

Natural Language Access to Visual Data: Dealing with Space and Movement

Elisabeth André, Gerd Herzog,* Thomas Rist

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

Abstract

Combining vision/image understanding and natural language generation the following issues arise: *How can we extract information concerning time, space and movement from visual data and how should the information be represented to permit the generation of natural language descriptions?* In this paper, we will report on practical experience gained in the project *Vitra* (VIsual TRAnslator). In order to design an interface between image understanding and natural language systems, we have examined different domains of discourse and communicative situations. Two natural language systems are presented which are combined with a concrete vision system. To give an impression which input data are supplied to the NL systems, the levels of image analysis will be briefly sketched. By means of the four orientation-dependent relations *right*, *left*, *in front of* and *behind*, we demonstrate how a purely geometrical description can be transformed into a propositional description of the spatial arrangement. After that, we present our approach to event recognition. In contrast to previous approaches, we don't start from a completely analyzed image sequence. Rather, events are to be represented in such a way that they can be simultaneously recognized and described in natural language as the scene progresses.

This paper appeared as: Report 63, Universität des Saarlandes, SFB 314 (VITRA), November 1989. It has been presented at the 1st Workshop on Logical Semantics of Time, Space and Movement in Natural Language, Toulouse, France, 1989.

1 Motivation

Connecting vision and natural language systems is both of practical and of theoretical interest. The visual data supplied by a vision system are often unreadable for humans. A great practical advantage of natural language is that it offers the possibility to condense information and to present it in a more comprehensible manner. Since most NL systems only consider the relationship between expressions of a knowledge representation language and natural language expressions, the underlying semantics remains incomplete. The access to visual data, however, allows the definition of a referential semantics that is perceptually anchored.

Although image understanding and natural language processing constitute two major areas of research within AI, there are only a few approaches to bring them together. For the description of static scenes algorithms have been developed which describe spatial arrangements of objects by means of spatial prepositions (cf. Fürnsinn et al. [1984], Hußmann and Scheffe [1984], Carsten and Janson [1985] and Bajcsy et al. [1985]). The recognition and verbalization of motion concepts is focused when describing dynamic scenes (cf. Badler [1975], Okada [1979], Wahlster et al. [1983] and Neumann and Novak [1986]). All these approaches have in common, that they start from a complete geometrical description of the scene.

Natural language access to visual data also forms the research background for the *Vitra* project, which is concerned with the development of knowledge-based systems for the integration of vision and natural language processing. In this paper, the following issues will be investigated:

- connecting vision systems with natural language access systems
- defining a computational semantics for spatial prepositions and verbs of motion and action which is based on visual data

2 The Systems CITYTOUR and SOCCER

To verify the domain-independence of the developed concepts and methods, two communicative situations and different domains of discourse have been investigated.

The system *Citytour* (cf. André et al. [1987] and Schirra et al. [1987]) answers natural language questions about spatial relations between objects in a scene. Fig. 1 shows an example scene: a map of the city of Saarbrücken from a bird's eye view with dynamic and static objects. Other scenes are a map of the University of Saarbrücken campus and a traffic scene in Karlsruhe. When answering questions, *Citytour* can take into account the current position of the dialog partners who are assumed to be at the same place in the scene.

The system *Soccer* (cf. André et al. [1988]) describes short image sequences of soccer games. The listener is assumed not to be watching the scene, but to have prototypical knowledge about the static background. In contrast to the approaches men-

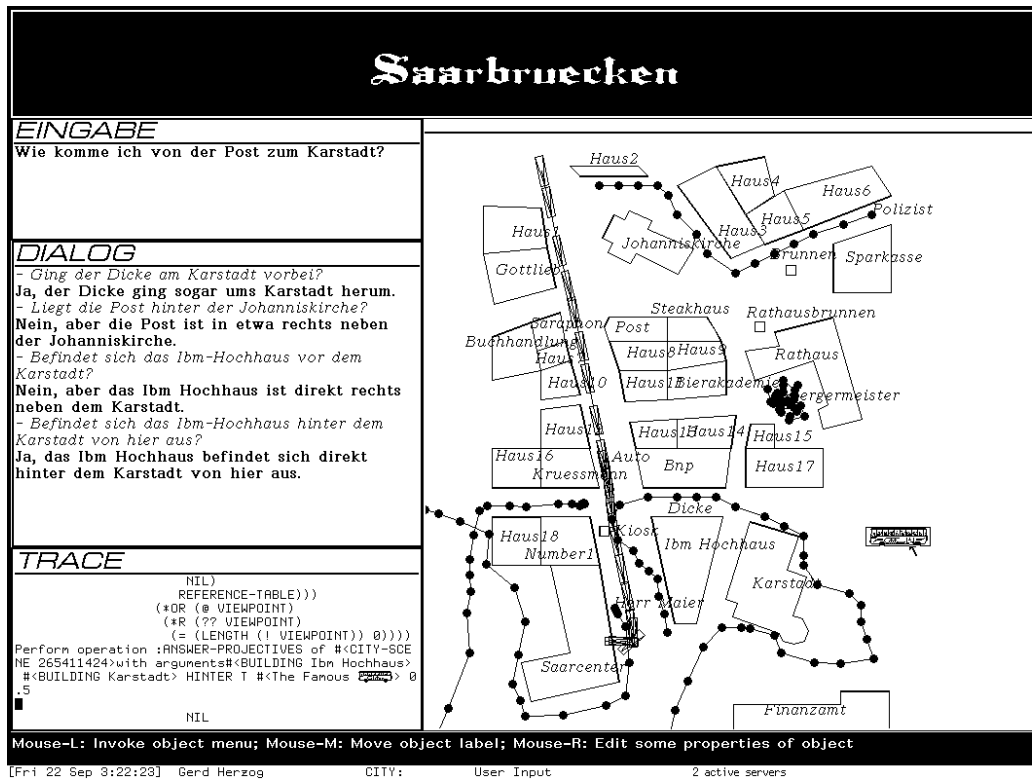


Figure 1: The basic windows of the system *Citytour*

tioned in the last section, *Soccer* doesn't require that the geometrical scene data are supplied all at once. Rather, the system should be able to process incoming information simultaneously as the scene progresses (cf. Herzog et al. [1989]). Fig. 2 shows the static background and the trajectories of a typical scene. While the scene is replayed in the graphics window, its natural language description appears in the output window.

