Sonderforschungsbereich 378
Ressourcenadaptive kognitive Prozesse

KI-Labor am Lehrstuhl für Informatik IV

Leitung: Prof. Dr. W. Wahlster

REAL

Universität des Saarlandes
FB 14 Informatik IV
Postfach 151150
D-66041 Saarbrücken
Germany

Tel. 0681 / 302-2363

Bericht Nr. 140

VITRA: Verbalisierung visueller Information

Gerd Herzog, Anselm Blocher, Klaus-Peter Gapp, Eva Stopp, Wolfgang Wahlster

Januar 1998

Sonderforschungsbereich 378
Ressourcenadaptive
Kognitive Prozesse

ISSN 0944-7814

140

VITRA: Verbalisierung visueller Information

Gerd Herzog, Anselm Blocher, Klaus-Peter Gapp,
Eva Stopp und Wolfgang Wahlster
Universität des Saarlandes
Fachbereich Informatik
Sonderforschungsbereich 314
Postfach 15 11 50
D-66041 Saarbrücken

Zusammenfassung

Das Projekt VITRA (Visual Translator) beschäftigt sich mit Grundfragen der Beziehung zwischen Sprache und Sehen. Ziel der experimentellen Studien ist die Entwicklung wissensbasierter Systeme zur Integration von visueller Wahrnehmung und der Verarbeitung natürlicher Sprache. Hierbei konnten erstmals automatische sprachliche Beschreibungen für aus realen Bildfolgen gewonnene Trajektorien erzeugt werden. Die vorliegende Arbeit stellt den in VITRA verfolgten Ansatz zur simultanen Auswertung und natürlichsprachlichen Beschreibung zeitveränderlicher Szenen genauer vor. Bei dieser Konzeption wird die Verarbeitung auf allen Stufen in inkrementeller Weise durchgeführt, eine wichtige Voraussetzung für die langfristig angestrebte Echtzeitverarbeitung.

Abstract

The project VITRA (Visual Translator) deals with the relationship between language and vision. Experimental studies are being carried out with the aim of developing knowledge-based systems for the integration of visual perception and natural language processing. For the first time, the automatic generation of natural language descriptions for recognized trajectories of objects in a real world image sequence has been achieved. In this article, the approach towards simultaneous evaluation and natural language description of time-varying scenes as it has been pursued in VITRA will be presented in more detail. With this conception, processing on all levels is carried out on an incremental basis, an important prerequisite for real-time performance.

In: Informatik – Forschung und Entwicklung, 11(1):12-19, 1996

1 Einleitung

Nachdem die beiden Teilgebiete Bild- und Sprachverstehen in der Künstlichen Intelligenz (KI) lange nahezu unabhängig voneinander behandelt wurden, gewinnt die Entwicklung wissensbasierter Systeme zur Integration von maschinellem Sehen und der Verarbeitung natürlicher Sprache zunehmend an Bedeutung [1], [7], [20], [29], [43]. Dabei ergeben sich vielfältige Anwendungsperspektiven in den unterschiedlichsten Forschungsfeldern: Robotik, Medizintechnik, Fernerkundung, Fahrerassistenzsysteme, intelligente Leitstände und Intellimediasysteme stellen einige wichtige Beispiele aus der Vielzahl potentieller Anwendungsbereiche dar.

Die Beziehung zwischen Sprache und Wahrnehmung bildet den Forschungshintergrund für das Projekt VITRA (Visual Translator), das sich mit dem Entwurf und der Realisierung von wissensbasierten Systemen für den natürlichsprachlichen Zugang zu visueller Information auseinandersetzt [19]. Langfristiges Hauptziel der in VITRA verfolgten Forschungsrichtung ist es, die komplexen Informationsverarbeitungsprozesse, welche der Interaktion von visueller Wahrnehmung und Sprachproduktion zugrunde liegen, algorithmisch zu beschreiben und zu erklären [42]. Neben diesem kognitionswissenschaftlichen Ansatz geht es aus der ingenieurwissenschaftlichen Perspektive darum, den Benutzern die Resultate eines bildverstehenden Systems besser zugänglich und leichter verständlich zu machen.

In VITRA werden unterschiedliche Diskursbereiche und verschiedenartige Kommunikationssituationen im Hinblick auf den natürlichsprachlichen Zugang zu visueller Information näher untersucht. In enger Kooperation mit dem Fraunhofer-Institut für Informations- und Datenverarbeitung (IITB), Karlsruhe, [30] sowie dem Institut für Prozeßrechentechnik und Robotik (IPR) der Universität Karlsruhe werden dabei folgende Szenarien betrachtet:

- Beantwortung von Fragen über Beobachtungen in einer Straßenverkehrsszene [3], [37]
- Generierung von Simultanbeschreibungen anhand kurzer Ausschnitte aus Videoaufnahmen eines Fußballspiels [4], [18] sowie verschiedener Straßenverkehrsszenen mit Fahrzeugen und Fußgängern [17]
- Erzeugung inkrementeller, multimodaler Wegbeschreibungen basierend auf einem 3D-Modell des Saarbrücker Campus [15], [28]
- Natürlichsprachliche Interaktion mit einem autonomen mobilen Robotersystem [27], [39]

Der vorliegende Beitrag konzentriert sich auf die automatische Auswertung und Beschreibung zeitveränderlicher Szenen ausgehend von Kamerabildfolgen.

2 Von der Bildfolge zur intentionalen Simultanbeschreibung

Beim natürlichsprachlichen Zugang zu bildverstehenden Systemen müssen drei Verarbeitungsebenen unterschieden werden [16]. Auf der sensorischen Ebene besteht die Aufgabe der Bildanalyse in der Identifizierung und Klassifizierung der beobachtbaren Objekte. Sie leistet damit den Übergang von der Bildfolge zur Szenenfolge im 3D-Raum. Die Szenenfolgenanalyse realisiert den Schritt von der rekonstruierten *geometrischen Szenenbeschreibung* [31] zu einer konzeptuellen Beschreibung des Szenengeschehens. Diese konzeptuelle Ebene dient der referenzsemantischen Verankerung sprachlicher Einheiten [6], [8], [21]. Im Hinblick auf den variierbaren Detaillierungsgrad sprachlicher Beschreibungen sind hierbei, wie in Abb. 1 angedeutet, konzeptuelle Einheiten auf unterschiedlichen Abstraktionsebenen bereitzustellen. Auf der dritten, der linguistischen Ebene werden die an der

Ziel: Verbalisierung visueller Wahrnehmungen

Ausdruck:	**Erkennung von:**
Sprachliche Beschreibung	
• erkannter Objekte:	**Objekten**
Da sind zwei PKW und ein Bus.	
• der räumlichen Lage von Objekten:	**Räumlichen Relationen**
Der Bus steht vor der Kirche.	
• von Objektbewegungen:	**Vorgängen**
Der Bus fährt an.	
• vermuteter Ziele und Pläne der beobachteten Agenten:	**Plänen**
Der Bus wartet vor der Ampel.	

