Erschienen in:
Tagungsband zum Workshop des Forschungsverbundes NRW – Die virtuelle Wissensfabrik
[22./23. Sept. 1999, Schloß Birlinghoven, St. Augustin: GMD]

Final

Sprachgestützte gestische Interaktion zur Steuerung Virtueller Konstruktion

Marc Erich Latoschik, Ipke Wachsmuth

AG Wissensbasierte Systeme
Technische Fakultät, Universität Bielefeld
Postfach 100131, D-33501 Bielefeld, Deutschland
e-mail: {marcl,ipke}@TechFak.Uni-Bielefeld.DE

Zusammenfassung In diesem Beitrag stellen wir unsere Ergebnisse zur Erforschung sprachlich/gestischer Eingaben zur Steuerung und Manipulation von Virtuellen Umgebungen vor. Konkret werden die Resultate an einem System für die multimodale Interaktion in der Virtuellen Konstruktion erläutert. Nach einem Überblick über bisherige Arbeiten auf dem Gebiet der multimodalen Steuerung, wird die benötigte Interaktionsfunktionalität anhand der realen Anwendung beschrieben. Darauf aufbauend folgt neben einer Beschreibung möglicher sprachlicher Handlungsanweisungen eine Klassifikation verschiedener Gestentypen, um mögliche Kandidaten für die Umsetzung der unterschiedlichen Manipulationsaufgaben in dieser Domäne zu identifizieren, und dieses an Interaktionsbeispielen zu erläutern. Als Resultat werden sprachgestützte deiktische und mimetische Gesten des Benutzers betrachtet. Erstere dienen der Referenzanalyse, letztere machen gewünschte Veränderungen „vor". Diese Manipulationen werden durch sprachliche oder gestische Trigger eingeleitet und bewirken eine Anpassung in den Funktionsmodi der Auswertung, wobei im Abschnitt der technischen Realisierung zwischen diskreten und kontinuierlichen Interaktionen unterschieden wird. Für die Umsetzung einer kontinuierlichen Modifikationen der virtuellen Szene werden neben dem Konzept der Manipulatoren sogenannte Aktuatoren als Repräsentanten für Benutzermodalitäten eingeführt. Diese koppeln während der Interaktion an sogenannte Motion-Modifikatoren um die unscharfen Sensor-Eingaben zu korrigieren.

1 Multimodale Interfaces für Virtuelle Umgebungen

Virtuelle Umgebungen bieten ein Medium, welches neuartige Formen der Mensch-Maschine Interaktion erfordert. Entgegen der Verbesserung der 3D-Ausgabesysteme in diesem Bereich, wurden viele Eingabemethoden den bekannten 2D-Eingabemetaphern konventioneller Computer-Arbeitsplätze entlehnt. In Systemen in welchen die Grenzen zwischen virtueller und realer Umgebung immer undeutlicher werden, erscheint eine „natürlichere" Benutzer-Interakation wünschenswert. Das übergeordnete Thema der hier dargestellten Arbeiten sind Sprach- und Gesten-Interfaces für Virtual Reality und Multimedia-Anwendungen. Das Ziel ist es, Techniken zu entwickeln, die dem Benutzer

den Einsatz grober, auf Körper-, Arm-, Hand- und Fingerstellung basierender gestischer Kommunikation ermöglichen. Damit sollen die Begrenzungen von üblichen Bildschirm-Displays überwunden werden und durch sprachlich-gestische Interaktionstechniken für den Einsatz mit Groß-Displays (Wandprojektionen, Workbenches, Caves) ersetzt werden. Als Resultat soll ein freistehendes, komfortables Agieren, und dadurch eine möglichst natürliche Form der Mensch-Maschine-Kommunikation (MMK) ermöglicht werden. Die Benutzung auf bildschirmorientierte Arbeitsplätze bezogener Eingabegeräte und Interaktionsmetaphern ist mit diesen neuen Ausgabegeräten nicht mehr adäquat. Versuche, die Eingabemetaphern dieser bisherigen WIMP (Windows, Icons, Mouse, Pointer) Interfaces in die dritte Dimension zu transportieren, führten zu der Entwicklung diverser Pointingdevices wie dem Stylus oder der 3D-Space-Mouse. Gerade eine herausragende Qualität VR-gestützter Anwendungen macht jedoch alternative Interaktionsmöglichkeiten wünschenswert: Durch die Art der Simulation steht nicht mehr der Computer, bzw. Desktop-orientierte Metaphern als Werkzeug, im Zentrum der Interaktion, sondern die Anwendung selbst. Ziel ist damit der Verzicht auf eine vermittelnde Schicht zwischen Benutzer und Benutztem. Tastatur und Maus weichen den natürlichen Modalitäten Gestik und Sprache.

Erste Bestrebungen zur Verwendung der Modalitäten Sprache und Gestik in der Mensch-Maschine-Kommunikation reichen bis in die 80er Jahre zurück. Das Put-That-There System [3] war ein früher Versuch, Gestik und Sprache als Eingabemodalitäten auszuwerten. Als „Gestenerkennung" wurde hier die Zeigerichtung einer Extremität (eines Armes) auf eine zweidimensionale Projektionsfläche mit statischen Objekten ausgewertet; unberücksichtigt blieben zusätzliche Informationen über Körper-, Kopf, Hand- und Fingerstellung. Der Zeigevektor wurde mit den Ergebnissen eines Wort-basierten Spracherkenners integriert; dabei wurden Plätze verbal unterspezifizierter Referenzen („...this...", „...there...") durch die Auswertung der Position eines ständig präsenten, per Armstellung gesteuerten Cursors ausgefüllt. Die Umsetzung der Benutzerinstruktionen nach der Eingabeanalyse erfolgte ausschließlich als diskrete Zustandsänderungen. Viele der in den 90er Jahren entstandenen Arbeiten konzentrieren sich ganz speziell auf die multimodale Integration. Bei gleichzeitiger graphischer Repräsentation von Objekten steht hier vor allem die Benutzerdeixis, also gestisches Zeigen auf Objekte, deren verbale Benennung oder Blickrichtung im Interesse [6][9][14]. Andere Arbeiten konzentrieren sich zwar konkret auf den Einsatzzweck in VR-Umgebungen [1][2], betrachten aber nur eingeschränkte Gestentypen, zum Beispiel symbolische Gesten, und bilden diese auf Systemkommandos ab (vgl. Übersicht in [9]). Diese Einschränkung wird auch in [21] kritisch bemerkt, wobei der Aspekt der Multimodalität jedoch nicht weiter verfolgt wird. Ein Ansatz, ikonische (formbeschreibende) Gesten zu berücksichtigen, findet sich in [20]. Hier dient eine Hand dazu, Kurven im Raum zu beschreiben und zu verändern. Einige der genannten Arbeiten befassen sich zwar als Randerscheinung mit Einzelaspekten der Gesten-Dynamik während der Erkennung, le-

gen aber keine generellen Lösungsansätze für die umfassende Bearbeitung dieser fundamentalen Eigenschaft gestischer Artikulation vor.

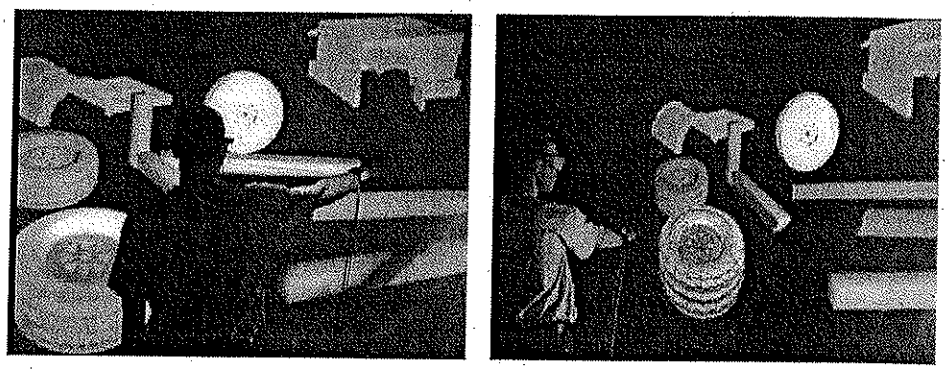

Abbildung 1: Seiten- und Rückansicht während des virtuellen Konstruierens an einer interaktiven Wand.

