basisfunktionen gesteuert. Sobald dem *Anchor* eines Bestandteils ein Perzept mit entsprechender Signatur zugeordnet werden konnte, werden das Perzept und die Signatur auch im multimodalen *Anchor* verarbeitet.

Im multimodalen *Anchoring* sind somit zwei Ebenen zu unterscheiden (siehe auch Abbildung 5.9): Auf der unteren Ebene befinden sich mehrere unimodale *Anchoring*-Prozesse im traditionellen Sinn, die die Bestandteile eines Objekts beobachten. Auf der übergeordneten Ebene gibt es einen multimodalen *Anchoring*-Prozess, durch den das Gesamtobjekt repräsentiert wird.

Diese Ebene unterscheidet sich von einem herkömmlichen *Anchoring*-System im Wesentlichen durch das Fehlen eines eigenen Sensorsystems. Der *Anchoring*-Vorgang wird daher nicht durch den Aufruf von Basisfunktionen, sondern durch Zuordnung von Perzepten aus den unimodalen *Anchoring*-Prozessen gesteuert.

Für den vorgestellten Ansatz müssen die folgenden Punkte geklärt werden:

- Neben den unimodalen *Anchor*-Funktionen muss auch die multimodale *Anchor*-Funktion bezugsmäßig korrekt sein. Damit dem multimodalen *Anchor* nur Perzepte zugeordnet werden, die alle von demselben Objekt stammen, muss die multimodale Betrachtung das unimodale *Anchoring* in entsprechender Weise einschränken.
- Die unimodalen *Anchoring*-Prozesse arbeiten parallel und asynchron und haben, zumindest wenn sie auf verschiedenen Sensoren basieren, jeweils einen eigenen Zeittakt. Dadurch können dem multimodalen *Anchor* potenziell Perzepte in nicht fortlaufender zeitlicher Reihenfolge zugeordnet werden.
- Die Prädikate auf Ebene des multimodalen *Anchoring*-Prozesses beschreiben Eigenschaften des Gesamtobjekts. Die Attribute, mit denen diese Prädikate in Beziehung gesetzt werden, können dabei nicht wie im herkömmlichen *Anchoring* auf einzelnen Perzepten berechnet werden, sondern es muss die Gesamtheit der zugeordneten Perzepte berücksichtigt werden.

Diese offenen Punkte werden in den folgenden drei Abschnitten diskutiert.

### 5.3.2 Objektmodelle

Um die bezugsmäßige Korrektheit des multimodalen *Anchor* zu gewährleisten, muss in den unimodalen *Anchoring*-Prozessen dafür Sorge getragen werden, dass die ausgewählten Perzepte über alle Modalitäten hinweg von demselben Objekt stammen. Wie kann die Zugehörigkeit von Perzepten zu demselben Objekt bestimmt werden? Jedes Perzept resultiert aus der Beobachtung eines Bestandteils des Objekts. Die Bestandteile stehen durch ihre Zugehörigkeit zum Objekt untereinander in Beziehung. Diese drückt sich in der Regel durch räumliche Relationen aus. Für einen stehenden Menschen zum Beispiel gilt, dass sich der Kopf über den Beinen befindet. Beziehungen können sich auch über Eigenschaften ausdrücken, die unabhängig von der räumlichen Position sind. Zum Beispiel können die Bestandteile eines Objekts ihre Zusammengehörigkeit durch eine einheitliche Farbe ausdrücken. Die Beziehungen zwischen den Bestandteilen eines Objekts sind in der Regel nicht starr. Es gibt viele zulässige Konfigurationen. So befindet sich zum Beispiel der Kopf eines stehenden Menschen nicht immer exakt senkrecht über den Beinen, je nachdem ob die Person aufrecht steht oder ihren Oberkörper neigt. Dennoch sind die Konfigurationen für ein Objekt begrenzt.

Die Konfiguration $\kappa_t$ des beobachteten Gesamtobjekts zum Zeitpunkt $t$ ist im multimodalen *Anchoring* durch die $N$ Signaturen $\bar{\alpha}_{t1}^{sig}, \ldots, \bar{\alpha}_{tN}^{sig}$ des multimodalen *Anchor* $\bar{\alpha}(t)$ bestimmt. Das

Problem der bezugsmäßigen Korrektheit des multimodalen *Anchor* lässt sich über die Konfiguration des Objekts lösen: Es dürfen in den unimodalen *Anchoring*-Prozessen nur Perzepte ausgewählt werden, die bei Zuordnung zum multimodalen *Anchor* eine zulässige Konfiguration ergeben. Um in den unimodalen *Anchoring*-Prozessen die Konfiguration berücksichtigen zu können, werden im multimodalen *Anchoring* drei Modelle eingeführt:

**Kompositionsmodell:** Das Kompositionsmodell spezifiziert den Raum der zulässigen Konfigurationen $K$.

**Bewegungsmodell:** Das Bewegungsmodell beschreibt die zulässigen Veränderungen sowohl der Konfiguration als auch der Position des Gesamtobjekts im Raum. Über das Bewegungsmodell können die Konfiguration und die Position im Raum für einen gegebenen Zeitpunkt vorhergesagt werden.

**Fusionsmodell:** Das Fusionsmodell dient dazu, bei Zuordnung eines Perzepts zum multimodalen *Anchor* aus den neu vorliegenden Messwerten die Konfiguration und die Position des Objekts im Raum zu schätzen.

Die drei Modelle werden in die unimodalen *Anchoring*-Prozesse über die Unterfunktionen $\mathrm{Predict}(\cdot)$, $\mathrm{Verify}(\cdot)$ und $\mathrm{Update}(\cdot)$ innerhalb der Basisfunktionen integriert. In der Funktion $\mathrm{Predict}(\cdot)$ wird zur Bestimmung der zu erwartenden Signatur die durch das Bewegungsmodell vorhergesagte Konfiguration und Position des Gesamtobjekts berücksichtigt. In der Funktion $\mathrm{Verify}(\cdot)$ werden Perzepte bestimmt, die dem *Anchor* zugeordnet werden können. Dabei wird unter anderem geprüft, ob das jeweilige Perzept bei Zuordnung zum multimodalen *Anchor* eine zulässige Konfiguration ergibt (Kompositionsmodell) und ob es sich dabei um eine zulässige Konfigurationsänderung handelt (Bewegungsmodell). In der Funktion $\mathrm{Update}(\cdot)$ wird die perzeptuelle Signatur aktualisiert. Gleichzeitig wird durch Zuordnung des Perzepts zum multimodalen *Anchor* über das Fusionsmodell die aktuelle Konfiguration und Position neu geschätzt.

Die Zugehörigkeit der unimodalen *Anchoring*-Prozesse zum multimodalen *Anchoring*-System hat Auswirkung auf den jeweiligen Aufbau der *Anchor*. Sobald einer der unimodalen *Anchor* durch die Basisfunktion FIND beziehungsweise ACQUIRE aufgebaut wurde, liegt die erste Schätzung der Konfiguration des Gesamtobjekts vor und muss bei dem Aufbau der verbleibenden unimodalen *Anchor* berücksichtigt werden.

### 5.3.3 Zeitkonsistente Zuordnung von Perzepten

Die unimodalen *Anchoring*-Prozesse arbeiten parallel und asynchron. Jeder der *Anchoring*-Prozesse besitzt ein Sensorsystem, das Perzepte generiert. Wie bei der Beschreibung der zeitabhängigen Bestandteile von *Anchoring* bereits erläutert, vergeht zwischen einem Messvorgang des Sensors und der anschließenden Bereitstellung von Perzepten und zugehörigen Signaturen eine entsprechende Zeit. Perzepte beziehen sich also immer auf einen Zeitpunkt in der Vergangenheit. Die benötigten Zeiten können im multimodalen *Anchoring* zwischen verschiedenen unimodalen *Anchoring*-Prozessen um Größenordnungen variieren. Dies hat unmittelbar zur Folge,

dass die dem multimodalen *Anchor* nacheinander zugeordneten Perzepte potenziell nicht der zeitlich korrekten Abfolge entsprechen. Bei der Fusion der Perzeptdaten zur Schätzung der neuen Konfiguration des Objekts durch das Fusionsmodell muss die zeitliche Reihenfolge eingehalten werden. Um dies gewährleisten zu können, wird im multimodalen *Anchoring* der durch ein Sensorsystem vorgegebene Zeittakt neu spezifiziert. Der Zeittakt ist nicht, wie auf Seite 46 definiert, durch die Zeitpunkte gegeben, zu denen das Sensorsystem Perzepte generiert, sondern zu denen die zu den Perzepten jeweils zugehörigen Messungen stattgefunden haben.

Wenn dem multimodalen *Anchor* ein Perzept zugeordnet wird, das im Vergleich zu anderen, bereits verarbeiteten Perzepten zeitlich weiter zurück liegt, wird wie folgt vorgegangen: Es sei angenommen, dass das betreffende Perzept den Zeitstempel $t$ trägt. Die Konfiguration wird zunächst für den Zeitpunkt $t$ unter Berücksichtigung des neuen Perzepts über das Fusionsmodell neu geschätzt. Dadurch werden allerdings die bereits berechneten Konfigurationen für darauf folgende Zeitpunkte ungültig und müssen ebenfalls neu geschätzt werden. Dazu werden die nach dem Zeitpunkt $t$ dem multimodalen *Anchor* zugeordneten Perzepte so behandelt, als würden sie erneut zugeordnet und verarbeitet. Durch dieses Vorgehen wirkt sich auch die in einem nicht in zeitlich korrekter Reihenfolge ausgewählten Perzept enthaltene Information auf die aktuell geschätzte Konfiguration aus. Während Positionsinformationen nur geringen Einfluss haben, wirken sich zeitunabhängige Informationen unmittelbar auf die aktuelle Konfiguration aus.

Das Vorgehen, Konfigurationen bei Vorliegen eines älteren Perzepts für nachfolgende Zeitpunkte neu zu schätzen, birgt die Gefahr, dass sich durch Berücksichtigung des alten Perzepts die Konfiguration derart ändert, dass zeitlich darauf folgende und bereits zugeordnete Perzepte nicht hätten zugeordnet werden können. Diese Gefahr ist bei realen Anwendungen jedoch sehr gering und wird hier nicht explizit behandelt.

### 5.3.4 Attribute des multimodalen *Anchor*

Prädikate auf Ebene des multimodalen *Anchor* beschreiben Eigenschaften des Gesamtobjekts. Da der multimodale *Anchoring*-Prozess über kein eigenes Sensorsystem verfügt, sind entsprechende Attribute, die mit den Prädikaten in Beziehung gesetzt werden können, zunächst nicht vorhanden. Es wird daher festgelegt, dass auch der multimodale *Anchoring*-Prozess eine Menge von Attributen besitzt. Ein Attribut ist in diesem Fall eine messbare Eigenschaften des Gesamtobjekts. Die Attributwerte berechnen sich aus der Konfiguration $\kappa_t$, die durch die im multimodalen *Anchor* enthaltenen perzeptuellen Signaturen zum Zeitpunkt $t$ bestimmt ist. Die Menge der Attribute des Gesamtobjekts wird ebenfalls als perzeptuelle Signatur bezeichnet. Zur Prüfung der Konsistenz von symbolischer Beschreibung und perzeptueller Signatur des Gesamtobjekts dient eine entsprechende Prädikat-*Grounding*-Relation.

Die symbolische Beschreibung des Gesamtobjekts muss auch innerhalb der unimodalen *Anchoring*-Prozesse bei der Auswahl geeigneter Perzepte berücksichtigt werden. Zu diesem Zweck wird in der Unterfunktion $\mathrm{Verify}(\cdot)$ innerhalb der Basisfunktionen überprüft, ob die sich durch Zuordnung des Perzepts ergebende perzeptuelle Signatur des Gesamtobjekts konsistent mit seiner symbolischen Beschreibung ist.

Die Ebene des multimodalen *Anchor* weist im Wesentlichen wieder die Struktur eines herkömmlichen *Anchor* auf. Die ursprüngliche Definition für *grounded* lässt sich jedoch nicht direkt auf den multimodalen *Anchor* übertragen, da der multimodale *Anchor* von den Perzepten mehrerer gleichzeitig arbeitender Sensorsysteme abhängt. Die Definition wird daher wie folgt angepasst: Ein multimodaler *Anchor* heißt *grounded* genau dann, wenn wenigstens einer der zugehörigen unimodalen *Anchor grounded* ist. Da der multimodale *Anchor* die Struktur eines herkömmlichen *Anchor* aufweist, kann über multimodales *Anchoring* eine Hierarchie von *Anchoring*-Prozessen aufgebaut werden. Dabei wird ein multimodaler *Anchor* zum Bestandteil eines übergeordneten multimodalen *Anchoring*-Prozesses.

### 5.3.5 Zusammenfassung

Multimodales *Anchoring* ist eine Erweiterung von *Anchoring*, die es ermöglicht, mehrere Perzepte verschiedener Modalitäten an ein Symbol zu knüpfen. Damit kann das Prinzip von *Anchoring* auch für solche Fälle eingesetzt werden, bei denen sich Objekte nur durch gleichzeitige Verwendung mehrerer, potenziell verschiedenartiger Sensoren vollständig und robust erfassen lassen. Im multimodalen *Anchoring* dienen mehrere herkömmliche *Anchoring*-Prozesse zur Beobachtung der Bestandteile eines Objekts. Die dabei selektierten Perzepte werden auf einer übergeordneten Ebene, dem multimodalen *Anchoring*-Prozess, fusioniert. Über ein Kompositionsmodell, welches die Anordnung der Bestandteile des Objekts beschreibt, und ein Bewegungsmodell, welches die Bewegung der Bestandteile zueinander und die Bewegung des Objekts im Gesamten beschreibt, kann gewährleistet werden, dass die ausgewählten Perzepte immer von demselben Objekt stammen. Multimodales *Anchoring* ist in der Lage, die von verschiedenen Sensoren asynchron erzeugten Daten zu fusionieren. Der modulare Aufbau erlaubt es in einfacher Weise, neue unimodale *Anchoring*-Prozesse hinzuzufügen oder bestehende zu ersetzen.

Multimodales *Anchoring* ist ebenso wie *Anchoring* ein *Tracking*-Konzept. Es gibt eine Reihe von Bestandteilen, die für eine bestimmte Anwendung zunächst spezifiziert werden müssen. Für jeden *Anchoring*-Prozess sind die Prädikate, die Attribute und die Prädikat-*Grounding*-Relationen anzugeben. Die multimodale Erweiterung erfordert die Spezifikation von Kompositions-, Bewegungs- und Fusionsmodell. Darüber hinaus müssen die Unterfunktionen der Basisfunktionen definiert werden. Dabei spielen insbesondere die Funktionen $\mathrm{Predict}(\cdot)$, $\mathrm{Verify}(\cdot)$ und $\mathrm{Update}(\cdot)$ eine wichtige Rolle, da sie im multimodalen *Anchoring* auf die drei Modelle zurückgreifen und somit die Konsistenz zwischen den unimodalen und dem multimodalen *Anchoring*-Prozess realisieren. Die Effizienz und Robustheit des geplanten Verfahrens hängt im Endeffekt von der Auswahl entsprechender Sensorsysteme und Objektmodelle ab. Im folgenden Kapitel wird die entsprechende Umsetzung für das Verfolgen von Personen von einem mobilen Roboter aus beschrieben.

# Kapitel 6

# Multimodales *Anchoring* für Personen

In diesem Kapitel wird beschrieben, wie das Verfolgen von Personen mit Hilfe von multimodalem *Anchoring* realisiert wird. Das Ziel dabei ist es, die Menschen in der Nähe des Roboters zu beobachten, um grundlegende Informationen für die Aufmerksamkeitssteuerung zu erlangen. Da jede Person potenziell die Aufmerksamkeit des Roboters erlangen können soll, müssen immer alle Individuen im Wahrnehmungsbereich der Sensoren verfolgt werden. Es gibt somit beim Aufbau von *Anchor*-Funktionen keine Einschränkung durch eine symbolische Beschreibung. Der Aufbau erfolgt daher bottom-up gerichtet durch die Basisfunktion ACQUIRE (Abbildung 5.5 auf Seite 50).

Die Gestaltung der Sensorsysteme im multimodalen *Anchoring* ist abhängig von den zur Verfügung stehenden Sensoren. Das Verfahren kommt auf dem Roboter *BIRON* zum Einsatz, der mit einer Farbkamera, zwei Mikrofonen, einem Laser-Entfernungsmesser, Sonarsensoren sowie Anstoßsensoren ausgestattet ist (siehe auch Kapitel 3.1). Die beiden letztgenannten Sensortypen werden aus den folgenden Gründen nicht berücksichtigt. Die Anstoßsensoren sind primär eine Sicherheitsvorrichtung, die den Kontakt mit Hindernissen melden, um daraufhin die Ansteuerung des Roboters zu unterbinden. Sie liefern daher keine nützliche Information für das Personen-*Tracking*. Die Sonarsensoren können Hinweise auf den Standort von Menschen geben (siehe zum Beispiel [Wil02]). Die Rate der fälschlich positiven Detektionen ist allerdings hoch, da die Unterscheidung von Menschen zu anderen Objekten in den Sonarmessdaten aufgrund ihrer geringen räumlichen Auflösung schwer zu treffen ist. Da es im *Anchoring*, wie in Abschnitt 5.1.3 erläutert, Probleme bei der Verarbeitung unsicherer Daten gibt, eignen sich die Sonarsensoren nicht für das Verfolgen von Personen mittels multimodalem *Anchoring*.

Zum Einsatz kommen folglich die Kamera, die Mikrofone und der Laser-Entfernungsmesser. Mit der Kamera werden Gesichter und die Kleidung des Oberkörpers detektiert. Über die Mikrofone wird der aktuelle Sprecher lokalisiert. Der Laser-Entfernungsmesser erlaubt es, Beinpaare zu erkennen. Für die Aufmerksamkeitssteuerung sind insbesondere das Gesicht und die daraus ableitbare Blickrichtung einer Person und die Position des aktuellen Sprechers von Bedeutung. Beide Informationen zusammen ermöglichen es zu entscheiden, ob gesprochen wird, wo gesprochen wird und ob zum Roboter gesprochen wird. Auf Basis dieser Daten wird die Aufmerksamkeit

der Roboters gesteuert. Die Erkennung von Beinpaaren und die Lokalisation der Kleidung des Oberkörpers dienen dagegen primär dazu, den Prozess des Verfolgens von Personen robust zu gestalten. Da der Laser-Entfernungsmesser ein aktiver Sensor ist, ist er weitgehend unabhängig von äußeren Bedingungen. Er ist auch dann noch in der Lage, Beine verlässlich zu detektieren, wenn die Kamera (die Mikrofone) aufgrund ungünstiger Beleuchtung (starker Störgeräusche) keine brauchbaren Ergebnisse mehr liefert. Die Lokalisation der Kleidung unterstützt darüber hinaus das Verfolgen von Personen, deren Gesicht und Stimme nicht erkannt werden können, da sie sich zum Beispiel vom Roboter abgewendet haben und nicht sprechen. Sind in dieser Situation auch die Beine durch ein Hindernis verdeckt, ist die Kleidung die letzte verfügbare Informationsquelle im *Tracking*-Prozess.

Die folgenden vier Abschnitte 6.1 bis 6.4 beschreiben zunächst die Mustererkennungsverfahren, die innerhalb der unimodalen *Anchoring*-Prozesse aus den Sensordaten Perzepte generieren. Der daran anschließende Abschnitt 6.5 behandelt die weitere Ausgestaltung des Konzepts vom multimodalen *Anchoring* für das Verfolgen von Personen. Dabei wird insbesondere auf die Berechnung der Perzeptattribute auf Ebene der unimodalen *Anchor*, die Spezifikation der Personenmodelle im multimodalen *Anchor* und die Gestaltung der im *Anchoring*-Prozess aufgerufenen Unterfunktionen eingegangen. Das Kapitel schließt mit einer Zusammenfassung.

## 6.1 Detektion von Beinpaaren

In diesem Abschnitt wird beschrieben, wie aus den Messdaten des Laser-Entfernungsmessers Beinpaare extrahiert werden. Der auf dem Roboter *BIRON* eingesetzte Laser-Sensor erfasst Objekte innerhalb eines 180° Öffnungswinkels auf einer Höhe von 30 cm mit einer Winkelauflösung von 0,5°. Jeder Datensatz besteht folglich aus 361 Messwerten. Ein Beispiel für einen Messdatensatz ist in Abbildung 6.1 gegeben. Innerhalb einer Messung zeichnen sich große ebene Flächen, wie zum Beispiel Wände, durch Folgen linear angeordneter Messpunkte ab. Diese linearen Anordnungen sind häufig dort unterbrochen, wo ein vorgelagertes Objekt die Fläche aus Sicht des Sensors verdeckt. Objektränder zeigen sich in der Regel durch abrupte Änderungen der Abstandswerte benachbarter Messpunkte. Entsprechend deuten aufeinander folgende Messpunkte mit geringen Differenzen der Abstandswerte darauf hin, dass sie von demselben Objekt stammen. Menschen zeichnen sich in den Messdaten in der Regel als einzelne Objekte ab. Es hängt dabei von der Einbauhöhe des Sensors ab, welcher Teil vom Körper erfasst wird. In einigen Ansätzen wird der Oberkörper detektiert (siehe zum Beispiel [Top04] und [Fod02]). Häufiger jedoch ist es das Ziel, in den Messdaten Beine zu erkennen.

In Szenarien mit festem Aufbau können Beine als bewegliche Objekte vor einem statischen Hintergrund extrahiert werden. Zhao und Kollegen [Zha02b] detektieren Beine über Differenzbildung der aktuellen Messung mit einem anfänglich generierten Hintergrundbild. Diese Technik wird auch von Blanco und Kollegen [Bla03] auf einem mobilen Roboter eingesetzt. Voraussetzung ist hier allerdings, dass der Roboter steht. Differenzbildung wird aber auch bei mobilen Roboteranwendungen verwandt. Es wird dabei die Differenz aufeinander folgender Messungen berechnet, wobei die Messpunkte einer der Datensätze so transformiert werden, dass die Bewegung

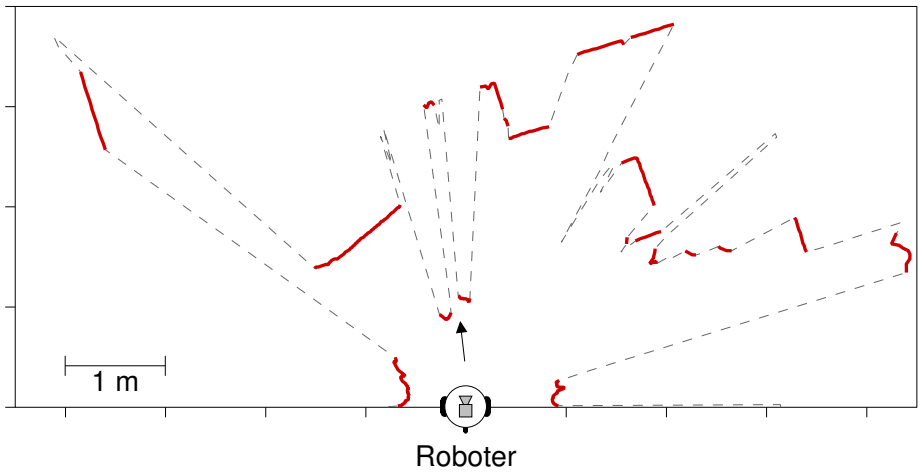


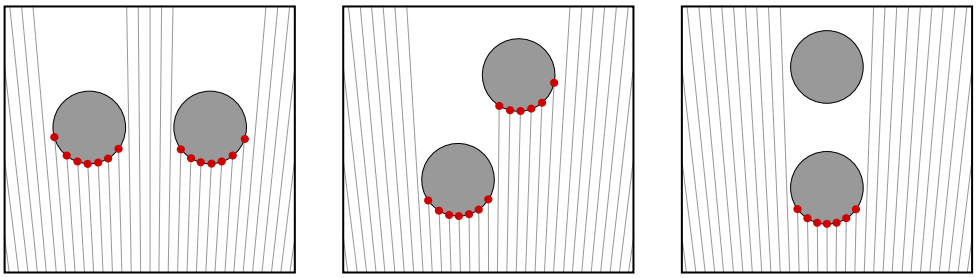

Abbildung 6.1: Beispiel für eine Messung mit dem Lasersensor. Die aufeinander folgenden Messpunkte sind durch eine gestrichelte Linie miteinander verbunden. Der Pfeil zeigt die Stelle, an der sich die beiden Beine eines Menschen abzeichnen.

des Roboters kompensiert wird (siehe zum Beispiel [Jen03, Sch01c]). Ohne Bewegungsinformation kommen die Ansätze von Schraft und Kollegen [Sch01b] und Brooks und Williams [Bro03] aus. Schraft und Kollegen berechnen für lokalisierte Objekte, die sich in der Nähe des Roboters befinden, Merkmale wie Durchmesser, Form und Abstand und verwenden *Fuzzy* Logik um Beinpaare zu erkennen. Brooks und Williams modellieren ein Bein als Halbkreis mit einem Radius innerhalb eines bestimmten Wertebereichs, wobei sich der Halbkreis vom Hintergrund um einen Mindestabstand abhebt. Ein Mensch wird über ein Paar von Beinen mit entsprechendem Abstand detektiert. Wenn ein Bein das andere verdeckt, wird ein Mensch nicht erkannt.

In dieser Arbeit wurde ein heuristisches Verfahren zur Detektion von Beinpaaren entwickelt. Beine werden in einzelnen Messungen ohne Bewegungsinformation detektiert. Im Gegensatz zum Ansatz von Brooks und Williams umfasst das Ergebnis sowohl Beinpaare als auch einzelne Beine.

### 6.1.1 Detektionsprozess

Der Laser-Entfernungsmesser des Roboters *BIRON* erfasst Objekte auf einer Höhe von ungefähr 30 cm. Die Laserstrahlen treffen damit auf Beine in einem Bereich zwischen den Knöcheln und den Knien. Die Beine einer Person ergeben ein charakteristisches Muster innerhalb der Messdaten. Abbildung 6.2 zeigt drei schematische Beispiele, die sich bezüglich der Orientierung der Person zum Roboter unterscheiden. Solange die Person zum Roboter ausgerichtet ist (a), können beide Beine separat erkannt werden. In den anderen Fällen kann das vordere Bein das hintere teilweise (b) oder vollständig (c) verdecken. Wenigstens ein Bein lässt sich jedoch in der Regel als Folge von Messpunkten mit ähnlichem Abstand erkennen.



(a) (b) (c)

Abbildung 6.2: Die Messpunkte für ein Beinpaar ergeben ein charakteristisches Muster. Je nach Orientierung der Person zum Roboter kann ein Bein das andere teilweise oder vollständig verdecken.

Die Detektion von Beinpaaren in den Daten einer Messung geschieht in drei Schritten. Zuerst werden die Messdaten segmentiert, das heißt es werden aufeinander folgende Messwerte gesucht, die aufgrund von ähnlichen Abstandswerten offensichtlich dasselbe Objekt repräsentieren. Die einander zugeordneten Messwerte werden als Segmente bezeichnet. Im folgenden Schritt werden Segmente aussortiert, die aufgrund von gewissen Eigenschaften kein menschliches Bein darstellen können. Die nach der Aussortierung verbleibenden Segmente stellen die Menge der detektierten Beine dar. Im dritten und letzten Schritt werden einzelne Beine in Abhängigkeit ihres jeweiligen Abstands zu Paaren gruppiert.

### Segmentierung

Ein Messvorgang des Laser-Entfernungsmessers resultiert in einer Folge von $N$ Messwerten $(m_1, \ldots, m_N)$. Als Segment $s(i, j)$ wird eine Teilfolge von Messwerten

$$s(i,j) = (m_i, \ldots, m_j) \quad \text{mit} \quad 1 < i < j < N$$

bezeichnet. Im ersten Schritt der Detektion von Beinpaaren werden alle Segmente gesucht, bei denen sich die Abstandswerte zweier benachbarter Messpunkte innerhalb des Segments nur geringfügig voneinander unterscheiden, während die Differenzen an den Segmentgrenzen zu den angrenzenden Messpunkten groß ausfallen. Sei der Schwellwert für die Abstandsdifferenzen $d$, so ergibt sich die Menge der gesuchten Segmente $S$ wie folgt:

$$\begin{aligned} S = \{ s(i,j) \mid\ & \forall k \in \{i+1, \ldots, j\} \big( |m_k - m_{k-1}| \leq d \big) \quad \wedge \\ & |m_i - m_{i-1}| > d \quad \wedge \quad |m_j - m_{j+1}| > d \,\} \end{aligned}$$

Für die Wahl von $d$ und allen folgenden Parametern wurde eine Beispielmenge von 38 Datensätzen des Laser-Entfernungsmessers analysiert, die verschiedene Situationen mit Personen vor dem Roboter erfasst. Für den Schwellwert $d$ wurde ein Wert von 7,5 cm als geeignet ermittelt. Damit werden in der Regel sämtliche Beine im Einzugsbereich des Sensors durch einzelne Segmente repräsentiert. Die Menge aller Segmente umfasst neben den Beinen jedoch auch andere

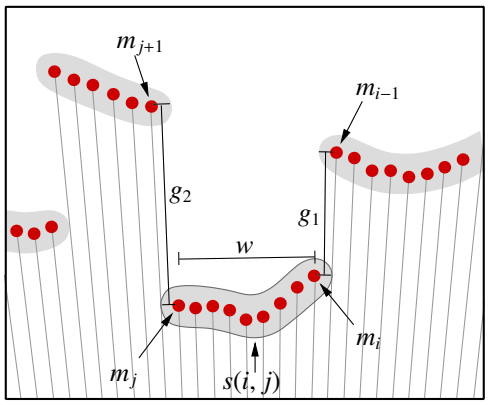



Abbildung 6.3: Aufeinander folgende Messpunkte mit geringen Abstandsdifferenzen werden zu Segmenten zusammengefasst (grau hinterlegt). Die Klassifikation als Bein oder Nicht-Bein erfolgt anhand von bestimmten Merkmalen. Dargestellt sind die Abstandsdifferenzen an den Segmentgrenzen $g_1$ und $g_2$ und die Breite $w$ für das umrandete Segment $s(i, j)$.

Objekte. Diese werden im folgenden Klassifikations-Schritt anhand entsprechender Merkmale aussortiert.

### Klassifikation

Segmente, die von menschlichen Beinen stammen, weisen charakteristische Merkmalwerte auf, die sie von anderen Segmenten unterscheiden lassen. So liegt zum Beispiel der Durchmesser eines Beins innerhalb eines eingeschränkten Wertebereichs, den viele andere Objekte nicht aufweisen. Im Folgenden wird die Menge der verwendeten Ausschlusskriterien spezifiziert.

- Da der Laser-Entfernungsmesser nur eine begrenzte Winkelauflösung hat, werden Objekte bei größerem Abstand durch weniger Messstrahlen erfasst. Bei einer sehr geringen Anzahl von Messpunkten innerhalb eines Segments lassen sich keine aussagekräftigen Merkmale bestimmen. Aus diesem Grund werden zunächst nur solche Segmente berücksichtigt, die aus wenigstens fünf Messpunkten bestehen.

- Da Menschen in der Regel einen gewissen Abstand zu Objekten, wie zum Beispiel Wänden, einhalten, heben sich die Beine bezüglich der Entfernungswerte vom Hintergrund ab. An den beiden Grenzen des Segments $s$[1] ergeben sich entsprechend große Abstandsdifferenzen $g_1(s)$ und $g_2(s)$ (siehe auch Abbildung 6.3):

$$g_1(s) = m_{i-1} - m_i$$
$$g_2(s) = m_{j+1} - m_j$$

[1] Der Übersichtlichkeit halber wird im Folgenden ein Segment anstelle von $s(i, j)$ nur mit $s$ bezeichnet.

Einen Ausnahmefall bildet die Situation, wenn die Person nicht frontal vor dem Roboter positioniert ist, sondern, wie in Abbildung 6.2 (b) dargestellt, leicht gedreht steht. Dann grenzen die Segmente eines Beinpaars direkt aneinander an. Die Abstandsdifferenz an den aneinanderstoßenden Segmentgrenzen nimmt folglich für das weiter hinten befindliche Segment einen negativen Wert an. Dieser kann jedoch absolut betrachtet nicht beliebig groß werden, da er durch die Schrittweite eines Menschen begrenzt ist. Zusammenfassend werden nur solche Segmente weiter betrachtet, deren Abstandsdifferenzen $g_1(s)$ und $g_2(s)$ einen Mindestwert (20 cm) annehmen, wobei einer der Werte auch einen entsprechenden negativen Wert (–5 cm) aufweisen darf.

- Ein charakteristisches Merkmal für Segmente ist die Breite des entsprechenden Objekts in der Szene. Da die Durchmesser von menschlichen Beinen innerhalb eines kleinen Intervalls liegen, können viele Segmente mit größeren Breitewerten aussortiert werden. Die Breite $w(s)$ eines Segments $s$ wird aus dem mittleren Abstand $\tilde{m}(s)$ des Segments und dem Winkel $\alpha(s)$, der durch die Laserstrahlen aufgespannt wird, die den jeweils äußersten Messpunkten $m_i$ und $m_j$ entsprechen, wie folgt bestimmt:

  $$w(s) = 2\,\tilde{m}(s)\,\tan(\alpha(s))$$

  Dabei berechnet sich der mittlere Abstand $\tilde{m}(s)$ eines Segments $s$ durch:

  $$\tilde{m}(s) = \frac{\sum_{k=i}^{j} m_k}{j - i + 1}$$

  Der Winkel $\alpha(s)$ beträgt bei einer Auflösung des Laser-Entfernungsmessers von 0,5°:

  $$\alpha(s) = 0,5°(j - i)$$

  Das beschriebene Breiten-Merkmal $w$ ist auch in der Abbildung 6.3 dargestellt. Als Beine werden nur Segmente mit 5 cm $\leq w(s) \leq$ 25 cm betrachtet.

- Beine besitzen näherungsweise eine zylindrische Form. Die Abstandswerte der einzelnen Messpunkte weichen innerhalb eines zugehörigen Segments nur geringfügig von dem mittleren Abstand $\tilde{m}$ ab. Häufig findet man jedoch Segmente, die nicht von Beinen stammen, bei denen die Abstandswerte stark um den mittleren Abstand variieren. Solche Segmente resultieren zum Beispiel von Wänden oder ebenen Oberflächen, die aus einem flachen Blickwinkel vom Laser erfasst werden. Ein Maß für die Abweichung vom mittleren Abstand $\tilde{m}(s)$ ist die Standardabweichung $d(s)$ der Messwerte:

  $$d(s) = \sqrt{\frac{\sum_{k=i}^{j}(m_k - \tilde{m})^2}{j - i + 1}}$$

  Segmente mit einer Standardabweichung $d(s) >$ 4 cm werden ausgeschlossen.

Die Ausschlusskriterien lassen sich wie folgt zusammenfassen. Die Funktion $\mathrm{Bein}(\cdot)$ klassifiziert dabei ein Segment $s$ als Bein oder Nicht-Bein:

$$\begin{aligned}
\mathrm{Bein}(s) \Leftrightarrow\ & j-i+1 && \geq\ 5 && \wedge \\
& \max\big(g_1(s), g_2(s)\big) && \geq 20\ \text{cm} && \wedge \\
& \min\big(g_1(s), g_2(s)\big) && \geq -5\ \text{cm} && \wedge \\
& w(s) && \geq\ 5\ \text{cm} && \wedge \\
& w(s) && \leq 25\ \text{cm} && \wedge \\
& d(s) && \leq\ 4\ \text{cm}
\end{aligned}$$

**Gruppierung**

Im letzten Schritt werden aus der Menge der verbleibenden Segmente Paare gebildet. Da der Abstand zwischen den Beinen eines Menschen aus anatomischen Gründen begrenzt ist, werden nur solche Segmente gruppiert, deren Abstand in der Szene einen gewissen Schwellwert (50 cm) unterschreitet. Es wird dabei der Abstand zwischen den Schwerpunkten der Messpunkte der jeweiligen Segmente betrachtet. Die Gruppierung von Beinen zu Paaren kann widersprüchlich sein, wenn beispielsweise aus drei detektierten Beinen zwei Paare gebildet werden und damit ein Bein Bestandteil verschiedener Paare ist. Darüber hinaus bleiben nach dem Gruppieren Segmente übrig, die keinem anderen Segment zugeordnet werden können. Dies tritt zum Beispiel dann ein, wenn, wie in Abbildung 6.2 (c) dargestellt, ein Bein einer einzeln stehenden Person das andere verdeckt. Um keine Information zu verwerfen, werden sämtliche, zu Paaren gruppierte Segmente, und die verbleibenden Einzelsegmente als Hypothesen für Beinpaare aufgefasst.

## 6.2 Detektion von Gesichtern

Die Gesichtsdetektion ist Bestandteil der meisten Roboteranwendungen, die auf eine natürliche Interaktion mit Menschen abzielen. Das Szenario mit einem mobilen Roboter stellt jedoch hohe Anforderungen an das entsprechende Erkennungssystem. Da ein Benutzer keinen festen Standort relativ zum Roboter hat, sind Position und Größe der Abbildung des Gesichts im Kamerabild unbekannt. Auch ist die Ansicht nicht notwendigerweise frontal. Des Weiteren sind bei einem mobilen System die Beleuchtungsverhältnisse nicht kontrollierbar. Hinzu kommt das grundsätzliche Problem, dass sich Gesichter verschiedener Individuen aufgrund von Geschlecht, Alter, Mimik, Haarbewuchs oder Accessoires, wie zum Beispiel Brille, unterscheiden. Beispielbilder für Faktoren, die die automatische Gesichtserkennung erschweren, sind in Abbildung 6.4 dargestellt.

Zahlreiche Forscher haben sich mit dem Problem der Gesichtsdetektion auseinandergesetzt. Es ist eine Vielzahl von Techniken vorgeschlagen worden, wie zum Beispiel Neuronale Netze [Row98], deformierbare Schablonen [Yui92], Hautfarbendetektion [Yan96] oder Hauptkomponenten-Analyse, die so genannte *Eigenface*-Methode [Tur91]. Einen Überblick verschaffen die Aufsätze von Yang [Yan02] und Hjelmås [Hje01].

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

Abbildung 6.4: Beispiele für variierende Faktoren, die die automatische Gesichtsdetektion erschweren. (a) Abbildungsgröße, (b) Kopfstellung, (c) Beleuchtungssituation, (d) Individuum

Nach dem heutigen Stand der Forschung zeigen insbesondere die so genannten ansichtsbasierten Methoden (engl. *appearance-based methods*) gute empirische Ergebnisse. Diese Methoden detektieren Gesichter anhand von zuvor aus Trainingsdaten gelernten Modellen. Die Trainingsdaten umfassen in der Regel eine repräsentative Menge von Positivbeispielen, die die Variabilität von Gesichtern abdeckt, und Negativbeispielen, die Bildausschnitte ohne Gesichter enthalten. Um die Modelle zu generieren, werden verschiedene maschinelle Lernalgorithmen verwendet, wie zum Beispiel neuronale Netze [Row98], Hauptkomponentenanalyse [Tur91], Hidden-Markov Modelle [Nef98] oder Boosting [Vio04].

Um Gesichter mit beliebiger Größe und Position im Bild zu detektieren, wird mit ansichtsbasierten Methoden eine ausgiebige Suche über alle Skalierungen und Positionen durchgeführt. Die Anzahl der zu klassifizierenden Bildausschnitte pro Bild kann dabei recht groß ausfallen.

**Beispiel:** Die Größe des zu untersuchenden Bildes sei $256 \times 192$ Bildpunkte. Es werden quadratische Bildausschnitte betrachtet. Die Kantenlänge des kleinsten Ausschnitts betrage 20 Pixel. Darauf aufbauend werden 12 Skalierungsstufen betrachtet, wobei sich die Kantenlänge der Bildausschnitte jeweils um den Faktor 1,25 erhöht. Innerhalb einer Skalierungsstufe werde der Bildausschnitt jeweils um das 1,5-fache der jeweiligen Kantenlänge in horizontaler und vertikaler Richtung verschoben. Bei vollständiger Suche ergeben sich 44635 Klassifikationen. ◇

Um trotz der hohen Anzahl möglicher Bildausschnitte die Bearbeitungszeit für ein einzelnes Bild gering zu halten, muss entweder der Suchraum stark eingeschränkt werden oder jeder einzelne Klassifikationsschritt sehr effizient sein. Im Rahmen dieser Arbeit wurden zwei Ansätze realisiert, die jeweils eine der beiden Strategien verfolgen. Das zuerst entwickelte Verfahren setzt auf der *Eigenface*-Methode von Turk und Pentland [Tur91] auf. Der Suchraum wird zunächst über Hautfarbendetektion eingeschränkt. Innerhalb der segmentierten hautfarbenen Regionen wird die Suche darüber hinaus durch ein Gradientenabstiegsverfahren gesteuert. Der zweite realisierte Ansatz basiert auf dem Klassifikator von Viola und Jones [Vio01]. Dabei ist der Klassifikationsschritt so effizient gestaltet, dass selbst bei vollständiger Suche hohe Verarbeitungsraten erreicht werden.

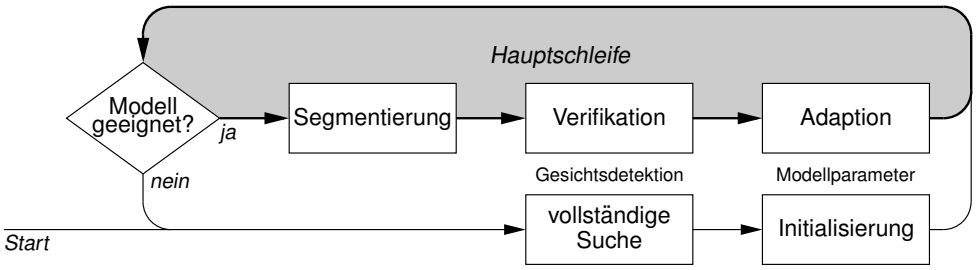



Abbildung 6.5: Ablauf der Gesichtsdetektion

### 6.2.1 Detektion über Hautfarbe und *Eigenfaces*

Das in dieser Arbeit entwickelte Verfahren zur Detektion von Gesichtern [Fri02] ist ein mehrstufiger Prozess. Abbildung 6.5 veranschaulicht den Ablauf. Zunächst wird in einem gegebenen Bild die Suche auf hautfarbene Bereiche beschränkt (Segmentierung). Die sich ergebenden Regionen werden anschließend mit der so genannten *Eigenface*-Methode auf Gesichter hin überprüft (Verifikation). Da die im ersten Schritt durchgeführte Hautfarben-Segmentierung sehr von äußeren Einflüssen, wie zum Beispiel Beleuchtungsverhältnissen oder Hauttyp, abhängt, muss das verwendete Hautfarbenmodell diesen Einflüssen fortwährend angepasst werden. Wenn für eine Region im Verifikationsschritt ein Gesicht detektiert wurde, so liegen im entsprechenden Bildbereich hautfarbene Bildpunkte vor. Diese werden sodann verwendet, um das Hautfarbenmodell zu adaptieren (Adaption).

Das Hautfarbenmodell kann für die aktuell gegebenen Bedingungen ungeeignet sein. Dies ist zum Beispiel dann der Fall, wenn bei Start des Verfahrens das Modell noch nicht initialisiert ist oder wenn die äußeren Einflüsse sich ändern und gleichzeitig keine Adaption stattgefunden hat, weil kein Gesicht gefunden wurde. Wenn das Hautfarbenmodell ungeeignet ist, wird die Hautfarben-Segmentierung, also der erste Schritt des Verfahrens, nicht durchgeführt. Da damit keine Einschränkung des Suchbereichs vorliegt, wird eine vollständige Suche im Gesamtbild durchgeführt. Im Folgenden wird das Verfahren im Detail erklärt.

#### Hautfarbensegmentierung

Die Aufgabe der Hautfarbensegmentierung ist es, hautfarbene Bildpunkte zu finden und aneinander grenzende Bereiche detektierter Bildpunkte zu Regionen zusammenzufassen.

