Ersch. in: W.-D. Miethling & J. Perl (Hrsg), Sport und Informatik VI (pp. 167-178).
Köln: Sport und Buch Strauss, 1999.

Mensch-Maschine-Kommunikation mit Gestik und Sprache

Ipke Wachsmuth
Technische Fakultät, Universität Bielefeld

1 Motivation: Anthropomorphe Schnittstellen

Die Leichtigkeit und Robustheit in der Kommunikation zwischen Menschen gab Anlaß, auch in der Wechselwirkung mit maschinellen Systemen natürlichere Modalitäten anzustreben. Hier sind Sprachschnittstellen zu nennen, die sich beim Zugriff auf Datenbanken und bei der Steuerung komplexer technischer Systeme zu bewähren beginnen. Ein Spitzenbedarf besteht in der Verbesserung der Interaktionsmöglichkeiten mit Multimedia-Systemen, welche bislang überwiegend auf die reine Präsentation von multimedialen Informationen ausgelegt sind. Mit Hilfe intelligenter Funktionen soll es möglich werden, gesprochene Sprache, textuelle und visuelle Eingaben zu verarbeiten und zur Manipulation von Text, Grafik, Bildern und Geräuschen bis hin zur Interaktion mit Virtual-Reality-Umgebungen (Barfield & Furness, 1995), in denen Tastatur und Maus keinen Platz mehr haben, zu mobilisieren. Auch im Bereich der Hilfesysteme für Behinderte werden heute Schnittstellen gefordert, die sich intuitiv bedienen lassen. Der 'state of the art' in der Interface-Technik schickt sich an, über Kommandoschnittstellen und die allgegenwärtigen Graphical User Interfaces (GUIs) hinauszugelangen (van Dam, 1997).

Die damit unternommenen Anstrengungen haben vordringlich mit der 'Anthropomorphisierung' von Schnittstellen, also der besseren Anpassung an den Menschen zu tun. Hier gibt es noch vielfältige Herausforderungen für die Forschung über Mensch-Maschine-Kommunikation und an die Ergonomie von Schnittstellen. Es ist zu erwarten, daß komfortable Schnittstellen der Zukunft stark durch die Techniken der Multimedia und Virtual Reality sowie durch intuitive Eingabehilfsmittel, die den Gebrauch natürlicher Ausdrucksformen (direkter Zugriff mit den Händen, Sprache, Gestik, Mimik) erlauben, geprägt sein werden. Besonders im Multimedia-Bereich finden in jüngerer Zeit Gesten und Zeichensprachen als Mittel der Informationsübermittlung an maschinelle Systeme starkes Interesse, sowie auch mehrmodale Schnittstellen, die Gestik und Sprache kombinieren.

2 Gestenerkennung: Forschungsansätze

Einfach gesagt sind Gesten Körperbewegungen, die in irgendeiner Form bedeutsame Information übermitteln. Dabei werden typischerweise die oberen Gliedmaßen – Arme und Hände – und der Kopf betrachtet. Bei der Informationsübermittlung spielen nicht nur statische Konfigurationen dieser Körperteile (Posturen), sondern insbesondere dynamisch wechselnde Körperkonfigurationen eine wichtige Rolle. Neben Gestenfunktionen, die das Herstellen und Ändern von Umgebungsbedingungen durch direkte Manipulation oder die Exploration der

Beschaffenheit von Gegenständen durch taktile Eindrücke vermitteln (ergotische bzw. epistemische Funktion) (Crowley & Coutaz, 1995) ist in der gestischen Mensch-Maschine-Kommunikation besonders die Übermittlung sinnbehafteter Information durch Zeichen von Bedeutung (semiotische Funktion); diese *kommunikative* Funktion soll im Weiteren im Vordergrund stehen. Der im englischen Sprachraum verwendete Begriff "gesture" betrifft dabei einerseits Gesten im engeren Sinne, als mehr oder weniger spontane Formen körperlicher Äußerung, wie andererseits auch intendierte Bewegungen mit ausgeprägtem Zeichencharakter (Gebärden), deren Bedeutung auf Konventionen beruht. Die beiden Typen kommunikativer Gestik – spontane Bewegungen und Zeichensprachen – bilden die Extrempunkte einer kontinuierlichen Einteilung, durch die Gesten ihrer Sprachlichkeit nach angeordnet werden (Kendon, 1988).

In der maschinellen Gestenerkennung findet sich ein weites Spektrum an Aufgabenstellungen von mehr grundsätzlichem und mehr anwendungsspezifischem Charakter. Erst in Ansätzen wird bislang die syntaktische und semantische Struktur von Gestenäußerungen behandelt. Es mangelt an allgemein anerkannten Verfahren der formalen Charakterisierung von Gesten, und es ist bei weitem noch nicht die linguistisch-semiotische Grundlagenforschung erbracht worden, wie dies z.B. für die gesprochene Sprache der Fall ist. Auf der anderen Seite sind etliche anwendungsgetriebene Vorhaben unterwegs, die die signaltechnische Erfassung von Hand- und Körpergesten und ihre Kopplung in Anwendungssysteme für beschränkte, abgegrenzte Gestenrepertoires untersuchen.