Die im folgenden dargestellten Arbeiten erweitern bisherige sprachlich-gestische Interfaces mit dem Ziel einer möglichst natürlichen Mensch-Maschine-Kommunikation. Dazu werden neben deiktischen auch mimetische („vormachende") Gesten zugelassen; neben diskreten Interaktionen sind auch kontinuierlich ausgewertete Manipulationen möglich; die Interaktionssemantik wird dabei durch den jeweiligen sprachlichen Kontext moduliert. Dynamikaspekte der Gestenerkennung fließen dabei an vielen Stellen sowohl auf der Erkennerseite, als auch bei der Umsetzung der durchgeführten Manipulationen in das Konzept ein. Die vorgestellten Arbeiten sind eingebettet in das SGIM-Projekt (Speech and Gesture Interfaces for MultiMedia), einem Teilprojekt des Multimedia-NRW Verbundprojekts „Virtuelle Wissensfabrik"[1]. Einige Grundlagen für die technische Realisierung der verschiedenen Komponenten, so zum Beispiel die der Spracherkennung [5], sind die Ergebnisse der anderen Teilprojekte innerhalb des Verbundes. Auf diesen integrativen Aspekt wird in Abschnitt 6 weiter eingegangen. Im folgenden Beitrag wird die Konzeption des SGIM-Systems für die multimodale Steuerung in der Domäne der Virtuellen Konstruktion erläutert.

2 Funktionale Beschreibung der Manipulationsaufgaben

Als Anwendungsszenario für die dargestellten Forschungsarbeiten zur sprachlich-gestischen MMK dient die Steuerung eines Systems zur interaktiven Montagesimulation in virtuellen Umgebungen, des „Virtuellen Konstrukteurs" [8]. Im diesem werden CAD-basierte Grundbausteine dreidimensional auf einer virtuellen Montagefläche präsentiert; die Aufgabe des Systems liegt in der wissensbasierten Unterstützung des Benutzers beim Zusammenbau dieser Bauteile. Im Virtuellen Konstrukteur bisher verfügbare

[1] Die Forschungsarbeiten in der Virtuellen Wissensfabrik werden unterstützt vom MSWWF des Landes Nordrhein-Westfalen unter OZ IV A3 -107 032 96

Interaktionstechniken, sprachliche Instruktionen und (Maus-basierte) direkte Manipulation, beziehen sich auf konventionelle Bildschirm-orientierte Arbeitsplätze. Die hier beschriebenen Arbeiten zielen auf eine sprachlich-gestische Steuerung an Großbild-Displays. Neben der Benutzer-Navigation stellt die Manipulation von Objekten eine zentrale Klasse von Interaktionsaufgaben in virtuellen Umgebungen dar. Im allgemeinen betreffen Manipulationen die folgenden Veränderungen visueller Objektattribute:

- Positionsänderung (Translation)
- Ausrichtungsänderung (Rotation)
- Größenänderung (Skalierung)
- Formänderung (Deformation)
- Erscheinung (Färbung, Texturierung)

Im Zusammenhang der Virtuellen Konstruktion sind insbesondere die ersten drei Manipulationsaufgaben von Interesse. Sie gehören zu den Standardoperationen in CAD- und anderen graphischen Modellierungs-Systemen. Im Virtuellen Konstrukteur, der speziell die interaktive Montagesimulation in virtuellen Umgebungen unterstützt, werden zusätzlich folgende Manipulationsaufgaben betrachtet:

- Verbinden von Bauteilen
- Trennen von Bauteilen
- Modifikation von Aggregaten durch Relativbewegung von Komponenten entlang zulässiger Freiheitsgrade von Objektverbindungen

Die Umsetzung dieser Manipulationen basiert im Virtuellen Konstrukteur auf einer wissensbasierten Modellierung der Verbindungsstellen („Ports") und Verbindungsarten zwischen diesen Ports. Abbildung 2 zeigt Beispiele der bisher modellierten Taxonomie von Port-Typen. Eine weitere Taxonomie klassifiziert mögliche Verbindungsarten bezüglich der zulässigen Freiheitsgrade bei eingegangenen Verbindungen [10]; z.B. sind bei Steckverbindungen Translation und Rotation ungekoppelt, während sie bei Schraubverbindungen gekoppelt sind, wodurch Hinein- und Hinausbewegung von Schrauben mit entsprechender Drehung erfolgt.

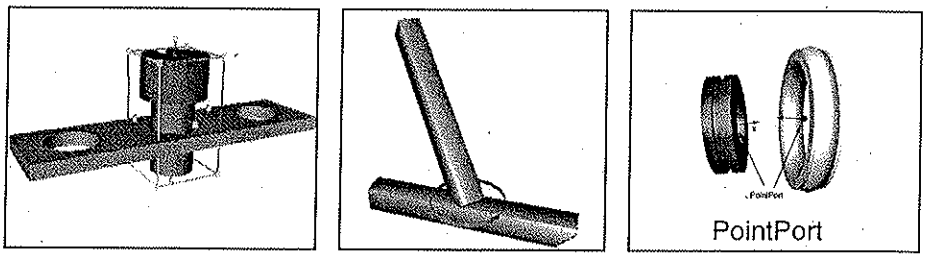

Abbildung 2: Typen von Verbindungsports beim Virtuellen Konstruieren: Extrusion ports (links), plane ports (mitte) und point ports (rechts). Bei Verbindungen zwischen den jeweiligen Ports besteht jeweils ein Freiheitsgrad bzgl. Rotation sowie bis zu zwei Freiheitsgrade bzgl. Translation (Abbildungen nach [8]).

Bei den Manipulationsaufgaben der Virtuellen Konstruktion, wie dem Verbinden oder Trennen von Bauteilen, bestehen somit im Vergleich zu den allgemeinen Manipulati-

onsaufgaben – wie Translation und Rotation von Bauteilen – zusätzliche Randbedingungen die im Virtuellen Konstrukteur explizit modelliert und bei der Auswertung von Benutzerinteraktionen zugänglich sind. So kann z.B. das Verbinden von Bauteilen als Spezialfall der Transformation (Translation und Rotation) eines Bauteils betrachtet werden, bei welcher die Zielposition des transformierten Objekts durch die Verbindungsports beider Bauteile eingeschränkt ist. Auf ähnliche Weise kann die Rotation von Teilaggregaten als Spezialfall der Rotation freier, d.h. unverbundener Bauteile betrachtet werden, wobei jedoch die Rotationsachse durch den Typ der Verbindung festgelegt ist. Diese Interaktionsaufgaben sind zunächst Eingabe-unabhängig. Sie können jeweils durch verschiedene Modalitäten wie Sprache oder Gestik, oder mit Hilfe spezieller Eingabegeräte, beispielsweise Maus-basierter Manipulationen, erfolgen. Bei Verarbeitung gestischer Benutzerinteraktionen werden die geschilderten Konstruktionsrandbedingungen ausgenutzt in dem Ungenauigkeiten bei der Gestenerkennung durch systemseitiges Wissen über den Anwendungsbereich ausgeglichen werden, bzw. ungenau erfolgende gestische Benutzerinteraktionen wissensgestützt justiert werden.