3 Image Sequence Analysis

The main task of computer vision is the construction of a symbolic computer-internal description of a scene from images. In the case of image sequence analysis, the focus lies on the detection and interpretation of changes which are caused by motion. In the narrow sense, the intended output of a vision system would be an explicit, meaningful description of physical objects (cf. Ballard and Brown [1982]). This kind of processing which results in a geometrical representation of a scene is called low-level analysis. One goal of approaches towards the integration of computer vision and natural language processing is to extend the scope of image analysis beyond the level of object recognition. Natural language access to vision systems requires processes which lead

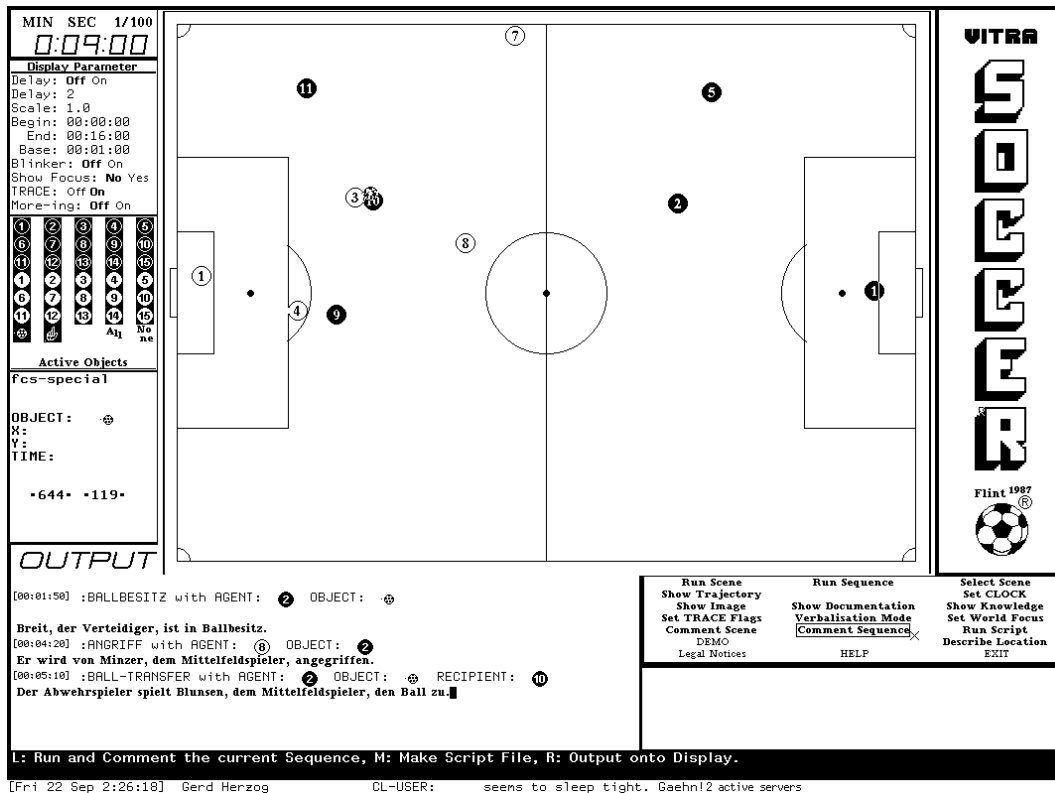


Figure 2: The basic windows of the system *Soccer*

to conceptual units of a higher level of abstraction. This so-called high-level analysis allows the description of a scene in terms of spatial relations and interesting events. Fig. 3 summarizes these two separate levels of image analysis.

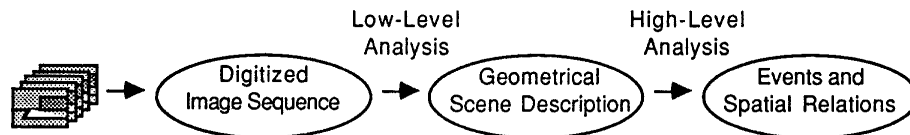


Figure 3: Low-level and high-level vision

The following section deals with low-level analysis in more detail. In order to show how information concerning moving objects can be extracted from a sequence of digitized images, we present the vision system *Actions* which is connected to the systems *Citytour* and *Soccer*. After that, the interface between low-level analysis and high-level analysis, i.e. between *Actions* and the *Vitra* systems, is motivated and described.

3.1 The System ACTIONS

The system *Actions*¹, which has been developed at the 'Institut für Informations- und Datenverarbeitung (IITB) der Fraunhofergesellschaft' in Karlsruhe, constitutes an approach for the automatic segmentation and cueing of moving objects in image sequences. The main goal for the work on *Actions* is the development of robust, generally useful methods for the analysis of real world scenes. Results achieved so far are summarized in Sung and Zimmermann [1986] und Sung [1988].

Fig. 4 gives an overview of the different steps of processing in the *Actions* system. The image sequences that are currently being investigated include snapshots of an intersection in Karlsruhe, made from a building about 35m high, and in addition material from a game of the German professional soccer league. In both cases, the video tapes have been recorded with a stationary, monocular camera. At present, sequences of up to 3300 selected frames (132 seconds) are digitized (512 by 512 pixels, 8-bit grayscale) and processed by the *Actions* system.

Moving objects are separated from the stationary ones by computing and analyzing displacement vector fields. The computation of the displacement vectors is based on characteristic local gray value distributions, so-called *features*. Features are determined by comparing each pixel with a fixed number (in our case 8) of surrounding pixels. Depending on the number of comparative points with a lower or higher gray value, the pixel considered will be assigned to the corresponding class. Connected points of the same class can be combined to *blobs*. Only those blobs that belong to one of the two extreme classes, i.e. those representing local maxima and minima, are considered for subsequent processing.

For the determination of the displacement vectors, the center of gravity of each blob is cued through several frames in order to rule out errors caused by accidental variation of feature positions. This leads to local displacement vectors from the n^{th} to the $(n + 4)^{th}$ frame. The obtained displacement vectors are clustered with respect to absolute value, direction and position within the picture domain. For each cluster, a frame is drawn parallel and perpendicular to the average displacement through the most distant vectors. These frames can be seen as candidates for the picture of moving rigid bodies. The geometrical center point of each frame serves as representative for the moving object in the picture domain.