Eine erste Frage betrifft die Art und Weise, wie die Gestik eines Benutzers erfaßt wird. Hier sind videobasierte Ansätze (Erfassung über eine oder mehrere Kameras) zu unterscheiden von Ansätzen, die Hand-, Arm- und Kopfbewegungen über Sensoren (Daten-Handschuh, Körper-Tracker) ermitteln. Mit kamerabasierten Verfahren läßt sich das Tragen unbequemer technischer Geräte am Körper vermeiden, jedoch erfordert die Rekonstruktion dreidimensionaler Konfigurationen aus 2D-Bildern erheblichen Aufwand, so daß Daten-Handschuhe und Tracking-Verfahren derzeit noch Geschwindigkeitsvorteile bieten. Eine Übersicht der relevanten Sensortechniken ist (Astheimer et al., 1994) zu entnehmen.

Unabhängig davon ist die Frage zu klären, wie die Signalauswertung vorgenommen wird; hier gibt es implizite Ansätze (Hidden-Markov-Modelle oder künstliche neuronale Netze), bei denen grundsätzlich alle verwendeten Gesten vorher zu trainieren sind, und explizite (wissensbasierte) Ansätze, die atomare Formelemente der Gestik beschreiben und zu größeren Einheiten zusammensetzen. Bereits recht erfolgreich sind Methoden mit trainierten 'emblematischen' Gesten, deren Bedeutung per Konvention festgelegt ist (wie Gebärdensprachen der Taubstummen oder Zeichensprachen der Kranführer). Die eingesetzten Mustererkennungsmethoden basieren grundsätzlich auf einem Vergleich von Merkmalsvektoren, die durch geeignete Vorverarbeitung der Rohdaten gewonnen werden, mit Referenzvektoren der zu erkennenden Gestenklassen. Diejenige Geste, deren Referenzvektor dem Eingabevektor am ähnlichsten ist, wird als erkannte Geste ausgegeben.

Als aktueller Überblick über den engeren Stand der Forschung sei der Ergebnisband des im letzten Jahr in Bielefeld durchgeführten internationalen Gesten-Workshops empfohlen (Wachsmuth & Fröhlich, 1998); hier sind u.a. auch Typologien dargestellt, mit denen unterschiedliche Formen und Funktionen von Gesten klassifiziert werden können.

3.1 Eigene Arbeiten

Unsere eigenen Arbeiten verfolgen einen zweistufigen, formbasierten Ansatz der Gestenerkennung. In einer ersten – applikationsunabhängigen – Stufe wird die gestische Äußerung eines Benutzers als reine Formbeschreibung der Körperbewegung analysiert und in expliziter symbolischer Notation verfügbar gemacht. In einer zweiten – applikationsspezifischen – Stufe wird die Formbeschreibung in einem aufgabenorientierten Szenario (und z.Tl. im Kontext sprachlicher Modalität; s.u.) auf verfügbare Handlungsmöglichkeiten des Anwendungssystems abgebildet. Auch wenn durch statistische bzw. trainingsbasierte Verfahren gute Ergebnisse bei fest definiertem Gestenrepertoire erzielbar sind, wird darauf in unserem Ansatz (im Anschluß an Wexelblat, 1995) zugunsten einer klaren Trennung von Gestenanalyse und -interpretation verzichtet. Hierdurch soll einerseits eine weitgehende Applikationsunabhängigkeit angestrebt und andererseits berücksichtigt werden, daß dieselbe Form einer Geste in unterschiedlichen Kontexten verschiedene Bedeutungen haben kann. Zudem sind Gestik und Sprache bezüglich ihrer Bedeutung, Funktion und ihres zeitlichen Auftretens eng verbunden (McNeill, 1992), so daß ein expliziter Ansatz auch im Hinblick auf eine Integration von Gestik und Sprache Vorteile verspricht. Für die Gestenerkennung wird in unseren Arbeiten ein symbolbasierter Ansatz verfolgt. Als Eingabe dienen nicht Sensor-Rohdaten, sondern in der Signalvorverarbeitung abstrahierte, zusammengesetzte Symbole, die Teilaspekte einer Geste als Ereignis in Raum und Zeit repäsentieren, unabhängig von der verwendeten Sensortechnik (merkmalsbasierte Erkennung).

Als Grundlage für eine Form- und Bewegungsbeschreibung von Gesten wurde das aus der Gebärdensprachenlehre hervorgegangene Hamburger Notations-System 'HamNoSys' (Prillwitz et al., 1989) ausgewählt. Es vereint eine strukturierte Symbolik mit recht einfach herleitbaren Symbolen und einem Fokus auf der Darstellung der oberen Gliedmaßen. Gesten werden in HamNoSys als Wörter notiert, die aus Grundsymbolen zusammengesetzt sind. Da der Ursprung von HamNoSys in der Notation von Gebärden liegt, wurde beim Design sinnvoll zwischen Formmerkmalen und Bedeutungseinheiten unterschieden; zudem sind die Formmerkmale in einem technisch traktablen System hierarchisch untergliedert, was einen systematischen Ansatz unterstützt.