3 Sprachgestützte gestische Interaktion

Nachdem wir im Kontext der Anwendung die Manipulationsmöglichkeiten funktional beschrieben haben, werden nun Sprache und Gestik als Eingabemedium evaluiert. Die sprachliche Ausdrucksfähigkeit gehört dabei wohl zu unseren mächtigsten Fähigkeiten; eine vollständige Interpretation von verbalen Handlungsanweisungen steht hier aber nicht im Mittelpunkt unseres Interesses. Vielmehr sind es die multimodalen Benutzerexpressionen welche gerade im Kontext von Aufgaben mit räumlichen Bezugssystemen und Relationen die Gestik als Ausdrucksform stärker in den Mittelpunkt stellen. Die einzelnen Aspekte der Sprache sowie der Gestik erfüllen in dem gewählten Kommunikationskontext ganz spezifische Rollen. Je nach Art des zu kommunizierenden Faktums in einer Handlungsanweisung ist es einmal die Sprache, ein anderes Mal die Gestik, welche eine adäquatere Vermittlung der Information ermöglicht.

3.1 Multimodalität in Handlungsanweisungen

An Beispielen soll der grundsätzliche Zusammenhang zwischen Art der zu kommunizierenden Information und der benutzten Modalitäten verdeutlicht werden. In der Anweisung „*...nimm* <Zeigegeste> *dieses Teil...*“ vermittelt das Verb adäquat die Art der gewünschten Manipulation, die Zeigegeste aber ist ein schnelles und zuverlässiges Mittel in einer virtuellen Szene mit räumlich angeordneten Objekten das Ziel-Objekt der Handlung zu bestimmen. Dieses kann natürlich ebenfalls rein sprachlich oder gemischt sprachlich/gestisch erfolgen, doch gerade in dem gewählten Kontext der Handlungsinstruktionen dient die Zeigegeste häufig der konkreten „Bedeutung“. Neben diesen Referenzäußerungen wird die Gestik auch zur Beschreibung von Zustandsveränderungen benutzt: „*Dreh das Rad* <Drehbewegung der Hand> *so herum*“ ist ohne eine Interpretation der Geste unterspezifiziert. Weiterhin hilft die zeitliche Koppelung [15] zwischen beiden Modalitäten Gestik und Sprache bei der Analyse der Äußerungen. So kann beispielsweise der konkrete Zeitpunkt der Dereferenzierung der Gestik bestimmt werden. Die Bestimmung des Indexpunktes (dem durch die Gestik bedeuteten Objekt) kann als „Gemeintes“ (das zu manipulierende Objekt) interpretiert werden, wenn die Gestik in

temporalem Zusammenhang mit verbalen Referenzen geäußert wird. Die Zeigegeste liegt immer kurz vor oder auf dem Beginn der verbalen Referenzierung, im ersten Beispiel also dem Ausdruck *„dieses Teil"*. Dieser Vorgriff deutet schon auf eine Unterscheidung verschiedener Gestentypen und ihrer Charakteristika hin. Im folgenden Abschnitt wollen wir weitere mögliche Gestenarten vorstellen und anschließend jene identifizieren, welche als nützlich in der Virtuellen Konstruktion erscheinen.

3.2 Gestentypen

Um zu einer Übersicht weiterer möglicher gestischer Interaktionsformen zu gelangen, ist in der folgenden Tabelle (n. [11]) eine zusammenfassende Gesten-Klassifikation nach verschiedenen bestehenden Schemata erfolgt. Diese versucht übersichtsartig die verschiedenen in der Literatur (s. [4][16][21]) gebräuchlichen Typen gemäß ihrer kommunikativen Funktion zusammenzufassen.

TABELLE 3. Gestentypen

Typ	Klassifikationsname	Charakteristika
I	1. Ikonisch 2. Mimetisch 3. Objektbezogen	Benutzen die Extremitäten als Platzhalter um das Verhalten eines beschriebenen Objektes oder Zustands nachzubilden. Bsp.: Eine sich öffnende Hand demonstriert das Öffnen einer Blüte.
II	4. Physiographisch 5. Kinetographisch 6. Pantomimisch	Repräsentieren und verbildlichen das Zusammenspiel mit einem Objekt. Zeigen die Interaktion bei der Benutzung. Bsp.: Jemand demonstriert die Benutzung von Hammer und Nagel ohne Hilfsmittel.
III	1. Symbolisch 2. Moduseinstellend 3. Emblematisch	Haben eine eindeutige Semantik als alleinstehende Geste und verändern ggf. den Modus in welchem eine gleichzeitig verbale Äußerung interpretiert wird. Bsp.: - Zeigefinger und Daumen bilden einen Kreis und symbolisieren OK. - Die Handfläche zeigt parallel zum Boden und wackelt um die Längsachse als Zeichen das eine parallele Aussage unsicher ist.
IV	1. Ideographisch 2. Metaphorisch 3. Ikonisch	Veranschaulichen eine räumliche metaphorische Manifestation eines internen Zustands. Beziehen sich auf eine Interpretation. Bsp.: Jemand sagt ihr/ihm ist schwindelig und dreht dabei den Zeigefinger in der Luft.

TABELLE 3. Gestentypen

Typ	Klassifikationsname	Charakteristika
V	1. Beats 2. Gestikulation 3. Sprachmarkierend 4. Selbstregulierend	Geben einen Sprachrythmus an. Betonungen und gestische Expression fallen in den gleichen Takt. Bsp.: Redner(in) unterstreicht die entscheidenden Punkte ihrer Aussage mit einem Klopfen des Zeigefingers auf den Tisch.
VI	1. Deiktisch	Referenzieren auf Objekt(e), Ort(e) und Richtung(en) im Raum. Bsp.: Jemand zeigt auf einen Stuhl und sagt: „Hol bitte diesen Stuhl...", zeigt dann auf eine freie Stelle neben den Tisch und sagt: „...und stell ihn da hin!".

Betrachten wir die Charakteristika der unterschiedlichen Gestentypen und vergleichen wir diese mit den gewünschten Manipulationsmöglichkeiten, so erscheinen zwei Klassen als besonders nützlich im Kontext der Virtuellen Konstruktion. Wie wir bereits gesehen haben sind deiktische Gesten (Typ VI) besonders nützlich, um Objekte und Orte zu bezeigen. In vielen Fällen ist eine Zeigegeste ein schnelles und eindeutiges Mittel die Referenz (das Objekt oder den Ort) der Handlungsanweisung zu bestimmen [11]. Dieses geschieht im Zusammenspiel mit der Sprache, welche weitere nützliche Hinweise für eine erfolgreiche Dereferenzierung geben kann, beispielsweise die Benennung einer Objekteigenschaft, einer Farbe bei zwei ansonsten gleichen, nah angeordneten Objekten. Zu den deiktischen Gesten kann man auch die relative Kopfstellung und Blickrichtung des Benutzers einordnen. Dieses erweist sich als aufschlußreich, um die relative Betrachterausrichtung zu bestimmen und ggf. eine unerwünschte Systemreaktion, etwa bei einem von der Szene abgewendeten Benutzer, zu vermeiden. Weiterhin gibt die Kopfrichtung als grobe Approximation der Blickrichtung Hinweise über den betrachteten Szenenbereich und die zu einem gegebenen Zeitpunkt zu berücksichtigenden Objekte.

Mit Gesten des Typ I, mimetisch/ikonischen Gesten, können gewünschte Veränderung von mindestens vier der in Abschnitt 2 aufgeführten Objekt-Attribute direkt simuliert werden. Soll beispielsweise im virtuellen Raum ein Objekt eine Lageveränderung durchführen, so wird die relative Verschiebung mit den Extremitäten beschrieben, also mit den Händen vorgemacht. Ebenso wird die Rotation eines Objektes durch eine angedeutete Rotation einer Hand in Drehrichtung und -umfang vorgeführt. Offensichtlich sind weiterhin auch Form- und Größenänderungen durch Beschreibungen mit dieser Art Gestik möglich.

