Architekturen situierter Kommunikatoren: Von Perzeption über Kognition zum Lernen

Gernot A. Fink, Jannik Fritsch, Nadine Leßmann,
Helge Ritter, Gerhard Sagerer, Jochen J. Steil, Ipke Wachsmuth

Universität Bielefeld, SFB 360 / Technische Fakultät,
Postfach 100 131, 33501 Bielefeld

Abstract: Charakteristisches Merkmal intelligenter Systeme ist das Ineinandergreifen zahlreicher Teilfunktionen. Während in der Vergangenheit in erster Linie die Realisierung eines geeigneten Umfangs von tragfähigen Teilfunktionalitäten angestrebt wurde, verschieben die Fortschritte auf diesem Feld die Herausforderung mehr und mehr zur Frage einer übergreifenden Architektur, die eine große Anzahl von Teilfunktionen integrieren und zu einem "intelligenten" Zusammenwirken bringen kann.

Die Entwicklung integrierter Architekturkonzepte ist eines der wesentlichen Ziele des Bielefelder SFB 360. Dabei entstanden drei auf jeweils einen zentralen Aspekt fokussierte Teildemonstratoren, die wir in diesem Beitrag vorstellen werden. Diese Teilsysteme mit den Schwerpunkten Perzeption, Kognition bzw. Lernen sind wechselseitig koppelbar und arbeiten auf einer realitätsnahen Komplexitätsebene. Die entwickelten Konzepte können somit einen wesentlichen Beitrag zur Erforschung der Architektur künstlicher kognitiver Systeme leisten.

1 Einleitung

Körperlich verankerte Kognition spielt heute in zahlreichen Gebieten eine zunehmend wichtigere Rolle, sei es in der Entwicklung Mensch-zentrierter robotischer Systeme, der Modellierung und dem Test von Verarbeitungsmechanismen in der Kognitionswissenschaft, oder für virtuelle Humanoide in Szenarien der virtuellen Realität. Sie ist auch ein zentraler Aspekt im Sonderforschungsbereich 360 *Situierte Künstliche Kommunikatoren*, wo situationsbezogene Kommunikationsfähigkeiten künstlicher Systeme in der Mensch-Maschine-Interaktion erforscht werden. Im Fokus steht dabei die Entwicklung maschineller Systeme, die das Verhalten und die Kompetenz natürlicher Kommunikatoren in relevanten Aspekten nachbilden. Angesichts der Komplexität dieser Aufgabe wird von einem begrenzten Basis-Szenario ausgegangen, in dem Sprache, Gestik, Wissen, Planung, Handlung und Sensomotorik von Mensch und Maschine zusammenwirken müssen, um eine gemeinsame Montageaufgabe kooperativ zu lösen.

Das maschinelle System muss dafür in der Lage sein, seine Umwelt und insbesondere seinen menschlichen Kommunikationspartner visuell und akustisch wahrzunehmen und das Wahrgenommene situativ und kognitiv zu verarbeiten. Dies erfordert eine Integration von Perzeption, Dialogfähigkeiten, Diskurswissen, Turntaking-Verhalten bis hin zu reflexiven

Fähigkeiten, wie etwa eine sprachliche Beschreibung des aktuellen Teilziels zu generieren oder Aussagen über den eigenen perzeptuellen Status zu formulieren und laufende Aktionen handlungsübergreifend zu kommentieren.

Die Realisierung solcher Systeme verschiebt die Herausforderung von der Lösung einzelner Teilprobleme hin auf die Ebene der Architektur: Zur Schlüsselfrage wird die Entwicklung leistungsfähiger Mechanismen zur dynamischen und inkrementellen Koordination eines großen Spektrums an Teilfähigkeiten. Dabei liegt nur selten eine einfache kompositionelle Struktur vor; vielmehr sind wesentliche Teilfunktionen oft in komplexer Weise ineinander verschränkt und softwaretechnisch heterogen implementiert.

Architekturforschung für derartige Systeme steht daher unweigerlich zwischen zwei Polen: die Konstruktion idealisierter Systeme zur möglichst weitgehenden Implementierung und Analyse übergreifender Konzepte auf der einen Seite, und die Entwicklung von Systemen, deren Fokus auf der Verankerung in einer Realweltsituation liegt. Unsere Fähigkeit, uns Handlungen und Bilder in "innerer Simulation" vorzustellen, deutet darauf hin, daß beide Pole für die Realisierung künstlicher kognitiver Systeme verbunden werden müssen, und motiviert vieles von dem Weg, der im SFB 360 zur Erforschung der damit einhergehenden Architekturfragen eingeschlagen wurde.

Forschungsleitend war das Ziel, Methoden der modernen VR-Modellierung, Erkenntnisse über kognitive Architekturen, heutige Realisierungsmöglichkeiten von Perzeptionskomponenten und Forschungen im Bereich situierten maschinellen Lernens zusammenzuführen, um in prototypischer Form wesentliche Architekturausschnitte in Form von drei Teildemonstratoren zu realisieren und diese zur Verwirklichung einer umfassenden Mensch-Maschine-Schnittstellenfunktionalität miteinander zu koppeln.

Der erste Teildemonstrator beinhaltet ein perzeptives Front-End-System, in dem visuelle Wahrnehmung, Sprachverarbeitung und -Verstehen sowie die Integration visueller und sprachlicher Ebenen stattfinden. Wir werden dabei besonders auf die integrierte Verarbeitung audiovisueller Sinneseindrücke eingehen. Mit Hilfe einer engen Verschränkung der Verarbeitungsschritte sowie dem Einsatz probabilistischer Verfahren zur Fusion unimodaler Perzeptionshypothesen zu multimodalen Interpretationsstrukturen wird die Grundlage für einen robusten multimodalen Mensch-Maschine-Dialog gelegt.

Der zweite Teildemonstrator hat eine Anzahl kognitiv motivierter Faktoren zum Gegenstand. Im Zentrum steht die Präsentation eines anthropomorphen "Gegenübers" zur Verkörperung der realisierten Interaktions- und Kommunikationsfähigkeiten. Als Realisierungsgrundlage wurde ein VR-Ansatz gewählt, um einen Agenten darzustellen, der sich weitgehend natürlich bewegen kann, innerhalb des Dialogszenarios menschenähnliches Verhalten approximiert und dazu über mehrkanalige Ausgabemöglichkeiten verfügt. Durch Mimik und Körpersprache kann er Auskunft über seinen internen Zustand geben und gleichzeitig in verbalen Äußerungen Auskünfte über Konstruktionschritte erteilen. Die Steuerung dieses Verhaltensrepertoires wurde in Form einer kognitiv motivierten Verhaltensarchitektur realisiert.

