Multimodale Interaktion in Mensch-Maschine-Systemen

Ipke Wachsmuth

Schlüsselwörter: Künstliche Intelligenz, Virtuelle Realität, Virtual Prototyping, Multimodale Interaktion, Intelligente virtuelle Assistenten.

Zusammenfassung

In diesem Beitrag werden aktuelle Arbeiten im Bielefelder Labor für Künstliche Intelligenz und Virtuelle Realität aus zwei Bereichen multimodaler Mensch-Maschine-Systeme im Überblick vorgestellt. Im Projekt „Virtuelle Werkstatt“ geht es um eine integrierte Plattform zur multimodal-interaktiven Erstellung, Modifikation und funktionalen Überprüfung von Prototypen mechanischer Objekte in der virtuellen Realität. Dabei werden hochaufgelöste räumliche Visualisierungen CAD-basierter Bauteilmodelle in realistischer Größe projiziert und über multimodale Eingaben (vermittelt durch Datenhandschuhe, Positionssensoren, Spracherkenner) zu komplexen Aggregaten zusammengefügt oder modifiziert. Erprobungsdomäne ist das virtuelle Konstruieren von CAD-Teilen und Baugruppen eines Kleinfahrzeugs. Im zweiten Teil wird der computeranimierte artikulierte Agent „Max“ vorgestellt – ein intelligenter virtueller Assistent, der mit Benutzern Dialoge über das Bauen mit Baukastenteilen führt, sprachliche und gestische Instruktionen versteht und sich selbst mehrmodal in synthetischer Sprache, Gestik und Mimik äußert. Methoden der Künstlichen Intelligenz kommen sowohl zur Unterstützung der Interaktion mit computergraphischen Objekten wie auch bei der Auswertung bzw. Erzeugung multimodaler Ein- und Ausgaben zum Einsatz.

Neue Formen der Mensch-Maschine-Interaktion

Die Interaktion des Menschen mit rechnergestützten Systemen erfordert Informationsübertragung in zwei Richtungen: vom Rechner zum menschlichen Benutzer und umgekehrt. In der ersten Richtung hat die moderne Multimedia-Technologie erhebliche Fortschritte gebracht. Information wird nicht mehr nur als Text angeboten, sondern auch in Form von Grafik, Bildern, Bildsequenzen und Ton. Die verschiedenen Sinnesmodalitäten, in denen Menschen Informationen aufnehmen können, werden immer besser genutzt und kombiniert. Der Mensch sollte aber in der Interaktion mit dem Rechner nicht nur seine (passiven) visuellen und auditiven Fähigkeiten, sondern auch seine (aktiven) kommunikativen Fähigkeiten voll entfalten können. Dazu müssten die Rechner Informationen nicht nur multimedial präsentieren, sondern auch multimodal aufnehmen können. Die meisten Rechner reagieren aber nur auf Tastatureingaben und Mausbewegungen. Die natürlichen kommunikativen Fähigkeiten der Benutzer liegen nicht nur brach, sondern die Benutzer werden gezwungen, sich kommunikativer Techniken zu bedienen, die sie in der gewöhnlichen Interaktion mit anderen Menschen nicht benötigen und die sie gesondert erlernen müssen.

Daher ist es ein Gesamtziel unserer Arbeiten im Bielefelder Labor für Künstliche Intelligenz und Virtuelle Realität, menschgerechte Formen der Interaktion mit technischen Systemen zu entwickeln. Es sollen natürlichere, „anthropomorphe“ Mensch-Maschine-Schnittstellen konzipiert werden, die aus kontextbezogenen und sogar unscharfen Eingaben ihrer Benutzer verwertbare Informationen beziehen können. Dazu müssen die traditionellen Eingabemöglichkeiten (Maus, Menüs, Kommandosprachen) um intuitiv naheliegendere Mittel wie natürliche gesprochene Sprache und Gestik erweitert werden. Mit dem Ziel, auch die systemseitige Generierung mehrmodaler Ausgaben einzubeziehen, sind in menschlicher Gestalt verkörperte virtuelle Agenten eine weitere Stoßrichtung, um damit anthropomorphe Assistenzsysteme mit vielfältigen Anwendungsperspektiven zu entwickeln.