Weniger im Interesse unserer Forschung sind die symbolischen Gesten des Typus III. Diese stellen zwar auf Erkennerseite die geringeren Anforderungen, sind aber prinzipiell als gering intuitiv zu betrachten, da sie vom sozialen Kontext geprägt (Daumen hoch als OK-Zeichen), oder schlicht erlernt sind (z. Bsp. Zeichensprache bei Tauchern).

3.3 Gesten im sprachlichen Kontext

In Abschnitt 3.1 wurden bereits mögliche sprachlich/gestische Interaktionsbeispiele beschrieben. Zwar lassen sich einige Aktionen rein gestisch kommunizieren, durch ein konkretes Zeigen und ein Zugreifen etwa, aber man gelangt offensichtlich schnell in Situationen in denen die Gestik als rein unimodale Äußerung zu ausdrucksschwach ist. So lassen sich die in Abschnitt 2 beschriebenen Manipulationen nur schwerlich in Gänze, ohne Zurhilfenahme symbolischer, zu erlernender Gesten vom Typus III, gestisch beschreiben.
Weiterhin haben wir auf die temporalen Zusammenhänge zwischen Sprache und Gestik hingewiesen. Eine Schwierigkeit bei der Gestenerkennung ist die Erkennung der Klimax, dem Zeitpunkt der maximalen Expression einer Geste. Auch haben die im Rahmen einer Diplomarbeit durchgeführten Experimente bei vielen Probanden eine solche Klimax beim Zeigen nur undeutlich erkennen lassen. Der sprachliche Kontext hilft in solchen Fällen bei deren zeitlicher Zuordnung und Interpretation.
Neben den beschriebenen Eigenschaften der Sprache, die gestische Expression zu konkretisieren und zeitlich einzuordnen, dient sie in der Konstruktionsdomäne natürlich auch einem primären Zweck. Untersucht man Dialoge in einem Instrukteur/Konstrukteur-Szenario, so fällt ein großer Anteil der Instruktionsanweisungen nach dem folgendem Schema aus. Sprachlich werden

1. Handlungen und Manipulationen initiiert.
2. Manipulationen modifiziert/konkretisiert.
3. die zu den entsprechenden Handlungen gehörenden Referenzen beschrieben.

Auch hier soll dies an Beispielen verdeutlicht werden. Die einzelnen Bereiche der Manipulationsanweisungen lassen sich gemäß obigem Schema zuordnen (gekennzeichnet durch die Klammernnotation):

Führe(1) *dieses*(3) *Rohr*(3).
Ich möchte dieses(3) *Rad*(3) *führen*(1).
Verbinde(1) *das*(3) *rote*(3) *Rohr*(3) *mit*(1) *der*(3) *Lenkstange*(3).
Nimm(1) *dieses*(3) *rote*(3) *Rad*(3) *und dreh*(1) *es*(3) *so*(2*) *um*(2) *die*(3) *Achse*(3).
Dreh(1) *es*(3) *um*(1) *30*(2) *Grad*(2) *nach*(2) *rechts*(2).

Gewisse Schlüsselworte deuten dabei bestimmte modale Wechsel an. Das Wort „*so*" besetzt hier eine Sonderstellung da es deutlich auf eine nicht im verbalen Ausdruck befindliche Beschreibung deutet. Der Großteil der Information ist aber verläßlich direkt aus dem Sprachkanal zu ermitteln. Verbale Handlungsanweisungen sind eine primäre Methode der gewünschten Manipulation Ausdruck zu verleihen. Je nach Art der Interaktion fällt deren Interpretation und die darauf folgende Umsetzung in einen neuen Systemzustand auf prinzipiell unterschiedliche Arten aus. Die folgenden Abschnitte tragen diesem Umstand durch die Identifikation von zwei verschiedenen Interaktionsmodi besonders Rechnung.

3.4 Sprachlich/gestische Interpretation und Integration

Das Ziel einer sprachlich/gestischen Auswertung und Integration ist immer eine mögliche Manipulation der aktuellen Szene. Diese unterscheiden sich je nach Art durch die zur Ausführung benötigten Informationen. Die aus dem Sprachkanal emittierten Worte werden gemäß dem Schema in Abschnitt 3.3 klassifiziert [12]. Das Ergebnis der Klassifikation steuert die drauf folgende Interpretation. Worte des Typus 1 (Handlung/Manipulation) identifizieren einen Manipulationsframe. Worte des Typus 3 (Referenzen) dienen zusammen mit der Benutzerdeixis zur Bestimmung von Referenzen. Während des laufenden Systems werden kontinuierlich die relativen Abstände der einzelnen Objekte zu den oberen Benutzer-Extremitäten (Arme und Kopf) protokolliert. Diese Informationen geben ein Abbild der deiktischen Ausrichtung des Benutzers relativ zu den visualisierten Objekten und ordnen diese in sogenannten Proximitätslisten an. In diesen stehen die Objekte absteigend nach dem Abstand zu einer hypothetischen, quasi unendlichen Verlängerung der entsprechenden Extremität geordnet. Weiterhin werden die Arm-Proximitätslisten durch den „gestischen Zustand" angereichert. Unter diesem verstehen wir die zusätzlichen Gestenerkenner-Resultate, also Informationen ob der Anwender zu einem bestimmten Zeitpunkt eine typische Zeigegeste gemäß unserer Definition ausgeführt hat [11]. Werden nun für einen Handlungsframe Referenzen benötigt, und treffen Worte gemäß Typus 3 (Referenzen) ein, so kann in diesen Listen ein Abgleich mit den aus den Worten zu ermittelnden Informationen vollzogen werden. Die kontinuierliche Auswertung ermöglicht hier auch das „Zurückschauen", also die Berücksichtigung des zeitlichen Vorlaufs deiktischer Zeigegesten vor dem verbalen Kanal. Auf diese Weise können multimodale Referenzen analysiert und identifiziert werden.
Eine etwas andere Aufgabenstellung wird durch das schon erläuterte Schlüsselwort „*so*" eingeleitet. Dieses Wort eröffnet eine gestisch-mimetische Spezifizierung der Manipulation. Die Art der Manipulation wird nach wie vor über die Sprache mitgeteilt. Ihre Ausprägung aber wird gestisch beschrieben. In einem solchen Fall wird die Kontrolle über die laufende Manipulation an einen speziellen Modus in der Anwendung übertragen. In diesem wird die Analyse der beschreibenden Gestik ausgeführt. Der folgende Abschnitt wird die Unterscheidung in die zwei Verarbeitungsmodi erläutern.

4 Diskrete vs. kontinuierliche multimodale Benutzerinteraktion

Bisher wurden Manipulationsaufgaben in Virtual Reality Systemen im allgemeinen sowie bei der Virtuellen Konstruktion im speziellen betrachtet. Aufbauend darauf wurden gestisch/sprachliche Ausdrucksformen als Instrukteur in einem solchen Szenario untersucht und solche identifiziert und analysiert, welche prinzipiell zur Steuerung solcher Systeme geeignet erscheinen. Wie im Abschnitt über die Integrationsaufgabe dargelegt, stellen die Manipulationsaufgaben die Interaktionsziele für die in unserem System behandelten sprachlich-gestischen Benutzereingaben dar. Zur Durchführung einer Manipulationsaufgabe werden, je nach Art der Manipulation, unterschiedliche Informationen benötigt. Soll zum Beispiel ein Objekt rotiert werden, so müssen das Objekt, die Rotationsachse und der Rotationsumfang bestimmt werden. Die Art und Weise, wie diese Informationen kommuniziert werden, bzw. wie Benutzereingaben in Änderungen des Systemzustands umgesetzt werden, legt eine Unterscheidung in zwei

unterschiedliche Interaktionsmodi nahe: *diskrete* und *kontinuierliche* Interaktionen. Die Abbildungen in diesem Abschnitt zeigen Beispielinteraktionen mit dem SGIM-Demonstrator.