Hautfarbene Bildpunkte können verschiedene Farbwerte besitzen. Diese hängen im Wesentlichen vom Hauttyp und von der Beleuchtungssituation ab. Je nach verwendetem Farbraum stellen die Farbwerte von Haut einen mehr oder weniger gut abgrenzbaren Bereich dar. Als besonders geeignet hat sich der rg-Farbraum herausgestellt [Yan98], der auch als normalisierter RGB-Farbraum bezeichnet wird. Die zwei Dimensionen des rg-Farbraums ergeben sich aus dem Rot-, Grün- und

Blauanteil der RGB-Darstellung wie folgt:

$$r = \frac{R}{R+G+B} \qquad g = \frac{G}{R+G+B}$$

Störring und Kollegen [Stö01a] haben auf Basis eines theoretischen Modells gezeigt, dass die Gesamt-Hautfarbenverteilung im rg-Farbraum einen schalen- oder bananenförmigen Bereich einnimmt, den sie als *skin locus* bezeichnen. Der *skin locus* eignet sich bei der Hautfarbensegmentierung dazu, nichthautfarbene Bildpunkte auszuschließen. Da aber viele andere Objekte, wie zum Beispiel Möbel aus Holz, Farbwerte ergeben, die innerhalb des *skin locus* liegen, bedarf es für die Segmentierung von Hautfarbe einer stärkeren Einschränkung. Für den Fall, dass ein einzelnes Gesicht zu modellieren ist, hat sich eine Gaußverteilung als ausreichend herausgestellt [Raj98].

Die hier verwendete Hautfarbensegmentierung modelliert jedes verfolgte Gesicht durch eine eigene Gaußverteilung. Für jedes Pixel eines gegebenen Bilds wird die Wahrscheinlichkeit für Hautfarbe als das Maximum aller Wahrscheinlichkeiten der verschiedenen Gaußmodelle bestimmt. Das resultierende Hautfarben-Wahrscheinlichkeitsbild wird mit einem empirisch bestimmten Schwellwert von 0,2 binarisiert und mit einem $3\times3$-Medianfilter nachbearbeitet. Es ergibt sich eine Menge von Regionen, die als Hypothesen für Gesichter aufgefasst werden. Nach der Hautfarbensegmentierung folgt der Verifikationsschritt, bei dem die Regionen auf das Vorhandensein von Gesichtern überprüft werden (siehe folgender Abschnitt).

Der abschließende Bearbeitungsschritt für ein Bild ist die Adaption der verwendeten Hautfarbenmodelle. Eine Adaption eines Modells wird nur durchgeführt, wenn für die zugehörige Region ein Gesicht detektiert wurde. Die Zuordnung eines Hautfarbenmodells zu einer Region ist dabei über ein Kalman-*Tracking* realisiert. An der Position einer Gesichtsdetektion werden alle hautfarbenen Bildpunkte innerhalb eines ellipsenförmigen Bereichs genutzt, um das Gaußmodell zu adaptieren. Es werden nur die Farbwerte berücksichtigt, die innerhalb des *skin locus* liegen. Ähnlich zu Soriano und Kollegen [Sor00] wurde der *skin locus* für das verwendete Kameramodell vorab anhand von handsegmentierten Trainingsbildern bestimmt [Fri02]. Die Adaption der Modellparameter geschieht nach folgenden Gleichungen:

$$\begin{aligned} \vec{\mu}_{neu} &= \gamma\vec{\mu}_{lokal} + (1-\gamma)\vec{\mu}_{alt} \\ \Sigma_{neu} &= \gamma\Sigma_{lokal} + (1-\gamma)\Sigma_{alt} \end{aligned}$$

Bei dem realisierten System, welches mit einer Rate von 3 Hz arbeitet, hat sich ein Wert von $\gamma = 0{,}6$ als geeignet herausgestellt.

### *Eigenface*-Methode

Die *Eigenface*-Methode wird im Verifikationsschritt verwendet, um innerhalb einer hautfarbenen Region ein Gesicht zu detektieren. Die Detektion erfolgt auf den Intensitätswerten des Originalbilds, unabhängig von Farbinformationen.

**Detektion** Bei der *Eigenface*-Methode werden Bildausschnitte konstanter Größe ($n \times m$) als Punkte im $nm$-dimensionalen Raum aufgefasst. Gesichtsbilder weisen dabei, trotz variierender Beleuchtungsverhältnisse und Unterschiede zwischen verschiedenen Menschen (siehe Abbildung 6.4 (c) und (d)) Ähnlichkeiten auf, wie zum Beispiel Position und Abstand von Augen, Nase und Mund. Innerhalb des hochdimensionalen Gesamtraums sind die Punkte zu Gesichtsbildern daher nicht willkürlich verteilt, sondern liegen in einem Unterraum, dem so genannten Gesichtsraum. Mit der Hauptkomponentenanalyse können die Hauptachsen dieser Verteilung ermittelt werden. Dies sind die Eigenvektoren der Kovarianzmatrix für die mittelwertfreien Bildvektoren. Die Eigenvektoren zu den größten Eigenwerten spannen den niedrigdimensionalen Gesichtsraum auf. Sie werden auch als Eigengesichter (engl. *eigenface*) bezeichnet [Tur91].

Die Verwendung der Hauptkomponentenanalyse auf Gesichtsbildern geht auf Kirby und Sirovich [Kir90] zurück. Sie nutzen die Technik, um eine effiziente Repräsentation von Gesichtsbildern zu realisieren. Turk und Pentland [Tur91] haben diese Methode dann eingesetzt, um Gesichter zu detektieren. Um für einen Bildausschnitt der Größe $n \times m$ zu entscheiden, ob es sich um ein Gesichtsbild handelt, wird eine Rekonstruktion des Bilds über die Eigengesichter durchgeführt. Da der Gesichtsraum eine deutlich geringere Dimension als der Gesamtraum hat, ergibt sich bei der Rekonstruktion ein Fehler $\varepsilon$. Dieser ist klein, wenn es sich bei dem betrachteten Bildausschnitt um ein Gesicht handelt und andernfalls groß. Die Klassifikation erfolgt daher auf Basis des Rekonstruktionsfehlers und eines geeigneten Schwellwerts.

Voraussetzung für den effektiven Einsatz der *Eigenface*-Methode ist, dass sich Gesichter immer in einer genormten Lage im Bildausschnitt befinden. Dies gilt sowohl für die Trainingsbilder, auf denen die Hauptkomponentenanalyse durchgeführt wird, als auch für die Bildausschnitte, die klassifiziert werden sollen. Zunächst werden die Konsequenzen für den Verifikationsschritt des hier vorgestellten Verfahrens erläutert.

Die Position und die Größe einer von der Hautfarbensegmentierung generierten Region geben einen ersten Anhaltspunkt für die Suche nach einem Gesicht. Aufgrund von Ungenauigkeiten bei der Segmentierung stimmt der Flächenschwerpunkt der Region nicht notwendigerweise mit dem Mittelpunkt des Gesichts überein. Auch die Schätzung der Gesichtsgröße aus der Fläche der Region ist fehlerbehaftet. Der optimale Bildausschnitt, bei dem sich das Gesicht in Normlage befindet, ist unbekannt. Eine mögliche Vorgehensweise, um ein Gesicht zu finden, ist eine erschöpfende Suche über verschiedene Positionen und Skalierungen bezüglich der initialen Position (siehe zum Beispiel [Li00]). Um den Suchaufwand zu reduzieren, wird hier eine andere Strategie vorgeschlagen.

Lemieux und Parizeau [Lem02] haben in Experimenten herausgefunden, dass der Rekonstruktionsfehler bei horizontaler und vertikaler Verschiebung aus der Normlage innerhalb eines begrenzten Bereichs monoton zunimmt. Unter der Annahme, die Schätzung der initialen Position aus den Regionendaten befinde sich innerhalb des entsprechenden Bereichs, kann die Suche über die Veränderung des Rekonstruktionsfehlers bei Verschiebung des Bildausschnitts gesteuert werden. Die Suche wird über einen Gradientenabstieg realisiert, bei dem die nächste zu klassifizierende Position sich aus der Richtung ergibt, in die der Rekonstruktionsfehler abnimmt. Die

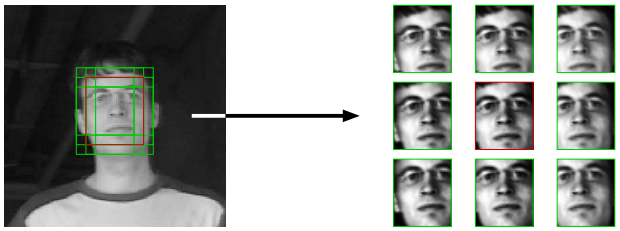


Abbildung 6.6: Generierung von neun Varianten durch horizontale und vertikale Translation des manuell selektierten Bildausschnitts.

Suche endet, wenn der minimale Rekonstruktionsfehler gefunden wurde.[2] Liegt dieser unter dem Schwellwert, so wurde ein Gesicht detektiert. War die Suche erfolglos, wird sie mit verschiedenen Skalierungsstufen ausgehend von der initialen Position wiederholt.

**Training** Eine Voraussetzung für den effektiven Einsatz der *Eigenface*-Methode ist die Auswahl geeigneter Gesichtsbilder für die Hauptkomponentenanalyse. Da Eigengesichter nur die Variationen im Aussehen der Gesichter von verschiedenen Menschen repräsentieren sollen, müssen andere Einflüsse möglichst vollständig reduziert werden. Unterschiede im Bildhintergrund werden ausgeschlossen, indem die Bildausschnitte nur den zentralen Gesichtsbereich mit Augen, Nase und Mund enthalten. Wechselnden Beleuchtungsverhältnissen wird durch eine Vorverarbeitung über Histogramm-Ausgleich entgegengewirkt. Um Variationen bezüglich Translation und Skalierung zu reduzieren, die bei der manuellen Selektion der Bildausschnitte entstehen, wurde ein iteratives Optimierungsverfahren entwickelt, das im Folgenden beschrieben wird.

Für jedes manuell ausgeschnittene Gesichts-Beispiel $f_i \;\; (i = 1, \ldots, N)$ wird eine Menge $V_i$ von $K$ Varianten erzeugt, wobei das manuell ausgeschnittene Bild $f_i$ selbst als Variante betrachtet wird (siehe auch Abbildung 6.6):

$$V_i = \{v_{i1}, \ldots, v_{iK}\} \quad \text{mit} \quad v_{i1} := f_i$$

Jede Variante $v_{ik} \;\; (k > 1)$ wird mit einer geringen Verschiebung und Skalierung bezüglich der initialen Auswahl $v_{i1}$ ausgeschnitten. Jede Auswahl $S$, die aus je einer Variante von jedem Gesicht besteht, dient als Trainingsmenge für die Hauptkomponentenanalyse:

$$S = \{s_1, \ldots, s_N\} \quad \text{mit} \quad s_i \in V_i$$

Um automatisch eine gute Auswahl zu extrahieren, wird der Algorithmus $\mathrm{SelectVariants}$ (Abbildung 6.7) eingeführt, der iterativ die Menge der Trainings-Bilder optimiert. Der Algorithmus startet zum Zeitpunkt $t = 0$ mit einer initialen Auswahl $S_0$, die zum Beispiel aus der Menge der manuell ausgeschnittenen Gesichts-Beispiele besteht. Um auf Basis einer Auswahl $S_t$ eine neue Auswahl $S_{t+1}$ zu bestimmen, wird die Funktion $\mathrm{NewSelection}(\cdot)$ eingeführt. Diese tauscht die

[2] Die Suche ist endlich, da sie auf dem diskreten Raster des Bilds erfolgt.

Gegeben sei eine Menge von $N$ Trainingsbildern mit zugehörigen Varianten $\{V_1, \ldots, V_N\}$. Jede Variantenmenge besteht aus $K$ Elementen: $V_i = \{v_{i1}, \ldots, V_{iK}\}$

Initialisierung:

(1) $S_0 \leftarrow \{s_{01}, \ldots, s_{0N}\}$ mit $s_{0i} = v_{i1}$

(2) $t \leftarrow 0$

Iterative Optimierung:

(3) **repeat**

(4) $r \leftarrow \mathrm{rand}(N)$

(5) $S_{t+1} \leftarrow \mathrm{NewSelection}(S_t, r)$

(6) **until** $S_{t+1} = S_t$

Abbildung 6.7: Algorithmus SelectVariants zur Optimierung der Trainingsstichprobe für die Hauptkomponentenanalyse.

$r$-te Variante $s_{tr}$ aus, während alle anderen ausgewählten Varianten $s_{ti}$ $(i \neq r)$ unverändert bleiben. Für den Austausch werden alle $K$ Varianten $v_{rj} \in V_r$, die zu dem entsprechenden Gesicht $f_r$ gehören, über den Gesichtsraum $F(S_t \setminus \{s_{tr}\})$ rekonstruiert, der aus den restlichen ausgewählten Varianten bestimmt wird. Die Variante mit dem kleinsten Rekonstruktionsfehler $\varepsilon(\cdot)$ wird zur neuen ausgewählten Variante $s_{t+1\,r}$:

$$\mathrm{NewSelection}(S_t, r) = S_{t+1} = \{s_{t+1\,1}, \ldots, s_{t+1\,N}\} \text{ mit}$$

$$s_{t+1\,i} = \begin{cases} \mathrm{argmin}_{v_{ij} \in V_i} \varepsilon\left(F(S_t \setminus \{s_{ti}\}),\, v_{ij}\right) & \text{, wenn } (i = r) \\ s_{ti} & \text{, sonst} \end{cases}$$

Der Algorithmus ersetzt iterativ zufällig ausgewählte Varianten mit Hilfe der Funktion NewSelection bis ein stabiler Zustand erreicht ist. Die mit Hilfe des Algorithmus erreichte Optimierung lässt sich qualitativ an dem Beispiel in Abbildung 6.8 beurteilen. Das linke Bild zeigt das Mittelbild, welches sich aus den manuell selektierten Bildausschnitten ergibt. Das rechte Bild zeigt das Mittelbild der optimierten Selektion. Das Mittelgesicht tritt klarer hervor, was auf eine bessere Ausrichtung der Bildausschnitte hindeutet.

### Zusammenfassung

Der in dieser Arbeit realisierte Gesichtsdetektor kombiniert adaptive Hautfarbensegmentierung mit Gesichtsdetektion basierend auf der *Eigenface*-Methode. Der Segmentierungsschritt schränkt den Suchraum ein, sodass nur Bildausschnitte, die sich bei hautfarbenen Regionen befinden, mit der *Eigenface*-Methode verifiziert werden müssen. Um variierende Beleuchtungsverhältnisse zu berücksichtigen, wird das Hautfarbenmodell kontinuierlich mit Bildpunkten, die bei der Gesichtsdetektion extrahiert werden, aktualisiert. Dieser zirkuläre Prozess muss initialisiert

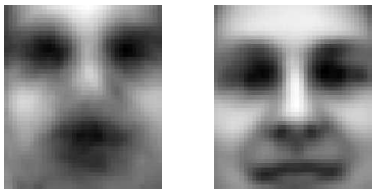


Abbildung 6.8: Mittleres Gesichtsbild der manuell selektierten Bildausschnitte (links) und der optimierten Selektion (rechts). Konturen (Augenbrauen, Nasenlöcher, Oberlippe) zeichnen sich deutlicher ab.

werden, da anfänglich kein geeignetes Hautfarbenmodell zur Verfügung steht. Dazu wird die Gesichtsdetektion mit der *Eigenface*-Methode auf dem gesamten Bild durchgeführt. Der vorgeschlagene Ansatz hat zwei wesentliche Nachteile: Er ist sehr anfällig gegenüber falsch-positiven Detektionen, da dann die Gefahr besteht, dass sich das Hautfarbenmodell an die Farbe des Hintergrunds adaptiert. Daneben ist die Initialisierung sehr zeitaufwändig. Es wurde daher ein neuer Ansatz von Viola und Jones verwendet, der im folgenden Abschnitt erläutert wird.

### 6.2.2 Viola-Jones-Detektor

Viola und Jones haben einen neuen Ansatz für einen Objektdetektor vorgeschlagen [Vio01], der in letzter Zeit immer mehr Verbreitung findet. Er setzt auf relativ einfachen und effizient zu berechnenden Merkmalen auf und verwendet eine hierarchisch aufgebaute Detektorarchitektur. Die Methode ist damit in der Lage, Bilder sehr schnell zu verarbeiten, und liefert insbesondere bei der Gesichtsdetektion gute Erkennungsraten. Es ist daher, im Gegensatz zu der im vorangegangenen Abschnitt vorgestellten Methode, weder eine zeitaufwändige Initialisierung noch eine Einschränkung der Suche durch ein Hautfarbenmodell notwendig.

#### Merkmale

Für ein gegebenes Gesamtbild werden mit dem Viola-Jones-Detektor Bildausschnitte über alle Skalierungen und Translationen untersucht. Die Klassifikation eines Bildausschnitts basiert dabei auf einfachen Merkmalen. Jedes Merkmal setzt sich aus Helligkeitsdifferenzen rechteckiger Bereiche innerhalb des betrachteten Bildausschnitts zusammen. Für die Berechnung eines Merkmalswerts werden die Summen der Intensitätswerte von Bildpunkten innerhalb der rechteckigen Bereiche bestimmt und gewichtet aufsummiert. Viola und Jones verwenden in [Vio01] Merkmale, die ähnlich zu *Haar-Wavelets* sind. Sie bestehen aus zwei bis vier aneinander grenzenden Rechtecken (siehe Abbildung 6.9 (a) bis (c)). Der Wert eines Zwei-Rechteck-Merkmals ist zum Beispiel die Differenz der Pixelsummen der zwei rechteckigen Bereiche.

Die Berechnung von Merkmalswerten lässt sich äußerst effizient und in konstanter Zeit mit Hilfe eines so genannten Integralbilds realisieren. Dieses wird einmal vorab für das zu verarbeitende

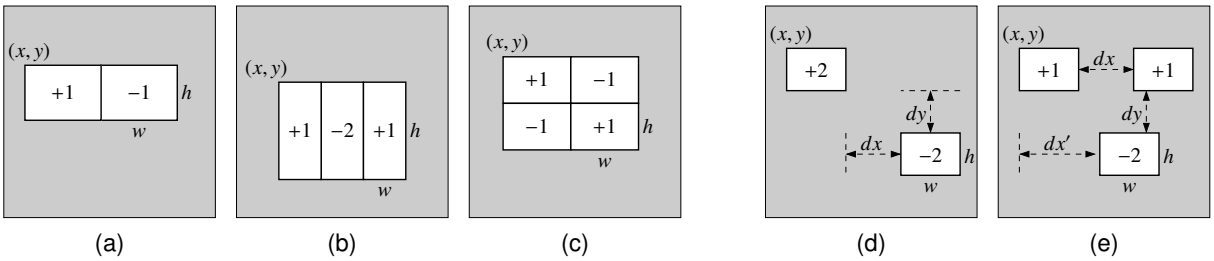



Abbildung 6.9: Merkmale für den Viola-Jones-Detektor. Die Zahlen geben die Gewichtung der Pixelsumme der jeweiligen Bereiche an. Merkmale (a) bis (c) sind in [Vio01] angegeben, Merkmale (d) und (e) werden in dieser Arbeit verwendet (nach [Zha02a]).

Gesamtbild berechnet. Jedes Pixel $i_{xy}$ des Integralbilds gibt die Summe der Pixel $g_{ij}$ des Gesamtbilds innerhalb des vom Ursprung $(0, 0)$ und dem Punkt $(x, y)$ aufgespannten Rechtecks an:

$$i_{xy} = \sum_{i=0}^{x} \sum_{j=0}^{y} g_{ij}$$

Die Pixelsumme $S$ innerhalb eines beliebigen Rechtecks mit den Eckpunkten $a$, $b$, $c$ und $d$ (siehe Abbildung 6.10) lässt sich über die Pixelsummen $A$, $B$, $C$ und $D$ der vier Rechtecke im Gesamtbild bestimmen, die durch den Ursprung und die jeweiligen Eckpunkte definiert sind:

$$S = D - B - C + A$$

Die Berechnung der Pixelsumme eines beliebigen Rechtecks erfordert somit lediglich vier Zugriffe auf das Integralbild, eine Addition und zwei Subtraktionen.

Bei der Berechnung der Merkmalswerte wird zusätzlich eine Varianznormierung der zu klassifizierenden Bildausschnitte durchgeführt. Die Varianz eines Bildausschnitts kann unter Verwendung eines zweiten Integralbilds, welches auf dem Gesamtbild mit quadrierten Pixelwerten erstellt wird, effizient berechnet werden. Sei $\sigma^2 = \frac{1}{N} \sum g^2 - \mu^2$, wobei $\sigma$ die Standardabweichung, $\mu$ der Mittelwert und $g$ die Pixel des Bildausschnitts sind. Der Mittelwert $m$ kann über das erste Integralbild ermittelt werden, während die Summe der Quadrate der Pixel im Bildausschnitt durch das zweite Integralbild bestimmt werden kann.

Die Merkmale nach Viola und Jones können bezüglich des linken oberen Eckpunkts $(x, y)$ sowie der Breite $w$ und Höhe $h$ der einzelnen Rechtecke variiert werden. Neben diesen Merkmalstypen haben andere Autoren weitere Arten vorgeschlagen (siehe zum Beispiel [Lie02b]). In dieser Arbeit finden die von Zhang und Kollegen [Zha02a] eingeführten Zwei- und Drei-Rechteck-Merkmale Verwendung (siehe Abbildung 6.9 (d) und (e)). Die Rechtecke grenzen dabei nicht notwendigerweise aneinander. Bei den Drei-Rechteck-Merkmalen liegen zwei Rechtecke immer auf derselben Höhe. Für die beiden Merkmalsarten ergeben sich zwei $(dx, dy)$ beziehungsweise drei $(dx, dy, dx')$ zusätzliche Freiheitsgrade.

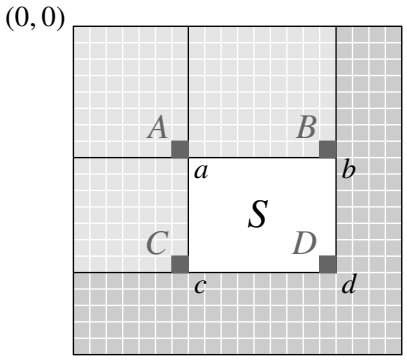



Abbildung 6.10: Berechnung der Pixelsumme $S$ eines rechteckigen Bereichs.

### Klassifikatoren

Die Klassifikation eines Bildausschnitts $x$ auf Basis der Merkmale erfolgt mit einfachen Schwellwertklassifikatoren. Zunächst wird zu jedem Merkmal $m_j$ ein Schwellwertklassifikator $w_j$, bestehend aus dem Merkmal, einem Schwellwert $\theta_j$ und einem Vorzeichenfaktor $p_j$, welcher die Richtung des Ungleichheitszeichens angibt, konstruiert:

$$w_j(x) = \begin{cases} 1 & \text{, wenn } p_j m_j(x) < p_j \theta_j \\ 0 & \text{, sonst} \end{cases}$$

Der Schwellwert $\theta_j$ wird auf Basis einer repräsentativen Stichprobe, bestehend aus Positiv- und Negativbeispielen, so bestimmt, dass die Fehlerrate minimal ist. Da einzelne Merkmale in der Regel nur eine geringe Aussagekraft mit Fehlerraten nahe bei 0,5 besitzen, spricht man von so genannten schwachen Klassifikatoren. Diese werden nun über eine gewichtete Summe zu starken Klassifikatoren kombiniert:

$$s_k(x) = \begin{cases} 1 & \text{, wenn } \sum_{i=1}^{T} \alpha_i w_i(x) > \phi_k \\ 0 & \text{, sonst} \end{cases} \qquad \text{mit } \phi_k = \frac{1}{2} \sum_{i=1}^{T} \alpha_i$$

Die Auswahl geeigneter Merkmale $m_j$ und die Bestimmung zugehöriger Gewichte $\alpha_j$ geschieht mit Hilfe von *Boosting* [Fre97]. Viola und Jones verwenden eine Variante des AdaBoost-Algorithmus. Dieser selektiert iterativ denjenigen schwachen Klassifikator, welcher die geringste Fehlerrate auf der Trainingsstichprobe aufweist. Die Gewichte $\alpha_i$ der schwachen Klassifikatoren hängen von der jeweils zugehörigen Fehlerrate ab und sind größer bei kleineren Fehlerraten. Nach jedem Selektionsschritt gewichtet der AdaBoost-Algorithmus die Trainingsbeispiele in Abhängigkeit von dem Klassifikationsergebnis des selektierten Klassifikators neu. Falsch klassifizierte Beispiele werden im weiteren Trainingsverlauf stärker gewichtet, während die Gewichtung korrekt klassifizierter Beispiele abnimmt. Durch diese Vorgehensweise konzentriert sich der Auswahlprozess auf die schwierigen Trainingsbeispiele.

Der Gesamtdetektor setzt sich schließlich aus einer Hierarchie von $N$ starken Klassifikatoren zusammen (siehe Abbildung 6.11). Ein Bildausschnitt wird zunächst an den ersten Klassifika-

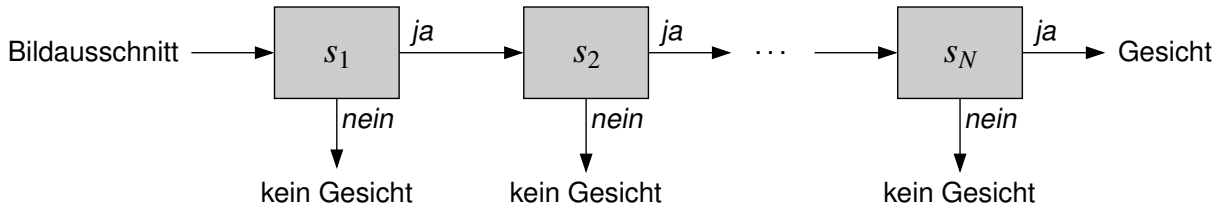



Abbildung 6.11: Detektorkaskade aus $N$ Einzelklassifikatoren. Ein Bildausschnitt wird nur dann als Gesicht erkannt, wenn er von allen Klassifikatoren als Gesicht klassifiziert wird.

tor $s_1$ übergeben. Wird der Bildausschnitt als Nicht-Gesicht bewertet, bricht die Klassifikation mit entsprechendem Ergebnis ab. Im anderen Fall wird der Bildausschnitt an den nachfolgenden Klassifikator weitergereicht. Dieser Prozess setzt sich solange fort, bis einer der Klassifikatoren den Bildausschnitt als Nicht-Gesicht bewertet. Nur wenn alle Klassifikatoren der Kaskade durchlaufen werden und auch der letzte Klassifikator $s_N$ eine positive Bewertung abgibt, wird der Bildausschnitt als Gesicht eingestuft.

Damit diese Vorgehensweise erfolgreich ist, muss die Rate der falsch-negativen Klassifikationsergebnisse der ersten $N-1$ Klassifikatoren der Kaskade annähernd Null sein. Ansonsten würden zu viele Positivbeispiele frühzeitig aussortiert. Die Rate der falsch-negativen Klassifikationen eines Klassifikators $s_k$ wird durch entsprechende Anpassung des zugehörigen Schwellwerts $\phi_k$ auf Basis einer Validierungsstichprobe erzielt. Zwar steigt dabei die Rate der falsch-positiven Ergebnisse, aber jeder Klassifikator sortiert dennoch eine Vielzahl von zu untersuchenden Bildausschnitten korrekt aus. Nur wenige Bildausschnitte gelangen in der Kaskade weit nach hinten.

Die Anzahl der Merkmale, aus denen die einzelnen starken Klassifikatoren zusammengesetzt sind, ist innerhalb der Kaskade nicht konstant. Sie nimmt, beginnend mit dem ersten Klassifikator $s_1$, zu. Für die Detektion frontaler Gesichter zum Beispiel basiert der erste Klassifikator $s_1$ häufig nur auf zwei Merkmalen. Dennoch können etwa 70% der Bildausschnitte schon auf dieser Stufe aussortiert werden. Durch die Kaskadenstruktur wird so im Mittel eine sehr effiziente Klassifikation erzielt.

### Verarbeitung eines Gesamtbilds

Das komplette Training geschieht auf einer festen Bildausschnittsgröße. Um bei der Verarbeitung eines Gesamtbilds Bildausschnitte mit anderen Größen zu klassifizieren, wird nicht das Bild skaliert, sondern die Merkmale. Die Eckpunkte der Rechtecke der skalierten Merkmale liegen dabei nicht notwendigerweise auf dem Raster des Integralbilds. Nach Lienhart und Kollegen [Lie02a] ist es eine geeignete Vorgehensweise, die Eckpunkte auf die jeweils nächstgelegenen Rasterpunkte des Integralbilds zu runden. Dabei wird den veränderten Rechteckgrößen entgegengewirkt, indem die Pixelsummen mit dem jeweiligen Größenverhältnis zwischen ursprünglichem und gerundetem Rechteck multipliziert werden.

Der Viola-Jones-Detektor ist robust gegenüber geringen Variationen in Translation oder Skalierung. Daher werden beim Abtasten des Gesamtbilds im Bereich von Gesichtern in der Regel mehrere Bildausschnitte positiv bewertet. Sich überlappende, positiv bewertete Bildausschnitte werden daher zu einem positiven Detektionsergebnis zusammengefasst. Die Position und Größe des Bildausschnitts des zusammengefassten Ergebnisses ergibt sich aus den jeweiligen Mittelwerten von Größe und Position der Einzeldetektionen. In der Regel findet man bei korrekten positiven Klassifikationen Überlappungen, während diese bei Fehlklassifikationen selten sind. Folglich kann die Falsch-Positiv-Rate gesenkt werden, wenn ein positives Detektionsergebnis nur für mindestens zwei sich überlappende Bildausschnitte ausgegeben wird.

**Anwendung in dieser Arbeit**

Die Trainingsstichprobe besteht aus Bildern der Größe 20×20 Pixel. Sie umfasst 5000 Positiv- und 7000 Negativbeispiele. Die Positivbeispiele zeigen Gesichter in frontaler Ansicht. Etwa 70% stammen von sieben Personen, die mit der Kamera des Roboters unter verschiedenen Beleuchtungsbedingungen aufgenommen wurden. Die übrigen 30% wurden aus Bildern aus dem Internet extrahiert. Für jedes Beispielgesicht wurden zusätzlich leicht um den Mittelpunkt des Ausschnitts rotierte Varianten in die Menge der Positivbeispiele übernommen. Es hat sich in der Praxis gezeigt, dass sich dadurch die Robustheit des Detektors gegenüber leicht seitlich geneigten Köpfen erhöht. Die Negativbeispiele sind zufällig ausgewählte Bildausschnitte ohne Gesichter.

Bei einer Bildausschnittsgröße von 20×20 Pixel sind knapp 36 Millionen verschiedene Zwei- und Drei-Rechteck-Merkmale möglich. Um den Trainingsaufwand zu reduzieren, wurde die Menge der Merkmale durch folgende Einschränkungen verkleinert: Die Werte für Breite und Höhe der Rechtecke sind aus dem Bereich $\{2, 3, 4, 5, 6\}$; bei der kleinsten Rechteck-Größe (2×2) sind die positions-bestimmenden Parameter ($x$, $y$, $dx$, $dy$, $dx'$) Vielfache von zwei; der maximale Abstand zwischen Rechtecken des Zwei-Rechteck-Merkmals beträgt 3 und zwischen Drei-Rechteck-Merkmalen 5. Die Anzahl der Merkmale wird dadurch um zwei Größenordnungen auf etwa 310.000 verkleinert.

Die Anzahl der Merkmale pro Klassifikator innerhalb der Kaskade ist den Werten aus [Vio04] angelehnt. Die Kaskade besteht aus 16 Klassifikatoren mit jeweils 2, 5, 30, zwei Mal 50, zehn Mal 100 und einmal 200 Merkmalen. Abbildung 6.12 zeigt die Merkmale der beiden ersten starken Klassifikatoren der trainierten Kaskade.

Bei schwacher Beleuchtung weisen zu klassifizierende Bildausschnitte einen sehr geringen Kontrastumfang auf. Bei der Varianznormierung wird insbesondere das Bildrauschen verstärkt. Da dieser Fall nicht durch die Trainingsstichprobe abgedeckt ist, werden Bildausschnitte mit einer geringen Standardabweichung der Helligkeitsverteilung der Pixel nicht klassifiziert. In der Anwendung hat sich ein Schwellwert von $\sigma_{max} = 13{,}8$ als geeignet erwiesen.

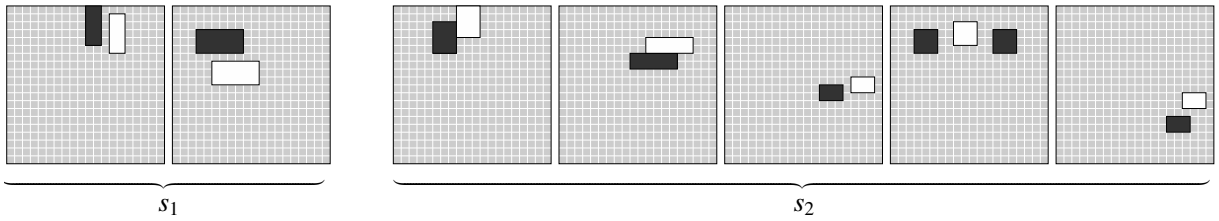



Abbildung 6.12: Merkmale der beiden ersten starken Klassifikatoren $s_1$ und $s_2$ der trainierten Kaskade. Auffällig ist, dass vorwiegend Bereiche in der oberen Gesichtshälfte berücksichtigt werden.

## 6.3 Lokalisation des Oberkörpers

Für die Lokalisation des Oberkörpers gibt es verschiedene Ansätze. Zum einen bildet der Kopf-Schulter-Bereich eine charakteristische Kontur, die sich im Bild detektieren lässt (vgl. zum Beispiel [Bla98, Wan99, Wil02]). Zum anderen sind Farbe und Textur der Kleidung geeignete Merkmale, über die der Oberkörper lokalisiert werden kann [Wal00, Doi01]. In dieser Arbeit wird ein farbbasierter Ansatz verwendet (siehe auch [Fri04]).

Zur Lokalisation des Oberkörpers werden im Bild Bereiche gesucht, deren Farbe zu einem zuvor gelernten Farbmodell passen. Als Farbmodell wird eine Gauß-Mischverteilung verwendet. Gauß-Mischverteilungen sind bereits erfolgreich zum Verfolgen von Gesichtern [Oli00] oder Getränkedosen [McK99] bei variierenden Beleuchtungsverhältnissen angewandt worden. Um die Farben typischer Kleidung zu modellieren, hat sich im praktischen Einsatz eine Gauß-Mischverteilung mit drei Komponenten als geeignet herausgestellt. Für sehr farbige Kleidung müsste die Anzahl der Komponenten entsprechend erhöht werden. Zur Repräsentation von Farbwerten wird der LUV-Farbraum [Wys82] verwendet. Dieser zeichnet sich dadurch aus, dass geometrische Abstände annähernd dem menschlichen Empfinden von Farbunterschieden entsprechen. Die Verwendung des LUV-Farbraums erlaubt es daher, ein homogenes Abstandsmaß einzusetzen, um die Übereinstimmung zwischen einer beobachteten Farbe und dem Farbmodell zu beurteilen.

Um im Detektionsprozess die Farbe der Kleidung vom Hintergrund zu unterscheiden, muss entweder ein eigenes Farbmodell für den Hintergrund oder ein geeignetes Rückweisungskriterium vorhanden sein. Da im Szenario mit dem mobilen Roboter der Hintergrund stark variieren kann, lässt sich kein einfaches, allgemeines Hintergrund-Farbmodell definieren. Es wird folglich ein Rückweisungskriterium verwendet. Dieses ist durch einen automatisch bestimmten Schwellwert realisiert. Bildpunkte, deren Farbwert im Farbmodell einen Wahrscheinlichkeitsdichtewert über (unter) dem Schwellwert ergeben, werden als Kleidung (Hintergrund) klassifiziert.

### 6.3.1 Ablauf des Verfahrens

Für ein gegebenes Bild der Größe $256 \times 192$ werden zunächst alle Bildpunkte bezüglich des Schwellwerts $\vartheta_s$ als Kleidung oder Hintergrund klassifiziert. Um isolierte Punkte zu entfernen

wird auf dem Ergebnis ein $5\times5$-Medianfilter angewendet. Anschließend werden benachbarte Bildpunkte zu Regionen zusammengefasst. Diese werden durch Polygone approximiert, auf denen die Merkmale Fläche und Flächenschwerpunkt berechnet werden. Regionen, deren Flächen außerhalb eines vorgegebenen Wertebereichs liegen, werden verworfen. Die größte verbleibende Region ist der detektierte Oberkörper.

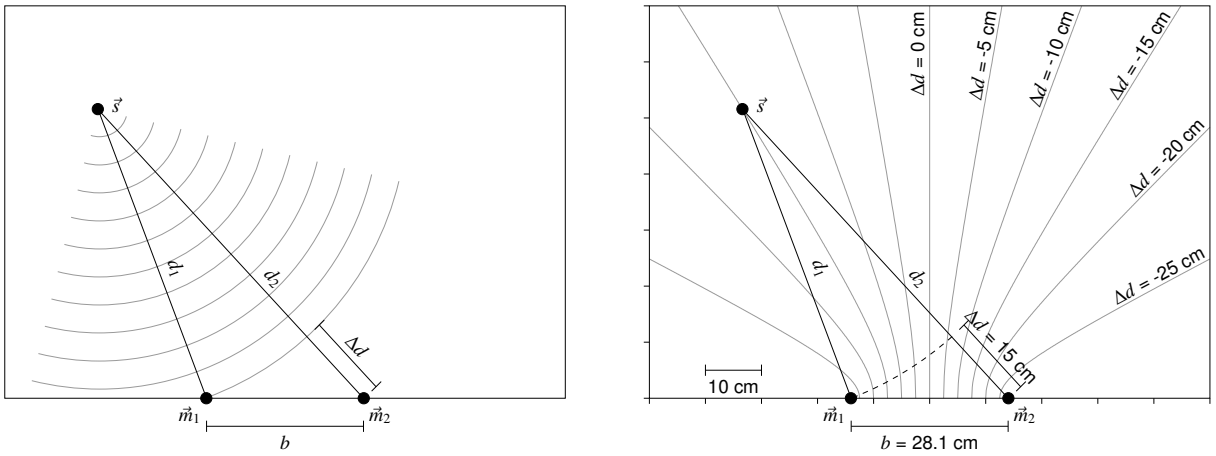
Da sich beim Einsatz des Verfahrens auf einem mobilen Roboter die Beleuchtung ändern kann, muss das Farbmodell der variierenden visuellen Erscheinung der Farbe der Kleidung angepasst werden. Dazu ist es notwendig, aus dem Bild geeignete Pixel zur Neuschätzung der Mischverteilung zu bestimmen. Es wird dabei der Bildbereich betrachtet, der sich durch Skalierung der polygonalen Beschreibung der Region bezüglich des Flächenschwerpunkts auf das 1,5-fache der Ursprungsfläche ergibt. Innerhalb dieses Bereichs werden Bildpunkte für das Training wiederum durch Klassifikation über das aktuelle Farbmodell bestimmt. Die Klassifikation erfolgt jedoch bezüglich eines Schwellwerts $\vartheta_t$, der größer als der bei der Segmentierung verwendete Schwellwert $\vartheta_s$ ist. Durch den großzügigeren Schwellwert werden mehr Bildpunkte gefunden, die ähnlich zur aktuellen Farbe der Kleidung sind. Auf der Menge der Trainingspixel wird die Gauß-Mischverteilung über den *k-means*-Algorithmus [Mac67] neu geschätzt.

#### Initialisierung

Beim Start des Verfahrens liegt kein Farbmodell vor. Folglich ist die oben beschriebene Vorgehensweise zur Bestimmung von Bildpunkten für die initiale Schätzung der Gauß-Mischverteilung nicht anwendbar. Um einen Bildbereich zu bestimmen, der sich im Bereich der Kleidung befindet, wird daher auf multimodale Informationen des Personen-*Trackings* zurückgegriffen. Wenn für eine verfolgte Person das erste Mal ein Gesicht detektiert wurde, ist die Größe der Person bekannt. Aus der Höhe des Gesichts kann die Position der Kleidung geschätzt werden. Es wird ein ellipsenförmiger Bereich betrachtet, der sich 35 cm unterhalb des Gesichts befindet. Die Größe des Bildbereichs hängt vom Abstand der Person zur Kamera ab. Alle Pixel in diesem Bildausschnitt werden zur initialen Schätzung des Farbmodells verwendet.

#### Bestimmung der Schwellwerte

Die Schwellwerte $\vartheta_s$ und $\vartheta_t$ werden nach jedem Adaptionsschritt neu bestimmt. Dazu wird zunächst eine diskrete Verteilung der Wahrscheinlichkeitsdichtewerte für die in der Trainingsmenge vorkommenden Farbwerte durch ein entsprechendes Histogramm erstellt. Jeder Eintrag des Histogramms stellt einen kleinen Bereich der möglichen Dichtewerte dar. Die Schwellwerte werden so festgelegt, dass ein vorgegebener Anteil $f$ der Trainingspixel positiv klassifiziert wird, das heißt, die Wahrscheinlichkeit dafür, dass die Bewertung der Trainingspixel durch das Farbmodell über dem Schwellwert $\vartheta$ liegt, ist $f$. In der Anwendung auf dem Roboter *BIRON* im gewählten Szenario haben sich Werte von $f_s = 0{,}98$ und $f_t = 0{,}99$ als geeignet herausgestellt.




(a) (b)

Abbildung 6.13: (a) Ausbreitung von Schallwellen von einer Schallquelle $\vec{s}$. Der Abstand der Schallquelle zu den Mikrofonen $\vec{m}_1$ und $\vec{m}_2$ unterscheidet sich im Allgemeinen um den Wert $\Delta d$. (b) Alle Schallquellen, die zu einem Laufstreckenunterschied $\Delta d$ führen, liegen auf einer Hälfte eines zweischaligen Hyperboloids.

## 6.4 Sprecherlokalisation

Ein Sprecher kann als punktförmige Geräuschquelle $\vec{s}$ im Raum aufgefasst werden (siehe Abbildung 6.13 (a)). Der Abstand von $\vec{s}$ zu einem Mikrofon $\vec{m}_i$ sei $d_i = ||\vec{s} - \vec{m}_i||$. Ein akustisches Ereignis $e(t)$, das zum Zeitpunkt $t$ von der Schallquelle ausgeht, trifft nach einer Zeitspanne $\tau_i = \frac{d_i}{c}$ auf das Mikrofon $\vec{m}_i$, wobei $c$ die Ausbreitungsgeschwindigkeit von Schall ist. Das Ereignis findet sich im aufgezeichneten Signal $s(t)$ mit entsprechender Verzögerung wieder:

$$s_i(t) = a_i e(t - \tau_i) + n_i(t)$$

Dabei ist $a_i$ ein durch die Ausbreitung bestimmter Dämpfungsfaktor, $n_i(t)$ fasst alle zusätzlich aufgezeichneten Störgeräusche zusammen.

Im Allgemeinen unterscheiden sich die Abstände $d_1$ und $d_2$ bei zwei Mikrofonen $\vec{m}_1$ und $\vec{m}_2$ um die Differenz $\Delta d = d_2 - d_1$. Im Vergleich der Signale von linkem und rechtem Mikrofon sind die aufgezeichneten akustischen Ereignisse der Schallquelle $\vec{s}$ folglich um $\delta = \tau_2 - \tau_1$ zueinander verschoben. Wenn es gelingt, aus den Signalen die Differenz $\delta$ zu ermitteln, so lässt sich wiederum mit Hilfe von Geometrie die Position der zugehörigen Schallquelle abschätzen. Allerdings kann aus der resultierenden Abstandsdifferenz $\Delta d$ die Position im Raum nicht auf einen Punkt oder eine Richtung eingeschränkt werden. Selbst im ebenen Fall, wenn sich alle Geräuschquellen auf derselben Höhe wie die Mikrofone befinden, kann die genaue Position von $\vec{s}$ nur bestimmt werden, wenn der Abstand bekannt ist oder andere zusätzliche Annahmen gemacht werden. In einer vereinfachten Geometrie kann der Abstand der Mikrofone $b = ||\vec{m}_2 - \vec{m}_1||$ als

hinreichend klein gegenüber dem Abstand der Geräuschquelle angesehen werden. Unter dieser Annahme sind die Winkel, unter denen der Schall auf die beiden Mikrofone trifft, annähernd gleich und können direkt aus $\Delta d$ berechnet werden.

