Informatik Forsch. Entw. (1986) 1: 83–92

Informatik
Forschung und Entwicklung

NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenen

B. Neumann und H.-J. Novak

Universität Hamburg, Fachbereich Informatik, Schlüterstr. 70, D-2000 Hamburg 13

Zusammenfassung. Bild- und Sprachverstehen sind bisher in der Künstlichen Intelligenz weitgehend unabhängig voneinander bearbeitet worden. Im NAOS-System werden beide Bereiche durch eine gemeinsame interne Repräsentation einer Szene verbunden, die im wesentlichen geometrische Informationen enthält. Verschiedene darauf aufbauende wissensbasierte Prozesse, von der Erkennung von Ereignissen bis hin zu Entscheidungsprozessen für die Textgenerierung werden motiviert und vorgestellt.

Schlüsselwörter: Natürlichsprachliche Schnittstellen, Wissensrepräsentation, natürlichsprachliche Generierung, Kontrollstrukturen, zeitveränderliche Bildfolgen

Abstract. Image sequence understanding and understanding natural language are two areas of artificial intelligence that have been investigated almost independently of each other. In the NAOS system both areas are combined via a common internal representation of a scene which basically contains geometric information. Several knowledge-based processes, including event recognition in the visual domain and decision procedures for text generation, are based on this representation and are introduced in this paper.

Key words: Natural language interfaces, Representations, Language generation, Control structures, Time-varying imagery

CR Subject Classifications: I.2.1, I.2.4, I.2.7, I.2.10, I.4.8

1. Einführung und Überblick

Die Untersuchungen, über die im folgenden berichtet wird, befassen sich mit der Aufgabe, eine zeitveränderliche Szene zu beobachten und in natürlicher Sprache zu beschreiben. Dazu werden mit Methoden der Künstlichen Intelligenz Berechnungsmodelle entwickelt, die sowohl eine Theorie der Informationsverarbeitung darstellen, als auch in einem konkreten System implementiert worden sind. Das Thema verbindet zwei bisher im wesentlichen unabhängig voneinander bearbeitete Teilbereiche der KI: Bildverstehen und natürlichsprachliche Systeme. Bildverstehen befaßt sich damit, anhand von Bildern eine symbolische, computerinterne Beschreibung einer Szene zu erzeugen. Das zentrale Problem ist dabei, bedeutungsvolle Objekte zu lokalisieren und zu erkennen. Die Untersuchungen, über die hier berichtet wird, konzentrieren sich jedoch auf darüber hinausgehende Interpretationsaufgaben, die dem Bereich der „Höheren Bilddeutung" zuzurechnen sind. Es geht insbesondere um das Erkennen zeitübergreifender Zusammenhänge in Bildfolgen, speziell um Objektbewegungen, die eine bedeutungsvolle Einheit darstellen, z. B. einen Überholvorgang im Straßenverkehr. Die Bilddeutung erreicht damit einen Abstraktionsgrad, der sprachlichen Begriffen entspricht, in diesem Fall Bewegungsverben, räumlichen Präpositionen etc.

Prozesse, die aus einer symbolischen Beschreibung natürlichsprachliche Sätze generieren, gehören traditionell in den Bereich natürlichsprachlicher Systeme. Obwohl es hierzu bereits einige Lösungsansätze gibt – sie werden weiter unten diskutiert – mußte in verschiedener Hinsicht Neuland betreten werden. Ein Aspekt betrifft die Semantik sprachlicher Begriffe. Sie ist in NAOS letztendlich durch Referenz auf konkrete Szenen definiert, im Gegensatz zu anderen Ansätzen, bei denen konzeptuelle Grundbausteine als Primitive verwendet werden [20]. Die Semantik von NAOS ist formal der Situationssemantik [4] zuzuordnen.

Ein zweiter Problemkreis, für den in NAOS neue Lösungen entwickelt worden sind, ist die Sprechplanung für eine konkrete Szenenbeschreibung. Der Ansatz stützt sich auf ein Partnermodell, in dem der

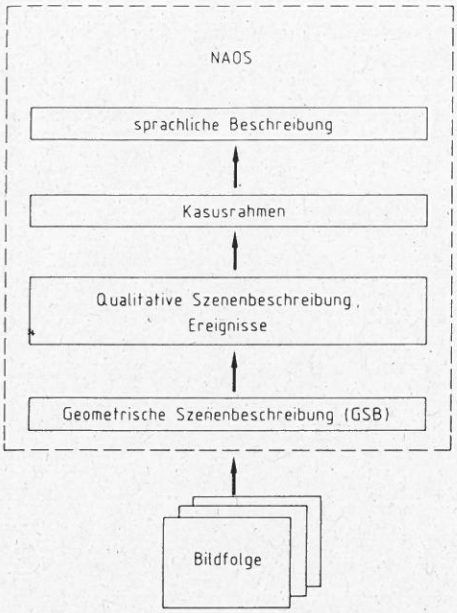

Abb. 1. Die wichtigsten Repräsentationsebenen in NAOS

sukzessive Aufbau einer vom Hörer visualisierten Szene nachvollzogen wird.

Die Untersuchungen sind an einer sehr konkreten Aufgabe orientiert, der Beschreibung von Straßenverkehrsszenen. Dabei werden jedoch auch Zusammenhänge deutlich, die über diese spezielle Anwendung hinaus gültig sind und grundsätzliche Aspekte im Grenzbereich von Bild- und Sprachverstehen betreffen. Es zeigt sich, daß eine interne Szenenrepräsentation in Gestalt der dreidimensionalen Szenengeometrie eine zentrale Rolle sowohl für die Bilddeutung, als auch für Sprachgenerierung und Sprachverstehen spielt, ebenso auch für einige weitere kognitive Prozesse wie Bewegungsplanung und räumliches Schließen. Eine solche gemeinsame Datenbasis ist beispielsweise für zukünftige Robotersysteme von Bedeutung, in denen zahlreiche kognitive Prozesse integriert werden müssen.

Abbildung 1 zeigt den schematischen Aufbau von NAOS anhand der wesentlichen Repräsentationsebenen. NAOS setzt voraus, daß eine Bildfolge bis zur Ebene der geometrischen Szenenbeschreibung (GSB) analysiert worden ist. Damit stehen Angaben über die räumliche Lage aller Objekte zu jedem Zeitpunkt in quantitativer Form zur Verfügung. Eine Begründung und genauere Spezifikation dieser intermediären Repräsentation erfolgen in Abschnitt 2.

Der erste Verarbeitungsschritt überführt diese Daten in eine qualitative Beschreibung, die an sprachlichen Konzepten orientiert ist. Insbesondere werden „Ereignisse" erkannt, die den Bewegungsverben der natürlichen Sprache entsprechen. Ereigniserkennung und verschiedene, damit verbundene Probleme (Ereignishierarchie, Zeitlogik, effektive Berechnungsverfahren), werden in Abschnitt 3 behandelt.