4.1 Diskrete Interaktion

Bei diskreten Interaktionen werden Änderungswünsche des Benutzers als instantane Zustandsänderungen der virtuellen Umgebung umgesetzt. Diskrete Interaktionen können unimodal geäußert werden, z.B. *„Drehe das gelbe Rad um 45 Grad nach hinten"*, oder multimodal, z.B. *„Stecke <Zeigegeste> dieses Rohr <Zeigegeste> da dran"*. Gestische Interaktionen sind dabei zumeist auf Zeigegesten beschränkt, welche Hinweise auf die auszuwählenden Objekte liefern ([11]; vgl. [3][9]). In multimodalen Konstruktionsdialogen sind die (diskreten und kontinuierlichen) Interaktionen des Benutzers oft unterspezifiziert, so daß eine Ergänzung der Eingaben um Kontextwissen über den Anwendungsbereich – bei der Virtuellen Konstruktion etwa Wissen über die Verbindungsmöglichkeiten der Bauteile – und Vorannahmen notwendig ist. Dies bedingt, daß bei der systemseitigen Interpretation von unterspezifizierten Benutzereingaben Systemzustände erzeugt werden können, die nicht den ursprünglichen Intentionen des Benutzers entsprechen. Bei diskreten Interaktionen, deren Auswirkungen sofort in der virtuellen Szene angezeigt werden, sind Korrekturen nur in folgenden Interaktionsschritten möglich. Der Benutzer hat jedoch keine Möglichkeit, die Interpretation einer Anweisung noch während deren Auswertung zu beeinflussen. Der Einsatz von VR-Techniken zielt jedoch oft gerade darauf, den Benutzer in die Szene zu integrieren und ihm unmittelbare Kontrolle der Manipulationen zu ermöglichen. In vielen Bildschirm-orientierten Anwendungen hat sich dafür der Einsatz direkter Manipulationen mittels Maus-Steuerung als nützlich erwiesen. Im folgenden werden dazu analog kontinuierliche Interaktionen im Kontext der natürlichen Modalitäten Gestik und Sprache betrachtet.

4.2 Kontinuierliche Interaktion

In der menschlichen Kommunikation kommen neben den schon oben behandelten deiktischen Gesten u.a. auch mimetische Gesten vor, die dem „Vormachen" einer beabsichtigten Änderung dienen (vgl. funktionale Klassifikation von Gestentypen in Tabelle 3). Dabei werden die Extremitäten (vor allem die Hände) als Platzhalter gebraucht, um dem Kommunikationspartner die Art und Weise der gewünschten Manipulation vorzumachen. Diese Interaktionsart bedingt eine andere Form der Umsetzung.

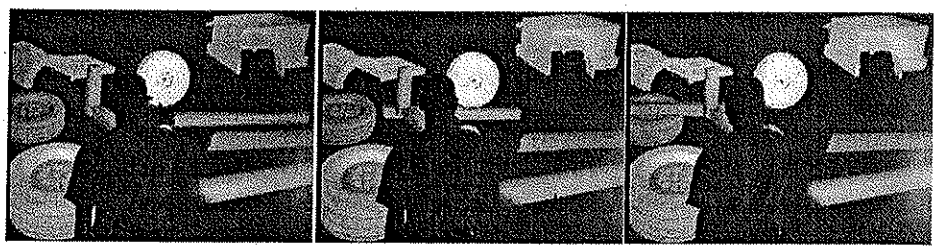

Abbildung 4: Auswahl und Drag & Drop I - Ein selektiertes Objekt wird durch den Benutzer mittels kontinuierlicher Interaktion von rechts nach links an einen neuen Platz geführt.

Auf virtuelle Umgebungen bezogen legen mimetische Gesten – im Gegensatz zu diskreten Interaktionen – eine über die Dauer des „Vormachens" folgende kontinuierliche Veränderung der virtuellen Umgebung nahe (Abb. 4, 5 u. 6).

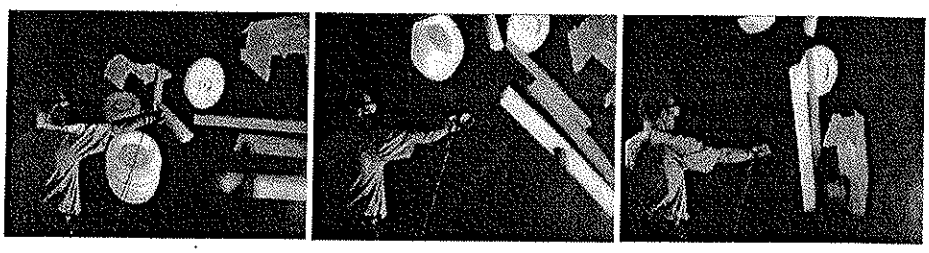

Abbildung 5: Auswahl und Drag & Drop II - Szenenmanipulation: Alle selektierten Objekte werden mittels kontinuierliche Interaktion rotiert.

In kontinuierlichen multimodalen Interaktionen ist mimetische Benutzer-Gestik oft begleitet durch spezifische Schlüsselworte in der sprachlichen Äußerung, z.B. *„so"* wie in *„Drehe das Rad <Beginn Rotation> so herum <Ende Rotation>"* (Abb. 6).

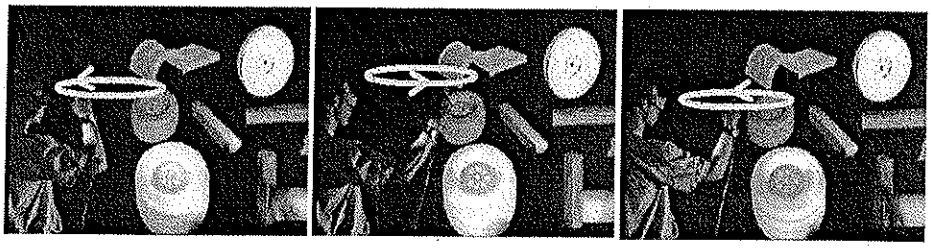

Abbildung 6: Rotationsbeschreibung: Kontinuierliche Interaktion über mimetische Gestik (Trajektorie der Hand wurde zur Veranschaulichung hinzugefügt).

Auswertung, Interpretation und Umsetzung kontinuierlicher Interaktionen erfolgen im SGIM-Demonstrator schritthaltend, wobei Benutzereingriffe zur unmittelbaren Korrektur möglich sind. Bei der Analyse dynamischer Gesten muß i.a. eine zeitbezogene Filterung körperbezogenen Daten erfolgen [12]. Die technische Realisierung diskreter und kontinuierlicher Interaktionen ist im folgenden Abschnitt beschrieben.

5 Systemmodellierung der Interaktionsformen

Die beiden Interaktionsmodi, diskret und kontinuierlich, erfordern konzeptionelle Unterschiede in ihrer technischen Realisierung. Ihre gemeinsame Auswertung in einem realen System zur Virtuellen Konstruktion bedingt ebenfalls zwei Ausführungsmodi. Bei beliebigen Eingaben, also gesprochenen und klassifizierten Worten oder einzelnen erkannten Gesten, wird während der multimodalen Integration versucht, benötigte Integrationsschemata zu füllen. Der Informationsfluß wird von einzelnen parallelen Erkennermodulen getrieben, welche ihre Resultate als singuläre Events an eine Integrationskomponente weiterleiten (s. Abb. 7).