Der dritte Teildemonstrator fokussiert die Thematik situiertes Lernen. Erst wenn wir Roboter unter Verbindung von sprachlichem Dialog, Gestik und visueller Demonstration zu gewünschten Aktionen anleiten können, werden sie ihre bis heute enge Spezialistenrolle

verlassen und dem Menschen eine vielseitige Unterstützung im Alltag bieten können. Ziel des dritten Teildemonstrators ist daher die Entwicklung einer geeigneten Lernarchitektur, die einen Roboter in die Lage versetzt, Aktionen zu beobachten, erfolgreich zu *imitieren* und — als Voraussetzung dazu — einen *gemeinsamen Aufmerksamkeitsfokus* mit dem menschlichen Partner herzustellen und aufrechtzuerhalten. Für eine multimodale Kommunikation mit dem Benutzer sind darüberhinaus perzeptive Fähigkeiten mindestens in den Bereichen des Sprachverstehens, des aktiven Sehens und in der Interpretation non-verbaler Hinweise wie z.B. Gestik zu realisieren und geeignet zu koordinieren. Daraus ergibt sich die Notwendigkeit einer engen Kopplung mit den beiden anderen Teildemonstratoren, insbesondere in Hinblick auf die Perzeptions- und Sprachkomponenten.

Alle drei Teildemonstratoren basieren auf einer größeren Anzahl von Funktionsmodulen, die in den zurückliegenden Förderphasen des SFB 360 entwickelt, evaluiert und optimiert wurden [KW02, BFF$^+$01, SHJ$^+$01, BPFWS99, WJ96]. Ihre Verfügbarkeit eröffnet die Möglichkeit, nunmehr auch auf der Ebene komplexer Architekturen Konzepte in konkreten Implementierungen auf ihre Tragfähigkeit zu überprüfen. Die im folgenden beschriebenen Teildemonstratoren sind wichtige Schritte auf diesem Weg, der für die Erforschung und Realisierung intelligenter Systeme auch künftig noch viele herausfordernde Forschungsfragen bereithalten wird.

2 Perzeption

Der perzeptionsorientierte Prototyp eines künstlichen Kommunikators ist in einem Konstruktionsszenario situiert und unterstützt die flexible Aggregierung einfacher Elemente eines Spielzeugbaukastens zu komplexeren Einheiten: Den Anweisungen eines menschlichen Instrukteurs folgend können in der Arbeitsumgebung vorhandene Objekte manipuliert und zu komplexeren Aggregaten verbunden werden. Diese Aktionen können sowohl in einer virtuellen Szenenrepräsentation als auch durch den realen Manipulator eines Roboters ausgeführt werden.

Die Architektur des perzeptionsorientierten Teildemonstrators zeigt Abbildung 1. Sie besteht aus zwei Strängen signalverarbeitender Module, die in einer Integrationskomponente zusammengeführt werden. Im links abgebildeten Sprachverarbeitungsstrang erfolgt eine integrierte Erkennung und Interpretation sprachlicher Äußerungen. Als Ergebnis werden Merkmalsstrukturen erzeugt, die die Bedeutung domänenspezifischer Konstituenten repräsentieren. Visuelle Daten werden im Strang der rechts gezeigten Module segmentiert und interpretiert. Dabei werden Hypothesen über einzelne Objekte sowie komplexe Objektaggregate erzeugt. Die zusätzlich zwischen diesen Ergebnissen über die Zeit etablierten Relationen erlauben außerdem die Erkennung von Handlungen und Bauplänen für Aggregate. Die Resultate aus Bild- und Sprachverarbeitung laufen in einem Modul zusammen, das Bayes-Netze einsetzt, um integrierte Interpretationen zu berechnen. Gleichzeitig realisiert eine zustandsbasierte Dialogkomponente robuste Strategien zur Dialogführung und ist verantwortlich für die Generierung von Rückfragen im Falle unverständlicher oder mehrdeutiger Anweisungen an das System [BPFWS99].

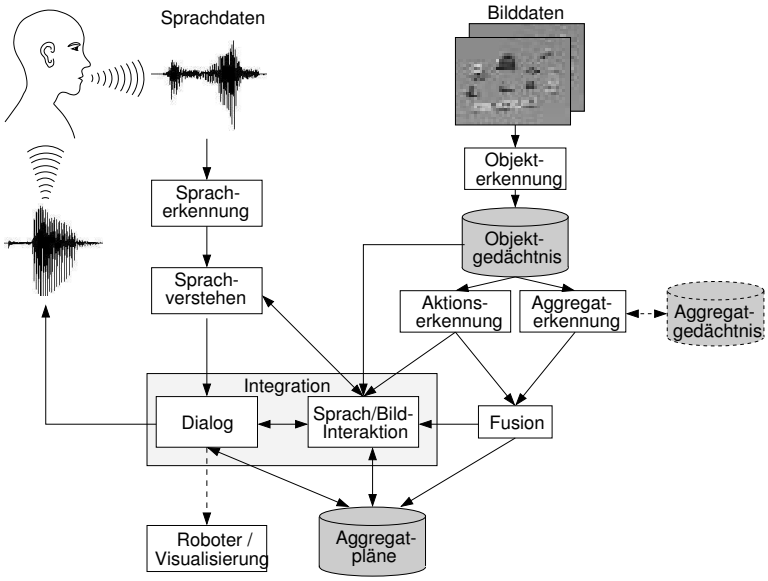

Abbildung 1: Module und Modulinteraktionen des perzeptiven Front-Ends.

Das Integrationsmodul, das Bild- und Sprachverarbeitungsresultate zusammenführt und aufeinander bezieht, nimmt in der Architektur eine zentrale Rolle ein. Aufgrund von Schnittstellen zur Dialogkomponente sowie zu einer Datenbank, die aus Bildern extrahierte Aggregatstrukturen speichert, ist es möglich, sowohl auf explizite wie auch auf implizite Art und Weise Benennungen für Aggregate einzuführen, die im Verlauf eines Konstruktionsdialoges entstehen. Um Tests des perzeptiven Front-Ends zu erleichtern, ist es möglich, die Robotikkomponente durch ein Visualisierungsmodul zu ersetzen, das die aktuelle Interpretation, die das Systems über Objekte und Ereignisse in seiner Umgebung hergeleitet hat, mit Mitteln der virtuellen Realität veranschaulicht.

2.1 Integrierte Sprach- und Bildverarbeitung

Um eine möglichst robuste und flexible sprachliche Kommunikation zu gewährleisten, verwenden wir ein sprecherunabhängiges Erkennungssystem für natürlichsprachliche Eingaben [Fi99]. Der Erkennungsprozeß wird direkt durch einen partiellen Parser beeinflußt, der linguistische und durch das Szenario gegebene Restriktionen auf Wortsequenzen einbringt. Da der Instruktor jedoch weder auf eine spezielle Kommandosyntax noch auf die Verwendung bestimmter Objektbezeichnungen eingeschränkt sein soll, müssen die nachgeordneten Sprachverstehensmodule auf einen hohen Grad an *referentieller Ungenauigkeit* ausgelegt sein. Im Ergebnis werden im Gegensatz zu einfachen Worthypothesenfolgen typischer Erkennungssysteme partielle syntaktische Strukturen generiert, z.B. Objektbeschreibungen (“der rote Würfel”) oder räumliche Beziehungen (“links von”) [BPFWS99].