Die Entwicklung multimodaler Interaktionstechniken erfordert die Verknüpfung der Ergebnisse unterschiedlicher Forschungsrichtungen, und zwar in Anwendungsfeldern, in denen die Vorteile der neuartigen Methoden sinnvoll zum Tragen kommen können. Im Konstruktionsbereich etwa ist die rechnergestützte Darstellung synthetischer Geometriedaten durch Techniken der Virtuellen Realität (VR) zunehmend wichtig. Für die Anwendung sind besonders VR-Systeme interessant, mit denen sich Modelle realer Objekte und deren Herstellungsprozesse bereits in der Konzeptphase realistisch darstellen und explorieren lassen, vor dem Bau eines physikalischen Produktmodells (Physical Mock-Up). Die Modellierung am

Computer ermöglicht einerseits die Übernahme vorhandener CAD-Modelldatenbanken und andererseits ein Probehandeln ohne Materialverbrauch und mit leichter Veränderbarkeit des (immateriellen) Modells.

Hier bildet die Unterstützung der VR-Technik durch Methoden der Künstlichen Intelligenz, insbesondere durch semantisch verarbeitbare Repräsentationen, den Ausgangspunkt für ein „virtuelles Konstruieren“, d.i. die Erstellung und Erprobung computergraphisch visualisierter 3D-Modelle geplanter mechanischer Konstruktionen in sog. virtuellen Prototypen (Digital Mock-Ups). Über die üblichen Interaktionen zur Navigation und Objektbewegung hinaus sind dabei gestische und sprachliche Eingaben als Mittel der komfortablen Anwendungssteuerung von Interesse. Kommen des weiteren anthropomorphe Assistenzsysteme hinzu, die dem Benutzer kontextbezogene Hilfen anbieten können sollen, so erfordert ein solches Gesamtszenario die Zusammenführung mindestens folgender Teilbereiche:

- Anwendung (technisches System),
- Systemausgabe/Feedback für den Benutzer,
- Erkennung und Generierung von Gestik,
- Erkennung und Generierung von Sprache,
- Integration multimodaler Eingaben und Ausgaben.

In den folgenden Abschnitten werden unsere Arbeiten in diesen Bereichen im Einblick erläutert.

Konstruieren in der virtuellen Werkstatt

In unserem mit Unterstützung der Deutschen Forschungsgemeinschaft (DFG) im Jahr 2001 begonnenen Projekt „Virtuelle Werkstatt“ [1] wurde als Anwendung und Erprobungsdomäne der Zusammenbau eines „Citymobils“ gewählt, eines elektrobetriebenen Kleinfahrzeugs für den Innenstadtverkehr. Bei der realen Konstruktion eines solchen Mobils gibt es eine erhebliche Anzahl unterschiedlicher Varianten. Abhängig vom speziellen Einsatzzweck oder von nutzerspezifischen Anpassungen sind diverse Modifikationen möglich, die sich auf die Gesamtauslegung des Fahrzeugs auswirken. Die Entwurfsvarianten des Citymobils werden in unserem Szenario nicht real, sondern virtuell hergestellt: Die Konstruktion wird mit CAD-Modellen der Bauteile simuliert und kann – ohne physischen Materialverbrauch – immer wieder geändert werden. Die Veränderung der Entwürfe erfolgt mittels zweihändiger manipulativer Gestik, ergänzt durch Möglichkeiten der kommunikativen Gestik und sprachlicher Eingaben. Hinzu kommen u.a. wissensgestützte, physikrekonstruierende Einpasshilfen zur Manipulation virtueller Objekte, die etwa ein passgenaues Zusammenschnappen zweier virtueller Bauteile leisten.

Grundsätzlich werden im Demonstratorsystem die Besonderheiten der VR wie 1:1-Darstellung, visuell-auditive Präsentation und Immersion ausgenutzt. Unsere Cave-artige 3-Seitenprojektion (zwei Wände und ein Boden bilden eine gemeinsame Ecke) der Firma 3Dims arbeitet auf Basis von 6 D-ILA Projektoren mit einer Auflösung von je 1365x1024. Zur Erzeugung eines von der Kopfneigung des Betrachters unabhängigen Stereoskopieeffektes kommen Zirkularpolarisationsfilter zum Einsatz. Ein optisches Tracking-System der Firma ART arbeitet Marker-basiert und gewährleistet eine effektive Datenrate von 60Hz mit hoher Güte. Als weitere Eingabegeräte werden zwei kabelfreie Datenhandschuhe sowie ein Funkmikrophon eingesetzt.