The correspondence chains of the frame positions represent the trajectories of the clusters of vectors and thus of the object candidates in the picture domain. Two frames in consecutive images correspond to each other if they have almost the same direction of motion and if the distance between them is less then two times the absolute value of the average displacement vector. In the case of ambiguity the candidate with the lowest distance is taken. Under consideration of the camera position and the geometry of the static background, the coordinate data of the resulting correspondence chains are retransformed into scene coordinates and put together in the (partial) geometrical scene description. The automatic classification of object candidates and the identification of

¹Automatic 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

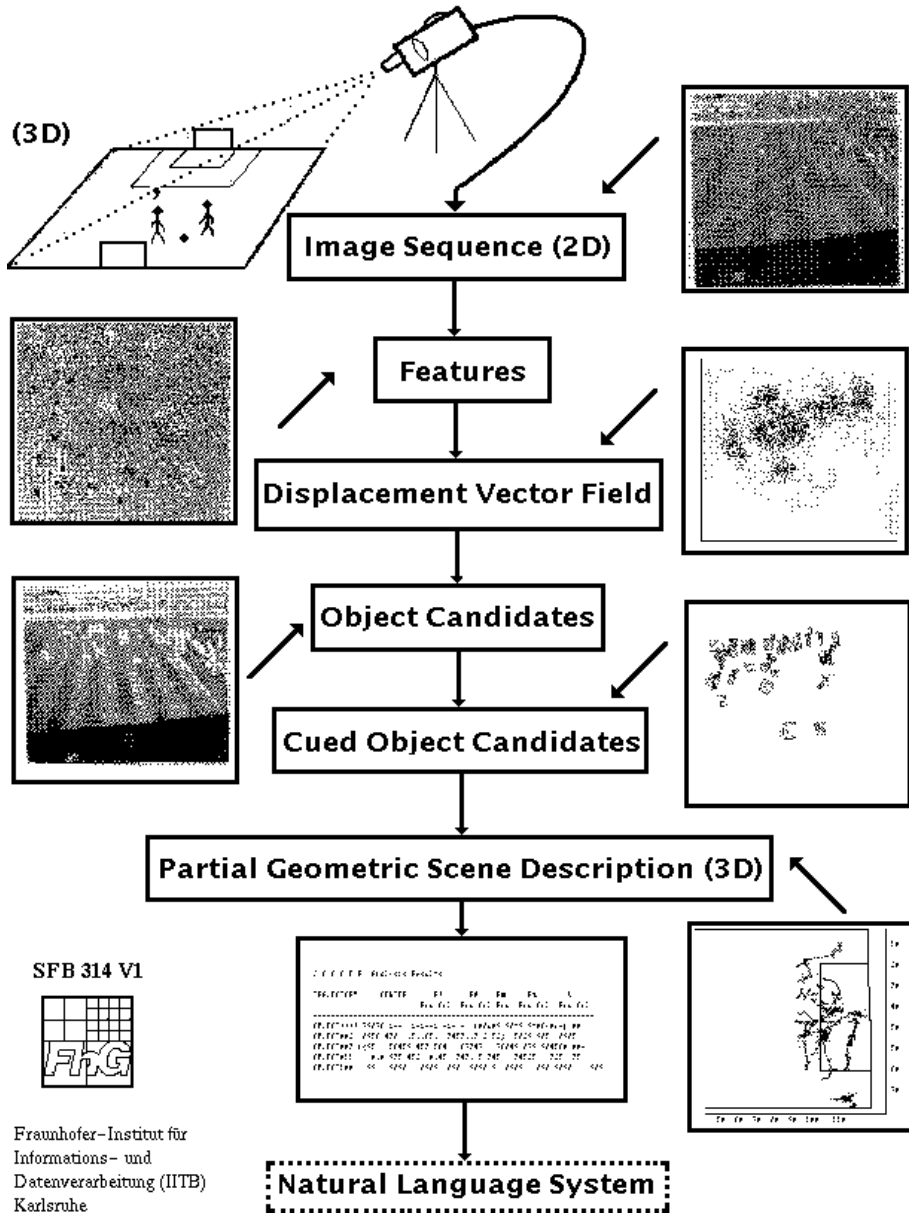


Figure 4: Processing phases in *Actions*

previously known objects is not yet possible in the *Actions* system; at present this is still done interactively.

3.2 The Interface between Low-Level and High-Level Analysis

In Neumann [1984], the *geometrical scene description* (GSD) has been introduced as a representation for the output of the low-level vision process. The aim of this description format is to represent the original image sequence completely and without loss of information, i.e. the data in the GSD suffice (in principle) to reconstruct the raw images. The geometrical scene description contains:

- for each frame of the image sequence:
 - instant of time
 - visible objects
 - viewpoint
 - illumination data
- for each object:
 - identity (i.e. frame to frame correspondence)
 - 3D-position and orientation in world coordinates in each frame
 - 3D-shape and surface characteristics (e.g. color)
 - class membership and possibly identity with respect to *a priori* knowledge (and thus additional properties that can be verbalized, e.g. names)

Please note that for the classification and identification of previously known objects the purely visual information is not sufficient; in this case, additional knowledge sources are required.

The concept of geometrical scene description constitutes an idealized interface between low-level vision and a natural language access system. In applications like *Naos*, a system for the description of street scenes developed by Neumann and Novak, the GSD is restricted according to the practical needs. In the *Vitra* systems, we do not consider the viewpoint, illumination data, and the complete 3D-shape of the objects, for example. The information concerning the static background of the scene is not supplied by the vision system but instead provided as an instantiated model of the considered part of the visual world. At present such restrictions are characteristic for all approaches since we are still far from a universally applicable AI system capable of completely analyzing arbitrary sequence of images.

Also, for reasons of efficiency, it is better not to start from a complete geometrical scene description for the computation of the applicability of spatial prepositions or motion verbs, but to use idealisations, e.g. those suggested by Herskovits (cf. Herskovits [1986]). Herskovits introduces the term *geometrical description* in order to

represent those object attributes that are relevant for the semantics of spatial relations. Formally, geometric descriptions are functions which map an object onto a situation-specific geometrical representative. Instead of using expensive 3D-reconstructions one is typically content with representations like the objects centroid or the contour of a projection of the object, approximated by a polygon.

The contents of the GSD are accessed in a functional way. A localisation function can be defined and used to determine the 3D-world-coordinates of an object or its idealized geometrical representative at a certain instant of time. The data of the GSD can be used to deduce physical quantities like distance between objects, speed and acceleration of moving objects. The transition from such physical quantities to spatial concepts and motion concepts happens by predicates. These predicates are defined with respect to the natural language use of spatial prepositions and motion verbs.