Zur Erkennung von elementaren Objekten und Objektaggregaten in Bildfolgen werden semantische Netzwerke eingesetzt, die domänenspezifisches Objektwissen modellieren [KFSB98]. Gleichzeitig erkennt ein regelbasierter Algorithmus elementare Montageoperationen; durch Fusion der dabei anfallenden Ergebnisse mit denen aus der Aggregaterkennung lassen sich Baupläne der in der Szene sichtbaren Aggregate ableiten, so dass umfangreiches Wissen über zuvor unbekannte komplexe Objekte aus Bilddaten erlernt werden kann [BFKS99].

Im Allgemeinen wird die automatische Integration von Sprach- und Bildverarbeitungsergebnissen durch verschiedene Unsicherheiten wie z.B. fehlerhafte Erkennungsergebnisse oder die Benutzung partieller oder unspezifischer Objektreferenzen beeinträchtigt. Dieser Tatsache trägt unser Integrationsansatz Rechnung, indem die Integration verschiedener perzeptiver Modalitäten als ein *statistischer Dekodierungsprozess* interpretiert wird, der sich mit Hilfe von Bayes-Netzen modellieren lässt. Genauer gesagt wird jede im Sprachsignal erkannte Objektbeschreibung und jedes im Bild erkannte Objekt als ein eigenes Subnetz repräsentiert. Die einzelnen Modellknoten stehen für sprachlich benennbare oder visuell erfassbare Objekteigenschaften sowie für mögliche Relationen zwischen diesen. Die statistischen Abhängigkeiten zwischen den möglichen Belegungen dieser Knoten mit konkreten Merkmalsausprägungen repräsentieren sowohl Beziehungen zwischen sprachlichen Bezeichnungen und bestimmten Objekten als auch Unsicherheiten oder Fehler, die durch den jeweiligen Analyseprozess entstehen. Sind diese Wahrscheinlichkeiten gegeben, lässt sich zu einer Äußerung über eine Szene ein Netz ableiten und anschließend relaxieren. Die wahrscheinlichste Abbildung zwischen den visuell erkannten Objekten und den im Sprachsignal erkannten Bezeichnern ist dann durch die maximalen *a posteriori* Hypothesen der entsprechenden Relationsknoten definiert. Sobald diese berechnet worden sind, lassen sich weitere Inferenzen ziehen, z.B. über die mit maximaler Wahrscheinlichkeit ermittelte Objektklasse [WS02]. Somit wird es möglich, zu entscheiden, ob und welches Objekt einer Szene verbal bezeichnet wurde.

2.2 Zeitliches Verhalten

Für die technische Evaluation der Systemleistung sowie für die Optimierung der Systemreaktionszeit wurde das Zeitverhalten aller Module ermittelt. Diese Daten wurden analysiert und einzelne Module wurden dahingehend optimiert, dass eine möglichst kurze Zeitspanne zwischen Instruktion und Dialogantwort liegt.

Das Zeitverhalten des Gesamtsystems für eine Instruktion mit einem intendierten Objekt und einem Referenzobjekt ("Nimm das Heck links neben der roten Schraube") ist in Abbildung 2 als UML-Sequenzdiagramm auf der Granularitätsebene der in Abbildung 1 gezeigten Module sichtbar. Die Darstellung zeigt deutlich sowohl die Asynchronität der Eingabemodalität Sprache relativ zur Bildverarbeitungsschleife als auch die asynchrone Verarbeitung dieser Modalitäten. Abhängig von der Anzahl der Objekte in Instruktionen ergibt sich eine durchschnittliche Reaktionszeit von ca. 500-850 ms bis das System eine Antwort generiert.

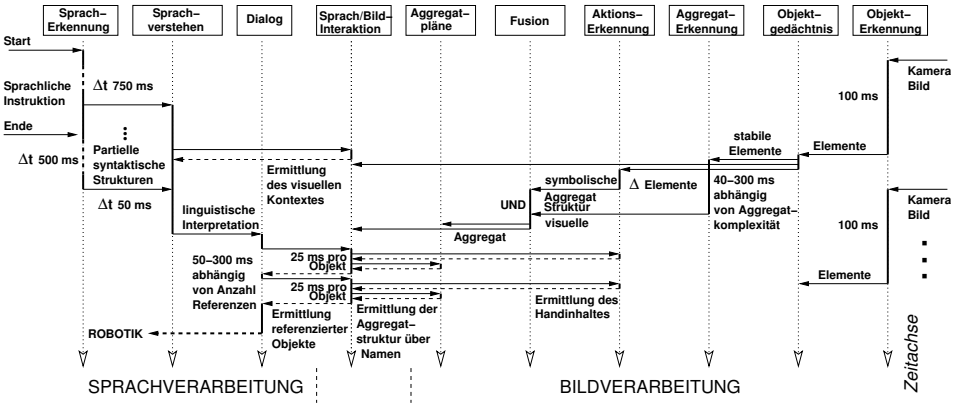

Abbildung 2: UML-Diagramm zum Zeitverhalten der Module des perzeptiven Front-Ends.

Neben der Länge der Systemreaktionszeit ist auch die Qualität der Interaktion zwischen Benutzer und System hochrelevant. Um diese Aspekte des Perzeptionsprototyps zu untersuchen, wurden umfangreiche Evaluierungsexperimente durchgeführt, bei denen naive Benutzer in Kooperation mit dem System einfache Konstruktionsaufgaben zu lösen hatten [BFR+02]. Die Ergebnisse dieser Evaluation haben gezeigt, daß mit zunehmender Komplexität der Konstruktionsaufgabe, und damit auch der Anzahl und Komplexität der Benuterinstruktionen, die Anzahl der Systemrückfragen zunimmt. Obwohl damit die Dauer der Interaktion stark zunimmt, sinkt die Interaktionsqualität nur leicht, d.h. das Konstruktionsziel wird oft erreicht, das Gesamtsystem zeigt also robustes Interaktionsverhalten.