HamNoSys gestattet u.a. die Notierbarkeit von allen Zeichen in allen Zeichensprachen, eine logische und anatomisch konsistente Klassifikation von Handformen, die Reduktion von Grundsymbolen auf ein Minimum sowie eine formalsprachliche Charakterisierung von Symbolstrukturen, die eine maschinelle Verwendung begünstigt. Das Alphabet der Grundsymbole ist festgelegt und umfaßt etwa 200 Zeichen. Der Zeichenvorrat läßt sich in diverse Klassen einteilen, die jeweils einen Aspekt der Geste, z.B. die Handform oder eine Bewegung erfassen. Mit den Kombinationsmöglichkeiten der Grundsymbole lassen sich statische Konfigurationen (Posturen) und dynamische Gestenteile (Aktionen) beschreiben. Bewegungen können translatorisch oder stationär sein. Bei einer translatorischen Bewegung bewegt sich die Hand als Ganzes, bei einer stationären Bewegung ändert sich die Position der Hand nicht, sondern nur ihre Form. Aktionen lassen sich kombinieren (so können selbst solch komplexe Gesten notiert werden wie die Andeutung eines Quadrats durch "Nachzeichnen" seiner Begrenzungslinien). In Abb. 1 ist beispielsweise eine Rechtszeigegeste notiert, bestehend aus einem statischen Anteil der Hand-Arm-Konfiguration (Hand mit Zeigefinger nach vorn gestreckt, Handfläche nach links orientiert, Arm in Schulterhöhe ganz gestreckt) und einem dynamischen Anteil (Hand erst nach vorn, dann nach rechts bewegt).

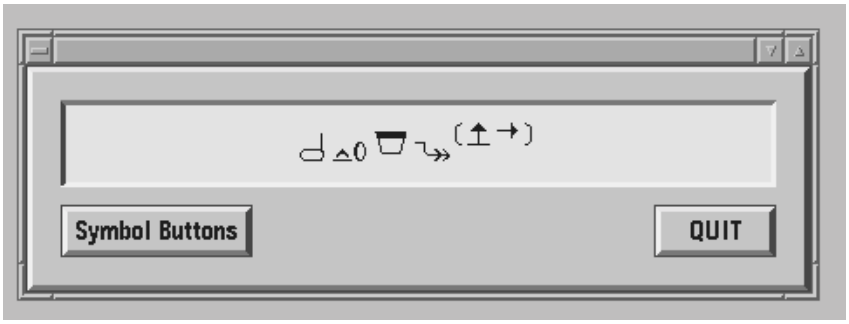

Abbildung 1: HamNoSys-Darstellung einer dynamischen Rechtszeigegeste (ausgestreckte Zeigehand, nach vorn und nach rechts bewegt)

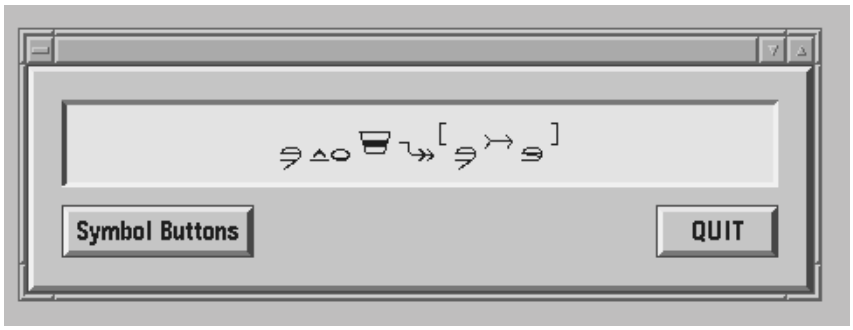

Abbildung 2: HamNoSys-Darstellung einer dynamischen Greifgeste (offener Griff → geschlossener Griff)

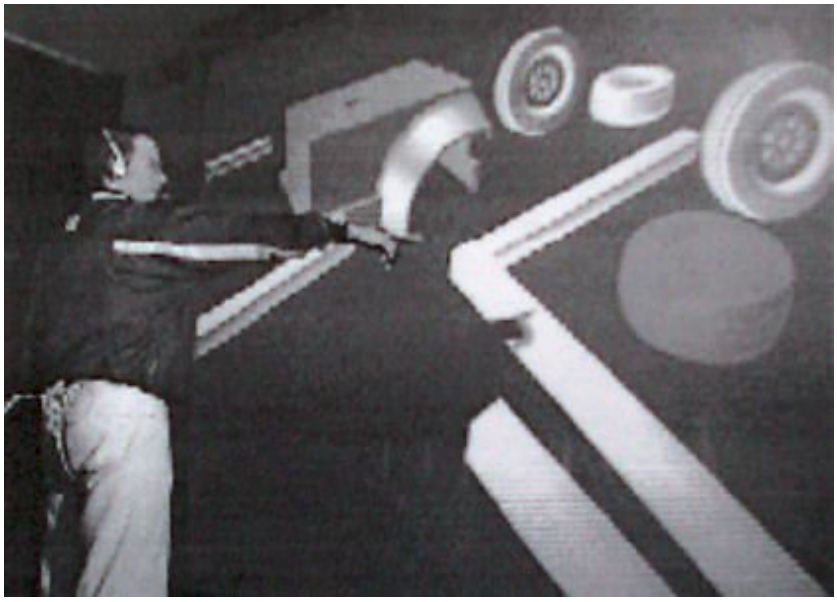

Abbildung 3: Zeigegeste zur Objektauswahl im Szenario der Virtuellen Konstruktion, HamNoSys-Darstellung:

In Abb. 2 ist eine Greifgeste notiert (Hand-Arm-Konfiguration mit geöffnetem Griff nach vorn, Handfläche nach unten orientiert, Arm in Brusthöhe ganz gestreckt, stationäre Bewegung vom geöffneten zum geschlossenen Griff). Abbildung 3 zeigt die Verwendung einer Zeigegeste in unserem Forschungsszenario der Virtuellen Konstruktion (siehe Abschnitt 4); die Interpretation der Geste führt hier zur Auswahl eines virtuellen Objektes, auf das sich nachfolgende Instruktionen beziehen können.

Für unsere Arbeiten wurde eine Teilmenge HNS' des verfügbaren HamNoSys-Zeichenvorrats, die sich mit der in unserem Labor verwendeten Sensorik (s.u.) gut erfassen läßt, ausgewählt. Mit Hilfe eines Symboleditors, der über die Taste "Symbol Buttons" (Abb. 1; 2) aktiviert wird, lassen sich Gestenwörter als Symbolketten über HNS' definieren, mit denen komplexe Gesten – wie oben angedeutet – beschrieben werden. Die Syntax wurde mit einer Backus-Naur-Grammatik spezifiziert, für die mit dem in der Unix-Welt verfügbaren Parser-Generator "bison" ein Syntax-Parser erzeugt wurde. Die Semantik für HNS'-Wörter ist kompositionell und besteht aus Symbolausdrücken, die mit einem Regelkonstruktor in CLIPS (C Language Integrated Production System)-Regeln überführt werden. Ein Software-Agent HamNoSysIO für die Definition in HNS' ausdrückbarer Gesten wurde implementiert und steht für experimentelle Anwendung zur Verfügung (Sowa, 1998). Als Beispiel zeigt Abbildung 4 die CLIPS-Regeln, die für die in Abb. 1 dargestellte dynamische Zeigegeste automatisch generiert wurden; die – in Abb. 1-3 mit Pretty-Printer dargestellten – HamNoSys-Symbole werden intern durch sprechende Namen als ASCII-strings notiert.

```
(defrule rule32
?tmp28 <- (object (is-a MoveA))          ;      eine Handvorwärtsbewegung
?tmp31 <- (object (is-a MoveR))          ;      eine Handrechtsbewegung
=>
(SEQUENCE tmpcl25 ?tmp28 ?tmp31))        ;      in Sequenz
```

Hier ist automatisch eine CLIPS-Regel `rule32` generiert worden, die ein Token der Klasse `tmpcl25` assertieren kann, das eine erkannte Sequenz von bestimmten Handbewegungen registriert.

```
(defrule rule36
?tmp16 <- (object (is-a BSifinger))      ;      Zeigefinger gestreckt
?tmp19 <- (object (is-a EFinA))          ;      nach vorn gestreckt
?tmp22 <- (object (is-a PalmL))          ;      Handfläche links
?tmp23 <- (object (is-a LocShoulder))    ;      in Schulterhöhe
?tmp24 <- (object (is-a LocStretched))   ;      ausgestreckt
?tmp35 <- (object (is-a tmpcl 25))       ;      bewegt wie oben
=>
(PARALLEL DUMMY ?tmp16 ?tmp19 ?tmp22 ?tmp23 ?tmp24 ?tmp35))
```

Hier ist automatisch eine Regel `rule36` generiert worden, die ein Token der Klasse `DUMMY` assertieren kann, das das Auftreten einer bestimmten Handpostur in Verbindung mit oben beschriebener Bewegung registriert (für `DUMMY` kann ein beschreibender Gestenname gesetzt werden).

Abbildung 4: Automatische Generierung von Regeln für HamNoSys-Wörter

Die Erfassung von Körpergesten geschieht in unserem Labor über mehrere elektromagnetische Positionsensoren, die die sechs räumlichen Freiheitsgrade der Bewegung und Rotation messen. Um Unzulänglichkeiten der Sensorik (zuwenige Sensoren; nur 6D-Position, keine weiteren Momente) ausgleichen zu können und eine definierte, normierte Schnittstelle zu nachfolgenden Verarbeitungsstufen zu haben, wurde ein abstraktes Körpermodell entwickelt. Es besteht aus einem geometrischen Modell auf der Basis von DIN-Normdaten (DIN, 1990) und weiteren Quellen und einem dynamischen Modell, das sich eines rekurrenten künstlichen neuronalen Netzes bedient; im einzelnen siehe (Fröhlich & Wachsmuth, 1998).