4 A Computational Semantics for Spatial Prepositions

The semantic analysis of spatial prepositions leads to the term spatial relation as a target-language independent meaning concept. Conversely, the definition and representation of the semantics of spatial relations is an essential condition for the synthesis of spatial reference expressions in natural language (cf. fig. 5). Spatial relations can

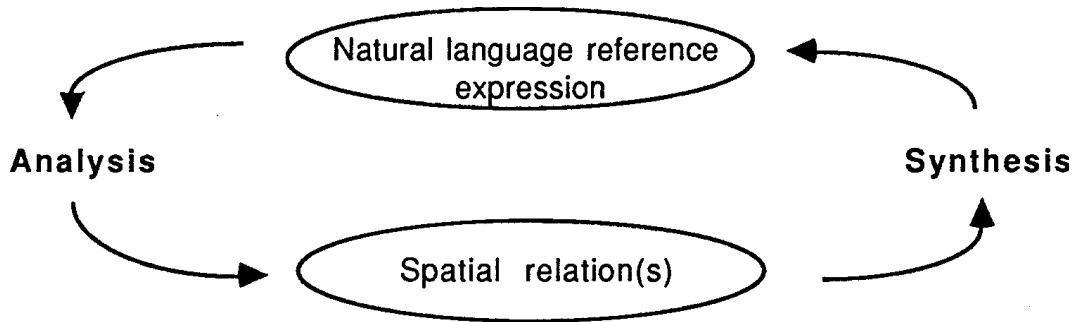


Figure 5: Representing locative expressions by spatial relations

be defined by specifying conditions for object configurations, such as the distance between objects or the relative position of objects. In this sense, a spatial relation characterizes a class of object configurations. Strictly speaking, spatial relations represent relations between spatial entities. It is, however, common practice to refer to these entities via the corresponding objects. Spatial relations are represented by relation tuples of the following form:

$$(rel\text{-}name\ subject\ ref - ob_1 \dots ref - ob_n \{orientation\})$$

The first argument of a tuple denotes the spatial relation; the second argument is called subject. It is the object that is to be located (if forced, according to an orientation) in relation to one or more reference objects.

4.1 Applicability of Relation Tuples

We say that a relation tuple is applicable if it can be used to characterize an object configuration. To determine the applicability of a relation tuple, we assign to each relation tuple an area of applicability and test whether the subject is located within this area. In general, the computation of applicability areas is a non-trivial task. Besides the size, the orientation and the shape of objects, it has to be taken into account that nearby objects can lead to deformations of the area of applicability.

In many cases, it is insufficient to distinguish only between relations that are applicable and those that are not. One solution to this problem is to introduce a measure of degrees of applicability which expresses the extent to which a spatial relation is applicable. As suggested among others in Hanßmann [1980], we represent degrees of applicability internally as values from the real interval $[0, 1]$. The value 0 indicates that a relation is not applicable, the value 1 that it is fully applicable, and values between 0 and 1 that it is more or less applicable. As shown in fig. 6, degrees of applicability correspond to partitions of the applicability area. The advantage of degrees of appli-

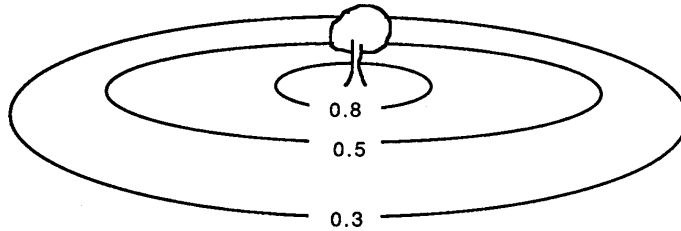


Figure 6: Degrees of applicability of the relation at

applicability becomes obvious when generating natural language scene descriptions. Since different degrees of applicability can be expressed by so called linguistic hedges (cf. Lakoff [1973]), such as *directly* or *more or less*, more exact scene descriptions are possible. Furthermore, if an object configuration can be described by several spatial relations, the degree of applicability is used to select the most appropriate preposition for verbalization.

In the next section, we will demonstrate how the systems *Citytour* and *Soccer* determine degrees of applicability for the orientation-dependent relations *right*, *left*, *front* and *behind* in the two-dimensional space.

4.2 Computing the Applicability of Orientation-Dependent Relations

When analyzing orientation-dependent prepositions, it becomes obvious that the orientation of the space depends on the use of the preposition. In the following, the common distinction between the intrinsic and extrinsic use of spatial prepositions (e.g., cf. Wunderlich [1985]) is transferred to spatial relations. We speak of *intrinsic use* if the

orientation is given by an inherent organization of the reference object. An inherent organization of an object can be given by perceptual organs (of humans and animals), by the characteristic direction of movement (e.g. of vehicles) and by functional properties (e.g. of the two goals in the soccer domain). Further aspects are discussed in Miller and Johnson-Laird [1976], Sondheimer [1976], Vandeloise [1984] and Wunderlich [1985]. As Vandeloise points out, conflicts can occur when two or more criteria contradict each other. For example, think of a crab whose direction of movement doesn't coincide with the line of sight. If the orientation is given by contextual factors, such as the accessibility of the reference objects or objects in its vicinity, we speak of *extrinsic use*. An example is *'From here, the post office is behind the church'*. In this example, the orientation is determined by the position of a probably imaginary observer. If the observer coincides with the speaker or hearer, we speak of *deictic use*.

In order to compute the applicability of an orientation-dependent relation, the following steps have to be carried out:

1. Determining the orientation
2. Computing the applicability of the relation tuple

An orientation of a three-dimensional space can be represented by a set of orthogonal vectors, a front-back (a_{FB}), a left-right- (a_{LR}) and an above-below-vector (a_{AB}). Since *Citytour* and *Soccer* operate on a two-dimensional representation of the scene, we don't consider an above-below-vector. The front-back-vector is determined as follows:

1. If the orientation is given by an object's inherent sides, then the front-back-vector is orthogonal to the front side and points out of the object. The left-right-vector is determined by the front-back-vector because the two vectors form an orthogonal right-handed-system.
2. If the object is localized with respect to an observer and the reference object and the observer coincide, then the orientation of the observer is transferred to the reference object (cf. fig. 7 (a)). The front-back-vector and the left-right-vector form an orthogonal left-handed-system. If the observer and the reference object are spatially separated, the orientation follows from the mirror principle (cf. fig. 7 (b)). In this case, the front-back-vector and the left-right-vector form an orthogonal right-handed-system.
3. If an orientation is induced by an object's actual movement, the front-back-vector coincides with the direction of movement. The left-right-vector is determined as in 1.

The reference object and the context essentially determine how an orientation-dependent relation is used. A strategy to figure out the appropriate use is proposed in André [1988].

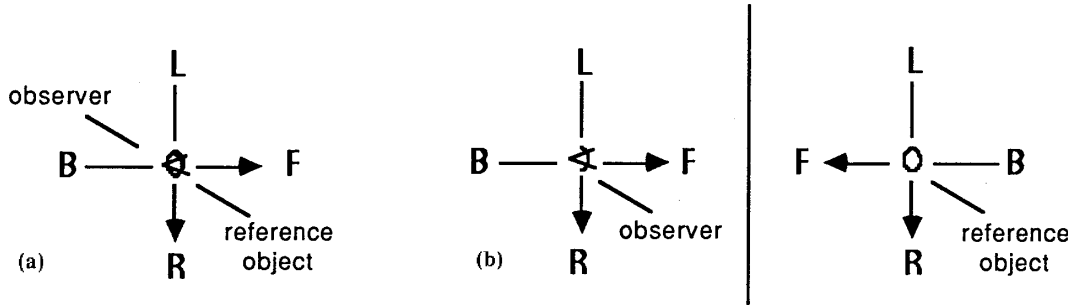


Figure 7: Reference frame selection based on (a) the coincidence principle and on (b) the mirror principle

For a given orientation, Herskovits [1980] proposes a method for computing the applicability of orientation-dependent relations in case both the subject and the reference object are represented by a point:

1. Construct a coordinate system whose origin coincides with the position of the reference object and whose axes are a_{FB} and a_{LR} .
2. Determine the location of the subject relative to this coordinate system.

