

Sprache und Raum: Natürlichsprachlicher Zugang zu visuellen Daten

Gerd Herzog,* Elisabeth André, Thomas Rist

SFB 314, Project VITRA, Universität des Saarlandes
D-66041 Saarbrücken, Germany
German Research Center for Artificial Intelligence (DFKI)
D-66123 Saarbrücken, Germany

Zusammenfassung

Der vorliegende Beitrag befaßt sich mit der Repräsentation und Verarbeitung räumlich-zeitlichen Wissens im Hinblick auf den natürlichsprachlichen Zugang zu visuellen Daten. Im Kontext des Projektes *Vitra*, das sich mit der Entwicklung wissensbasierter Systeme zur Integration von maschinellem Sehen und der Verarbeitung natürlicher Sprache auseinandersetzt, soll diese Problemstellung näher diskutiert werden. Zunächst wird am Beispiel des Bildfolgenanalyse-Systems *Actions* demonstriert, auf welche Weise Trajektorien bewegter Objekte aus Realweltbildfolgen extrahiert werden können. Daran anschließend wird die in *Vitra* zugrundegelegte Schnittstelle zur Bildanalyse, die geometrische Szenenbeschreibung, vorgestellt. Am Beispiel der lokalen orientierungsabhängigen Relationen *rechts*, *links*, *vor* und *hinter* wird dann gezeigt, wie sich eine rein geometrische Darstellung in eine propositionale Beschreibung der räumlichen Anordnung überführen läßt. Schließlich geht es um die Interpretation und Repräsentation von Ereignissen, d.h. räumlich-zeitlichen Konzepten. Im Gegensatz zu bisherigen Ansätzen soll dabei jedoch nicht von einer bereits vollständig analysierten Bildfolge ausgegangen werden; vielmehr sind Ereignisse so zu repräsentieren, daß sie simultan zu ihrem Auftreten in der Szene detektiert und natürlichsprachlich beschrieben werden können.

Dieser Beitrag ist erschienen in: C. Freksa und C. Habel (Hrsg.), Repräsentation und Verarbeitung räumlichen Wissens, pp. 207–220. Berlin, Heidelberg: Springer, 1990.

1 Motivation

Die Bedeutung, die der geeigneten Repräsentation und Verarbeitung räumlich-zeitlichen Wissens im Bereich der künstlichen Intelligenz zukommt, erklärt sich aus der besonderen Rolle von Raum und Zeit als zentrale Konzepte in der `realen' Welt.¹ Die sprachorientierte KI-Forschung setzt sich mit räumlich-zeitlichen Konzepten aus zwei unterschiedlichen Perspektiven auseinander.

Zum einen geht es bei dem Verstehen von Text um die Zuordnung von natürlichsprachlichen Äußerungen zu entsprechenden visuellen Vorstellungen (vgl. hierzu u.a. Waltz und Boggess (1979), Adorni et al. (1983), Mohnhaupt (1987) und Pribbenow (1988)). Die besondere Schwierigkeit liegt hierbei in der prinzipiellen Mehrdeutigkeit dieser Abbildung. Die Vagheit von Sprache in bezug auf exakte Geometrie und genauen zeitlichen Verlauf bedingt, daß einer sprachlichen Beschreibung im allgemeinen unendlich viele räumliche Konstellationen bzw. räumlich-zeitliche Vorgänge zugeordnet werden können.

Zum anderen wird die duale Fragestellung, der natürlichsprachliche Zugang zu visuellen Daten, untersucht, die in dieser Arbeit im Vordergrund stehen wird. Der Rückgriff auf visuelle Information erlaubt, im Sinne einer Referenzsemantik, eine in der Perzeption verankerte Definition der Bedeutung sprachlicher Begriffe. Kennzeichnend ist, daß dabei von exakten räumlichen Beschreibungen, d.h. geometrischen Repräsentationen, ausgegangen wird (vgl. u.a. Fürnsinn et al. (1984), Hußmann und Scheffe (1984), Carsten und Janson (1985) und Bajcsy et al. (1985)). Ansätze, mit denen auch räumlich-zeitliche Konzepte behandelt werden sollen (vgl. u.a. Badler (1975), Okada (1979), [Wahlster et al. 83] und Neumann und Novak (1986)), setzen weiterhin voraus, daß der genaue zeitliche Verlauf einer Szene bekannt ist. Es ergibt sich hier eine enge Verzahnung mit einem anderen Kerngebiet der künstlichen Intelligenz, dem Bildverstehen, in dem traditionell untersucht wird, wie sich solche geometrischen Repräsentationen einer Szene anhand von Bildern aufbauen lassen.

Der natürlichsprachliche Zugang zu visuellen Daten bildet auch den Forschungshintergrund für das Projekt *Vitra*², das sich mit der Entwicklung wissensbasierter Systeme zur Integration von maschinellem Sehen und der Verarbeitung natürlicher Sprache auseinandersetzt. Die Untersuchungen in *Vitra* konzentrieren sich dabei zur Zeit hauptsächlich auf zwei Diskursbereiche: Das System *Citytour* (vgl. André et al. (1987) und Schirra et al. (1987)) leistet die Beantwortung natürlichsprachlicher Anfragen über räumliche Relationen und abgeschlossene Bewegungsverläufe in einer Straßenverkehrsszene. Im System *Soccer* (vgl. André et al. (1988)) liegt der Schwerpunkt auf der simultanen Berichterstattung über beobachtbare Ereignisse während des Ablaufs einer Bildsequenz (kurze Ausschnitte aus Fußballübertragungen). Im Kontext von *Vitra* sollen in diesem Beitrag die folgenden Punkte konkretisiert werden:

¹Eine gute Darstellung dieses Zusammenhangs und der vielfältigen Einsatzbereiche räumlichen Wissens findet sich in Habel (1988).

²Visual TRANslator

- Zusammenspiel von Bildfolgenanalyse und natürlichsprachlichem Zugangssystem

Ein wichtiges Ziel in *Vitra* ist die Kopplung mit einem Bildfolgenanalyzesystem. Daher wird im nachfolgenden Abschnitt zunächst an einem konkreten System gezeigt, wie sich Trajektorien bewegter Objekte aus einer Realweltbildfolge extrahieren lassen, um daran anschließend den in *Vitra* verfolgten Ansatz für die Schnittstelle zur Bildanalyse vorzustellen.

- Semantik räumlicher Präpositionen

Ein erster möglicher Schritt für einen weitergehenden Interpretationsprozeß besteht in der Analyse räumlicher Beziehungen zwischen den Szenenobjekten, d.h. in der Überführung einer rein geometrischen Darstellung in eine explizite Beschreibung der räumlichen Anordnung. In bezug auf die sprachliche Beschreibung einer Szene heißt das, eine Referenzsemantik für räumliche Präpositionen zu definieren.

- Semantik von Bewegungs- und Handlungsverben

Unter zusätzlicher Berücksichtigung des zeitlichen Verlaufs bietet sich die Möglichkeit, auch Bewegungskonzepte bzw. einfache Handlungskonzepte aus der geometrischen Beschreibung der sichtbaren Objekte und ihrer Trajektorien zu abstrahieren. Im Gegensatz zu anderen Ansätzen soll hier jedoch nicht von einer bereits vollständig analysierten Bildfolge ausgegangen werden. Es geht vielmehr darum zeitübergreifende Vorgänge so zu repräsentieren, daß sie simultan zu ihrem Auftreten in der Szene detektiert und natürlichsprachlich beschrieben werden können.

2 Bildfolgenanalyse

Die Aufgabe der Bildanalyse ist es, anhand von Bildern eine symbolische computerinterne Beschreibung einer Szene zu erzeugen. Bei der Bildfolgenanalyse steht hierbei insbesondere die Analyse und Interpretation bewegungsbedingter Änderungen im Vordergrund. Am Beispiel des für die Kopplung mit *Vitra* eingesetzten *Actions* Systems wird im nachfolgenden Abschnitt gezeigt, auf welche Weise Information über bewegte Objekte aus einer Folge digitalisierter Bilder abgeleitet werden kann. Hieran anschließend wird die Schnittstelle zwischen Bildfolgenanalyse und *Vitra*, d.h. weitergehender Szeneninterpretation, motiviert und vorgestellt.