Ereignisse bilden den Grundstock für die sprachliche Szenenbeschreibung. In Abschnitt 4 wird zunächst gezeigt, wie die Komponenten eines Ereignisses mit den Tiefenkasus des zugehörigen Bewegungsverbs in Beziehung stehen. Sodann wird die Planungskomponente beschrieben, die durch Auswahl geeigneter Ereignisse und Tiefenkasus einen kohärenten, informativen Text erzeugt.

Im letzten Abschnitt werden die Grenzen der hier vorgestellten Verfahren diskutiert. Es werden einige weiterführende Ansätze aufgezeigt, die einen Ausbau der allgemeinen Wissensbasis voraussetzen. Zum einen geht es um typische Objektbewegungen im Kleinen, zum anderen um charakteristische Ereignisfolgen, also ausgedehntere zeitliche Zusammenhänge.

2. Geometrische Szenenbeschreibung

Im folgenden werden die Eingabedaten von NAOS näher beschrieben. Sie stellen eine kanonische Zwischenrepräsentation dar, die in der Literatur bisher nicht explizit diskutiert worden ist und deshalb hier besonders begründet wird.

Entsprechend der in Abb. 1 dargestellten Verarbeitungskette sind als Datenquellen Bildfolgen vorgesehen, die z. B. durch Kameraaufnahmen einer zeitveränderlichen Szene entstanden sein können. Die ersten Verarbeitungsschritte bis hin zur Objekterkennung sind die Aufgabe eines bildverstehenden Systems im engeren Sinn: Es geht darum festzustellen, „wo wann was ist", also welche Objekte zu welchem Zeitpunkt an welcher Stelle sind. Diese Aufgabe kann beim heutigen Stand der Technik für natürliche Szenen noch nicht zufriedenstellend gelöst werden, obwohl der weitaus größte Teil aller Forschungsanstrengungen im Bereich Bildverstehen damit befaßt ist.

Um uns von diesen Schwierigkeiten zu lösen, gehen wir im folgenden davon aus, daß alle interessierenden Objekte der Szene bereits erkannt und ihre dreidimensionale Form sowie ihre Lage im Raum exakt bestimmt sind. Diese Zwischenrepräsentation enthält im wesentlichen geometrische Informationen, jedoch keinerlei Konzepte der höheren Bilddeutung. In [15] wurde dafür die Bezeichnung „Geometrische Szenenbeschreibung" (GSB) eingeführt. Es ist klar, daß ein Bildanalysesystem dieses Ergebnis genau genommen kaum jemals erbringen kann,

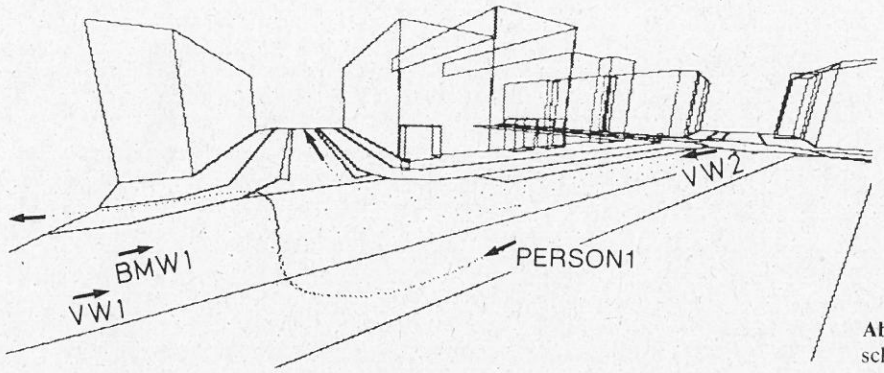

Abb. 2. Objektbewegungen in einer synthetischen Szene

denn schließlich sind Objekte im allgemeinen nicht vollständig sichtbar. Höhere Bilddeutung basiert jedoch vorwiegend auf dem, was man über eine Szene weiß, nicht auf dem, was man sieht. Deshalb scheint es angemessen, bei einer kanonischen Zwischenrepräsentation von der vollständigen 3D-Geometrie auszugehen. Hierzu sollen auch Betrachterstandpunkt und Beleuchtungsgeometrie gerechnet werden.

Aus Gründen der Vollständigkeit soll die GSB auch photometrische Objekteigenschaften (z. B. Farbe) enthalten. Obwohl diese Daten für NAOS nicht von großer Bedeutung sind, garantieren sie doch, daß eine GSB im folgenden Sinn vollständig ist: Man kann aus ihr die Bildfolge regenerieren – zumindest im Prinzip. Eine GSB impliziert also keinen Informationsverlust gegenüber den ursprünglichen Bilddaten.

Zusammengefaßt enthält die GSB also

- für jeden Zeitpunkt:
  - alle in der Szene sichtbaren Objekte
  - Betrachterstandpunkt
  - Beleuchtungsdaten
- für jedes Objekt:
  - 3D-Position und -Orientierung je Zeitpunkt
  - 3D Form
  - physikalische Oberflächeneigenschaften (Farbe)
  - Klassenzugehörigkeit
  - Identität.

‚Klassenzugehörigkeit' entspricht der traditionellen ISA-Beziehung, während ‚Identität' physikalisch identische Objekte in Beziehung setzt. Die Identität von bewegten Objekten zu verschiedenen Zeitpunkten (Korrespondenz) wird stets als bekannt angenommen. Darüberhinaus kann die Identität einzelner Objekte auf Grund entsprechenden Vorwissens bekannt sein.

Eine GSB für NAOS beschreibt Verkehrsszenen, die von unserem Labor aus zu beobachten sind. Dazu wird ein detailliertes, rechnerinternes Modell der stationären Umgebung verwendet (Häuser, Straßen, Bäume etc.). Objektbewegungen können aus tatsächlichen Bildfolgen extrahiert oder durch künstliche Daten vorgegeben werden. Abbildung 2 zeigt eine synthetische Szene mit eingeblendeten Objekttrajektorien.

Für die Ereigniserkennung spielen die Trajektorien der bewegten Objekte eine besondere Rolle. Sie werden relational in folgendem Format repräsentiert:

```
(LAGE ⟨Objektbezeichner⟩ ⟨xyz-Koordinaten⟩
      ⟨Orientierungsvektor⟩ ⟨Zeitpunkt⟩)
```

z. B.
```
(LAGE VW1 (93 158 0) (.9912 -.1322 .0) 18)
(LAGE VW1 (77 159 0) (.9981 -.0624 .0) 19)
(LAGE VW1 (60 159 0) (1.0000 .0000 .0) 20)
etc.
```

Die relationalen Tupel beschreiben Schwerpunktslage und Orientierung eines Objektes für fortlaufende Zeitpunkte. Objekteigenschaften wie Form und Zusammengehörigkeit sind in objektzentrierten Schemata (frames) enthalten, die hier nicht im Einzelnen beschrieben werden.