An extension of this method has been developed for the system *Citytour* (cf. André et al. [1987]). In contrast to Herskovits, we also consider polygonal object representations and compute degrees of applicability. Both methods, which have been designed for the two-dimensional space, partition the area around the reference object into four half-planes. With each half-plane, one of the relations `front`, `back`, `right` and `left` is associated. If the reference object is represented by a polygon, our algorithm determines a delineative rectangle which serves as an extended origin. It is the smallest rectangle whose sides are parallel to the vectors a_{FB} and a_{LR} . The axes of the coordinate system have degenerated into bands. In the case of intrinsic use, the delineative rectangle is determined by the prominent front (cf. fig. 8). If the object is located with respect to an observer, the delineative rectangle is oriented by the observer's position (cf. fig. 8).

In Herskovits' system of reference, a relation is only applicable, if the subject is located exactly on one of the four axis sections. In our case, a relation is applicable if the subject is within the corresponding half-plane. As mentioned above, different degrees of applicability can be determined by partitioning of the half-planes into regions of the same degree of applicability. In the current version of the system, this partition depends on the size of the reference object and the shape of the delineative rectangle. Other aspects, such as the exact shape of the reference object or objects in its vicinity, are neglected.

In fig. 9, `Obj2` is located on the right-axis of the coordinate system. Thus, the applicability of the relation tuple (`rel-right` `Obj2` `Obj1` $\{a_{FB}, a_{LR}\}$) is accord-

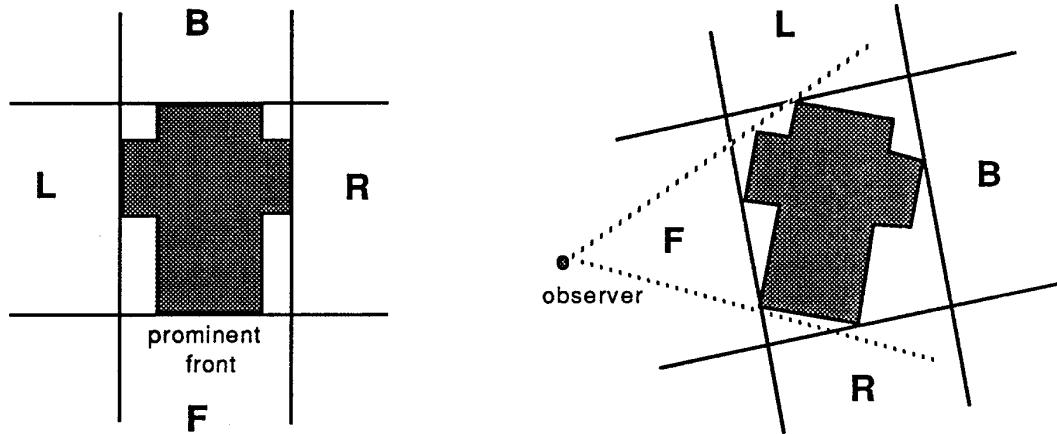


Figure 8: Delineative rectangle oriented by the prominent front and by the observer's position

ingly high. A lower value is assigned to the tuple $(rel-right\ Ob3\ Ob1\ \{a_{FB}, a_{LR}\})$. For Ob4 and Ob5, the relation tuple is not applicable.

5 A Computational Semantics for Verbs of Motion and Action

When analyzing image sequences instead of single frames, we can also extract spatio-temporal concepts from the geometrical scene description. These conceptual units, which we will call *events*, serve for the symbolic abstraction of the temporal aspects of the scene. With respect to the natural language description of image sequences events are meant to capture the meaning of motion and action verbs; i.e. in our case events are those changes in the world people usually talk about (cf. Miller and Johnson-Laird [1976]).

5.1 Simultaneous versus Retrospective Interpretation

Besides the question of which concept are to be extracted out of the geometrical scene description, it is decisive how the recognition process is realized. The aim of previous attempts at connecting vision systems and natural language systems has been to provide a retrospective description of the analyzed image sequence. In the systems *Naos* (cf. Neumann and Novak [1986]) and *Epex* (cf. Walter [1989]) for the interpretation of traffic scenes an *a posteriori* strategy is used, which requires a complete geometrical scene description as soon as the analysis process starts. As opposed to this, in the system *Alven* (cf. Tsotsos [1981]), which treats left ventricular heartwall motion, scene analysis happens data-driven and successively. But even in this case, recognized events

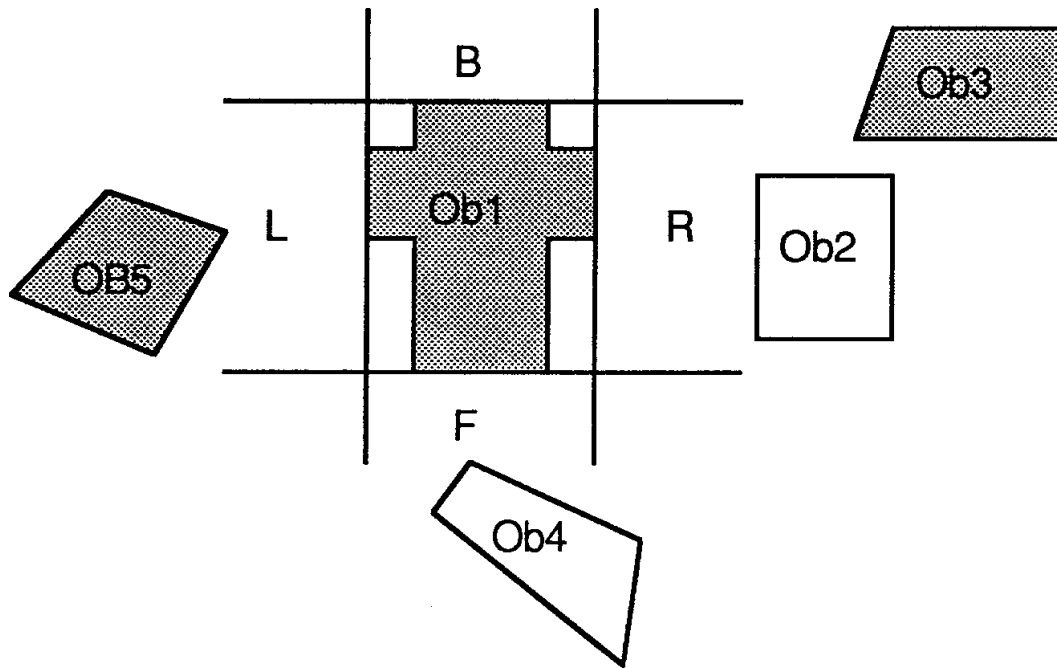


Figure 9: Degrees of applicability of orientation-dependent relations

are to be supplied only after the scene under discussion has been completely analyzed.