Im dreidimensionalen Fall hängt die beobachtete Zeitdifferenz nicht nur von der Richtung und dem Abstand, sondern auch von der relativen Höhe der Geräuschquelle in Bezug zu den Mikrofonen ab. Wenn nur $\Delta d$ gegeben ist, dann ist das Problem der Lokalisierung unterbestimmt. Alle Geräuschquellen, die zum selben $\Delta d$ führen, liegen auf einer Hälfte eines zweischaligen Hyperboloids, gegeben durch

$$\frac{s_x^2 + s_z^2}{\frac{1}{4}\left(b^2 - (\Delta d)^2\right)} - \frac{s_y^2}{\frac{1}{4}(\Delta d)^2} = -1$$

wobei $(s_x, s_y, s_z)$ die Position der Geräuschquelle in karthesischen Koordinaten ist. Der Mittelpunkt des Hyperboloids befindet sich bei $\frac{\vec{m}_1+\vec{m}_2}{2}$ und seine Symmetrieachse fällt mit der Verbindungslinie der beiden Mikrofone zusammen. Abbildung 6.13 (b) zeigt einen Schnitt der Ebene, die durch die Mikrofone $\vec{m}_1$ und $\vec{m}_2$ und die Schallquelle $\vec{s}$ aufgespannt wird, mit Hyperboloiden zu verschiedenen $\Delta d$.

Um die Position des Sprechers im Raum auf einen Punkt einschränken zu können, müssen mindestens drei Zeitdifferenzen für entsprechend viele verschiedene Mikrofonpaare bestimmt werden. Auf dem Roboter *Lino* [Krö03] werden zum Beispiel drei jeweils senkrecht zueinander stehende Mikrofonpaare zur Lokalisation des Sprechers eingesetzt. Auf einigen Robotern wird jedoch auch nur ein einzelnes Mikrofonpaar für die Sprecherlokalisation verwendet (siehe zum Beispiel [Mat03, Oku02]). Hier wird aber offensichtlich die Annahme gemacht, dass sich alle Sprecher auf derselben Höhe befinden.

Auch in dieser Arbeit besitzt der verwendete Roboter *BIRON* lediglich ein Mikrofonpaar. Um dennoch den aktuellen Sprecher im Raum zu lokalisieren, werden Informationen aus dem multimodalen *Tracking* über potenzielle Geräuschquellen hinzugezogen. Befindet sich eine Person, beziehungsweise ihr Mund, nahe zu dem aus dem $\Delta d$ bestimmten Hyperboloid, dann wird angenommen, dass diese Person der Sprecher ist (vgl. [Lan03]).

### 6.4.1 Schätzung der Zeitverzögerung

Schnelle und robuste Techniken für die Lokalisierung von Geräuschen sind zum Beispiel die generalisierte Kreuzkorrelationsmethode [Kna76] oder die daraus abgeleitete *Crosspower-Spectrum Phase (CSP)*-Analyse [Giu94], die jeweils sowohl für Mikrofonfelder als auch für einzelne Mikrofonpaare eingesetzt werden können. Komplexere Algorithmen zur Lokalisation von Sprechern, wie zum Beispiel *Spectral Separation and Measurement Fusion* [Ber02] oder *Linear-Correction Least-Squares* [Hua01] sind ebenfalls sehr robust und können zusätzlich den Abstand und die Höhe des Sprechers schätzen oder mehrere Audioquellen separieren. Diese komplexen Algorithmen benötigen mehr als ein Mikrofonpaar, um angemessene Ergebnisse zu liefern und sind darüber hinaus sehr rechenintensiv.

Um die Zeitverschiebung zu bestimmen, wird in dieser Arbeit die *CSP*-Analyse [Giu94] verwendet. Zur Schätzung der Zeitverschiebung $\delta$ zum Zeitpunkt $t$ wird ein spektrales Korrelationsmaß

$$C(t,\tau) = FT^{-1}\left(\frac{\hat{S}_L(t,f)\ \hat{S}_R^*(t,f)}{|\hat{S}_L(t,f)|\ |\hat{S}_R(t,f)|}\right)$$

berechnet, wobei $\hat{S}_L(t,f)$ und $\hat{S}_R(t,f)$ die Kurzzeitspektren des linken und rechten Kanals um den Zeitpunkt $t$ sind. Wenn nur eine einzelne Geräuschquelle vorhanden ist, ist die Zeitverzögerung $\delta$ zum Zeitpunkt $t$ durch das Argument $\tau$ gegeben, welches das spektrale Korrelationsmaß maximiert:

$$\delta = \underset{\tau}{\operatorname{argmax}}\ C(t,\tau)$$

Werden neben dem Hauptmaximum auch Nebenmaxima berücksichtigt, ist es möglich, mehrere Geräuschquellen gleichzeitig zu detektieren.

#### Genauigkeit der Lokalisation

Auf dem Roboter *BIRON* erfolgt die Signalaufzeichnung mit einer Abtastrate von 16 kHz oder 48 kHz. Die minimal messbare Verschiebung zweier Signale entspricht bei 16 kHz (48 kHz) folglich einem Laufläng enunterschied von 21,25 mm (7,08 mm).[3] Da der Abstand $b$ der Mikrofone 281 mm beträgt, gibt es 27 (79) verschiedene Messwerte, die den möglichen Verschiebungen im Bereich zwischen $-b$ und $+b$ entsprechen. Die Winkelauflösung der Sprecherlokalisation beträgt damit in dem Bereich vor dem Roboter etwa 5° (1,5°) und nimmt für äußere Positionen deutlich ab (siehe Abbildung 6.14).

## 6.5 Bestandteile im multimodalen *Anchoring* für Personen

Nachdem in den vorangegangenen Abschnitten die Verfahren zur multimodalen Erfassung von Personen beschrieben wurden, beschäftigt sich der weitere Teil des Kapitels mit der Ausgestaltung der Bestandteile im multimodalen *Anchoring* für Personen. Dies umfasst die Definition der Attribute für die Perzepte auf Ebene der unimodalen *Anchor*, die Spezifikation der drei Modelle für Komposition, Bewegung und Fusion vom multimodalen *Anchor* sowie die Gestaltung der Unterfunktionen, die im *Anchoring*-Prozess verwendet werden.

Einen Überblick über den Gesamtaufbau vom multimodalen *Anchoring* für Personen verschafft die Abbildung 6.15. Drei Sensoren des Roboters liefern die Daten für vier unimodale *Anchoring*-Prozesse. Es gibt je einen *Anchoring*-Prozess für die Beine, das Gesicht, den Oberkörper und die Stimme einer Person. Die Generierung von Perzepten in den Sensorsystemen erfolgt mit Hilfe der in den vier vorangegangenen Abschnitten beschriebenen Mustererkennungsverfahren.

[3] Die Schallgeschwindigkeit wird dabei als 340 m/s angenommen.

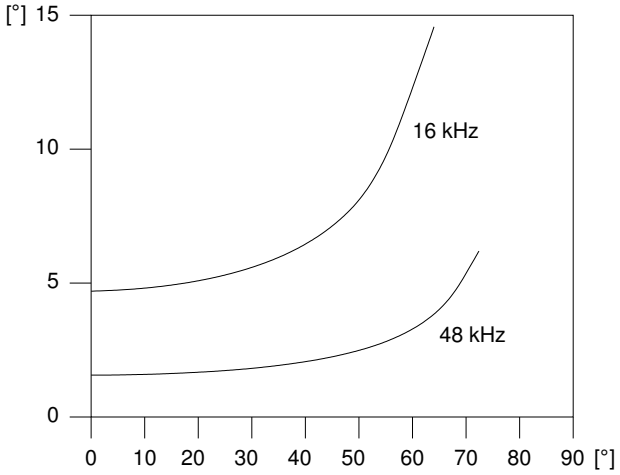



Abbildung 6.14: Zusammenhang zwischen dem Winkel (Abszisse) und der Winkelauflösung der Sprecherlokalisation (Ordinate) für eine typische Position eines Sprechers (1,5 m Abstand und 0,6 m über den Mikrofonen).

Die verschiedenen Perzepte sind dabei wie folgt definiert: Beinperzepte sind alle zu Paaren gruppierte Segmente (Beinpaarperzepte) und die Einzelsegmente, die mit keinem anderen Segment zu einem Paar gruppiert werden konnten (Einzelbeinperzepte). Bei der Gesichtsdetektion gelten die quadratischen Bildausschnitte, die als Gesicht klassifiziert wurden, als Perzepte. Analog sind bei der Lokalisation des Oberkörpers die segmentierten Regionen im Bild die Perzepte. Als Perzepte der Sprecherlokalisation werden die Laufzeitdifferenzen der detektierten Geräuschquellen betrachtet.

### 6.5.1 Attribute für Perzepte in den unimodalen *Anchoring*-Prozessen

Im *Anchoring* werden mit den Attributen die aus den Perzepten extrahierbaren Eigenschaften beschrieben. Die Attributwerte können über die Prädikat-*Grounding*-Relation mit der symbolischen Beschreibung verglichen werden, um zu entscheiden, ob ein Perzept an ein vorgegebenes Symbol geknüpft werden kann. Bei dem hier realisierten multimodalen *Anchoring* von Personen wird die Möglichkeit, bestimmte Perzepte durch Vorgabe einer symbolischen Beschreibung auszuschließen, nicht genutzt. Für die geplante Aufmerksamkeitssteuerung ist es notwendig, alle Personen zu verfolgen, die sich in den Wahrnehmungsbereichen der Sensoren des Roboters aufhalten. Es werden daher alle generierten Perzepte beim Verfolgen von Personen verwertet. Die Attribute, die in den unimodalen *Anchoring*-Prozessen verwendet werden, dienen vorwiegend zur Beschreibung der Position der beobachteten Personenbestandteile. Die bezugsmäßige Korrektheit der multimodalen *Anchor* wird über einen Vergleich der geschätzten Positionswerte realisiert.

Die Schätzung der Position aus den Perzepten ist im Allgemeinen fehlerbehaftet. Um Unsicher-

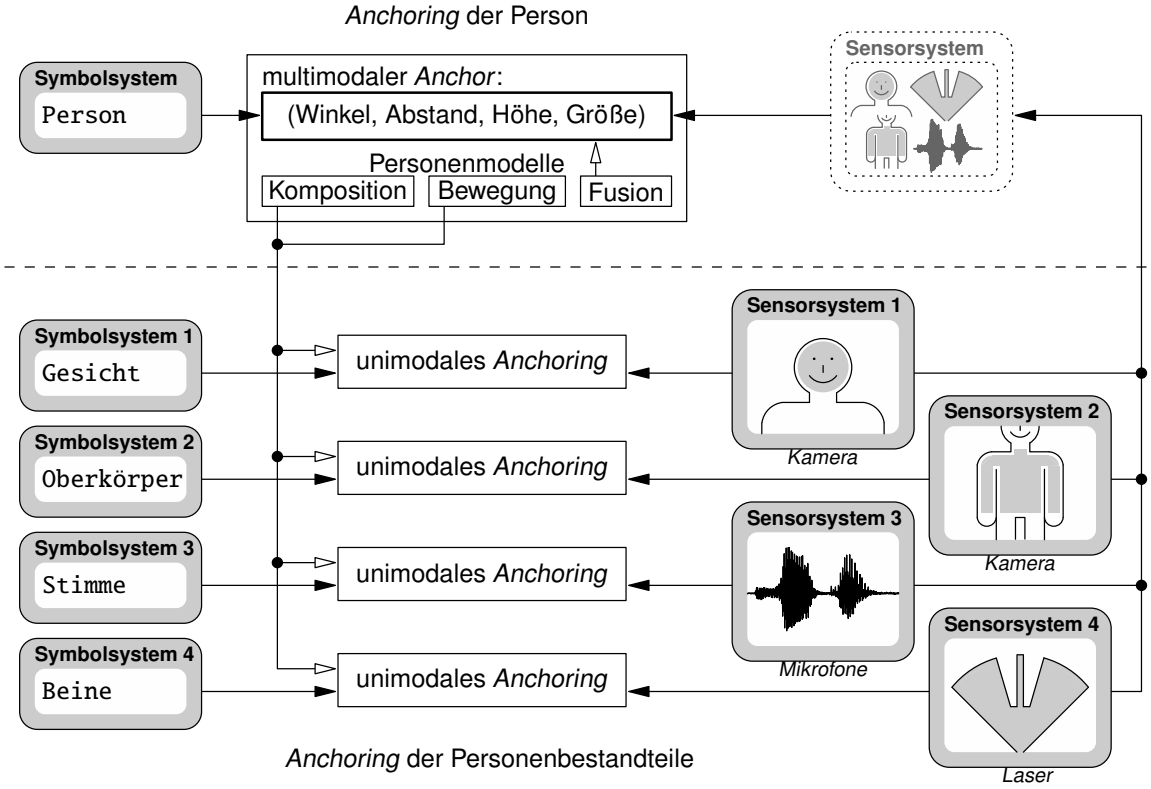



Abbildung 6.15: Multimodales *Anchoring* von Personen

heit zu modellieren, werden Attributwerte nicht durch skalare Werte, sondern durch Gaußverteilungen beschrieben. Für Attribute werden also Wertepaare $(z, R)$ angegeben, die aus der gemessenen Position $z$ und der Unsicherheit $R$ (Varianz) bestehen. Die Varianz der Gaußverteilungen kann aus den Messgenauigkeiten der Sensoren und den Genauigkeiten der Verfahren zur Generierung von Perzepten bestimmt werden.

Alle Positionsangaben erfolgen im Zylinderkoordinatensystem des Roboters durch Winkel, Abstand und Höhe. Im Folgenden werden die Attributmengen der vier unimodalen *Anchor* beschrieben.

**Beine:** Aus jedem Beinperzept kann die Position der entsprechenden Person geschätzt werden. Für jedes in den Perzepten vorkommende Segment wird zunächst der Schwerpunkt der jeweiligen Messpunkte berechnet. Für Einzelbeinperzepte, die nur aus einem Segment bestehen, ist die Position direkt durch den entsprechenden Schwerpunkt des Segments definiert. Bei Beinpaarperzepten, die aus einem Paar von Segmenten bestehen, wird der Mittelpunkt der zwei Schwerpunkte der Segmente als Position der Person definiert. Die Position wird im Zylinderkoordinatensystem des Roboters durch Abstand und Winkel angegeben. Die zugehörigen Attribute sind dementsprechend `Abstand` und `Winkel`.

Zusätzlich zu den positionsbeschreibenden Attributen wird das Attribut `ist_Paar` einge-

führt, welches beschreibt, ob es sich um ein Einzelbein- oder ein Beinpaarperzept handelt. Diese Information wird beim Neuaufbau eines multimodalen *Anchor* berücksichtigt. Da bei einzelnen Segmenten, die mit keinem anderen Segment zu einem Paar gruppiert werden konnten, die Falsch-Positiv-Rate relativ hoch ist, da zum Beispiel Objekte wie Vasen und Tischbeine ebenfalls in einzelnen Segmenten resultieren können, werden Einzelbeinperzepte nicht zum Neuaufbau eines multimodalen *Anchor* verwendet. Für bereits bestehende multimodale *Anchor* ist die Position der entsprechenden Person bereits durch andere Modalitäten bekannt. In diesem Fall werden auch Einzelbeinperzepte zugeordnet, da anzunehmen ist, dass es sich bei diesen Perzepten an der geschätzten Position tatsächlich um ein Bein der Person handelt.

Die Attributmenge $\Phi_{leg}$ im unimodalen *Anchor* für Beine sieht damit wie folgt aus:

$$\Phi_{leg} = \{\texttt{Abstand}, \texttt{Winkel}, \texttt{ist\_Paar}\}$$

**Gesicht:** Aus einem Gesichtsperzept wird eine Person im Raum lokalisiert und ihre Größe geschätzt. Aus der Position des Gesichts im Bild kann unter Berücksichtigung der Stellung der Kamera ein Strahl bestimmt werden, der von der Kamera ausgeht, auf dem sich der Mittelpunkt des Gesichts im Raum befindet. Der Winkel zwischen der $x$-Achse und der Projektion des Strahls auf die horizontale $xy$-Ebene gibt die Winkelkomponente der in Zylinderkoordinaten beschriebenen Position der Person an. Um auch noch den Abstand und die Höhe bestimmen zu können, ist es notwendig, den Abstand des Gesichts von der Kamera zu ermitteln, also den Punkt auf dem Strahl, an dem sich das Gesicht befindet. Unter der Annahme, dass die Größe von Gesichtern von ausgewachsenen Menschen näherungsweise konstant ist, ist der gesuchte Abstand umgekehrt proportional zu der Größe der Abbildung des Gesichts im Kamerabild. Aus dem Abstand des Gesichts zur Kamera ergibt sich die Abstands- und die Höhenkomponente der in Zylinderkoordinaten beschriebenen Position der Person. Die Attributmenge $\Phi_{face}$ für Gesichter ist damit:

$$\Phi_{face} = \{\texttt{Abstand}, \texttt{Winkel}, \texttt{Höhe}\}$$

**Oberkörper:** Bei der Lokalisation von Kleidung werden die segmentierten Regionen als Perzepte aufgefasst. Um daraus Information über die Position einer Person abzuleiten, wird ähnlich wie bei den Gesichtsperzepten vorgegangen. Als zentraler Punkt der Region wird der Flächenschwerpunkt betrachtet. Unter Berücksichtigung der Stellung der Kamera lässt sich ein Strahl im Raum bestimmen, der bei der Kamera beginnt und in Richtung des zentralen Punkts des Oberkörpers der Person verläuft. Daraus lässt sich die Winkelkomponente der Position im Zylinderkoordinatensystem des Roboters berechnen. Um auch die Abstands- und Höhenkomponente zu erlangen, müsste der Abstand des Oberkörpers von der Kamera bestimmt werden. Bei Gesichtern konnte ausgenutzt werden, dass der Abstand eines Gesicht umgekehrt proportional zur Größe der Abbildung ist. Bei der Lokalisation des Oberkörpers gibt es jedoch etliche weitere Faktoren, die die Größe der segmentierten Region stark beeinflussen. Dazu gehören zum Beispiel die Art der Kleidung und der Umstand, dass der Oberkörper nicht immer vollständig von der Kamera erfasst wird. Der

Tabelle 6.1: Extrahierbare Positionsinformationen der verschiedenen Perzeptarten:

| Perzeptart | Winkel | Abstand | Höhe | sonstiges |
|---|---|---|---|---|
| Beine | ja | ja | – | – |
| Gesicht | ja | ja | ja | – |
| Oberkörper | ja | – | – | – |
| Stimme | – | – | – | Hyperboloid |

Abstand kann folglich nicht bestimmt werden. Die Attributmenge für die Perzepte der Lokalisation des Oberkörpers enthält somit nur das Attribut `Winkel`:

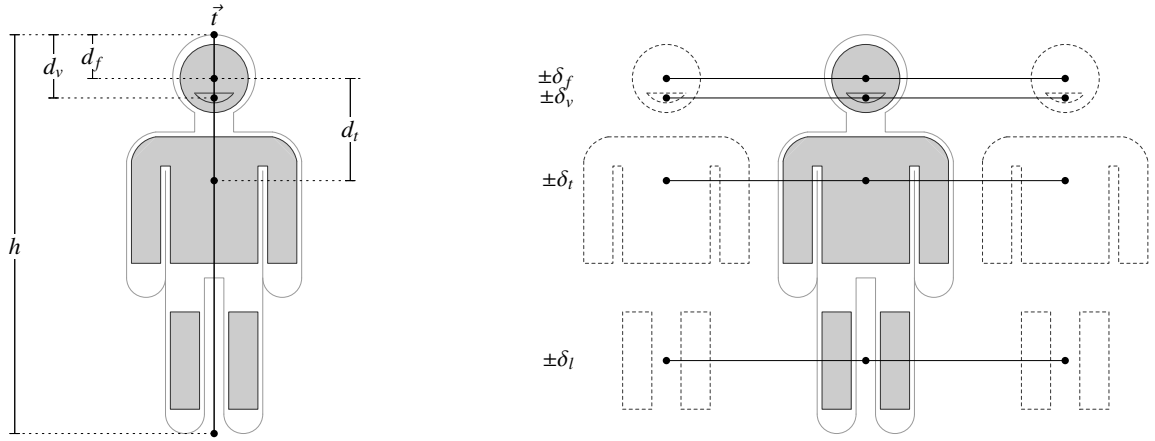
$$\Phi_{torso} = \{\texttt{Winkel}\}$$

**Stimme:** Die Perzepte der Sprecherlokalisation sind die Laufzeitdifferenzen der lokalisierten Geräuschquellen. Wie in Abschnitt 6.4 erläutert, kann aus den Laufzeitdifferenzen die Position der Person auf eine Hälfte eines zweischaligen Hyperboloids eingegrenzt werden. Damit ist es jedoch nicht möglich, eine der Komponenten der Position der Person in Zylinderkoordinaten anzugeben. Die Attributmenge für Perzepte der Sprecherlokalisation beinhaltet damit nur die Laufzeitdifferenz:

$$\Phi_{voice} = \{\texttt{Laufzeitdifferenz}\}$$

Dass keine direkte Positionsinformation abzuleiten ist, hat insbesondere zur Folge, dass ein multimodaler *Anchor* über ein Perzept der Sprecherlokalisation nicht neu aufgebaut werden kann. Es ist jedoch möglich, Perzepte den bereits existierenden multimodalen *Anchor*-Funktionen zuzuordnen: Aus der zu einem multimodalen *Anchor* zugehörigen Konfiguration kann die Position des Munds beziehungsweise der Geräuschquelle im Raum vorhergesagt werden. Der minimale Abstand zwischen dem Mund und dem durch die Laufzeitdifferenz bestimmten Hyperboloid stellt dabei das Bewertungskriterium dar. Ist der Abstand gering, wird das Perzept der Person zugeordnet, das heißt es wird angenommen, dass die Person spricht. Konnte für keine der verfolgten Personen ein kleiner Abstand ermittelt werden, wird das Perzept verworfen.

Die Tabelle 6.1 gibt einen Überblick über die Positionsinformationen, die aus den verschiedenen Perzeptarten extrahiert werden können. Information über die Höhe lässt sich lediglich aus Gesichtsperzepten ableiten, und ist daher für einen multimodalen *Anchor* solange unbekannt, bis diesem zum ersten Mal ein Gesichtsperzept zugeordnet werden konnte. Die Kenntnis der Höhe ist jedoch Voraussetzung dafür, dass auch Stimmenperzepte zugewiesen werden können, da dafür die Position des Munds bekannt sein muss. Um Stimmenperzepte für einen multimodalen *Anchor* auch dann verwerten zu können, wenn noch kein Gesicht detektiert wurde, wird die Höhe zunächst als eine für ausgewachsene Menschen durchschnittliche Höhe angenommen.




(a) (b)

Abbildung 6.16: Das Kompositionsmodell: (a) Standardkonfiguration, (b) zulässige Abweichungen von der Standardkonfiguration.

### 6.5.2 Modelle im multimodalen *Anchor*

Im multimodalen *Anchoring* bilden das Kompositions- und das Bewegungsmodell die Grundlage für die multimodale Datenassoziation. Das Kompositionsmodell gibt die räumlichen Relationen der einzelnen Bestandteile an, die durch die unimodalen *Anchoring*-Prozesse beobachtet werden. Das Bewegungsmodell beschreibt zum einen die zeitliche Positionsveränderung der einzelnen Bestandteile im Bezug zueinander und zum anderen die Bewegung des Gesamtobjekts im Raum. Beide Modelle zusammen erlauben die Vorhersage der Position von einzelnen Bestandteilen. Über einen Vergleich der vorhergesagten Position mit der für ein neu generiertes Perzept gemessene Position wird entschieden, ob das Perzept dem multimodalen *Anchor* zugeordnet werden kann. Das Fusionsmodell realisiert schließlich das *Tracking*. Es beschreibt, wie die Positionsinformation der zugeordneten Perzepte genutzt wird, um die aktuelle Konfiguration des beobachteten Objekts zu schätzen. Im Folgenden werden die drei Modelle für das multimodale *Anchoring* für Personen beschrieben.

**Kompositionsmodell:** Das Kompositionsmodell beschreibt die räumlichen Relationen von Beinen, Gesicht, Oberkörper und Stimme einer Person. Mit Hinblick auf die geplante Anwendung, bei der der Roboter seine Aufmerksamkeit auf Personen richtet, die an ihn herantreten, um eine Interaktion zu beginnen, werden Menschen als aufrecht stehend modelliert. Da keines der Perzepte einen direkten Hinweis auf die Orientierung der Person in Bezug auf den Roboter liefert, wird ein einfaches, rotationssymmetrisches Modell verwendet. In der Standardkonfiguration befinden sich alle Bestandteile auf einer vertikalen Achse (siehe Abbildung 6.16). Den Hauptbezugspunkt bildet der Punkt $\vec{t}$, der sich an der Oberkante des Kopfs befindet und durch seine Höhe $h$ über dem Boden die Größe der Person festlegt. Zulässige Werte für die Größe werden durch einen Wertebereich eingeschränkt, der die Körpergrößen sehr kleiner und sehr großer Menschen umfasst. In konstantem Abstand unter

dem Bezugspunkt befinden sich das Gesicht ($d_f$ = 16 cm) und der Mund beziehungsweise die Stimme ($d_v$ = 20 cm). Obwohl es für die Zuordnung von Perzepten nicht notwendig ist, wird auch der Abstand des Oberkörpers vom Gesicht spezifiziert ($d_t$ = 35 cm). Dieser Wert findet bei der Initialisierung der Lokalisation des Oberkörpers Verwendung (siehe Abschnitt 6.3.1).

Alle vier Bestandteile können von der Standardkonfiguration um einen jeweiligen maximalen Betrag abweichen. Dadurch werden unterschiedliche Körperhaltungen berücksichtigt, zum Beispiel, wenn eine Person leicht geneigt steht. Das Gesicht darf um $\delta_f$ = 60 cm von der Standardposition abweichen, der Mund um $\delta_v$ = 20 cm. Für den Oberkörper ist eine maximale Abweichung von $\delta_t$ = 30 cm festgelegt. Dies betrifft jedoch nur die seitliche Richtung aus Sicht des Roboters, da aus dem Oberkörperperzept lediglich die Winkelkomponente der Zylinderkoordinaten gewonnen werden kann. In entsprechender Weise ist mit $\delta_l$ = 50 cm die Abweichung der Beine von der Standardposition nur innerhalb der horizontalen Ebene gemeint.

Das erste Perzept, mit dem ein multimodaler *Anchor* neu aufgebaut wird, bestimmt die anfängliche Position des Hauptbezugspunkts. Der Aufbau des multimodalen *Anchor* kann nur durch ein Bein- oder ein Gesichtsperzept erfolgen. Während bei einem Gesichtsperzept die Personengröße sofort bekannt ist, bleibt dieser Wert bei einem Beinperzept unbestimmt. In diesem Fall wird solange eine provisorische Personengröße (180 cm) verwendet, bis die erste Information über die tatsächliche Personengröße durch ein Gesichtsperzept vorliegt.

**Bewegungsmodell:** Das Bewegungsmodell erlaubt es, zwei Arten von Bewegung zu modellieren: Zum einen die Veränderung der Konfiguration des Kompositionsmodells mit der Zeit, das heißt die relative Lageveränderung der einzelnen Bestandteile einer Person, und zum anderen die Bewegung von Personen als Gesamtobjekt im Raum. Da sich durch Konfigurationsveränderungen in dem einfachen, rotationssymmetrischen Personenmodell keine aussagekräftigen, menschentypischen Bewegungen beschreiben lassen, die für die beabsichtigte Anwendung relevant sind, wird diese Art von Bewegung nicht explizit modelliert. Das bedeutet, dass zu aufeinander folgenden Zeitpunkten beliebige Konfigurationsänderungen im Rahmen des Kompositionsmodells zulässig sind. Das Bewegungsmodell spezifiziert folglich nur die Bewegung der Person als Ganzes.

Die Position der Person wird im Zylinderkoordinatensystem des Roboters durch Winkel, Abstand und Höhe angegeben. Da die genaue Personenposition nicht bekannt ist, werden die einzelnen Komponenten jeweils durch Gaußverteilungen beschrieben. Ein Mittelwert $x$ gibt den geschätzten Positionswert an, die Varianz $P$ modelliert die Unsicherheit über die Schätzung. Das Bewegungsmodell definiert nun, wie sich aus der Position $(x_{t_{i-1}}, P_{t_{i-1}})$ zum Zeitpunkt $t_{i-1}$ die Position $(\hat{x}_{t_i}, \hat{P}_{t_i})$ zum Zeitpunkt $t_i$ vorhersagen lässt. Es wird angenommen, dass die Person sich nicht bewegt. Daher bleibt der Mittelwert unverändert. Gleichzeitig steigt die Varianz linear mit der Zeit und beschreibt so die wachsende Unsi-

cherheit:

$$\hat{x}_{t_i} = x_{t_{i-1}} \tag{6.1}$$
$$\hat{P}_{t_i} = P_{t_{i-1}} + Q(t_i - t_{i-1}) \tag{6.2}$$

**Fusionsmodell:** Das Fusionsmodell definiert, wie bei Zuordnung eines Perzepts mit dem Zeitstempel $t_i$ zu einem multimodalen *Anchor* die neue Personenposition für den entsprechenden Zeitpunkt $t_i$ geschätzt wird. Diese ergibt sich aus einer Kombination der Vorhersage der Position $(\hat{x}_{t_i}, \hat{P}_{t_i})$ für den Zeitpunkt $t_i$ durch das Bewegungsmodell (Gleichungen 6.1 und 6.2) und dem entsprechenden Attributwert des Perzepts $(z, R)$ nach folgenden Formeln:

$$x_{t_i} = \frac{\hat{x}_{t_i} R + z\hat{P}_{t_i}}{R + \hat{P}_{t_i}} \tag{6.3}$$
$$P_{t_i} = \frac{R\hat{P}_{t_i}}{R + \hat{P}_{t_i}} \tag{6.4}$$

Die Berechnungsvorschriften im Bewegungsmodell (Gleichungen 6.1 und 6.2) und im Fusionsmodell (Gleichungen 6.3 und 6.4) entsprechen zusammen einem Kalman-Filter [Kal60] mit einem einfachen Übergangsmodell. Ein komplexeres Modell, dass auch Geschwindigkeit berücksichtig, ist nicht geeignet, da, wie Brooks und Williams in [Bro03] feststellen, sich ein Mensch nur schwer als ein Punkt mit konstanter Geschwindigkeit beschreiben lässt. In dem hier beschriebenen Verfahren kommt erschwerend hinzu, dass die Eigenbewegung des Roboters die relative Bewegung der Personen überlagert.

### 6.5.3 Unterfunktionen in den Basisfunktionen

Die Basisfunktionen vom *Anchoring* (siehe Abschnitt 5.1.2) rufen eine Reihe domänenabhängiger Unterfunktionen auf, die für die jeweilige Anwendung spezifiziert werden müssen. Aus dem herkömmlichen *Anchoring* stammen die Unterfunktionen $\mathrm{Predict}(\cdot)$, $\mathrm{Verify}(\cdot)$ und $\mathrm{Update}(\cdot)$. Die Funktion $\mathrm{Predict}(\cdot)$ dient zur Vorhersage der zu erwartenden Signaturen und bedient sich dazu des Bewegungsmodells. In der Funktion $\mathrm{Verify}(\cdot)$ werden die Perzepte herausgefiltert, die sich einem *Anchor* zuordnen lassen. Im multimodalen *Anchoring* von Personen wird dazu geprüft, ob der Abstand der gemessenen Position zur vorhergesagten Position im Rahmen des durch das Kompositionmodell erlaubten Bereichs liegt. Da im Symbolsystem keine Prädikate verwendet werden, ist ein Abgleich der perzeptuellen Signatur mit einer symbolischen Beschreibung hier nicht erforderlich. In der Funktion $\mathrm{Update}(\cdot)$ wird schließlich bei erfolgreicher Zuordnung eines Perzepts zu einem multimodalen *Anchor* die perzeptuelle Signatur auf Basis des Fusionsmodells neu geschätzt.

Für das *Anchoring* von mehreren Objekten (siehe Abschnitt 5.2) wurden zwei weitere Unterfunktionen eingeführt. Die Funktion $\mathrm{Assign}(\cdot)$ realisiert innerhalb der Basisfunktion MULTIREACQUIRE (Seite 55) die konsistente Zuordnung von den zu einem Zeitpunkt durch ein Sensorsystem generierten Perzepten zu mehreren *Anchor*-Funktionen. Die Funktion $\mathrm{IsIrrelevant}(\cdot)$ entscheidet, wann ein *Anchor* für die geplante Anwendung nicht mehr relevant ist und von dem weiteren *Tracking*-Prozess ausgeschlossen werden kann. Die Gestaltung der Unterfunktionen für das multimodale *Anchoring* von Personen wird in den folgenden zwei Unterabschnitten genauer beschrieben.

### Unterfunktion $\mathrm{Assign}(\cdot)$

Beim *Anchoring* von mehreren Personen müssen die Perzepte $V_t$, die von einem Sensorsystem in einem Zeitpunkt $t$ generiert werden, optimal auf die aktuell verwendeten multimodalen *Anchor* $A$ verteilt werden. Jedes Perzept darf höchstens einem *Anchor* zugeordnet werden. Gleichzeitig kann es vorkommen, dass einem *Anchor* kein Perzept, sprich das Nullperzept $\perp$, zugewiesen wird. Die Aufgabe der Unterfunktion $\mathrm{Assign}(\cdot)$ ist es, die bezüglich eines vorgegebenen globalen Optimalitätskriteriums beste Kombination herauszusuchen.

Um das Problem formal zu beschreiben, werden zunächst zwei Abbildungen eingeführt. Die Abbildung $z : A \rightarrow (V_t \cup \perp)$ wird als Zuordnungskombination bezeichnet. Sie bildet jeden *Anchor* $\alpha \in A$ auf ein Perzept $\pi \in (V_t \cup \perp)$ ab. Eine Zuordnungskombination ist konsistent, wenn jedes Perzept höchstens einem *Anchor* zugewiesen wird:

$$\alpha_i \neq \alpha_j \;\Rightarrow\; z(\alpha_i) \neq z(\alpha_j) \;\vee\; z(\alpha_i) = z(\alpha_j) = \perp$$

Als Bewertung wird die Abbildung $b : A \times (V_t \cup \perp) \rightarrow [0 \ldots 1]$ bezeichnet. Sie weist der Zuordnung einer *Anchor*-Funktion $\alpha$ zu einem Perzept $\pi$ eine Zahl aus $[0 \ldots 1]$ zu und beschreibt damit, wie gut das Perzept zu dem *Anchor* passt. Beim multimodalen *Anchoring* von Personen basiert die Bewertung auf dem Abstand zwischen der Position des Perzepts und der durch das Bewegungsmodell vorausgesagten Position. Stimmen diese überein, dann erhält die Kombination die maximale Bewertung von 1. Mit zunehmendem Abstand sinkt die Bewertung linear, bis sie bei dem Abstand, bei dem das Kompositionsmodell verletzt wird, den Wert 0 erreicht. Kombinationen mit größerem Abstand werden ebenfalls mit 0 bewertet. Wird einem *Anchor* kein Perzept zugewiesen, dann ist die Bewertung ebenfalls minimal: $b(\alpha, \perp) = 0$

Das Ziel der $\mathrm{Assign}(\cdot)$-Funktion ist es nun, die Zuordnungskombination $z^\star$ zu finden, für die die Summe aller zugehörigen Bewertungen maximal ist:

$$z^\star = \operatorname*{argmax}_{z \in Z} \sum_{\alpha \in A} b(\alpha, z(\alpha))$$

Bei $a$ *Anchor*-Funktionen und $p$ Perzepten ergibt sich die Anzahl möglicher Zuordnungskombinationen $n = |Z|$ zu:

$$n = \sum_{i=0}^{\min(a,p)} \binom{i}{a} \frac{p!}{(p-i)!}$$

Diese steigt mit wachsender Anzahl von *Anchor*-Funktionen und Perzepten stark an. Daher ist es notwendig, ein effizientes Verfahren zur Bestimmung der optimalen Zuordnung zu nutzen. In [Fri03b] wird die Verwendung des A$^\star$-Algorithmus vorgeschlagen. Zu diesem Zweck werden Zuordnungskombinationen betrachtet, die auch teilweise undefiniert sein können, bei denen also die Abbildungen einzelner *Anchor* nicht definiert sind. Der A$^\star$-Algorithmus startet mit einer vollständig undefinierten Zuweisungsabbildung und erweitert die Definition iterativ auf alle *Anchor*. Die Auswahl der Zuordnung wird durch den A$^\star$-Algorithmus gegeben. Die Gesamtbewertung einer Zuordnungskombination ist die Summe der Bewertungen für die *Anchor*, auf denen die Abbildung definiert ist. Die optimale Schätzung der restlichen Bewertung ergibt sich unter der Annahme, dass alle folgenden Zuordnungen maximale Bewertung erhalten. In der Praxis zeigt sich jedoch, dass nur wenige Zuweisungen eine positive Bewertung erhalten, da in der Regel die Positionsabstände zwischen Perzepten und *Anchor*-Funktionen zu groß sind. Daher handelt es sich bei der optimistischen Schätzung um eine ungeeignete Wahl. Das führt dazu, dass die Effizienz des A$^\star$-Algorithmus nicht ausgenutzt werden kann.

Eine andere und effiziente Möglichkeit der Auswahl einer geeigneten Zuordnung ist die Verwendung eines *Greedy*-Algorithmus. Dazu wird die zunächst vollständig undefinierte Zuordnungsabbildung sukzessive um die Zuordnung erweitert, die unter den verbleibenden Zuordnungen die maximale Bewertung aufweist. Dieser Prozess wird, unter Berücksichtigung der Vorgaben für eine Zuordnungsabbildung, so lange wiederholt, bis die Abbildung vollständig definiert ist. Dieses Verfahren liefert nicht notwendigerweise die optimale Zuordnung, ist in der Praxis jedoch im Allgemeinen ausreichend.

#### Unterfunktion $\mathbf{IsIrrelevant}(\cdot)$

Die Funktion $\mathrm{IsIrrelevant}(\cdot)$ entscheidet für einen *Anchor* zu einem gegebenen Zeitpunkt, ob er aus der Menge der im *Anchoring*-Prozess verwendeten *Anchor* entfernt werden soll, da er für die Anwendung nicht mehr relevant ist. Da für die Aufmerksamkeitssteuerung des Roboters im Prinzip alle Personen berücksichtigt werden müssen, darf zunächst keiner der aufgebauten *Anchor* verworfen werden. Wenn eine zuvor beobachtete Person durch keinen der verwendeten Sensoren erfasst werden kann, weil zum Beispiel die Person den Raum verlassen hat, dann steigt die Unsicherheit bei der Schätzung der Konfiguration beziehungsweise der Position der Person ununterbrochen an. Wenn die Unsicherheit zu groß wird, macht es keinen Sinn, dem entsprechenden *Anchor* weiterhin Perzepte zuzuordnen, da die Gefahr einer falschen Zuordnung dann sehr groß ist. Ein multimodaler *Anchor* wird daher durch die Funktion $\mathrm{IsIrrelevant}(\cdot)$ genau dann aus dem *Anchoring*-Prozess entfernt, wenn die Unsicherheit in der Vorhersage der Position der entsprechenden Person einen Schwellwert überschreitet.

### 6.5.4 Behandlung spezieller Fälle in der realen Anwendung

Abschließend werden in diesem Abschnitt zwei Maßnahmen beschrieben, die in das multimodale *Anchoring* von Personen integriert werden, um Probleme mit falsch-positiven Detektionen und

schlechter Zuordnung von Perzepten zu verringern.

Die Rate der Fehlklassifikationen ist in einer realen Anwendung nicht Null. Sie ist aber deutlich kleiner als die Rate der korrekten Detektionen. Die Fehler können unterteilt werden in Falsch-Negative (keine Detektion für eine Person, obwohl sie sich im Messbereich des Sensors befindet) und Falsch-Positive (Detektion an einer Stelle, an der sich keine Person befindet). Falsch-Negative sind insofern unproblematisch, als dass sie die Zustandsschätzungen von Personen nicht verfälschen. Das Ausbleiben von Detektionen führt nur zu einer größeren Unsicherheit beim Verfolgen. Solange genug Detektionen von zum Beispiel anderen Modalitäten vorliegen, ist das Verfolgen von Personen im Allgemeinen möglich.

Die Auswirkung von Falsch-Positiven ist dagegen besonders zu betrachten. Bei Zuordnung von falsch-positiven Detektionen zu einer momentan verfolgten Person verschlechtert sich in der Regel die Schätzung des aktuellen Zustands. Der multimodale Ansatz kann dieses Problem jedoch in der Regel abfangen, da die Wahrscheinlichkeit einer gleichzeitigen Zuordnung von falsch-positiven Perzepten anderer Modalitäten gering ist. Werden Falsch-Positive keiner Person zugeordnet, dann führen sie allerdings zum Aufbau neuer Personenhypothesen beziehungsweise neuer multimodaler *Anchor*. Dieser Aspekt hat große Bedeutung für die personengerichtete Aufmerksamkeitssteuerung. In diesem Fall besteht die Gefahr, dass der Roboter seine Aufmerksamkeit in eine Richtung lenkt, in der sich keine Person befindet. Dies könnte zur Irritation anderer Personen führen. Um dieses Problem zu beheben, wird dem vorgestellten Verfahren zum Verfolgen von Personen eine weitere Komponente zur Beurteilung der Glaubwürdigkeit multimodaler *Anchor* hinzugefügt. Es wird dabei ausgenutzt, dass falschen Hypothesen im Vergleich zu richtigen Hypothesen wenig Perzepte zugeordnet werden, da die Rate von Fehlklassifikationen niedrig ist. Die Rate der zugeordneten Perzepte zu einem multimodalen *Anchor* ist ein Maß für die Vertrauenswürdigkeit. *Anchor* mit einer geringen Zuordnungsrate werden zwar weiter verfolgt (in der Regel werden sie aufgrund längeren Ausbleibens von Perzepten durch die Funktion $\mathrm{IsIrrelevant}(\cdot)$ wieder verworfen), aber den nachfolgenden Prozessen, zum Beispiel der Aufmerksamkeitssteuerung, nicht zur Verfügung gestellt.

Der Aufbau falscher Hypothesen erfolgt nicht nur durch falsch-positive Perzepte, sondern kann auch durch korrekte Detektionen ausgelöst werden. Denn durch zu große Ungenauigkeit der positionsbestimmenden Attribute eines Perzepts oder durch sehr schnelle Positionsveränderungen einer Person wird ein korrekt detektiertes Perzept dem entsprechenden multimodalen *Anchor* der Person aufgrund zu großen Abstands nicht zugeordnet. Dies führt zum Aufbau einer neuen Hypothese. Im ungünstigen Fall werden die nachfolgend generierten Perzepte auf beide Hypothesen verteilt, sodass für eine Person zwei multimodale *Anchor* aufgebaut wurden. In solchen Fällen ist der Abstand (in der horizontalen Ebene) zwischen den Personenhypothesen aber geringer, als sich die beobachteten, realen Personen einander nähern. Diese Tatsache wird ausgenutzt, um das Problem zu behandeln. Am Ende jeder Zuordnung von Perzepten werden alle momentan verwendeten *Anchor* paarweise betrachtet. Wenn sich der Abstand zweier Hypothesen unter einem Schwellwert befindet, wird der später aufgebaute *Anchor* wieder verworfen.

## 6.6 Zusammenfassung

In diesem Kapitel wurde die Umsetzung des Konzepts vom multimodalen *Anchoring* für das Verfolgen von Personen von einem mobilen Roboter aus beschrieben. Der gesamte Ansatz basiert auf vier unimodalen *Anchoring*-Prozessen, die durch die Daten von Kamera, Mikrofonen und Laser-Entfernungsmesser gespeist werden und daraus Bein-, Gesichts-, Oberkörper- und Stimmenperzepte generieren. Das Kompositionsmodell ist ein einfaches, rotationssymmetrisches Modell, bei dem die verfolgten Bestandteile auf einer vertikalen Achse übereinander angeordnet sind. Bewegungs- und Fusionsmodell realisieren zusammen ein Kalman-*Tracking*. Es wurden zusätzliche Überlegungen angestellt, um negative Auswirkungen durch falsch-positive Detektionen und falsche Zuordnungen von Perzepten zu den *Anchor*-Funktionen für die auf die Daten des *Anchoring*-Systems aufbauende Aufmerksamkeitssteuerung zu verringern.

Nicht alle Möglichkeiten, die das *Anchoring* bietet, sind in der vorgestellten Umsetzung ausgeschöpft worden. Insbesondere werden in den Symbolsystemen der unimodalen *Anchor* bisher keine Prädikate genutzt. Eine Möglichkeit zur Erweiterung des bestehenden Verfahrens ist es, Identifikation von Personen zum Beispiel über das Gesicht oder die Stimme in die Sensorsysteme zu integrieren. In diesem Fall können Prädikate genutzt werden, um eine identifizierte Person auf symbolischer Ebene als das entsprechende Individuum zu beschreiben. Die Integration der Identifikation könnte helfen, falsche Zuordnungen von Perzepten zu *Anchor*-Funktionen zu vermeiden. Darüber hinaus kann die Information über die Identität genutzt werden, um personenbezogene *Anchor* aufzubauen, in denen im Lauf der Zeit Wissen über die referenzierte Person gesammelt wird, das dann von anderen Prozessen abgerufen und verwertet werden kann.