2.1 Das Bildfolgenanalyzesystem ACTIONS

Das am Institut für Informations- und Datenverarbeitung (IITB) der Fraunhofergesellschaft, Karlsruhe, entwickelte *Actions*³ System leistet die automatische Detektion und Verfolgung von Objekten in Bildfolgen. Kernziel der Arbeit an *Actions* war und ist es, robuste, allgemein anwendbare Methoden zur Analyse von Realweltszenen zu entwickeln. Eine zusammenfassende Darstellung der bisherigen Untersuchungsergebnisse und des daraus entwickelten Verfahrens findet sich in Sung und Zimmermann (1986) und Sung (1988).

Abb. 1 bietet einen Überblick zu den einzelnen Verarbeitungsschritten. Bei dem betrachteten Bildmaterial handelt es sich um Aufnahmen von einer Straßenkreuzung, die von einem ca. 35m hohen Gebäude aus gemacht wurden, sowie um Aufnahmen von einer Begegnung aus der Fußballbundesliga. Das Material wurde jeweils mit einer stationären, monokularen Kamera aufgenommen. Für die Analyse werden zur Zeit bis zu 132 Sekunden dauernde Ausschnitte von 3300 Vollbildern mit $512 * 512$ Bildpunkten zu je 8 Bit Grauwertaufösung verwandt.

Bewegte Objekte werden durch die Berechnung und Analyse von Verschiebungsvektorfeldern erkannt. Die Berechnung der Verschiebungsvektoren basiert dabei auf der Zuordnung von charakteristischen lokalen Grauwertverteilungen, sogenannten *Merkmalen*. Zur Bestimmung der Merkmale wird der *Monotonie-Operator* verwandt (vgl. Zimmermann und Kories (1984)), der besonders stabile, d.h. leicht wiederauffindbare, Merkmale liefert. Dabei wird jedes Pixel mit einer festen Anzahl umliegender Pixel verglichen. Je nach Anzahl der Vergleichspunkte, die ein bestimmtes Kriterium erfüllen, wird das betrachtete Pixel einer entsprechenden Klasse zugeteilt. Im konkreten Fall werden 8 Vergleichspunkte und die Relation *kleiner bzgl. des Grauwertes* als Kriterium verwandt. Jedes Pixel läßt sich damit eindeutig einer der 9 möglichen Klassen zuordnen. Zusammenhängende Punkte aus derselben Klasse können zu Flecken zusammengefaßt werden. Im weiteren werden dann nur noch Flecken aus einer der beiden extremen Klassen, d.h. Kuppen und Senken des Grauwertgebirges, berücksichtigt.

Um bei der Ermittlung von Verschiebungsvektoren Fehler durch zufällige Schwankungen der Merkmalspositionen zu verringern, werden die Schwerpunkte der Flecken über mehrere Bilder hinweg verfolgt. Hieraus ergeben sich lokale Verschiebungsvektoren vom n -ten zum $(n + i)$ -ten Bild. Die Werte 4 bzw. 5 für die Variable i haben sich dabei am besten bewährt. Die so gewonnenen Verschiebungsvektorfelder werden hinsichtlich Betrag, Richtung und Bildposition auf Ballungen ähnlicher Vektoren untersucht. Die Ballungen werden jeweils durch ein entsprechend dem mittleren Bewegungsvektor ausgerichtetes umschreibendes Rechteck markiert. Diese *Rahmen* können als Kandidaten für das Abbild bewegter starrer Objekte gelten. Der geometrische Mittelpunkt eines Rahmens dient als Repräsentant des bewegten Objektes in der Bildebene.

Die Korrespondenzlinien der Rahmenpositionen entsprechen den Trajektorien der Vektorballungen und somit der Objektkandidaten in der Bildebene. Zwei Rahmen aus

³Automatic Cueing and Trajectory estimation in Imagery of Objects in Natural Scenes

ACTIONS : "Fußball"

Automatic Cueing and Trajectory estimation in Imagery of Objects in natural Scenes

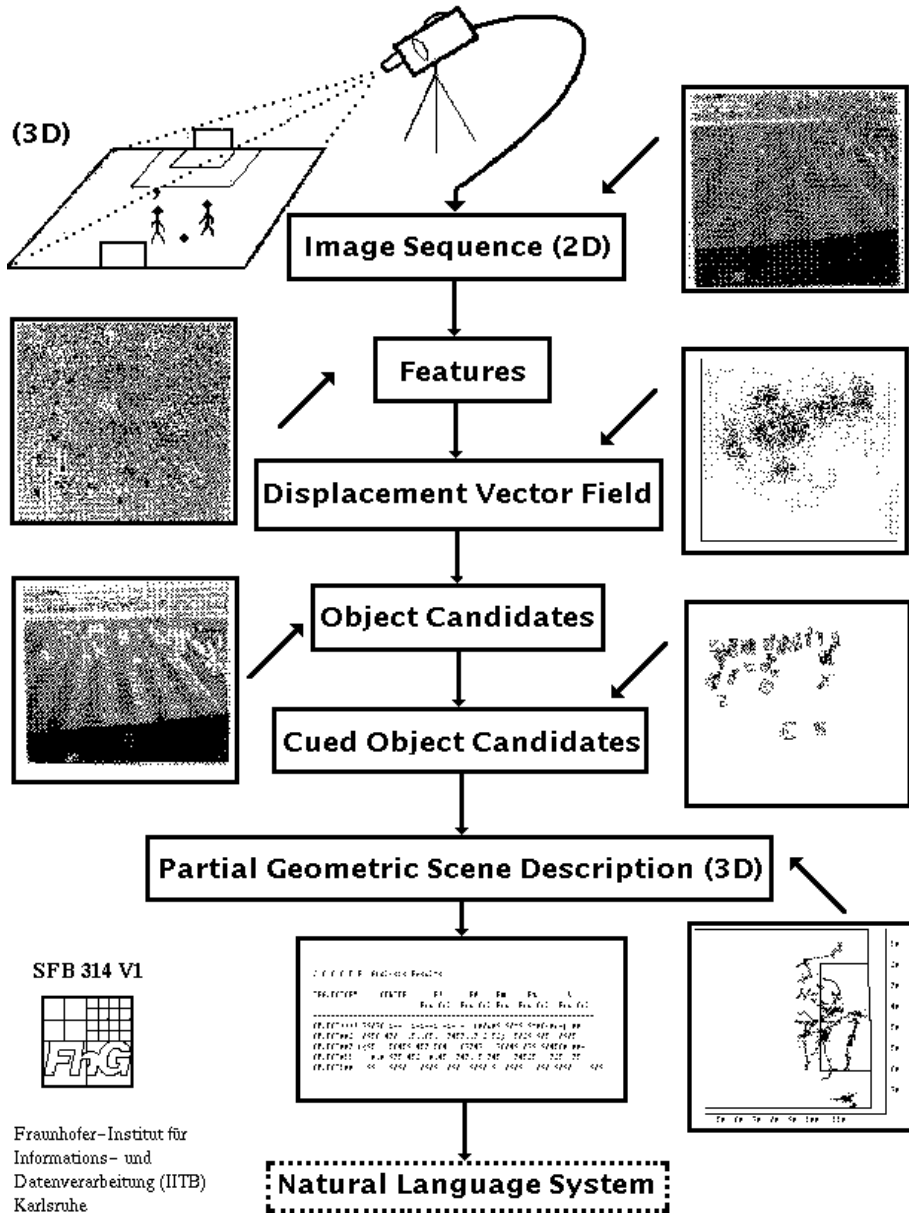


Abbildung 1: Verarbeitungsschritte in *Actions*

aufeinanderfolgenden Aufnahmen werden einander zugeordnet, falls ihre Bewegungsrichtung annähernd übereinstimmt und ihr räumlicher Abstand kleiner ist als der doppelte Betrag des mittleren Verschiebungsvektors. Bei noch verbleibender Mehrdeutigkeit wird der Zuordnungskandidat mit dem kleinsten räumlichen Abstand gewählt. Die Koordinatenangaben der so gewonnenen Korrespondenzlinien werden unter Berücksichtigung von Kameraposition, Aufnahmewinkel und der Geometrie des statischen Szenenhintergrundes in Szenenkoordinaten rücktransformiert und in der (eingeschränkten) geometrischen Szenenbeschreibung zusammengefaßt. Die Klassifikation von Objektkandidaten und deren Zuordnung zu a priori bekannten Objekten kann derzeit von *Actions* noch nicht geleistet werden und erfolgt daher interaktiv.