A totally new task results if events are to be recognized simultaneously as they occur in the scene. With respect to the generation of simultaneous scene descriptions in natural language, the following problem becomes obvious: If the description is to be focused on what is currently happening, it is very often necessary to verbalize events even while they are currently happening and not yet completed. Examples for this would be the description of an overtaking maneuver just taking place in a traffic scene or the portrayal of an attack in a soccer game during a live broadcast of a radio reporter. In general this problem always occurs if further reactions of an image understanding system are to be based on simultaneous recognition. One might think of a robot capable of visually *'perceiving'* its environment and compelled to be able to react immediately on the stimuli. Here the following question has to be asked: *How can partly recognized events be represented in order to make them available for further processing?*

The specific requirements for the modelling of events with respect to this task can not even be met by formalisms like the temporal logics in Allen [1984] and McDermott [1982] since they only distinguish between events that have occurred and those that have not. In order to allow for a more detailed description of the occurring of an event, it seems reasonable to consider different phases of an event, namely the beginning, the progression and the termination of the occurrence. For this reason, we introduce additional predicates:

TRIGGER (t_i event) for the beginning,
PROCEED (t_i event) for the progression,
STOP (t_i event) for the termination and
SUCCEED (t_i event) for the continuation of an occurring event.

The **SUCCEED**-predicate is used to model events that have already been completely recognized, but still continue to occur. This special group includes for example event concepts that correspond to durative motion verbs, such as *drive*, *run* or *walk*.

Contrary to the interval-based predicates **OCCUR** in Allen's and **OCC** in McDermott's approach the predicates introduced here apply to discrete instances of time. They enable us to characterize the state of an event at a particular moment. To clarify this, consider an overtaking maneuver in a traffic scene. Fig. 10 shows four significant frames from a real world image sequence. Each frame corresponds to one of the discrete instances of time T_1 to T_4 . At the moment T_1 , the car is approaching the delivery

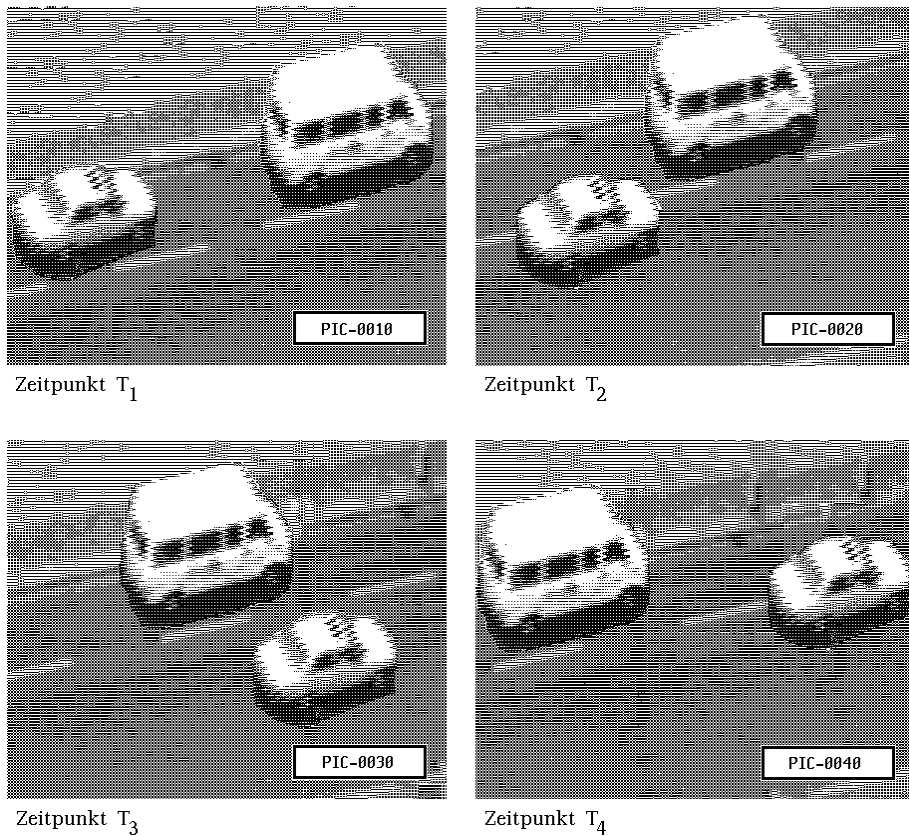


Figure 10: An 'overtake' event in a traffic scene

van. At the moment T_2 , the car swings out; it starts overtaking the delivery van. The fact that the event has started is formally represented by $\text{TRIGGER}(T_2$ (overtake

CAR1 VAN1)). At the moment T_4 , the car swings into line in front of the delivery van and thus finishes the overtaking. The event (overtake CAR1 VAN1) has now completely occurred and the following predication holds: STOP(T_4 (overtake CAR1 VAN1)). Between T_2 and T_4 the overtake event can be observed in the scene. Hence, the event is just occurring, but is not yet completely recognized during this interval. If, for example, the car would turn into a side street at the moment T_3 one could not speak of an overtaking maneuver any longer. The predicate PROCEED is used to describe events which have not yet been completely recognized. The predication PROCEED(t_i (overtake CAR1 VAN1)) holds in the situation shown in our example for all t_i with $T_2 < t_i < T_4$. One example for the continuation of an event with respect to the image sequence shown above would be the fact that CAR1 is moving during the whole interval. For all t_i with $T_1 \leq t_i \leq T_4$ the predication SUCCEED(t_i (move CAR1)) holds.

The predicates just introduced allow a more detailed modelling of events. In particular, interval-based descriptions of events, like the fact that an overtake event occurred in the interval $[T_2 T_4]$, can be inferred from such a finer description.