Abbildung 1: Ebenen der sprachlichen Szenenbeschreibung

Wahrnehmung orientierten konzeptuellen Strukturen in natürlichsprachliche Äußerungen überführt.

Frühere natürlichsprachliche Zugangssysteme, wie HAM-ANS [44] und NAOS [31], beschränken sich auf eine *a posteriori* Analyse, die lediglich retrospektive Szenenbeschreibungen zuläßt. Im Gegensatz dazu wird in VITRA eine *inkrementelle* Analyse durchgeführt, bei der die Verarbeitung der visuellen Daten simultan zum Fortschreiten der Szenenfolge geschieht. Hierdurch werden Informationen zur aktuellen Szene bereitgestellt und unmittelbare Systemreaktionen, wie z.B. eine natürlichsprachliche Simultanbeschreibung, ermöglicht.

Aufgrund des eingesetzten, flexibleren Analysemodus ergibt sich in unserem Ansatz die in Abb. 2 skizzierte Verarbeitungskaskade. Die Digitisierung und Auswertung der Rohbilder erfolgt mit den durch unsere Projektpartner am IITB entwickelten Systemen [24], [25], [35], [40].

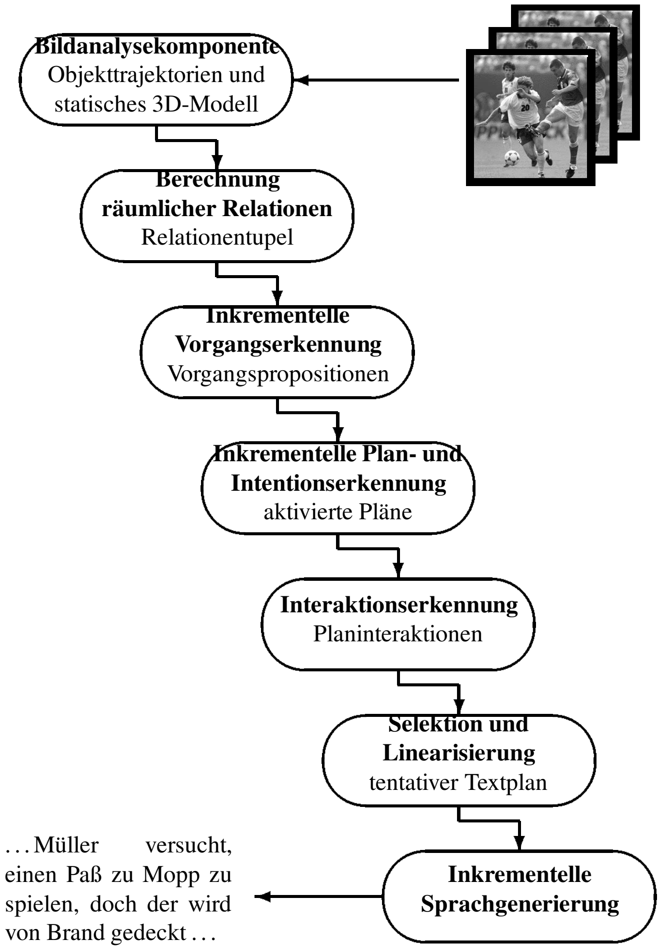


Abbildung 2: Die Verarbeitungskaskade in VITRA

Die Szenenfolgeanalyse in VITRA umfaßt neben der Extraktion räumlicher Beziehungen und interessanter Bewegungsvorgänge auch die Erkennung vermuteter Handlungsintentionen sowie von Planinteraktionen der beobachteten Agenten. Bei der Sprachproduktion müssen die mitteilungswürdigen Propositionen ausgewählt und in einem vorläufigen Textplan angeordnet werden. Unter Rückgriff auf den ständig aktualisierten Textplan werden simultan zum Ablauf der Szene natürlichsprachliche Äußerungen generiert und ausgegeben.

3 Inkrementelle Szenenfolgenanalyse

Ausgehend von der sukzessive durch das Bildanalysesystem bereitgestellten geometrischen Szenenbeschreibung erfolgt die weitergehende Interpretation der Szenenfolge, welche die für eine sprachliche Beschreibung notwendigen höheren konzeptuellen Einheiten, wie räumliche Relationen, auftretende Bewegungsvorgänge, aktivierte Handlungspläne und Planinteraktionen, zur Verfügung stellt.

3.1 Auswertung räumlicher Relationen

Die semantische Analyse räumlicher Präpositionen führt auf den Begriff *räumliche Relation* als einzelsprachunabhängige Bedeutungseinheit. Man definiert räumliche Relationen, indem man Bedingungen über räumliche

Gegebenheiten einer Objektkonfiguration spezifiziert, wie z.B. Distanz zwischen Objekten (topologisch) oder ihre relative Lage bezüglich einer Orientierung (projektiv) [13], [26], [41].

Die Berechnung der Semantik räumlicher Relationen wird in VITRA in mehreren Stufen vollzogen. Die Basis bildet die *Grundbedeutung*, welche nur die geometrische Objektbeschreibung und die Gebrauchsart der Relation — intrinsisch, extrinsisch oder deiktisch — berücksichtigt [32]. Auf höherer Ebene wird mittels zusätzlichem kontextspezifischem konzeptuellem Wissen eine erweiterte Semantik räumlicher Relationen ausspezifiziert [9].