3 Kognition

In diesem Abschnitt wird eine kognitiv motivierte Architektur für einen virtuellen anthropomorphen Agenten vorgestellt. Sie dient dazu einen künstlichen Kommunikator zu schaffen, der sich auf eine natürliche, direkte Interaktion mit dem Benutzer konzentriert. Mit diesem Ansatz wird nicht allein eine Konzeption für einen humanoiden Dialogpartner verfolgt, sondern es soll damit auch ein theoretisches Modell für die Integration verschiedener Ansätze kognitiver Modellierung entwickelt werden, das Körperlichkeit und Kognition stimmig zusammenführen kann. Einsatz findet die Architektur in dem anthropomorphen künstlichen Kommunikator *Max* (Abb. 3) in virtueller Realität. Sie ermöglicht sowohl Fähigkeiten eines Dialoges mit geplanten Äußerungen, als auch die Fähigkeit zu spontaner reaktiver Äußerung, beispielsweise in Form von Turntaking- und Feedback-Signalen. Zusätzlich können verschiedene spezialisierte Planer, z.B. mit Wissen über die Konstruktion von Baufix-Flugzeugen, und spezialisierte Gedächtnisse, z.B. mit dynamischen Konzeptmodellen für strukturierte Aggregate ([WJ96]), integriert und in die Kommunikation einbezogen werden. Konzipiert wird die Kommunikation auf Basis der Sprechakttheorie

Abbildung 3: Interaktion mit dem virtuellen anthropomorphen Agenten *Max*

nach Searle [SV85] bzw. der Theorie kommunikativer Akte nach Poggi und Pelachaud [PP00]. Das Dialogsystem ist planbasiert; kommunikative Akte werden als Aktion-Plan-Operatoren dargestellt. Dabei kann die kognitive Komponente für die Dialog- wie für die Handlungsplanung eingesetzt werden. Die Darstellung von Bewegungen des Körpers von Max erfolgt durch Echtzeit-Computeranimation eines kinematischen Modells [KW02].

3.1 Struktureller Aufbau

Das für den anthropomorphen Agenten Max konzipierte Kernsystem eines situierten künstlichen Kommunikators integriert symbolverarbeitende und verhaltensbasierte Ansätze in einer hybriden Systemarchitektur, die Wahrnehmung und reaktives Verhalten, höhere mentale Prozesse wie Schlußfolgern und planvolles Handeln bis hin zum Einbezug von Aufmerksamkeit und motivationaler Handlungsbewertung (über "Desires") betreffen. Abbildung 4 skizziert den strukturellen Aufbau des Agenten. Der Kreis, unterteilt in eine *Perceive-Reason-Act*-Triade, stellt die interne Verarbeitung des Agenten dar und grenzt ihn von seiner Umwelt ab. Dabei hebt die Dreiteilung die Verzahnung und das enge Zusammenspiel des klassischen *Perceive-Reason-Act*-Zyklus hervor. Der direkte Informationsfluß zwischen den Sektoren Perceive und Act berücksichtigt jedoch, daß reaktives Verhalten entstehen kann, ohne daß zuvor eine Deliberation stattgefunden haben muß und die kognitive Schleife durchlaufen wurde. Damit kann der vorliegende Ansatz als eine Hybrid-Architektur charakterisiert werden, die reaktives und deliberatives Verhalten in einer Struktur vereinigt.

Die Kreissektoren *Perceive* und *Act* repräsentieren die Physis des Agenten. Durch seine Körperlichkeit ist der Agent in der Umwelt verankert, erhält er Weltbezug. Sie dient weiterhin als Ausdrucksmöglichkeit in Form der Multimodalität (Gestik, Sprechmimik wie

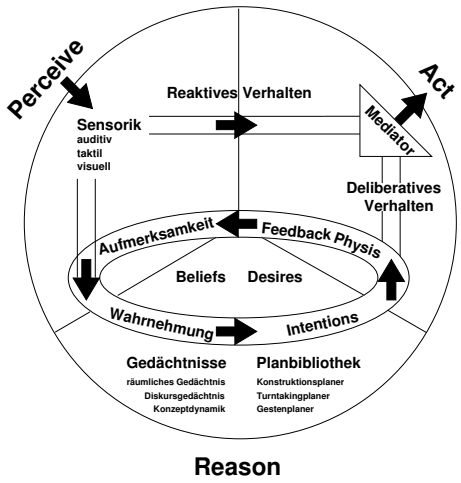

Abbildung 4: Struktureller Aufbau der kognitiv motivierten Architektur

auch emotionale Mimik). Die Sensorverarbeitung sowie die ausführende Aktorik sind durch körperliche Eigenschaften des Agenten geprägt. Dies wird auch bei der Modalitätenwahl einer auszuführenden Aktion berücksichtigt.

3.2 Zentraler Ablauf

Die Wahrnehmung-, Schlußfolgerungs- und Handlungskomponente sind nebenläufig realisiert und es existiert eine parallele Verarbeitung in der deliberativen und reaktiven Komponente. Auf der einen Seite können die Sensordaten (von Spracheingabe/auditiv, Körpersensorik/taktil, Szenenwahrnehmung/visuell) direkt ein reaktives Verhalten auslösen, welches schnell und auf einem niedrigen Abstraktionsniveau operiert und sich dabei durch eine enge Kopplung an die Sensorik auszeichnet. Reaktive Verhaltensweisen dienen in der Form von Reflexen mit hohen Prioritätswerten unmittelbaren Systemreaktionen wie Ausweichverhalten bei drohender Kollision; im Bereich der Kommunikation treten reaktive Verhaltensweisen z.B. beim Turntaking auf. Zusätzlich wird der reaktiven Komponente die Aufgabe der *Secondary Behaviors* wie z.B. Augenzwinkern zuteil.

Auf der anderen Seite präsentiert die *deliberative Schleife* einen Kreislauf, der die interne kognitive Verarbeitung des Agenten betrifft und das Wechselspiel zwischen Datenakquisition und Informationsverarbeitung aufzeigt. Wahrnehmung besteht hier nicht aus der starren Erfassung sensorischer Daten, sondern aus einer situationssensitiven Verarbeitung perzipierter Sensoreindrücke. Kognition wird damit nicht als abgelöster interner Vorgang betrachtet, sondern vielmehr als stark an die Physis gekoppelter Prozeß mit einer stärkeren Betonung der prozeduralen Komponente. Die Sensordaten finden Eingang in die kognitive Schleife, werden dabei durch eine *Aufmerksamkeitssteuerung* gefiltert und wechselwirken in Form einer interpretierten und analysierten Wahrnehmung mit verschiedenen speziali-

sierten Gedächtnissen. Diese arbeiten auf verschiedenartigen Repräsentationen, legen aber vereinheitlicht jeweils relevante Fakten auf einem hohen Abstraktionsniveau in den *Beliefs* ab, die das Arbeitsgedächtnis des Agenten darstellen.