2.2 Schnittstelle zwischen Bildfolgenanalyse und Interpretation

Die *geometrische Szenenbeschreibung* (GSB) wurde in Neumann (1984) als Repräsentation für die Ausgabe des Bildanalyseprozesses eingeführt. Ziel dieser Repräsentation ist es, die ursprüngliche Bildfolge prinzipiell vollständig und ohne Informationsverlust darzustellen. Das bedeutet, daß die mit der Kamera aufgenommene Bildfolge (im Prinzip) aus der GSB rekonstruiert werden könnte. Die geometrische Szenenbeschreibung enthält:

- Für jedes Einzelbild der Bildfolge:
 - Zeitpunkt
 - Betrachterstandpunkt
 - alle in der Szene sichtbaren Objekte
 - Beleuchtungsdaten
- für jedes Objekt:
 - Identität (über die Bildfolge)
 - 3D-Position in Weltkoordinaten für jedes Einzelbild
 - Orientierung für jedes Einzelbild
 - 3D-Form
 - Physikalische Oberflächeneigenschaften (Farbe)
 - Klassenzugehörigkeit
 - Identität (in bezug auf mögliches Vorwissen) und somit weitere verbalisierbare Eigenschaften wie z.B. Namen

Hierbei ist zu beachten, daß die Klassifizierung und die Identifizierung bereits vorher bekannter Objekte nur mit Hilfe zusätzlicher Wissensquellen möglich ist und nicht allein aus der reinen Bildinformation gewonnen werden kann.

Das Konzept der geometrischen Szenenbeschreibung stellt eine idealisierte Schnittstelle zwischen der Bildanalyse und einem darauf aufsetzenden natürlichsprachlichen System dar. In Anwendungen, wie z.B. dem von Neumann und Novak entwickelten *Naos* System zur natürlichsprachlichen Beschreibung von Straßenverkehrsszenen, wird die GSB jedoch entsprechend den konkreten Anforderungen eingeschränkt. In den *Vitra* Systemen werden z.B. Betrachterstandpunkt, Beleuchtungsdaten sowie die vollständige 3D-Form der Objekte nicht berücksichtigt. Die Information über den statischen Szenenhintergrund wird nicht von der Bildanalyse geliefert, sondern liegt vielmehr als instantiiertes Modell des betrachteten Weltausschnitts vor. Derartige Einschränkungen finden sich derzeit in allen Ansätzen, da man von einem universell einsetzbaren KI-System, das beliebige Bildfolgen vollständig analysieren kann, noch sehr weit entfernt ist.

Bei der Berechnung der Anwendbarkeit räumlicher Präpositionen und Bewegungs-
verben wird man auch aus Effizienzgründen nicht von einer vollständigen Szenen-
beschreibung ausgehen, sondern auf Idealisierungen, etwa im Sinne von Herskovits
(vgl. Herskovits (1986)), zurückgreifen. Herskovits führt den Begriff *geometrische
Beschreibung* ein, um die für die Semantik räumlicher Relationen relevanten Eigen-
schaften zu repräsentieren. Formal gesehen sind geometrische Beschreibungen Funk-
tionen, die einem Objekt einen situationsabhängigen geometrischen Repräsentanten
zuordnen. Anstelle von aufwendigen 3D-Rekonstruktionen begnügt man sich typi-
scherweise mit Repräsentationen durch den Objektschwerpunkt oder der Kontur einer
Objektprojektion, angenähert durch einen Polygonzug.

Der Zugriff auf die Daten der GSB erfolgt funktional. So wird man etwa eine
Lokalisierungsfunktion definieren, mit der die 3D-Weltkoordinaten eines Objekts
oder seines idealisierten geometrischen Repräsentanten zu einem vorgegebenen Zeit-
punkt in Erfahrung gebracht werden können. Aus den Daten der GSB ableitbar sind
quantitative Meßgrößen wie Abstand zwischen Objekten, Geschwindigkeit und Be-
schleunigung von bewegten Objekten. Der Übergang von quantitativen Meßgrößen
zu Raum- und Bewegungskonzepten geschieht über Prädikate, deren Definitionen am
natürlichsprachlichen Gebrauch von räumlichen Präpositionen und Bewegungs-
verben orientiert sind.

3 Semantik räumlicher Präpositionen

Die semantische Analyse von räumlichen Präpositionen führt auf den Begriff *räumliche
Relation* als einzelsprachunabhängige Bedeutungseinheit. Man kann räumliche Rela-
tionen dadurch definieren, daß man Bedingungen über räumliche Gegebenheiten einer
Objektkonfiguration spezifiziert, wie z.B. Abstand zwischen Objekten, relative Lage
bezüglich einer Orientierung usw.; d.h. man kennzeichnet eine bestimmte Klasse von
Objektkonfigurationen. Räumliche Beziehungen zwischen Objekten lassen sich pro-
positional durch Relationentupel folgender Form repräsentieren:

$$(\text{Rel-Name } \text{Subjekt } \text{Bezugsobjekt}_1 \dots \text{Bezugsobjekt}_n \{ \text{Orientierung} \})$$

Das erste Argument bezeichnet die entsprechende räumliche Relation. Das als Subjekt benannte Argument steht für dasjenige Objekt, das relativ zu einem oder mehreren Objekten, den Bezugsobjekten, (bezüglich einer Orientierung) lokalisiert werden soll.

3.1 Anwendbarkeit und Anwendbarkeitsraum

Wir sprechen von der *Anwendbarkeit eines Relationentupels*, falls es sich zur Charakterisierung einer Objektkonstellation eignet. Zur Bestimmung der Anwendbarkeit erweist sich die Verwendung von *Anwendbarkeitsräumen* als hilfreich. Der grundlegende Gedanke ist dabei der, daß man jedem Relationentupel einen Anwendbarkeitsraum zuordnet und dann prüft, in welcher mengentheoretischen Beziehung (Inklusion, Exklusion oder Überlappung) sich der vom Subjekt eingenommene Raum zu dem durch die restlichen Argumente des Relationentupels bestimmten Anwendbarkeitsraums befindet. Bei der zweistelligen Relation in wäre etwa zu untersuchen, ob der vom Subjekt eingenommene Raum ein Teilraum des Anwendbarkeitsraums ist, der in diesem Fall gerade mit dem Innenraum des Bezugsobjekts zusammenfällt. Die Bestimmung der Anwendbarkeitsräume ist im allgemeinen nicht trivial, weil u.a. Orientierung, Ausdehnung und Form der Bezugsobjekte zu berücksichtigen sind. Weitere Schwierigkeiten ergeben sich aufgrund benachbarter Objekte, die, anschaulich gesprochen, zu einer *Deformierung* des Anwendbarkeitsraumes führen können.

Die Unterscheidung zwischen Anwendbarkeit und Nichtanwendbarkeit eines sprachlichen Ausdrucks reicht nicht aus, um eine räumliche Situation adäquat zu beschreiben. Vielmehr müssen die Grenzen als fließend angesehen werden. Um dieser Tatsache Rechnung zu tragen, wird jedem Relationentupel ein *Anwendbarkeitsgrad* zugeordnet. In Hanßmann (1980) wird z.B. ein Wert zwischen 0 und 1 vorgeschlagen, wobei 0 für nicht anwendbar und 1 für voll anwendbar steht. Veranschaulichen läßt sich graduierte Anwendbarkeit durch Partitionierung eines Anwendbarkeitsraums in Regionen gleicher Anwendbarkeit, mit denen dann verschiedene *linguistische Hecken* (vgl. Lakoff (1973)) wie z.B. *'unmittelbar'* oder *'genau'* assoziiert werden können.

Im folgenden wird demonstriert, wie in den Systemen *Citytour* und *Soccer* der Anwendbarkeitsgrad der orientierungsabhängigen Relationen *rechts*, *links*, *vor* und *hinter* im zweidimensionalen Raum berechnet wird.