Zur Berechnung der Grundbedeutung räumlicher Relationen wird ein von der Ausdehnung des Referenzobjekts und der Gebrauchsart der Relation abhängiges *lokales Koordinatensystem* erstellt (Abb. 3). Dieses lokale Referenzsystem dient der Skalierung des Raums zur Messung von Distanz und Winkel zwischen zu lokalisierendem und Referenzobjekt Der *Grad* der Anwendbarkeit einer räumlichen Relation wird durch geeignete Abbildung von Distanz bzw. Winkel mittels relationenspezifischer Bewertungsfunktionen auf das Anwendbarkeitsintervall $[0..1]$ ermittelt. Die Durchführung erster experimenteller Untersuchungen bezüglich der Bewertung projektiver Relationen zeigte eine lineare Korrelation zwischen Richtungsabweichung und Anwendbarkeitsgrad bei quadratischen Referenzobjekten [10, 11].

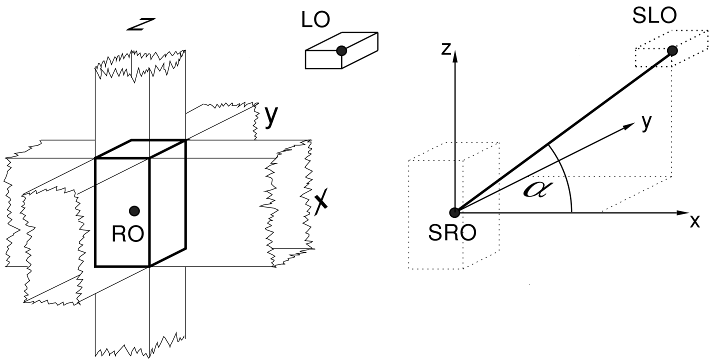


Abbildung 3: Lokales Koordinatensystem und Winkelabweichung

Die graphische Darstellung sogenannter *Anwendbarkeitsfelder* ermöglicht die Visualisierung der Anwendbarkeitsstruktur räumlicher Relationen bezüglich eines bestimmten Referenzobjekts und Kontexts. Abb. 4 gibt die graphische Darstellung zweier Anwendbarkeitsräume. Man beachte, daß hier die durch die projektive Relation bedingte Richtungsabweichung mit einem Distanzkonzept kombiniert wurde.

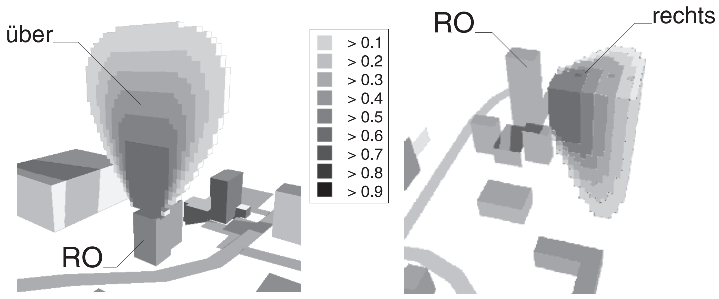


Abbildung 4: Ausschnitte der 3D Anwendbarkeitsstruktur von *über* und *rechts* zur Objektlokalisierung

Auch die Kernsemantik von Relationen, welche sich nicht unmittelbar in Distanz- bzw. Richtungsrelationen kategorisieren lassen, wie z.B. *auf*, *neben* und *zwischen*, lassen sich im vorgestellten Modell formalisieren. So kann die Berechnung der räumlichen Relation *zwischen* beispielsweise auf eine zweimalige Berechnung der Relation *vor* zurückgeführt werden. Die Kombination elementarer Relationen durch die Bildung gewichteter Minima erlaubt die Berechnung zusammengesetzter Relationen, wie z.B. *rechts-hinter* oder *links-bei* [9].

3.2 Charakterisierung von Objektbewegungen

Neben räumlichen spielen zeitliche Aspekte bei der Auswertung einer Szenenfolge eine zentrale Rolle. Aus den erkannten Objekttrajektorien müssen konzeptuelle Beschreibungen abgeleitet werden, welche die relevanten Objektbewegungen in geeigneter Weise charakterisieren. Im Hinblick auf die natürlichsprachliche Beschreibung einer Bildsequenz dienen derartige konzeptuelle Einheiten dazu, die referentielle Bedeutung der korrespondierenden Bewegungs- und Handlungsverben zu erfassen.

Generische Vorgangsmodelle, d.h. deklarative Beschreibungen von Klassen von Bewegungsvorgängen, bilden die Grundlage für die Erkennung relevanter Objektbewegungen. Diese Vorgangskonzepte sind in einer Abstraktionsheterarchie angeordnet, die auf zeitlicher Dekomposition in Teilbewegungen und auf der Spezialisierungsrelation (`laufen` ist z.B. eine spezielle Art von `bewegen`) basiert.

Header: (BALL-TRANSFER ?p1*player ?b*ball ?p2*player)
Conditions: (same (TEAM ?p1) (TEAM ?p2))
Subconcepts: (BALLBESITZ ?p1 ?b) [I1]
(FREI-BEWEGEN ?b) [I2]
(BALLBESITZ ?p2 ?b) [I3]
Temporal-Relations: [I1] :meets [BALL-TRANSFER]
[I1] :meets [I2]
[I2] :equal [BALL-TRANSFER]
[I2] :meets [I3]

Abbildung 5: Vorgangsmodell `Ball-Transfer`

Abb. 5 zeigt eine vereinfachte Definition des Konzepts `ball-transfer`, bei dem auf den Ballbesitz eines Spielers eine freie Bewegung des Balles folgt, die wiederum mit dem Ballbesitz eines Mitspielers abschließt. Zeitliche Beziehungen werden hierbei mittels der qualitativen Relationen zwischen je zwei Zeitintervallen [2] ausgedrückt.