3. Ereigniserkennung

Eine zentrale Aufgabe von NAOS besteht darin, die quantitativen Daten der GSB in eine qualitative, sprachbezogene Beschreibung zu überführen. In diesem Abschnitt werden zunächst Ereignismodelle vorgestellt, die a priori Wissen über interessante Bewegungsabläufe in einer Szene repräsentieren. Sie beruhen auf semantischen Primitiven, die perzeptuell motiviert sind. Dann wird der Prozeß der Ereigniserkennung erläutert. Es handelt sich dabei um ein relationales Vergleichsverfahren, bei dem zusätzlich zeitlogische Beziehungen durch Propagieren von Beschränkungen ausgewertet werden.

Die Bewegungsverben der natürlichen Sprache stellen ein naheliegendes Repertoire von Konzepten dar, mit denen Objektbewegungen in einer Szene qualitativ beschrieben werden können, zumal dann, wenn letztendlich eine sprachliche Beschreibung angestrebt wird. Es ist jedoch auch durchaus möglich, nicht-sprachliche Konzepte zu verwenden. Ein Beispiel sind die Arbeiten zur automatischen Klassifizierung von Herzkammerbewegungen in Röntgenaufnahmen [6, 21], bei denen medizinisch relevante Bewegungskonzepte zugrundegelegt wurden. Man kann sagen, daß die Konzeptauswahl im allgemeinen dadurch bestimmt wird, welche Bewegungsabläufe in dem jeweiligen Anwendungsfeld bedeutungsvolle Einheiten darstellen.

Das aus KI-Sicht wichtigere Problem ist die Strukturierung der Bewegungskonzepte im Hinblick auf eine transparente Repräsentation und effektive Berechnungsverfahren. Rein prozedurale Lösungen, bei denen die Bedeutung eines Konzeptes in einem Programm verborgen ist, befriedigen nicht, denn hierbei können weder die Beziehungen der Konzepte untereinander noch ihre Bedeutung für das menschliche Verständnis explizit gemacht werden. Die semantische Dekomposition in [21] ist eine überzeugende Lösung. Sie führt auf eine Hierarchie von physikalisch unterscheidbaren Bewegungstypen, z. B. ‚allgemeine Bewegung', ‚nichtstarre Bewegung', ‚Schrumpfen', etc. (Eine Diskussion weiterer Ansätze findet sich in [13]).

Uns geht es weniger darum, reine Bewegung zu strukturieren, als vielmehr Konzepte zu definieren, die zwar Bewegungen implizieren aber darüberhinaus noch viele zusätzliche Aspekte einer Situation enthalten können. Z. B. vermittelt das Bewegungsverb „bremsen" im Kontext einer Verkehrsszene nicht nur eine Aussage über eine negative Beschleunigung, sondern auch Informationen über den Agenten (ein Fahrzeug). Dies wird noch deutlicher bei komplexeren Verben wie z. B. „ausweichen", wo eine Situation komplexe Voraussetzungen erfüllen muß, bevor das Verb anwendbar ist.

Weil Bewegungsverben in verschiedener Hinsicht über reine Bewegungskonzepte hinausgehen, wird im folgenden von *Ereignissen* gesprochen. Ein Ereignis ist ein vierdimensionaler Unterraum des Ort-Zeit-Kontinuums, der mit einem Bewegungsverb beschreibbar ist, also ein Ausschnitt aus einer zeitveränderlichen Szene. Ein *Ereignismodell* ist eine generische Beschreibung von Ereignissen. Ereignismodelle definieren die Semantik von Bewegungsverben und dienen der Ereigniserkennung.

Ein Ereignismodell in NAOS besteht aus einem Kopf, der das Ereignis in Form einer Proposition ausdrückt, und einem Rumpf, in dem Prämissen zusammengefaßt sind, die für das Ereignis erfüllt sein müssen. Wir benutzen hier die Notation:

⟨Kopf⟩ ← ⟨Rumpf⟩

Das folgende Ereignismodell definiert ABBIEGEN. (Weitere Beispiele finden sich in [14]).

```
(ABBIEGEN *OBJ1 *OBJ2 *T1 *T2) ←
        ((DREHEN *OBJ1 *T1 *T2)
         (PARALLEL *OBJ1 *OBJ2 *T1 *T3)
         (AUF *OBJ1 *OBJ2 *T1 *T1 *T4)
         (NICHT AUF *OBJ1 *OBJ2 *T4 *T2))
```

Alle mit * versehenen Bezeichner sind Variable, die für ein konkretes Abbiegeereignis instantiiert werden müssen. Die letzten zwei Komponenten eines Tupels stehen für Start- und Endzeit eines Zeitintervalls, für das das jeweilige Prädikat gültig ist. DREHEN ist seinerseits ein Ereignis, das weiter zerlegt werden kann:

```
(DREHEN *OBJ *T1 *T2) ←
        ((ODER (LINKS-DREHEN *OBJ *T1 *T2)
               (RECHTS-DREHEN *OBJ *T1 *T2))
```

Alle anderen oben genannten Prädikate sind primitiv, d. h. sie werden durch eine zugeordnete Prozedur direkt auf der GSB evaluiert.

Die Beispiele lassen erkennen, daß Ereignisse eine Spezialisierungshierarchie bilden: Spezielle Ereignisse werden aus allgemeineren aufgebaut oder - genauer gesagt - implizieren allgemeinere Ereignisse. Z. B. impliziert LOSFAHREN sowohl STEHEN als auch FAHREN. Die Hierarchie basiert auf semantischen Primitiven, die interessante Gesetzmäßigkeiten offenbaren.

- Primitive Prädikate sind durativ, d. h. sie beschreiben Zeitintervalle, in denen bestimmte Szeneneigenschaften ununterbrochen gültig sind.
- Primitive Prädikate sind qualitative Prädikate über beobachtbaren Größen, die durch die GSB quantitativ vorgegeben sind.

Es gibt nur eine kleine Zahl von beobachtbaren Größen, die hierfür in Betracht kommen. (Das ist auch plausibel angesichts der Dimension eines physikalischen Zustandsvektors.) Sie sind: relative Position und Orientierung eines Objektes bezüglich geeigneter Referenzgrößen, sowie die zugehörigen zeitlichen Ableitungen, jeweils als Funktion der Zeit. Diese Größen stellen gewissermaßen die perzeptuelle Basis der Szenenbeschreibung dar.