3 Multimodale Mensch-Maschine-Kommunikation

Eine Erhöhung der Interaktionsbandbreite mit technischen Systemen ist erreichbar, wenn die Systemschnittstelle den parallelen Einsatz mehrerer Interaktionstechniken zuläßt, die auch als 'Modalitäten' der Kommunikation bezeichnet werden. Im technischen Sinn ist eine Modalität durch ein physikalisches Gerät und eine Interaktionssprache gekennzeichnet, z.B. Mikrofon und natürliche Sprache im Falle der sprachlichen Modalität (Coutaz et al., 1995). Der ebenfalls in diesem Zusammenhang benutzte Begriff Medium bezieht sich auf einen Informationsträger oder -übertragungskanal. Systeme, die über mehr als eine Modalität verfügen, heißen multimodale Systeme. Als physikalisches Gerät für die gestische Modalität können Tracker oder auch Videokameras herangezogen werden. Interaktionssprachen für gestische Kommunikation sind bislang kaum betrachtet worden, jedoch könnte das im Abschnitt 2 betrachtete HamNoSys als solche gewertet werden.

Von zunehmenden Anwendungsinteresse sind derzeit 'koverbale' Gesten, also Gesten, die sprachliche Äußerungen mehr oder weniger spontan begleiten, z.B. wenn auf einen Gegenstand gezeigt wird ("dieses Rad") oder eine Drehrichtung ("so herum") signalisiert wird. Als Herausforderung stellt sich dabei die Koordination multimodaler Eingaben, insbesondere die zeitliche Kopplung der beiden Modalitäten natürlicher Sprache und Gestik.

Unbefriedigend ist in der bisherigen Forschung in der multimodalen Mensch-Maschine-Kommunikation vor allem die bislang geringe Flexibilität bei der Interaktion mit multimodalen Systemen, die weitgehend durch eine starr vorgegebene Zuweisung von Interaktionsinhalten an Interaktionsmodalitäten gekennzeichnet sind. Auf der anderen Seite haben psycholinguistische Untersuchungen gezeigt, daß Gestik und Sprache bezüglich ihrer Bedeutung, Funktion und ihres zeitlichen Auftretens eng verbunden sind (McNeill, 1992). Die zeitliche Ausführung einer Geste unterteilt sich in mehrere Phasen, von denen die expressive Phase ('Stroke') die wichtigste ist. Der Stroke ist häufig durch einen abrupten Halt gekennzeichnet, der mit koexpressiven Wörtern zeitlich in enger Beziehung steht. Experimente haben ergeben, daß ein Wort nicht vor dem Stroke der koverbalen Geste geäußert wird, sondern entweder gleichzeitig oder eine Silbe oder ein Wort verspätet.

Eine zentrale Forschungsfrage ist somit die zeitliche Integration der unterschiedlichen Modalitäten Sprache und Gestik, die technisch als nebenläufige Sprach- und Gestenperzepte auf getrennten Kanälen registriert werden und für die Steuerung von Anwendungen integriert und interpretiert werden müssen. Ähnlich wie bei der Spracherkennung besteht die typische Verar-

beitungssequenz für Gestenerkennungssysteme aus drei Phasen. Zunächst wird ein Datenstrom sensortechnisch aufgenommen, daraufhin durch Software analysiert und anschließend interpretiert. Die bei der Signalerfassung aufgenommenen Meßdaten werden von verschiedenen Arbeitsstationen weiterverarbeitet. Dabei kommt es zu Zeitverzögerungen, die den Zeitpunkt, an dem die Meßergebnisse vorliegen, und den Meßzeitpunkt voneinander abweichen lassen. Für die Interpretation der so erhaltenen Meßergebnisse ist es aber wichtig, den inhaltlichen Zusammenhang wieder herzustellen. Dafür ist es nötig, zunächst einmal einen zeitlichen Zusammenhang zu ermitteln. Technische Verfahren müssen das Zeitverhalten schon deshalb erfassen, damit die Integration des Zeichenhaften (z.B. Zeigegeste) mit dem Signalgehalt (z.B. Zeigevektor im Moment der maximalen Gestenexpression) rekonstruiert werden kann.