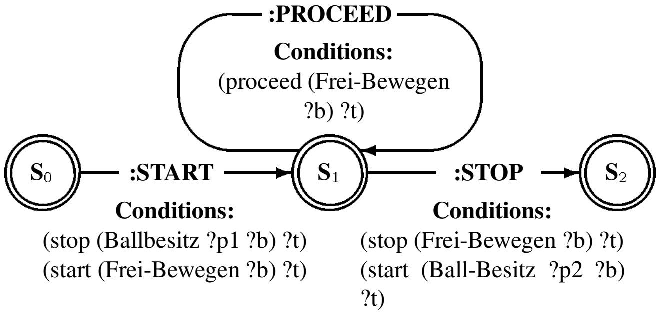


Abbildung 6: Ablaufschema für `Ball-Transfer`

Um einen Simultanbericht für eine gerade ablaufende Szenenfolge zu ermöglichen, müssen interessante Bewegungsabläufe schritthaltend erkannt und bereits während ihres Auftretens beschrieben werden. Zu diesem Zweck werden die intervall-basierten Vorgangsmodelle in eine für eine inkrementelle Erkennung besser geeignete Darstellungsform, sogenannte *Ablaufschemata*, transformiert. Ein Ablaufschema ist ein markierter, gerichteter Graph mit getypten Kanten, der den prototypischen Verlauf eines Vorgangs modelliert, indem die für jeden einzelnen Zeitpunkt zu erfüllenden Bedingungen formuliert werden. Die drei Verlaufsprädikate `start`, `proceed` und `stop` dienen dazu, den aktuellen Zustand eines gerade auftretenden Bewegungsvorgangs zu charakterisieren. Die inkrementelle Erkennung einer Vorgangsausprägung entspricht der schrittweisen Traversierung des zugehörigen Ablaufschemas. Die automatische Konstruktion der Ablaufschemata erfolgt mit Methoden des zeitlichen Schließens [14]. Abb. 6 zeigt das für das Konzept `Ball-Transfer` erzeugte Ablaufschema.

3.3 Intentionale Interpretation

Beim Beschreiben von Handlungen und Geschehnissen beschränkt sich ein menschlicher Beobachter im allgemeinen nicht auf visuelle Informationen, sondern bezieht auch Absichten und Ziele der Akteure in seine Schilderung mit ein. Diese intentionale Interpretation spiegelt sich in Sätzen wie „A will X tun.“, „A tat X, um Y zu erreichen.“ oder einfach „A verfolgt B.“ wider.

Ein Kriterium für die Wahl der Fußballdomäne in VITRA war die Tatsache, daß der Einfluß der bei den Agenten vermuteten Intentionen auf die sprachliche Beschreibung besonders offenkundig erscheint. Unter Berücksichtigung von Spielposition, Mannschaftszugehörigkeit und der Rollenverteilung in Standardsituationen können in jeder Situation Annahmen über stereotypische Intentionen gemacht werden.

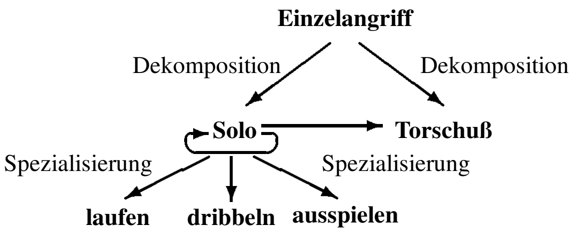


Abbildung 7: Ausschnitt einer Planhierarchie in VITRA

Für die intentionale Interpretation wird in VITRA auf Planerkennungsmethoden zurückgegriffen [33], [34]. Das Wissen über Ziele und Pläne ist in einer Planhierarchie repräsentiert, die durch Aktionen und deren hierarchische sowie zeitliche Beziehungen gebildet wird. Innere Knoten stellen abstrakte Aktionen dar. Die Blätter repräsentieren elementare, d.h. wahrnehmbare, generische Aktionen. Die Knoten enthalten zusätzlich Informationen über notwendige Vorbedingungen und den intendierten Effekt der korrespondierenden Aktion. Jeder zusammenhängende Teilgraph, der von einer zeitlich dekomponierbaren Aktion bis zu deren Blättern reicht, entspricht einem Plan. Zeitliche Dekomposition und Spezialisierungsrelation erlauben in Plänen die Verwendung von Aufeinanderfolge, Alternativen und Wiederholungen. Abb. 7 zeigt einen beispielhaften Ausschnitt einer Planhierarchie. Die unmarkierten Zeitkanten repräsentieren dabei eine unmittelbare zeitliche Aufeinanderfolge.

Unter Einsatz von Fokussierungsheuristiken werden potentielle Pläne inkrementell erkannt, wobei drei Arten von vermuteten Intentionen abgeleitet werden können. Der intendierte Effekt des aktuellen Planschrittes kann als zustandsgerichtete Intention aufgefaßt werden („Bommer greift Mopp an, um den Ball zu bekommen“). Die aufgrund einer Planhypothese erwarteten zukünftigen Aktionen stellen handlungsgerichtete Intentionen dar („Mopp will Maier den Ball zuspielen.“). Die Wurzel des aktuellen Plans definiert das vom Agenten verfolgte Oberziel („Mopp startet einen Einzelangriff.“).

Ausgehend von den beobachteten Plänen können in einem weiteren Verarbeitungsschritt kooperative („Während
Mopp einen Einzelangriff startet, läuft sich Maier frei.“) und antagonistische Interaktionen („Als Mopp nach vorne läuft, greift Bommer ihn an.“) zwischen den Plänen mehrerer Akteure erkannt werden. Neben diesen jeweils simultanen oder sequentiellen Interaktionen gibt es die Möglichkeit, kollektive Aktionen darzustellen.

Aufbauend auf der Intentions- und Interaktionserkennung lassen sich auch Planfehlschläge interpretieren. Eine Reihe von Fehlschlagsursachen können in VITRA bereits erkannt werden [34]. Ein Plan kann fehlschlagen aufgrund falscher Annahmen des Agenten hinsichtlich der Erfülltheit von Vorbedingungen oder bedingt durch einen antagonistischen Plan eines Kontrahenten. Im Falle einer kooperativen Interaktion kann ein Versagen des Partners vorliegen.

4 Generierung natürlichsprachlicher Simultanbeschreibungen

Eine Simultanbeschreibung ist im Gegensatz zu einer retrospektiven Beschreibung dadurch geprägt, daß der vollständige Verlauf der betrachteten Szenenfolge zum jeweiligen Zeitpunkt der Textgenerierung noch nicht bekannt ist. Da sich die Beschreibung auf das aktuelle Geschehen konzentrieren soll, muß über Vorgänge und Aktionen bereits berichtet werden während diese ablaufen und eventuell noch nicht vollständig erkannt wurden. Diese Besonderheiten der Kommunikationssituation machen eine inkrementelle Verarbeitungsstrategie erforderlich, bei der mit der Verbalisierung begonnen wird, bevor der Inhalt einer Äußerung bis ins letzte Detail geplant werden kann.