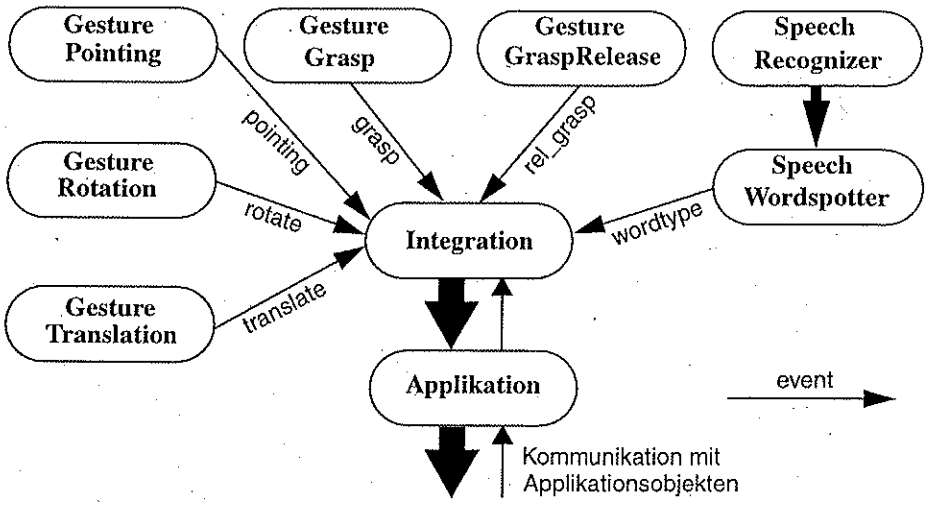

Abbildung 7: Event-getriebene Erkenner und Integrationsstruktur

Die einzelnen Ergebnisse der Erkennermodule werden in der Integration in eine gemeinsame Struktur gebracht. Signifikante Gesten, zum Beispiel ein Zeigen, oder spezielle Aktionswörter, wie „...drehe...", „...schiebe..." oder „...verbinde...", aktivieren jeweils einen speziellen Integrationsframe. Jeder dieser Frames hat spezifische Slots, welche durch die einkommenden Events gefüllt werden. Ein Objekt-Referenzframe benötigt diverse Spezifikationen. Verbal können dieses neben Benennungen auch die visuellen Objekt-Attribute Farbe, Lage oder Form sein. Gestisch wird das bedeutete Objekt durch die Richtung während einer Zeigegeste ermittelt. Ziel dieses Referenzframes ist es eine eindeutige Objekt-Instanz zu ermitteln. Dagegen benötigt ein Rotationsframe den Rotationsmittelpunkt, eine Rotationsachse und den Winkel der Änderung. Ist die Integration abgeschlossen, so wird eine mit dem Frametyp assoziierte Funktion ausgeführt. Ein Referenz-Frame aktiviert eine Selektion des referenzierten Objektes, ein Rotationsframe führt zu einer entsprechenden Objektlage oder -positionsänderung.

5.1 Umsetzung diskreter und kontinuierlicher Interaktionen

Der Frame-Abschluß kann durch drei verschiedene Ereignisse ausgelöst werden: Im einfachsten Fall ist der Frame vollständig spezifiziert und die assoziierte Aktion kann im diskreten Ausführungsmodus umgesetzt werden. Ein unterspezifizierter Frame kann durch zwei Arten von Ereignissen in verschiedene Ausführungsmodi gesetzt werden. Detektiert die Spracherkennung das Ende einer Äußerung, und liegt kein Ergebnis eines Gestenerkenners vor, so wird - unter Ergänzung der Eingabe durch Vorannahmen - die Benutzereingabe ebenfalls diskret umgesetzt. Wird ein Triggerwort („...*so*...") erkannt und eine entsprechende mimetische Geste ausgeführt, so wird in den kontinuierlichen Modus umgeschaltet und versucht, die unterspezifizierten Werte aus der Gestik zu ermitteln. Kann dieses nicht erfolgen, wird die Interaktion abgebrochen.

5.2 Umsetzung diskreter Interaktionen über Manipulatoren

Die Modellierung graphischer Szenen erfolgt i.a. durch die hierarchische Anordnung der Objekte als Knoten in einem Szenengraphen. Veränderungen werden durch sogenannte *Manipulatoren* ausgeführt. Diese können je nach Art bestimmte Attribute dieser Objekt-Knoten verändern. Für die technische Realisierung einer Selektion und Drehung benötigen wir mindestens zwei Manipulatoren, einen zur Suche des entsprechenden Knotens und anschließender Hervorhebung (Selektion und Highlighting), einen für die Manipulation der Knoten-Transformationsmatrix (Rotation). Abbildung 8 illustriert ein Beispiel für die diskrete Umsetzung einer Rotationsanweisung.

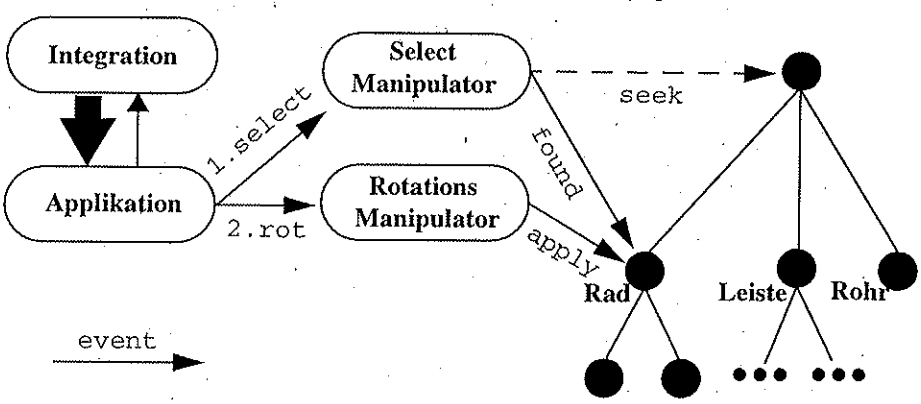

Abbildung 8: Einfache Manipulator-Szenengraphmodifikation bei instantaner Ausführung („Dreh das Rad")

5.3 Umsetzung kontinuierlicher Interaktionen über Aktuatoren und Motion-Modifikatoren

Auf die Funktion bestimmter Trigger als Einleitung einer mimetischen Beschreibung wurde bereits mehrfach hingewiesen. Auf der Anwendungsebene bewirkt ein Trigger einen Moduswechsel. Nach einem solchen Modustrigger wird die Interaktion nicht in einem kompletten Schritt mittels eines Manipulators umgesetzt; stattdessen wird die gewünschte Manipulation kontinuierlich aus den Bewegungsänderungen des Benutzers ermittelt. Für diesen Vorgang wird ein mehrstufiges Konzept benutzt.