3.2 Berechnung der Anwendbarkeit im Falle orientierungsabhängiger Relationen

Bei der Analyse von Richtungspräpositionen wird deutlich, daß die zu betrachtende Orientierung des Raumes von der Gebrauchsart der Präposition abhängt. Im folgenden wird die in Wunderlich (1985) verwendete Unterscheidung zwischen *intrinsischem* und *extrinsischem* Gebrauch von Präpositionen auf die korrespondierenden Relationen übertragen. In diesem Sinne wird eine Relation *intrinsisch* gebraucht, wenn die Orientierung durch die inhärente Organisation des Bezugsobjekts festgelegt

wird. Faktoren, die inhärente Seiten eines Objekts festlegen, werden in Miller und Johnson-Laird (1976), Sondheimer (1976), Vandeloise (1984) und Wunderlich (1985) diskutiert. Genannt werden u.a. die Standardbewegungsrichtung (z.B. bei Fahrzeugen), Lage der Wahrnehmungsorgane (bei Menschen oder Tieren) sowie funktionale Eigenschaften (z.B. bei Sitzmöbel). In der *Soccer*-Domäne sind beispielsweise die beiden Tore inhärent organisiert, nicht aber der Ball. Hat ein Objekt keine inhärenten Seiten, muß durch den Kontext eine Orientierung induziert werden. In diesem Fall spricht man vom *extrinsischen Gebrauch* einer Relation. Konflikte treten auf, wenn ein Objekt inhärente Seiten hat, gleichzeitig aber kontextuell eine Orientierung induziert wird, wie beispielsweise bei einem rückwärtsfahrenden Auto. In Sätzen wie “*Der Linksaußen steht vor dem Libero vom Elfmeterpunkt aus gesehen*” wird die Orientierung kontextuell durch die Position eines, möglicherweise virtuellen, Beobachters vorgegeben. Stimmt der Beobachter mit dem Sprecher (oder Hörer) überein, dann spricht man vom *deiktischen Gebrauch* einer Relation.

Die Berechnung der Anwendbarkeit einer orientierungsabhängigen Relation läßt sich in folgende Schritte unterteilen:

1. Bestimmung der Orientierung
2. Berechnung der Anwendbarkeit des Relationentupels

Eine Orientierung des zweidimensionalen Raums läßt sich durch Angabe zweier zueinander orthogonaler, jeweils vom Nullvektor verschiedener Vektoren a_{lr} und a_{hv} festlegen (hierbei steht a_{lr} für Links-Rechts-Achse und a_{hv} für Hinten-Vorne-Achse). Die Richtung der Hinten-Vorne-Achse wurde in den Systemen *Citytour* und *Soccer* wie folgt festgelegt:

- Wird die Orientierung durch die inhärente Vorderseite des Bezugsobjekts vorgegeben, dann ist a_{hv} orthogonal zur Frontseite und zeigt aus dem Objekt heraus.
- Wird ein Objekt in bezug auf einen Beobachterstandpunkt lokalisiert, dann wird die Orientierung des Beobachters auf das Bezugsobjekt übertragen (vgl. Abb. 2 (a)), falls Bezugsobjekt und Beobachter koinzidieren. Sind Beobachter und Bezugsobjekt örtlich getrennt, dann ergibt sich die Orientierung nach dem Spiegelbildprinzip (vgl. Abb. 2 (b)).
- Wird durch die Bewegung eines Objekts eine Orientierung induziert, dann stimmt a_{hv} mit der Bewegungsrichtung des Objekts überein.

Die Links-Rechts-Achse wird so bestimmt, daß sie mit der zuvor festgelegten Hinten-Vorne-Achse ein orthogonales Rechtssystem bildet, sofern das Koinzidenzprinzip angewandt wurde. In allen anderen Fällen bilden Hinten-Vorne-Achse und Links-Rechts-Achse ein orthogonales Rechtssystem. Ob eine Orientierungsrelation intrinsisch oder extrinsisch gebraucht wird, hängt im wesentlichen vom Bezugsobjekt und dem sprachlichen Kontext ab. Die hierzu in *Soccer* verwendete Strategie wird in André (1988) beschrieben.

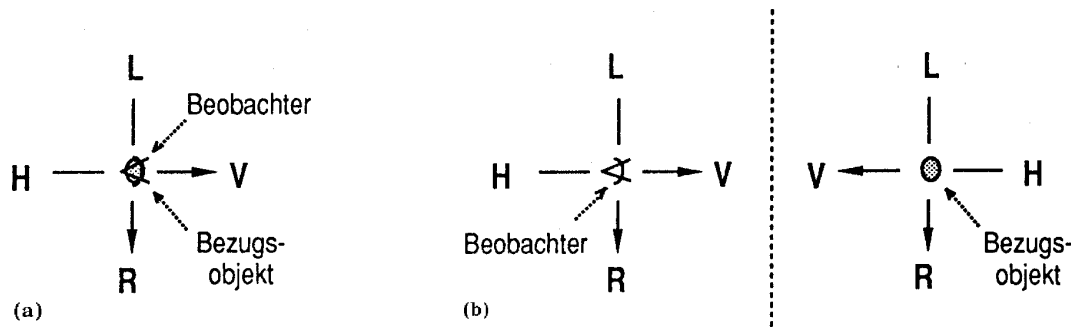


Abbildung 2: Wahl des Referenzsystems (a) nach dem Koinzidenzprinzip und (b) nach dem Spiegelbildprinzip

Zur Berechnung der Anwendbarkeit von orientierungsabhängigen Relationen bei vorgegebener Links-Rechts- bzw. Hinten-Vorne-Achse wurde in Herskovits (1980) für den Fall, daß sowohl Subjekt als auch Bezugsobjekt punktförmig sind, folgendes Verfahren vorgeschlagen:

1. Konstruiere ein Bezugssystem, dessen Ursprung mit der Lokation des Bezugsobjekts übereinstimmt und dessen Abszisse die Links-Rechts-Achse und dessen Ordinate die Hinten-Vorne-Achse ist (vgl. Abb. reffig:sys (a)).
2. Bestimme die Lokation des Subjektes relativ zu dem in Schritt 1 konstruierten Bezugssystem.

In André et al. (1987)) wird eine Erweiterung dieses Verfahrens beschrieben, die für das System *Citytour* entwickelt wurde. Zum einen können dort polygonal repräsentierte Bezugsobjekte herangezogen werden, zum anderen wird die Anwendbarkeit einer Relation graduiert. Beide Verfahren, die jeweils nur für den zweidimensionalen Raum entwickelt wurden, teilen den Raum um das Bezugsobjekt in Halbebenen auf, mit denen eine der Relationen *vorne*, *hinten*, *rechts* und *links* assoziiert wird. Bei einem nicht punktförmigen Bezugsobjekt wird bei dem erweiterten Verfahren ein umschreibendes Rechteck bestimmt, das dann als ausgedehnter Ursprung dient. Die Achsen des in Abb. 3 (b) dargestellten Bezugssystems sind zu Bändern entartet. Als Rechteck wurde das kleinste, den das Bezugsobjekt repräsentierenden Polygonzug umschreibende Rechteck gewählt, das achsenparallel zu den Basisvektoren a_{hv} und a_{lr} orientiert ist. Während im Bezugssystem bei Herskovits eine der vier Relationen nur dann anwendbar ist, falls das Subjekt genau auf einem der vier Achsenabschnitte lokalisiert wird, ist in *Citytour* und *Soccer* eine Orientierungsrelation immer dann anwendbar, falls sich das Subjekt innerhalb einer mit der Relation assoziierten Halbebene befindet. Der Anwendbarkeitsraum sowie eine Zerlegung in Regionen gleicher Anwendbarkeit wird durch Hinzunahme weiterer Bedingungen festgelegt. Zum einen wird eine von der Ausdehnung des Bezugsobjekts abhängige

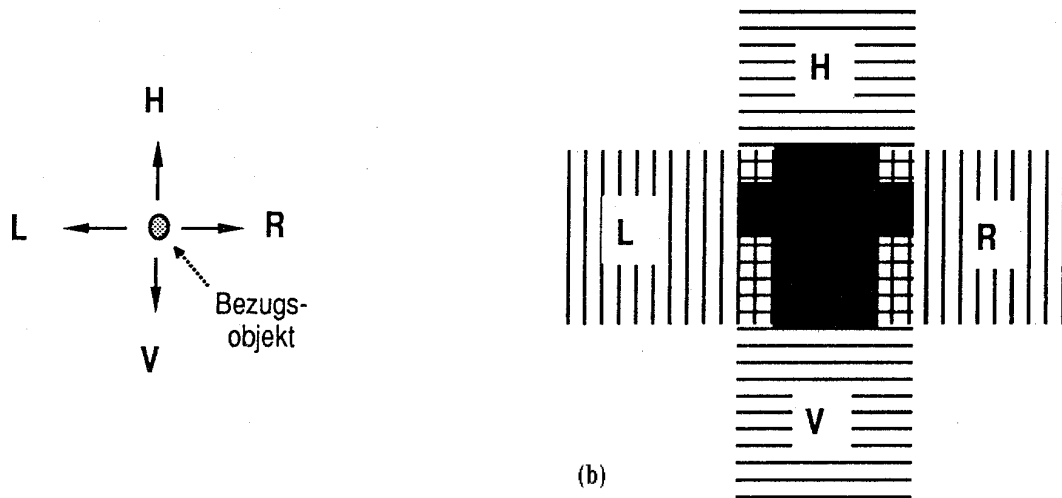


Abbildung 3: Bezugssystem bei (a) Herskovits und (b) in *Citytour*

Skalierung des Koordinatensystems zugrunde gelegt, zum anderen wird die Lage des Subjekts in diesem Koordinatensystem genauer berücksichtigt (durch die Bewertung des Abstandes zwischen Subjekt und Koordinatenursprung bzw. zwischen Subjekt und den angrenzenden Halbachsen). In Abb. 4 liegt Ob2 auf der zu einem Band entarteten