Die Sprachproduktion in VITRA umfaßt Prozesse zur Selektion, Linearisierung und Enkodierung von Propositionen. Das hierbei verwendete Hörermodell stellt eine Imaginationskomponente bereit, die dazu dient, die beim Hörer vermutete visuelle Konzeptualisierung der beschriebenen Szene zu antizipieren, damit eine möglichst adequate Beschreibung erzeugt werden kann.

4.1 Selektion und Linearisierung von Propositionen

Da die betrachtete zeitveränderliche Szene schritthaltend beschrieben werden soll, unterliegt die Sprachproduktion starken zeitlichen Restriktionen. Das System kann folglich nicht über alle Vorgänge und Aktionen berichten, die es erkannt hat, sondern muß sich für diejenigen entscheiden, die es dem Hörer ermöglichen, dem Szenengeschehen zu folgen. Entsprechend dem Kooperationsprinzip von Grice [12] sollte der Hörer dabei unter Vermeidung redundanter Information über alle relevanten Fakten informiert werden.

Die Relevanz einer Proposition wird durch vielfältige Faktoren bestimmt und ändert sich dynamisch mit dem Fortschreiten der Szene. Zu den besonders wichtigen Aspekten zählen hierbei der Informationsgehalt, der mit der Komplexität des zugehörigen generischen Konzepts korreliert, die Auftrittshäufigkeit, die Aktualität und der Erkennungszustand. Um Redundanz zu vermeiden, werden solche Propositionen, die durch andere schon erwähnte Propositionen impliziert werden, nicht selektiert. Beispielsweise ist für den Hörer evident, daß nach einem Zuspiel von Mopp zu Maier letzterer in Ballbesitz ist. Die taxonomische Organisation der konzeptuellen Wissensbasis bildet eine wichtige Voraussetzung zur Ableitung derartiger Zusammenhänge.

Weitere Selektionsprozesse im Rahmen der Enkodierung betreffen die Wahl geeigneter Deskriptionen für Objekte, Lokationen und temporale Angaben, um systeminterne Repräsentationsstrukturen in natürlichsprachliche Äußerungen überführen zu können. Entscheidend ist dabei der stetige Rückgriff auf ein Textgedächtnis und das Hörermodell, deren Inhalte die Auswahlprozesse maßgeblich beeinflussen.

Der Linearisierungsprozeß bestimmt die Reihenfolge, in der die ausgewählten Propositionen im Text erwähnt werden sollen und legt damit einen vorläufigen Textplan fest. Die Linearisierung erfolgt primär unter dem Gesichtspunkt der zeitlichen Abfolge der korrespondierenden Vorgänge und Aktionen, wobei zusätzlich Fokuskriterien zur Steigerung der Textkohärenz Berücksichtigung finden.

4.2 Antizipation der Hörervorstellung

Im Hörermodell werden die geplanten Äußerungen mit angenommenen Hörervorstellungen in einer Antizipationsrückkopplungsschleife [22] ausgehend von bereits mitgeteilten Szenenabschnitten verglichen, um Verständnisprobleme bzw. redundante Informationen auszuschließen.

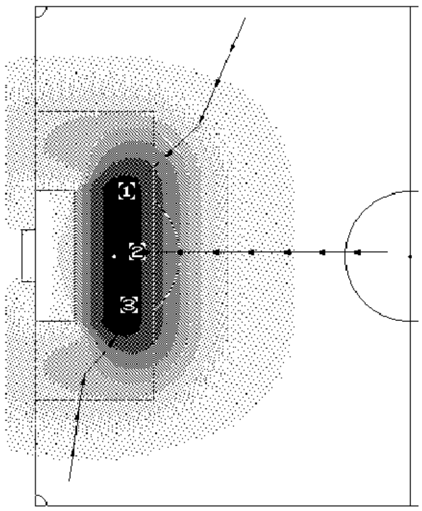
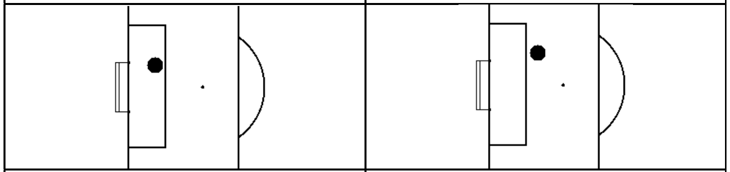


Abbildung 8: Ein Beispiel zu kontextsensitiver Approximation: `(vor Spieler linker_Torraum)`

Das Hörermodell muß daher in der Lage sein, im aktuellen Kontext eine möglichst typische Interpretation der geplanten Äußerung zu finden [36], [38]. Dazu werden räumliche Propositionen mit Hilfe des Gradientenverfahrens auf die in Abschnitt 3.1 beschriebenen Anwendbarkeitsräume – hier interpretiert als Typikalitätsverteilungen – approximiert, so daß sie die gegebene Beschreibung im aktuellen Kontext maximaltypisch erfüllen (Abb. 8). Die Vorgehensweise bei zeitlichen Relationen ist analog.

Die zur Verbalisierung anstehende Proposition wird anhand des zugehörigen Vorgangsmodells in ihre *propositionale Elementarstruktur* überführt. Die entstehenden zeitpunktweise zusammengefaßten Mengen von räumlichen Propositionen werden mittels des Approximationsalgorithmus in die geometrische Repräsentation mentaler Bilder umgewandelt. Die durch die Abfolge der konstruierten Bildvorstellungen entstandene *imaginierte* Szenenfolge wird dann analog zum ursprünglichen Szenenfolge reanalysiert.

Visuelle Perzeption	**Imaginierte Bildvorstellung**



Proposition	A-Grad	Proposition	A-Grad
(in Mopp linke_Spielhälfte)	1.0	(in Mopp linke_Spielhälfte)	1.0
(in Mopp linker_Torraum)	1.0	(vor Mopp linker_Torraum)	0.9
...		...	