Als Simulations- und Render-System dient ein über Hochgeschwindigkeitsnetz (Myrinet) verbundenes, unter Linux betriebenes Clustersystem der Firma Artabel. Dieses besteht aus insgesamt 6 Doppelprozessor-Compute-Servern und 8 synchronisierten Render-Nodes (mit Nvidia-Grafik) auf der Basis von PC-Technologien. Die Verteilung der Graphikprimitive auf die Render-Nodes erfolgt über einen speziellen OpenGL Layer.

Zielsetzung beim virtuellen Konstruieren ist es, dass alle im Realen aus einer gegebenen Menge von Grundbauteilen physikalisch konstruierbaren Aggregate auch in der virtuellen Umgebung herstellbar sind. Neben den Standardinteraktionen herkömmlicher Graphiksysteme wie Navigation und Objekttranslation erlaubt unser System die Echtzeitsimulation montagebezogener Manipulationen: passgenaues Fügen und Trennen von Bauteilen und Aggregaten und die Modifikation erzeugter Aggregate durch Relativbewegung (Rotation und Translation) von Bestandteilen gemäß verbindungsartspezifischer Freiheitsgrade.

Die Benutzerinteraktion mit der visualisierten 3D-Szene erfolgt multimodal mit Hilfe von sprachbegleiteten Gesteneingaben. Dazu wurden Grundlagentechniken entwickelt, die mittels 6DOF- und Bimetallsensoren präzise Informationen über die Bewegungsrichtung der oberen Extremitäten und die Position eines Benutzers bei der Interaktion in der VR-Umgebung vermitteln. Sie betreffen die signaltechni-

sche Erfassung und Bedeutungsanalyse von Körpergestik (vor allem Hände, Arme und Kopfstellung des Benutzers), die Analyse von Spracheingaben, die Integration der Gesten und Spracheingaben sowie die Kopplung in das Echtzeit-Anwendungssystem. Bei der Umsetzung dieser Aufgaben wurde insbesondere Wert auf eine kompositionelle Zerlegung der benötigten Funktionen unter Berücksichtigung und Erweiterung aktueller Standards zur VR-Modellierung gelegt. Die entwickelten Konzepte wurden als Baukastensystem für die Modellierung multimodaler Interaktion in der VR realisiert.

Die Gestenerkennung erfolgt mittels in die Szenenstruktur eingebetteter Detektornetze. Dabei bilden sog. Aktuatoren eine Benutzer- und Sensorikabstraktionsebene. Spezielle Szenengraphknoten führen eine Voranalyse der veränderlichen Szene im Hinblick auf eine zeitversetzte Analyse gestischer Eingaben durch. Andere Knotentypen etablieren eine Kommunikation mit weniger zeitkritischen Komponenten zur Wissensunterstützung. Sog. Attributsequenzen erlauben einfache Datenflussverschaltungen ähnlich den Feldkonzepten bei VRML, arbeiten aber anders als diese asynchron zur Renderpipeline.