5.2 Event Models for Incremental Event Recognition

In analogy to object models, events are described conceptually using so-called *event models* (cf. Neumann [1984]). Event models represent *a priori* knowledge concerning typical occurrences in a scene, especially about interesting object motions. They constitute the connecting link between the event predicates introduced above and the event concepts that are to be inferred from the geometrical scene description. It is an important principle to order event models into a conceptual hierarchy consisting of different levels of abstraction. On the lowest level concepts such as exist or move can be found, which are defined directly through the geometrical data. These concepts can then be utilized in order to define more complex events, such as overtake. Recognized events are instantiations of the respective generic event models; subsequently they will also be called event instances.

Simultaneous recognition of events, can only be carried out by means of an incremental recognition strategy in which detection is controlled by temporal relations between subevents and in which events are recognized 'stepwise', as they progress. Event instances must be explicitly represented in the knowledge base of the system right from the moment they are first noticed. The methodology for the modelling of events developed in the framework of the *Soccer* system (cf. Rist et al. [1987], Herzog and Rist [1988]) was designed to take into account these demands. An event model in *Soccer* includes:

- roles

Roles are existentially quantified variables for the objects involved in an event. In the event instances these roles are filled with the respective designations for real objects in the scene.

- role restrictions

An additional specification is used to restrict the set of possible role fillers for the instantiation of an event model. Type restrictions are obligatory; they determine the object class to which role fillers have to belong to. In addition, role restrictions can be used to formulate conditions that refer to dependencies between the single role fillers. An example for such a restriction is: *'If the filler of role A has the attribute p, then the filler of role B has to possess the attribute q'*.

- course diagram

The core of an event model is its course diagram. It serves for specifying the prototypical course of an event.

The course diagram of an event specifies the sub-concepts and situational context which characterize the instances of the particular event model. Formally, course diagrams are defined as finite labeled directed graphs. The basic idea is to represent the temporal aspects in such a way that the recognition of an event corresponds to traversing the according course diagram. Such a traversal then happens step by step with respect to a certain measure, i.e. incrementally. To demonstrate this, consider the concept `running_pass` as an example. It describes a situation in which a player passes the ball to another member of his team who must be running at the same time. Using Allen's formalism (cf. Allen [1984]) this concept could be defined as:

```
OCCUR(timeinterval1 (running_pass PLAYER1 BALL PLAYER2))
<==>
  EXIST timeinterval2
    DURING(timeinterval2 timeinterval1)
    OCCUR(timeinterval2 (run PLAYER2))
    OCCUR(timeinterval1 (pass PLAYER1 BALL PLAYER2))
```

Fig. 11 shows the corresponding course diagram; it can be derived by projecting the interval-based validity conditions onto a discrete time axis (cf. Herzog and Rist [1988]). The edges of the course diagram are labelled with type markers which are used for the definition of the elementary event predicates. Event recognition is based on a data-driven bottom-up-strategy. Each recognition cycle starts at the lowest level of the event hierarchy: first, the traversal of course diagrams corresponding to basic events is attempted; later, more complex event instances can look at those lower levels to verify the existence of their necessary subevents.

6 Summary

When combining our natural language systems *Citytour* and *Soccer* with the vision system *Actions*, we distinguish between several representational levels (cf. fig. 12). The processes on the sensory level take a scene as input and provide a geometrical

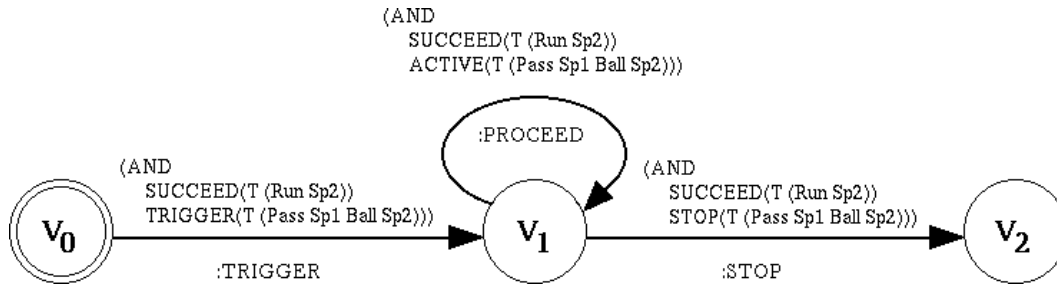


Figure 11: Course diagram for running_pass

scene description. Starting from this, the processes on the cognitive level extract spatial relations and event concepts. These conceptual structures bridge the gap between visual data and natural language concepts, such as spatial prepositions and motion verbs. They are passed on to the processes on the linguistic level which transform them into natural language utterances. In the sense of a referential semantics, we have established explicit links between sensory data and natural language expressions. Eventually, the meaning of such expressions is perceptually anchored.

The approach pursued in the soccer domain emphasizes concurrent image sequence analysis and natural language processing. This distinguishing feature is an important prerequisite for real-time performance, which is one long-term goal of our research (cf. Herzog et al. [1989] and Nagel [1988]). In order to improve the capabilities of the vision and natural language components, current research concentrates on the model-based 3D-reconstruction of non-rigid bodies (cf. Rohr [1989]), the recognition and description of intentions and plans behind the observed actions and movements (cf. Retz-Schmidt [1988]), and the development of a pictorial partner model representing the listener's temporal and spatial conceptualization of the scene (cf. Schirra [1989]).

7 Technical Notes

Image processing has been done with a VTE Digital Video Disk and a VAX-11/780, programmed in Pascal. The current versions of the *Vitra* systems have been implemented in Commonlisp and Flavors on Symbolics 3600 and 3640 Lispmachines running Release 7.1. One of the machines is equipped with an OP36-C108 Color Option, which includes an additional high-resolution color screen. Special versions of *Citytour* and *Soccer* which utilize the color screen are available. TCP/IP is used to connect the Symbolics machines to a Siemens 7.570 mainframe that serves as a gateway to the German Research Net (DFN) and the VAX in Karlsruhe.