Analysierender Vergleich	
Differenz-Propositionenpaar	$\left.\begin{array}{l}((\textit{in}\ \text{Mopp linker\_Torraum})\ 1.0)_{vp}\\((\textit{vor}\ \text{Mopp linker\_Torraum})\ 0.9)_{im}\end{array}\right\}\ncong$
Zu explizierende Proposition	(in Mopp linker_Torraum)
Mögliche Unterspezifikation	linker_Torraum $\rightarrow$ Torraum
Zu generierende Proposition	(in Mopp Torraum)

Abbildung 9: Ein Beispiel der Verfeinerung des Textplans durch Antizipation der Hörervorstellung

Zum Vergleich der intendierten und imaginierten Äußerungsgehalte werden im folgenden – wie in Abb. 9 dargestellt – verschiedene Mengen von Propositionen aus den konzeptuellen Beschreibungen von realer und imaginierter Szene gegenübergestellt [5]. Dieser Vergleich findet *nicht* auf der Ebene absoluter Koordinaten statt, da in diesem Fall der Kontext nicht berücksichtigt werden könnte. Ziel ist es, den vorläufigen Textplan hinsichtlich der folgenden drei pragmatischen Kriterien zu überprüfen und gegebenenfalls Verbesserungen vorzuschlagen:

- *Referenz:* Die verwendeten Objektbeschreibungen müssen für Sprecher und Hörer gleichermaßen verständlich sein. Redundanz soll vermieden werden.
- *Plausibilität:* Die neue Information muß in den gegebenen Kontext integriert werden können.
- *Korrektheit:* Das intendierte und das antizipierte Verständnis müssen übereinstimmen. Falsche Inferenzen durch den Hörer sollen vermieden werden.

Die in Abb. 10 noch einmal schematisch dargestellte Antizipationsrückkopplungsschleife resultiert nötigenfalls in einer Änderung des Textplans. Hierunter fallen insbesondere die Modifikation optionaler Tiefenkasus und die Generierung von Klärungseinschüben.

4.3 Inkrementelle Verbalisierung

Bei der Enkodierung einer ausgewählten Proposition, d.h. der Umwandlung einer konzeptuellen Beschreibung in natürlichsprachliche Äußerungen, wird zunächst ein geeignetes Verb ausgewählt und der damit assoziierte Kasusrahmen instantiiert. Die Lexikalisierung erfolgt unter Zugriff auf ein Konzeptlexikon, das die Verbindung zwischen außersprachlichen und sprachlichen Einheiten realisiert. Ergänzende Selektionsprozesse entscheiden, welche Information bezogen auf die einzelnen Kasusfüller mitgeteilt werden soll. Die selektierte Infomation wird in natürlichsprachliche Ausdrücke umgewandelt, die auf Zeit, Ort und Objekte referieren. In Abhängigkeit vom

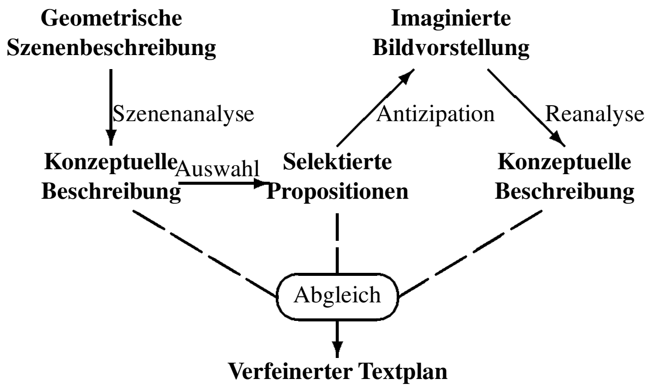


Abbildung 10: Antizipationsrückkopplungsschleife

Textgedächtnis ist dabei gegebenenfalls auch Anaphernbildung möglich. Durch Zeitformen bzw. mittels Temporaladverbien ist es möglich, temporale Information auszudrücken; räumliche Information wird analog dazu durch geeignete Präposiţonalphrasen beschrieben. Die geeignete Selektion von Attributen für interne Objektbezeichner ermöglicht dem Hörer eine eindeutige Identifikation des intendierten Referenten. Charakterisierende Attribute können im Hörermodell eingetragenes Vorwissen, die räumliche Position oder auch vorerwähnte Vorgänge und Aktionen sein, an denen das Objekt beteiligt war.

Die aus der Wortwahl und der Bestimmung morphosyntaktischer Information resultierende präverbale Struktur wird Stück für Stück an den Oberflächengenerator übergeben, der für die grammatische Enkodierung, die Linearisierung und die Flexion verantwortlich ist. Für die syntaktische Beschreibungsebene wird eine lexikalisierte Baumadjunktionsgrammatik mit Merkmalsunifikation (LTAG) verwendet [23]. Der erweiterte Lokalitätsbereich sowie die flexible Expansion partieller Strukturen durch Substitution und Adjunktion machen LTAGs zu geeigneten Kandidaten für die inkrementelle syntaktische Generierung.

5 Zusammenfassung und Ausblick

Die sprachorientierte KI-Forschung im Bereich der Integration von Sprache und Wahrnehmung zielt auf eine (extreme) operationale Form einer Referenzsemantik ab, die bis zur sensorischen Ebene hinunterreicht. VITRA stellt das bislang einzige System dar, mit dem natürlichsprachliche Beschreibungen für Objekttrajektorien, die aus Realweltbildfolgen abgeleitet wurden, automatisch generiert werden können. Die Szenenfolgenanalyse in VITRA beschränkt sich dabei nicht allein auf rein visuelle Aspekte, sondern schließt auch die Ableitung vermuteter Handlungsintentionen der beobachteten Agenten mit ein.

In enger Kooperation mit unseren Projektpartnern am IITB wird in VITRA ein Ansatz zur Kopplung von bildverstehenden und sprachverstehenden Systemen verfolgt, bei dem die Verarbeitungsprozesse auf allen Ebenen simultan zu der betrachteten Bild- bzw. Szenenfolge ablaufen. Langfristig wird damit eine Echtzeitverarbeitung angestrebt, die es ermöglichen soll, visuelle Information unmittelbar zu interpretieren, um Systemreaktionen, wie z.B. sprachliche Äußerungen oder motorische Aktionen eines Robotersystems, basierend auf den jeweils vorliegenden Analyseergebnissen sofort durchführen zu können.