Der Kern des deliberativen Moduls folgt dem *Belief-Desire-Intention* (BDI)-Ansatz und setzt auf JAM [Hu99] auf. Als verhaltensauslösender Antrieb dienen explizit repräsentierte Ziele (*Desires*), die sowohl durch interne Verarbeitung als auch von außen aufgeworfen werden können. Die Intentionsbildung der kognitiven Schleife wird durch einen BDI-Interpreter vorgenommen, welcher aufgrund der vorliegenden *Beliefs*, den aktuellen Wünschen und Zielen des Agenten sowie seinen alternativen Handlungsmöglichkeiten eine aktuelle *Intention* bestimmt. Handlungsoptionen liegen in Form von Plänen vor, die durch Vorbedingungen, Kontextbedingungen, erreichbaren Konsequenzen und eine Prioritätsfunktion beschrieben werden. Die Planbibliothek besteht zum einen aus simplen Plankonstrukten, die einfache Aktionen direkt in entsprechende Behaviors umsetzen können. Zum anderen können jedoch auch dynamische eigenständige Planer bei Bedarf angestoßen werden, um einen konkreten, komplexeren Plan auszuarbeiten. Aus der Priorität des *Desires* sowie der Kompetenzbewertung des Planers und eventuell weiteren Parametern wird eine Gesamtpriorität bestimmt, mit der ein Plan darum konkurriert, aktiv zu werden. Verfügt er über die höchste Priorität, so wird er zur aktuellen Intention und erhält die Möglichkeit, interne Variablen und Beliefs zu beeinflussen sowie Behaviors zu instantiieren, die dann wiederum im Mediator um den Zugriff auf die Aktoren konkurrieren.

Sowohl reaktives als auch deliberatives Verhalten wird durch Behaviors und Motorskills verschiedener Komplexitätsstufen umgesetzt. Der Mediator schlichtet zwischen den Verhaltensweisen und zieht dabei in Betracht, welche Modalitäten gerade frei bzw. im Rahmen anderer Verhaltensweisen bereits im Einsatz sind. Die Entscheidungsgrundlage des Mediators besteht aus Prioritätswerten, die die Dringlichkeit und Angemessenheit eines Verhaltens in einer vorliegenden Situation ausdrücken und von den Verhaltensweisen und Intentionen selbst lokal berechnet werden [BG95]. Planselektion findet somit einerseits auf der Ebene der kognitiven (bewußten) Intentionsbildung statt, andererseits und ebenfalls auch durch den Mediator auf der Ebene der direkten Aktionsausführung. Sowohl die aktiv ausgeführten Intentionen als auch die aktuell anliegenden und möglicherweise konkurrierenden Verhaltensweisen werden bei den zurückfließenden *Feedbackinformationen* berücksichtigt. Die Rückkopplung der erfolgten Aktionen und Aktorzustände wirken sich wiederum in Form einer *Aufmerksamkeitssteuerung* auf die Sensorik und Wahrnehmung aus und schließen somit den Zyklus. Die Schleife verdeutlicht eine der zentralen Kernideen der Architektur, nämlich daß ein ständiger Strom von Informationen zwischen den Sektoren umläuft, der sowohl aktuelle Sensor- und Aktorinformationen als auch interne Zustände einbezieht.

3.3 Max als Konstruktionspartner

Die aktuelle Implementierung realisiert zentrale Aspekte der Architektur und versetzt Max in die Lage auf Anfrage des Benutzers die Konstruktion verschiedener Aggregate zu erläutern oder interaktiv mit dem Benutzter Konstruktionen vorzunehmen. Das heißt Max

beschreibt mittels seines Konstruktionswissens in synthetischer Sprache und unter Zuhilfenahme verschiedener Gesten, welche Baufixteile miteinander verbunden werden müssen, und erläutert so entweder schrittweise den gesamten Bauplan eines Aggregats, oder aber er erklärt, welcher Konstruktionsschritt als nächster vollzogen werden soll, überläßt jedoch die Ausführung dem Benutzer, der entsprechende multimodale Instruktionen absetzen kann. Die Erkennung des Benutzers in der realen Welt (Bewegung, Blickrichtung, Gestik) erfolgt über getrackte Marker, Datenhandschuhe und durch ein Mikrophon, das Daten an eine Sprachverarbeitungskomponente liefert. Nach erfolgter Benutzeraktion liefert Max Feedback. Wurde die Konstruktion korrekt vorgenommen, so stimmt er zu und setzt seine Erläuterungen fort. Im Falle einer falschen Handlung jedoch macht er die Aktion des Benutzers in der virtuellen Umgebung rückgängig und erklärt den Konstruktionsschritt erneut.

Für die Beurteilung des Erfolgs der Benutzeraktion greift Max auf das Szenenwissen von COAR zurück [WJ96]. Die schritthaltend aktualisierte COAR-Beschreibung der Szene enthält Informationen über die Objekte und deren eingegangene Verbindungen und weitere Objekteigenschaften. Zusätzlich verfügt Max über eine einfache visuell-räumliche Wahrnehmung, die Eingang in ein räumliches Gedächtnis findet. Während des gesamten Dialogs berücksichtigt Max den Diskursverlauf und weist Turntakingverhalten auf, indem er beispielsweise nur dann spricht, wenn er im Besitz des Turns ist aber andererseits auch jederzeit unterbrochen werden kann. Reaktives Verhalten existiert in der aktuellen Version in Form von *Secondary Behaviors*, die durch Atmungsbewegungen und Augenblinzeln zu einem lebendigen Erscheinungsbild von Max beitragen. Ferner existiert ein Behavior für die Fixation des Blickpunktes auf den Benutzer. Vollführt Max gerade keine Aktionen, bei denen er Sichtüberwachung benötigt, so er schaut er dem Benutzer in die Augen und verfolgt ihn mit seinem Blick. Sobald Max jedoch anfängt, etwas zu erklären, wird dieses Verhalten überstimmt und Max fixiert die Objekte, die er gerade referenziert.

4 Auf dem Weg zum situierten Lernen

Um Lernen in künstlichen kognitiven Systemen praktikabel zu machen, wird seit einigen Jahren der Ansatz des *Imitationslernens* diskutiert [Sc99, AGM$^+$01]. Die grundlegende Idee besteht dabei darin, ein "Vorbild" für eine erfolgreiche Trajektorie im entsprechenden Suchraum durch Beobachtung eines Instrukteurs zu finden. Basis dafür ist ein künstliches Robotiksystem mit perzeptuellen, kognitiven, sowie aktorischen Komponenten, wie es im Rahmen des aktuellen Aktorik-Teildemonstrators GRAVIS (Gesture Recognition Active Vision System) gegeben ist. Abbildung 5 gibt einen Grobüberblick über die wichtigsten Verarbeitungspfade. Die Sprachverarbeitungsmodule und die Aufmerksamkeitssteuerung liefern situationsbezogene Dialoginformation und nichtverbale (visuelle und gestische) Information an ein Integrationsmodul. Die daraus erkannte Instruktion wird an die Aufmerksamkeitssteuerung übermittelt, die nach einer Neufokussierung geeignete Befehle für die Robotersteuerungskomponenten (Bewegung, Greifen) generiert. Darüberhinaus sind weitere Steuerungsfunktionalitäten wie Kalibrierungsfunktionen für die visuelle Verarbei-