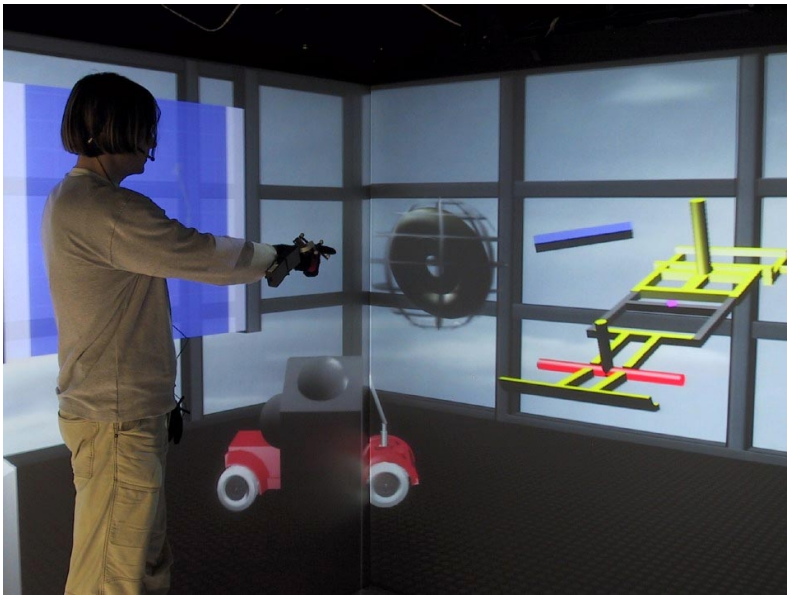

Abb.1: Gestik- und Spracheingaben: Konstruktion einer Citymobil-Variante in der virtuellen Werkstatt

Die Erkennung von Gesten basiert auf der Detektion definitorischer Merkmale, die sowohl die Form als auch den zeitlichen Verlauf einer Geste betreffen. Als Formmerkmale werden Fingerstellung, Handorientierung und -position betrachtet. Expressive Elemente, die auf das Vorliegen einer bedeutungstragenden Geste hinweisen, sind Ruhepunkte, hohe Beschleunigungen, Symmetrien und Abweichungen von Ruhestellungen bei Handspannung und der Handposition. Realisiert wurden u.a. Erkenner für die universellen Basisinteraktionen (Zeigen, Greifen, Loslassen, Rotation, Translation).

Bei der sprachlich-gestischen Interaktion werden drei Typen kommunikativer Gesten ausgewertet: Deiktische Gesten („nimm <Zeigegeste> dieses Teil“) spezifizieren ein Objekt oder einen Ort der virtuellen Umgebung, mimetische Gesten („drehe es <kreisender Zeigefinger> so herum“) qualifizieren die Ausführung einer Aktion, und ikonische Gesten („das so <Andeutung eines Zylinders durch die Handform> geformte Objekt ...“) werden zur Objektreferenz verwendet. Die Integration multimodaler Eingaben beruht auf temporalen ATNs [2].

Mit den skizzierten Methoden kann z.B. ein Citymobil aus seinen Hauptkomponenten in verschiedenen Varianten in VR zusammengesetzt werden. Über reine Montageprüfungen hinausgehend, zielen die Forschungsarbeiten in der virtuellen Werkstatt auf die umfassende wissensbasierte Unterstützung von Konstruktionsaufgaben durch eine Wissensrepräsentationsschicht [3]. Dazu wurden Techniken entwickelt, die eine interaktive Skalierung einzelner Bauteile sowie ganzer Baugruppen unter Aufrechterhaltung ihrer semantisch repräsentierten Verbindungseigenschaften bei sofortigem Feedback ermöglichen. Weitere Arbeiten betreffen u.a. funktionale Überprüfungen (z.B. Freigangprüfungen) von interaktiv modellierten Variantenkonstruktionen sowie die Netzwerkverteilung der virtuellen Umgebung zur Erschließung von Anwendungen im Concurrent Engineering. Einzelheiten sind der angegebenen Literatur zu entnehmen, wo auch die Einbettung in den allgemeinen Stand der Forschung beschrieben ist.

Kommunikation mit einem virtuellen Assistenten

Die natürliche Interaktion zwischen Menschen profitiert von der engen Kopplung mehrerer Modalitäten – wie Sprechen, Zeigen, Blickrichtung, Gesichtsausdruck etc. –, die man simultan äußern und umgekehrt ohne Mühe verstehen kann. Mit dem künstlichen Agenten „Max“, den wir in mehrjähriger Arbeit im DFG-Sonderforschungsbereich 360 „Situierte Künstliche Kommunikatoren“ an der Universität Bielefeld entwickelt haben, untersuchen wir multimodale Interaktion sowohl Eingabe- als auch Ausgabe-seitig mit einem anthropomorphen Assistenten in virtueller Realität [4]. Max verfügt über ein menschliches Aussehen und wird in der VR-Umgebung in Lebensgröße projiziert. Mit synthetischer Stimme und einem computeranimierten Körper kann Max sprechen, gestikulieren und Gesichtsausdrücke zeigen. Über Mikrophon und ein Tracking-System kann Max sein Gegenüber auch „hören“ und „sehen“ und Sprache, Gestik und Blickrichtung des Menschen als Eingaben verarbeiten.