Als Einschränkung werden zur Zeit nur Aufnahmen mit stationärer Kamera untersucht, sodaß eine Rückkoppelung von der Szenenfolgenanalyse zum Bildauswertungssystem bislang nicht realisiert werden kann. Im Hinblick auf die geplante aktive Sensorsteuerung bietet der in VITRA konsequent realisierte Ansatz der inkrementellen Verarbeitung jedoch die notwendigen Voraussetzungen. Die kontextabhängige Auswahl relevanter Objektklassen, partiell erkannte Vorgänge, aktivierte Handlungspläne und imaginierte Folgeszenen können notwendige Parameter zur Sensorsteuerung liefern.

Technische Anmerkungen

Bei den experimentellen Arbeiten zur Realisierung des natürlichsprachlichen Zugangssystems VITRA wird vorwiegend objekt-orientierte Programmierung in der KI-Programmiersprache Common Lisp und dem Common Lisp

Object System (CLOS) eingesetzt. Die Realisierung der graphischen Systemoberfläche basiert auf dem Common Lisp Interface Manager (CLIM). Die Systementwicklung erfolgt auf Symbolics 36xx Lisp-Maschinen, Symbolics UX1200S Lisp-Coprozessoren und auf Hewlett Packard 9720 sowie SPARC-basierten Arbeitsplatzrechnern.

Danksagung

Die hier beschriebenen Forschungsarbeiten wurden von der Deutschen Forschungsgemeinschaft im Rahmen des Sonderforschungsbereichs 314 „Künstliche Intelligenz und wissensbasierte Systeme“, Projekt N2: VITRA, gefördert.

Literatur

[1] *Artificial Intelligence Review Journal, 8, Special Volume on the Integration of Natural Language and Vision Processing*, 1994.

[2] ALLEN, J. F.: *Towards a General Theory of Action and Time.* Artificial Intelligence, 23(2):123–154, 1984.

[3] ANDRÉ, E., G. BOSCH, G. HERZOG und T. RIST: *Characterizing Trajectories of Moving Objects Using Natural Language Path Descriptions.* In: *Proc. of the 7th ECAI*, Vol. 2, S. 1–8, Brighton, UK, 1986.

[4] ANDRÉ, E., G. HERZOG und T. RIST: *On the Simultaneous Interpretation of Real World Image Sequences and their Natural Language Description: The System SOCCER.* In: *Proc. of the 8th ECAI*, S. 449–454, München, 1988.

[5] BLOCHER, A. und J. R. J. SCHIRRA: *Optional Deep Case Filling and Focus Control with Mental Images: ANTLIMA-KOREF.* In: *Proc. of the 14th IJCAI*, S. 417–423, Montreal, Canada, 1995.

[6] CARSTEN, I. und T. JANSON: *Verfahren zur Evaluierung räumlicher Präpositionen anhand geometrischer Szenenbeschreibungen.* Fachbereich für Informatik, Univ. Hamburg, 1985.

[7] CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE: *Images et Langages: Multimodalité et Modélisation Cognitive, Colloque Interdisciplinaire du Comité National de la Recherche Scientifique*, Paris, 1993.

[8] FÜRNSINN, M., M. N. KHENKHAR und B. RUSCHKOWSKI: *GEOSYS — Ein Frage-Antwort-System mit räumlichem Vorstellungsvermögen.* In: ROLLINGER, C.-R. (Hrsg.): *Probleme des (Text-) Verstehens, Ansätze der künstlichen Intelligenz*, S. 172–184. Niemeyer, Tübingen, 1984.

[9] GAPP, K.-P.: *Basic Meanings of Spatial Relations: Computation and Evaluation in 3D Space.* In: *Proc. of AAAI-94*, S. 1393–1398, Seattle, WA, 1994.

[10] GAPP, K.-P.: *Angle, Distance, Shape, and their Relationship to Projective Relations.* In: MOORE, J. D. und J. F. LEHMAN (Hrsg.): *Proc. of the 17th Annual Conference of the Cognitive Science Society*, S. 112–117. Lawrence Erlbaum, Mahwah, NJ, 1995.

[11] GAPP, K.-P.: *An Empirically Validated Model for Computing Spatial Relations.* In: WACHSMUTH, I., C.-R. ROLLINGER und W. BRAUER (Hrsg.): *KI-95: Advances in Artificial Intelligence. 19th Annual German Conference on Artificial Intelligence*, S. 245–256. Springer, Berlin, Heidelberg, 1995.

[12] GRICE, H. P.: *Logic and Conversation.* In: COLE, P. und J. L. MORGAN (Hrsg.): *Speech Acts*, S. 41–58. Academic Press, London, 1975.

[13] HERSKOVITS, A.: *Language and Spatial Cognition. An Interdisciplinary Study of the Prepositions in English.* Cambridge University Press, Cambridge, London, 1986.

[14] HERZOG, G.: *Utilizing Interval-Based Event Representations for Incremental High-Level Scene Analysis.* In: AURNAGUE, M., A. BORILLO, M. BORILLO und M. BRAS (Hrsg.): *Proc. of the 4th International Workshop on Semantics of Time, Space, and Movement and Spatio-Temporal Reasoning*, S. 425–435, Château de Bonas, France, 1992. Groupe “Langue, Raisonnement, Calcul”, Toulouse.

[15] HERZOG, G., W. MAASS und P. WAZINSKI: *VITRA GUIDE: Utilisation du Langage Naturel et de Représentation Graphiques pour la Description d'Itinéraires*. In: *Images et Langages: Multimodalité et Modélisation Cognitive, Colloque Interdisciplinaire du Comité National de la Recherche Scientifique*, S. 243–251, Paris, 1993.

[16] HERZOG, G., T. RIST und E. ANDRÉ: *Sprache und Raum: Natürlichsprachlicher Zugang zu visuellen Daten*. In: FREKSA, C. und C. HABEL (Hrsg.): *Repräsentation und Verarbeitung räumlichen Wissens*, S. 207–220. Springer, Berlin, Heidelberg, 1990.

[17] HERZOG, G. und K. ROHR: *Integrating Vision and Language: Towards Automatic Description of Human Movements*. In: WACHSMUTH, I., C.-R. ROLLINGER und W. BRAUER (Hrsg.): *KI-95: Advances in Artificial Intelligence. 19th Annual German Conference on Artificial Intelligence*, S. 257–268. Springer, Berlin, Heidelberg, 1995.