Die qualitativen Prädikate erfassen verschiedene Arten von Konstantheiten, die geeignet sind, den zeitlichen Verlauf der beobachteten Größen zu charakterisieren:

- Konstanter Wert
  z. B. Stillstand, Geradeausbewegung, konstante Geschwindigkeit
- eingeschränkter Wertebereich
  z. B. parallel, nahe, neben, auf
- größer/kleiner
  z. B. Komparative, ungewöhnliche Werte (rasen)
- monotoner Verlauf
  z. B. beschleunigen, drehen, sich nähern

Im Gegensatz zu bisherigen Ansätzen [12, 18, 20] sehen wir die semantischen Primitive nicht als atomare konzeptuelle Grundbausteine sondern als qualitative Operatoren an, die eine perzeptuelle Basis erschließen. Zur Zeit sind in NAOS die folgenden 19 Primitive implementiert:

EXISTIEREN, STEHEN, BEWEGEN, SCHNELLER-WERDEN, LANGSAMER-WERDEN, LINKS-DREHEN, RECHTS-DREHEN, SYM-ENTFERNEN (=wachsender Abstand), SYM-NÄHERN, LANGSAM (=langsamer als Normalwert), SCHNELL, PARALLEL, QUER, VOR, HINTER, NEBEN, BEI, AUF, NAHE.

Es ist bemerkenswert, daß räumliche Primitive wie z. B. AUF oder NAHE in dieser Sichtweise genauso wie Bewegungsprimitive behandelt werden, eine strenge Unterscheidung von örtlichen und zeitlichen Konzepten erscheint also nicht sinnvoll. Jedes Primitiv beschreibt Zeitintervalle, in denen eine räumliche Beziehung gültig ist. Auf diese Weise ist es möglich, eine heterogene Folge von Prädikaten (wie z. B. in ABBIEGEN) auf homogene Weise zu evaluieren.

Die aus diesen Primitiven aufgebaute Ereignishierarchie umfaßt ca. 50 Verben, die alle zur Beschreibung von Verkehrsszenen in Frage kommen. Eine vollständige Liste findet sich in [14].

Wir wenden uns jetzt dem Prozess der Ereigniserkennung zu. Zum Aufbau einer kompletten Szenenbeschreibung werden alle Ereignisse gesucht, die in der Szene instantiiert werden können. Dies geschieht unter Ausnutzung der Spezialisierungshierarchie, die von der Wurzel (EXISTIEREN) zu spezielleren Ereignissen hin abgearbeitet wird. Das Verfahren für ein einzelnes Ereignis ist Hierarchischer Vergleich mit Backtracking [3]. Dabei werden durch rekursiven Abstieg in die Komponenten eines Ereignisses alle möglichen Kombinationen von Variableninstantiierungen erzeugt und geprüft. Durch dynamisches Umordnen der Komponenten wird dabei der Verzweigungsgrad des Suchbaumes klein gehalten. Das Verfahren ist in [14] ausführlich beschrieben.

Im folgenden soll auf die zeitlogischen Aspekte eingegangen werden, die für die Ereigniserkennung von grundsätzlicher Bedeutung sind. Das verwendete Verfahren beruht entscheidend darauf, daß die Ausgangsdaten in der GSB mit quantitativen Zeitangaben, bezogen auf eine zusammenhängende Zeitachse, versehen sind. Dadurch ist es möglich, Gleichzeitigkeit, Nacheinander und andere zeitlogische Konzepte algebraisch zu behandeln. Andere Zeitlogiken wie z. B. in [2] sind darauf angelegt, einen sprachlich ausgedrückten Sachverhalt, für den es im allgemeinen keine explizite Zeitachse gibt, adäquat zu repräsentieren. Dazu müssen eine Reihe von logischen Primitiven eingeführt werden, z. B. MEETS (Ereignis1, Ereignis2) für zeitlichen Anschluß. In NAOS können diese Primitive auf Grund der explizit genannten Zeitvariablen entfallen.

Die algebraische Behandlung von Zeitbeziehungen kann durch Propagieren von Beschränkungen sehr effektiv erfolgen. Dies wird nun näher erläutert. Zunächst wird die Instantiierung eines primitiven Prädikates betrachtet, z. B.

(BEWEGEN *OBJ *T1 *T2)

Sei (BEWEGEN VW1 13 42) Teil der Datenbasis, dann kann *OBJ1 mit VW1 instantiiert werden, *T1 und *T2 dagegen können keine festen Werte erhalten. Vielmehr müssen sie wegen der Durativität von BEWEGEN alle möglichen Teilintervalle innerhalb von (13 42) repräsentieren. Anstelle einer Instantiierung erfolgt also eine Beschränkung, die bei primitiven Prädikaten stets die Form hat:

$$T1_{min} \leq *T1 < *T2 \leq T2_{max}$$

Bei einem Ereignismodell, das mehrere Primitive konjunktiv verknüpft, akkumulieren die Beschränkungen, und es ergibt sich das Problem, ein System von linearen Ungleichungen zu lösen. Man kann leicht sehen, daß der Lösungsbereich für jede Zeitvariable zusammenhängend ist, so daß Ereignisse im allgemeinen über einem Zeitintervall (*T1 *T2) gültig sind, das wie folgt beschränkt ist:

$$T1_{min} \leq *T1 \leq T1_{max}$$
$$T2_{min} \leq *T2 \leq T2_{max}$$
$$*T1 < *T2$$

Ereignisse dieser Art, bei denen Start- und Endzeit individuell beschränkt sind, sind z. B. ÜBERHOLEN und ÜBERQUEREN. Drei Sonderfälle sind von Bedeutung, bei denen die Beschränkungen eine besondere Form haben. *Durative* Ereignisse bzw. Prädikate sind bereits eingeführt worden, sie sind beschränkt wie weiter oben angegeben. *Inchoative* Ereignisse haben einen festen Startzeitpunkt und genügen den Beschränkungen:

$$T1_{min} = *T1 < *T2 \leq T2_{max}$$

Analog haben *resultative* Ereignisse einen festen Endzeitpunkt und genügen den Beschränkungen:

$$T1_{min} \leq *T1 < *T2 = T2_{max}$$

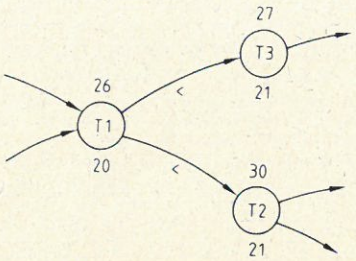

Abb. 3. Beschränkungsnetz für Zeitvariable

Die Berechnung von Ereignisgrenzen aus akkumulierenden Beschränkungen erfolgt mit jeder Instantiierung eines zeitbezogenen Prädikates, um ggf. frühzeitiges Backtracking zu ermöglichen, und muß deshalb möglichst effektiv sein. In NAOS wird ein Beschränkungsnetz (Abb. 3) verwendet, dessen Knoten Zeitvariable und dessen Kanten Ungleichungsbeziehungen zwischen Zeitvariablen repräsentieren.