Das hier realisierte Verfahren zum Verfolgen von Personen bildet die Grundlage für die multimodale Aufmerksamkeitssteuerung. Die Detektion von Gesichtern und die Lokalisation des Sprechers liefern dabei die grundlegenden Daten für das System. Bein- und Oberkörperperzepte dienen dagegen vorwiegend dazu, das Verfahren robuster zu gestalten. Da der Laser-Entfernungsmesser zum einen ein aktiver Sensor ist und damit von äußeren Einflüssen wie beispielsweise der Beleuchtungssituation unabhängig ist, und zum anderen mit seinem Einzugsbereich von 180° den gesamten Bereich vor dem Roboter abdeckt, steuert er die meisten Perzepte für das Verfolgen von Personen bei. Der große Einzugsbereich des Laser-Sensors kompensiert den relativ begrenzten Blickwinkel der Kamera und realisiert, zumindest für das Problem des Verfolgens von Personen, das periphere Sehen des Roboters. Dies wird in anderen Anwendungen häufig durch den Einsatz von omnidirektionalen Kameras erreicht (vgl. zum Beispiel [Doi01, Wil02, Shi04]). Das Verfolgen des Oberkörpers ist im Wesentlichen für die Situation gedacht, in der ein Benutzer den Roboter zu einem neuen Ort führt und dabei dem Roboter den Rücken zukehrt. In diesem Fall können das Gesicht und die Stimme nicht mehr detektiert werden, sodass zusätzliche Information durch den Oberkörper neben den Beinen die Robustheit erhöhen kann.

# Kapitel 7

# Das Aufmerksamkeitssystem

Das Verfahren zum Verfolgen von Personen aus dem vorangegangenen Kapitel realisiert die Fähigkeit des Roboters, Menschen multimodal wahrzunehmen. Diese Fähigkeit ist eine Voraussetzung für eine natürlich gestaltete Interaktion mit Menschen. Für ein effektives, zielgerichtetes Verhalten muss der Roboter darüber hinaus in der Lage sein, auch bei Anwesenheit mehrerer Personen innerhalb seines Wahrnehmungsbereichs, die für ihn relevante Person zu erkennen, seine Aufmerksamkeit auf diese zu richten und situationsabhängig zu halten. Die Aufmerksamkeitszuwendung versetzt den Roboter in die Lage, die für die Interaktion relevanten audio-visuellen Daten der fokussierten Person über seine Sensoren optimal zu erfassen. Außerdem stellt die Aufmerksamkeitszuwendung eine Signalwirkung für die beobachtende Person dar und liefert damit einen wichtigen Beitrag für eine natürliche Kommunikationssituation.

Dieses Kapitel stellt das in dieser Arbeit entwickelte Aufmerksamkeitssystem vor. Zunächst wird in Abschnitt 7.1 die prinzipielle Vorgehensweise der Aufmerksamkeitssteuerung beschrieben, die aus den zwei Schritten Selektion einer Person und Fokussierung der Aufmerksamkeit besteht. Ein wichtiger Aspekt der Fokussierung ist die Ausrichtung der Sensoren auf die selektierte Person. In Abschnitt 7.2 wird diskutiert, wie die Sensoren des Roboters im Allgemeinen auszurichten sind, um sowohl die Wahrnehmung zu optimieren als auch eine eindeutige und intuitiv verständliche Rückmeldung für die beobachtenden Personen zu erreichen. Die darauf folgenden drei Abschnitte beschreiben das Aufmerksamkeitsverhalten des Roboters. Es wird dabei die Bereitschaftsphase (Abschnitt 7.3), in der der Roboter reaktiv auf Sprechaktivität von umstehenden Personen reagiert, um einen neuen Kommunikationspartner zu finden, von der Interaktionsphase (Abschnitt 7.4), in der sich die Aufmerksamkeit alleine auf den Benutzer richtet, unterschieden. Die beiden Verhaltensweisen werden in Abschnitt 7.5 zu einer gesamten Aufmerksamkeitssteuerung zusammengeführt. Während die Bereitschaftsphase unabhängig von einer konkreten Anwendung ist, wurde die Interaktionsphase für den Einsatz des Roboters im *Home-Tour*-Szenario modelliert. Für diese Anwendung werden neben der Aufmerksamkeitssteuerung andere Komponenten wie Dialogsteuerung oder Gestenerkennung benötigt. Abschnitt 7.6 beschäftigt sich mit der Intergration aller Komponenten in eine Gesamt-Softwarearchitektur für den Roboter. Im Anschluss wird in Kapitel 7.7 die auditive Aufmerksamkeitssteuerung dargestellt. Sie dient dazu, die auditive Aufmerksamkeit des Roboters auf die selektierte Person zu fokussieren

und aktiviert darüber hinaus die Sprachverarbeitung, immer dann, wenn die Person zum Roboter spricht. Abschließend wird in Abschnitt 7.8 beschrieben, wie durch Anzeige eines Gesichts auf dem Flachbildschirm des Roboters die Rückmeldung für die beobachteten Personen unterstützt werden kann. Das Kapitel schließt mit einer Zusammenfassung.

## 7.1 Prinzipielle Vorgehensweise

Die Aufmerksamkeit des Roboters wird durch seine Wahrnehmung bedingt. Die Wahrnehmung ist in diesem Fall durch die multimodale Personenverfolgung gegeben. Da die beobachtete Szene ständigen Veränderungen unterworfen ist, handelt es sich bei der Aufmerksamkeitssteuerung um einen dynamischen Prozess. Mit jeder Aktualisierung der Personenverfolgung durch Messungen der Sensorsysteme führt die Aufmerksamkeitssteuerung folgende Schritte durch:

**Selektion:** Die Aufmerksamkeitssteuerung selektiert aus der Menge der beobachteten Personen diejenige, die für den Roboter die höchste Relevanz hat. Die Bestimmung der Relevanz hängt von der augenblicklichen Situation ab. Es lassen sich zwei Situationen voneinander abgrenzen: die Interaktionsphase und die Bereitschaftsphase. Während der Interaktionsphase hat der Roboter einen festen Kommunikationspartner. Dieser hat die höchste Relevanz. Die Aufmerksamkeit wird allein ihm gewidmet, während andere Personen, die eventuell auch anwesend sein könnten, ignoriert werden. Hat der Roboter dagegen keinen Kommunikationspartner, befindet er sich in der Bereitschaftsphase. Das Ziel des Roboters ist es dann, wieder in die Interaktionsphase zu wechseln. Es wird angenommen, dass der Benutzer den Interaktionsaufbau initiiert, indem er den Roboter anspricht. Um schnell neue Kommunikationspartner erkennen zu können, selektiert die Aufmerksamkeitssteuerung vornehmlich Personen, die entsprechende Kennzeichen, wie „Sprechen“ oder „zum Roboter Schauen“ aufweisen. Das heißt, in der Bereitschaftsphase wird die Selektion durch äußere Reize gesteuert. Dadurch wird eine reaktive Aufmerksamkeit realisiert, die als bottom-up gesteuerte Aufmerksamkeit bezeichnet wird. Im Gegensatz dazu ist in der Interaktionsphase nur der Benutzer selektiert. Äußere Reize („Sprechen“ und „Schauen“) anderer Personen werden ignoriert und führen nicht zu einer Neuselektion. In diesem Fall wird eine willentlich gesteuerte Aufmerksamkeit realisiert, die als top-down gesteuerte Aufmerksamkeit bezeichnet wird.

**Fokussierung:** Die Aufmerksamkeitssteuerung fokussiert die Aufmerksamkeit des Roboters auf die selektierte Person. Dies geschieht durch aktives Ausrichten der Sensoren. Dieser Vorgang dient dazu, die Wahrnehmung des Roboters zu optimieren und ermöglicht es zum Beispiel, die selektierte Person in das Blickfeld der Kamera zu bekommen. Das Ausrichten der Sensoren stellt darüber hinaus den Aufmerksamkeitsfokus des Roboters transparent für die beobachtenden Personen dar. Es signalisiert, worauf der Roboter seine Aufmerksamkeit gerichtet hält. Dieser Effekt trägt dazu bei, die Natürlichkeit des Verhaltens des Roboters in der Interaktion zu unterstützen und muss folglich bei der Ausrichtung der Sensoren explizit mit berücksichtigt werden.

Neben dem nach außen hin sichtbaren Vorgang des Ausrichtens der Sensoren findet auch eine verdeckte Verschiebung der Aufmerksamkeit statt. Dieser Aufmerksamkeitswechsel betrifft die Verarbeitung von Sprache. Die Aufmerksamkeitssteuerung richtet die auditive Aufmerksamkeit auf die selektierte Person und verarbeitet dadurch nur sprachliche Äußerungen dieser Person. Zudem erkennt die Aufmerksamkeitssteuerung, wann eine Äußerung an den Roboter gerichtet ist und entscheidet demnach, welche Äußerung zu interpretieren ist.

## 7.2 Ausrichten der Sensoren

Eine wesentliche Aufgabe der Aufmerksamkeitssteuerung ist das Ausrichten der Sensoren des Roboters auf Personen. Die folgenden Aspekte spielen dabei eine Rolle:

- Jeder Sensor erfasst in der Regel nur einen begrenzten räumlichen Bereich. Darüber hinaus hängt die Erkennungsleistung davon ab, wo sich die zu erkennenden Objekte innerhalb des Messbereichs des jeweiligen Sensors befinden.
- Die Stellung der Sensoren zeigt den umstehenden Menschen den Fokus der Aufmerksamkeit des Roboters an.

Das Ausrichten der Sensoren dient folglich sowohl der Optimierung der Wahrnehmung des Roboters als auch einer intuitiv verständlichen Rückmeldung für die beobachtenden Personen. In diesem Abschnitt wird diskutiert, wie die Sensoren unter Berücksichtigung der beiden Aspekte im allgemeinen Fall auszurichten sind.

Die Anforderungen an die Ausrichtung der Sensoren hängen direkt von der verwendeten Hardware ab. Die folgenden Überlegungen beziehen sich auf den Roboter *BIRON* (siehe Abschnitt 3.1). Dieser ist mit einer Kamera, zwei Mikrofonen und einem Laser-Entfernungsmesser ausgestattet. Die Sensoren sind auf der Roboterbasis montiert. Sei können folglich durch Bewegung der Roboterbasis (Rotation und Translation) ausgerichtet werden. Die Kamera verfügt darüber hinaus über eine eigenständige Schwenk-Neige-Einheit und kann daher separat bewegt werden.

Im Folgenden werden zunächst für jeden Sensor die individuellen Einzugsbereiche angegeben und die jeweiligen Erkennungsleistungen in Abhängigkeit von der Position der zu erkennenden Objekte erläutert. Aus diesen Fakten werden die Anforderungen an die Ausrichtung der Sensoren für die Interaktion abgeleitet.

**Kamera:** Die Kamera von *BIRON* besitzt eine Zoom-Funktion. Um einen möglichst großen Bereich zu erfassen, wird sie im Weitwinkelmodus betrieben. Der horizontale (vertikale) Öffnungswinkel beträgt dabei 48,8° (37,6°).

Die Kamera wird von der Personenverfolgung zur Detektion von Gesichtern und zur Lokalisation des Oberkörpers eingesetzt. Für die Interaktion ist zusätzlich die Erkennung

von Gesten des Benutzers geplant. Voraussetzung für die jeweiligen Detektionen ist, dass das Gesicht oder die Hände vollständig, beziehungsweise der Oberkörper in ausreichender Größe im Kamerabild abgebildet werden. Da der Einzugsbereich der Kamera relativ klein ausfällt, sollte die Kamera immer möglichst direkt auf die selektierte Person ausgerichtet werden.

Die Detektion von Gesichtern ist für Personen in einem Bereich von ca. 0,23 m bis 2,5 m Entfernung zum Roboter möglich. Für die Initialisierung der Lokalisation von Kleidung muss ein ellipsenförmiger Ausschnitt, dessen Position sich an dem Gesicht der entsprechenden Person orientiert, vollständig im Bild liegen (siehe Abschnitt 6.3.1). Zentriert die Kamera das Gesicht, beträgt der resultierende Mindestabstand der Person 1,3 m. Um neben dem Gesicht auch die Hände im Blickfeld der Kamera zu halten, muss der Abstand mindestens 1,2 m betragen.[1] Für die vollständige Erfassung der selektierten Person muss folglich ein Mindestabstand von 1,3 m eingehalten werden.

**Mikrofone:** Die Mikrofone erfassen den gesamten Bereich vor dem Roboter. Sie werden von der Personenverfolgung zur Lokalisation von Sprechern eingesetzt. Während der Interaktion liefern sie ferner das Sprachsignal für die Sprachverarbeitung.

Die Genauigkeit der Lokalisation hängt vom relativen Winkel und Abstand des Sprechers zum Roboter ab (siehe Abschnitt 6.4.1). Sie verschlechtert sich mit jeweils zunehmendem Winkel und Abstand. Für die Personenverfolgung ist die Genauigkeit jedoch im gesamten Bereich ausreichend. Die Wortfehlerrate der Spracherkennung hängt in analoger Weise von der Position des Sprechers ab und verschlechtert sich mit zunehmendem Winkel und Abstand des Sprechers [Fin04]. Damit die Sprachverarbeitung brauchbare Ergebnisse liefert, sollte der Winkel nicht größer als 30° sein. Die Mikrofone sollten folglich möglichst direkt auf die selektierte Person ausgerichtet werden. Der darüber hinaus zulässige Winkelbereich von $\pm 30°$ gibt dabei eine gewisse Flexibilität, die dazu genutzt werden kann, bei Anwesenheit mehrerer Personen möglichst viele im Einzugsbereich der Sensoren zu halten.

**Laser-Entfernungsmesser:** Der Laser-Entfernungsmesser erfasst Objekte im vorderen Bereich des Roboters auf einer Höhe von 0,3 m. Die Messgenauigkeit ist innerhalb des gesamten Bereichs konstant. Jeder Messdatensatz besteht aus 361 Einzelmessungen, wobei die Winkeldifferenz zweier aufeinander folgender Messrichtungen 0,5° beträgt. Je weiter ein Objekt folglich vom Laser entfernt ist, desto geringer ist die Anzahl der Messstrahlen, von denen es erfasst wird.

Die Personenverfolgung setzt den Laser zur Lokalisation von Beinen ein. Um ein Bein detektieren zu können, sollte es wenigstens von fünf Messstrahlen erfasst werden (siehe Abschnitt 6.1). Infolgedessen können Beine nur von Personen erkannt werden, die sich höchstens 3 m vom Roboter entfernt aufhalten.[2] Der Laser-Entfernungsmesser gibt somit

[1] Der vertikale Abstand der Hände zum Gesicht wird dabei als 0,8 m angenommen.
[2] Der Durchmesser eines Beins wird dabei als 0,13 m angenommen.

für die Ausrichtung der Sensoren keine Richtung, sondern lediglich einen Maximalabstand vor.

Für die Aufmerksamkeitssteuerung ergeben sich zusammengefasst folgende Vorgaben bezüglich der Ausrichtung der Sensoren: Die Kamera sollte immer direkt auf die selektierte Person ausgerichtet werden. Die Roboterbasis sollte so gedreht werden, dass möglichst alle anwesenden Personen im Einzugsbereich des Laser-Entfernungsmessers gehalten werden. Wenn die selektierte Person spricht, sollte der Winkel zu dieser Person möglichst gering und nicht größer als 30° sein. Um die Initialisierung der Lokalisation von Kleidung zu ermöglichen und in der Interaktion gleichzeitig den Kopf und die Hände im Blickfeld der Kamera halten zu können, muss der Abstand mindestens 1,3 m betragen.

Im Folgenden wird der zweite Aspekt diskutiert, der das Ausrichten der Sensoren als Mittel zur intuitiv verständlichen Rückmeldung für die umstehenden Personen betrifft. Es wird diskutiert, welche Signalwirkung die Bewegung und Stellung der Sensoren auf den Beobachter hat. Da die Kamera separat, die Mikrofone und der Laser-Entfernungsmesser dagegen nur in Verbindung mit der Roboterbasis bewegt werden können, werden Kamera und Roboterbasis getrennt voneinander betrachtet.

**Kamera:** Die Kamera fällt durch ihre hohe Position auf dem Roboter und ihre schnellen Bewegungen besonders ins Auge. Es ist zu erwarten, dass auch Laien die Kamera, zum Beispiel aufgrund ihrer Linse, als solche erkennen und den prinzipiellen Nutzen eines solchen Geräts einschätzen können. Die meisten nehmen daher an, dass sie dem Roboter zum Sehen dient. Die Kamera kann als „das Auge des Roboters“ interpretiert werden.

In der zwischenmenschlichen Kommunikation spielen die Augen beziehungsweise die Blickrichtung eine wichtige Rolle. Die Blickrichtung zeigt in der Regel den Fokus der visuellen Aufmerksamkeit an. Da der Bereich des fovealen Sehens beim Menschen mit ein bis zwei Grad sehr klein ist, befindet sich das Objekt der Aufmerksamkeit ziemlich genau in Blickrichtung. Wenn Menschen einander in der Kommunikation anschauen, sehen sie in der Regel auf die Augen oder auf den Mund des Gesprächspartners und signalisieren damit gleichzeitig, dass sie ihm ihre Aufmerksamkeit schenken. Kleine Abweichungen vom direkten Anschauen können Menschen mühelos erkennen und interpretieren diese als mangelnde Aufmerksamkeit.

Folglich muss auch beim Roboter die entsprechende Signalwirkung der Stellung der Kamera im Besonderen berücksichtigt werden. Die Kamera sollte direkt auf das Gesicht der von der Aufmerksamkeitssteuerung selektierten Person zentriert werden. Der Roboter zeigt dadurch dem Benutzer in eindeutiger Weise seine Aufmerksamkeit.

**Roboterbasis:** Der symmetrische Aufbau und die Positionen von Flachbildschirm und Laser-Entfernungsmesser ordnen der Roboterbasis eindeutig eine vordere Seite zu. Die Roboterbasis kann als „der Körper des Roboters“ interpretiert werden. Die Stellung des Körpers zeigt beim Menschen ebenfalls die Richtung der Aufmerksamkeit an, wenn auch nicht so

eindeutig wie dies bei der Blickrichtung der Fall ist. Die Roboterbasis sollte sich daher tendenziell auch auf die selektierte Person ausrichten. Neben der momentanen Stellung muss die Signalwirkung der Drehbewegung berücksichtigt werden. Je nachdem, ob dabei der relative Winkel einer Person zum Roboter kleiner oder größer wird, kann die Bewegung als Zuwenden oder Abwenden der Aufmerksamkeit interpretiert werden. Ein widersprüchliches Verhalten des Roboters, das aus dem Ausrichten der Kamera auf eine Person bei gleichzeitigem Wegdrehen der Roboterbasis resultiert, sollte vermieden werden, da dies sonst zu Irritationen bei den beobachtenden Personen führen kann.

Neben der Rotation des Roboters spielt auch die Translation und damit der Abstand zu Personen eine Rolle. Der Roboter darf nicht zu nah an den Benutzer heranfahren. Ein zu geringer Abstand stört die Kommunikation, da sich der Gesprächspartner bedrängt fühlen könnte. Der Roboter darf sich auch nicht unaufgefordert schnell auf eine Person zu bewegen, da dies bedrohlich auf den Benutzer wirkt. Folglich sollte sich der Roboter weitgehend passiv verhalten. Darüber hinaus wird der Roboter aus Gründen der Sicherheit nicht rückwärts fahren, da er über keine Sensoren verfügt, die den Bereich hinter dem Roboter abdecken.

Die Berücksichtigung der Signalwirkung durch die Stellung der Sensoren führt weitestgehend zu den gleichen Anforderungen an die Ausrichtung der Sensoren, die sich zuvor auch für die Optimierung der Wahrnehmung ergeben haben.

Zusammengefasst sollte die Kamera, wenn möglich, auf das Gesicht der selektierten Person gerichtet werden. Um in der Interaktion auch die Hände zu erfassen, ist es allerdings zeitweilig erforderlich, die Kamera etwas zu senken. Ein der Interaktionssituation angepasstes Verhalten des Roboters wird im Abschnitt 7.4 über die top-down gesteuerte Aufmerksamkeit beschrieben. Die Roboterbasis sollte sich in Richtung der selektierten Person drehen, dabei aber andere Personen nicht aus dem Wahrnehmungsbereich verlieren. Wenn die selektierte Person spricht, muss der Winkel weniger als 30° betragen, um die Wortfehlerrate der Spracherkennung niedrig zu halten. Kamera und Roboterbasis dürfen sich, der Signalwirkung wegen, nicht in entgegengesetzter Richtung bewegen.

## 7.3 Bottom-up gesteuerte Aufmerksamkeit

Nachdem im vorangegangenen Abschnitt die Ausrichtung der Sensoren für den allgemeinen Fall bestimmt wurde, wird in den folgenden Abschnitten das genaue Verhalten des Roboters in verschiedenen sich ergebenden Situationen festgelegt. Zunächst wird die Bereitschaftsphase betrachtet, in der der Roboter nach neuen Kommunikationspartnern Ausschau hält. Es wird angenommen, dass eine neue Interaktion durch den Benutzer aufgebaut wird, indem dieser den Roboter anspricht. Den Akt des Ansprechens erkennt der Roboter daran, dass die betreffende Person zur selben Zeit spricht und den Roboter anschaut. Um schnell auf einen Kommunikationswunsch reagieren zu können, muss die Aufmerksamkeit des Roboters bevorzugt auf Personen gerichtet werden, die sprechen oder schauen. Sprechaktivität und Blick zum Roboter lassen sich aus der

multimodalen Personenverfolgung extrahieren. Dies sind die Stimuli, die den Fokus der Aufmerksamkeit in der Bereitschaftsphase lenken. Personen sind damit in der Lage, durch Sprechen und Schauen die Aufmerksamkeit des Roboters auf sich zu lenken. Der Roboter wendet keinen Mechanismus an, um bestimmten Personen dauerhaft mehr oder weniger Aufmerksamkeit zu widmen. Es ist keine willentliche Steuerung der Aufmerksamkeit beteiligt. In der Bereitschaftsphase ist folglich eine rein bottom-up gesteuerte Aufmerksamkeit aktiv.

Im folgenden Unterabschnitt wird spezifiziert, wie die relevanten Stimuli aus der multimodalen Personenverfolgung extrahiert werden. Anschließend wird der Selektionsmechanismus vorgestellt, der auf Basis der vorliegenden Stimuli eine Person für den Fokus der Aufmerksamkeit auswählt. Abschließend wird das situationsabhängige, aktive Verhalten des Roboters, sprich die Ausrichtung von Roboterbasis und Kamera, beschrieben.

### 7.3.1 Relevante Stimuli

Die bottom-up gesteuerte Aufmerksamkeit ermöglicht es dem Roboter, seine Aufmerksamkeit auf potenzielle Kommunikationspartner zu richten. Die für die Steuerung relevanten Stimuli sind Sprechaktivität und Blick zum Roboter der verfolgten Personen. Für jede Person können die Eigenschaften `spricht` und `schaut` aus dem Verfahren zum multimodalen Anchoring von Personen (Abschnitt 6.5) ermittelt werden. Je nachdem, ob einem multimodalen Anchor, der eine bestimmte Person repräsentiert, Stimmen- oder Gesichtsperzepte zugeordnet werden konnten, hat die Person zum entsprechenden Zeitpunkt gesprochen oder zum Roboter geschaut.

Während die Zuordnung (Nichtzuordnung) eines Stimmenperzepts direkt darauf schließen lässt, dass eine Person gesprochen (nicht gesprochen) hat, ist die Schlussfolgerung bei Gesichtsperzepten nicht so eindeutig. Die Zuordnung eines Gesichtsperzepts bedeutet, dass das Gesicht der betreffenden Person detektiert worden ist. Damit wird allerdings keine Aussage über die Blickrichtung getroffen. Da die Ausrichtung des Kopfs und die Blickrichtung jedoch häufig übereinstimmen, wird vereinfachend angenommen, dass die Zuordnung eines Gesichtsperzepts das Schauen der Person zum Roboter anzeigt. Aus der Nichtzuordnung eines Gesichtsperzepts kann wiederum nicht direkt geschlossen werden, dass die betreffende Person nicht zum Roboter geschaut hat. Nur wenn sich die Person zum Zeitpunkt der Messung auch im Einzugsbereich der Kamera befand, ist die Schlussfolgerung zulässig. Im anderen Fall kann keine Entscheidung bezüglich der Eigenschaft `schaut` getroffen werden.

Die Bestimmung der relevanten Stimuli anhand einzelner Zuordnungen von Perzepten ist fehleranfällig, da Fehler in der Perzeptgenerierung oder der Zuordnungen direkt zu falschen Ergebnissen für die Eigenschaften `spricht` und `schaut` führen. Um die Auswirkung einzelner Fehler zu reduzieren, wird nicht nur der letzte Zeitschritt berücksichtigt, zu dem das jeweilige Sensorsystem gearbeitet hat. Stattdessen werden alle potenziellen Zuordnungen innerhalb eines Zeitfensters der Dauer $\Delta t$ betrachtet. Abbildung 7.1 veranschaulicht das Vorgehen an einem Beispiel. Zur Bestimmung der Eigenschaften `spricht` und `schaut` für eine einzelne Person werden innerhalb des Zeitfensters die Anzahl der dieser Person zugeordneten Perzepte $n^+$ und

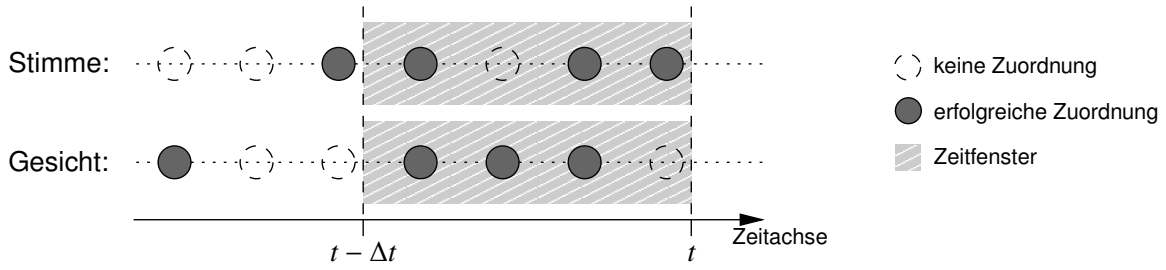



Abbildung 7.1: Beispiel für die Berechnung der Eigenschaften `spricht` und `schaut`. In beiden Modalitäten ist innerhalb des betrachteten Zeitfensters die Anzahl der zugeordneten Perzepte größer als die Anzahl nicht zugeordneter Perzepte. Beide Eigenschaften treffen folglich zu.

die Anzahl der nicht zugeordneten Perzepte $n^-$ gezählt. Es wird festgelegt, dass die untersuchte Eigenschaft genau dann zutrifft, wenn $n^+ \geq n^-$.

### 7.3.2 Selektionsmechanismus

Die Steuerung der Aufmerksamkeit in der Bereitschaftsphase soll den Roboter befähigen, potenzielle Kommunikationspartner zu erkennen. Um dieses Ziel effizient zu erreichen, darf der Fokus der Aufmerksamkeit nicht wahllos auf Personen gerichtet werden, sondern sollte bevorzugt auf solche gerichtet werden, die die Merkmale eines Kommunikationspartners (`spricht` und `schaut`) aufweisen. Der zugehörige Selektionsmechanismus berechnet zu jedem Zeitpunkt $t$ für jede Person $p$ einen so genannten Relevanzwert $\mathrm{Relevanz}_t(p)$. Je größer der Wert, desto höher die Relevanz der Person für den Roboter, und desto eher wird die Person selektiert. Der Relevanzwert fasst die beiden berücksichtigten Stimuli in einer Zahl zusammen. Dazu werden die Funktionen $\mathrm{spricht}(\cdot)$ und $\mathrm{schaut}(\cdot)$ eingeführt. Die Funktionen nehmen den Wert $1$ ($0$) an, wenn die entsprechende Eigenschaft für die Person zutrifft (nicht zutrifft). Der Relevanzwert einer Person berechnet sich als gewichtete Summe der Funktionswerte:

$$\mathrm{Relevanz}_t(p) = c_{spricht}\mathrm{spricht}_t(p) + c_{schaut}\mathrm{schaut}_t(p) \tag{7.1}$$

Der Selektionsmechanismus wählt zum Zeitpunkt $t$ aus der Menge aller verfolgten Personen $P_t$ immer die Person $p_t^\star$ aus, die den höchsten Relevanzwert aufweist:

$$p_t^\star = \operatorname*{argmax}_{p \in P_t} \mathrm{Relevanz}_t(p)$$

Eine geeignete Wahl der Gewichte $c_{spricht}$ und $c_{schaut}$ ergibt sich aus der folgenden Überlegung: Während die Eigenschaft `spricht` für alle verfolgten Personen zu jedem Zeitpunkt ermittelt werden kann, ist die Bestimmung der Eigenschaft `schaut` nur für die Personen möglich, die

sich im Blickfeld der Kamera befinden. Der Wechsel der Aufmerksamkeit muss daher primär durch akustische Reize bestimmt werden. Durch die Fokussierung der Aufmerksamkeit auf einen Sprecher gelangt dieser ins Blickfeld der Kamera, so dass dann auch die Eigenschaft `schaut` bestimmt und somit diese Person als ein potenzieller Kommunikationspartner erkannt werden kann. Damit die Aufmerksamkeit primär durch akustische Reize gelenkt wird, muss die Eigenschaft `spricht` den höheren Einfluss auf den Relevanzwert bekommen. Es gilt daher:

$$c_{spricht} > c_{schaut}$$

Der Relevanzwert einer Person kann demzufolge vier verschiedene Werte annehmen. Personen, für die der höchstmögliche Relevanzwert ermittelt wurde, sprechen und schauen zum Roboter und werden als potenzielle Kommunikationspartner angesehen.

### Modifikation des Selektionsmechanismus

Die Selektion von Personen anhand von Relevanzwerten nach Formel 7.1 beschreibt in bestimmten Situationen nicht immer die geeignete Vorgehensweise. Im Folgenden werden Spezialfälle und die jeweils notwendigen Modifikationen des Selektionsmechanismus behandelt.

- Aufgrund der geringen Anzahl verschiedener Relevanzwerte kommt es häufig vor, dass zu einem Zeitpunkt $t$ mehrere Personen den aktuell höchsten Wert aufweisen. In diesem Fall wird für die Auswahl der Person eine ähnliche Strategie wie in [Was99] verwendet. Es wird die Person selektiert, die für die längste Zeit nicht mehr im Fokus der Aufmerksamkeit war. Dieses Vorgehen sorgt dafür, dass die Informationen über die verfolgten Personen regelmäßig aktualisiert werden können.

- Die Eigenschaften von Personen werden aus dem multimodalen Anchoring extrahiert. Mit jeder Messung eines der verwendeten Sensorsysteme werden neue Perzepte generiert, wodurch sich die Eigenschaften der Personen ändern können. Da die Taktfrequenz, mit denen Sprecherlokalisation und Gesichtsdetektion arbeiten, zusammen bei mehr als 15 Hz liegt, besteht die Gefahr, dass der Fokus der Aufmerksamkeit sehr schnell zwischen Personen hin und her wechselt. Dies hätte die folgenden Nachteile:

  – Schnelle Wechsel des Aufmerksamkeitsfokus erfordern ständige Neuausrichtungen der Sensoren und lassen den Roboter hektisch wirken.

  – Wenn der Aufmerksamkeitsfokus zu kurz auf einer Person verharrt, kann die Eigenschaft `schaut` nicht erkannt werden (siehe Abschnitt 7.3.1).

  Das Vorgehen des Selektionsmechanismus wird daher wie folgt abgewandelt: Wenn eine Person neu ausgewählt wird, bleibt der Fokus der Aufmerksamkeit für eine festgelegte Mindestzeit $d_{min}$ auf dieser Person, selbst wenn zwischenzeitig eine andere Person einen höheren Relevanzwert erlangt.

- Solange sich die Relevanzwerte aller Personen nicht ändern, und genau eine Person den aktuell höchsten Wert aufweist und damit ausgewählt ist, verharrt der Fokus der Aufmerksamkeit auf dieser Person. Dieser Fall tritt zum Beispiel ein, wenn mehrere Personen um den Roboter herumstehen und ihn anschauen. Nur der ausgewählten Person kann die Eigenschaft `schaut` zugewiesen werden, da sie sich im Blickfeld der Kamera befindet. Für die Person im Fokus der Aufmerksamkeit ergibt sich ein Relevanzwert von $c_{schaut}$, während die anderen Relevanzwerte Null sind. Solange die Person nicht gleichzeitig spricht und dadurch als Kommunikationspartner höchste Relevanz für den Roboter hätte, ist es sinnvoll, die Aufmerksamkeit trotz höchstem Relevanzwert nach einer gewissen Zeit zu lösen und auf eine andere Person zu wechseln.

  Zu diesem Zweck wird der Effekt der Habituation in den Selektionsmechanismus integriert. Die Habituation setzt den Relevanzwert einer Person, die für eine gewisse Zeitspanne $d_{sel}$ im Fokus der Aufmerksamkeit war, auf Null. Dieser Zustand hält für eine begrenzte Zeitspanne $d_{hab}$ an. Danach wird die Herabsetzung des Relevanzwerts für die entsprechende Person wieder aufgehoben. Die Habituation wird durch eine Funktion $\mathrm{Habituation}_t(\cdot)$ realisiert, die den Wert $0$ ($1$) annimmt, wenn der Relevanzwert der betrachteten Person herabgesetzt (nicht herabgesetzt) werden soll. Die Berechnung des Relevanzwerts $\mathrm{Relevanz}_t(p)$ einer Person $p$ zum Zeitpunkt $t$ nach Formel 7.1 wird damit wie folgt erweitert:

  $$\mathrm{Relevanz}_t(p) = \mathrm{Habituation}_t(p)\,(c_{spricht}\mathrm{spricht}_t(p) + c_{schaut}\mathrm{schaut}_t(p)) \qquad (7.2)$$

  Da es das wesentliche Ziel der bottom-up gesteuerten Aufmerksamkeit ist, den Roboter zu befähigen, auf Wunsch einer Person in eine Kommunikation einzutreten, haben Personen, die den Roboter ansprechen, Vorrang gegenüber anderen Personen. Solange für eine Person die beiden Eigenschaften `spricht` und `schaut` zutreffen, muss sie im Fokus der Aufmerksamkeit bleiben. Während dieser Zeit wird daher die Habituation für den potenziellen Kommunikationspartner außer Kraft gesetzt.

Den gesamten Selektionsmechanismus beschreibt der in Abbildung 7.2 angegebene Algorithmus $\mathrm{SelectPerson}$. Für die vom Selektionsmechanismus verwendeten Zeitkonstanten haben sich folgende Werte aus dem praktischen Einsatz des Systems als geeignet erwiesen:

$$d_{min} = 1\mathrm{s}, \quad d_{sel} = 4\mathrm{s}, \quad d_{hab} = 8\mathrm{s}$$

Die Selektion einer Person hat die Ausrichtung der Sensoren zur Folge. Die Auswahl ist ein interner Prozess, der von außen nicht zu beobachten ist. Er entspricht einem verdeckten Aufmerksamkeitswechsel (*covert shift of attention*). Die Ausrichtung der Sensoren zeigt dagegen den offenen Wechsel der Aufmerksamkeit (*overt shift of attention*) an. Die Ausrichtung ist Thema des folgenden Abschnitts.

Gegeben seien die zuletzt selektierte Person $p_{selected}$ und die Menge der zum aktuellen Zeitpunkt $t$ verfolgten Personen $P_t$. Die Funktion $\mathrm{TimeOfSelection}(\cdot)$ gibt für jede Person den Zeitpunkt der jeweils letzten Selektion an.

Eine Neuselektion erfolgt frühestens nach einer Zeitspanne $d_{min} = 1s$:

(1) **if** $t - \mathrm{TimeOfSelection}(p_{selected}) < d_{min}$ **then**
(2) **return**
(3) **fi**

Wenn die zuletzt selektierte Person spricht und schaut, erfolgt keine Neuselektion:

(4) **if** $\mathrm{Relevanz}_t(p_{selected}) = c_{spricht} + c_{schaut}$ **then**
(5) $\mathrm{TimeOfSelection}(p_{selected}) \leftarrow t$
(6) **return**
(7) **fi**

Gegebenenfalls Habituation für die selektierte Person beginnen:

(8) **if** $t - \mathrm{TimeOfSelection}(p_{selected}) > d_{sel}$ **then**
(9) $\mathrm{Habituation}_{t'}(p_{selected}) \leftarrow 0$ für $t' \in [t, t + d_{hab}]$
(10) **fi**

Die Person mit dem höchsten Relevanzwert selektieren. Wenn mehrere Personen den momentan höchsten Relevanzwert haben, dann diejenige auswählen, die am längsten nicht selektiert war:

(11) $r_{max} \leftarrow \max_{p \in P_t} \mathrm{Relevanz}_t(p)$
(12) $P_{max} \leftarrow \{p \in P_t | \mathrm{Relevanz}_t(p) = r_{max}\}$
(13) $p^\star \leftarrow \mathrm{argmin}_{p \in P_{max}} \mathrm{TimeOfSelection}(p)$

Gegebenenfalls Selektion einer neuen Person:

(14) **if** $p^\star \neq p_{selected}$ **then**
(15) $p_{selected} \leftarrow p^\star$
(16) $\mathrm{TimeOfSelection}(p_{selected}) \leftarrow t$
(17) **fi**.

Abbildung 7.2: Algorithmus $\mathrm{SelectPerson}$ zum Selektieren einer Person.

### 7.3.3 Das Verhalten des Roboters

Nachdem eine Person selektiert wurde, richtet der Roboter seine Sensoren neu aus. In Abschnitt 7.2 wurde diskutiert, wie das Verhalten des Roboters im allgemeinen Fall auszusehen hat. Während die Kamera immer direkt auf die selektierte Person zu richten ist, bleiben für die Ansteuerung der Roboterbasis in Abhängigkeit der Sprechaktivität der selektierten Person Variationsmöglichkeiten. Offen ist darüber hinaus, wie vorgegangen werden soll, wenn gar keine Person anwesend ist. Das Verhalten des Roboters ist folglich von der gegebenen Situation ab-

hängig.

In der Bereitschaftsphase lassen sich vier charakteristische Situationen ausmachen:

- Es ist keine Person anwesend.
- Der Roboter hat Geräusche wahrgenommen, verfolgt aber im Moment keine Person.
- Mehrere Personen sind anwesend. Keine von ihnen spricht.
- Es gibt mindestens einen Sprecher.

Jede der genannten Situationen führt zu einem anderen Verhalten des Roboters. Die verschiedenen Verhaltensweisen werden in der Aufmerksamkeitssteuerung durch so genannte Aufmerksamkeitszustände spezifiziert. Das gesamte Verhaltensmuster des Roboters wird durch einen endlichen Automaten beschrieben, der vier Aufmerksamkeitszustände umfasst.

Die vier Aufmerksamkeitszustände werden mit *AS:Sleeping*[3], *AS:Awake*, *AS:Alert* und *AS:Listening* bezeichnet. Das Verhalten des Roboters in den Zuständen ist wie folgt:

***AS:Sleeping*:** Der Roboter befindet sich im Ruhezustand, da sich für längere Zeit keine Personen in seiner Nähe aufgehalten haben. Die Bildverarbeitung wird nicht benötigt und ausgeschaltet, um die begrenzten Ressourcen des Roboters zu schonen. Die Aufmerksamkeit des Roboters kann folglich nur noch durch akustische Reize erlangt werden. Der Laser wird weiterhin zur Detektion von Beinen und zum Verfolgen von Personen eingesetzt. Um das Nichtverwenden der Kamera zu demonstrieren, wird sie nach unten gesenkt und dann nicht weiter bewegt. Auch die Roboterbasis wird nicht bewegt. Der Roboter soll den Eindruck erwecken, dass er schläft. Der Schlafzustand *AS:Sleeping* ist zugleich der Startzustand des endlichen Automaten.

***AS:Awake*:** In diesem Zustand befindet sich der Roboter, wenn er aufgrund akustischer Reize neue Personen erwartet, obwohl momentan keine Personen verfolgt werden. Alle Sensorsysteme sind aktiv. Der Roboter ist bereit, schnell auf herankommende Personen zu reagieren. Die Roboterbasis und die Kamera werden nicht bewegt. Die Kamera ist nach vorne ausgerichtet.

***AS:Alert*:** Dies ist ein Zustand erhöhter Wachsamkeit, wenn sich Personen in der Nähe des Roboters aufhalten, von denen aber keine spricht. In diesem Fall ist noch keine Interaktionsabsicht der Personen zu erkennen. Daher verhält sich der Roboter weitgehend passiv, das heißt die Basis wird nicht bewegt. Die Kamera wird allerdings immer auf das Gesicht der selektierten Person ausgerichtet. Dies signalisiert ein neugieriges Umherschauen des Roboters und ermöglicht zugleich, Informationen über die anwesenden Personen zu erlangen, zum Beispiel deren Identität.

[3] Das Präfix *AS:* zeigt an, dass es sich um einen Zustand des endlichen Automaten der Aufmerksamkeitssteuerung handelt.

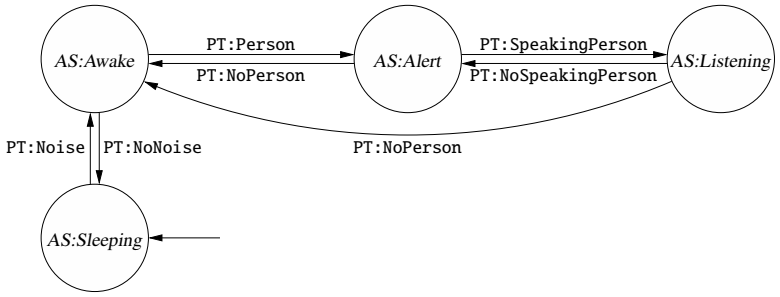



Abbildung 7.3: Endlicher Automat, der das Verhalten des Roboters im Bottom-up-Modus modelliert.

***AS:Listening*:** Diesen Zustand nimmt der Roboter ein, wenn er wenigstens einen Sprecher registriert. Die Kamera wird auf das Gesicht der selektierten Person ausgerichtet. Die Roboterbasis richtet sich durch Rotation so aus, dass möglichst viele der anwesenden Personen im Wahrnehmungsbereich des Lasers gehalten werden. Gleichzeitig wird beachtet, dass der relative Winkel zur selektierten Person (dem Sprecher) klein gehalten wird, um die Wortfehlerrate der Spracherkennung gering zu halten. Die Roboterbasis dreht sich nie von der selektierten Person weg, da dieses Verhalten als Abwendung der Aufmerksamkeit interpretiert werden könnte.

Die Wechsel zwischen den Aufmerksamkeitszuständen werden durch Ereignisse ausgelöst, die aus dem multimodalen Anchoring für Personen abgeleitet werden. Die vier Zustände und die im Folgenden beschriebenen Ereignisse und Zustandswechsel sind in der Abbildung 7.3 dargestellt.

Aus dem Schlafzustand *AS:Sleeping* kann der Roboter aufgrund der deaktivierten Bildverarbeitung nur durch akustische Reize geweckt werden. Der Wechsel zum Wachzustand *AS:Awake* erfolgt, wenn innerhalb einer gewissen Zeitspanne $d_{noise}$ ununterbrochen Stimmenperzepte durch das entsprechende Sensorsystem generiert werden (Ereignis `PT:Noise`[4]). Es ist dabei nicht erforderlich, dass gleichzeitig Personen verfolgt werden. Der Roboter wacht also durch Geräusche auf. Er wacht dagegen nicht auf, wenn sich Personen vor dem Roboter aufhalten, die nicht reden. Die Einschränkung auf akustische Reize lässt den Schlafzustand auf den Beobachter authentischer wirken. Sobald wenigstens eine Person erfasst wird (Ereignis `PT:Person`), wechselt der Roboter vom Wachzustand *AS:Awake* in den Zustand erhöhter Aufmerksamkeit *AS:Alert*. Wenn sich unter den Personen ein Sprecher befindet (Ereignis `PT:SpeakingPerson`), erfolgt der Übergang zum Aufmerksamkeitszustand *AS:Listening*. Die Folge der beschriebenen Wechsel führt zu einer stetigen Erhöhung der Aufmerksamkeit des Roboters, beginnend mit dem Schlafzustand und endend mit aktivem Zuhören.