Datenfluß zwischen den Komponenten

Die multimodale Integration arbeitet auf Event-Basis. Das bedeutet, daß die Kommunikation zwischen Integration, Applikation und den Manipulatoren auf dem Vorhandensein von Nachrichten als diskreten Signalen beruht. Im Gegensatz dazu arbeitet die Visualisierung der virtuellen Szene in einer Schleife, der sogenannten Rendering-Loop. Diese ist treibende Kraft und impliziter Taktgeber, um eine stetige Framerate (Anzahl der gerenderten Bilder/Zeiteinheit) zu gewährleisten. Kommunikation mit der Anwendung geschieht in den Zeiträumen zwischen den Berechnungen der einzelnen Bilder. Events übertragen nur den Wechsel zwischen verschiedenen Zuständen und treten vergleichsweise selten auf, die Rendering-Loop wird kaum belastet. Würde auch die Umsetzung einer kontinuierlichen Interaktion und die Auswertung der Benutzergestik vor der Integration erfolgen, so müßte jede erkannte Geste über die Integrationskomponente

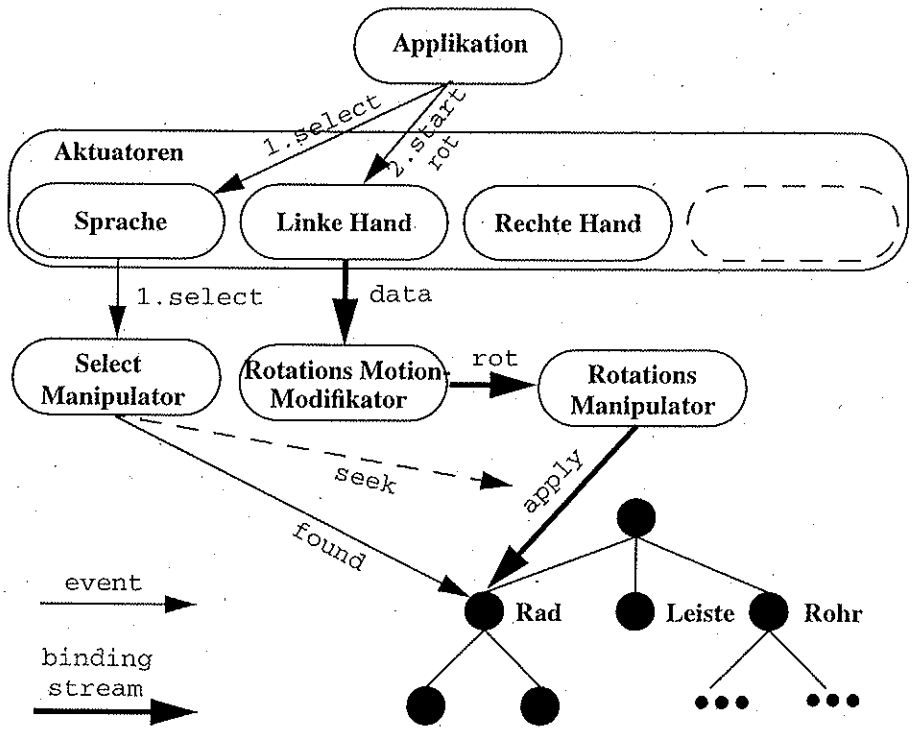

Abbildung 9: Kontinuierliche Interaktion mittels Aktuatoren und Modifikatoren („Dreh das Rad so herum")

und die Applikation als Event weitergegeben werden. Dieser Ansatz würde in einem Datenstreaming-Konzept resultieren und widerspräche grundlegend der Struktur einer Eventauswertung. Weiterhin läuft die Gestenerkennung parallel ab, zu einem beliebigen Zeitpunkt können für die gleiche Extremität gültige, unterschiedlich gewichtete Resultate vorliegen. Diese würden übermittelt, obwohl sie für eine stattfindende kontinuierliche Manipulation nicht relevant wären. Sie erforderten eine vorgeschaltete Filterung der Erkennerresultate, selbst wenn die entsprechende Extremität gerade eine Manipulation ausführen würde. Die asynchrone Kommunikation fände immer statt. Ein derartiges Event-basiertes Modell ist aus Performanzgründen für flüssige kontinuierliche Interaktion in einem virtuellen Szenario nicht akzeptabel. Wir setzen daher einen in Abbildung 9 illustrierten speziellen Anwendungsmodus ein. Die speziellen Trigger-Events veranlassen die Umschaltung in den kontinuierlichen Modus. Kontrolliert und umgesetzt wird die Benutzer-Interaktion von dann aktivierten Modulen, deren Datenfluß synchron und parallel dem der virtuellen Umgebung ist. Gegenseitige Bindungen zwischen diesen Modulen etablieren über den Zeitraum mehrerer Frames hinweg die kontinuierliche Manipulation.

Die Interaktionen des Benutzers werden durch *Aktuatoren* in der Repräsentation der virtuellen Umgebung vermittelt. Für eine Hand sind dies zum Beispiel die aktuelle Lage

und Position des Handgelenkmittelpunktes in Weltkoordinaten. *Motion-Modifikatoren* binden an diese Daten und testen, ob der jeweilige Aktuator zwischen jedem Renderschritt die Modifikator-eigenen Bedingungen erfüllt, zum Beispiel weiterhin eine Drehung ausführt [12]. Ist dieses der Fall, so versorgen sie entsprechende Manipulatoren über den Zeitraum der Bindung mit kontinuierlichen, durch Template-Intervallvergleich geglättete Manipulationsanweisungen. Sind die Bedingungen nicht mehr erfüllt, signalisieren die Modifikatoren den Abbau der Bindungen, die Interaktion wird insgesamt beendet. Aktuatoren und Motion-Modifikatoren sind, im Gegensatz zu der Event-getriebenen Erkennung, synchronisiert mit dem Datenfluß der virtuellen Umgebung. Für jedes neue zu berechnende Bild wird ein Update der eingebetteten Objekte durchgeführt. Solange Bindungen zwischen Aktuatoren, Motion-Modifikatoren und Manipulatoren bestehen, werden bei jedem neuen Frame die internen Aktionen der gebundenen Objekte durchgeführt; gebundene Aktuatoren können keine weitere Aktion als die gerade aktuelle ausführen. Da die einzelnen Erkenner ihre Ergebnisse in Form von gewichteten Hypothesen an das System weitergeben, und da gewisse Formanteile einer Geste denen einer anderen zu erkennenden Geste entsprechen können (die Merkmalsvektoren der einzelnen Gesten sind nicht vollständig orthogonal), kann es bei den Erkennern zu Überschneidungen kommen. Ist aber der durch die Erkenner referenzierte Aktuator bereits in einer kontinuierlichen Manipulation gebunden, so werden alle anderen diesen Aktuator möglicherweise betreffenden Erkenner-Events ignoriert. Inkonsistenzen im Interaktionsfluß werden so vollständig vermieden.

Die strikte Unterscheidung zwischen diskreter und kontinuierlicher Manipulation stellt die Voraussetzung zur Auswertung der in Abschnitt 3.3 beschriebenen multimodalen Äußerungen während der Interaktion zur Verfügung. Ein weiteres Beispiel: Die verbale Anweisung „*Dreh die Leiste* <Drehbewegung der Hand> ***so*** *um dieses Rad*" ist durch zwei Moduswechsel gekennzeichnet. Nach der Spezifikation der Manipulation („*Dreh*") und des zu manipulierenden Objektes („*die Leiste*") folgt ein Wechsel zur kontinuierlichen Interaktion (Trigger „***so***"), anschließend folgen weitere Informationen über die Manipulation, in diesem Fall die Spezifikation des Rotationsmittelpunktes („*um das Rad*"). Die Modalität wird hier mehrfach gewechselt um die Interaktion zu beschreiben. Erfolgt die Umsetzung der Manipulation bereits nach dem Trigger, so kann ein Korrekturansatz (wie in [13] erarbeitet) eingesetzt werden. Die schon erfolgende kontinuierliche Manipulation kann modifiziert, und die neuen Informationen über die Rotationsachse berücksichtigt werden.

6 Das System

Ziel unserer Forschung ist die Erprobung neuer Interaktionsformen für Multimedia- und VR-Systeme. Anhand einer Anwendung zur Virtuellen Konstruktion haben wir das SGIM-System, welches eine multimodale, sprachlich/gestische Steuerung vor einer Großbildprojektion ermöglicht, vorgestellt. In diesem werden neben deiktischen und einigen symbolischen Gesten, insbesondere auch Gesten mit mimetischem Charakter sprachgestützt verarbeitet. Erste Experimente bestätigen die Nützlichkeit dieser Gestentypen in Fällen, wo ausschließlich sprachliche Äußerungen zu komplex, unpräzise oder unnatürlich sind.