In jedem Knoten ist der aktuelle Minimal- und Maximalwert der entsprechenden Zeitvariablen gespeichert. Kommt eine neue Beschränkung dazu, so wird sie entlang der Kanten zu allen betroffenen Knoten propagiert. Sobald der Minimalwert eines Knotens den Maximalwert übersteigt, ist keine Lösung mehr möglich und Backtracking muß erfolgen. Der Algorithmus ist in [16] ausführlich beschrieben.

4. Verbalisierung

Die bisher diskutierten Prozesse und Repräsentationen können als Erweiterung eines bildverstehenden Systems um eine zusätzliche Abstraktionsebene angesehen werden, die der Erkennung zeitübergreifender Konzepte dient. In diesem Abschnitt wird der Zusammenhang zwischen diesen Konzepten und korrespondierenden natürlichsprachlichen Beschreibungen aufgezeigt.

Zunächst wird ein einzelnes Ereignis betrachtet, z. B.

```
(ÜBERHOLEN BMW1 LKW1 (40 47) (52 58))
```

Hiermit wird ausgedrückt, daß ein Objekt mit dem internen Namen BMW1 ein zweites Objekt LKW1 in einem Zeitintervall überholt, dessen Startzeit zwischen 40 und 47, und dessen Endzeit zwischen 52 und 58 liegt. Ein natürlichsprachlicher Satz, der dieses Ereignis beschreibt (und dazu auch weitere Szenendaten einbezieht), könnte z. B. so lauten:

„Ein roter BMW überholte den LKW neben der Post, nachdem der Fußgänger die Straße überquert hatte."

Wir beschreiben jetzt die wichtigsten Verarbeitungsschritte, die zu so einem Satz führen. Als erster Schritt muß das Ereignis in eine für die Verbalisierung geeignete Tiefenstruktur überführt werden. Da Ereignisse verborientiert sind, liegt es nahe, den Kasusrahmen des Verbs als formale Repräsentationsform zu wählen. Es zeigt sich, daß die einzelnen Tiefenkasus eines Kasusrahmens mit den Komponenten des zugehörigen Ereignisses bzw. mit den dadurch beschriebenen Bestandteilen der bewegten Szene in enger Beziehung stehen. Für das Verb „überholen" beispielsweise gibt der Kopf des Ereignismodells die beteiligten Objekte *OBJ1 und *OBJ2 an, die offensichtlich Agent und objektiver Kasus sind. Der Lokativ ist das räumliche Volumen, das der Agent während des Ereignisses durchstreicht. Es ist unmittelbar durch die GSB vorgegeben. Quell- und Zielkasus sind analog als räumliche Volumina definiert. Die zeitliche Lage bildet einen weiteren Tiefenkasus, der durch das Ereignisintervall definiert ist. Zusammenfassend kann man sagen, daß die Tiefenkasus die an einem Ereignis beteiligten Objekte sowie Zeit und Ort des Geschehens spezifizieren. Man beachte, daß die Tiefenkasus hier perzeptuell definiert werden, d. h. sie beziehen sich auf szenische Daten. Diese Sichtweise stellt eine Konkretisierung der von Fillmore [9] vorgeschlagenen „Scenes-and-Frames" Semantik dar.

Die Hauptaufgabe der Verbalisierung besteht darin, die durch interne Bezeichner bzw. Szenenkoordinaten repräsentierten Tiefenkasus sprachlich zu referenzieren. Es können drei Aufgaben unterschieden werden:

- Referenzierung von Objekten (REF)
  z. B. „ein roter BMW"
- Referenzierung von Orten (LOC-REF, PATH-REF, u. a.)
  z. B. „neben der Post", „in Richtung Hartungstraße"
- Referenzierung von Zeiten (TIME-REF)
  z. B. „nachdem der Fußgänger die Straße überquert hatte".

Jede dieser Referenzierungsaufgaben unterliegt ihren eigenen Gesetzmäßigkeiten. Im folgenden werden die in NAOS realisierten Referenzierungskomponenten in groben Zügen erläutert.

REF erzeugt eine Nominalphrase oder ein Pronomen, mit dem ein Szenenobjekt sprachlich identifiziert werden kann. Zunächst wird geprüft, ob das Objekt dem Hörer bekannt ist. Dies kann der Fall sein entweder, weil es a priori bekannt ist (in NAOS kennt der Hörer einige stationäre Objekte) oder weil es schon vorher eingeführt worden ist. Bei Vorerwähntheit im vorhergehenden Satz generiert REF ein Pronomen. War dort allerdings von zwei bewegten Objekten die Rede, wird eine Nominalphrase generiert, um Mehrdeutigkeiten zu vermeiden.

Nominalphrasen für bekannte Objekte sind *definit*. Sie werden mit Hilfe der Klassenzugehörigkeit, die in der GSB verzeichnet ist, und möglicherweise zusätzlichen diskriminierenden Merkmalen aufgebaut. Falls ein Eigenname existiert, wird dieser verwendet. Als diskriminierendes Merkmal kommt auch ein das Objekt betreffendes vorerwähntes Ereignis in Frage, das zu einer attributierten Nominalphrase führen kann, z. B. „der LKW, der überholt worden ist". Diese Möglichkeit, die auf der Bewegungskompetenz von NAOS beruht, findet sich bisher in keinem anderen System.

Objekte, die dem Hörer unbekannt sind, werden durch *indefinite* Nominalphrasen eingeführt. Dabei wird zwischen stationären und bewegten Objekten unterschieden. Stationäre Objekte werden durch Klassenzugehörigkeit und - soweit erforderlich - diskriminierende Merkmale referenziert, z. B. „ein großer Baum". Alternativ können Positionsangaben zur Diskriminierung verwendet werden, z. B. „ein VW vor dem Fachbereich Informatik". Bewegte Objekte brauchen nicht mit diskriminierenden Merkmalen eingeführt zu werden, falls das mit ihnen assoziierte Ereignis zur Unterscheidung ausreicht. Andernfalls müssen diskriminierende Merkmale gefunden werden. Wenn dies nicht möglich ist, bleiben Standardphrasen wie z. B. „ein weiterer roter BMW".