In analoger Weise können die Aufmerksamkeitszustände in entgegengesetzter Richtung durchlaufen werden. Wenn für eine gewisse Zeitspanne $d_{quiet}$ keine der verfolgten Personen mehr spricht (Ereignis `PT:NoSpeakingPerson`), wechselt der Roboter vom Zustand *AS:Listening*

[4]Das Präfix *PT:* zeigt an, dass es sich um ein Ereignis handelt, das sich aus dem Personen-*Tracking* ergibt.

zum Zustand *AS:Alert*. Die Verzögerung des Wechsels garantiert, dass der Roboter auch während kurzer Sprechpausen im Zustand des aktiven Zuhörens bleibt und dadurch jederzeit die Mikrofone auf den aktuellen Sprecher ausrichtet. Aus dem Zustand *AS:Alert* erfolgt der Übergang zum Zustand *AS:Awake*, wenn sich keine Person mehr im Wahrnehmungsbereich des Roboters befindet (Ereignis `PT:NoPerson`). Dasselbe Ereignis kann auch im Zustand *AS:Listening* eintreten und führt ebenfalls zum direkten Übergang zum Wachzustand *AS:Awake*. Aus dem Wachzustand gelangt der Roboter schließlich wieder in den Schlafzustand *AS:Sleeping*, wenn über einen längeren Zeitraum $d_{silence}$ das Ereignis `PT:Noise` nicht auftritt, der Roboter also keine Geräusche wahrnimmt. Dieses Ereignis wird entsprechend mit `PT:NoNoise` bezeichnet.

Für die in den Ereignissen verwendeten Zeitkonstanten haben sich folgende Werte aus dem praktischen Einsatz des Systems als geeignet erwiesen:

$$d_{noise} = 1\text{s}, \quad d_{silence} = 1\text{s}, \quad d_{quiet} = 3\text{s}$$

## 7.4 Top-down gesteuerte Aufmerksamkeit

Während der Interaktionsphase benötigt der Roboter die Fähigkeit, seine Aufmerksamkeit durchgehend auf den Benutzer zu richten. Andere Personen werden ignoriert, selbst wenn diese nach dem Bewertungskriterium aus Abschnitt 7.3.2 einen höheren Relevanzwert aufweisen. In der Interaktionsphase wird die Aufmerksamkeit sozusagen willentlich auf dem Benutzer gehalten. In diesem Fall handelt es sich folglich um eine top-down gesteuerte Aufmerksamkeit.

Abhängig von der geplanten Anwendung und der jeweiligen Situation in der Interaktion mit dem Benutzer ist es notwendig, dass der Roboter auch in der top-down gesteuerten Aufmerksamkeit unterschiedliche Verhaltensweisen zeigt. In manchen Situationen ist es zum Beispiel notwendig, dass der Roboter den Benutzer von einer festen Position aus beobachtet, während es in anderen Situationen erwünscht ist, dass der Roboter dem Benutzer hinterherfährt. Die erforderlichen Verhaltensweisen können analog zur bottom-up gesteuerten Aufmerksamkeit durch entsprechende Aufmerksamkeitszustände realisiert werden. Während die bottom-up gesteuerte Aufmerksamkeit weitgehend unabhängig von der beabsichtigten Anwendung ist, müssen bei der Gestaltung der top-down gesteuerten Aufmerksamkeit die geplanten Fähigkeiten des Roboters berücksichtigt werden.

Die hier entwickelte Aufmerksamkeitssteuerung kommt auf dem Roboter *BIRON* zum Einsatz. *BIRON* soll mit Interaktionsfähigkeiten ausgestattet werden, die für das so genannte *Home-Tour*-Szenario benötigt werden. Ein solcher Roboter wird von Anwendern im privaten Haushalt eingesetzt. Neu gekauft und ausgepackt können dem Roboter in einem interaktiven Prozess Gegenstände und Orte gezeigt und beschrieben werden, die für spätere Interaktionen und Aufgabenstellungen relevant sind. Dem Benutzer soll es möglich sein, mit dem Roboter natürlichsprachlich zu kommunizieren. Um auf Objekte zu verweisen, können neben Sprache auch deiktische Gesten verwendet werden. Der Roboter erfasst benannte Objekte mit seiner Kamera und speichert neben der Ansicht des Objekts Informationen wie Position, Größe und zusätzliche sprachliche

Angaben des Benutzers. Um neue Objekte zu zeigen, ist es dem Benutzer möglich, den Roboter zu einer neuen Position im Haus zu führen.

Aus der geplanten Anwendung kristallisieren sich für die personenbasierte Aufmerksamkeit drei charakteristische Situationen heraus, in denen der Roboter unterschiedliche Verhaltensweisen zeigen muss:

- Der Roboter wartet auf neue sprachliche Anweisungen.
- Der Benutzer zeigt neue Objekte, wobei potenziell deiktische Gesten eingesetzt werden.
- Der Benutzer führt den Roboter an eine neue Position.

Die zweite Situation unterscheidet sich von der ersten im Wesentlichen dadurch, dass der Roboter zusätzlich auf Gesten achten muss. Dazu muss er die Kamera so ausrichten, dass neben dem Gesicht des Benutzers auch dessen Hände erfasst werden können.

Entsprechend der aufgezeigten Situationen werden die jeweils erforderlichen Verhaltensweisen des Roboters durch drei Aufmerksamkeitszustände realisiert. Die Zustände werden mit *AS:Person*, *AS:Show* und *AS:Follow* bezeichnet.

***AS:Person*:** Dies ist der Grundzustand der Interaktionsphase, in dem der Roboter neue Anweisungen des Benutzers erwartet. Die Roboterbasis richtet sich durch Rotation auf den Benutzer aus. Die Kamera ist auf das Gesicht fokussiert. Durch dieses Verhalten signalisiert der Roboter seine Bereitschaft und Aufmerksamkeit gegenüber dem Benutzer.

***AS:Show*:** Sobald sich aus der Kommunikationssituation erkennen lässt, dass der Benutzer auf Objekte durch deiktische Gesten verweisen wird, muss die Kamera etwas nach unten geneigt werden, um neben dem Gesicht auch die Hände in das Blickfeld der Kamera zu bekommen. Bis auf die veränderte Ausrichtung der Kamera ist das Verhalten dasselbe wie im Zustand *AS:Person*.

***AS:Follow*:** Dieser Zustand ist dafür gedacht, dass der Benutzer den Roboter durch Vorausgehen zu einer neuen Position führen kann. Zu diesem Zweck wird die Roboterbasis, wie in den beiden anderen Zuständen, auf den Benutzer ausgerichtet. Zusätzlich fährt der Roboter in Richtung des Benutzers und versucht einen optimalen Abstand einzuhalten. Wenn der Benutzer beginnt, sich vom Roboter zu entfernen, fängt der Roboter an, dem Benutzer hinterher zu fahren. Somit kann der Roboter komfortabel über längere Strecken zu einem anderen Ort geführt werden.

Wenn der Benutzer vorausgeht, wird er in der Regel in Gehrichtung schauen. Das Gesicht ist dann nicht mehr der Kamera zugewandt und kann nicht detektiert werden. In dieser Situation kann die Kamera nur noch zum Verfolgen des Oberkörpers vom Benutzer eingesetzt werden. Die Kamera fokussiert in diesem Zustand den Oberkörper, indem sie nach unten geneigt wird. Sie wird jedoch nur so weit geneigt, dass der Kopf des Benutzers noch erfasst wird und das Gesicht erneut detektiert werden kann.

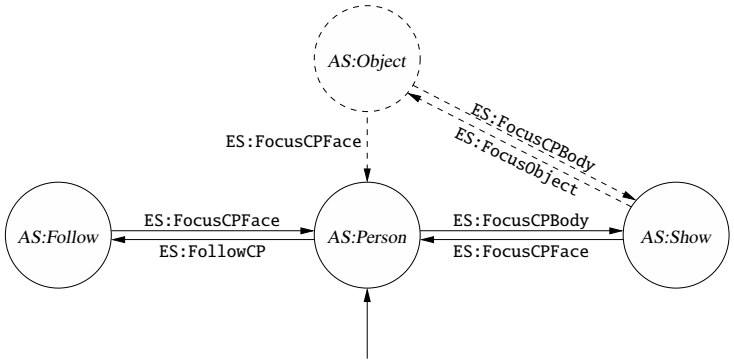



Abbildung 7.4: Endlicher Automat, der das Verhalten des Roboters im Top-down-Modus modelliert. Der Zustand *AS:Object* ist gestrichelt gezeichnet, da er eine Sonderrolle einnimmt: In diesem Zustand gibt die Personenaufmerksamkeit die Kontrolle über die Ansteuerung der Sensoren temporär an die Objektaufmerksamkeit ab.

Alle bisher beschriebenen Aufmerksamkeitszustände realisieren eine personenbasierte Aufmerksamkeitssteuerung. Der Fokus richtet sich ausschließlich auf Menschen. Während der Interaktion kann der Benutzer jedoch auch auf Objekte verweisen, zum Beispiel sprachlich oder durch deiktische Gesten. In dieser Situation werden für die Aufmerksamkeit des Roboters neben dem Benutzer auch Objekte relevant. Der Roboter muss dann zwischenzeitig den Fokus der Aufmerksamkeit von dem Benutzer abwenden und auf die verwiesenen Objekte richten. Dies geschieht vornehmlich durch das Ausrichten der Kamera. Die objektbasierte Aufmerksamkeitssteuerung geht über die Zielsetzung dieser Arbeit hinaus und wird deshalb hier nicht behandelt. Dennoch wird in der hier vorgestellten Aufmerksamkeitssteuerung die Möglichkeit zur Fokussierung von Objekten durch einen weiteren Aufmerksamkeitszustand *AS:Object* berücksichtigt. In diesem Zustand hat nicht mehr die personenbasierte Aufmerksamkeitssteuerung Kontrolle über die Ansteuerung von Roboterbasis und Kamera. Die Kontrolle wird temporär an eine entsprechende objektbasierte Aufmerksamkeitssteuerung übergeben. Die top-down gesteuerte Aufmerksamkeit umfasst damit die vier Zustände *AS:Person, AS:Show, AS:Follow* und *AS:Object.* Der zugehörige endliche Automat ist in Abbildung 7.4 dargestellt.

Die Wechsel zwischen den vier Aufmerksamkeitszuständen ergeben sich im Wesentlichen durch den sprachlichen Dialog mit dem Benutzer. Das Verhalten des Roboters kann durch Anweisungen, wie zum Beispiel *„Folge mir!“* (Wechsel zum Zustand *AS:Follow*) oder *„Ich zeige dir etwas!“* (Wechsel zum Zustand *AS:Show*), bestimmt werden. Die Wechsel zwischen Zuständen können aber auch aufgrund anderer Ereignisse erfolgen, die sich nicht aus dem sprachlichen Dialog ergeben. Wenn der Roboter sich zum Beispiel im Zustand *AS:Show* befindet und eine deiktische Geste erwartet, hängt der Zustandswechsel davon ab, ob eine Geste erkannt wird, oder nicht. Im Erfolgsfall wird der Zustand für die Objektaufmerksamkeit *AS:Object* eingenommen, im anderen Fall wechselt die Aufmerksamkeitssteuerung wieder zum Grundzustand *AS:Person.* Das heißt, Ereignisse, die in der top-down gesteuerten Aufmerksamkeit zu Zustandswechseln

führen, können durch unterschiedliche Komponenten ausgelöst werden.

Für die Anwendung von *BIRON* im *Home-Tour*-Szenario werden neben der Aufmerksamkeitssteuerung diverse andere Komponenten benötigt, wie Sprachverarbeitung, Dialogsteuerung, Gestenerkennung, objektbasierte Aufmerksamkeitssteuerung und ein Szenemodell, in dem Informationen über die gezeigten Objekte gespeichert werden (vgl. [Haa04]). Alle Komponenten werden in einer entsprechenden Software-Architektur miteinander verknüpft [Kle05]. Die ausführlichere Beschreibung geschieht im folgenden Abschnitt 7.6. Für die Architektur von *BIRON* wurde eine neue, zentrale Komponente, der so genannte *Execution Supervisor* (siehe Abschnitt 7.6.1), entwickelt, der sowohl den Datenaustausch zwischen Komponenten koordiniert als auch einzelne Komponenten konfiguriert. Der *Execution Supervisor* ist auch dafür zuständig, die Aufmerksamkeitssteuerung zu konfigurieren. Das bedeutet, die Ereignisse, die zu Wechseln von Aufmerksamkeitszuständen des Top-down-Modus führen, kommen nicht direkt von anderen Modulen, wie Dialogsteuerung oder Gestenerkennung, sondern werden vom *Execution Supervisor* an die Aufmerksamkeitssteuerung übermittelt. Dies ist Thema des Abschnitts 7.6.2.

## 7.5 Das Gesamtverhalten des Roboters

Das Verhalten des Roboters in der bottom-up und top-down gesteuerten Aufmerksamkeit wird jeweils durch endliche Automaten modelliert. Durch Kombination beider Automaten zu einem Gesamtautomaten kann das Verhalten des Roboters im ganzen Interaktionsszenario beschrieben werden. Durch das Zusammenführen der Aufmerksamkeitszustände ergeben sich weitere Transitionen (siehe Abbildung 7.5):

- Übergänge zwischen den Zuständen der bottom-up zur top-down gesteuerten Aufmerksamkeit entsprechen dem Vorgang, dass der Roboter aus der Beobachtung der anwesenden Personen ein Individuum als neuen Kommunikationspartner identifiziert hat. Es beginnt eine Phase der Interaktion mit dem ausgewählten Benutzer.
- Übergänge zwischen den Zuständen der top-down zur bottom-up gesteuerten Aufmerksamkeit kennzeichnen das Ende der Interaktion mit dem Benutzer. Der Roboter steht für einen neuen Kommunikationspartner bereit und beobachtet dazu wieder alle Personen in seiner Nähe.

Transitionen zwischen den Aufmerksamkeitszuständen beider Modi werden ebenfalls durch Ereignisse ausgelöst, die vom *Execution Supervisor* empfangen werden. Alle Ereignisse, welche die top-down gesteuerte Aufmerksamkeit betreffen, werden im folgenden Abschnitt spezifiziert.

## 7.6 Integration in die Software-Architektur des Roboters

Die auf dem Roboter *BIRON* für das *Home-Tour*-Szenario eingesetzten Software-Komponenten sind in einer Drei-Schichten-Architektur [Gat98] organisiert. Dabei handelt es sich um eine

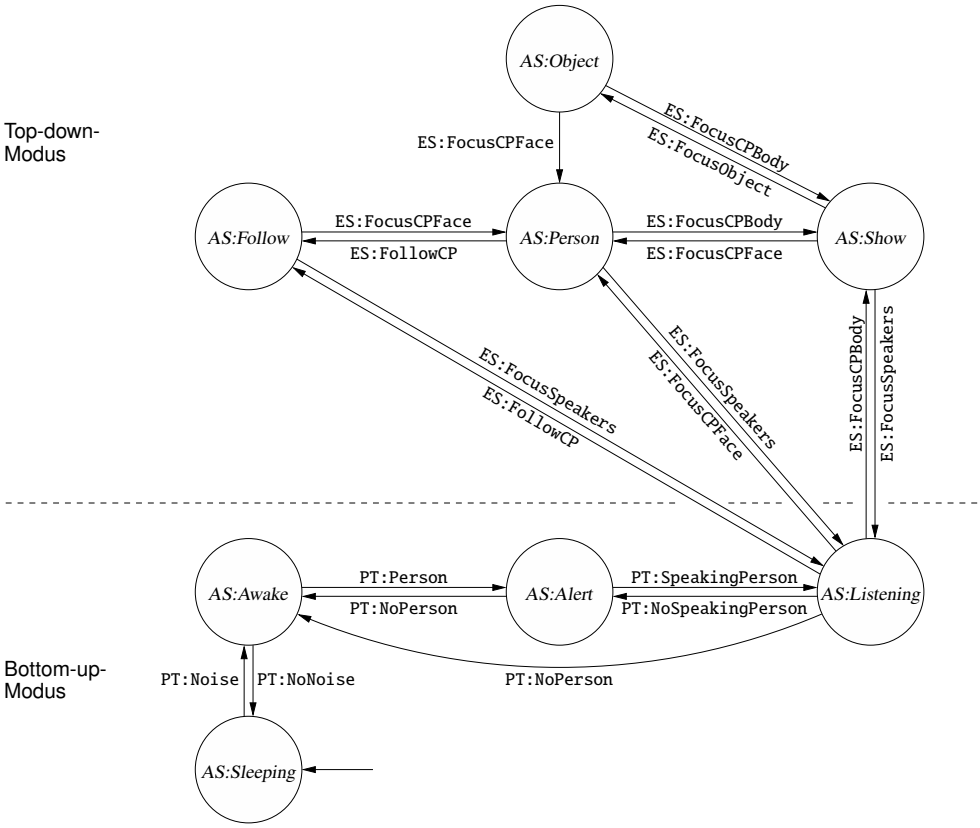



Abbildung 7.5: Endlicher Automat, der das Gesamtverhalten des Roboters modelliert. Transitionen, bei denen Zustände des Top-down-Modus beteiligt sind, werden durch Ereignisse vom *Execution Supervisor* ausgelöst (Präfix `ES:`). Die übrigen Transitionen werden durch Ereignisse ausgelöst, die sich aus der Personenverfolgung ableiten lassen (Präfix `PT:`).

hybride Architektur. Hybride Architekturen kombinieren die Vorteile reaktiver und deliberativer Steuerung. Module einer reaktiven Steuerung verarbeiten Sensordaten und berechnen daraus ein direktes Antwortverhalten des Roboters. Sie sind durch enge Sensor-Aktions-Schleifen charakterisiert. Die Module einer reaktiven Steuerung eignen sich besonders, um in dynamischen Umgebungen auf Veränderungen zu reagieren (Reflex, Reizreaktion). Sie arbeiten auf einer kleinen Zeitskala. Module einer deliberativen Steuerung dagegen realisieren die Planung in einem Robotersystem. Die typische deliberative Vorgehensweise besteht aus den drei Schritten Wahrnehmen–Planen–Ausführen [Nil82]. Da Planung zeitintensiv ist, arbeiten die Module einer deliberativen Steuerung auf einer größeren Zeitskala.

In einer hybriden Architektur arbeiten reaktive und deliberative Komponenten nicht vollständig unabhängig voneinander. Die deliberative Steuerung muss die reaktive Steuerung so konfigurieren, dass gestellte Aufgaben effizient gelöst werden. Die reaktive Steuerung wiederum muss

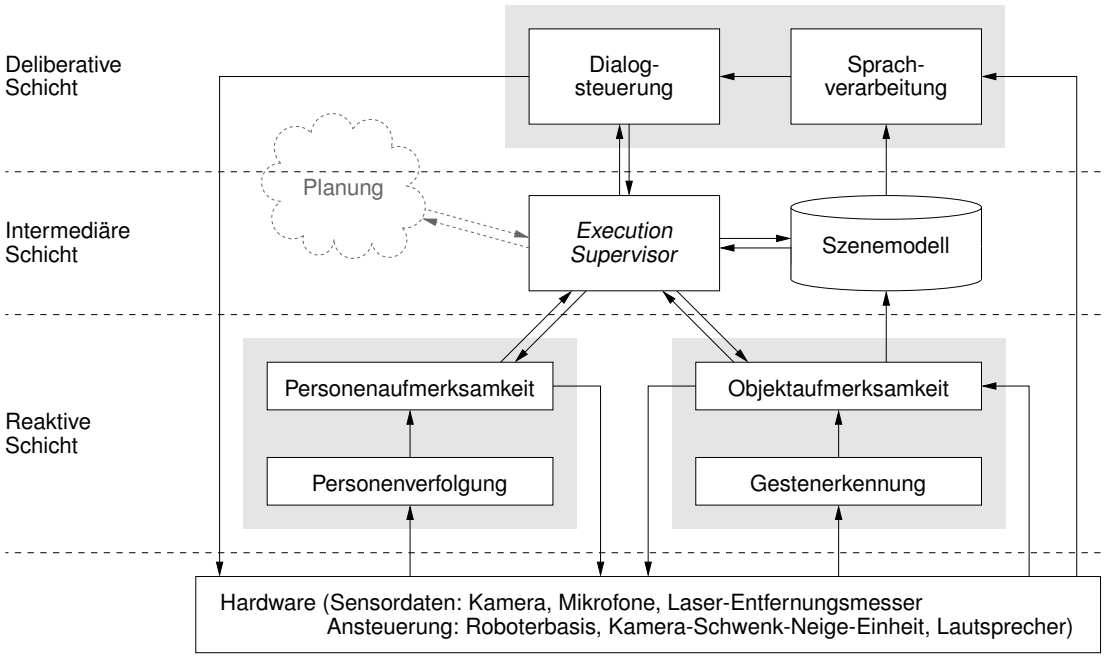



Abbildung 7.6: Drei-Schichten-Architektur

bei plötzlich und unerwartet auftretenden Schwierigkeiten bei der Ausführung von Plänen die deliberative Steuerung benachrichtigen, damit die Pläne entsprechend modifiziert werden. Der Datenaustausch zwischen reaktiver und deliberativer Steuerung ist aufgrund der verschiedenen Zeitskalen, auf denen die Prozesse arbeiten, in der Regel die zentrale Herausforderung hybrider Architekturen. Häufig werden zwischengelagerte Module eingeführt, die die Vermittlung übernehmen. In diesem Fall kann man die Module in drei Schichten unterteilen: deliberative, reaktive und intermediäre Schicht. Man spricht in diesem Fall von Drei-Schichten-Architekturen.

In der Drei-Schichten-Architektur von *BIRON* (siehe Abbildung 7.6) stellen die personenbasierte Aufmerksamkeitssteuerung in Verbindung mit der Personenverfolgung Komponenten der reaktiven Schicht dar. Sensordaten von Kamera, Mikrofonen und Laser werden verarbeitet und führen zu direkten Steuerungsanweisungen von Roboterbasis und Kamera. In analoger Weise handelt es sich bei der objektbasierten Aufmerksamkeitssteuerung, die Informationen von der Gestenerkennung bezieht, um eine reaktive Komponente. Die Dialogsteuerung in Verbindung mit der Sprachverarbeitung führt dagegen planerische Aufgaben durch und arbeitet auf einer größeren Zeitskala. Sie ist daher auf der deliberativen Schicht verortet. Die zentrale Komponente der Architektur von *BIRON* ist der so genannte *Execution Supervisor.*

### 7.6.1 *Execution Supervisor*

Der *Execution Supervisor* [Kle05] kontrolliert den Zustand des Gesamtsystems, steuert sequenzielle Abläufe, synchronisiert Ereignisse verschiedener Module und kontrolliert den modulwei-

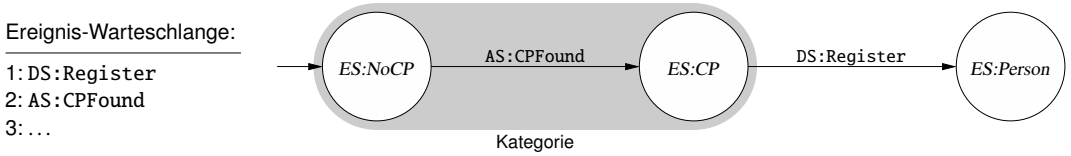



Abbildung 7.7: Funktionsweise von Zustandskategorien: Der aktuelle Zustand sei *ES:NoCP*. Das erste in der Warteschlange anliegende Ereignis `DS:Register` kann nicht angewandt werden. Da aber in derselben Kategorie ein Zustand *ES:CP* existiert, auf den das Ereignis angewandt werden könnte, wird es nicht verworfen sondern zurückgestellt. Nachdem das folgende Ereignis `AS:CPFound` nun zum Wechsel in den Zustand *ES:CP* führt, kann das zurückgestellte Ereignis anschließend verarbeitet werden.

ten Datenaustausch, wenn Informationen mit begrenzter Gültigkeitsdauer versendet werden. Der *Execution Supervisor* konfiguriert Module der reaktiven und der intermediären Schicht und sendet Statusmeldungen an die Module der deliberativen Schicht. Die Module wiederum senden an den *Execution Supervisor* Ereignisse, die eine zeitlich begrenzte Gültigkeit haben. Gültige Ereignisse können zu Zustandswechseln des *Execution Supervisor* führen, nicht mehr gültige Ereignisse werden zurückgewiesen.

Der *Execution Supervisor* ist als erweiterter deterministischer endlicher Automat realisiert. Jeder Zustand des Automaten repräsentiert einen Zustand des Gesamtsystems. Der Automat modelliert die sequenziellen Abläufe im Gesamtsystem. Transitionen werden durch Ereignisse ausgelöst, die von anderen Modulen gesendet wurden. Jedes Ereignis trägt einen Zeitstempel, der angibt, wann es erzeugt wurde, und es enthält Informationen darüber, wie lange das Ereignis gültig ist. Zudem können jedem Ereignis Daten angehängt werden, die jedoch vom *Execution Supervisor* nicht interpretiert werden beziehungsweise keinen Einfluss auf Transitionen haben. Der *Execution Supervisor* speichert die angehängten Daten zwischen, um sie zu einem späteren Zeitpunkt an andere Module weitersenden zu können.

Eintreffende Ereignisse werden zunächst anhand ihres Zeitstempels in eine Ereignis-Warteschlange einsortiert. Die ältesten Ereignisse werden immer zuerst verarbeitet. Ein Ereignis führt zu einer Transition, sofern eine solche für den momentanen Zustand des endlichen Automaten vorgesehen ist. Kann das Ereignis nicht angewendet werden, wird es ignoriert und an den Sender zurückgewiesen. Eine Ausnahme wird gemacht, wenn Ereignisse synchronisiert werden müssen, wenn also erst das Vorliegen mehrerer Ereignisse zu einem Wechsel führt. Die Ereignissynchronisation wird im *Execution Supervisor* durch so genannte Zustandskategorien realisiert. Eine Kategorie fasst mehrere Zustände zusammen. Wenn ein Ereignis vorliegt, das im momentanen Zustand nicht anwendbar ist, es aber bei einem anderen Zustand, der sich in derselben Kategorie wie der momentane Zustand befindet, zu einer Transition führen kann, so wird es nicht verworfen, sondern zurückgestellt. Abbildung 7.7 erläutert die Vorgehensweise anhand eines Beispiels.

Mit jeder Transition wird eine Aktion durchgeführt. Eine Aktion besteht aus dem Versenden von

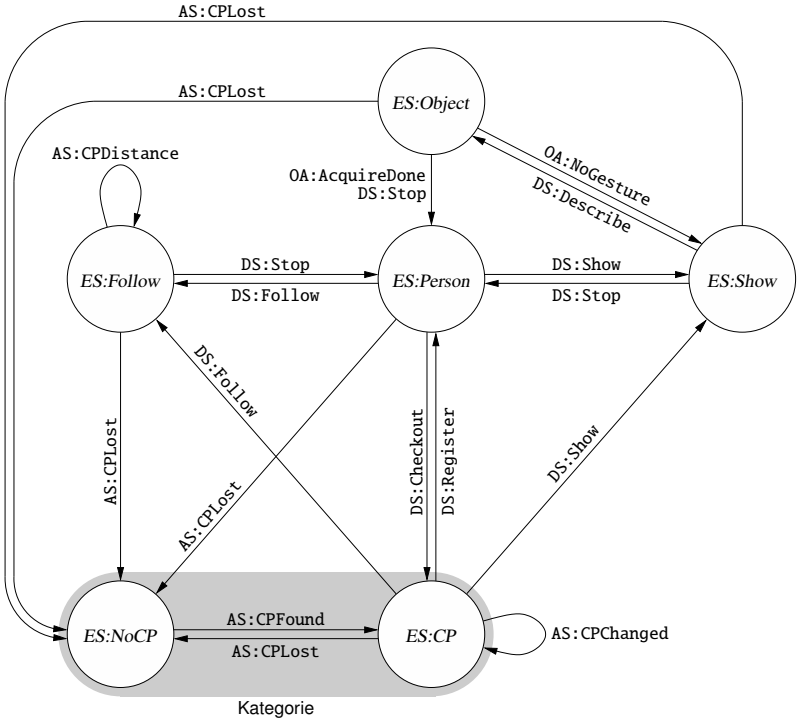



Abbildung 7.8: Erweiterter endlicher Automat des *Execution Supervisor.*

Nachrichten an andere Module. Nachrichten, die an Module der deliberativen Schicht gesendet werden, heißen Statusmeldungen. Diese informieren die Module über Zustandsänderungen, die daraufhin ihre Vorgehensweise abstimmen. Nachrichten, die an Module der reaktiven Schicht gesendet werden heißen Kommandos. Sie konfigurieren die Module, das heißt sie schreiben deren Vorgehensweise vor. Statusmeldungen und Kommandos können wiederum mit Daten versehen sein, die vorher mit Ereignissen eingetroffen waren und zwischengespeichert wurden.

Die Realisierung des erweiterten endlichen Automaten des *Execution Supervisor* für die Anwendung im *Home-Tour*-Szenario ist in Abbildung 7.8 angegeben. Im folgenden Abschnitt wird der Datenaustausch zwischen *Execution Supervisor* und Aufmerksamkeitssteuerung betrachtet.

### 7.6.2 Konfiguration der Aufmerksamkeitssteuerung

Die Aufmerksamkeitssteuerung sendet als Modul der reaktiven Schicht Nachrichten in Form von Ereignissen an den *Execution Supervisor.* Sie informiert im Wesentlichen über ihren aktuellen Zustand, zum Beispiel ob potenzielle Kommunikationspartner beobachtet werden konnten, oder ob der aktuelle Benutzer bei der Personenverfolgung verloren gegangen ist. Der *Execution Supervisor* wiederum sendet Kommandos an die Aufmerksamkeitssteuerung und löst damit dort Transitionen im endlichen Automaten aus. Dies gilt zumindest dann, wenn Aufmerksamkeitszustände des Top-down-Modus beteiligt sind. Im Top-down-Modus erfolgen die Wechsel von

Tabelle 7.1: Die Tabelle gibt an, welche Zustände die Aufmerksamkeitssteuerung in Abhängigkeit vom Zustand des *Execution Supervisor* einnehmen kann. Im Top-down-Modus gibt es eine eindeutige Korrespondenz.

| Zustand des *Execution Supervisor* | Mögliche Zustände der Aufmerksamkeitssteuerung |
|---|---|
| Bottom-up-Modus | |
| *ES:NoCP* | *AS:Sleeping*, *AS:Awake*, *AS:Alert*, *AS:Listening* |
| *ES:CP* | *AS:Alert*, *AS:Listening* |
| Top-down-Modus | |
| *ES:Person* | *AS:Person* |
| *ES:Follow* | *AS:Follow* |
| *ES:Show* | *AS:Show* |
| *ES:Object* | *AS:Object* |

Zuständen im *Execution Supervisor* und in der Aufmerksamkeitssteuerung synchron. Es gibt eine direkte Korrespondenz zwischen den Zuständen (siehe auch Tabelle 7.1). Mit jeder Transition im *Execution Supervisor* wird ein Kommando an die Aufmerksamkeitssteuerung gesendet, welches dort die entsprechende Transition auslöst. Der *Execution Supervisor* kontrolliert folglich im Top-down-Modus das Verhalten der Aufmerksamkeitssteuerung.

Im Folgenden wird der Datenaustausch zwischen Aufmerksamkeitssteuerung und *Execution Supervisor* anhand von Standardsituationen im *Home-Tour*-Szenario genauer beleuchtet.

**Beginn einer Interaktion:** Der Beginn einer Interaktion geht mit dem Wechsel zwischen Bottom-up- und Top-down-Modus der Aufmerksamkeitssteuerung einher. Es müssen dazu zwei Bedingungen erfüllt sein:

- Eine Person ist von der Aufmerksamkeitssteuerung als potenzieller Kommunikationspartner eingestuft worden, hat also zur selben Zeit zum Roboter geschaut und gesprochen.
- Die Dialogsteuerung hat aufgrund einer von der Sprachverarbeitung erkannten Äußerung den Beginn eines Dialogs registriert, zum Beispiel durch die Äußerung *„Hallo BIRON!“*.

Jede Bedingung für sich alleine ist zu schwach, um eine Kommunikation zu beginnen. Wenn zum Beispiel eine Person mit einer anderen über den Roboter spricht, wird sie sicherlich zwischendurch auch mal den Roboter anschauen. In diesem Fall würde sie gleichzeitig zum Roboter schauen und sprechen, obwohl sie nicht mit dem Roboter in Kontakt treten möchte. Ebenso kann es passieren, dass der Roboter aus dem Gespräch zweier Personen Äußerungen wahrnimmt, die als Beginn eines Dialogs mit dem Roboter verstanden werden könnten. Auch in diesem Fall darf der Roboter nicht reagieren. Um die Robustheit

des Anmeldeprozesses zu erhöhen, müssen also beide Bedingungen vorliegen. Die entsprechenden Ereignisse werden nicht notwendigerweise zeitgleich registriert. Der Wechsel erfolgt daher dann, wenn beide Ereignisse in einem begrenzten zeitlichen Intervall eintreffen. Diese Vorgabe wird durch den *Execution Supervisor* über zwei in einer Kategorie zusammengefasste Zustände realisiert (*ES:NoCP* und *ES:CP*).

Die Zustandswechsel in den jeweiligen endlichen Automaten von Aufmerksamkeitssteuerung und *Execution Supervisor* ergeben sich wie folgt: Wenn die Aufmerksamkeitssteuerung einen potenziellen Kommunikationspartner registriert hat, sendet sie ein entsprechendes Ereignis an den *Execution Supervisor*. Sie sendet das Ereignis `AS:CPFound`, wenn zuvor kein potenzieller Kommunikationspartner verfolgt wurde beziehungsweise das Ereignis `AS:CPChanged`, wenn eine andere Person die Eigenschaften des potenziellen Kommunikationspartners aufweist. Der *Execution Supervisor* wechselt daraufhin vom Zustand *ES:NoCP* direkt in den Zustand *ES:CP*. In entsprechender Weise sendet die Aufmerksamkeitssteuerung das Ereignis `AS:CPLost`, wenn die zuvor als potenzieller Kommunikationspartner eingestufte Person in der Personenverfolgung verloren gegangen ist. Der *Execution Supervisor* wechselt daraufhin wieder direkt in den Zustand *ES:NoCP*. Mit jedem Ereignis werden Daten über die betreffende Person, insbesondere eine eindeutige *ID* an den *Execution Supervisor* übermittelt.

Die zweite Bedingung, das Registrieren einer sprachlichen Äußerung, die von der Dialogsteuerung als Beginn eines Dialogs interpretiert wird, führt dazu, dass die Dialogsteuerung ein entsprechendes Ereignis an den *Execution Supervisor* sendet. Je nach Äußerung, wie zum Beispiel *„Hallo BIRON!“*, *„Folge mir!“* oder *„Ich zeige dir etwas.“*, wird eines der Ereignisse `DS:Register`, `DS:Follow` oder `DS:Show` gesendet. Jedes Ereignis ist mit einer begrenzten Gültigkeitsdauer versehen. Befindet sich der *Execution Supervisor* bereits im Zustand *ES:CP*, führt das Ereignis direkt zum Wechsel in einen der Zustände *ES:Person*, *ES:Follow* oder *ES:Show*. Die Anmeldung eines Benutzers ist damit erfolgt. Mit der Transition benachrichtigt der *Execution Supervisor* wiederum die Aufmerksamkeitssteuerung durch ein entsprechendes Kommando, das als Daten die *ID* der betreffenden Person enthält. Dieses veranlasst die Aufmerksamkeitssteuerung in den Top-down-Modus zu wechseln und die Aufmerksamkeit auf den neuen Benutzer, also die Person mit entsprechender *ID*, zu richten. Im anderen Fall, wenn der *Execution Supervisor* sich noch im Zustand *ES:NoCP* befand, kann das Ereignis der Dialogsteuerung nicht angewandt werden. Es wird aber, so lange es noch gültig ist, in der Ereignis-Warteschlange gehalten, da es im anderen Zustand *ES:CP*, welcher derselben Kategorie angehört, eine Transition auslösen kann. Die Kategorie mit den Zuständen *ES:NoCP* und *ES:CP* realisiert somit die Synchronisation der zwei Ereignisse, die für den Beginn einer Interaktion erforderlich sind.

**Ende einer Interaktion:** Das Ende einer Interaktion resultiert entweder aus der sprachlichen Abmeldung des Benutzers oder dadurch, dass die Personenverfolgung den Benutzer verliert, wenn zum Beispiel der Benutzer vom Roboter weggeht. Im ersten Fall sendet die Dialogsteuerung das Ereignis `DS:Checkout` an den *Execution Supervisor*. Dieser wech-

selt, sofern er sich im Zustand *ES:Person* befand, in den Zustand *ES:CP*. Mit der Transition wird ein Kommando an die Aufmerksamkeitssteuerung übermittelt, die den Übergang vom Zustand *AS:Person* in den Zustand *AS:Listening* veranlasst. Im zweiten Fall benachrichtigt die Aufmerksamkeitssteuerung den *Execution Supervisor* durch Senden des Ereignisses `AS:CPLost`, dass die Personenverfolgung den Benutzer verloren hat. Der *Execution Supervisor* wechselt daraufhin in den Ausgangszustand *ES:NoCP*, wobei mit der Transition ein Kommando an die Aufmerksamkeitssteuerung gesendet wird, das diese veranlasst, in den Zustand *AS:Listening* zu wechseln.

Beide Fälle führen bei der Aufmerksamkeitssteuerung zum Wechsel vom Top-down-Modus in den Bottom-up-Modus. Der Roboter löst seine Aufmerksamkeit vom Benutzer und beginnt erneut die Personen in seiner Nähe zu beobachten.

**Anweisungen:** Ausgehend vom Standardzustand der Interaktion kann der Benutzer über sprachliche Äußerungen das Verhalten des Roboters ändern beziehungsweise den Roboter in einen anderen Zustand führen. Die Anweisungen veranlassen die Dialogsteuerung dazu, entsprechende Ereignisse an den *Execution Supervisor* zu senden. Anweisungen wie *„Folge mir!“* oder *„Schau mal her!“* führen zum Wechsel in die Zustände *ES:Follow* beziehungsweise *ES:Show*. Mit einer Abbruchanweisung, wie zum Beispiel *„Stopp!“* kann der Roboter in analoger Weise immer wieder in den Grundzustand gebracht werden. Mit den Transitionen im *Execution Supervisor* gehen entsprechende Wechsel in der Aufmerksamkeitssteuerung einher (siehe auch Tabelle 7.1). Der Roboter zeigt somit auf Anweisungen des Benutzers gewünschte Verhaltensweisen.

**Hinterherfahren:** Wenn der Roboter im Folgen-Modus hinter dem Benutzer herfährt und der Benutzer schneller geht, als der Roboter folgen kann, dann droht der Roboter beziehungsweise die Personenverfolgung den Benutzer zu verlieren. Um dies zu verhindern, sendet die Aufmerksamkeitssteuerung bei Überschreiten eines Abstands zwischen Benutzer und Roboter das Ereignis `AS:CPDist` an den *Execution Supervisor*. Dies hat einen Selbstübergang des Zustands *ES:Follow* zur Folge. Bei der Transition wird eine Statusmeldung an die Dialogsteuerung übermittelt, die auf das Problem hinweist. Die Dialogsteuerung kann daraufhin dem Benutzer das Problem mitteilen und ihn zum Beispiel bitten, langsamer zu gehen.

**Zeigen eines Objekts:** Das Zeigen von Objekten läuft nach einem festen Schema ab. Die Wechsel der entsprechenden Zustände im *Execution Supervisor* und in der Aufmerksamkeitssteuerung laufen synchron ab (siehe Tabelle 7.1). Das heißt, mit jeder Transition im *Execution Supervisor* sendet dieser ein Kommando an die Aufmerksamkeitssteuerung, das dort die entsprechende Transition auslöst. Die Transitionen im *Execution Supervisor* werden wiederum durch Ereignisse von der Dialogsteuerung und der Objektaufmerksamkeit ausgelöst. Die beim Zeigen von Objekten beteiligten Zustände sind der Ausgangszustand *ES:Person*, der Zustand *ES:Show*, in dem der Roboter sprachliche und gestische Eingaben über das Objekt entgegen nimmt, und der Zustand *ES:Object*, der die Objektaufmerksamkeit aktiviert, um das Objekt mit der Kamera zu fokussieren.

## 7.7 Auditive Aufmerksamkeit

Als weiterer Aspekt der multimodalen Aufmerksamkeitssteuerung für mobile Roboter wird in diesem Abschnitt die auditive Aufmerksamkeit behandelt.

Im angestrebten Szenario soll es den Benutzern erlaubt sein, in möglichst natürlicher Weise mit dem Roboter zu kommunizieren. Da Sprache das gebräuchlichste Mittel in der menschlichen Kommunikation ist, muss der Roboter natürlichsprachliche Äußerungen erkennen und verstehen können. Im Gegensatz zu einfachen Szenarien, in denen der Sprecher mit geringem Abstand in ein Mikrofon spricht, der Signal-Rausch-Abstand groß ist und nur das Sprachsignal aufgezeichnet wird, das tatsächlich verarbeitet werden soll, ergibt sich für die Sprachverarbeitung von einem mobilen Roboters aus ein deutlich schwierigeres Bild. Die Verwendung von Nahbesprechungsmikrofonen wird prinzipiell ausgeschlossen, da von den Benutzern nicht gefordert werden soll, spezielle Hardware zu tragen. Die Mikrofone sind folglich auf dem Roboter montiert. Im Szenario mit einem mobilen Roboter ist die Position des Sprechers relativ zum Roboter variabel. Die Verwendung eines Richtmikrofons ist problematisch, da bei Anwesenheit mehrerer Personen nicht garantiert werden kann, dass das Mikrofon bereits dann auf den Sprecher gerichtet ist, wenn dieser zu sprechen beginnt. Es würden Teile der sprachlichen Äußerung nicht erfasst. Es müssen daher Mikrofone eingesetzt werden, die den gesamten vorderen Bereich des Roboters abdecken, in dem sich Benutzer in der Regel aufhalten, um mit dem Roboter zu kommunizieren. Auf dem Roboter *BIRON* werden zwei handelsübliche Grenzflächenmikrofone eingesetzt (siehe Abschnitt 3.1).

Bei der Sprachverarbeitung von einem mobilen Roboter aus ergeben sich insbesondere die folgenden Probleme:

- Die Position des Sprechers ist variabel und sein Abstand zu den Mikrofonen ist groß. Daher besitzt das Sprachsignal eine geringe Energie. Zusätzlich werden neben dem Sprachsignal Störgeräusche aus anderen Richtungen aufgezeichnet.
- Nicht jede sprachliche Äußerung einer Person in der Nähe des Roboters ist an den Roboter gerichtet. Da das von der Sprachverarbeitung verwendete Lexikon in der Regel auf das Aufgabengebiet des Roboters begrenzt ist, besteht die Gefahr, dass nicht aufgabenbezogene Äußerungen falsch interpretiert werden.

Die genannten Probleme erfordern zusätzliche eine auditive Aufmerksamkeitssteuerung, die es dem Roboter erlaubt

- seine auditive Aufmerksamkeit auf den aktuellen Sprecher zu richten und gleichzeitig Störgeräusche aus anderen Richtungen zu ignorieren, und
- zu entscheiden, wann er dem aktuellen Sprecher zuhört und das aufgezeichnete Sprachsignal verarbeitet und wann nicht.

Diese beiden Aspekte werden in den folgenden zwei Abschnitten behandelt.

### 7.7.1 Selektive auditive Aufmerksamkeit

Der Mensch ist in der Lage, die Äußerungen eines einzelnen Sprechers zu verstehen, selbst wenn gleichzeitig andere Personen im selben Raum sprechen oder weitere Störgeräusche vorhanden sind. Diese Fähigkeit zur selektiven auditiven Aufmerksamkeit wird in der Literatur gemeinhin als Cocktailparty-Effekt bezeichnet [Che53]. Würde man von derselben Situation eine Mono-Aufnahme anfertigen und diese einer Person vorspielen, so würde es der Person deutlich schwerer fallen, einem einzelnen Sprecher zu folgen. Das räumliche Hören und die Fähigkeit zur Lokalisation von Geräuschquellen spielt offenbar eine wichtige Rolle für die selektive auditive Aufmerksamkeit.