Zur Verarbeitung von sowohl diskreten wie auch kontinuierlichen multimodalen Interaktionen wurde ein gemischtes Event/Binding-basiertes Architekturkonzept vorgestellt, das unterschiedlich getriebene und synchronisierte Programmkomponenten umfaßt. Die Umsetzung kontinuierlicher Interaktionen erfolgt dabei über Aktuatoren, Motion-Modifikatoren und Manipulatoren, welche durch sprachlich oder gestisch getriggerte Events mit der VR-spezifischen Rendering-Loop synchronisiert, und für die Dauer einer Interaktion aneinander gebunden werden. Das vorgeschlagene Architekturkonzept leistet über die Manipulatoren auch die Integration in das Anwendungssystem zur Virtuellen Konstruktion.

Das gegenwärtige Demonstrator-System verbindet die verschiedenen Komponenten der AG1 des Verbundes „Virtuelle Wissensfabrik". Als Grundlage dient der Virtuelle Konstrukteur, ein wissensbasiertes System zur interaktiven Konstruktion, welches in unserer AG entwickelt wird und im Rahmen des SGIM-Projekts an die neuen Anforderungen angepaßt wurde. Die inkrementelle Spracherkennung [5] ist eine weitere grundlegende Komponente um eine flüssige Interaktion zu gewährleisten. Ihre Resultat, die produzierten zeitgestempelten Einzelwörter, werden in SGIM analysiert, klassifiziert, und dem Sprach/Gesten-Integrationsmechanismus zugeführt. Die Gestenerkennung wird mit den im Rahmen des Projektes erstellten Erkennermodulen bewerkstelligt. Neben Erkennern für verschiedene Formen der Zeigegeste wurden auch solche integriert, die das Öffnen und Schließen der Hände registrieren. Für die formale Definition einer Geste wurde eine Spezifikationssprache entwickelt, deren erheblich erweiterte Version aktuell umgesetzt wird. Aus einem weiteren Verbundteilprojekt wurden auch experimentell Armposturdaten, welche Kamera-basiert und mittels neuronaler Netze ermittelt werden [17], angebunden. Die Implementation der Visualisierung und den für die Interaktionsumsetzungen erforderlichen Komponenten, den Aktuatoren, Motion-Modifikatoren und Manipulatoren, wurde mit Hilfe des AVOCADO-Systems [18] realisiert. Mit diesen Mitteln wurden Interaktionen zur Deixis-Auswertung, zur Führung von Objekten und deren Verbindungen vollständig implementiert. Die Auswertung weiterer Interaktionen befindet sich im Erprobungsstadium.

7 Literatur

1. K. Böhm, W. Hübner & K. Väänänen: *GIVEN: Gesture Driven Interactions in Virtual Environments, A Toolkit Approach to 3D Interactions*. In Interfaces to Real and Virtual Worlds, Montpellier, France, 1992.
2. K. Böhm, W. Broll & M. Sokolewicz: *Dynamic Gesture Recognition Using Neural Networks; A Fundament for Advanced Interaction Construction*. In SPIE Conference Electronic Imaging Science & Technology, San Jose California, USA, 1994.
3. R. A. Bolt: *„Put-That-There": Voice and Gesture at the Graphics Interface*, Computer Graphics 14(3), S. 262-270, 1980.
4. D. Efron: *Gesture and Environment*, King's Crown Press, New York, 1941; current ed.: Gesture, race and culture, The Hague: Mouton, 1972
5. G. A. Fink, C. Schillo, F. Kummert & G. Sagerer: *Incremental speech recognition for multimodal interfaces*. In IECON'98: Proceedings of the 24th Annual Conference of the IEEE Industrial Electronics Society, Vol. 4, IEEE, 1998.

6. A. G. Hauptmann & P. McAvinney: *Gestures with speech for graphic manipulation.* In International Journal of Man-Machine Studies, Vol. 38, S. 231-249, 1993.
7. C. Huls, E. Bos & W. Claassen: *Automatic Referent Resolution of Deictic and Anaphoric Expressions.* In Computational Linguistics, Vol. 21, No 1, S.59-79, 1995.
8. B. Jung, M. Latoschik & I. Wachsmuth: *Knowledge-Based Assembly Simulation for Virtual Prototype Modeling.* In IECON'98 -Proceedings of the 24th Annual Conference of the IEEE Industrial Electronics Society, Vol. 4, IEEE, S. 2152-2157, 1998.
9. D. B. Koons, C. J. Sparrell & K. R. Thorisson: *Integrating Simultaneous Input from Speech, Gaze, and Hand Gestures.* In M. Maybury (Ed.): Intelligent Multimedia Interfaces, AAAI Press, S. 257-276, 1993.
10. S. Kopp: *Ein wissensbasierter Ansatz zur Modellierung von Verbindungen zur virtuellen Montage*, Diplomarbeit an der Technischen Fakultät der Universität Bielefeld, 1998.
11. M. Latoschik & I. Wachsmuth: *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), Lecture Notes in Artificial Intelligence, Volume 1371, Springer-Verlag, 1998.
12. M. Latoschik, M. Fröhlich, B. Jung & I. Wachsmuth: *Utilize Speech and Gestures to Realize Natural Interaction in a Virtual Environment.* In IECON'98 - Proceedings of the 24th Annual Conference of the IEEE Industrial Electronics Society, Vol. 4, IEEE, S. 2028-2033, 1998.
13. B. Lenzmann: *Benutzeradaptive und Multimodale Interface-Agenten.* Dissertation an der Technischen Fakultät der Universität Bielefeld, Infix Verlag, DISKI 184, 1998.
14. M. T. Maybury: *Research in Multimedia and Multimodal Parsing and Generation.* In P. McKevitt (Eds.): Journal of Artificial Intelligence Review: Special Issue on the Integration of natural Language and Vision Processing, Vol. 9, Kluwer, 1995.
15. D. McNeill: *Hand and Mind - What Gestures Reveal about Thought.* The University of Chicago Press, Chicago, USA, 1992.
16. J.-L. Nespoulous & A. R. Lecours: *Gestures: Nature and Function.* In J.-L. Nespoulous, P. Péron & A. R. Lecours (ed.): The Biological Foundations of Gestures: Motion and Semiotic Aspects, Lawrence Erlbaum Associates, New Jersey, London, 1986
17. C. Nölker &H. Ritter: *Illumination Independent Recognition of Deictic Arm Postures.* IECON'98 -Proceedings of the 24th Annual Conference of the IEEE Industrial Electronics Society, Vol. 4, IEEE, S. 2006-2011, 1998.
18. H. Tramberend: *Avocado: A Distributed Virtual Reality Framework.* In L. Rosenblum, P. Astheimer & D. Teichmann (Eds.): Proceedings of the Virtual Relity'99 IEEE Conference, Houston, USA, S. 14-21, 1999.
19. B. Rimé & L. Schiaratura: *Gesture and Speech.* In Feldmann & Rimé (eds.): Fundamentals of Nonverbal Behaviour, Press Syndicate of the University of Cambridge, New York, 1991.
20. D. Weimer & S. K. Ganapathy: *Interaction Techniques using Hand Tracking and Speech Recognition.* In M. M. Blattner & R. B. Dannenberg (Eds.): Multimedia Interface Design, ACM Press, S. 109-126, 1992.
21. A. D. Wexelblat: *An Approach to Natural Gesture in Virtual Environments.* In ACM Transactions on Computer-Human Interaction, Special Issue on Virtual Reality Software and Technology, Vol. 2 #3, 1995.