LOC-REF erzeugt eine sprachliche Referenzierung für Lokative. Ähnliche Prozeduren existieren für die anderen örtlichen Tiefenkasus (Quellort, Pfad). Der Algorithmus ist weniger differenziert als REF aber dennoch sehr aufwendig. Als Ausgangsdaten wird eine Ortsspezifikation erwartet, die beim Lokativ durch eine Objekttrajektorie innerhalb eines bestimmten Zeitintervalls gegeben ist. Als erstes wird eine für den Lokativ geeignete Präposition ausgewählt. Dann wird geprüft, ob es ein geeignetes Bezugsobjekt gibt, dessen Lage zum Lokativ in der entsprechenden präpositionalen Beziehung steht. Falls dies nicht der Fall ist, wird das Verfahren für eine andere Präposition wiederholt. Die Auswahl eines geeigneten Bezugsobjektes bei vorgegebener Präposition ist eine komplexe Aufgabe, bei der zahlreiche Gesichtspunkte (z. B. Nähe, relative Größe, Beweglichkeit, Orientierung) berücksichtigt werden müssen. Hierzu sind im Zusammenhang mit NAOS vertiefende Untersuchungen angestellt worden [8].

Es bleibt anzumerken, daß die Referenzierung des Bezugsobjektes wiederum eine Aufgabe für REF ist. Dabei können weitere rekursive Aufrufe von LOC-REF erforderlich werden, wenn das Bezugsobjekt durch seinen Standort diskriminiert werden muß. Auf diese Weise entstehen geschachtelte Ausdrücke wie z. B. „neben dem Baum vor dem Fachbereich Informatik".

Die dritte Referenzierungskomponente, TIME-REF, hat die Aufgabe, die zeitliche Lage eines Ereignisses sprachlich auszudrücken. Von der Ereigniserkennung her liegt das Ereignisintervall in Form von (beidseitig beschränkten) Zeitmarken vor, die sich auf die Zeitachse der Szene beziehen. Prinzipiell wird für Beschreibungen als Verbzeit das Präsens gewählt, ähnlich wie für immer-wahre Aussagen (Galileo stell*te* fest: Die Erde *dreht* sich um die Sonne). Um dem Hörer eine genauere zeitliche Einordnung des Ereignisses zu ermöglichen, können zeitliche Nebensätze erzeugt werden, die auf andere Ereignisse Bezug nehmen (s. das Beispiel oben). Das hierzu verwendete Verfahren ist in mancher Hinsicht analog zur örtlichen Referenzierung. Zunächst wird ein dem Hörer bekanntes Bezugsereignis aus der zeitlichen Nachbarschaft gesucht. Dabei kommt es darauf an, daß Start- oder Endzeitpunkt dieses Ereignisses eng eingegrenzt ist, also als Referenzzeitmarke geeignet ist. In der Regel kommen demnach keine durativen Ereignisse in Frage. Schließlich wird eine zeitliche Konjunktion gesucht, die die Beziehung zwischen Ereignis und Bezugszeit korrekt ausdrückt, z. B. „nachdem", „bevor", „während".

Damit sind die Referenzierungskomponenten von NAOS vorgestellt. Es verbleibt die Aufgabe, aus den erzeugten Satzbausteinen einen grammatikalisch korrekten Satz aufzubauen. Wir benutzen dazu eine geringfügig modifizierte Version des Programms SUTRA [7], das für das Dialogsystem HAM-ANS [10] entwickelt wurde. Es verfügt über ein umfangreiches Lexikon mit Informationen über Wortarten, Flexion, etc., so daß NAOS für diesen Teil der Generierung keine Sorge tragen muß.

5. Sprechplanung

Als Ausgabe von NAOS soll eine informative, kohärente Beschreibung der in der Szene sichtbaren Objektbewegungen erzeugt werden. Dazu reicht es offenbar nicht, alle erkannten Ereignisse in einer beliebigen Reihenfolge zu verbalisieren. Vielmehr bedarf es einer Sprechplanung, die zahlreiche Optionen und Entscheidungsmöglichkeiten beim Verbalisieren im Hinblick auf das angestrebte Ziel zum Gegenstand hat.

Bevor das Planungsverfahren näher erläutert wird, soll das damit verfolgte Ziel schärfer gefaßt werden. Beim „Informieren" eines Hörers geht es darum, ihn möglichst in denselben Kenntnisstand über eine Sache zu versetzen, den der Sprecher hat. Das bedeutet konkret für NAOS: Der Hörer muß

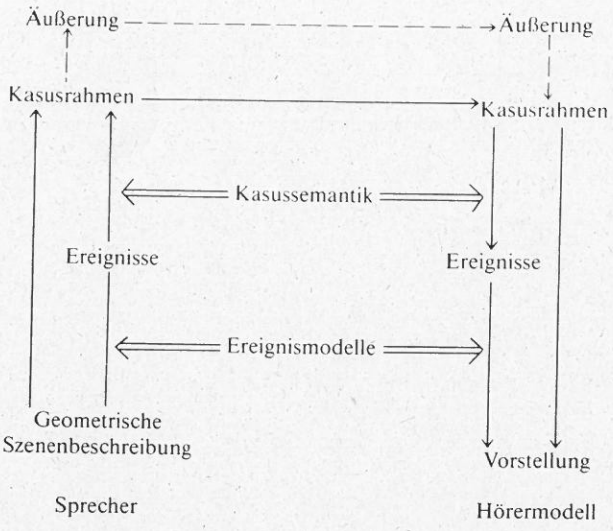

Abb. 4. Schematischer Ablauf der antizipierten Visualisierung

Kenntnis von den Objektbewegungen einer Szene erhalten. Dabei darf auf zahlreiche, als a priori bekannt vorausgesetzte stationäre Szenenbestandteile Bezug genommen werden.

Um dieses Ziel zu erreichen, muß die Sprechplanungskomponente berücksichtigen, welche „Vorstellungen" sich im Hörer bei fortschreitender Szenenbeschreibung entwickeln. Aus dem Vergleich dieser Vorstellungen mit der konkreten Szene kann dann im Prinzip erkannt werden, wo Informationen zu ergänzen sind, und ob eine geplante Äußerung informativ ist. Dies bedeutet letzten Endes, daß der Verstehensprozeß des Hörers sprecherseitig vorweggenommen werden muß. Wir bezeichnen diese Aufgabe als „antizipierte Visualisierung". Abbildung 4 zeigt schematisch den Ablauf.