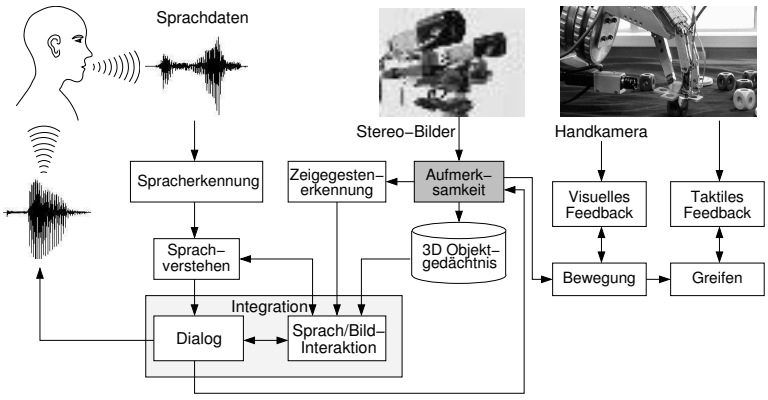

Abbildung 5: Schemabild des aktuellen Aktorik-Teildemonstrators.

tung, den Kamerakopf und die "Hand-Auge"-Koordination im System integriert, weitere Details finden sich in [SHJ+01, MFS+02].

Die *Aktorikkomponenten* bilden zur Zeit ein 6-DOF PUMA-Arm, der mit der Echtzeitbibliothek RCCL betrieben wird, und eine mehrfingrige Greifhand. Der Arm verfügt neben der üblichen Positionssensorik über eine Endeffektorkamera, um lokales visuelles Feedback während der Annäherungs- und Greifphasen aufzunehmen und auszuwerten. Das Greifen wird durch eine dreifingrige 9-DOF Roboterhand mit hydraulischem Antrieb ausgeführt, die an der Universität München entwickelt wurde (TUM-Hand). Sie hat drei anthropomorphe Finger und wird hydraulisch betrieben. Die Fingerspitzen tragen in Eigenentwicklung hergestellte Kraftsensoren, durch die ein Kraftfeedback zur Kontrolle und Evaluation von Griffen zur Verfügung steht. Zur Verbesserung der Greiffähigkeiten und um eine größere Anzahl verschiedener Griffe zu erlauben, haben wir die Originalkonfiguration der TUM-Hand um eine Handfläche so erweitert, dass ein Finger als "Daumen" genutzt werden kann (siehe auch Abb. 7). Auf die so gewonnene Handinnenfläche wurde zusätzlich eine taktile Sensormatrix von aufgebracht, welche insbesondere für Kraftgriffe wertvolles Feedback liefern kann und deren Auflösung ausreicht, um verschiedene Objekte duch ihr taktiles Profil unterscheiden zu können.

Die *visuelle Perzeption* wird durch einen aktiven Kamerakopf geleistet, der mit zwei 3-Chip-CCD Farbkameras ausgestattet ist und die üblichen Freiheitsgrade in Pan, Tilt und Vergenz hat. Zur Echtzeitverarbeitung der Stereobilder verwenden wir ein DATACUBE-System, welches Bilder mit einer Rate von 25 Hz in einer Pipeline-Architektur verarbeiten kann. Zur Integration verschiedener visueller Kanäle kommt eine Aufmerksamkeitssteuerung zum Einsatz, welche über die gewichtete Summation topographischer Merkmalskarten saliente Fixationspunkte auswählt [SHJ+01]. Im Bildausschnitt um den momentanen Aufmerksamkeitsfokus gewinnt ein auf neuronalen Netzen basierender holistischer Objekterkenner [HLR00] Hypothesen über Lage und Identität von Objekten, die der in Abschnitt 2.1 beschriebenen Komponente zur Bild-Sprach Integration zur Verfügung gestellt werden, falls sie im Laufe einer Explorationssequenz wiederholt fokussiert und korrekt klassifiziert werden.

4.1 Modularität und Adaptivität

In der Architektur des Aktorik-Teildemonstrators sind viele der oben beschriebenen Teilfähigkeiten als funktionale Module implementiert. Meist können sie auch für andere Zwecke verwendet werden und wurden zunächst unabhängig vom Gesamtsystem entwickelt und getestet. Dadurch sind sie im Einzelbetrieb in der Regel wesentlich mächtiger als im integrierten GRAVIS-System, welches aufgrund wechselseitiger Abhängigkeiten, weniger spezialisierter Hardware und häufig schlechterer Qualität der Eingangsdaten die Möglichkeiten der Teilmodule nur teilweise nutzen kann.

Ein System der Komplexität von GRAVIS, welches in der Realwelt operiert und dabei mit einem Benutzer interagiert, kann auf vielfältige Weise gestört und in seiner Funktionalität beeinträchtigt werden. In unserem Fall sind die häufigsten Probleme sich verändernde Lichtbedingungen, variierende Hautfarben, das Auftreten unbekannter Objekte und technische Probleme der Robotikkomponenten. Darüberhinaus verlangt die Architektur in vielfältiger Weise nach Robustheit der Teilmodule, denn das aktive Sehsystem, der Roboterarm und die Hand operieren in nur grob kalibrierten Koordinatensystemen, 3D-Objektpositionen sowie 3D-Zeigerichtungen werden aus 2D-Pixelkoordinaten geschätzt und alle darauf aufbauenden Berechnungen sind entsprechend ungenau. Daher ist eine entscheidende Voraussetzung für ein robustes Funktionieren der Gesamtarchitektur der Einsatz lokaler Adaptivität bereits auf der Ebene von Einzelmodulen und von lokalem Feedback insbesondere bei der Ausführung von Aktionen.

Hier nur einige der wichtigsten Beispiele: Die in der Aufmerksamkeitssteuerung verwendeten topographischen Merkmalskarten verfügen über einen Normalisierungsmechanismus, der adativ alle berechneten Merkmale zu berücksichtigen versucht. Dieses führt z.B. zu einer graduellen Rekalibration aufgrund von Änderung der Lichtbedingungen. Ebenfalls passt die verwendete Hautfarbensegmentierung ihr Farbmodell relativ zu den erkannten Händen adaptiv an. Die Gestenerkennung und –klassifikation, sowie die gesamte Objekterkennung sind als neuronale Netze implementiert, die offline trainiert wurden und teilweise über eine online ausführbare schnelle Farbkalibration verfügen. Schließlich verwendet der Roboterarm visuelles Feedback von der Handkamera um die Annäherungsbewegung bezüglich Position und Orientierung des Objektes auszuführen und eine automatische Rekalibration der Fingerspitzensensorik vorzunehmen.