Bei technischen Systemen kann man sich das über mehrere Mikrofone gewonnene Wissen über die relative Position des Sprechers ebenfalls zu Nutze machen, um bessere Ergebnisse bei der automatischen Sprachverarbeitung zu erzielen. Mit Hilfe von *Beamforming* [Joh93] ist es möglich, das Sprachsignal zu verstärken und Störgeräusche aus anderen Richtung zu dämpfen. Auf *BIRON* erfolgt die Sprecherlokalisation über die multimodale Personenverfolgung. Aus der geometrischen Anordnung von Sprecher und Roboter werden die zwei Abstände zwischen dem Sprecher und den jeweiligen Mikrofonen bestimmt. Schallwellen erreichen das näher gelegene Mikrofon eher als das weiter entfernte. Es ergibt sich ein Laufzeitunterschied $\delta$, der in zeitlich zueinander verschobenen Signalen resultiert. *Beamforming* kompensiert diesen Effekt, indem es die aufgezeichneten Signale zunächst entsprechend des berechneten Laufzeitunterschieds in entgegengesetzter Richtung verschiebt und dann zu einem einzelnen Signal durch Summation zusammenfasst. Schallwellen, die mit dem Laufzeitunterschied $\delta$ von den Mikrofonen erfasst wurden, werden im resultierenden Signal durch Überlagerung verstärkt, während Geräusche aus anderen Richtungen keine Verstärkung durch Überlagerung erfahren. Das *Beamforming* realisiert somit die selektive auditive Aufmerksamkeit des Roboters.

### 7.7.2 Aktivierung der Sprachverarbeitung

Die situationsabhängige Aktivierung der Sprachverarbeitung soll garantieren, dass nur Äußerungen, die an den Roboter gerichtet sind, verarbeitet werden, während andere Äußerungen, die sich zum Beispiel im Gespräch zweier Personen vor dem Roboter ergeben, ignoriert werden. Die zentrale Aufgabe ist zu entscheiden, wann eine Äußerung an den Roboter gerichtet ist und über welchen Zeitraum sie sich erstreckt. Für die Aktivierung und die Deaktivierung der Sprachverarbeitung müssen jeweils die Zeitpunkte des Anfangs und des Endes der Äußerung an das Sprachverarbeitungsmodul übermittelt werden.

Zunächst werden sprachliche Eingaben nur von der Person $p^\star$ berücksichtigt, die sich im Fokus der Aufmerksamkeit des Roboters befindet. Die Frage, woran zu erkennen ist, wann die beobachtete Person den Roboter anspricht, wurde bereits bei der Definition des Relevanzwerts für den Auswahlmechanismus der bottom-up gesteuerten Aufmerksamkeit in Abschnitt 7.3.2 diskutiert. Demnach wurde die Annahme gemacht, dass eine Person den Roboter genau dann anspricht, wenn sie zur selben Zeit spricht und den Roboter anschaut. Das heißt, die Spracherkennung wird

zu dem Zeitpunk $t$ aktiviert, wenn für den Relevanzwert $\mathrm{Relevanz}_t(p^\star)$ der Person $p^\star$ der höchstmögliche Wert $c_{spricht} + c_{schaut}$ ermittelt wurde. Die Sprachverarbeitung bleibt so lange aktiviert, wie ohne größere Unterbrechung gesprochen wird. Bei der Deaktivierung muss berücksichtigt werden, dass der Blick des Sprechers in der zwischenmenschlichen Kommunikation im Mittel nur zu etwa 40% der Zeit auf den Zuhörenden gerichtet ist (vgl. [Arg75], Seite 229). Das bedeutet, dass die Deaktivierung nicht dadurch erfolgen darf, dass die Eigenschaft `schaut` für den Sprecher nicht mehr zutrifft. Der Zeitpunkt $t$ der Deaktivierung richtet sich alleine danach, wann die Person $p^\star$ aufhört zu sprechen oder – anders ausgedrückt – der Relevanzwert $\mathrm{Relevanz}_t(p^\star)$ einen kleineren Wert als $c_{spricht}$ einnimmt.

Die Bedingungen können wie folgt zusammengefasst werden. Die Sprachverarbeitung wird zum Zeitpunkt $t_i$ aktiviert, wenn die Person $p^\star$ im Fokus der Aufmerksamkeit den Roboter anspricht, und bleibt so lange aktiviert, wie die Person ununterbrochen spricht:

$$\begin{aligned} \mathrm{Aktivierung}(t_i) \quad \Leftrightarrow \quad & (\mathrm{Relevanz}_{t_i}(p^\star) = c_{spricht} + c_{schaut}) \quad \vee \\ & (\mathrm{Relevanz}_{t_i}(p^\star) \geq c_{spricht}) \quad \wedge \quad \mathrm{Aktivierung}(t_{i-1}) \end{aligned} \tag{7.3}$$

Die Entscheidung über die Aktivierung der Sprachverarbeitung geschieht anhand der Eigenschaften `spricht` beziehungsweise `schaut`. Wie in Abschnitt 7.3.1 auf Seite 103 beschrieben, werden die Eigenschaften auf Basis des Verhältnisses zwischen zugeordneten und nicht zugeordneten Perzepten innerhalb eines Zeitfensters der Dauer $\Delta t$ bestimmt. Da dabei unumgänglicherweise ein Zeitbereich in der Vergangenheit betrachtet wird, werden die Eigenschaften immer erst verspätet erkannt. Diese Verzögerung muss bei der Aktivierung der Sprachverarbeitung berücksichtigt werden, da genau der Zeitpunkt relevant ist, der den tatsächlichen Beginn der Äußerung im Sprachsignal kennzeichnet. Wenn die Formel 7.3 den Zeitpunkt $t_i$ als Beginn der Aktivierung liefert, dann wird der Sprachverarbeitung ein um die Hälfte der Länge des Zeitfensters korrigierter Wert $t_i - \frac{\Delta t}{2}$ als Zeitpunkt des Beginns der Äußerung übermittelt.

## 7.8 Zusätzliche Rückmeldung über ein animiertes Gesicht

Um die Rückmeldung des Roboters an den Benutzer zu unterstützen, die aus dem Ausrichten der Kamera und der Bewegung der Roboterbasis entsteht, wird ein animiertes Gesicht auf dem Flachbildschirm dargestellt (siehe Abbildung 7.9). Das Gesicht kann darüber hinaus verwendet werden, um neben der Richtung des Aufmerksamkeitsfokus auch den Aufmerksamkeitszustand intuitiv verständlich zu visualisieren. Das Gesicht enthält mit Augen, Augenbrauen und Mund nur die Gesichtsmerkmale, die in der Kommunikation die größte Bedeutung haben. Zusätzlich verfügt es über Haare, um ein gefälliges Erscheinungsbild zu erzeugen.

Die Augen geben durch die Position der Pupillen und durch geöffneten oder geschlossenen Zustand zweierlei Art von Rückmeldung an den Benutzer. Die Pupillen bewegen sich analog zur Ausrichtung der Kamera und stellen dadurch den Fokus der Aufmerksamkeit des Roboters nach außen dar. Die Augen sind im Aufmerksamkeitszustand *AS:Sleeping* geschlossen und zeigen

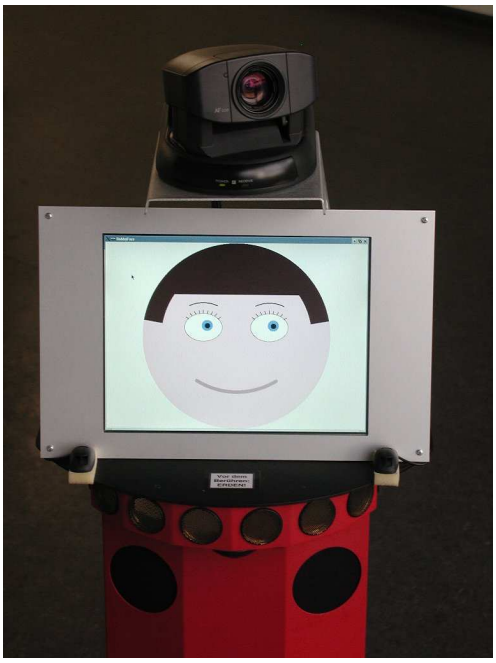


Abbildung 7.9: Ein Gesicht auf dem Flachbildschirm des Roboters zeigt neben der Ausrichtung der Sensoren über die Augen den Fokus der Aufmerksamkeit an.

damit an, dass sich der Roboter im Schlafmodus befindet und die bildverarbeitenden Module deaktiviert sind. In allen anderen Zuständen sind die Augen geöffnet. Sie blinzeln von Zeit zu Zeit, um anzuzeigen, dass das System arbeitet und der Roboter weiterhin die Personen registriert.

Die Augenbrauen und der Mund werden eingesetzt, um verschiedene Gesichtsausdrücke, wie neutral, angestrengt, freundlich und freudestrahlend darzustellen. Der neutrale Gesichtsausdruck wird in den Zuständen *AS:Sleeping* und *AS:Awake* eingesetzt. Wenn der Fokus der Aufmerksamkeit auf eine Person gerichtet ist, hängt der Gesichtsausdruck davon ab, ob die Person zum Roboter schaut oder nicht. Bei Blickkontakt zeigt das Gesicht ein Lächeln, um die Person zu ermutigen, eine Interaktion zu beginnen. Ansonsten nimmt das Gesicht den angestrengten Ausdruck mit hochgezogenen Brauen und einem schmalen Mund ein. Wenn die Person zum Roboter spricht, wird der freudestrahlende Gesichtsausdruck mit breitem, lächelndem Mund und hochgezogenen Augenbrauen angezeigt.

## 7.9 Zusammenfassung

In diesem Kapitel wurde das in dieser Arbeit entwickelte multimodale Aufmerksamkeitssystem für mobile Roboter beschrieben. Es realisiert die personenbasierte selektive Aufmerksamkeit des Roboters. Die Aufmerksamkeitssteuerung ist eine wesentliche Voraussetzung für die natürliche

Interaktion mit Menschen und bildet damit einen Grundbaustein für anspruchsvolle Anwendungen, wie zum Beispiel das *Home-Tour*-Szenario.

Die Aufmerksamkeitssteuerung modelliert sowohl das Verhalten des Roboters in der Bereitschaftsphase, in der er aufgrund audio-visueller Stimuli nach neuen Kommunikationspartnern Ausschau hält, als auch das Verhalten in der Interaktionsphase, in der die Aufmerksamkeit des Roboters allein auf den Benutzer gerichtet wird. Das Ausrichten der Aufmerksamkeit beinhaltet zum einen ein nach außen hin sichtbares Verhalten, bei dem die Sensoren auf die selektierte Person ausgerichtet werden, um sowohl die Wahrnehmung zu optimieren als auch eine Rückmeldung an die beobachtenden Personen zu geben. Letztere wird noch mittels eines auf dem Flachbildschirm des Roboters dargestellten Gesichts unterstützt, das durch die Stellung der Pupillen den Aufmerksamkeitsfokus des Roboters anzeigt und durch verschiedene Gesichtsausdrücke den Zustand des Roboters intuitiv verständlich darstellt. Zum anderen beinhaltet das Ausrichten der Aufmerksamkeit einen nach außen hin nicht sichtbaren Prozess, der die selektive auditive Aufmerksamkeit realisiert. Dabei wird die Sprachverarbeitung so angesteuert, dass nur Äußerungen der fokussierten Person verarbeitet werden, die an den Roboter gerichtet sind.

Die Aufmerksamkeitssteuerung geht in ihrer Gesamtheit über andere bisher entwickelte Verfahren für mobile Roboter hinaus, da hier sowohl die Bereitschaftsphase als auch die Interaktionsphase berücksichtigt werden. Viele Ansätze ermöglichen dem jeweiligen Roboter zwar aufgrund audio-visueller Reize ihre Aufmerksamkeit entsprechend auszurichten, sie verfügen aber über keinen expliziten Mechanismus, der den Beginn einer längerfristigen Interaktion mit einem Benutzer erkennt, um währenddessen Stimuli anderer Personen zu ignorieren. Beispiele hierfür sind *SIG* [Oku01] und *Lino* [Krö03]. Andere Ansätze, bei denen die Roboter aus der Beobachtung von Personen einen Benutzer für eine Interaktion erkennen, fällen die entsprechende Entscheidung anhand von Merkmalen, die für eine natürliche Interaktion im allgemeinen Fall ungeeignet sind. Beispiel für ein solches Merkmal ist der Abstand einer Person in [Doi02].

Die hier vorgestellte Aufmerksamkeitssteuerung ist auf einem mobilen Roboter realisiert, der mit handelsüblichen Sensoren ausgestattet ist. Da audio-visuelle Stimuli die Aufmerksamkeit des Roboters lenken, sind die Kamera und die Mikrofone unverzichtbar. Entsprechende Sensoren wird man wahrscheinlich auch in Zukunft zum Beispiel bei weit entwickelten humanoiden Robotern finden. Der Laser-Entfernungsmesser ist für die Aufmerksamkeitssteuerung insofern wichtig, als dass er die periphere Wahrnehmung des Roboters realisiert und dadurch ein robustes Verfolgen von Personen über einen großen Bereich ermöglicht. Eine ähnliche Leistung ließe sich auch durch geeignete Weitwinkelkameras erzielen, mit denen Personen zum Beispiel über ihre Silhouette erkannt werden. Das vorgestellte Verfahren lässt sich daher bei geringen Modifikationen auch auf anderen Robotern einsetzten.

# Kapitel 8

# Evaluation

Dieses Kapitel präsentiert Ergebnisse von Experimenten, die im Laufe der Entwicklung des Aufmerksamkeitssystems durchgeführt wurden. Mit den Experimenten sollte die Effektivität der erarbeiteten Ansätze überprüft werden. Einerseits wurde das multimodale *Anchoring* zum Verfolgen von Personen im Einsatz auf dem Roboter getestet, andererseits stand die Aufmerksamkeitssteuerung im Zentrum der Untersuchung. Im Hinblick auf eine der Zielsetzungen dieser Arbeit, ein lauffähiges System zu entwickeln, wurden alle Experimente auf dem Roboter unter realen Bedingungen durchgeführt, wobei die Versuchspersonen einfache Interaktionsaufgaben zu bewältigen hatten.

Im folgenden Abschnitt wird zunächst ein kurzer Überblick über die Implementierung des eingesetzten Gesamtsystems gegeben. Die daran anschließenden Abschnitte beschreiben die Experimente zum multimodalen *Anchoring* (Abschnitt 8.2) und zur Aufmerksamkeitssteuerung (Abschnitt 8.3).

## 8.1 Realisierung

Das multimodale *Anchoring* und die Aufmerksamkeitssteuerung sind objekt-orientiert unter Verwendung der Programmiersprache C++ implementiert worden. Die Komponenten, die in den Sensorsystemen der unimodalen *Anchoring*-Prozesse für die Extraktion von Perzepten aus den Sensordaten verantwortlich sind, stellen eigenständige Prozesse dar und arbeiten daher parallel und asynchron. Eine Ausnahme bilden die bildverarbeitenden Komponenten, die als *Plugins* für das Programm *iceWing* [Löm04] realisiert wurden. Da *iceWing* die *Plugins* mit Bilddaten versorgt und sequenziell aufruft, laufen Gesichts- und Oberkörperdetektion synchron.

Die Aufmerksamkeitssteuerung, die die Bewegung des Roboters bestimmt, hat keinen direkten Zugriff auf die Roboterhardware. Um eine höhere Flexibilität zu erreichen, ist eine Robotersteuerungssoftware zwischengeschaltet, die eine geeignete Schnittstelle zu den Aktuatoren und Sensoren des Roboters darstellt. In einer früheren Version war die Aufmerksamkeitssteuerung als ein *Behavior* in der *ISR*-Software [And99] realisiert. Die neuere Version verwendet die

*Player/Stage*-Software [Ger03] als Schnittstelle zwischen Aufmerksamkeitssteuerung und Roboterhardware.

Die beteiligten Prozesse sind wie folgt auf die beiden Rechner des Roboters verteilt: Die Sprecherlokalisation und die Robotersteuerungssoftware laufen auf dem 850-MHz-Rechner. Der zweite Rechner (500 MHz) wird für das multimodale *Anchoring*, die Aufmerksamkeitssteuerung, die Bildverarbeitung, die Darstellung des animierten Gesichts auf dem Flachbildschirm und den *Execution Supervisor* genutzt. Sprachverarbeitung und Dialogsteuerung sind auf einen externen Laptop ausgelagert.

## 8.2 Experimente zum multimodalen *Anchoring*

Die Experimente zum multimodalen *Anchoring* wurden in verschiedenen Entwicklungsphasen des Systems durchgeführt. Das bedeutet, dass nicht immer alle der vier beschriebenen unimodalen *Anchoring*-Prozesse zur Verfügung standen. In einem ersten Experiment wurde das *Tracking*-Konzept unter Verwendung der Bein- und Gesichtsdetektion untersucht. Der Roboter sollte dabei einer Versuchsperson hinterherfahren, ohne von anderen, gleichzeitig anwesenden Personen abgelenkt zu werden. In dem schwierigen dynamischen Szenario sollte die grundlegende Effektivität des *Tracking*-Verfahrens gezeigt werden. In einer zweiten Experimentphase wurde die Einbindung der Sprecherlokalisation getestet. Es sollte analysiert werden, wie gut die Lokalisation unter der Verwendung eines einzelnen Mikrofonpaars im multimodalen Gesamtverfahren funktioniert. In einem weiteren Experiment wurde die Erweiterung um die Fähigkeit zur Lokalisation des Oberkörpers getestet. In einem entsprechenden Szenario, in dem der Roboter auf die Oberkörperperzepte angewiesen war, sollte die Effektivität des Verfahrens demonstriert werden.

Alle Experimente wurden mit Versuchspersonen durchgeführt. Die Versuchspersonen hatten einfache Interaktionsaufgaben mit dem Roboter zu bewältigen. Zu diesem Zweck wurden, aufbauend auf dem multimodalen *Anchoring*, einfache Anwendungen konzipiert, die jeweils Teilaspekte des Aufmerksamkeitssystems widerspiegelten. Das heißt, Bestandteil aller Experimente war es, dass Versuchspersonen die Aufmerksamkeit des Roboters erlangen. In den folgenden Unterabschnitten werden die einzelnen Experimente detailliert beschrieben.

### 8.2.1 Folgen einer Person

Das Ziel des hier beschriebenen Experiments ist es, die generelle Leistungsfähigkeit des *Tracking*-Verfahrens in einem dynamischen Szenario zu testen, wobei sich sowohl der Benutzer als auch der Roboter bewegen (siehe auch [Fri03b]). Für den Versuch wurde der Roboter so programmiert, dass er einer vorausgehenden Person hinterherfährt. Dieses Verhalten findet sich auch im Top-down-Modus der Aufmerksamkeitssteuerung in dem Aufmerksamkeitszustand *AS:Follow* wieder (siehe Abschnitt 7.4). Um das Folgen zu aktivieren, musste sich die Versuchs-

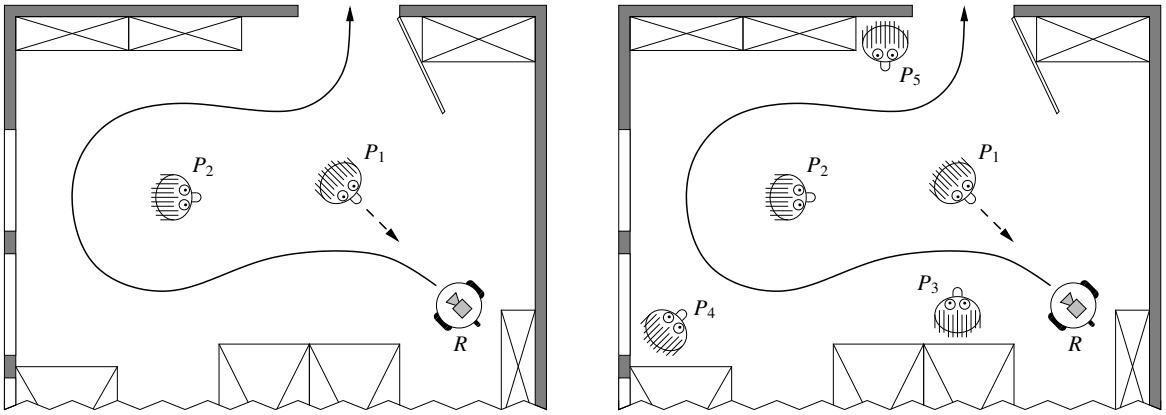



Abbildung 8.1: Verschiedene Versuchsanordnungen mit zwei Personen (a) und mit fünf Personen (b). Die Aufgabe von $P_1$ ist es, den Roboter um $P_2$ herum aus dem Raum zu führen.

person dem Roboter auf unter einen Meter Abstand nähern. Von diesem Zeitpunkt an richtete sich der Roboter immer direkt auf den Benutzer aus und versuchte, den Abstand konstant zu halten.

Im *Tracking*-Verfahren fanden die unimodalen *Anchoring*-Prozesse für Beine und Gesichter Verwendung, wobei zur Generierung von Gesichtsperzepten die weniger leistungsfähige Variante zur Verfügung stand, die die Detektion über adaptive Hautfarbensegmentierung, kombiniert mit der *Eigenface*-Methode, realisiert (siehe Abschnitt 6.2.1). Neben der generellen Effektivität des *Tracking*-Verfahrens bei dieser Aufgabe sollten die Erkennungsraten der Sensorsysteme in Form von Zuordnungen von Perzepten zu den *Anchor*-Funktionen getestet werden.

### Experimenteller Aufbau

Das Experiment wurde in einem Büroraum in einem freien Bereich mit einer Fläche von ungefähr 4,6 m$\times$3,4 m durchgeführt. Der Raum war mit Möbeln aus hellem Holz ausgestattet. Dieser Umstand stellt für die Gesichtsdetektion eine Herausforderung dar, da Holz und Haut eine ähnliche Farbe aufweisen. Es wurden zwei Einzelexperimente durchgeführt. Die unterschiedlichen Versuchsanordnungen sind in Abbildung 8.1 dargestellt.

Im ersten Szenario (Abbildung 8.1 (a)) waren zwei Personen anwesend, von denen eine ($P_2$) relativ mittig im Raum positioniert war und während der Durchführung auf ihrer Position verharrte. Die Aufgabe der anderen Person ($P_1$) war es, den Roboter durch den Raum zu führen. Dazu musste $P_1$ zunächst die Aufmerksamkeit des Roboters erlangen, indem sie sich dem Roboter wie oben beschrieben näherte. Dann sollte sie den Roboter, wie in der Abbildung 8.1 eingezeichnet, um $P_2$ herumführen und mit dem Roboter den Raum durch die offene Tür verlassen. Während des Folgens sollte die führende Person $P_1$ so lange wie möglich zum Roboter schauen, damit dieser in der Lage war, das Gesicht zu detektieren. Der Aufbau des zweiten Szenarios unterschied

Tabelle 8.1: Ergebnisse des ersten Experiments mit Personen $P_1$ und $P_2$

| | | | | *grounded* (%) | | | Perzepte/Arbeitszyklus | |
|---|---|---|---|---|---|---|---|---|
| Nr. | $t$ (s) | $v_\varnothing$ (m/s) | verloren | Person | Beine | Gesicht | Beine | Gesicht |
| 1 | 39 | 0,19 | 0 | 99,7 | 98,9 | 63,1 | 1,78 | 0,76 |
| 2 | 62 | 0,12 | 0 | 96,6 | 93,8 | 36,4 | 1,71 | 0,90 |
| 3 | 52 | 0,14 | 1 | 95,4 | 83,5 | 51,0 | 1,72 | 0,57 |
| 4 | 56 | 0,13 | 0 | 99,3 | 93,8 | 54,1 | 1,79 | 0,59 |
| 5 | 81 | 0,09 | 1 | 96,4 | 95,7 | 34,7 | 1,63 | 0,40 |
| 6 | 32 | 0,23 | 0 | 99,2 | 98,7 | 51,1 | 1,87 | 0,78 |
| 7 | 90 | 0,08 | 1 | 80,9 | 73,2 | 22,0 | 1,94 | 0,49 |
| 8 | 51 | 0,15 | 0 | 99,2 | 98,8 | 56,5 | 1,75 | 0,70 |
| 9 | 42 | 0,18 | 0 | 98,0 | 97,9 | 35,7 | 1,79 | 0,45 |
| 10 | 44 | 0,17 | 0 | 88,6 | 87,1 | 16,3 | 1,60 | 0,39 |
| ∅ | 55 | 0,14 | – | 95,3 | 92,1 | 42,1 | 1,76 | 0,60 |

sich vom ersten durch die Anwesenheit von drei zusätzlichen Personen $P_3$–$P_5$ (siehe Abbildung 8.1 (b)). Diese waren an vorgeschriebenen Positionen im Raum platziert, ohne jedoch die aus dem ersten Versuchsaufbau abzulaufende Bahn zu beeinflussen. Die sich ergebende Strecke hatte eine Länge von ungefähr 7,5 m.

Für den Fall, dass der Roboter die Person beim Folgen verlor, wurde $P_1$ angewiesen, die Aufmerksamkeit des Roboters wiederzuerlangen und die Aufgabe fortzuführen. Wenn dies nicht möglich war, weil der Roboter direkt versuchte, einer der anderen im Raum anwesenden Personen zu folgen, weil der Abstand zu dieser Person weniger als einen Meter betrug, wurde der entsprechende Durchlauf als Fehlversuch gewertet. Für beide Versuchsanordnungen wurden zehn Durchläufe mit jeweils zehn verschiedenen Versuchspersonen durchgeführt.

### Versuchsergebnisse

Im ersten Szenario (siehe Tabelle 8.1) wurde die Aufgabe in einer durchschnittlichen Zeit von 55 s bewältigt. Der Roboter hat drei Versuchspersonen einmal verloren. Diese waren jedoch in der Lage, die Aufmerksamkeit des Roboters wiederzuerlangen und den Durchlauf erfolgreich zu beenden. Der multimodale *Anchor* für die Person $P_1$ war im Mittel zu 95,3% der Zeit im Zustand *grounded*, wobei der unimodale *Anchoring*-Prozess für Beine mit 92,1% der Zeit im Zustand *grounded* eine relativ hohen, und der für das Gesicht mit 42,1% der Zeit im Zustand *grounded* einen vergleichsweise niedrigen Beitrag leistete. Im Durchschnitt wurden in jedem Arbeitszyklus des jeweiligen Sensorsystems 1,76 Beinperzepte und 0,60 Gesichtsperzepte generiert.

Die Zeit, die benötigt wurde, um die Aufgabe in dem zweiten, komplexeren Versuchsaufbau erfolgreich zu bewältigen, war im Mittel nur zwei Sekunden länger (siehe Tabelle 8.2). Für den zweiten Aufbau wurde erwartet, dass wegen der drei zusätzlich anwesenden Personen mehr Perzepte generiert werden. Dies traf tatsächlich auf die Beinperzepte zu, jedoch nicht auf die

Tabelle 8.2: Ergebnisse des zweiten Experiments mit Personen $P_1$ – $P_5$

| | | | | *grounded* (%) | | | Perzepte/Arbeitszyklus | |
|---|---|---|---|---|---|---|---|---|
| Nr. | $t$ (s) | $v_\varnothing$ (m/s) | verloren | Person | Beine | Gesicht | Beine | Gesicht |
| 1 | 60 | 0,13 | 2 | 93,6 | 91,5 | 27,7 | 2,63 | 0,41 |
| 2 | 43 | 0,17 | 0 | 96,7 | 95,0 | 20,7 | 2,61 | 0,32 |
| 3 | Der Roboter hat $P_1$ verloren und versuchte $P_3$ zu folgen. | | | | | | | |
| 4 | 51 | 0,15 | 0 | 98,7 | 90,4 | 66,0 | 2,49 | 0,74 |
| 5 | 47 | 0,16 | 0 | 96,2 | 94,5 | 7,1 | 2,52 | 0,20 |
| 6 | Der Roboter hat $P_1$ verloren und versuchte $P_2$ zu folgen. | | | | | | | |
| 7 | 77 | 0,10 | 0 | 99,8 | 97,5 | 72,0 | 2,59 | 0,85 |
| 8 | 74 | 0,10 | 0 | 93,4 | 92,6 | 20,3 | 2,63 | 0,22 |
| 9 | 61 | 0,12 | 0 | 97,7 | 96,1 | 36,4 | 2,55 | 0,56 |
| 10 | 42 | 0,18 | 0 | 86,1 | 84,2 | 11,9 | 2,73 | 0,26 |
| $\varnothing$ | 57 | 0,13 | – | 95,3 | 92,7 | 32,8 | 2,59 | 0,45 |

Gesichtsperzepte. Die Versuchspersonen, die den Roboter führten, achteten darauf, nicht mit einer der Personen $P_3$ – $P_5$ zu kollidieren und schauten daher weniger oft in Richtung Kamera. Dies führte zu einer entsprechend niedrigeren Rate bei der Gesichtsdetektion. Im Durchschnitt war der unimodale *Anchor* für das Gesicht für die Person $P_1$ zu 32,8% der Zeit im Zustand *grounded*. Die Werte für den unimodalen *Anchor* für Beine und für den multimodalen *Anchor* für die Gesamtperson waren mit 95,3% und 92,7% ähnlich zu denen im ersten Experiment. Die Durchläufe 3 und 6 mussten als Fehlversuch gewertet werden, da der Roboter, unmittelbar nachdem er $P_1$ verloren hatte, versuchte einer anderen Person zu folgen, die sich in dem Moment im Aufmerksamkeitsbereich von weniger als einem Meter Abstand befand.

Während der Tests stellte der Laser-Entfernungsmesser neue Messdaten mit einer Rate von 4,6 Hz zur Verfügung, wobei die Zeit, die zur Generierung der Beinperzepte notwendig war, vernachlässigbar klein war. Die adaptive Hautfarbensegmentierung verarbeitete Bilder der Größe 192×144. Für jede segmentierte Hautfarbenregion wurde die Gesichtsdetektion durchgeführt. Die Verarbeitungszeit eines Einzelbilds hängt daher von der Anzahl der segmentierten Regionen ab. Im Durchschnitt wurden Gesichtsperzepte mit einer Rate von 3,1 Hz generiert. Die Zeit, die zur Verarbeitung der Perzepte im multimodalen *Anchoring*-System benötigt wird, ist vernachlässigbar klein. Zusammengenommen wurden daher die multimodalen *Anchor* mit einer mittleren Rate von 7,7 Hz aktualisiert, die sich aus der asynchronen Verarbeitung der verschiedenen Einzelperzepte ergibt.

**Diskussion**

Das multimodale *Anchoring* hat sich zum Verfolgen von Personen in einem dynamischen Szenario bewährt. Dabei hat sich der unimodale *Anchoring*-Prozess, der für die Beobachtung der Beine zuständig ist, als unverzichtbarer Bestandteil herausgestellt. Dass Personen beim Verfol-

gen vom Roboter mehrfach verloren wurden, ist mit fehlerhaften Zuordnungen von Perzepten zu den bestehenden Personenhypothesen zu erklären. Ein vielversprechender Ansatz, die Zuordnungsfehler zu verringern, besteht darin, die Eigenbewegung des Roboters, die insbesondere bei Rotation zu schnellen Relativbewegungen der verfolgten Person führen, herauszufiltern. Die Fehlversuche im zweiten Experiment sind auch auf die Programmierung des Roboters zurückzuführen, einer Person zu folgen, die sich in einem Abstand von weniger als einem Meter vom Roboter aufhält. Wenn die führende Person verloren ging, während sich eine andere Person im Aufmerksamkeitsradius von einem Meter befand, wechselte der Roboter unwiderruflich seinen Aufmerksamkeitsfokus. Als Abhilfe bietet es sich an, Menschen nicht nur zu detektieren, sondern auch als Individuen zu identifizieren, zum Beispiel anhand ihres Gesichts. In diesem Fall hätte der Roboter den Irrtum selbst erkennen können und die vorausgehende Versuchsperson die Aufmerksamkeit wiedererlangen können.

### 8.2.2 Sprecherlokalisation

In einem zweiten Experiment sollte die Einbindung der Sprecherlokalisation in das multimodale *Anchoring*-Verfahren getestet werden (siehe auch [Lan03]). Wie in Abschnitt 6.4 beschrieben, kann die Position eines Sprechers unter Verwendung eines einzelnen Mikrofonpaars, wie es auf dem Roboter *BIRON* zur Verfügung steht, lediglich auf eine Hälfte eines zweischaligen Hyperboloids eingeschränkt werden. Erst unter Berücksichtigung zusätzlicher Information über potenzielle Positionen lässt sich der Sprecher eindeutig lokalisieren. Diese Information wird aus dem multimodalen *Anchoring* gewonnen. Beine geben dabei relativ genau den Standort der Person an; das Gesicht liefert zusätzliche Information über die Größe.

Um die Leistungsfähigkeit der Sprecherlokalisation mit *BIRON* zu testen, wurde zunächst ein Vorexperiment durchgeführt, bei dem die Genauigkeit des Verfahrens separat getestet wurde. Anschließend fand das Experiment statt, bei dem die Lokalisation des Sprechers im multimodalen Gesamtverfahren untersucht wurde.

#### Vorexperiment

Zunächst wurde die Genauigkeit des Verfahrens zur Sprecherlokalisation unabhängig vom multimodalen *Anchoring*-Verfahren analysiert. Es wurden die Mikrofone auf dem Roboter verwendet, wobei andere Geräte, wie zum Beispiel der Laser-Entfernungsmesser in Betrieb waren, um die Störgeräusche des normalen Betriebs des Roboters mit einzubeziehen. Der Versuchsaufbau wurde so arrangiert, dass sich die Mikrofone auf derselben Höhe wie die Geräuschquelle, das heißt dem Mund des Sprechers, befanden. Damit konnte das in Abschnitt 6.4 erwähnte Modell der vereinfachten Geometrie verwendet werden, um den Einfallswinkel des Schalls zu schätzen.

Jede Versuchsperson wurde nacheinander an zwölf verschiedenen Positionen platziert, die sich aus sechs verschiedenen Winkeln (0°, 10°, 20°, 40°, 60° und 80°) und zwei Abständen (1 m und 2 m) im Roboterkoordinatensystem ergaben (vgl. Abschnitt 3.2). Die Versuchsperson stand immer frontal zum Roboter. An jeder Position las sie einen Satz vor. Dies dauerte pro Satz etwa acht

Tabelle 8.3: Ergebnisse des Vorexperiments zur Sprecherlokalisation

| | | Winkel | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0° | 10° | 20° | 40° | 60° | 80° |
| Abstand | 1 m | –0,9° (0,56) | 9,1° (0,34) | 18,9° (0,21) | 38,2° (0,50) | 57,7° (0,40) | 74,0° (2,62) |
| | 2 m | –0,3° (0,81) | 9,2° (0,37) | 19,3° (0,27) | 38,8° (0,22) | 57,5° (0,64) | 73,3° (2,18) |

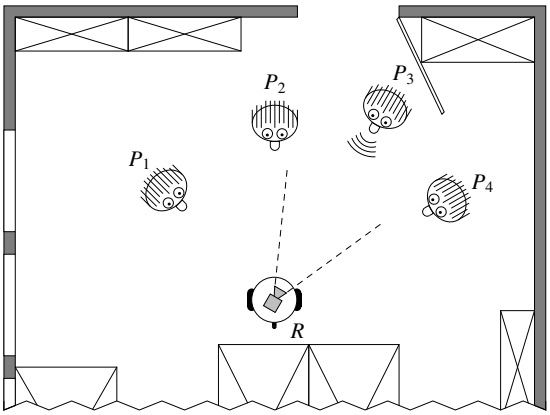
Sekunden. Während des Sprechens wurde mit einer Rate von 20 Hz der Winkel des Sprechers geschätzt. Es wurden Daten von fünf verschiedenen Versuchspersonen aufgezeichnet.

Aus den Ergebnissen des Lokalisationsverfahrens wurden für jede der zwölf Positionen der mittlere Winkel und die Varianz berechnet. Die Werte wurden wiederum über alle Versuchspersonen gemittelt. Tabelle 8.3 fasst alle Ergebnisse zusammen. Zunächst deuten die Messwerte darauf hin, dass der Roboter nicht korrekt ausgerichtet war, da besonders im Bereich kleiner Winkel der gemittelte gemessene Winkel durchgehend um knapp –1° vom Sollwert abweicht. Unter dieser gerechtfertigten Annahme liefert die Sprecherlokalisation im Bereich von 0° bis 40° gute Ergebnisse: Der geschätzte Winkel stimmt mit dem Sollwinkel sehr gut überein, und auch die Varianz ist niedrig. Darüber hinaus arbeitet die akustische Lokalisation für die beiden berücksichtigten Abstände von einem und zwei Metern gleich gut. Bei größeren Winkeln (60° und 80°) nimmt die Genauigkeit deutlich ab. Dies ist möglicherweise ein Effekt der Richtcharakteristik der Mikrofone. Insgesamt liefert das Verfahren zufrieden stellende Ergebnisse.

### Experimenteller Aufbau des Hauptexperiments

Im anschließenden Hauptexperiment wurde die Sprecherlokalisation im Rahmen des multimodalen Gesamtverfahrens getestet. Der zugehörige unimodale *Anchoring*-Prozess war dabei von zentraler Bedeutung. Der *Anchor* für Stimme wurde aufgebaut, wenn der zugehörigen Person zum ersten Mal ein Stimmenperzept zugeordnet werden konnte. Er befand sich sodann im Zustand *grounded*. Konnten der Person für einen Zeitraum von zwei Sekunden keine Stimmenperzepte mehr zugeordnet werden, wurde der unimodale *Anchor* für Stimme aus dem Gesamt-*Anchoring*-Prozess durch die Unterfunktion $\mathrm{IsIrrelevant}(\cdot)$ wieder entfernt. Der Roboter war so programmiert, dass er seine Aufmerksamkeit durch Ausrichten der Kamera immer auf die Person fokussierte, für die ein neuer *Anchor* für Stimme aufgebaut wurde, und seine Aufmerksamkeit so lange aufrecht hielt, bis der entsprechende *Anchor* wieder entfernt wurde. Neben der Sprecherlokalisation wurden im multimodalen *Anchoring* die Beindetektion und die effizientere Variante der Gesichtsdetektion nach Viola und Jones eingesetzt. Der Roboter stand fix, damit keine der Versuchspersonen bei Rotation aus dem Einzugsbereich des Laser-Entfernungsmessers gelangte und so verloren gewesen wäre.

Das Experiment wurde in demselben Büro durchgeführt, in dem auch das erste Experiment zum Folgen von Personen stattgefunden hatte. Vier Versuchspersonen standen um den Roboter herum




(a)

| Schritt / Person | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_1$ | ● | | | | | ● | | | ● | | | |
| $P_2$ | | ● | | | | | ● | | | | | ● |
| $P_3$ | | | ● | | ● | | | | | | ● | |
| $P_4$ | | | | ● | | | | ● | | ● | | |

(b)

Abbildung 8.2: (a) Versuchsanordnung beim Experiment zur Evaluation der Sprecherlokalisation. (b) Reihenfolge der Sprecher. In den Zeitschritten 1–4 und 9–12 sollte der Roboter angesprochen werden, in den Zeitschritten 5–8 eine benachbarte Person.

an vorgegebenen Positionen (siehe auch Abbildung 8.2 (a)). Mit Bezug zum lokalen Koordinatensystem des Roboters befand sich Person $P_1$ in einem Abstand von 1,2 m und bei einem Winkel von 45°, wobei 0° als die Geradeausrichtung des Roboters definiert ist. Person $P_2$ stand an Position (1,4 m, 0°), Person $P_3$ bei Position (1,8 m, -30°) und Person $P_4$ bei Position (1,6 m, -60°). Die Versuchspersonen hatten die Aufgabe, nacheinander für jeweils 10 Sekunden zu sprechen. Sie sollten während dieser Zeit entweder den Roboter oder eine der anderen Versuchspersonen ansprechen, indem sie den Kopf in die entsprechende Richtung drehten. Die Reihenfolge der Sprecher und der jeweilige Adressat waren festgelegt und sind in Abbildung 8.2 (b) angegeben. Es wurden keine Einschränkungen gemacht, wie die Versuchspersonen zu stehen haben. Das Experiment wurde dreimal durchgeführt. Insgesamt haben neun verschiedene Versuchspersonen daran teilgenommen.

### Versuchsergebnisse

Der Roboter war immer in der Lage, innerhalb der jeweils 10 Sekunden langen Abschnitte den Sprecher zu bestimmen und seine Aufmerksamkeit auf die korrekte Person zu fokussieren. In manchen Situationen jedoch wurde entweder der Fokus auf dem vorangegangenen Sprecher zu lange gehalten oder es wurde zwischenzeitlich eine falsche Person selektiert. Ein Diagramm, das den Aufmerksamkeitsfokus des Roboters über die Zeit darstellt, zeigt Abbildung 8.3. Ein fehlerhaftes Verhalten trat in 4 von 36 Zeitschritten auf: In diesen Fällen wechselte der Roboter seine Aufmerksamkeit zu einer Person, die in dem Moment nicht sprach (siehe die roten Pfeile beziehungsweise Spalte 5 in allen drei Experimentdurchläufen und Spalte 4 im letzten Durchlauf in Abbildung 8.3). Es fällt dabei auf, dass in allen Fällen die Person $P_2$, die frontal

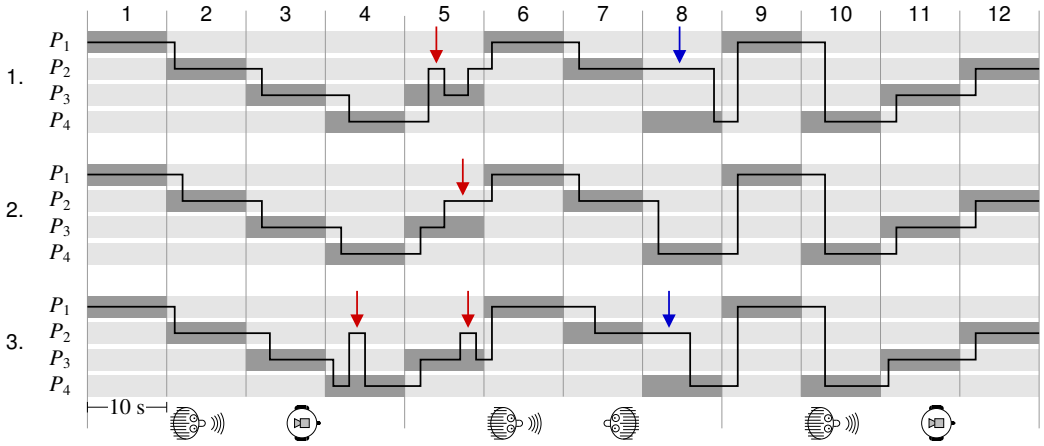



Abbildung 8.3: Diagramm für die drei Durchläufe des Experiments. Jeder Person ist eine hellgraue Spur zugeordnet, die in dem Zeitraum, in dem sie der Sprecher war, dunkel gefärbt ist. Die drei schwarzen Linien geben an, auf wen die Aufmerksamkeit des Roboters gerichtet war. Die roten und blauen Pfeile kennzeichnen fehlerhaftes Verhalten des Roboters.

vor dem Roboter stand, fälschlicherweise als Sprecher ausgewählt wurde. Des Weiteren gab es zwei Aufmerksamkeitswechsel, die im Prinzip richtig waren, jedoch erst mit einer erheblichen Verzögerung stattgefunden haben (blaue Pfeile beziehungsweise achter Zeitschritt im ersten und dritten Experimentdurchlauf). Wiederum war die direkt vor dem Roboter platzierte Person $P_2$ daran beteiligt. Darüber hinaus fällt auf, dass fünf der sechs genannten Fälle während der Phase auftraten, als die Sprecher nicht den Roboter, sondern eine der anderen Personen adressierten. Alle Fehler traten auf, weil eine Geräuschquelle in der Richtung von 0° lokalisiert wurde, obwohl die Person $P_2$ zu dem Zeitpunkt nicht gesprochen hatte. Dieses lässt sich mit den Geräuschen erklären, die der Roboter selbst verursacht, da diese als Geräuschquelle aus der entsprechenden Richtung interpretiert werden.

Wie das Diagramm in Abbildung 8.3 zeigt, hatte jeder korrekte Aufmerksamkeitswechsel eine Verzögerung von ungefähr zwei Sekunden. Dies war aus der Implementierung der Anwendung für dieses Experiment zu erwarten, da, wie oben beschrieben, der *Anchor* für die Stimme nach zwei Sekunden ohne zugeordnetes Sprachperzept aus dem *Anchoring*-Prozess entfernt wurde und erst danach ein Aufmerksamkeitswechsel erfolgen konnte.