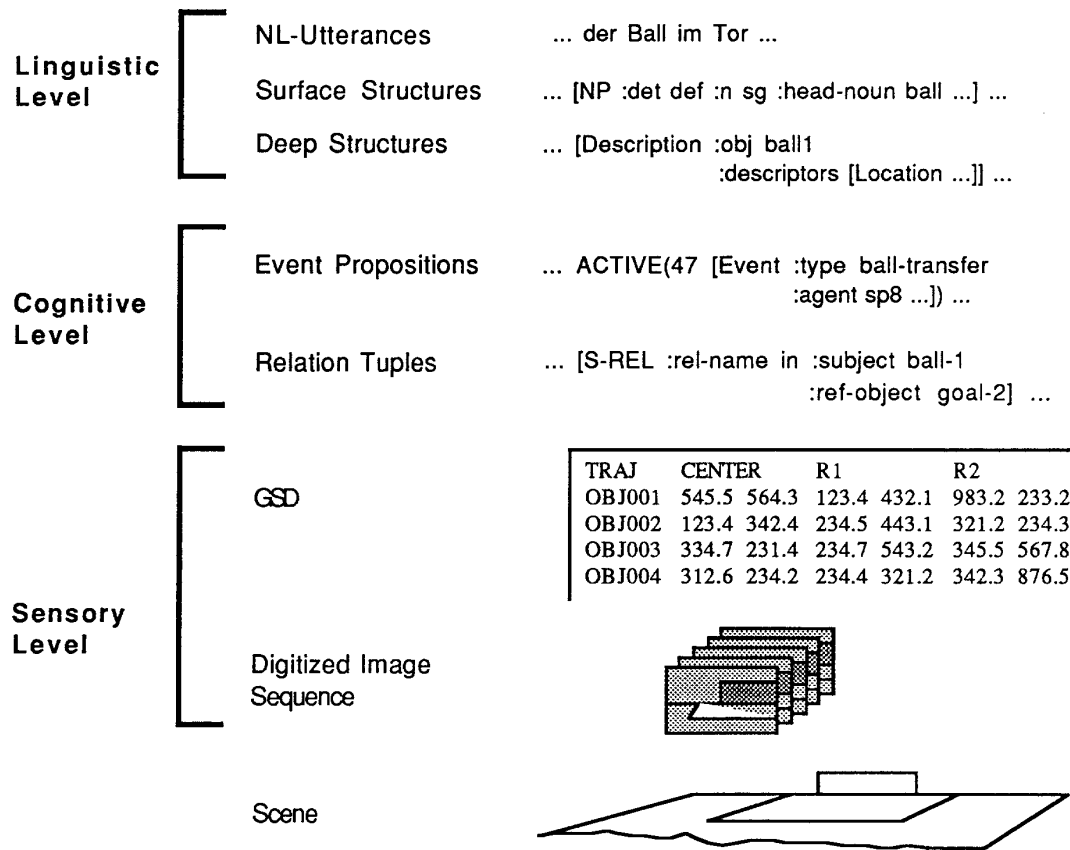


Figure 12: Levels of representation

Acknowledgements

We would like to thank our colleagues at the 'Institut für Informations- und Datenverarbeitung (IITB) der Fraunhofergesellschaft' in Karlsruhe, for their helpful comments on an earlier version of this paper and we gratefully acknowledge the permission to use the figure which shows how the *Actions* system works. The work described here was partly supported by the 'Sonderforschungsbereich 314 der Deutschen Forschungsgemeinschaft, Künstliche Intelligenz und wissensbasierte Systeme, projekt N2: VITRA (VIsual TRAnslator)'.

References

- 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, eds., *Artificial Intelligence II: Methodology, Systems, Applications*, pp. 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*, pp. 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*, pp. 919–921, Los Angeles, CA, 1985.
- D. H. Ballard, C. M. Brown.** *Computer Vision*. Prentice-Hall, Englewood Cliffs, NJ, 1982.
- I. Carsten, T. Janson.** *Verfahren zur Evaluierung räumlicher Präpositionen anhand geometrischer Szenenbeschreibungen*. Master's thesis, 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, ed., *Probleme*

- des (Text-) Verstehens, Ansätze der künstlichen Intelligenz*, pp. 172–184, Niemeyer, Tübingen, 1984.
- 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*, pp. 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.
- G. Herzog, C.-K. Sung, E. André, W. Enkelmann, H.-H. Nagel, T. Rist, W. Wahlster, G. Zimmermann.** Incremental Natural Language Description of Dynamic Imagery. In: C. Freksa, W. Brauer, eds., *Wissensbasierte Systeme. 3. Int. GI-Kongreß*, pp. 153–162, Springer, Berlin, Heidelberg, 1989.
- M. Hußmann, P. Scheffe.** The Design of SWYSS, a Dialogue System for Scene Analysis. In: L. Bolc, ed., *Natural Language Communication with Pictorial Information Systems*, pp. 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.
- 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.** NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenen. *Informatik Forschung und Entwicklung*, **1**, 83–92, 1986.
- N. Okada.** SUPP: Understanding Moving Picture Patterns Based on Linguistic Knowledge. In: *Proc. of the 6th IJCAI*, pp. 690–692, Tokio, Japan, 1979.

- G. Retz-Schmidt.** A REPLAI of SOCCER: Recognizing Intentions in the Domain of Soccer Games. In: *Proc. of the 8th ECAI*, pp. 455–457, Munich, 1988.
- T. Rist, G. Herzog, E. André.** Ereignismodellierung zur inkrementellen High-level Bildfolgenanalyse. In: E. Buchberger, J. Retti, eds., *3. Österreichische Artificial-Intelligence-Tagung*, pp. 1–11, Springer, Berlin, Heidelberg, 1987.
- K. Rohr.** Auf dem Wege zu modellgestütztem Erkennen von bewegten nicht-starren Körpern in Realweltbildfolgen. In: H. Burkhardt, K. H. Höhne, B. Neumann, eds., *Mustererkennung 1989, 11. DAGM Symposium*, pp. 324–328, Springer, Berlin, Heidelberg, 1989.
- J. R. J. Schirra.** Ein erster Blick auf ANTLIMA: Visualisierung statischer räumlicher Relationen. In: D. Metzger, ed., *GWAI-89: 13th German Workshop on Artificial Intelligence*, pp. 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, eds., *Mustererkennung 1988; 10. DAGM Symposium*, pp. 90–96, Springer, Berlin, Heidelberg, 1988.
- C.-K. Sung, G. Zimmermann.** Detektion und Verfolgung mehrerer Objekte in Bildfolgen. In: G. Hartmann, ed., *Mustererkennung 1986; 8. DAGM-Symposium*, pp. 181–184, Springer, Berlin, Heidelberg, 1986.
- J. K. Tsotsos.** Temporal Event Recognition: An Application to Left Ventricular Performance. In: *Proc. of the 7th IJCAI*, pp. 900–907, Vancouver, Canada, 1981.
- C. Vandeloise.** *Description of Space in French*. Ph.D. thesis, Univ. of California, San Diego, CA, 1984.
- 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*, pp. 643–646, Karlsruhe, FRG, 1983.
- I. Walter.** *Datenbankgestützte Repräsentation und Extraktion von Episodenbeschreibungen aus Bildfolgen*. Springer, Berlin, Heidelberg, 1989.
- D. Wunderlich.** Raumkonzepte. Zur Semantik der lokalen Präpositionen. In: T. T. Ballmer, R. Posner, eds., *Nach-Chomskysche Linguistik*, pp. 340–351, de Gruyter, Berlin, New York, 1985.