4.2 Ausblick auf eine integrierte Lernarchitektur

Imitationslernen auf der Architekturebene erfordert mindestens, (i) das Robotersystem mit genügend perzeptiven Fähigkeiten auszustatten, um die zu imitierende Aktion visuell zu erfassen; (ii) die gesehene Aktion in eine geeignete interne Repräsentation zu übersetzen, die den eigenen Systemzustand (wie z.B. die andere Lage im Raum), aber auch den gegenüber der Beobachtung des Instruktlors geänderten Zugriff auf Sensor und Aktordaten berücksichtigt; (iii) eine geeignete Aktion motorisch ausführen zu können. Dazu wollen wir einen Ansatz untersuchen, der in aufeinander aufbauenden Ebenen das Ziel verfolgt,

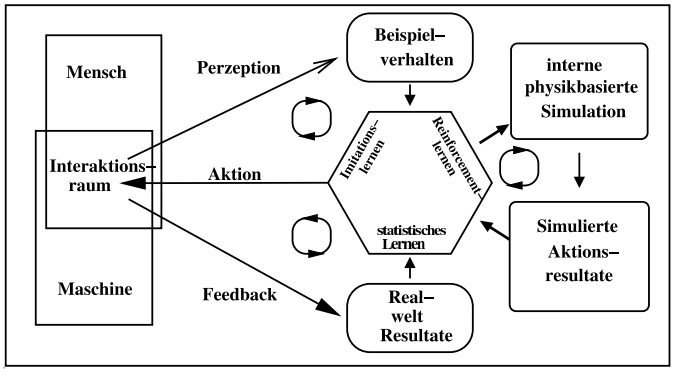

Abbildung 6: Mehrebenen-Lernarchitektur zur Realisierung von Imitationslernen

den für die zu lernende Handlung noch zu explorierenden Suchraum einer möglichst weitgehenden Einschränkung zu unterwerfen.

Die *Wahrnehmungs- und Imitationsebene* betrachtet Lernen aus der Perspektive der Beobachtung der Außenwelt ("Instrukteur") mit dem Ziel des Imitierens erfolgreicher Handlungssequenzen. Schlüsselaufgaben sind dabei die Extraktion relevanter Merkmale, Ereignisse und Abfolgen beobachteter Teilaktionen, ihre Übersetzung von der Beobachtungs- in die Eigenperspektive, sowie ihre Ausnutzung zur Fokussierung eigenen Explorierens auf vielversprechende Bereiche des i.d.R. a-priori sehr hochdimensionalen Handlungsraums. Hier soll ein suchraumbegrenztes Reinforcementlernen zum Einsatz kommen, welches die Exploration des Suchraumes auf die Nachbarschaft einer durch Beobachtung gewonnenen vorhandenen "Erfolgstrajektorie" im Zustandsraum konzentriert.

Die *Aktionsauswahl- und Explorationsebene* betrachtet Lernen aus der Perspektive der Beobachtung einer "Innenwelt" ("Simulation"). Ziel ist dabei, innerhalb der auf der ersten Ebene gewonnenen Suchraumeinschränkung Handlungsdetails zu explorieren und dabei mittels vorhandenen Modellwissens zu einer weiteren Fokussierung real noch zu verifizierender Handlungen zu gelangen. Hier sollen Reinforcementlernverfahren, welche auf Basis von Simulationen agieren, für das Lernen von Aktions-Zustandsübergängen genutzt werden. Auf der *sensomotorischen Ebene* schließlich erfolgt die tatsächliche Ausführung einer nach Durchlaufen der vorangegangenen Ebenen aussichtsreichsten Aktion.

Ein entscheidendes und verbindendes Element könnte in einer aufmerksamkeitsgesteuerten Plastizitätsfokussierung bestehen: Dabei soll Lernen durch flexibel vorgebbare, *kontextabhängige Parameterauswahlregeln* von vornherein auf in einer jeweiligen Situation besonders entscheidende Parameter fokussiert werden. Damit kann Vorwissen flexibel eingebracht werden, um das hochdimensionale Credit-Assignment-Problem herkömmlicher Lernverfahren zu umgehen. Ein solcher Mechanismus kann als eine Art "Aufmerksamkeitssteuerung" für den Lernprozess angesehen und interaktiv durch sprachliche Eingaben der Benutzer moduliert werden.

Abb. 6 zeigt das Ineinandergreifen der verschiedenen Ebenen. Die Interaktion mit dem

Abbildung 7: Angestrebtes Imitationsszenario und nach dem Vorbild menschlicher Griffposturen eingestellte, “wirkungsäquivalente” Posturen der Roboterhand des GRAVIS-Systems.

Benutzer geschieht durch *simulationsgetriebene Rückfragen:* Die nach der Filterung durch die Architekturebenen 1 und 2 verbleibende Restunsicherheit kann zur aktiven Generierung klärender Rückfragen genutzt werden. Dies macht Lernen zu einem aktiven Prozess, der Sprache zur wirkungsvollen Optimierung von Exploration nutzt. Abbildung 7 zeigt das angestrebte Imitationsszenario, in dem “wirkungsäquivalente” Posturen der Roboterhand nach dem Vorbild menschlicher Griffposturen eingestellt werden.

5 Zusammenfassung

Architekturforschung für intelligente Systeme ist der Gefahr ausgesetzt, zwischen den Polen über lange Jahre “gewachsener” und unbeweglich gewordener Großarchitekturen auf der einen Seite, und idealisierter, aber realitätsfern vereinfachter Miniaturarchitekturen auf der anderen Seite, zerrieben zu werden. Dieser Problematik haben wir uns im Bielefelder SFB 360 durch die Entwicklung dreier sorgfältig aufeinander abgestimmter Teildemonstratoren gestellt, die jeweils einen größeren, aber noch überblickbaren Ausschnitt eines künstlichen kognitiven Systems schwerpunktmäßig thematisieren, wechselseitig koppelbar sind und dabei durchgehend den Anspruch einer realistischen Komplexitätsebene aufrechterhalten.

Der perzeptive Teildemonstrator greift die Schlüsselaufgabe einer Verknüpfung von Sprache und Sehen auf und liefert uns wichtige Einblicke, wie daraus im Dialog dynamische Bedeutungskonstitution, Robustheit und natürliche Situiertheit erwachsen. Der kognitive Teildemonstrator verknüpft moderne VR-Techniken zur Realisierung eines anthropomorphen Agenten mit kognitiven Erkenntnissen über die Steuerung seiner Verhaltenskomponenten. An ihm kann untersucht werden, wie und unter welchen Voraussetzungen multi-

modales Feedback zur Übermittlung von Turntakingsignalen oder Emotionalität zu generieren ist. Der dritte Teildemonstrator fokussiert die wichtige Thematik situierten Lernens für ein Robotersystem. Mit ihm erforschen wir die Verknüpfung visueller Aufmerksamkeitssteuerung, Gestikerkennung und Integration von multimodalem Dialog zur Realisierung von Imitationslernen, d.h. der Fähigkeit, sprachlich-gestisch kommentierte Aktionsfolgen zu erfassen und geeignet generalisiert in einem Roboter nachahmen zu können.