Darüber hinaus konnten folgende Werte für den *Anchoring*-Prozess ermittelt werden: Die Gesichtsdetektion wurde auf Bildern der Größe 256×192 mit einer durchschnittlichen Rate von 9,6 Hz durchgeführt und lief damit mehr als dreimal so schnell, wie die weniger effiziente Variante über adaptive Hautfarbensegmentierung in Kombination mit der *Eigenface*-Methode. Die Sprecherlokalisation arbeitet mit einer Taktrate von 5,5 Hz. Beinperzepte standen, ähnlich wie im Experiment zum Folgen einer Person, mit einer Rate von 4,7 Hz zur Verfügung. Der *Anchoring*-Prozess der Person, die jeweils gerade zum Roboter sprach, wurde durchschnittlich durch

Perzepte mit einer Frequenz von 15,4 Hz aktualisiert. Gesichtsperzepte konnten dem entsprechenden *Anchor* zu 71,4% der Zeit zugeordnet werden. Dabei muss bedacht werden, dass es bei einem Aufmerksamkeitswechsel etwa eine Sekunde dauerte, bis die neu selektierte Person in das Blickfeld der Kamera gelangte. Während dieser Zeit konnte für diese kein Gesichtsperzept generiert werden. Stimmenperzepte wurden zu 69,5% der Zeit zugeordnet und Beinperzepte zu 99,9% der Zeit. Das *Anchoring*-Verfahren war in der Lage, die Größe aller Versuchspersonen mit einer Genauigkeit von +/–5 cm zu bestimmen.

#### Diskussion

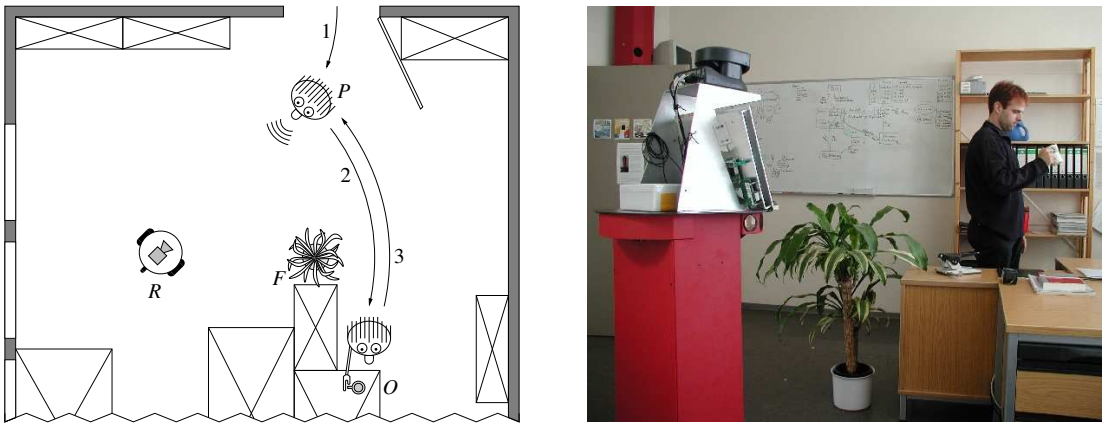
Das Experiment hat gezeigt, dass es auch bei alleiniger Verwendung eines einzelnen Mikrofonpaars möglich ist, über die Integration von Sensordaten verschiedener Modalitäten die Position von Geräuschquellen im Raum zu orten. Die Fehler, die durch die Störgeräusche des Roboters verursacht werden, könnten verringert werden, indem ein Verfahren zur Sprechaktivitätserkennung (engl. *voice activity detection*) eingesetzt würde, welches Sprache von anderen Geräuschen unterscheidet. Ein anderer, jedoch technisch aufwändigerer Ansatz wird von Nakadai und Kollegen [Nak00] verfolgt, bei dem ein zusätzliches Mikrofonpaar, das sich innerhalb der äußeren Hülle des Roboters befindet, die Eigengeräusche aufzeichnet. Diese können dann aus dem anderen Signal eliminiert werden.

### 8.2.3 Verfolgen des Oberkörpers

Im dritten Experiment wurde das komplette, in Kapitel 6 beschriebene multimodale *Anchoring*-System zum Verfolgen von Personen eingesetzt (siehe auch [Fri04]). Im Speziellen sollte die Erweiterung um die Fähigkeit zur Lokalisation des Oberkörpers getestet werden. In der für den Versuch realisierten Anwendung konnte die Aufmerksamkeit des Roboters durch akustische Reize, das heißt durch Ansprechen, gewonnen werden. Der Roboter verfolgte den Benutzer durch Ausrichtung der Kamera und Rotation der Roboterbasis.

#### Experimenteller Aufbau

Der Aufbau des Experiments ist in Abbildung 8.4 dargestellt. Der Roboter $R$ stand auf einer vorgegebenen Position und beobachtete die Tür. Eine Versuchsperson $P$ wurde angewiesen, den Raum durch die Tür zu betreten und durch Ansprechen die Aufmerksamkeit des Roboters zu erlangen. Dann sollte sie zum Schreibtisch gehen, mit einem dort befindlichen Objekt $O$ interagieren und anschließend wieder zum Ausgangspunkt zurückkehren. Der Versuchsaufbau war so gestaltet, dass der Roboter von der Person, nachdem sie die auf dem Boden stehende Pflanze $F$ passiert hatte, nur noch den Oberkörper wahrnehmen konnte: Die Beine waren aus Sicht des Laser-Entfernungsmessers von der Pflanze und dem Schrank verdeckt, das Gesicht war vom Roboter abgewendet und die Versuchsperson hatte nicht gesprochen. Wenn der Roboter in der Lage war, die Person erfolgreich über Oberkörperperzepte zu verfolgen, dann fokussierte er die




(a) (b)

Abbildung 8.4: Aufbau beim Experiment zum Verfolgen des Oberkörpers. (a) Vogelperspektive. (b) Blick aus einer Position hinter dem Roboter.

Versuchsperson auch zum Ende des Experiments. Es wurden 25 Durchläufe mit mehreren Versuchspersonen durchgeführt, die T-Shirts und Pullover verschiedener Farben und Muster trugen.

### Versuchsergebnisse

In 80% der Durchläufe konnte die jeweilige Person erfolgreich während der gesamten Interaktion verfolgt werden. Die Interaktion mit dem Objekt dauerte etwa 5 bis 10 Sekunden. Während dieser Phase standen zum Verfolgen der Person nur Oberkörperperzepte zur Verfügung. In allen Fehlversuchen sind die Personen dem Roboter verloren gegangen, weil die Oberkörperdetektion fehlgeschlagen war.

Die Gesichtsdetektion und die Lokalisation des Oberkörpers wurden auf Bildern der Größe $256 \times 192$ durchgeführt. Für die synchron laufenden Prozesse wurde eine Rate von 5,0 Hz ermittelt. Die Raten bei der Generierung von Stimmen- und Beinperzepten lagen bei 5,5 Hz und 4,7 Hz und bestätigen damit die Werte aus dem Experiment zur Sprecherlokalisation.

### Diskussion

Obwohl die aus den Oberkörperperzepten gewonnene Information das *Tracking* in Situationen, wie sie im Versuch betrachtet wurden, wesentlich verbessert, verlässt sich der *Tracking*-Algorithmus für eine nicht unbeträchtliche Zeitspanne auf nur ein einzelnes Merkmal. Für ein robustes Verfolgen unter einer größeren Variation an Situationen ist es daher erforderlich, mehrere perzeptuelle Merkmale zur Verfügung zu haben. Dies kann durch Erweiterung des bestehenden Systems um zusätzliche Detektoren für zum Beispiel die menschliche Kopf-Schulter-Kontur oder Gesichter in nicht-frontaler Ansicht erreicht werden.

Abbildung 8.5: Versuchspersonen interagieren mit *BIRON* während des ersten Experiments zum Gesamtsystem.

## 8.3 Experimente zum Aufmerksamkeitssystem

Für die Evaluation des Aufmerksamkeitssystems wurden zwei Benutzerexperimente durchgeführt. Da die Leistungsfähigkeit der vollständigen Aufmerksamkeitssteuerung, wie sie für das *Home-Tour*-Szenario entwickelt worden ist, analysiert und beurteilt werden sollte, wurde in den Versuchen jeweils das Gesamtsystem inklusive Sprachverarbeitung, Dialogsteuerung und *Execution Supervisor* eingesetzt. Eine Gestenerkennung und eine Aufmerksamkeitssteuerung für Objekte war nicht vorhanden. Im multimodalen *Anchoring* wurden Bein-, Gesichts- und Stimmenperzepte berücksichtigt.

### 8.3.1 Erstes Benutzerexperiment

Das erste Experiment, bei dem Versuchspersonen das Gesamtsystem getestet haben, ist von Li und Kollegen durchgeführt worden [Li04]. In einer Fragebogenstudie sollten der aktuelle Entwicklungsstand beurteilt, aber auch die Erwartungen und Wünsche der Benutzer an ein zukünftiges Robotersystem für das *Home-Tour*-Szenario analysiert werden. Die Darstellungen in diesem Abschnitt beschränken sich auf die Ergebnisse, die zum einen das Aufmerksamkeitssystem betreffen und zum anderen die Gestaltung des zweiten Benutzerexperiments (Abschnitt 8.3.2) beeinflusst haben.

#### Beschreibung des Experiments

Das Experiment wurde in einem Raum mit einer großen freien Fläche durchgeführt. Dort konnte der Roboter von den Versuchspersonen herumgeführt werden, ohne dass die Gefahr von Zusammenstößen mit Hindernissen bestand. Abbildung 8.5 zeigt einige der Versuchspersonen bei der Interaktion mit *BIRON* während des Experiments. Zur Einweisung wurde den Teilnehmern vom Versuchsleiter das Verhalten des Roboters anhand einer Grafik erläutert. Diese zeigte die

verschiedenen Zustände der Aufmerksamkeitssteuerung und Beispiele für Anweisungen, mit denen Zustandswechsel erreicht werden können. Den Zettel mit der Grafik durften die Versuchspersonen auch während des Experiments zur Hilfe nehmen. Jede Versuchsperson wurde angewiesen, in einer Interaktion mit dem Roboter über sprachliche Anweisungen die verschiedenen Aufmerksamkeitszustände in einer vorgegebenen Reihenfolge zu durchlaufen: Zunächst sollte der Roboter durch Ansprechen aus dem Schlafmodus geweckt und in den Interaktionsmodus gebracht werden. Dann war es die Aufgabe, den Roboter im Folgen-Modus zu einer anderen Stelle im Raum zu führen. Als nächstes war die Person angewiesen, dem Roboter ein Objekt unter Verwendung von Sprache und Gestik zu zeigen. Abschließend sollte die Interaktion mit *BIRON* durch eine Verabschiedung beendet werden. Die Reihenfolge der Aufmerksamkeitszustände in der Interaktionsphase war folglich: *AS:Person – AS:Follow – AS:Person – AS:Show – AS:Object – AS:Person.* Im Anschluss an die Interaktion wurden die Teilnehmer gebeten, einen Fragebogen auszufüllen.

An dem Versuch haben insgesamt 21 Personen im Alter von 22 bis 54 Jahren teilgenommen. Die meisten Teilnehmer verfügten über ein ausgeprägtes technisches Hintergrundwissen. Alle Versuchspersonen haben die vorgegebene Aufgabe erfolgreich bewältigt. Sie benötigten dazu zwischen drei und fünf Minuten.

### Ergebnisse

In dem Fragebogen sollten die Benutzer verschiedene Aspekte der Interaktion mit dem Roboter bewerten. Einerseits sollte herausgefunden werden, welche der Fähigkeiten des Roboters die Benutzer beeindruckt hat. Zu diesem Zweck diente die Frage *„Welche Fähigkeiten des Roboters fanden Sie besonders interessant?“*, bei der Antwortmöglichkeiten vorgegeben und Mehrfachnennungen zulässig waren. Andererseits sollten die Eigenschaften benannt werden, die als unerfreulich empfunden wurden. Dazu wurde die Frage *„Was hat Ihnen an BIRON nicht gefallen?“* gestellt, bei der die Antwortmöglichkeiten offen gelassen waren. Die Ergebnisse sind in der Abbildung 8.6 dargestellt.

Es stellte sich heraus, dass die Aufmerksamkeitssteuerung die meisten positiven Rückmeldungen der Versuchspersonen erhielt. Elf der 21 Personen hielten die Fähigkeit des Roboters, seinen Kommunikationspartner zu fokussieren, für interessant und sieben waren vom Folgen durch Hinterherfahren beeindruckt. Von vier Probanden wurde auch das Aufwecken des Roboters als interessant bewertet. Keiner der Befragten äußerte bei der zweiten Frage Unzufriedenheit über die Leistungsfähigkeit des Aufmerksamkeitssystems. Offensichtlich verhielt sich der Roboter entsprechend der Erwartungen der Benutzer an das System.

Neben der Aufmerksamkeitssteuerung war für die Benutzer die automatische Sprachverarbeitung ein wichtiger Aspekt. Diese fand jedoch, im Gegensatz zur Aufmerksamkeitssteuerung, ein geteiltes Echo. Acht Personen bewerteten die Interaktion mit dem Roboter über natürliche Sprache interessant. Gleichzeitig bemängelten zwölf Probanden die hohe Fehlerrate bei der Spracherkennung. Vier Benutzer beklagten einen zu unflexiblen Dialog, der sie zu sehr in der Wahl der Worte einschränkte. Offensichtlich hat die natürlichsprachliche Interaktion, neben der Fähigkeit des Ro-

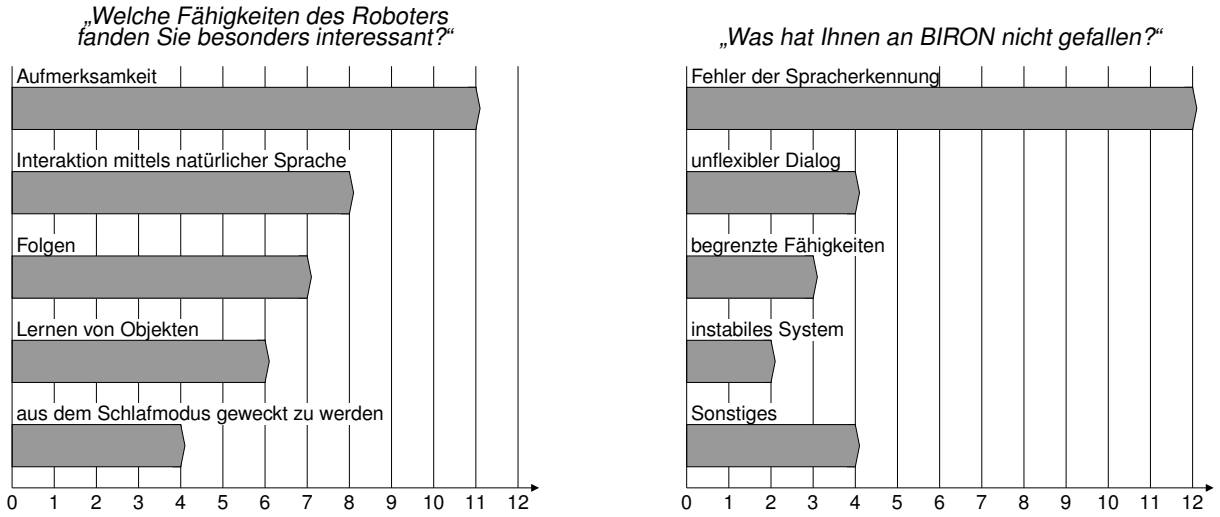



Abbildung 8.6: Histogramme der Antworten von 21 Versuchspersonen. Bei beiden Fragen waren Mehrfachnennungen erlaubt.

boters zur Aufmerksamkeit, einen wesentlichen Anteil an der Natürlichkeit der Benutzerschnittstelle und der Attraktivität des Gesamtsystems. Das heißt zugleich, dass die Leistungsfähigkeit der sprachverarbeitenden Komponente entscheidend für die Akzeptanz des Systems ist.

Ein weiterer Aspekt, der in der Studie untersucht werden sollte, war die Frage, wie nützlich die Ausgaben von Berechnungsergebnissen einzelner Prozesse und von internen Zuständen des Roboters für den Benutzer sind. Zu diesem Zweck wurden sowohl Ergebnisse der automatischen Spracherkennung als auch der Zustand der Aufmerksamkeitssteuerung angezeigt. Im Allgemeinen fanden die Versuchspersonen Rückmeldungen sehr hilfreich. Jedoch haben die Benutzer sehr individuelle Ansichten darüber, welche Art sie als geeignet empfinden. Während manche die Ergebnisse der Spracherkennung sehen wollten, fanden andere diese Angaben zu technisch und zu ablenkend von der eigentlichen Aufgabe. Die Rückmeldung über den Zustand der Aufmerksamkeitssteuerung wurde dagegen im Allgemeinen als hilfreich empfunden. Die Kenntnis über den internen Zustand des Roboters scheint nützlich zu sein, jedoch muss dieser in angemessener Weise dargestellt werden.

### 8.3.2 Zweites Benutzerexperiment

Mit der Durchführung des zweiten Benutzerexperiments wurden zwei Absichten verfolgt. Zum einen war es das Ziel, die Leistungsfähigkeit der auditiven Aufmerksamkeit festzustellen, wobei allein der Aspekt der Aktivierung der Sprachverarbeitung betrachtet wurde (Abschnitt 7.7.2). Zum anderen sollte das Gesamtsystem unter leicht veränderter Aufgabenstellung erneut von den Versuchspersonen beurteilt werden.

Als Folge der Erkenntnisse aus dem ersten Experiment, dass die Benutzer zum einen sehr häufig

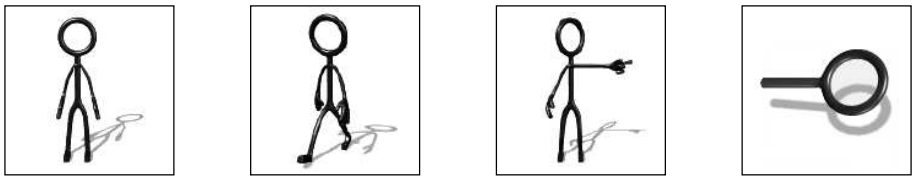



Abbildung 8.7: Verwendete Symbole zur Darstellung der Aufmerksamkeitszustände in der Interaktionsphase

Fehler bei der automatischen Spracherkennung bemängelten und zum anderen die Darstellung des internen Zustands des Roboters als hilfreich empfanden, wurden folgende Modifikationen am Gesamtsystem vorgenommen:

- Die Aufzeichnung der Daten für die Sprachverarbeitung erfolgte über ein Nahbesprechungsmikrofon, wodurch die Wortfehlerrate deutlich gesenkt wurde.
- In der oberen Ecke des Flachbildschirms vom Roboter wurde, als Ergänzung zum animierten Gesicht, der Zustand der Aufmerksamkeitssteuerung in der Interaktionsphase in Form eines Strichmännchen-Symbols dargestellt (siehe Abbildung 8.7).

**Beschreibung des Experiments**

Das Experiment wurde wiederum in einem Raum mit einer größeren freien Fläche durchgeführt. Zu Beginn jedes Testdurchlaufs stand der Roboter auf einer vorgegebenen Position, etwa drei Meter von einem Tisch entfernt, auf dem sich diverse Objekte befanden, unter anderem eine Tasse, Stifte, eine Pflanze und ein Telefon. Die Versuchspersonen wurden aufgefordert, den Roboter aus dem Schlafzustand aufzuwecken, durch Begrüßung in die Interaktionsphase zu gelangen, dann dem Roboter einige der Objekte zu zeigen und sich anschließend von ihm zu verabschieden. Im Gegensatz zum ersten Experiment gab der Versuchsleiter jedoch keine Erläuterung über die verschiedenen Aufmerksamkeitszustände, die der Roboter im Verlauf der Interaktion einnehmen kann. Vielmehr stand es den Versuchspersonen in der Interaktionsphase frei, wie und in welcher Reihenfolge sie die Fähigkeiten des Roboters nutzten. Der Versuchsaufbau erforderte es jedoch, dass die Probanden wenigstens einmal den Roboter baten, ihnen zu folgen, um ihn nah genug an den Tisch mit den zu lernenden Objekten zu führen. Wurde der Roboter dabei auf direkter Strecke zum Tisch geleitet, befand sich der Tisch mit den Objekten aus Sicht des Roboters schräg rechts. Eine Zeige- oder Objekterkennung war nicht vorhanden. Der Roboter bewegte die Kamera nach jeder Anweisung zum Lernen eines Objekts in eine vorher festgelegte Richtung und zwar nach rechts unten. Stand der Roboter wie oben beschrieben vor dem Tisch, drehte sich die Kamera in Richtung der Objekte. Einige Versuchspersonen haben den Roboter jedoch über längere Strecken hinter sich herfahren lassen, sodass der Roboter dann zum Teil in einer anderen

Position zu den Objekten zum Stehen kam und die Kamera dann nicht in Richtung des Tischs schwenkte.

Ein Teil der Versuchspersonen hatte nie zuvor mit Robotern oder Spracherkennungssystemen gearbeitet. Die Erfahrung zeigt, dass solche Personen anfänglich einige Schwierigkeiten haben, mit dem Roboter zu interagieren. Jedoch kann die Effizienz der Interaktion durch eine kurze Eingewöhnungsphase erheblich gesteigert werden. Aus diesem Grund durfte jede Versuchsperson vor der eigentlichen Aufzeichnung der Daten einen Testlauf unternehmen. Während dieser Phase gaben die Versuchsleiter Hinweise, wie der Erfolg der Interaktion verbessert werden kann. Das Training erstreckte sich über einen ähnlichen Zeitraum wie das anschließende Experiment. Während der Experimentphase wurden keine Hinweise mehr gegeben.

Nachdem die Versuchspersonen mit dem Roboter interagiert hatten, wurden sie gebeten, einen Fragebogen auszufüllen. An dem Experiment haben zwölf Versuchspersonen im Alter zwischen 21 und 31 Jahren teilgenommen. Es handelte sich um acht Studenten und vier wissenschaftliche Mitarbeiter der Universität Bielefeld verschiedener Fachrichtungen.

### Ergebnisse

Die Interaktionsdauer lag zwischen zwei und sechseinhalb Minuten. Jede Versuchsperson hatte während dieser Zeit dem Roboter erfolgreich mindestens drei verschiedene Gegenstände gezeigt. In jedem Experimentdurchlauf wurde das Folgen des Roboters genutzt.

Als erstes werden die durch den Fragebogen ermittelten Bewertungen der Versuchsteilnehmer vorgestellt. Um einen Vergleich zum ersten Benutzerexperiment zu ermöglichen, wurden die zwei dort gestellten Fragen unverändert übernommen. Abbildung 8.8 zeigt die Histogramme der Antworten. Die Antworten auf die Frage, welche Fähigkeiten des Roboters die Benutzer besonders interessant fanden, zeigen eine ähnliche Verteilung, wie in der ersten Studie. Wieder wurde die Aufmerksamkeit am häufigsten genannt. Neun der zwölf Versuchspersonen waren von der Fähigkeit des Roboters beeindruckt, sich auf den Benutzer zu fokussieren. Genauso häufig wurde das Folgen als interessanteste Fähigkeit genannt. Die Interaktion mittels natürlicher Sprache fanden sieben von zwölf Versuchspersonen interessant. Fünf Probanden bewerteten auch das Lernen von Objekten als positive Eigenschaft.

Auf die offene Frage, was den Versuchspersonen an *BIRON* nicht gefallen habe, zeichnete sich keine Antwort zahlenmäßig gegenüber anderen Antworten so klar ab, wie es in der ersten Studie der Fall war. Dort bemängelten viele Teilnehmer Fehler in der Spracherkennung. Unter Verwendung des Nahbesprechungsmikrofons konnten in diesem Experiment die Erkennungsfehler offensichtlich so weit gesenkt werden, dass keine der Versuchspersonen dieses Problem als erwähnenswert erachtete. Mehrfach wurde dagegen das äußere Erscheinungsbild als negativ empfunden, wobei zwei Versuchspersonen ein menschlicheres Aussehen wünschten und zwei weitere explizit die grafische Darstellung des Gesichts bemängelten. Vier Versuchspersonen erwähnten die undeutliche Sprachausgabe. Von drei Teilnehmern wurde die langsame Reaktion des Systems genannt, die ihnen am Gesamtsystem nicht gefallen hatte. Genauso wie im ersten Experiment

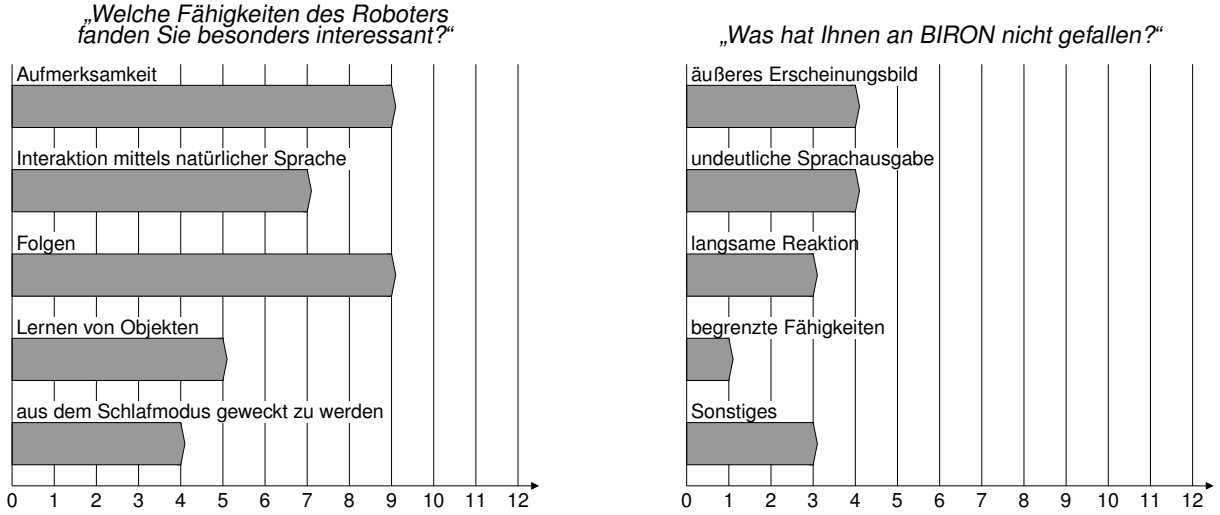



Abbildung 8.8: Histogramme der Antworten von zwölf Versuchspersonen. Bei beiden Fragen waren Mehrfachnennungen erlaubt.

äußerte keine der Versuchspersonen Unzufriedenheit über die Leistungsfähigkeit des Aufmerksamkeitssystems.

In einer weiteren offenen Frage wurden die Versuchspersonen gefragt, welche Fähigkeiten sie bei der Interaktion mit *BIRON* vermissten. Hier drehte sich die überwiegende Anzahl der Antworten um das Lernen von Objekten. Offensichtlich wünschten sich die Versuchspersonen eine Rückmeldung, die erkennen lässt, dass der Roboter tatsächlich das vom Benutzer gezeigte Objekt erfasst und in seinen Wissensspeicher aufgenommen hatte. Es wurde vermisst, dass der Roboter das genannte Objekt mit der Kamera korrekt fokussiert. Mehrere Teilnehmer wünschten sich, dass der Roboter die Objekte auch noch zu einem späteren Zeitpunkt der Interaktion wiederfinden und benennen konnte.

Zusätzlich sollten die Versuchspersonen einige Aussagen auf einer vierstelligen Skala bewerten. Die äußeren Punkte der Skala entsprachen den Antworten *„trifft voll zu“* und *„trifft gar nicht zu“*. Die Resultate sind in Abbildung 8.9 dargestellt. Demnach wurde die Interaktion mit dem Roboter überwiegend als einfach empfunden. Kritischer fiel die Beurteilung der Natürlichkeit der Interaktion aus. Keiner fand die Interaktion uneingeschränkt natürlich und fünf bewertetenn sie als wenig oder gar nicht natürlich. Die Aussagen, die sich auf die Aufmerksamkeit beziehen, wurden überwiegend positiv bewertet. So sagte die Mehrheit der Versuchsteilnehmer, dass es einfach war, die Aufmerksamkeit des Roboters zu erlangen. Auch beurteilten fast alle den Roboter als aufmerksam. Das animierte Gesicht konnte die Interaktion nur bei fünf der zwölf Teilnehmer in gewissem Maß erleichtern. Besser geeignet war offensichtlich das Strichmännchen-Symbol, das den internen Roboterzustand darstellte. Die Mehrheit empfand es als eindeutig hilfreich.

Der folgende Text beschäftigt sich mit der Fragestellung, wie leistungsfähig die Aktivierung der Sprachverarbeitung durch das Aufmerksamkeitssystem in dem Experiment war. Die Aktivierung hängt von einer erfolgreichen Erkennung der Attribute `schaut` und `spricht` ab. Da

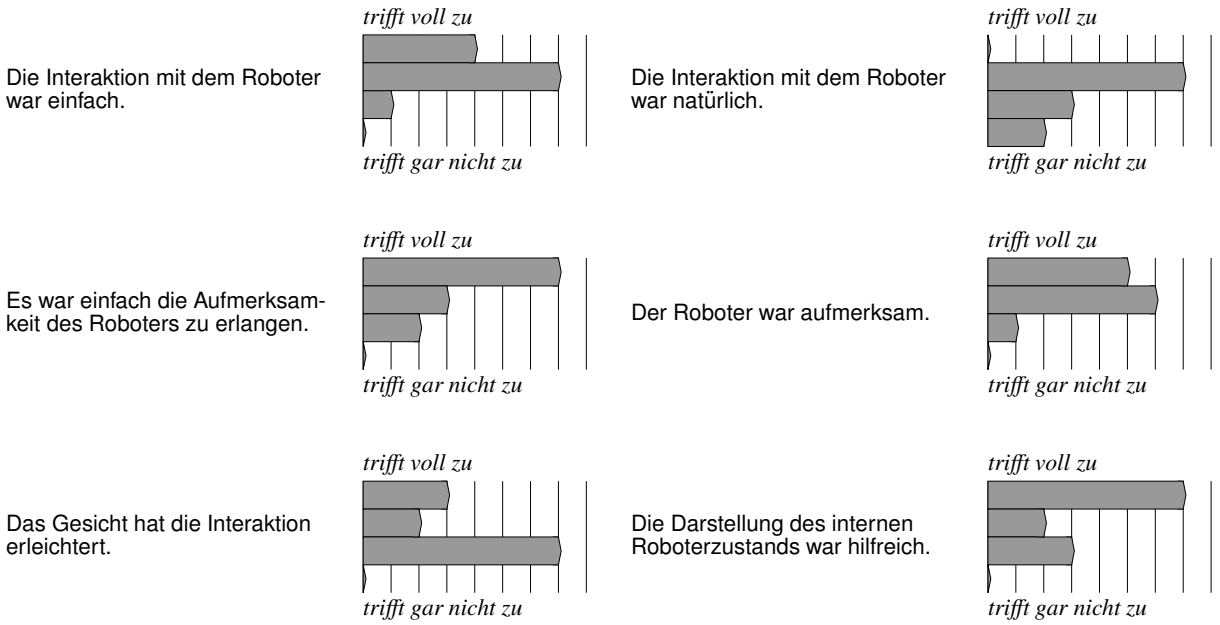



Abbildung 8.9: Histogramme zu den Bewertungen durch die zwölf Versuchspersonen.

sich darauf auch die Relevanzbewertung von Personen in der Bereitschaftsphase stützt, geben die Ergebnisse zusätzlich einen Hinweis auf die Leistungsfähigkeit der Aufmerksamkeitssteuerung in der Bereitschaftsphase.

Um die einzelnen Erkennungsraten quantifizieren zu können, wurden während des Experiments Informationen über die im multimodalen *Anchoring* verfolgten Personenhypothesen, die generierten und zugeordneten Perzepte und die Aktivierungsvorgänge der Sprachverarbeitung durch die auditive Aufmerksamkeit aufgezeichnet. Um die tatsächlich durchgeführten Aktivierungsvorgänge mit den erforderlichen vergleichen zu können, wurde ebenfalls der Beginn und das Ende von Äußerungen, die von den Probanden an den Roboter gerichtet wurden, vom Versuchsleiter per Tastendruck aufgezeichnet. Dabei ist zu beachten, dass die als erforderlich erachteten Aktivierungsvorgänge der subjektiven Meinung einer beobachtenden Person entsprechen, und damit fehleranfällig und ungenau sein können.

Abbildung 8.10 zeigt einen repräsentativen Ausschnitt der aufgezeichneten Daten aus dem Verlauf der Interaktion einer Versuchsperson mit dem Roboter. Auf der Abszisse ist die Zeit in Sekunden aufgetragen. Die blauen vertikalen Linien geben die Wechsel des Aufmerksamkeitssystems in einen neuen Zustand an. So erfolgt der erste Wechsel bei ungefähr drei Sekunden zum Zustand *AS:Alert*. Die Spracherkennungsergebnisse sind unten eingetragen, wobei die gestrichelte Linie den Zeitpunkt des Endes der Verarbeitung markiert. Die grüne Fläche im unteren Viertel der Grafik ist immer dann eingetragen, wenn der Roboter für die Person im Fokus der Aufmerksamkeit das Attribut `schaut` ermitteln konnte. Das Attribut `spricht` ist entsprechend durch die blauen Balken dargestellt. Die gelben Balken im dritten Viertel von unten geben

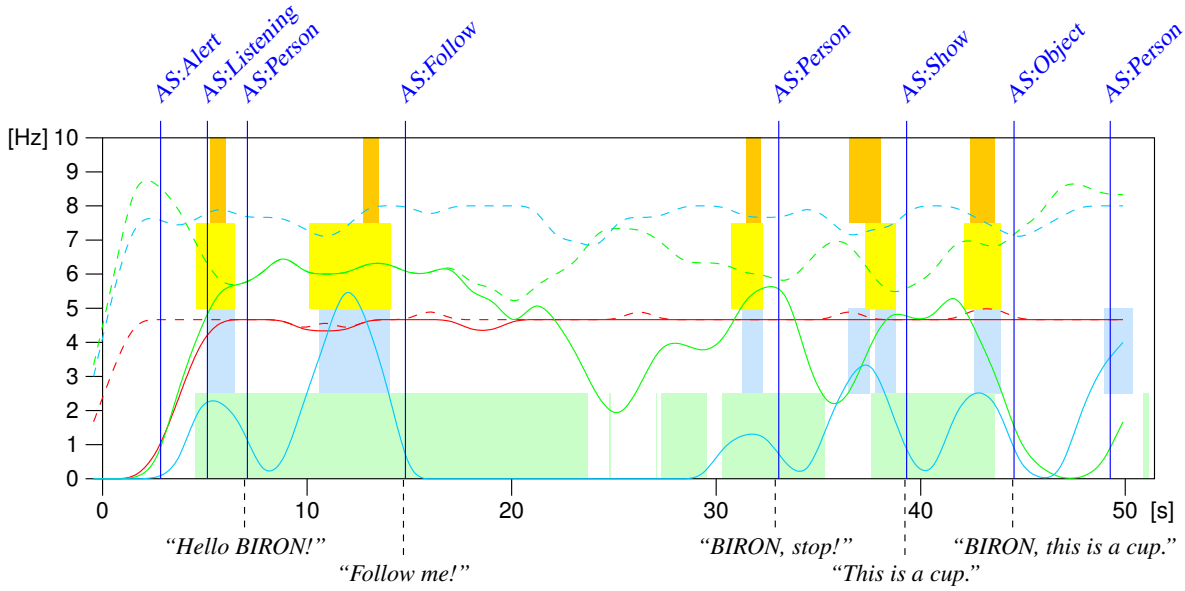



Abbildung 8.10: Aufgezeichnete Daten aus dem Verlauf der Interaktion einer Versuchsperson mit dem Roboter. Die Erläuterung der Grafik erfolgt im Text.

die Zeitbereiche an, zu denen die Sprachverarbeitung aktiviert war. Die Aktivierung geht in diesem Beispiel immer mit dem Auftreten des Attributs `spricht` einher. Es ist zu erkennen, dass die Aktivierung jeweils um eine halbe Sekunde in die Vergangenheit verschoben wurde (vgl. Abschnitt 7.7.2, Seite 123). Die vom Versuchsleiter als zu aktivieren erforderlich bewerteten Zeitspannen sind durch die orangefarbenen Balken dargestellt. Es zeigt sich, dass im dargestellten Ausschnitt die auditive Aufmerksamkeit (gelb) gut die Zeitbereiche der tatsächlichen sprachlichen Äußerungen (orange) abdeckt. Bei 36 Sekunden wurde der Anfang der Äußerung nicht erfasst, da das Gesicht der Versuchsperson nicht detektiert wurde. Dennoch konnte die Sprachverarbeitung in diesem Fall die wesentliche Aussage mit *„this is a cup“* extrahieren.

Auf der Ordinate sind die Taktraten für die Generierung und Zuordnung von Perzepten aufgetragen. Die eingezeichneten Kurven sind Mittelwerte über ein gleitendes Fenster von drei Sekunden Breite. Die gestrichelten Linien geben an, mit welcher Taktrate die jeweiligen Sensorsysteme gearbeitet haben. Die durchgezogenen Linien geben die Frequenz der Zuordnung von Perzepten zur Hypothese für die Versuchsperson an. Rot kennzeichnet die Linien für die Beinerkennung, Grün für die Gesichtsdetektion und Blau für die Sprecherlokalisation. Zu Beginn der Interaktion fällt die grüne durchgezogene Linie mit der gestrichelten zusammen, das heißt das Gesicht der Person konnte durchgehend erkannt werden. Erst bei 22 Sekunden wird das Gesicht seltener detektiert. Hier befand sich der Roboter im Zustand *AS:Follow*. Die vorausgehende Person hatte zeitweise das Gesicht abgewandt. Es ist deutlich zu erkennen, wie hohe Zuordnungsraten von Gesichts- und Stimmenperzepten mit erfolgreicher Bestimmung der Attribute `schaut` und `spricht` einhergeht.

Tabelle 8.4: Erkennungsraten der auditiven Aufmerksamkeit für die Aktivierung der Sprachverarbeitung auf zeitlicher Ebene.

| Nr. | korrekt (%) | falsch-positiv (%) | falsch-negativ (%) | Person sprach (%) | Verarbeitung aktiviert (%) |
|---|---|---|---|---|---|
| 1 | 78,1 | 19,3 | 2,6 | 22,6 | 39,1 |
| 2 | 69,9 | 29,3 | 0,8 | 12,7 | 41,2 |
| 3 | 76,7 | 21,3 | 1,9 | 14,5 | 33,9 |
| 4 | 82,3 | 14,1 | 3,6 | 14,5 | 25,0 |
| 5 | 82,2 | 15,9 | 2,0 | 11,2 | 25,1 |
| 6 | 75,4 | 23,0 | 1,6 | 15,3 | 36,7 |
| 7 | 74,5 | 19,5 | 6,0 | 22,5 | 36,0 |
| 8 | 79,0 | 16,9 | 4,1 | 13,2 | 26,0 |
| 9 | 83,7 | 12,7 | 3,6 | 13,2 | 22,3 |
| 10 | 80,2 | 17,1 | 2,7 | 16,5 | 30,9 |
| 11 | 80,4 | 16,9 | 2,7 | 7,0 | 21,2 |
| 12 | 77,6 | 21,2 | 1,2 | 12,6 | 32,6 |
| ∅ | 78,3 | 18,9 | 2,7 | 14,7 | 30,8 |

Für eine quantitative Analyse wurden die Erkennungsraten der auditiven Aufmerksamkeit für die Aktivierung der Sprachverarbeitung auf zeitlicher Ebene berechnet. Die Ergebnisse für alle Versuchspersonen sind in Tabelle 8.4 zusammengefasst. Die Aktivierung erfolgte über 78,3% der Zeit korrekt. Bei diesem Wert muss jedoch berücksichtigt werden, dass der Zeitanteil, zu dem gesprochen wurde, im Mittel nur bei 14,7% lag (vorletzte Spalte), und die Nichtaktivierung, wenn nicht gesprochen wird, erheblich leichter zu entscheiden ist, als die Aktivierung, wenn gesprochen wird. Eine größere Aussage über die Leistungsfähigkeit haben die Raten, mit denen das System falsch entschieden hat. Sie fallen sehr unterschiedlich aus: Im Durchschnitt wurde die Sprachverarbeitung für 18,9% der Zeit fälschlicherweise aktiviert, aber nur für 2,7% fälschlicherweise nicht aktiviert. Das bedeutet, dass die Aufmerksamkeitssteuerung im Vergleich zur Beurteilung des beobachtenden Versuchsleiters zu sensibel ausfällt (vgl. auch Abbildung 8.10). Da die Falsch-Positiv-Rate jedoch sehr gering ist, besteht selten die Gefahr, dass relevante Sprachdaten unberücksichtigt bleiben. Die Sprachverarbeitung war während des Experiments im Schnitt nur zu 30,8% der Zeit aktiviert. Mit dem Verfahren konnte folglich die für die Sprachanalyse erforderliche Rechenleistung deutlich reduziert werden.

Des Weiteren wurden die Fehlerursachen analysiert, die bei der Aktivierung der Sprachverarbeitung aufgetreten sind. Dies geschah auf der Ebene von einzelnen Instruktionen. Wie in Abschnitt 7.7.2 beschrieben ist, erfolgt die Aktivierung, wenn die beiden Attribute `schaut` und `spricht` für eine Person erkannt wurden. Sie bleibt solange aktiviert, bis das Attribut `spricht` nicht mehr vorliegt. Fehler treten entsprechend bei Ausbleiben der Attribute auf. Ein Beispiel ist in Abbildung 8.10 bei 36 Sekunden nach Beginn der Aufzeichnung zu sehen. Hier wurde die Sprachverarbeitung zu spät aktiviert, weil das Attribut `schaut` aufgrund der geringen Anzahl zugeordneter Gesichtsperzepte nicht ermittelt wurde. Die Aktivierung für eine Instruk-

Tabelle 8.5: Fehler bei der Aktivierung der Sprachverarbeitung auf Ebene einzelner Instruktionen.

| | Instruktionen | | Fehlerursache | | |
|---|---|---|---|---|---|
| Nr. | gesamt | korrekt | `schaut` | `spricht` | beides |
| 1 | 49 | 45 | 0 | 4 | 0 |
| 2 | 21 | 18 | 0 | 3 | 0 |
| 3 | 26 | 22 | 0 | 4 | 0 |
| 4 | 50 | 38 | 1 | 8 | 3 |
| 5 | 24 | 18 | 3 | 3 | 0 |
| 6 | 21 | 16 | 3 | 2 | 0 |
| 7 | 22 | 13 | 5 | 0 | 4 |
| 8 | 26 | 17 | 5 | 0 | 4 |
| 9 | 27 | 17 | 7 | 3 | 0 |
| 10 | 17 | 13 | 2 | 2 | 0 |
| 11 | 30 | 21 | 5 | 0 | 4 |
| 12 | 13 | 13 | 0 | 0 | 0 |
| total | 326 | 251 (77,0%) | 31 (9,5%) | 29 (8,9%) | 15 (4,6%) |

tion wurde als Fehler bewertet, wenn die Zeitspanne der Instruktion nicht vollständig innerhalb der Aktivierungszeitspanne liegt, selbst dann, wenn, wie in dem Beispiel, die Anweisung der Versuchsperson zu einem erfolgreichen Wechsel des Aufmerksamkeitszustands führte. Die Ergebnisse der Analyse sind in Tabelle 8.5 zusammengefasst. Von insgesamt 326 Instruktionen der zwölf Versuchsteilnehmer wurden 77% vollständig korrekt erkannt. In 9,5% der Fälle war das nicht vorliegende Attribut `schaut` die Fehlerursache, und in 8,9% der Fälle das nicht vorliegende Attribut `spricht`. Bei weiteren 4,6% der Instruktionen führte das Ausbleiben beider Attribute zum Fehler. Wenngleich die Raten der Fehlerursachen für die beiden Attribute sehr ähnlich sind, muss beachtet werden, dass das Attribut `schaut` nur beim Einschaltvorgang einen Fehler verursachen kann, wobei das Ausbleiben des Attributs `spricht` während der gesamten Instruktionsdauer zu einem Fehler führt. Folglich ist die Rate des Attributs `schaut` kritischer einzustufen als die des Attributs `spricht`.

## 8.4 Zusammenfassung

In diesem Kapitel wurden die Ergebnisse von fünf Experimenten präsentiert, die im Rahmen der Evaluation des entwickelten Aufmerksamkeitssystems durchgeführt wurden.