Die Notwendigkeit, den Effekt einer Äußerung zu antizipieren, überrascht sicherlich nicht und wurde auch vorher von anderen Forschern aufgezeigt [1, 11]. Das Besondere unseres Ansatzes liegt darin, daß der antizipierte Effekt eine visualisierte Szene ist, deren Erzeugung und interne Repräsentation besondere Probleme aufwerfen. Visualisierungen sind auch auf Grund der kognitiven Bezüge besonders interessant [5], eine Diskussion dieser Problematik würde jedoch den Rahmen dieses Beitrags sprengen. Im folgenden wird das Planungsverfahren vorgestellt, das als erste Näherung für den in Abb. 4 gezeigten Ablauf in NAOS implementiert ist. Es handelt sich um einen Standardplan, bei dem eine Beschreibung nach festen Regeln erfolgt, die die Visualisierbarkeit der Szene in gewissem Umfang sicherstellen, aber eine explizite Visualisierung nicht erfordern.

Standardplan zur Szenenbeschreibung

Ausgangsdaten sind die GSB sowie alle darauf instantiierbaren Ereignisse. Die Beschreibung beginnt mit einer standardisierten Übersicht über alle bewegten Objekte, z. B.

„Die Szene enthält fünf bewegte Objekte: drei Autos und zwei Fußgänger".

Daran schließen chronologische Beschreibungen für jedes einzelne Objekt an, die nach folgenden Regeln aufgebaut sind:

A Die sichtbaren Teile der Objekttrajektorie werden durch Ereignisse lückenlos beschrieben.

B Es werden die jeweils speziellsten Ereignisse entsprechend der Ereignishierarchie verbalisiert, es sei denn, die erforderlichen Ortsangaben (siehe C) machen es notwendig, ein anderes, allgemeineres Ereignis mit einer geeigneteren Tiefenkasusstruktur zu wählen.

C Orts- und Zeitangaben werden generiert, soweit sie zur räumlich-zeitlichen Einordnung eines Ereignisses in die Szene erforderlich oder für ein Ereignis obligat sind.

(i) Die räumliche Lage wird mit Hilfe von Quell-, Zeit-, Pfad- und Lokativangaben spezifiziert. Die dabei ggf. verbleibenden Unsicherheiten, z. B. bezüglich der Bewegungsrichtung, werden durch Folgeereignisse beseitigt.

(ii) Zeitspezifikationen (außer durch die Verbzeit) erfolgen nur in Ausnahmefällen. In der Regel ergibt sich die zeitliche Einbettung wie folgt:
- Das erstgenannte Ereignis beginnt am Anfang der Szene. Nachfolgend genannte Ereignisse schließen chronologisch an.
- Die Dauer eines Ereignisses folgt aus Standardwerten, bei durativen Ereignissen aus einer Standardgeschwindigkeit. Falls ein Zeitpunkt explizit etabliert werden muß (z. B. bei überlappenden Ereignissen), wird auf bereits genannte Ereignisse in der zeitlichen Nachbarschaft Bezug genommen.

Die folgende Beschreibung ist unter Verwendung des Standardplanes für die Beispielsszene in Abb. 2 generiert worden.

„Die Szene enthält vier bewegte Objekte: drei PKWs und einen Fußgänger.

Ein VW fährt von der alten Post vor den Fachbereich Informatik. Er hält an.

Ein anderer VW fährt in Richtung Dammtor. Er biegt von der Schlüterstraße ab. Er fährt in Richtung Grindelhof auf der Bieberstraße.

Ein BMW fährt in Richtung Hallerplatz. Dabei überholt er den VW, der angehalten hat, vor der Bieberstraße. Der BMW hält an der Ampel an.

Der Fußgänger geht in Richtung Dammtor. Dabei überquert er die Schlüterstraße vor dem Fachbereich Informatik."

6. Zusammenfassung und Ausblick

In den vorangegangenen Abschnitten wurden Repräsentationen und Prozesse vorgestellt, mit denen in NAOS eine zusammenhängende, natürlichsprachliche Beschreibung für eine zeitveränderliche Szene berechnet wird. Im folgenden werden die wichtigsten Aussagen zusammengefaßt, die über die konkrete Beispielswelt des Straßenverkehrs hinausgehen und für zukünftige Systeme relevant sein können, in denen Bildverstehen und Sprachverstehen integriert sind. Man kann dabei z. B. an zukünftige Robotersysteme denken.

Sprach- und Bildverstehen sind über eine gemeinsame interne Repräsentation des Gegenständlichen gekoppelt, in NAOS realisiert durch die GSB. Bildverstehen liefert eine solche Repräsentation als Ergebnis von Prozessen, die die Geometrie einer konkreten Szene anhand von Bildern rekonstruieren und die in der Szene sichtbaren Objekte erkennen. Sprache bezieht sich auf diese Repräsentation des Gegenständlichen mittels qualitativer Konzepte. Dies wurde für verschiedene Bestandteile von sprachlichen Aussagen über Szenen gezeigt, insbesondere für Bewegungsverben. Das Erkennen solcher Konzepte, z. B. das Erkennen von Ereignissen, kann als ein wissensbasierter Prozeß implementiert werden, bei dem die Wissensbasis die charakteristischen Eigenschaften der Konzepte in strukturierter und weitgehend deklarativer Form enthält. Dies ist für die Ereigniserkennung in NAOS durch eine Hierarchie von Ereignismodellen realisiert worden, die man sowohl als relationale als auch als logische Repräsentationen auffassen kann. Der Erkennungsprozeß ist je nach Sichtweise ein relationaler Vergleich oder eine Deduktion, beides paradigmatische Verfahren in der KI. Die Verknüpfung von Propositionen, die über Zeitintervallen gültig sind, hat allerdings eine grundsätzliche Erweiterung dieser Verfahren erforderlich gemacht. An die Stelle der „harten" Instantiierung muß eine „weiche" Akkumulation von Beschränkungen treten. Eine derartige Technik findet sich bisher noch nicht in gebräuchlichen KI-Sprachen.

Für die Verbalisierung eines Ereignisses bietet der Kasusrahmen eines Bewegungsverbs die geeignete Grundlage. Die internen Repräsentationen für die Komponenten eines Kasusrahmens gehen aus der Szenengeometrie und aus Ereignisinstanzen unmittelbar hervor, das Hauptproblem ist ihre sprachliche Referenzierung. Die in NAOS verwendeten Algorithmen wurden in einigem Detail erläutert. Sie versuchen, deutsche Sprachkonventionen zu modellieren, die z. B. den korrekten Gebrauch von örtlichen Präpositionen oder von Pronomen betreffen.

Um einen Hörer über die Szene zu informieren, muß außer stilistischen Vorgaben auch der Sprachverstehensprozeß beachtet werden. Sprechplanung in NAOS basiert auf dem Konzept der antizipierten Visualisierung, d. h. einer Vorausschau auf die Möglichkeit des Hörers, sich die Szene vorzustellen. Der Visualisierungsprozeß ist bisher nur ansatzweise realisiert. Er erfordert zusätzliche Wissensquellen, die quantitative Informationen ergänzen, wo qualitative sprachliche Aussagen keinen unmittelbaren Rückschluß auf die Szenengeometrie erlauben [17]. Mit dieser Problematik befassen sich weiterführende Arbeiten an NAOS.