Für diese Aufgabe kann man aufwendige temporale Logiken oder, wie in unseren Arbeiten favorisiert, zeitintervallbasierte Verfahren verwenden. Anhaltspunkte ergeben sich durch Forschungsbefunde aus den Humandisziplinen, die zeigen, daß menschliche Kommunikation durch signifikant 'rhythmische' Muster geprägt ist, die als koordinative Strategie des menschlichen Äußerungs- und Wahrnehmungsapparats gedeutet werden (Condon, 1986); (Fant & Kruckenberg, 1996); (Cummins & Port, 1996). Ferner lassen sich Untersuchungen anschließen, die sich mit der zeitlichen Segmentierung und der präsemantischen Integration perzeptorischer Einheiten im neuronalen Verarbeitungsapparat des Menschen auseinandersetzen (Pöppel, 1997). Hierauf aufbauend wurde in einer kürzlich abgeschlossenen Dissertation (Lenzmann, 1998) ein erster Ansatz für ein zeit- und ereignisgesteuertes Integrationsverfahren für multimodale (sprachlich-gestische) Eingaben realisiert. Das Verfahren beruht auf einer zeitzyklengetriebenen Modalitätsintegration, in welcher Eingaben in abgestimmten festen Zeitscheiben segmentiert werden, womit als Nebeneffekt die offene Eingabe (bei der Anfang und Ende einer Interaktion nicht bekannt sind) bewältigt werden kann. Dieser Ansatz ist ein wegweisender Ausgangspunkt für unsere weiteren Arbeiten zur multimodalen Integration.

4 Ein Applikationsszenario: Virtuelles Konstruieren

Eine Erprobungsgrundlage für unsere bisherigen Arbeiten ist der in unserem Labor entwickelte Virtuelle Konstrukteur, ein mit Maus- und Spracheingaben instruierbares Montagesimulationssystem. Der Virtuelle Konstrukteur ermöglicht auf einer bildschirmpräsentierten virtuellen Werkbank die interaktive Montage komplexer Aggregate aus einfachen Grundbausteinen. Durch eine Stereo-Brille kann der Betrachter die dreidimensionale Struktur der dargestellten Objekte erfassen und durch Mausmanipulationen und durch Anweisungen in natürlicher Sprache interaktiv verändern. Die Spracheingaben können auf alles Sichtbare – Farben, Formen, Positionen, aber auch Namen von Teilen und Baugruppen – bezugnehmen; mit der Maus kann das gezeigte Objekt gedreht und verschoben werden, oder es können die Teile von Aggregaten bewegt werden. Die Aktionen des Systems, beispielsweise beim Zusammenstecken von Teilen, werden durch gesampelten Stereo-Sound hörbar gemacht, wodurch nicht nur der Realitätseindruck verstärkt, sondern auch ein räumlich gerichtetes auditives Feedback für die Systemaktionen geschaffen wird.

Der Kern des Virtuellen Konstrukteurs ist ein wissensbasiertes Repräsentationssystem, das im DFG-Sonderforschungsbereich 360 (Rickheit & Wachsmuth, 1996) im Projekt CODY entwikkelt wurde. Er umfaßt einen einfachen Parser für sprachliche Instruktionen sowie verschiedene Wissensbasen, in denen Modellwissen für die jeweilige Anwendung bereitgestellt wird. Durch eine integrierte 'Konzept-Dynamik' kann das System zeitveränderliche konzeptuelle Strukturbeschreibungen von dargestellten Objekten und Aggregaten maschinell verarbeiten (Wachsmuth & Jung, 1996). Hierfür setzen wir Agenten- und Expertensystemtechniken ein; siehe z.B. (von Bechtolsheim, 1993). Das grundsätzliche Vorgehen ist es, die Anwendungsdomäne geeignet zu *konzeptualisieren*, d.h. die in der Benutzerinteraktion beeinflußbaren Stellgrößen begrifflich zu fassen, und ihre Zusammenhänge mit der Änderung von Systemparametern in Form von Wenn-Dann-Regeln zu beschreiben. Durch solche Inferenzregeln werden die die einzelnen Objekte und Aggregate beschreibenden Konzepte an den jeweiligen Montagestand dynamisch angepaßt; ausgewertet werden auch die aus den CAD-Modellen erschließbaren geometrischen und räumlichen Informationen (Jung & Wachsmuth, 1996). Mit der Integration und Nutzung von Objektwissen zielt diese Anwendung auf die intelligente Unterstützung von Entwurfsaktivitäten ab und erschließt die Herstellung virtueller Prototypen von CAD-basierten Konstruktionen, die eine Produktentwicklung "ohne Materialverbrauch" erlauben soll.

3.1 Virtuelles Konstruieren mit Gestik und Sprache

In dem Bielefelder Projekt SGIM ('Sprach- und Gesten-Interfaces für Multimedia') liegt seit 1996 der Forschungsschwerpunkt auf der Zusammenführung von gestischer und sprachlicher Kommunikation in einem Szenario des Virtuellen Konstruierens. Der Einsatz von Gesten in VR-Applikationen ist an sich nicht neu; so werden seit längerem bereits konventionalisierte Gesten zur Navigation in virtuellen Umgebungen benutzt; siehe etwa (Väänanen & Böhm, 1991) oder (Astheimer et al., 1994). Selbst die Kombination gestischer und sprachlicher Eingaben ist sehr frühzeitig erprobt worden (z.B. Weimer & Ganapathy, 1989). Wie bringen nunmehr unsere Arbeiten zur Sprach- und Gesteninteraktion, die eine weitgehende Unabhängigkeit von konventionalisierten Kommunikationsformen zum Ziel hat, in ein VR-Szenario ein. In unserem Labor werden dazu Stereoprojektionen auf der Großbildwand eingesetzt (siehe Abb. 3), die hochaufgelöste räumliche Darstellungen der CAD-Modelle eines Kleinfahrzeugs in realistischer Größe gestatten und die mit Eingabegeräten der Virtuellen Realität (Körper-Tracker, Datenhandschuhe, Spracherkennungssystem) manipuliert werden. Auf dem Virtuellen Konstrukteur wird die Erstellung von Entwurfsvarianten eines "City-Mobils" durch koverbale Gestik gesteuert.