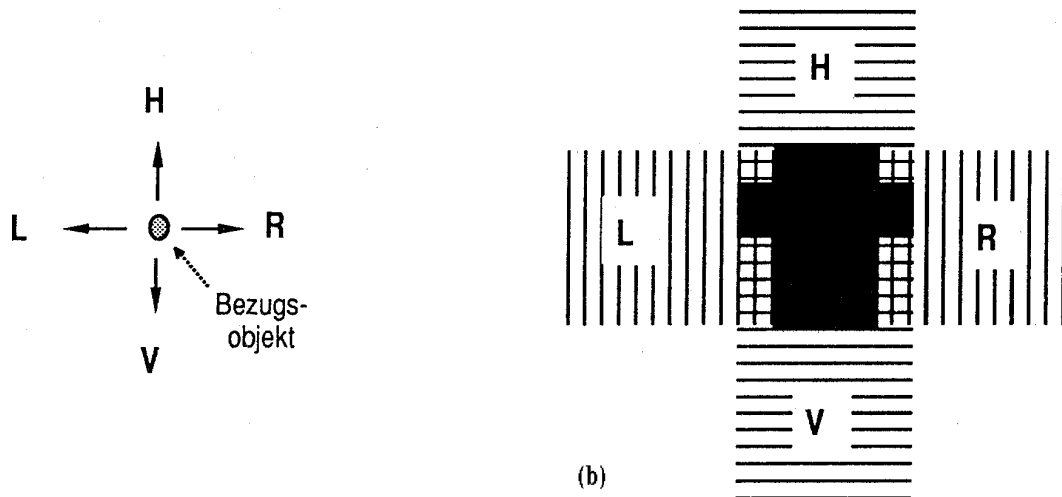


Abbildung 4: Anwendbarkeitsgrade lokaler Orientierungsrelationen

Rechts-Achse hinsichtlich des im Bezugsobjekt Ob1 aufgespannten Bezugssystems, so daß die Anwendbarkeit der Relation (rechts Ob2 Ob1 $\{a_{hv} a_{lr}\}$) entsprechend hoch ist. Einen geringeren Grad der Anwendbarkeit ergibt sich für das Relationentupel (rechts Ob3 Ob1 $\{a_{hv} a_{lr}\}$). Für die Objekte Ob4 und Ob5 ist die

Relation `rechts` schließlich gar nicht mehr anwendbar.

4 Semantik von Bewegungs- und Handlungsverben

Durch den Übergang von der Beschreibung statischer Szenen zur Analyse von Kamerabildfolgen und die Berücksichtigung temporaler Aspekte bietet sich die Möglichkeit neben räumlichen auch zeitabhängige Konzepte aus der von einem Bildfolgenanalyzesystem erzeugten geometrischen Beschreibung der sichtbaren Objekte und ihrer Trajektorien zu abstrahieren. Solche konzeptuellen Einheiten, die hier als Ereignisse bezeichnet werden, dienen zur Erfassung des Geschehens in einer zeitveränderlichen Szene. Im Hinblick auf die natürlichsprachliche Beschreibung einer Bildfolge stehen Ereignisse, in Analogie zum Begriff der räumlichen Relation, als einzelsprachunabhängige Bedeutungseinheiten zur Verfügung. Sie dienen dazu, die Semantik von Bewegungs- und Handlungsverben zu definieren. Ereignisse sind hier also diejenigen wahrnehmbaren Veränderungen der Welt, über die man üblicherweise spricht (vgl. Miller und Johnson-Laird (1976)).

4.1 Simultane versus retrospektive Interpretation

Neben der Frage, welche Konzepte aus einer geometrischen Szenenbeschreibung zu extrahieren sind, ist es entscheidend, wie der Erkennungsprozeß realisiert wird. Die Betrachtung der relevanten Forschungsarbeiten zeigt, daß bisher nur Ansätze verfolgt wurden, die eine retrospektive Beschreibung der zu analysierenden zeitveränderlichen Szene zum Ziel haben. Die Systeme *Naos* (vgl. Neumann und Novak (1986)) und *Epex* (vgl. Walter (1987)) bedienen sich zur Interpretation von Straßenverkehrsszenen einer *A-posteriori-Strategie*, bei der vor Analysebeginn eine vollständige geometrische Szenenbeschreibung vorhanden sein muß. Im Gegensatz dazu erfolgt die Szenenanalyse in dem zur Herzkammerdiagnose eingesetzten System *Alven* (vgl. Tsotsos (1981)) datengetrieben und sukzessive. Dennoch sollen auch in diesem Fall die erkannten Ereignisse erst nach der vollständigen Analyse einer Szene zur Verfügung gestellt werden.

Eine völlig neue Problemstellung ergibt sich, wenn es darum geht, Ereignisse simultan zu ihrem Auftreten in der Szene zu erkennen. Sind Ereignisse zu betrachten, deren Auftreten sich über vergleichsweise längere Zeiträume erstreckt, so stellt sich die Frage, wie teilweise erkannte Ereignisse zu repräsentieren sind, um sie zur weiteren Verarbeitung zur Verfügung zu stellen. Im Hinblick auf die Erzeugung simultaner Szenenbeschreibungen in natürlicher Sprache wird dieses Problem unmittelbar deutlich: Zur Zentrierung der Beschreibung auf das aktuelle Szenengeschehen ist es häufig sinnvoll, Ereignisse bereits dann zu verbalisieren, während sie ablaufen, und nicht erst dann, wenn sie vorbei sind. Beispiele hierfür wären die Beschreibung eines gerade auftretenden Überholvorgangs in einer Straßenverkehrsszene oder die Schilderung eines Angriffs in einem Fußballspiel bei einer Livereportage eines Radioreporters. Allgemein wird diese Problematik immer dann auftreten, wenn simultane Erkennung die

Grundlage weiterer Reaktionen eines bildverstehenden Systems bildet. Man denke beispielsweise an einen Roboter, der seine Umgebung visuell “*wahrnimmt*” und auf das Wahrgenommene unmittelbar reagieren muß.

Diese Problemstellung stellt besondere Anforderungen an die Modellierung von Ereignissen, denen auch zeitlogischen Formalisierungen, wie die in Allen (1984) und McDermott (1982) vorgestellten Ansätze, nicht gerecht werden, da sie formal nur vollständig auftretende Ereignisse betrachten. Um zu einer feineren Beschreibung des Auftretens eines Ereignisses zu gelangen, bietet es sich an, die verschiedenen Phasen eines Ereignisses, d.h. den Beginn, den Ablauf und das Ende des Auftretens, zu berücksichtigen. Hierzu führen wir zusätzlich die folgenden Prädikate ein:

TRIGGER ($t_i \text{ event}$) für den Beginn,

PROCEED ($t_i \text{ event}$) für das Fortschreiten,

STOP ($t_i \text{ event}$) für das Ende und

SUCCEED ($t_i \text{ event}$) für das Andauern eines auftretenden Ereignisses.

Das Prädikat dient zur Modellierung von Ereignissen, die als solche bereits vollständig erkannt sind, deren Auftreten jedoch andauert. Hierzu zählen beispielsweise die zu durativen Bewegungsverben korrespondierenden Ereigniskonzepte wie *fahren*, *laufen* oder *gehen*.