Die Versuche zum multimodalen *Anchoring* wurden in verschiedenen Entwicklungsphasen des Systems durchgeführt, was sich durch die unterschiedliche Anzahl von zur Verfügung stehenden *Anchoring*-Prozessen ausdrückte. Alle Experimente wurden mit Versuchspersonen in anwendungsnahen Szenarien durchgeführt. Beim Verfolgen einer vorausgehenden Person durch den

Roboter auf der Basis von Bein- und Gesichtsperzepten konnte die Effizienz und Effektivität des Verfahrens in einem dynamischen Szenario demonstriert werden. In einem zweiten Versuch wurde die neu hinzugekommene Sprecherlokalisation getestet. Der Roboter war gut in der Lage, sich auf den jeweiligen Sprecher zu richten. Vereinzeltes Fehlverhalten wurde offensichtlich durch Eigengeräusche verursacht, die der Roboter auf 0° ortete und damit der direkt vor ihm stehenden Person zuordnete. Im dritten Experiment wurde das *Tracking*-Verfahren unter Verwendung aller vier *Anchoring*-Prozesse getestet. In einem speziellen Szenario, in dem eine Person über eine längere Phase nur über Oberkörperperzepte verfolgt werden konnte, waren 80% der Testläufe erfolgreich. In allen Fällen hat sich das multimodale *Anchoring* als modulares und effizientes *Tracking*-Verfahren erwiesen.

Das Aufmerksamkeitssystem wurde schließlich im Rahmen des Gesamt-Interaktionssystems untersucht. Versuchspersonen sollten dabei die Fähigkeiten des Roboters im *Home-Tour*-Szenario testen und bewerten. Einmal war es die Aufgabe, alle Aufmerksamkeitszustände der Interaktionsphase zu durchlaufen. Ein anderes Mal war die Vorgabe weniger technisch: Den Versuchspersonen wurden lediglich die Fähigkeiten des Roboters erläutert. In Fragebögenstudien wurde die Aufmerksamkeit als interessanteste Fähigkeit des Roboters bewertet. Auch fanden die Probanden es einfach, die Aufmerksamkeit des Roboters zu erlangen. Die meisten waren der Meinung, dass der Roboter aufmerksam war. Die Aufmerksamkeitssteuerung wurde von den Versuchspersonen sehr positiv wahrgenommen. Darüber hinaus konnte die Effektivität der auditiven Aufmerksamkeit gezeigt werden, die durch ihren Einsatz den von der Sprachverarbeitung zu verarbeitenden Anteil auf 30,8% reduzierte. Das in dieser Arbeit entwickelte Konzept für eine multimodale Aufmerksamkeit konnte sich im praktischen Einsatz erfolgreich bewähren.

# Kapitel 9

# Zusammenfassung

Möglicherweise haben wir alle in Zukunft einen „*BIRON*“ bei uns zu Hause! Ob jedoch solche Serviceroboter tatsächlich einmal einen selbstverständlichen Bestandteil unseres täglichen Lebens darstellen, hängt sehr von ihrer Akzeptanz durch die zukünftigen Nutzer ab. Einen wichtigen Beitrag dazu leistet eine ergonomische Schnittstelle, die es dem Anwender erlaubt, in einer dem Menschen vertrauten Weise mit dem Roboter zu kommunizieren. Eine natürlich gestaltete Interaktion gestattet es dem Benutzer zudem, sich frei vor dem Roboter zu bewegen. Daraus folgt jedoch, dass sich mehrere Personen gleichzeitig in der Nähe des Roboters aufhalten können, die für ihn alle potenzielle Kommunikationspartner darstellen. Er muss folglich erkennen können, wann jemand zu ihm spricht und seine Aufmerksamkeit auf ihn gerichtet hat. Hat er einen Benutzer erkannt, muss der Roboter auch seine Aufmerksamkeit auf den Benutzer richten, um ihn optimal wahrzunehmen und ihm zu zeigen, dass er ihn registriert hat. Aufmerksamkeit spielt folglich eine zentrale Rolle in einer natürlich gestalteten Interaktion. Das Ziel dieser Arbeit war es daher, ein Aufmerksamkeitssystem für einen mobilen Roboter zu entwickeln, das ihm die Fähigkeit zu einer effektiven und natürlichen Interaktion mit Menschen verleiht. Als Anwendungsbeispiel wurde das *Home-Tour*-Szenario gewählt, bei dem es die Idee ist, dass ein Benutzer seinen neu gekauften Roboter durch sein Haus führt, um ihm Objekte und Räume zu zeigen, die für spätere Aufgaben relevant sind.

Die Grundlage für eine personenbasierte Aufmerksamkeitssteuerung ist die Fähigkeit des Roboters, Menschen in seiner Nähe zu detektieren und über den Verlauf der Zeit zu verfolgen. Um diese Voraussetzung zu schaffen, ist im Rahmen dieser Arbeit ein Verfahren zum Verfolgen von Personen entwickelt worden. Das Verfahren baut auf dem *Anchoring*-Ansatz von Coradeschi und Saffiotti auf [Cor00]. *Anchoring* stellt ein *Tracking*-Konzept dar, welches das Verfolgen eines Objekts mit einem einzelnen Sensor gestattet. Für die Aufmerksamkeitssteuerung sind jedoch multimodale Informationen wichtig (Wer spricht? Wer schaut?), die mit mehreren verschiedenen Sensoren erfasst werden müssen. Daher wurde der *Anchoring*-Ansatz so erweitert, dass Daten von verschiedenen Sensoren im *Tracking*-Prozess integriert werden können. Der neu entwickelte Ansatz wird entsprechend als multimodales *Anchoring* bezeichnet.

Im multimodalen *Anchoring* ergibt sich die Situation, dass die Sensoren potenziell verschiedene

Bestandteile oder Ausschnitte desselben Objekts erfassen, die einander im *Tracking*-Prozess zugeordnet werden müssen. Zu diesem Zweck wurden Objektmodelle eingeführt: Ein Kompositionsmodell beschreibt die räumliche Konstellation der mit den jeweiligen Sensoren beobachteten Bestandteile. Ein Bewegungsmodell definiert die zulässigen Änderungen von Konstellationen über die Zeit. Ein Fusionsmodell gibt schließlich an, wie neue Messungen der einzelnen Sensoren zu integrieren sind, um eine neue Schätzung der aktuellen Konstellation des beobachteten Objekts zu bekommen. Multimodales *Anchoring* erlaubt es die Daten asynchron zu verarbeiten. Es bietet zudem die Möglichkeit, eine variable Anzahl von Objekten zu verfolgen. Es erfüllt damit alle Anforderungen an ein Verfahren zum Verfolgen von Personen mit verschiedenen Sensoren.

Die Sensoren, die dem in dieser Arbeit eingesetzen Roboter zur Verfügung stehen, sind ein Laser-Entfernungsmesser, eine Kamera und zwei Mikrofone. Da der Laser Objekte in einer Höhe von 30 cm erfasst, also auf der Höhe von Beinen, wurde ein Verfahren zur Bestimmung der Position von Beinpaaren in den Laserdaten entwickelt. Da der Laser mit einem Öffnungswinkel von 180° den gesamten Halbraum vor dem Roboter abdeckt, bildet die Beinerkennung eine wichtige Grundlage für die Personenverfolgung. Für die Kamera wurde zunächst ein Verfahren zur Detektion von Gesichtern entwickelt, das über eine Kombination aus Hautfarbensegmentierung und *Eigenface*-Methode realisiert ist. Dieses wurde jedoch später durch den Viola-Jones-Detektor ersetzt, der leistungsfähiger und effizienter ist. Für die Kamera wurde ein weiteres Verfahren zur Lokalisation des Oberkörpers anhand von Farbinformation eingesetzt. Mit den Mikrofonen wird Sprache über die *CSP*-Analyse lokalisiert. Eine Sprachquelle lässt sich mit zwei Mikrofonen zwar nur auf eine Hälfte eines zweischaligen Hyperboloids eingrenzen. Unter Berücksichtigung multimodaler Information ist im multimodalen *Anchoring* jedoch eine Zuordnung zu Personenhypothesen möglich. Als Kompositionsmodell wurde ein einfaches, rotationssymmetrisches Modell definiert. Das Bewegungs- und Fusionsmodell stellen zusammen einen Kalman-Filter dar.

Das multimodale *Anchoring* realisiert die Wahrnehmung von Personen. Hierauf baut die in dieser Arbeit entwickelte personenbasierte Aufmerksamkeitssteuerung auf. Die grundlegende Aufgabe der Aufmerksamkeitssteuerung besteht aus den Schritten „Selektieren“ und „Fokussieren“. Die Selektion wählt die Person aus, die für den Roboter die höchste Relevanz hat. Auf diese richtet sich im Folgenden die Aufmerksamkeit des Roboters. In der Interaktion hat immer der Kommunikationspartner die höchste Relevanz. Es wird dadurch eine dauerhafte Aufmerksamkeit realisiert. Da es in der Bereitschaftsphase dagegen das Ziel des Roboters ist, einen neuen Interaktionspartner zu finden, werden alle vom multimodalen *Anchoring* verfolgten Personen in dem Selektionsprozess berücksichtigt. Es wird angenommen, dass Menschen, die zum Roboter sprechen, ihn gleichzeitig anschauen. Daher sind „sprechen“ und „schauen“ die Attribute, welche die Relevanz einer Person bestimmen. Die Merkmale werden für jede Person aus dem multimodalen *Anchoring*-Prozess gewonnen. Das Vorliegen eines Merkmals führt zu einer Erhöhung der Relevanz der jeweiligen Person um einen konstanten Wert. „Sprechen“ hat dabei einen höheren Wert als „schauen“. Damit reagiert der Roboter bevorzugt auf akustische Signale und schaut immer auf den aktuellen Sprecher.

Um neue Benutzer effektiv zu finden, beinhaltet der Selektionsmechanismus zusätzliche Funktionalitäten. So erfolgt ein Wechsel frühestens nach einer Sekunde, um den Roboter nicht hek-

tisch wirken zu lassen. Relevanzwerte von Personen, die für eine gewisse Zeitspanne im Fokus der Aufmerksamkeit waren, aber keinen Kommunikationspartner darstellen, werden für eine begrenzte Dauer auf das Minimum gesetzt, sodass auch andere Personen beobachtet werden können und der Fokus nicht auf einer Person „hängen“ bleibt. Bei gleichen, höchsten Relevanzwerten wird die Person selektiert, die für die längste Zeit nicht im Fokus der Aufmerksamkeit war.

Nach der Selektion erfolgt das Fokussieren. Dabei wird die Aufmerksamkeit des Roboters auf die selektierte Person gerichtet. Dies geschieht durch ein nach außen hin sichtbares, aktives Verhalten, das durch Ansteuerung der Roboterbasis und Ausrichtung der Kamera erzeugt wird. Es dient sowohl der Optimierung der Wahrnehmung als auch einer intuitiv verständlichen Rückmeldung der Aufmerksamkeit an den Benutzer. Das Verhalten des Roboters ist nicht immer gleich, sondern hängt von der Interaktionssituation ab. Es ist an bestimmte Aufmerksamkeitszustände gekoppelt. Es gibt vier Zustände in der Bereitschaftsphase, die den Roboter schlafend, wach, wachsam oder zuhörend darstellen. In der Interaktionsphase sind die Zustände an die Dialogsituation gebunden, wobei der Roboter auf eine Anweisung wartet, dem Benutzer folgt, eine Zeigegeste erwartet oder auf ein Objekt schaut, und damit die für das *Home-Tour*-Szenario benötigten Fähigkeiten realisiert.

Die zulässigen Zustandswechsel sind über einen deterministischen endlichen Automaten festgelegt. Transitionen zwischen Zuständen der Bereitschaftsphase werden durch Ereignisse ausgelöst, die sich aus dem multimodalen *Anchoring* ableiten (bottom-up). Alle anderen Zustandswechsel sind Folge von Ereignissen anderer Komponenten des Interaktionssystem, vorwiegend der Dialogsteuerung (top-down). Der Datenaustausch zwischen der Aufmerksamkeitssteuerung und den anderen Komponenten wird in einer Drei-Schichten-Architektur strukturiert. Zentrale Komponente der Architektur ist der so genannte *Execution Supervisor*, der den Datenaustausch kontrolliert.

Die Übergange zwischen Aufmerksamkeitszuständen der Bereitschaftsphase und der Interaktionsphase entsprechen dem Beginn beziehungsweise dem Ende einer Interaktion. Eine Person wird als neuer Benutzer registriert, wenn sie zum Roboter spricht und die Äußerung als Beginn eines Dialogs im *Home-Tour*-Szenario geeignet ist.

Neben dem aktiven Verhalten beim Fokussieren findet auch ein interner, nach außen hin nicht sichtbarer Aufmerksamkeitszuwendungsprozess statt. Das Aufmerksamkeitssystem aktiviert dabei die Sprachverarbeitung für den Zeitraum, der durch den Beginn und das Ende einer Anweisung des Benutzers an den Roboter bestimmt ist. Es übergibt dabei die relative Position der Person in Bezug zu den Mikrofonen, sodass das Sprachsignal durch *Beamforming* aus der entsprechenden Richtung verstärkt werden kann.

Als zusätzliche Rückmeldung an den Benutzer wird ein animiertes Gesicht auf dem Flachbildschirm des Roboters angezeigt. Die Pupillen der Augen sind dabei immer auf die selektierte Person gerichtet. Der Gesichtsausdruck, der aus unterschiedlichen Positionen von Mund und Augenbrauen resultiert, hängt vom Aufmerksamkeitszustand und der Interaktionssituation ab und gibt damit eine zusätzliche Rückmeldung über die internen Verarbeitungsprozesse.

Das entwickelte System ist im praktischen Einsatz in fünf verschiedenen Experimenten mit Versuchspersonen getestet worden. Dabei konnten die Robustheit des Verfahrens zur Personenver-

folgung und die Effektivität der Aufmerksamkeitssteuerung im *Home-Tour*-Szenario demonstriert werden. In Befragungen bewertete die Mehrheit der Versuchspersonen die Aufmerksamkeit als eine besonders interessante Fähigkeit des Roboters.

Das im Rahmen dieser Arbeit entwickelte Aufmerksamkeitssystem verleiht einem mobilen Roboter die Fähigkeit zu einer effektiven und natürlichen Interaktion mit Menschen im *Home-Tour*-Szenario. Das System ermöglicht es dem Roboter, Personen in seiner Umgebung aktiv zu beobachten, um in effizienter Weise zu erkennen, wann jemand beabsichtigt, mit ihm zu interagieren. In der Interaktion ist er in der Lage, sich allein auf den Benutzer zu fokussieren und andere zeitgleich anwesende Personen zu ignorieren. Durch das aktive Verhalten stehen den Komponenten des Interaktionssystems über die Sensoren jederzeit geeignete Daten zur Verfügung. Gleichzeitig zeigt der Roboter dem Benutzer in intuitiv verständlicher Weise den Fokus seiner Aufmerksamkeit an und schafft dadurch eine angenehme Interaktionssituation. Das Aufmerksamkeitssystem geht in seiner Vollständigkeit weit über bereits existierende Realisierungen hinaus.

# Literaturverzeichnis


[And99] M. Andersson, A. Orebäck, M. Lindstrom, H. I. Christensen: *ISR: An Intelligent Service Robot*, in H. I. Christensen, H. Bunke, H. Noltmeier (Hg.), *Sensor Based Intelligent Robots; International Workshop Dagstuhl Castle, Germany, September 28 – October 2, 1998. Selected Papers*, Springer-Verlag, New York, NY, USA, Bd. 1724 von *Lecture Notes in Computer Science*, 1999, S. 287–310.

[Arg75] M. Argyle: *Bodily Communication*, Methuen & Co Ltd, London, 1975.

[Arg76] M. Argyle, M. Cook: *Gaze and Mutual Gaze*, Cambridge University Press, 1976.

[Aso01] H. Asoh, Y. Motomura, F. Asano, I. Hara, S. Hayamizu, K. Itou, T. Kurita, T. Matsui, et al.: *Jijo-2: An Office Robot that Communicates and Learns*, *IEEE Intelligent Systems*, Bd. 16, Nr. 5, 2001, S. 46–55.

[Bar95] Y. Bar-Shalom, X. Li: *Multitarget-Multisensor Tracking: Principles and Techniques*, YBS Publishing, 1. Aufl., 1995.

[Bea03] M. J. Beal, N. Jojic, H. Attias: *A Graphical Model for Audiovisual Object Tracking*, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Bd. 25, Nr. 7, 2003, S. 828–836.

[Ber02] B. Berdugo, J. Rosenhouse, H. Azhari: *Speakers' Direction Finding using Estimated Time Delays in the Frequency Domain*, *Signal Processing*, Bd. 82, 2002, S. 19–30.

[Bla98] A. Blake, M. Isard: *Active Contours*, Springer, 1998.

[Bla03] J. Blanco, W. B. abd R. Sanz, J. L. Fernández: *Fast Face Detection for Mobile Robots by Integrating Laser Range Data with Vision*, in *Proc. 11th Int. Conf. on Advanced Robotics (ICAR)*, Coimbra, Portugal, 2003, Bd. 2, S. 953–958.

[Böh03] H.-J. Böhme, T. Wilhelm, J. Key, C. Schauer, C. Schröter, H.-M. Groß, T. Hempel: *An Approach to Multimodal Human-Machine Interaction for Intelligent Service Robots*, *Robotics and Autonomous Systems*, Bd. 44, Nr. 1, 2003, S. 83–96.

[Bre99] C. Breazeal, B. Scassellati: *A Context-Dependent Attention System for a Social Robot*, in D. Thomas (Hg.), *Proc. of the 16th Int. Joint Conf. on Artificial Intelligence (IJCAI-99-Vol2)*, Morgan Kaufmann Publishers, SF, 1999, S. 1146–1153.

[Bri98] F. Z. Brill, G. S. Wasson, G. J. Ferrer, W. N. Martin: *The Effective Field of View Paradigm: Adding Representation to a Reactive System*, *Engineering Applications of Artificial Intelligence*, Bd. 11, 1998, S. 189–201.

[Bro03] A. Brooks, S. Williams: *Tracking People with Networks of Heterogeneous Sensors*, in *Proc. Australasian Conf. on Robotics and Automation*, Brisbane, Australien, 2003.

[Bur98] W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, S. Thrun: *The Interactive Museum Tour-Guide Robot*, in *Proc. of the Fifteenth Nat. Conf. on Artificial Intelligence (AAAI-98)*, AAAI/MIT Press, Madison, WI, USA, 1998, S. 11–18.

[Cav99] K. R. Cave: *The FeatureGate Model of Visual Selection*, *Psychological Research*, Bd. 62, 1999, S. 182–194.

[Che53] E. C. Cherry: *Some Experiments on the Recognition of Speech, With One and Two Ears*, *Journal of the Acoustical Society of America*, Bd. 25, 1953, S. 975–979.

[Che01] G. Cheng, A. Nagakubo, Y. Kuniyoshi: *Continuous Humanoid Interaction: An Integrated Perspective — Gaining Adaptivity, Redundancy, Flexibility — In One*, *Robotics and Autonomous Systems*, Bd. 37, Nr. 2–3, 2001, S. 161–183.

[Che04] A. Chella, S. Coradeschi, M. Frixione, A. Saffiotti: *Perceptual Anchoring via Conceptual Spaces*, in *Proc. of the AAAI-04 Workshop on Anchoring Symbols to Sensor Data*, AAAI Press, Menlo Park, CA, USA, 2004, S. 40–45.

[Cor00] S. Coradeschi, A. Saffiotti: *Anchoring Symbols to Sensor Data: Preliminary Report*, in *Proc. of the 7th Conf. on Artificial Intelligence (AAAI-00)*, AAAI/MIT Press, Menlo Park, CA, USA, 2000, S. 129–135.

[Cor01a] S. Coradeschi, D. Driankov, L. Karlsson, A. Saffiotti: *Fuzzy Anchoring*, in *Proc. of the IEEE Int. Conf. on Fuzzy Systems*, Melbourne, Australien, 2001, S. 111–114.

[Cor01b] S. Coradeschi, A. Saffiotti: *Perceptual Anchoring of Symbols for Action*, in B. Nebel (Hg.), *Proc. of the Seventeenth Int. Conf. on Artificial Intelligence (IJCAI-01)*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, S. 407–416.

[Cor03] S. Coradeschi, A. Saffiotti: *An Introduction to the Anchoring Problem*, *Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, Bd. 43, Nr. 2–3, 2003, S. 85–96.

[Dén03] O. Déniz, M. Castrillón, J. Lorenzo, M. Hernández, J. Méndez: *Multimodal Attention System for an Interactive Robot*, in F. J. Perales López, A. J. C. Campilho, N. Pérez de la Blanca Capilla, A. Sanfeliu i Cortés (Hg.), *Pattern Recognition and Image Analysis; First Iberian Conf., IbPRIA 2003, Puerto de Andratx, Mallorca, Spain, June 4–6, 2003.*, Springer-Verlag, Heidelberg, Bd. 2652 von *Lecture Notes in Computer Science*, 2003, S. 212–220.

[Doi01] M. Doi, M. Nakakita, Y. Aoki, S. Hashimoto: *Real-time Vision System for Autonomous Mobile Robot*, in *Proc. of 2001 IEEE Int. Workshop on Robot and Human Interaction (ROMAN'01)*, IEEE Press, Bordeaux/Paris, Frankreich, 2001, S. 442–449.

[Doi02] M. Doi, K. Suzuki, S. Hashimoto: *Integrated Communicative Robot "BUGNOID"*, in *Proc. of 2002 IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN'02)*, IEEE Press, Berlin, 2002, S. 259–264.

[Dri98] J. A. Driscoll, R. A. Peters II, K. R. Cave: *A Visual Attention Network for a Humanoid Robot*, in *Proc. of the 1998 IEEE/RSJ Int. Conf. on Intelligent Robotic Systems*, B. C., Kanada, 1998.

[Eim96] M. Eimer, D. Nattkemper, E. Schröger, W. Prinz: *Unwillkürliche Aufmerksamkeit*, in O. Neumann, A. F. Sanders (Hg.), *Enzyklopädie der Psychologie*, Hogrefe-Verlag, Göttingen, Bd. 2, Aufmerksamkeit von *2*, *Kognition*, Kap. 5, 1996, S. 219–266.

[Fey00] S. Feyrer, A. Zell: *Robust Real-Time Pursuit of Persons with a Mobile Robot Using Multisensor Fusion*, in E. Pagello, F. Groen, T. Arai, R. Dillman, A. Stentz (Hg.), *Proc. Int. Conf. on Intelligent Autonomous Systems*, IOS Press, Venice, Italien, 2000, S. 710–715.

[Fin04] G. A. Fink, J. Fritsch, S. Hohenner, M. Kleinehagenbrock, S. Lang, G. Sagerer: *Towards Multi-Modal Interaction with a Mobile Robot*, *Pattern Recognition and Image Analysis*, Bd. 14, Nr. 2, 2004, S. 173–184.

[Fod02] A. Fod, A. Howard, M. J. Mataric: *Laser-Based People Tracking*, in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA'02)*, Washington DC, USA, 2002, S. 3024–3029.

[Fre97] Y. Freund, R. E. Schapire: *A Decision-theoretic Generalization of On-line Learning and an Application to Boosting*, *Journal of Computer and System Sciences*, Bd. 55, Nr. 1, 1997, S. 119–139.

[Fri02] J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink, G. Sagerer: *Improving Adaptive Skin Color Segmentation by Incorporating Results from Face Detection*, in *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, IEEE Press, Berlin, 2002, S. 337–343.

[Fri03a] S. Frintrop, E. Rome, A. Nüchter, H. Surmann: *An Attentive, Multi-modal Laser „Eye"*, in J. Crowley, J. H. Piater, M. Vincze, L. Paletta (Hg.), *Proc. of 3rd Int. Conf. on Computer Vision Systems (ICVS 2003)*, ECVision, Springer, Berlin, 2003, LNCS 2626, S. 202–211.

[Fri03b] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, G. Sagerer: *Multi-Modal Anchoring for Human-Robot-Interaction*, *Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, Bd. 43, Nr. 2–3, 2003, S. 133–147.

[Fri04] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, G. Sagerer: *Audiovisual Person Tracking with a Mobile Robot*, in F. Groen, N. Amato, A. Bonarini, E. Yoshida, B. Kröse (Hg.), *Proc. Int. Conf. on Intelligent Autonomous Systems*, IOS Press, Amsterdam, Niederlande, 2004, S. 898–906.

[Fuj98] M. Fujita, H. Kitano: *Development of an Autonomous Quadruped Robot for Robot Entertainment*, *Autonomous Agents*, Bd. 5, Nr. 1, 1998, S. 7–18.

[Gat98] E. Gat: *On Three-Layer Architectures*, in D. Kortenkamp, R. P. Bonasso, R. Murphy (Hg.), *Artificial Intelligence and Mobile Robots: Case Studies of Successful Robot Systems*, MIT Press, Cambridge, MA, USA, Kap. 8, 1998, S. 195–210.

[Ger03] B. P. Gerkey, R. T. Vaughan, A. Howard: *The Player/Stage Project: Tools for Multi-Robot and Distributed Sensor Systems*, in *Proc. Int. Conf. on Advanced Robotics*, Coimbra, Portugal, 2003, S. 317–323.

[Ghi00] S. S. Ghidary, Y. Nakata, T. Takamori, M. Hattori: *Human Detection and Localization at Indoor Environment by Home Robot*, in *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics*, Nashville, Tennessee, 2000, S. 1360–1365.

[Ghi02] S. S. Ghidary, Y. Nakata, H. Saito, M. Hattori, T. Takamori: *Multi-Modal Interaction of Human and Home Robot in the Context of Room Map Generation*, *Autonomous Robots*, Bd. 13, Nr. 2, 2002, S. 169–184.

[Giu94] D. Giuliani, M. Omologo, P. Svaizer: *Talker Localization and Speech Recognition using a Microphone Array and a Cross-PowerSpectrum Phase Analysis*, in *Int. Conf. on Spoken Language Processing*, Yokohama, Japan, 1994, Bd. 3, S. 1243–1246.

[Gon99] L. M. G. Gonçalves, D. S. Wheeler, A. A. F. Oliveira, R. A. Grupen: *Towards a Framework for Robot Cognition*, in *Proc. IEEE Int. Symp. on Computational Intelligence in Robotics and Automation (CIRA '99)*, Monterey, CA, USA, 1999.

[GP04] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, D. Moore: *Audio-Visual Speaker Tracking with Importance Particle Filters*, in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, Barcelona, Spanien, 2004.

[Haa04] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, et al.: *BIRON – The Bielefeld Robot Companion*, in E. Prassler, G. Lawitzky, P. Fiorini, M. Hägele (Hg.), *Proc. Int. Workshop on Advances in Service Robotics*, Fraunhofer IRB Verlag, Stuttgart, 2004, S. 27–32.

[Ham04] K. A. Hambuchen: *Multi-modal Attention and Event Binding in Humanoid Robots Using a Sensory Ego-Sphere*, Dissertation, Faculty of the Graduate School of Vanderbilt University, Nashville, Tennessee, USA, 2004.

[Has02] S. Hashimoto, S. Narita, H. Kasahara, K. Shirai, T. Kobayashi, A. Takanishi, S. Sugano, J. Yamaguchi, et al.: *Humanoid Robots in Waseda University—Hadaly-2 and WABIAN*, *Autonomous Robots*, Bd. 12, Nr. 1, 2002, S. 25–38.

[Hje01] E. Hjelmås, B. K. Low: *Face Detection: A Survey*, *Computer Vision and Image Understanding (CVIU)*, Bd. 83, Nr. 3, 2001, S. 236–274.

[Hua01] Y. Huang, J. Benesty, G. W. Elko, R. M. Mersereau: *Real-Time Passive Source Localization: A Practical Linear-Correction Least-Squares Approach*, *IEEE Trans. on Speech and Audio Processing*, Bd. 9, Nr. 8, 2001, S. 943–956.

[Itt98] L. Itti, C. Koch, E. Niebur: *A Model of Saliency-Based Visual Attention for Rapid Scene Analysis*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 20, Nr. 11, 1998, S. 1254–1259.

[Jen03] B. Jensen, R. Philippsen, R. Siegwart: *Narrative Situation Assessment for Human-Robot Interaction*, in *Proc. of the IEEE Int. Conf. on Robotics and Automation*, Taipei, Taiwan, 2003, S. 1503–1508.

[Joh93] D. Johnson, D. Dudgeon: *Array Signal Processing: Concepts and Techniques*, Prentice Hall, 1993.

[Kal60] R. E. Kalman: *A New Approach to Linear Filtering and Prediction Problems*, *Transactions of the ASME–Journal of Basic Engineering*, Bd. 82-D, 1960, S. 35–45.

[Kaw00] K. Kawamura, R. A. Peters II, D. M. Wilkes, W. A. Alford, T. E. Rogers: *ISAC: Foundations in Human-Humanoid Interaction*, *IEEE Intelligent Systems*, Bd. 15, Nr. 4, 2000, S. 38–45.

[Kir90] M. Kirby, L. Sirovich: *Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces*, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Bd. 12, Nr. 1, 1990, S. 103–108.

[Kle05] M. Kleinehagenbrock: *Interaktive Verhaltenssteuerung für Robot Companions*, Dissertation, Universität Bielefeld, Technische Fakultät, Angewandte Informatik, Bielefeld, Deutschland, 2005.

[Kna76] C. H. Knapp, G. C. Carter: *The Generalized Correlation Method for Estimation of Time Delay*, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Bd. ASSP-24, Nr. 4, 1976, S. 320–327.

[Kop02] L. Kopp, P. Gärdenfors: *Attention as a Minimal Criterion of Intentionality in Robots*, *Cognitive Science Quaterly*, Bd. 2, 2002, S. 302–319.

[Koz01] H. Kozima, H. Yano: *A Robot that Learns to Communicate with Human Caregivers*, in *The First Int. Workshop on Epigenetic Robotics*, Lund, Sweden, 2001.

[Krö03] B. J. A. Kröse, J. M. Porta, A. J. N. van Breemen, K. Crucq, M. Nuttin, E. Demeester: *Lino, the User-Interface Robot*, in E. H. L. Aarts, R. Collier, E. van Loenen, B. E. R. de Ruyter (Hg.), *Ambient Intelligence; First European Symp., EUSAI 2003, Veldhoven, The Netherlands, November 3–4, 2003. Proceedings*, Springer-Verlag, Heidelberg, 2003, Bd. 2875 von *Lecture Notes in Computer Science*, S. 264–274.

[Lan03] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, G. Sagerer: *Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot*, in *Proc. Int. Conf. on Multimodal Interfaces*, ACM, Vancouver, Kanada, 2003, S. 28–35.

[Lem02] A. Lemieux, M. Parizeau: *Experiments on Eigenfaces Robustness*, in *Proc. 16th Int. Conf. on Pattern Recognition (ICPR)*, 2002, Bd. 1, S. 421–424.

[Li00] Y. Li, S. Gong, H. Liddell: *Support Vector Regression and Classification Based Multi-view Face Detection and Recognition*, in *IEEE Int. Conf. on Automatic Face & Gesture Recognition*, Grenoble, Frankreich, 2000, S. 300–305.

[Li04] S. Li, M. Kleinehagenbrock, J. Fritsch, B. Wrede, G. Sagerer: *"BIRON, let me show you something": Evaluating the Interaction with a Robot Companion*, in W. Thissen, P. Wieringa, M. Pantic, M. Ludema (Hg.), *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics, Special Session on Human-Robot Interaction*, IEEE, The Hague, The Netherlands, 2004, S. 2827–2834.

[Lie02a] R. Lienhart, A. Kuranov, V. Pisarevsky: *Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection*, Techn. Ber., Intel Labs, 2002.

[Lie02b] R. Lienhart, J. Maydt: *An Extended Set of Haar-like Features for Rapid Object Detection*, in *IEEE Proceedings, ICIP 2002*, 2002, Bd. 1, S. 900–903.

[Löm04] F. Lömker: *Lernen von Objektbenennungen mit visuellen Prozessen*, Dissertation, Universität Bielefeld, Technische Fakultät, Angewandte Informatik, Bielefeld, Deutschland, 2004.

[Mac67] J. MacQueen: *Some Methods for Classification and Analysis of Multivariate Observations*, in L. M. L. Cam, J. Neyman (Hg.), *Proc. Fifth Berkeley Symp. on Mathematical Statistics and Probability*, 1967, Bd. 1, S. 281–296.

[Mat01] Y. Matsusaka, S. Fujie, T. Kobayashi: *Modeling of Conversational Strategy for the Robot Participating in the Group Conversation*, in *Proc. Europ. Conf. on Speech Communication and Technology (Eurospeech'01)*, Aalborg, Dänemark, 2001, S. 2173–2176.

[Mat03] Y. Matsusaka, T. Tojo, T. Kobayashi: *Conversation Robot Participating in Group Conversation, IEICE Transaction on Information and System*, Bd. E86-D, Nr. 1, 2003, S. 26–36.

[McG02] P. McGuire, J. Fritsch, H. Ritter, J. Steil, F. Röthling, G. A. Fink, S. Wachsmut, G. Sagerer: *Multi-Modal Human-Machine Communication for Instructing Robot Grasping Tasks*, in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'02)*, IEEE, 2002, S. 1082–1089.

[McK99] S. J. McKenna, Y. Raja, S. Gong: *Tracking Colour Objects using Adaptive Mixture Models, Image and Vision Computing*, Bd. 17, Nr. 3–4, 1999, S. 225–231.

[Mül02] H. J. Müller, J. Krummenacher: *Aufmerksamkeit*, in W. Prinz, J. Müsseler (Hg.), *Allgemeine Psychologie*, Spektrum Akademischer Verlag, Heidelberg/Berlin, Kap. 1c, 2002.

[Müs00] J. Müsseler: *Aufmerksamkeit*, in G. Wenninger (Hg.), *Lexikon der Psychologie*, Spektrum Akademischer Verlag, Heidelberg, Bd. 1, 2000, S. 154–156.

[Nak00] K. Nakadai, T. Lourens, H. G. Okuno, H. Kitano: *Active Audition for Humanoid*, in *Proc. of the Seventeenth Nat. Conf. on Artificial Intelligence (AAAI-2000)*, Austin, 2000, S. 832–839.

[Nak01] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, H. Kitano: *Real-Time Auditory and Visual Multiple-Object Tracking for Robots*, in B. Nebel (Hg.), *Proc. of 17th Int. Joint Conf. on Artificial Intelligence (IJCAI-01)*, Morgan Kaufmann Publishers Inc., Seattle, WA, USA, 2001, Bd. 2, S. 1425–1432.

[Nef98] A. V. Nefian, M. H. Hayes: *Face Detection and Recognition using Hidden Markov Models*, in *Int. Conf. on Image Processing (ICIP '98)*, Chicago, Illinois, USA, 1998, Bd. 1, S. 141–145.

[Nil82] N. J. Nilsson: *Principles of Artificial Intelligence*, Springer-Verlag, Berlin, 1982.

[Nou99] I. Nourbakhsh, J. Bobenage, S. Grange, R. Lutz, R. Meyer, A. Soto: *An Affective Mobile Robot Educator with a Full-time Job, Artificial Intelligence*, Bd. 114, 1999, S. 95–124.

[Oku01] H. G. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, H. Kitano: *Human-Robot Interaction through Real-Time Auditory and Visual Multiple-Talker Tracking*, in *Proc. IEEE Int. Conf. on Intelligent Robots and Systems*, Maui, Hawaii, 2001, Bd. 3, S. 1402–1409.

[Oku02] H. G. Okuno, K. Nakadai, H. Kitano: *Social Interaction of Humanoid RobotBased on Audio-Visual Tracking*, in T. Hendtlass, M. Ali (Hg.), *Proc. of 18th Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-2002)*, Springer-Verlag, Cairns, Australien, 2002, Bd. 2358 von *Lecture Notes in Artificial Intelligence*, S. 725–734.

[Oku03] H. G. Okuno, K. Nakadai, H. Kitano: *Realizing Personality in Audio-Visually Triggered Non-verbal Behaviors*, in *Proc. of IEEE-RAS Int. Conf. on Robotics and Automation (ICRA-2003)*, Taipei, Taiwan, 2003, Bd. 1, S. 392–397.

[Oli00] N. Oliver, A. Pentland, F. Bérard: *LAFTER: a Real-time Face and Lips Tracker with Facial Expression Recognition*, *Pattern Recognition*, Bd. 33, Nr. 8, 2000, S. 1369–1382.

[Par01] H. K. Park, H. S. Hong, H. J. Kwon, M. J. Chung: *A Nursing Robot System for the Elderly and the Disabled*, in *Proc. Int. Workshop on Human-friendly Welfare Robotic Systems*, Daejeon, Korea, 2001, S. 122–126.

[Pet01] R. A. Peters II, K. A. Hambuchen, K. Kawamura, D. M. Wilkes: *The Sensory Ego-Sphere as a Short-term Memory for Humanoids*, in *Proc. of the 2001 IEEE-RAS Int. Conf. on Humanoid Robots*, Waseda University, Tokyo, Japan, 2001, S. 451–459.

[Pos84] M. I. Posner, Y. A. Cohen: *Components of Visual Orienting*, in H. Bouma, D. G. Bouwhuis (Hg.), *Attention and Performance X: Control of Language Processes*, Erlbaum, London, 1984, S. 531–556.

[Raj98] Y. Raja, S. J. McKenna, S. Gong: *Segmentation and Tracking using Colour Mixture Models*, in *Proc. Asian Conf. on Computer Vision*, Hong Kong, 1998, Bd. 1, S. 607–614.

[Reh99] J. M. Rehg, K. P. Murphy, P. W. Fieguth: *Vision-Based Speaker Detection Using Bayesian Networks*, in *Proc. of the IEEE Computer Science Conf. on Computer Vision and Pattern Recognition (CVPR-99)*, IEEE, Los Alamitos, 1999, S. 110–116.

[Rog00] T. Rogers, M. Wilkes: *The Human Agent: A Work in Progress toward Human-Humanoid Interaction*, in *Proc. 2000 IEEE Int. Conf. on Systems, Man and Cybernetics*, Nashville, 2000.

[Row98] H. Rowley, S. Baluja, T. Kanade: *Neural Network-Based Face Detection*, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Bd. 20, Nr. 1, 1998, S. 23–38.

[Roy00] N. Roy, G. Baltus, D. Fox, F. Gemperle, J. Goetz, T. Hirsch, D. Magaritis, M. Montemerlo, et al.: *Towards Personal Service Robots for the Elderly*, in *Proc. Int. Workshop on Interactive Robotics and Entertainment*, Pittsburgh, PA, USA, 2000.

[Sca01] B. Scassellati: *Foundations for a Theory of Mind for a Humanoid Robot*, Dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA, USA, 2001.

[Sch01a] B. J. Scholl: *Objects and Attention: The State of the Art*, *Cognition*, Bd. 80, Nr. 1–2, 2001, S. 1–46.

[Sch01b] R. D. Schraft, B. Graf, A. Traub, D. John: *A Mobile Robot Platform for Assistance and Entertainment*, *Industrial Robot: An International Journal*, Bd. 28, Nr. 1, 2001, S. 29–34.

[Sch01c] D. Schulz, W. Burgard, D. Fox, A. B. Cremers: *Tracking Multiple Moving Objects with a Mobile Robot*, in *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2001)*, Kauwai, Hawaii, 2001, S. 371–377.

[Sek02] A. Ş. Sekmen, M. Wilkes, K. Kawamura: *An Application of Passive Human-Robot Interaction: Human Tracking Based on Attention Distraction*, *IEEE Trans. on Systems, Man, and Cybernetics – Part A: Systems and Human*, Bd. 32, Nr. 2, 2002, S. 248–259.

[Shi04] M. Shiomi, T. Kanda, N. Miralles, T. Miyashita, I. Fasel, J. Movellan, H. Ishiguro: *Face-to-face Interactive Humanoid Robot*, in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2004)*, Sendai, Japan, 2004, S. 1340–1346.

[Sid99] H. Sidenbladh, D. Kragić, H. I. Christensen: *A Person Following Behaviour for a Mobile Robot*, in *Proc. IEEE Int. Conf. on Robotics and Automation*, IEEE Press, Detroit, Michigan, 1999, Bd. 1, S. 670–675.

[Sor00] M. Soriano, B. Martinkauppi, S. Huovinen, M. Laaksonen: *Skin Detection in Video under Changing Illumination Conditions*, in *Proc. 15th Int. Conf. on Pattern Recognition*, Barcelona, Spanien, 2000, Bd. 1, S. 839–842.

[Stö01a] M. Störring, H. J. Andersen, E. Granum: *Physics-based Modelling of Human Skin Colour under Mixed Illuminants*, *Robotics and Autonomous Systems*, Bd. 35, Nr. 3–4, 2001, S. 131–142.

[Sto01b] A. Stoytchev, R. Arkin: *Combining Deliberation, Reactivity and Motivation in the Context of a Behavior-Based Robot Architecture*, in *Proc. of IEEE Int. Symp. on Computational Intelligence in Robotics and Automation (CIRA 2001)*, Banff, Kanada, 2001, S. 290–295.

[Top04] E. A. Topp, D. Kragic, P. Jensfelt, H. I. Christensen: *An interactive interface for service robots*, in *Proc. IEEE Int. Conf. on Robotics and Automation*, IEEE Press, New Orleans, LA, USA, 2004, Bd. 4, S. 3469–3475.

[Tur91] M. Turk, A. Pentland: *Eigenfaces for Recognition*, *Journal of Cognitive Neuroscience*, Bd. 3, Nr. 1, 1991, S. 71–86.

[Ver01] J. Vermaak, A. Blake, M. Gangnet, P. Pérez: *Sequential Monte Carlo Fusion of Sound and Vision for Speaker Tracking*, in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, IEEE Computer Society, Vancouver, Kanada, 2001, Bd. 1, S. 741–746.

[Vij01] S. Vijayakumar, J. Conradt, T. Shibata, S. Schaal: *Overt Visual Attention for a Humanoid Robot*, in *Proc. Int. Conf. on Intelligence in Robotics and Autonomous Systems (IROS 2001)*, Hawaii, 2001, S. 2332–2337.

[Vio01] P. Viola, M. Jones: *Robust Real-time Object Detection*, in *Proc. IEEE Int. Workshop on Statistical and Computational Theories of Vision*, Vancouver, Kanada, 2001.

[Vio04] P. Viola, M. J. Jones: *Robust Real-Time Face Detection, Int. Journal of Computer Vision*, Bd. 57, Nr. 2, 2004, S. 137–154.

[Wal00] S. Waldherr, R. A. F. Romero, S. Thrun: *A Gesture Based Interface for Human-Robot Interaction*, *Autonomous Robots*, Bd. 9, Nr. 2, 2000, S. 151–173.

[Wan99] C. Wang, M. S. Brandstein: *Multi-Source Face Tracking with Audio and Visual Data*, in *IEEE Int. Workshop on Multimedia Signal Processing*, Kopenhagen, Dänemark, 1999.

[Was99] G. Wasson, D. Kortenkamp, E. Huber: *Integrating Active Perception with an Autonomous Robot Architecture*, *Robotics and Autonomous Systems*, Bd. 29, Nr. 2–3, 1999, S. 175–186.

[Wil02] T. Wilhelm, H.-J. Böhme, H.-M. Groß: *Sensor Fusion for Vision and Sonar Based People Tracking on a Mobile Service Robot*, in *Proc. Int. Workshop on Dynamic Perception*, IOS Press, Infix, Bochum, 2002, S. 315–320.

[Wol94] J. M. Wolfe: *Guided Search 2.0: A Revised Model of Visual Search*, *Psychonomic Bulletin and Review*, Bd. 1, Nr. 2, 1994, S. 202–238.

[Wys82] G. Wyszecki, W. S. Stiles: *Color Science: Concepts and Methods, Quantitative Data and Formulae*, John Wiley and Sons, New York, NY, USA, 1982.

[Yan96] J. Yang, A. Waibel: *A Real-time Face Tracker*, in *Proc. of the Third IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, USA, 1996, S. 142–147.

[Yan98] J. Yang, W. Lu, A. Waibel: *Skin-color Modeling and Adaption*, in *Proc. Asian Conf. on Computer Vision*, Hong Kong, 1998, Bd. 2, S. 687–694.

[Yan02] M.-H. Yang, D. J. Kriegman, N. Ahuja: *Detecting Faces in Images: A Survey*, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Bd. 24, Nr. 1, 2002, S. 34–58.

[Yui92] A. L. Yuille, P. W. Hallinan, D. S. Cohen: *Feature Extraction from Faces Using Deformable Templates*, *Int. Journal of Computer Vision*, Bd. 8, Nr. 2, 1992, S. 99–111.

[Zha02a] Z. Zhang, L. Zhu, S. Z. Li, H. Zhang: *Real-Time Multi-view Face Detection*, in *Proc. of The 5th Int. Conf. on Automatic Face and Gesture Recognition*, Washington DC, USA, 2002, S. 149–154.

[Zha02b] H. Zhao, R. Shibasaki, N. Ishihara: *Pedestrian Tracking using Single-row Laser Range Scanners*, in *Proc. IAPR Workshop on Machine Vision Application*, Nara, Japan, 2002, S. 158–162.