Zu den bisher noch nicht vollständig realisierten Möglichkeiten von NAOS gehören auch erwartungsbasierte Aussagen. Sie finden sich in natürlichsprachlichen Beschreibungen in verschiedenen Ausprägungen, z. B. in Gestalt von negierten Aussagen („er hielt nicht an"). Um Erwartungen generieren zu können, bedarf es Wissens um typische Ereignisfolgen, z. B. wie Fahrzeuge sich an Ampeln verhalten. Eine Formalisierung derartigen Wissens in Form von Skripten sowie die dadurch ermöglichte Generierung von Erwartungen wird in [19] vorgestellt.

Literatur

1. Allen, J. F.: A plan based approach to speech act recognition. Ph. D. Thesis, Department of Computer Science, University of Toronto 1979
2. Allen, J. F.: A general model of action and time. Technical Report 97, University of Rochester, Rochester/NY 1981
3. Barrow, H. G., Ambler, A. P., Burstall, R. M.: Some techniques for recognizing structures in pictures. In: Watanabe, S. (Hrsg.) Frontiers of pattern recognition, pp. 1–29. New York: Academic Press 1972
4. Barwise, J., Perry, J.: Situations and attitudes. Bradford Books 1983
5. Block, N. (Hrsg.): Imagery. Cambridge/Mass.: MIT Press 1981
6. Bunke, H., Sagerer, G., Niemann, H.: Model based analysis of scintigraphic image sequences of the human heart. In: Huang, T. S. (Hrsg.) Image sequence processing and dynamic scene analysis, pp. 725–741. Berlin, Heidelberg, New York: Springer 1983
7. Busemann, S.: Surface transformations during the generation

of written german sentences. Report ANS-27, Forschungsstelle für Informationswissenschaft und Künstliche Intelligenz, Hamburg 1984
8. Carsten, I., Janson, T.: Überlegungen zur Evaluierung räumlicher Präpositionen anhand geometrischer Szenenbeschreibungen. Studienarbeit, FB Informatik, Universität Hamburg 1984
9. Fillmore, J.: Scenes-and-frames semantics. In: Zampolli, A. (Hrsg.) Universals in linguistic theory, pp. 55-81. Amsterdam: North-Holland 1977
10. Hoeppner, W., Christaller, T., Marburger, H., Morik, K., Nebel, M., O'Leary, M., Wahlster, W.: Beyond domain independence: Experience with the development of a german language access system to highly diverse background systems. Proc. International Joint Conference on Artificial Intelligence IJCAI-83, pp. 588-594 (1983)
11. Jameson, A., Wahlster, W.: User modelling in anaphora generation: Ellipsis and definite description. Proc. First European Conference on Artificial Intelligence ECAI-82, pp. 222-227 (1982)
12. Miller, G. A.: English verbs of motion: A case study in semantics and lexical memory. In: Melton, A. W., Martin, E., (Hrsg.) Coding processes in human memory, pp. 335-372. Washington/DC: V. H. Winston 1972
13. Neumann, B.: Towards natural language description of real-world image sequences. GI 12. Jahrestagung, Informatik Fachberichte 57, pp. 349-358. Berlin, Heidelberg, New York: Springer 1982
14. Neumann, B.: Natural language description of time-varying scenes. In: Waltz, D. (Hrsg.). Advances in natural language processes, Vol. 1. Hillsdale NJ: Lawrence Erlbaum (auch erschienen als FBI-HH-B-105/84, Fachbereich Informatik, Universität Hamburg 1984)
15. Neumann, B., Novak, H.-J., Event models for recognition and natural language description of events in real-world image sequences, Proc. International Joint Conference on Artificial Intelligence IJCAI-83, pp. 724-726 (1983)
16. Novak, H.-J.: A relational matching strategy for temporal event recognition. In: Laubsch, J. (Hrsg.) GWAI-84, pp. 109-118. Informatik Fachberichte 103, Berlin, Heidelberg, New York: Springer 1985
17. Novak, H.-J., Neumann, B.: Szenenbeschreibung und Imagination in NAOS. In: Rollinger, C.-R. (Hrsg.) Probleme des (Text-)Verstehens, pp. 192-206. Tübingen: Niemeyer 1984 (auch erschienen als IfI-HH-M-123/84, Fachbereich Informatik, Universität Hamburg 1984)
18. Okada, N.: Conceptual taxonomy of japanese verbs for understanding natural language and picture patterns. Proc. Conference on Computational Linguistics COLING-80, pp. 127-135 (1982)
19. Retz-Schmidt, G.: Script-based generation and evaluation of expectations in traffic scenes. Mitteilung FBI-HH-M-136/85, Fachbereich Informatik, Universität Hamburg 1985
20. Schank, R. C.: Identification of conceptualizations underlying natural language. In: Schank, R. C., Colby, K. M. (Hrsg.) Computer models of thought and language, pp. 187-247. New York: W. A. Freeman 1973
21. Tsotsos, J. K.: A framework for visual motion understanding. TR CSRG-114, University of Toronto 1980

Eingegangen am 3. Januar 1986
Angenommen am 21. Februar 1986

Bernd Neumann (geb. 1943 in Lüneburg) studierte Elektrotechnik in Berlin und Darmstadt. Nach dem Diplom im Jahre 1967 ermöglichte ihm ein ESRO/NASA-Stipendium das Studium von Informationstheorie und Nachrichtentechnik am MIT, Cambridge. Dort erhielt er 1968 den M. S. und 1971 den Ph. D. Während der Zeit am MIT wuchs das Interesse an digitaler Informationsverarbeitung und Künstlicher Intelligenz. Nach seiner Rückkehr wurde er Dozent für Informatik an der Universität Hamburg und Mitglied der Forschungsgruppe „Kognitive Systeme", deren Leiter er heute ist. Seine Forschungsschwerpunkte sind wissensbasierte Systeme, Künstliche Intelligenz und Bildverarbeitung.

Hans-Joachim Novak (geb. 1953) ist wiss. Mitarbeiter am Fachbereich Informatik der Universität Hamburg und arbeitet im Bereich Kognitive Systeme. Der Schwerpunkt seiner Arbeiten liegt in der Untersuchung und Algorithmisierung kognitiver Entscheidungsverfahren, mit deren Hilfe visuell wahrgenommene Daten natürlichsprachlich beschrieben werden können. Herr Novak studierte Informatik und Linguistik und erhielt 1982 sein Diplom in Informatik von der Universität Hamburg.