Die Erkennung von Zeige- und Rotationsgesten beim Zusammenfügen von Teilen und Baugruppen wird mit Hilfe von Agententechniken vorgenommen. Durch einfache Mikrofon-Spracheingabe, mit der Objekttypen oder Positionen spezifiziert werden, wird die Interaktion unterstützt, um bedeutete Objekte oder Richtungen auch dort analysieren zu können, wo direkte Gestik unnatürlich ist oder an technische Grenzen stößt. Die parallele Sprachanalyse dient in erster Linie zur Auflösung von Mehrdeutigkeiten des gestischen Eingabekanals. Am Beispielsatz *"Montiere dieses Rohr links an den Rahmen"* wird deutlich, daß zur Auflösung der Referenzen weitere Informationen notwendig sind. Eine Zeigegeste auf ein Objekt vom Typ Rohr sowie Kenntnis über den Standort des Benutzers im virtuellen Raum können hier die notwendigen Informationen liefern. Die Kombination der beiden Modalitäten Gestik und

Sprache kann in vielen Fällen Mehrdeutigkeiten auflösen, das heißt dazu beitragen, unscharfe, vage Informationen der einzelnen Modalitäten zu konkretisieren.

In den bisherigen Arbeiten wurde zunächst die technische Infrastruktur zur Rohdatenerfassung der Eingaben geschaffen, und es wurden Techniken entwickelt, welche über die Auswertung von elektromagnetischen wie auch elektrischen Sensoren Informationen über die räumliche Bewegungsrichtung der oberen Extremitäten und der Positionierung eines Benutzers ermitteln. Hierdurch ist dem System zu jeder Zeit der Standort des Betrachters bekannt, so daß bei Spracheingaben zum Beispiel Begriffe wie *links* oder *vorn* intuitiv benutzt werden können. Bereits ausgearbeitet wurden Verfahren für die Erkennung entfernter Zeigegestik (Latoschik & Wachsmuth, 1998) und die Übersetzung registrierter Körperbewegungen in symbolisch verarbeitbare Darstellungen; siehe (Fröhlich & Wachsmuth, 1998). Aufgrund der in Interaktionsszenarien der Virtuellen Realität erforderlichen geringen Latenzzeit für die Gesten- und Spracherkennung wird im SGIM-Projekt eine hochparallele integrierte Verarbeitung, ermöglicht wiederum durch eine Agentenarchitektur, eingesetzt. Hiermit wurde ein erster Prototyp eines lokalen Zeigegestenerkenners fertiggestellt, der mit dem Virtuellen Konstrukteur gekoppelt ist. Laufende Arbeiten, die zum Teil in Kooperation mit Partnern im Multimedia-Verbund NRW durchgeführt werden, betreffen unter anderem die Einbindung verbesserter Spracherkennungssysteme und die weitergehende Realisierung unserer exemplarischen Anwendung aus dem Bereich des virtuellen Konstruierens.

Dank und Hinweise

Die hier dargestellten Arbeiten profitieren wesentlich von den Forschungsbeiträgen der Mitarbeiter, Doktoranden und Diplomstudenten: Martin Fröhlich, Timo Sowa, Marc Latoschik, Bernhard Jung, Britta Lenzmann, Martin Hoffhenke, Sebastian Hübner, Stefan Kopp, Björn Knafla sowie Liudger Franzen. Ihnen allen sei herzlich gedankt.

Das SGIM-Projekt wird durch Zuwendungen im Forschungsverbund "Multimedia NRW: Die Virtuelle Wissensfabrik" und im Graduiertenkolleg "Aufgabenorientierte Kommunikation" seit 1996 gefördert. Das CODY-Projekt, erwähnt im Zusammenhang mit dem Virtuellen Konstrukteur, wird von der Deutschen Forschungsgemeinschaft im Sonderforschungsbereich 360 seit 1993 gefördert.

Literatur

Astheimer, P., Böhm, K., Felger, W., Göbel, M. & Müller, S. (1994). Die Virtuelle Umgebung – Eine neue Epoche in der Mensch-Maschine-Kommunikation; Teil I: *Informatik-Spektrum 17*(5), 281-290; Teil II: *Informatik-Spektrum 17*(6), 357-367.

Barfield, W. & Furness, T.A. (Eds.) (1995). *Virtual Environments and Advanced Interface Design*, Oxford University Press, 1995.

Condon, William S. (1986): Communication: Rhythm and Structure. In James Evans & Manfred Clynes (Eds.): *Rhythm in Psychological, Linguistic and Musical Processes (pp. 55-77)*. Springfield, Ill.: Thomas.