Im Gegensatz zu den intervallbasierten Prädikaten OCCUR bei Allen bzw. OCC bei McDermott beziehen sich die hier vorgestellten Prädikate auf diskrete Zeitpunkte. Sie ermöglichen es, den Zustand eines auftretenden Ereignisses zu einem vorgegebenen Zeitpunkt zu charakterisieren. Zur Veranschaulichung betrachte man einen Überholvorgang in einer Straßenverkehrsszene. Abb. 5 zeigt hierzu vier markante, aus einer Realbildfolge stammende Einzelbilder. Jedes Einzelbild korrespondiert zu je einem der diskreten Zeitpunkte T_1 bis T_4 . Zum Zeitpunkt T_1 nähert sich der Pkw dem Kleinbus. Zum Zeitpunkt T_2 schert der Pkw aus; er beginnt, den Kleinbus zu überholen. Die Tatsache, daß das Ereignis beginnt, wird formal durch **TRIGGER**(T_2 (Überholen PKW1 BUS1)) repräsentiert. Zum Zeitpunkt T_4 schert der Pkw vor dem Kleinbus ein und beendet somit den Überholvorgang. Das Ereignis (Überholen PKW1 BUS1) ist jetzt vollständig aufgetreten und es gilt: **STOP**(T_4 (Überholen PKW1 BUS1)). Zwischen T_2 und T_4 ist der Überholvorgang in der Szene zu beobachten. Das Ereignis tritt also gerade auf, ist jedoch während dieses Zeitraums noch nicht vollständig erkannt. Würde der Pkw beispielsweise zum Zeitpunkt T_3 in eine Seitenstraße abbiegen, so könnte nicht mehr von einem Überholvorgang gesprochen werden. Für noch nicht vollständig erkannte Ereignisse steht das Prädikat **PROCEED** bereit. In der im Beispiel dargestellten Situation gilt somit für alle t_i mit $T_2 < t_i < T_4$ die Prädikation **PROCEED**(t_i (Überholen PKW1 BUS1)).

Ein Beispiel für das Andauern eines Ereignisses bezüglich obiger Bildfolge wäre etwa die Tatsache, daß PKW1 fährt. Es gilt für alle t_i mit $T_1 \leq t_i \leq T_4$ die Prädikation **SUCCEED**(t_i (Fahren PKW1)).

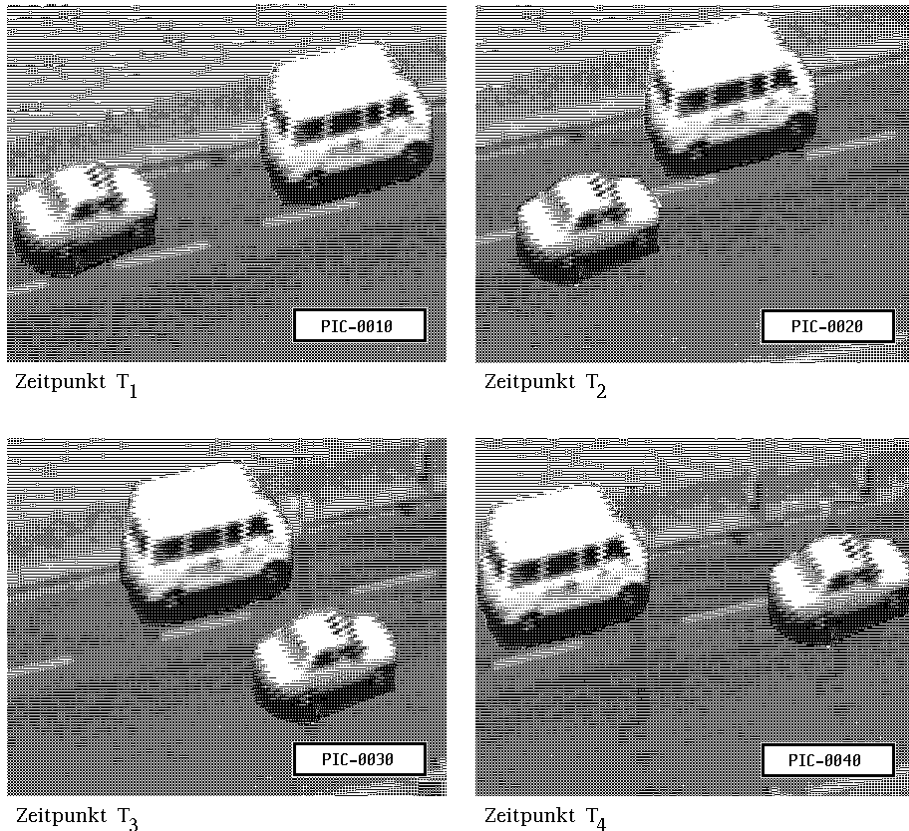


Abbildung 5: Ein Überholvorgang in einer Straßenverkehrsszene

Die soeben vorgestellten Prädikate erlauben also eine feinere Modellierung von Ereignissen. Insbesondere kann eine intervallbezogene Beschreibung von Ereignissen, wie z.B. die Tatsache, daß ein Überholvorgang über dem Zeitraum $[T_2 T_4]$ stattgefunden hat, aus dieser verfeinerten Beschreibung abgeleitet werden.

4.2 Ereignismodelle zur inkrementellen Ereigniserkennung

In Analogie zu Objektmodellen werden Ereignisse konzeptuell durch *Ereignismodelle* beschrieben (vgl. Neumann und Novak (1983)). Ereignismodelle repräsentieren A-priori-Wissen zur Erfassung des Szenengeschehens, insbesondere Wissen über interessante Bewegungsabläufe der Szenenobjekte. Sie dienen als Bindeglied zwischen den in 4.1 vorgestellten Ereignisprädikaten und denen aus einer geometrischen Szenenbeschreibung zu abstrahierenden Ereigniskonzepten. Ein wichtiges Prinzip besteht darin, die Ereignismodelle in einer aus unterschiedlichen Abstraktionsstufen bestehenden Hierarchie anzuordnen. Auf der untersten Stufe stehen dabei die direkt über den geometrischen Daten definierten Konzepte, wie z.B. *exist* oder *move*, die zur Definition komplexerer Ereignismodelle, wie beispielsweise *überholen*, herangezogen

werden. Erkannte Ereignisse sind Instantiierungen entsprechender Ereignismodelle; sie werden im folgenden auch als Ereignisinstanzen bezeichnet.

Die hier betrachtete Aufgabenstellung, die simultane Erkennung von Ereignissen, kann nur mit Hilfe einer inkrementell arbeitenden Erkennungsstrategie geleistet werden, bei der die zeitlichen Beziehungen zwischen den Subereignissen den Detektionsprozeß steuern und Ereignisse entsprechend ihrem Ablauf schrittweise erkannt und explizit in der Wissensbasis des Systems repräsentiert werden. Mit der im Rahmen des *Soccer* Systems entwickelten Methodik zur Modellierung von Ereigniskonzepten (vgl. Rist et al. (1987), Herzog und Rist (1988)) soll diesen Anforderungen Rechnung getragen werden. Ein Ereignismodell in *Soccer* umfaßt:

- Rollen

Rollen stehen als existenzquantifizierte Platzhalter für die an einem Ereignis beteiligten Objekte. In den Ereignisinstanzen sind diese Rollen mit entsprechenden Bezeichnern für konkrete Szenenobjekte gefüllt.

- Rollenrestriktionen

Rollenrestriktionen schränken die Menge der möglichen Rollenfüller bei der Instantiierung eines Ereignismodells ein. Obligatorisch sind hierbei Typrestriktionen, die angeben, welcher Objektklasse die Rollenfüller angehören müssen. Desweiteren können durch Rollenrestriktionen auch Bedingungen formuliert werden, die sich auf Abhängigkeiten zwischen den einzelnen Rollenfüllern beziehen. Solche Restriktionen sind typischerweise von der Form: 'Falls der Füller der Rolle *A* die Eigenschaft *p* besitzt, dann muß der Füller der Rolle *B* die Eigenschaft *q* besitzen'.

- Ablaufschema

Das Kernstück eines Ereignismodells ist sein Ablaufschema. Es dient dazu, den prototypischen Ablauf eines Ereignisses zu spezifizieren.

Das Ablaufschema eines Ereignismodells, formal als endlicher gerichteter markierter Graph definiert, spezifiziert die Sub-Ereignisse bzw. den situativen Kontext, der in einer Szene beobachtbar sein muß, um von einem Auftreten des entsprechenden Ereignisses sprechen zu können. Der zugrunde liegende Gedanke ist der, daß die temporalen Aspekte so repräsentiert werden, daß das Erkennen eines Ereignisses zu einer Traversierung des dazugehörigen Ablaufschemas korrespondiert. Eine solche Traversierung erfolgt dabei schrittweise innerhalb eines vorgegebenen Zeittaktes — also inkrementell.⁴ Zur Demonstration sei als Beispiel das Konzept `Pass_in_den_Lauf` gewählt.