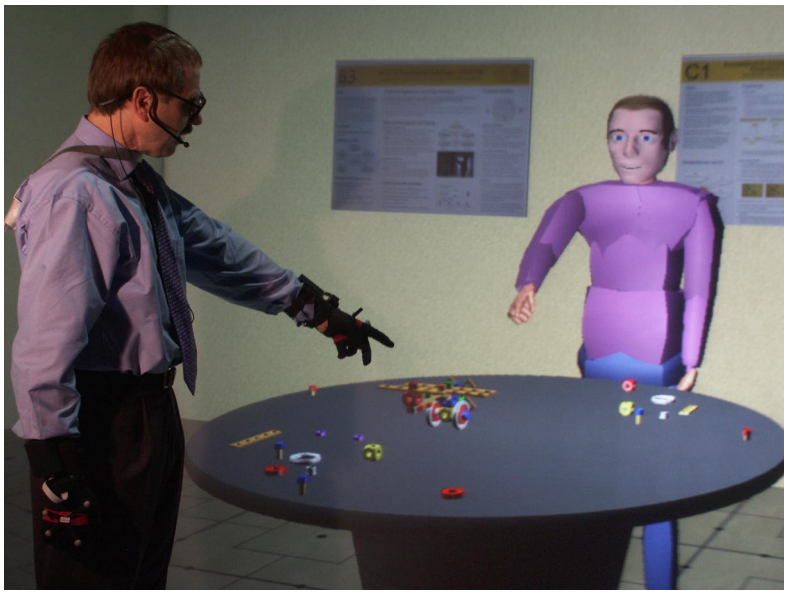

Abb.2: Multimodaler Dialog: Mensch und Max kooperieren beim Zusammenbau eines Flugzeugmodells

In unserem Forschungsszenario geht es um das Bauen von Objekten, zum Beispiel eines Flugzeugmodells, aus einem *Baufix*-Konstruktionsbaukasten. Hieran wird erprobt, ob Max sich in wechselnden Situationen soweit „verständig“ erweist, dass er im Dialog mit einem Menschen standhält.[1] Mensch und Max stehen sich dabei an einem Tisch gegenüber wie in Abb. 2 gezeigt; mit Ausnahme des Menschen ist die dargestellte Szene eine projizierte virtuelle Realität. Auf dem Tisch liegen verschiedene Bauteile, die im Verlauf des Dialogs zusammengesetzt werden. Sowohl Mensch als auch Max können durch natürlichsprachliche Instruktionen und Gesten den Zusammenbau einzelner Teile veranlassen, der in physikgerechter Simulation, unterlegt durch realistische Geräusche, ausgeführt wird. Die sprachlichen Äußerungen von Max werden, unter Anpassung von Parametern an die aktuelle Situation und inklusive der Generierung passender Gesten, aus einem Repertoire stereotyper Aussageformen erzeugt [5]. Mit simulierten Gesichtsmuskeln kann Max dabei auch „emotionale Zustände“ zum Ausdruck bringen, die unter anderem von dem Erreichen oder Misslingen kommunikativer Ziele beeinflusst werden.

Neben dem Aspekt der technischen Machbarkeit sind unsere Forschungsarbeiten auch mit der Erwartung verbunden, durch die Entwicklung und den Test operationaler Modelle detaillierte Erkenntnisse über menschliche Kommunikation zu gewinnen. Wie funktioniert beispielsweise das zeitliche Zusammenspiel von Sprechen und Zeigen? Wie wird das Abwechseln im Dialog gesteuert? So erfolgt die Entwicklung neuartiger Formen der Mensch-Maschine-Interaktion in enger Verbindung von erkenntnisgeleiteter Forschung und Anwendungserprobung.

[1] Weil unser Agent sich einerseits multimodal (mit Sprache, Gestik und auch Gesichtsmimik) äußern kann und er sich andererseits mit der Assemblierung virtueller Objekte auskennt, wurde er auf MAX – für „Multimodaler AssemblierungseXperte“ – getauft.