[18] HERZOG, G., C.-K. SUNG, E. ANDRÉ, W. ENKELMANN, H.-H. NAGEL, T. RIST, W. WAHLSTER und G. ZIMMERMANN: *Incremental Natural Language Description of Dynamic Imagery*. In: FREKSA, C. und W. BRAUER (Hrsg.): *Wissensbasierte Systeme. 3. Int. GI-Kongreß*, S. 153–162. Springer, Berlin, Heidelberg, 1989.

[19] HERZOG, G. und P. WAZINSKI: *VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions*. Artificial Intelligence Review, 8(2/3):175–187, 1994.

[20] HILDEBRANDT, B., R. MORATZ, G. RICKHEIT und G. SAGERER: *Integration von Bild- und Sprachverstehen in einer kognitiven Architektur*. Kognitionswissenschaft, 4(3):118–128, 1995.

[21] HUSSMANN, M. und P. SCHEFE: *The Design of SWYSS, a Dialogue System for Scene Analysis*. In: BOLC, L. (Hrsg.): *Natural Language Communication with Pictorial Information Systems*, S. 143–201. Hanser/McMillan, München, 1984.

[22] JAMESON, A. und W. WAHLSTER: *User Modelling in Anaphora Generation*. In: *Proc. of the 5th ECAI*, S. 222–227, Orsay, France, 1982.

[23] KILGER, A.: *Using UTAGs for Incremental and Parallel Generation*. Computational Intelligence, 10(4):591–603, 1994.

[24] KOLLER, D.: *Detektion, Verfolgung und Klassifikation bewegter Objekte in monokularen Bildfolgen am Beispiel von Straßenverkehrsszenen*. Infix, St. Augustin, 1992.

[25] KOLLNIG, H. W.: *Ermittlung von Verkehrsgeschehen durch Bildfolgenauswertung*. Infix, St. Augustin, 1995.

[26] LANDAU, B. und R. JACKENDOFF: *"What" and "Where" in Spatial Language and Spatial Cognition*. Behavioral and Brain Sciences, 16:217–265, 1993.

[27] LÄNGLE, T., T. C. LÜTH, G. HERZOG, E. STOPP und G. KAMSTRUP: *KANTRA - A Natural Language Interface for Intelligent Robots*. In: REMBOLD, U., R. DILLMAN, L. O. HERTZBERGER und T. KANADE (Hrsg.): *Intelligent Autonomous Systems (IAS 4)*, S. 357–364. IOS, Amsterdam, 1995.

[28] MAASS, W.: *From Visual Perception to Multimodal Communication: Incremental Route Descriptions*. Artificial Intelligence Review, 8(2/3):159–174, 1994.

[29] MCKEVITT, P. (Hrsg.): *Proc. of AAAI-94 Workshop on Integration of Natural Language and Vision Processing*, Seattle, WA, 1994.

[30] NAGEL, H.-H.: *From Image Sequences Towards Conceptual Descriptions*. Image and Vision Computing, 6(2):59–74, 1988.

[31] NEUMANN, B. und H.-J. NOVAK: *NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenen*. Informatik Forschung und Entwicklung, 1:83–92, 1986.

[32] RETZ-SCHMIDT, G.: *Various Views on Spatial Prepositions*. AI Magazine, 9(2):95–105, 1988.

[33] RETZ-SCHMIDT, G.: *Recognizing Intentions, Interactions, and Causes of Plan Failures*. User Modeling and User-Adapted Interaction, 1:173–202, 1991.

[34] RETZ-SCHMIDT, G.: *Die Interpretation des Verhaltens mehrerer Akteure in Szenenfolgen*. Springer, Berlin, Heidelberg, 1992.

[35] ROHR, K.: *Towards Model-based Recognition of Human Movements in Image Sequences*. Computer Vision, Graphics, and Image Processing (CVGIP): Image Understanding, 59(1):94–115, 1994.

[36] SCHIRRA, J. R. J.: *Bildbeschreibung als Verbindung von visuellem und sprachlichem Raum: Eine interdisziplinäre Untersuchung von Bildvorstellungen in einem Hörermodell*. Infix, St. Augustin, 1994.

[37] SCHIRRA, J. R. J., G. BOSCH, C.-K. SUNG und G. ZIMMERMANN: *From Image Sequences to Natural Language: A First Step Towards Automatic Perception and Description of Motions*. Applied Artificial Intelligence, 1:287–305, 1987.

[38] SCHIRRA, J. R. J. und E. STOPP: *ANTLIMA — A Listener Model with Mental Images*. In: *Proc. of the 13th IJCAI*, S. 175–180, Chambery, France, 1993.

[39] STOPP, E., K.-P. GAPP, G. HERZOG, T. LÄNGLE und T. C. LÜTH: *Utilizing Spatial Relations for Natural Language Access to an Autonomous Mobile Robot*. In: NEBEL, B. und L. DRESCHLER-FISCHER (Hrsg.): *KI-94: Advances in Artificial Intelligence. 18th German Annual Conference on Artificial Intelligence*, S. 39–50. Springer, Berlin, Heidelberg, 1994.

[40] SUNG, C.-K. und G. ZIMMERMANN: *Detektion und Verfolgung mehrerer Objekte in Bildfolgen*. In: HARTMANN, G. (Hrsg.): *Mustererkennung 1986*, S. 181–184. Springer, Berlin, Heidelberg, 1986.

[41] TALMY, L.: *How Language Structures Space*. In: PICK, H. und L. ACREDOLO (Hrsg.): *Spatial Orientation: Theory, Research and Application*, S. 225–282. Plenum, New York, London, 1983.

[42] WAHLSTER, W.: *One Word Says More Than a Thousand Pictures. On the Automatic Verbalization of the Results of Image Sequence Analysis Systems*. Computers and Artificial Intelligence, 8:479–492, 1989.

[43] WAHLSTER, W.: *Text and Images*. In: COLE, R. A., J. MARIANI, H. USZKOREIT, A. ZAENEN und V. ZUE (Hrsg.): *Survey on Speech and Natural Language Technology*. Kluwer, Dordrecht, 1995.

[44] WAHLSTER, W., H. MARBURGER, A. JAMESON und S. BUSEMANN: *Over-answering Yes-No Questions: Extended Responses in a NL Interface to a Vision System*. In: *Proc. of the 8th IJCAI*, S. 643–646, Karlsruhe, 1983.