⁴Für eine A-posteriori-Bildfolgenanalyse wird in Walter (1987) die Verwendung von ATN-Strukturen zur Modellierung von Ereignissen vorgeschlagen. Für die inkrementelle Erkennung kann dieser Ansatz nicht übernommen werden, da in der Szene gleichzeitig auftretende Ereignisse auch parallel erkannt werden müssen. Ein ATN-Interpreter, bei dem der Zugriff auf Teilkonzepte durch rekursiven Aufruf von Subnetzen erfolgt, kann aber gerade diese Aufgabe nicht leisten.

Es beschreibt die Situation, in der ein Spieler seinem laufenden Mannschaftskameraden den Ball zuspielt. Im Formalismus von Allen könnte dieses Konzept wie folgt definiert werden:

```
OCCUR(timeintervall1 (Pass_in_den_Lauf Sp1 Ball Sp2))
<==>
EXIST timeinterval2
  DURING(timeinterval2 timeintervall1)
  OCCUR(timeinterval2 (Laufen Sp2))
  OCCUR(timeintervall1 (Zuspiel Sp1 Ball Sp2))
```

Abb. 6 zeigt das entsprechende Ablaufschema, das sich durch Projektion der intervallweise angegebenen Gültigkeitsbedingungen auf diskrete Zeitpunkte ergibt (vgl. dazu Herzog und Rist (1988)). Die Typmarkierungen an den Kanten des Ablaufschemas werden zur Definition der elementaren Ereignisprädikate herangezogen.

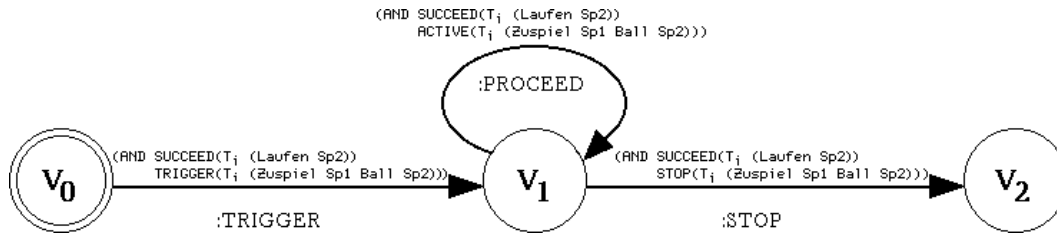


Abbildung 6: Ablaufschema des Konzeptes Pass_in_den_Lauf

5 Zusammenfassung und Ausblick

Bei der Konzeption, die bei der Kopplung von *Vitra* und *Actions* verfolgt wird, besteht die Aufgabe der Bildanalyse in der Erzeugung einer quantitativen Beschreibung grundlegender wahrnehmbarer Größen, wie z.B. Position und Geschwindigkeit von Objekten. Ausgehend von dieser geometrischen Szenenbeschreibung stellen weitergehende Interpretationsprozesse qualitative Beschreibungen räumlicher Anordnungen und zeitübergreifender Vorgänge in Form von räumlichen Relationen und Ereignissen zur Verfügung. Diese konzeptuellen Strukturen sind das Bindeglied zwischen visuellen Daten und sprachlichen Einheiten wie räumliche Präpositionen bzw. Bewegungs- und Handlungsverben. Eine Besonderheit des in *Soccer* verfolgten Ansatzes ist die Zielsetzung, eine Szene simultan zu ihrem Ablauf zu interpretieren und in natürlicher Sprache zu beschreiben. Langfristig betrachtet wird damit die Realzeitverarbeitung bei der Analyse, Interpretation und natürlichsprachlichen Beschreibung zeitveränderlicher Szenen angestrebt (vgl. auch Nagel (1988)).

Um die Interpretationsleistung des *Soccer* Systems und damit die Qualität der erzeugten natürlichsprachlichen Simultanbeschreibungen noch zu verbessern, sollen in einer weiteren Ausbaustufe die folgenden Erweiterungen durchgeführt werden:

- Objektgruppen als Rollenfüller

Läßt man als Rollenfüller auch Objektgruppen zu, dann könnten auch Konzepte wie beispielsweise *Angriff_des_Sturms* innerhalb des Formalismus definiert werden. Eine besondere Schwierigkeit hierbei liegt sicherlich darin, daß sich Objektgruppen, im Gegensatz zu einzelnen Objekten, neu formieren, vergrößern, verkleinern oder auflösen können; d.h. ihre Gestalt als auch ihre zeitliche Existenz ist variabel. Hinzu kommt noch, daß die Menge aller möglichen Objektgruppen exponentiell zur Menge der dynamischen Szenenobjekte wächst.

- Raumregionen als Rollenfüller

Bei der Definition von Konzepten wie etwa *Schuß_vors_Tor* wäre es wünschenswert, als Rollenfüller Regionen, wie z.B. den *Vor_Raum_des_Tores*, heranziehen zu können. Derzeit können Orte nur relational auf der Grundlage räumlicher Relationen beschrieben werden. Die Schwierigkeit bei dieser Repräsentationsform ist darin zu sehen, daß bereits während der Interpretation der Szenendaten zur Charakterisierung eines Ortes sowohl eine geeignete Relation als auch die Bezugsobjekte ausgewählt werden müssen. Da bei dieser Auswahl sehr unterschiedliche Faktoren, u.a. auch der sprachliche Kontext, zu berücksichtigen sind, ergäbe sich hierdurch eine ungewollt starke Vernetzung zwischen Ereigniserkennung und Sprachproduktion. Ein anderer Ansatz, der einer Trennung von Ereigniserkennung und Sprachproduktion entgegenkommt, basiert auf der Einführung eigenständiger Entitäten zur Repräsentation von Orten. Inwiefern sich dieser Ansatz als geeignet erweist, wird davon abhängen, wie Regionen repräsentiert werden können und ob sie sich mit vertretbarem Aufwand bestimmen und adäquat in natürlicher Sprache beschreiben lassen.

- Erkennung und Verbalisierung von Intentionen

Es hat sich gezeigt, daß es in vielen Fällen bei der Beschreibung eines Ereignisses nicht genügt, nur den Verlauf der Trajektorien zu betrachten. Um eine möglichst informative Beschreibung generieren zu können, ist es notwendig, auch die Intentionen der agierenden Objekte zu berücksichtigen. Ob es sich bei einem Ereignis beispielsweise um einen verfehlten Torschuß oder um ein Abwehren des Balls handelt, kann nicht allein aus visuellen Daten geschlossen werden. Vielmehr lassen sich in Abhängigkeit von erkannten Intentionen ein und derselben Trajektorie unterschiedliche Ereigniskonzepte zuordnen. Ein System zur Erkennung von Intentionen und Plänen wurde bereits im Rahmen des Projekts *Vitra* entwickelt (vgl. Retz-Schmidt (1988)) und soll in einer weiteren Ausbaustufe in *Soccer* integriert werden.

- Entwicklung eines visuellen Hörermodells

Im ersten Abschnitt wurden der Aufbau räumlich-zeitlicher Vorstellungen beim Textverstehen einerseits und die natürlichsprachliche Beschreibung von Szenen anhand visueller Daten andererseits als getrennte Aufgabenstellungen beschrieben. Es zeigt sich jedoch, daß es im Hinblick auf die Generierung adäquater Beschreibungen sinnvoll ist, die beim Hörer vermutlich erzeugten visuellen Vorstellungen explizit zu modellieren und bei der weiteren Sprachproduktion zu berücksichtigen (vgl. Wahlster (1987) und Novak (1988)). Durch Abgleich der Imagination des Hörers mit den tatsächlichen Szenendaten läßt sich dann verifizieren, ob eine Äußerung den intendierten Effekt hat. Eine entsprechende Komponente für eine solche Antizipation der Imagination des Hörers wird zur Zeit im Rahmen des *Soccer Systems* entwickelt (vgl. Schirra (1989)).

Danksagung

Die Autoren möchten sich an dieser Stelle bei den Teilnehmern des Workshops “Repräsentation und Verarbeitung räumlichen Wissens” für zahlreiche und wertvolle Anregungen bedanken. Die Abbildung zur Funktionsweise des *Actions Systems* wurde uns freundlicherweise von unseren Kollegen vom Institut für Informations- und Datenverarbeitung (IITB) der Fraunhofergesellschaft, Karlsruhe, zur Verfügung gestellt.

Die hier beschriebene Arbeit wurde teilweise unterstützt vom Sonderforschungsbereich 314 der Deutschen Forschungsgemeinschaft, “Künstliche Intelligenz und wissensbasierte Systeme”, Projekt N2: VITRA (VIsual TRAnslator).