Ein virtueller Museumsführer

Als erste Alltagsanwendung fungiert Max im Heinz Nixdorf MuseumsForum (HNF) in Paderborn als virtueller Museumsführer, wo er seit Anfang 2004 als fester Bestandteil der Dauerausstellung „KI und Robotik“ durchgehend läuft. In dieser Rolle gibt er beispielsweise Auskünfte über die Ausstellung und das Leben und Wirken von Heinz Nixdorf.

In der Museumsinstallation wird Max über einen Beamer in Lebensgröße auf eine Leinwand projiziert (siehe Abb. 3). Er kann seine direkte Umgebung mittels einer Videokamera visuell wahrnehmen und mehrere Gesprächspartner sowie Umgebungsobjekte gezielt anschauen. Er kann sowohl in den virtuellen Raum (z.B. auf sich) zeigen als auch auf Exponate im Realraum, deren Koordinaten ihm bekannt sind, und simuliertes emotionales Verhalten an den Tag legen [6].

Abb.3: Eine Alltagsanwendung: Max im Dialog mit Besuchern des Heinz Nixdorf MuseumsForums

Technische Grundlage des gesamten Systems ist ein Doppelprozessor-PC. Über eine Tastatur nimmt Max beliebige Texteingaben entgegen und ist vermöge seines Konversations- und Weltwissens und eines Dialogsystems in der Lage, sinnvolle Antworten zu generieren und zusammenhängende Gespräche zu führen. Auf diese Weise kann er „Smalltalk“-Dialoge, z.B. über das Wetter (mit Zugriff auf den Wetterbericht im Internet), Kinofilme oder Fussball führen, aber auch auf natürliche Weise Erklärungen über verschiedene vordefinierte Inhalte geben. Dabei baut Max ein Modell des Gegenübers auf und passt sich den Wünschen und Interessen des Gesprächspartners an. Durch Gesichtsausdrücke und Stimmlage kann er darüber hinaus auch Emotionen wie Freude oder Verärgerung ausdrücken, so dass ein sehr „menschlicher“ Gesamteindruck entsteht. In der Museumsumgebung konnten wir erste (positive) Erkenntnisse über die Akzeptanz gewinnen, die Menschen einem künstlichen Ansprechpartner entgegen bringen [7].

Auf Unterstützung unserer Forschungsarbeiten durch die Deutsche Forschungsgemeinschaft (DFG) und das Heinz Nixdorf MuseumsForum (HNF) wird hingewiesen.

Literatur

[1] Jung, B., Latoschik, M., Biermann, P. & Wachsmuth, I. (2002). Virtuelle Werkstatt. In J. Gausemeier & M. Grafe (Hrsg.), *1. Paderborner Workshop Augmented & Virtual Reality in der Produktentstehung* (S. 185-196). Paderborn: HNI.

[2] Latoschik, M.E. (2002). Designing transition networks for multimodal VR-interactions using a markup language. *Proceedings IEEE 4th International Conf. on Multimodal Interfaces (ICMI 2002)*, Pittsburgh, USA, 411-416.

[3] Latoschik, M. E., Biermann, P. & Wachsmuth, I. (2005). Knowledge in the loop: Semantics representation for multimodal simulative environments. *Proceedings 5th International Symposium on Smart Graphics* (S. 25-39). Berlin: Springer (LNCS 3638).

[4] Kopp, S., Jung, B., Lessmann, N. & Wachsmuth, I. (2003). Max – a multimodal assistant in virtual reality construction. *KI – Künstliche Intelligenz 4/03*, 11-17.

[5] Kopp, S. & Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Journal of Computer Animation and Virtual Worlds, 15* (1), 39-52.

[6] Becker, C., Kopp, S. & Wachsmuth, I. (2004). Simulating the emotion dynamics of a multimodal conversational agent. In E. André, L. Dybkjaer, W. Minker & P. Heisterkamp (Hrsg.), *Affective Dialogue Systems* (S. 154-165). Berlin: Springer (LNAI 3068).

[7] Kopp, S., Gesellensetter, L., Krämer, N.C. & Wachsmuth, I. (2005). A conversational agent as museum guide – Design and evaluation of a real-world application. *Proc. Intelligent Virtual Agents* (IVA 2005; Kos, Greece, Sep. 12-14, 2005), im Druck.