Coutaz, J., Nigay, L. & Salber, D. (1995). Multimodality from the User and System Perspectives. *Proceedings ERCIM-95 Workshop on Multimedia Multimodal User Interfaces.*

Crowley, J.L. & Coutaz, J. (1995). Vision for Man Machine Interaction. *Proceedings of Engineering Human Computer Interaction (EHCI'95)*, London: Chapman and Hall, 28-45.

Cummins, F. & Port, R.F. (1996). Rhythmic Commonalities Between Hand Gestures and Speech. In *Proceedings of the Eighteenth Meeting of the Cognitive Science Society (pp. 415-419)*, Lawrence Erlbaum Associates.

DIN (1990). DIN 33402 T2 Bbl. 1: Körpermaße des Menschen; Werte; Anwendung von Körpermaßen in der Praxis. In Deutsches Institut für Normung e.V. (Hrsg.): *Bildschirmarbeitsplätze: Normen und Sicherheitsregeln, Kap. 5 (pp. 197-224)*. Jgg. 12 d. DIN-Taschenbücher, Köln: Beuth.

Fant, G. & Kruckenberg, A. (1996). On the Quantal Nature of Speech Timing. *Proc. ICSLP 1996*, pp. 2044-2047.

Fröhlich, M. & Wachsmuth, I. (1998). Gesture Recognition of the Upper Limbs – From Signal to Symbol. In I. Wachsmuth & M. Fröhlich (eds.): *Gesture and Sign Language in Human-Computer Interaction (pp. 173-184).* Berlin: Springer (LNAI 1371).

Jung, B. & Wachsmuth, I. (1996). Ein wissensbasiertes System für die 3D-computergraphische Montage-Simulation. In D. Ruland (Ed.): *Verteilte und intelligente CAD-Systeme: Tagungsband CAD '96 (pp. 107-119)*. Bonn: Gesellschaft für Informatik; Kaiserslautern/Saarbrücken: DFKI.

Kendon, A. (1988). How Gestures Can Become Like Words. In F. Poyatos (Ed.): *Cross-cultural Perspectives in Nonverbal Communication.* Toronto: Hogrefe.

Latoschik, M. & Wachsmuth, I. (1998). Exploiting Distant Pointing Gestures for Object Selection in a Virtual Environment. In I. Wachsmuth & M. Fröhlich (eds.): *Gesture and Sign Language in Human-Computer Interaction (pp. 185-196).* Berlin: Springer (LNAI 1371).

Lenzmann, B. (1998). *Benutzeradaptive und multimodale Interface-Agenten.* Dissertation, Technische Fakultät, Universität Bielefeld; ersch. als DISKI-Band 184, Sankt Augustin: Infix.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* Chicago: University of Chicago Press.

Pöppel, E. (1997). A Hierarchical Model of Temporal Perception. *Trends in Cognitive Sciences 1*(2), 56-61.

Prillwitz, S., Leven, R., Zienert, H., Hamke, T. & Henning, J. (1989). *HamNoSys Version 2.0: Hamburg Notation System for Sign Languages: An Introductory Guide*, Vol. 5 of International Studies on Sign Language and Communication of the Deaf, Signum Press.

Rickheit, G. & Wachsmuth, I. (1996). Collaborative Research Centre "Situated Artificial Communicators" at the University of Bielefeld, Germany. *Artificial Intelligence Review 10*(3-4), 65-170.

Sowa, T. (1998). Ein wissensbasierter Ansatz zur Integration zeitsensitiver Information und seine Anwendung auf die Gestenerkennung. Diplomarbeit, Universität Bielefeld: Technische Fakultät.

Väänänen, K. & Böhm, K. (1991). Gesture-Driven Interaction as a Human Factor in Virtual Environments - An Approach with Neural Networks. In M. A. Gigante und H. Jones (Eds.): *Virtual Reality Systems*, Academic Press.

van Dam, A. (1997). Post-WIMP User Interfaces. *Communications of the ACM 40*(2), 63-67.

von Bechtolsheim, M. (1993). *Agentensysteme – Verteiltes Problemlösen mit Experten-systemen.* Braunschweig: Vieweg.

Wachsmuth, I. & Fröhlich, M. (eds.) (1998). *Gesture and Sign Language in Human-Computer Interaction (Proceedings International Gesture Workshop, Bielefeld, Germany, September 1997).* Berlin: Springer (LNAI 1371).

Wachsmuth, I. & Jung, B. (1996). Dynamic Conceptualization in a Mechanical-Object Assembly Environment. *Artificial Intelligence Review 10(3-4)*, 1996, 345-368.

Weimer, D. & Ganapathy, S.K. (1989). A Synthetic Visual Environment with Hand Gesturing and Voice Input. In: *Human Factors in Computing Systems – Proceedings CHI'89 Conference*, 235-240.

Wexelblat, A.D. (1995). An Approach to Natural Gesture in Virtual Environments. *ACM Transactions on Computer-Human Interaction 2*(3), 179-200.