Alle drei Teildemonstratoren sind darauf angelegt, sich wechselseitig zu ergänzen. Sie realisieren – bereits unter einer vergleichsweise losen Kopplung – einen situierten künstlichen Kommunikator, der in einem Instruktionsszenario auf weitgehend natürliche Weise mit einem menschlichen Partner interagieren kann. Die bislang verfolgte Forschungsstrategie hat damit eine wichtige und durchaus schwierige Bewährungsprobe bestanden. Die dadurch möglich gewordene Evaluierung komplexer Mensch-Maschine-Interaktionssequenzen wird uns als Ausgangpunkt dienen, die Thematik maschinellen Handlungsverstehens und -lernens auf kognitiv höheren Ebenen zu erforschen und – in einer längerfristigen Perspektive – auf portable kognitiv motivierte Architekturen hinzuarbeiten, die ein hochgradiges, aufgabenbezogenes "Alignment" zwischen künftigen, anthropomorph aufgebauten Perzeptions-Aktionssystemen und menschlichen Kooperationspartnern herstellen und damit zumindest einen Teil des Anspruchs an intelligente Systeme einlösen können.

Danksagung

Wir danken allen unseren Kollegen aus dem SFB 360, der Technischen Fakultät und der Fakultät für Linguistik und Literaturwissenschaft, die zur Entstehung der beschriebenen Teildemonstratoren beigetragen haben. Unser besonderer Dank gilt Elke Braun, Christian Bauckhage, Robert Haschke, Gunther Heidemann, Bernhard Jung, Stefan Kopp, Franz Kummert, Frank Lömker, Patrick McGuire, Frank Röthling und Sven Wachsmuth.

Die vorliegende Arbeit wurde im Rahmen des Sonderforschungsbereichs 360 "Situierte Künstliche Kommunikatoren" von der Deutschen Forschungsgemeinschaft gefördert.

Literatur

[AGM+01] Andry, P., Gaussier, P., Moga, S., Banquet, J. P., und Nadel, J.: Learning and communication via imitation: An autonomous robot perspective. *IEEE Trans. on Systems, Man, and Cybernetics*. 31(5):431–442. 2001.

[BFF+01] Bauckhage, C., Fink, G. A., Fritsch, J., Kummert, F., Lömker, F., Sagerer, G., und Wachsmuth, S.: An Integrated System for Cooperative Man-Machine Interaction. In: *IEEE Int. Symp. on Computational Intelligence in Robotics and Automation*. S. 328–333. Banff, Canada. 2001.

[BFKS99] Bauckhage, C., Fritsch, J., Kummert, F., und Sagerer, G.: Towards a vision system for supervising assembly processes. In: *Proc. Symp. on Intelligent Robotic Systems*. S. 89–98. Coimbra. 1999.

[BFR+02] Bauckhage, C., Fritsch, J., Rohlfing, K., Wachsmuth, S., und Sagerer, G.: Evaluating integrated speech- and image understanding. In: *Proc. Int, Conf. on Multimodal Interfaces (ICMI)*. S. 9–14. Pittsburgh, PA. 2002.

[BG95] Blumberg, B. M. und Galyean, T. A.: Multi-level direction of autonomous creatures for real-time virtual environments. *Computer Graphics*. 29:47–54. 1995.

[BPFWS99] Brandt-Pook, H., Fink, G. A., Wachsmuth, S., und Sagerer, G.: Integrated recognition and interpretaion of speech for a construction task domain. In: *Proc. Int. Conf. on Human-Computer Interaction*. volume 1. S. 550–554. München. 1999.

[Fi99] Fink, G. A.: Developing HMM-based recognizers with ESMERALDA. In: Matoušek, V., Mautner, P., Ocelíková, J., und Sojka, P. (Hrsg.), *Lecture Notes in Artificial Intelligence*. volume 1692. S. 229–234. Berlin – Heidelberg. 1999. Springer.

[HLR00] Heidemann, G., Lücke, D., und Ritter, H.: A system for various visual classification tasks based on neural networks. In: Sanfeliu, A. et al. (Hrsg.), *Proc. Int. Conf. on Pattern Recognition*. volume I. S. 9–12. Barcelona. 2000.

[Hu99] Huber, M. J.: JAM: A BDI-theoretic mobile agent architecture. In: *Proc. Int. Conf. on Autonomous Agents*. S. 236–243. Seattle, WA. 1999.

[KFSB98] Kummert, F., Fink, G. A., Sagerer, G., und Braun, E.: Hybrid Object Recognition in Image Sequences. In: *Proc. Int. Conf. on Pattern Recognition*. volume II. S. 1165–1170. Brisbane. 1998.

[KW02] Kopp, S. und Wachsmuth, I.: Model-based animation of coverbal gesture. In: *Proc. Computer Animation*. S. 252–257. IEEE Press, Los Alamitos, CA. 2002.

[MFS+02] McGuire, P., Fritsch, J., Steil, J. J., Röthling, F., Fink, G. A., Wachsmut, S., Sagerer, G., und Ritter, H.: Multi-modal human-machine communication for instructing robot grasping tasks. In: *Proc. Int. Conf. on Intelligent Robots and Systems (IROS)*. S. 1082–1089. Lausanne. 2002.

[PP00] Poggi, I. und Pelachaud, C.: Performative facial expression in animated faces. In: Cassell, J., Sullivan, J., Prevost, S., und Churchill, E. (Hrsg.), *Embodied Conversational Agents*. S. 155–188. Cambridge, MA. 2000. The MIT Press.

[Sc99] Schaal, S.: Is imitation learning the route to humaniod robots? *Trends in Cognitive Sciences*. 3(6):233–242. 1999.

[SHJ+01] Steil, J. J., Heidemann, G., Jockusch, J., R.Rae, Jungclaus, N., und Ritter, H.: Guiding attention for grasping tasks by gestural instruction: The GRAVIS-robot architecture. In: *Proc. Int. Conf. on Intelligent Robots and Systems (IROS)*. S. 1570–1577. Maui, Hawaii. 2001.

[SV85] Searle, J. R. und Vanderveken, D.: *Foundations of Illocutionary Logic*. Cambridge University Press. Cambridge. 1985.

[WJ96] Wachsmuth, I. und Jung, B.: Dynamic conceptualization in a mechanical-object assembly environment. In: *Artificial Intelligence Review*. volume 10. S. 345–368. 1996.

[WS02] Wachsmuth, S. und Sagerer, G.: Integrated Analysis of Speech and Images as a Probabilistic Decoding Process. In: *Proc. Int. Conf. on Pattern Recognition*. S. 588–592. Québec City, Québec, Canada. 2002.