Literatur

- G. Adorni, M. DiManzo, F. Giunchiglia.** Some Basic Mechanisms for Common Sense Reasoning about Stories Environments. In: *Proc. of the 8th IJCAI*, S. 72–74, Karlsruhe, FRG, 1983.
- J. F. Allen.** Towards a General Theory of Action and Time. *Artificial Intelligence*, **23** (2), 123–154, 1984.
- E. André.** Generierung natürlichsprachlicher Äußerungen zur simultanen Beschreibung von zeitveränderlichen Szenen: Das System SOCCER. Memo 26, Universität des Saarlandes, SFB 314 (VITRA), Saarbrücken, 1988.
- E. André, G. Bosch, G. Herzog, T. Rist.** Coping with the Intrinsic and the Deictic Uses of Spatial Prepositions. In: K. Jorrand, L. Sgurev (Red.), *Artificial Intelligence II: Methodology, Systems, Applications*, S. 375–382, North-Holland, Amsterdam, 1987.

- E. André, G. Herzog, 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, Munich, 1988.
- N. I. Badler.** Temporal Scene Analysis: Conceptual Description of Object Movements. Technical Report 80, Computer Science Department, Univ. of Toronto, 1975.
- R. Bajcsy, A. K. Joshi, E. Krotkov, A. Zwarico.** LandScan: A Natural Language and Computer Vision System for Analyzing Aerial Images. In: *Proc. of the 9th IJCAI*, S. 919–921, Los Angeles, CA, 1985.
- I. Carsten, T. Janson.** *Verfahren zur Evaluierung räumlicher Präpositionen anhand geometrischer Szenenbeschreibungen.* Diplomarbeit, Fachbereich für Informatik, Univ. Hamburg, 1985.
- M. Fürnsinn, M. N. Khenkhar, B. Ruschkowski.** GEOSYS – Ein Frage-Antwort-System mit räumlichem Vorstellungsvermögen. In: C.-R. Rollinger (Red.), *Probleme des (Text-) Verstehens, Ansätze der künstlichen Intelligenz*, S. 172–184, Niemeyer, Tübingen, 1984.
- C. Habel.** Repräsentation räumlichen Wissens. In: G. Rahmstorf (Red.), *Wissensrepräsentation in Expertensystemen*, S. 98–131, Springer, Berlin, Heidelberg, 1988.
- K.-J. Hanßmann.** Sprachliche Bildinterpretation für ein Frage-Antwort-System. Bericht 74, Fachbereich Informatik, Univ. Hamburg, 1980.
- A. Herskovits.** On the Spatial Uses of Prepositions. In: *Proc. of the 18th ACL*, S. 1–5, Philadelphia, PA, 1980.
- A. Herskovits.** *Language and Spatial Cognition. An Interdisciplinary Study of the Prepositions in English.* Cambridge University Press, Cambridge, London, 1986.
- G. Herzog, T. Rist.** Simultane Interpretation und natürlichsprachliche Beschreibung zeitveränderlicher Szenen: Das System SOCCER. Memo 25, Universität des Saarlandes, SFB 314 (VITRA), Saarbrücken, 1988.
- M. Hußmann, P. Scheffe.** The Design of SWYSS, a Dialogue System for Scene Analysis. In: L. Bolc (Red.), *Natural Language Communication with Pictorial Information Systems*, S. 143–201, Hanser/McMillan, München, 1984.
- G. Lakoff.** Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic*, **2**, 458–508, 1973.
- D. McDermott.** A Temporal Logic for Reasoning about Processes and Plans. *Cognitive Science*, **6**, 101–155, 1982.

- G. A. Miller, P. N. Johnson-Laird.** *Language and Perception*. Cambridge University Press, Cambridge, London, 1976.
- M. Mohnhaupt.** On Modelling Events with an 'Analogical' Representation. In: K. Morik (Red.), *GWAI-87. 11th German Workshop on Artificial Intelligence*, S. 31–40, Springer, Berlin, Heidelberg, 1987.
- H.-H. Nagel.** From Image Sequences Towards Conceptual Descriptions. *Image and Vision Computing*, **6** (2), 59–74, 1988.
- B. Neumann.** Natural Language Description of Time-Varying Scenes. Report 105, Fachbereich Informatik, Univ. Hamburg, 1984.
- B. Neumann, H.-J. Novak.** Event Models for Recognition and Natural Language Description of Events in Real-World Image Sequences. In: *Proc. of the 8th IJCAI*, S. 724–726, Karlsruhe, FRG, 1983.
- B. Neumann, H.-J. Novak.** NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenen. *Informatik Forschung und Entwicklung*, **1**, 83–92, 1986.
- H.-J. Novak.** What has Imagery to do with Natural Language Generation? In: *Proc. of the Fourth Int. Workshop on Natural Language Generation*, Santa Catalina, CA, 1988, Forthcoming.
- N. Okada.** SUPP: Understanding Moving Picture Patterns Based on Linguistic Knowledge. In: *Proc. of the 6th IJCAI*, S. 690–692, Tokio, Japan, 1979.
- S. Pribbenow.** Verträglichkeitsprüfungen für die Verarbeitung räumlichen Wissens. In: W. Hoepfner (Red.), *Künstliche Intelligenz, GWAI-88, 12. Jahrestagung*, S. 226–235, Springer, Berlin, Heidelberg, 1988.
- G. Retz-Schmidt.** A REPLAI of SOCCER: Recognizing Intentions in the Domain of Soccer Games. In: *Proc. of the 8th ECAI*, S. 455–457, Munich, 1988.
- T. Rist, G. Herzog, E. André.** Ereignismodellierung zur inkrementellen High-level Bildfolgenanalyse. In: E. Buchberger, J. Retti (Red.), *3. Österreichische Artificial-Intelligence-Tagung*, S. 1–11, Springer, Berlin, Heidelberg, 1987.
- J. R. J. Schirra.** Ein erster Blick auf ANTLIMA: Visualisierung statischer räumlicher Relationen. In: D. Metzger (Red.), *GWAI-89: 13th German Workshop on Artificial Intelligence*, S. 301–311, Springer, Berlin, Heidelberg, 1989.
- J. R. J. Schirra, G. Bosch, C.-K. Sung, 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.

- N. K. Sondheimer.** Spatial Reference and Natural Language Machine Control. *Int. Journal of Man-Machine Studies*, **8**, 329–336, 1976.
- C.-K. Sung.** Extraktion von typischen und komplexen Vorgängen aus einer langen Bildfolge einer Verkehrsszene. In: H. Bunke, O. Kübler, P. Stucki (Red.), *Mustererkennung 1988; 10. DAGM Symposium*, S. 90–96, Springer, Berlin, Heidelberg, 1988.
- C.-K. Sung, G. Zimmermann.** Detektion und Verfolgung mehrerer Objekte in Bildfolgen. In: G. Hartmann (Red.), *Mustererkennung 1986; 8. DAGM-Symposium*, S. 181–184, Springer, Berlin, Heidelberg, 1986.
- J. K. Tsotsos.** Temporal Event Recognition: An Application to Left Ventricular Performance. In: *Proc. of the 7th IJCAI*, S. 900–907, Vancouver, Canada, 1981.
- C. Vandeloise.** *Description of Space in French*. Doktorarbeit, Univ. of California, San Diego, CA, 1984.
- W. Wahlster.** Ein Wort sagt mehr als 1000 Bilder. Zur automatischen Verbalisierung der Ergebnisse von Bildfolgeanalyse-Systemen. *Annales, Forschungsmagazin der Univ. des Saarlandes*, **1** (1), 82–93, 1987.
- W. Wahlster, H. Marburger, A. Jameson, 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, FRG, 1983.
- I. Walter.** EPEX: Bildfolgendeutung auf Episodenebene. In: K. Morik (Red.), *GWAI-87. 11th German Workshop on Artificial Intelligence*, S. 21–30, Springer, Berlin, Heidelberg, 1987.
- D. L. Waltz, L. C. Boggess.** Visual Analog Representations for Natural Language Understanding. In: *Proc. of the 6th IJCAI*, S. 926–934, Tokio, Japan, 1979.
- D. Wunderlich.** Raumkonzepte. Zur Semantik der lokalen Präpositionen. In: T. T. Ballmer, R. Posner (Red.), *Nach-Chomskysche Linguistik*, S. 340–351, de Gruyter, Berlin, New York, 1985.
- G. Zimmermann, R. Kories.** Eine Familie von Bildmerkmalen für die Bewegungsbestimmung in Bildfolgen. In: W. Kropatsch (Red.), *Mustererkennung 1984; DAGM/ÖAGM Symposium*, S. 147–153, Springer, Berlin, Heidelberg, 1984.