

Sonderforschungsbereich 314
Künstliche Intelligenz - Wissensbasierte Systeme

KI-Labor am Lehrstuhl für Informatik IV

Leitung: Prof. Dr. W. Wahlster

Universität des Saarlandes
FB 14 Informatik IV
Postfach 151150
D-66041 Saarbrücken
Fed. Rep. of Germany
Tel. 0681 / 302-2363



Bericht Nr. 124

From Visual Input to Verbal Output in the Visual Translator

Gerd Herzog

Juli 1995

From Visual Input to Verbal Output in the Visual Translator*

Gerd Herzog
SFB 314, Project VITRA,
Universität des Saarlandes,
D-66041 Saarbrücken, Germany
herzog@cs.uni-sb.de

Abstract

The project VITRA (VISual TRANslator) deals with the relationship between natural language and vision. Experimental studies are being carried out in the way of designing an interface between image understanding and natural language systems, with the aim of developing systems for the natural language description of image sequences. During the last ten years several domains of application have been investigated and an approach for integrating language and vision has evolved, which allows incremental evaluation and simultaneous natural language description of real world image sequences. This contribution provides a compact description of the VITRA conception and presents some of the results obtained so far. Current limitations of the approach and future research tasks will be discussed as well.

*In: Proc. of the AAAI Fall Symposium on Computational Models for Integrating Language and Vision, Cambridge, MA, 1995.

1 Introduction

Despite the intrinsic difficulty of both natural language processing and image understanding, the integration of vision and language increasingly attracts attention in current research (see e.g. [Amsili et al. 95], [AI Review 94], [McKevitt 94]). Following Wahlster [Wahlster 89], the intended natural language access to vision systems can be motivated from different points of view:

A cognitive science perspective: A computational theory of the interaction between natural language and vision

An artificial intelligence perspective: Natural language as an efficient means to make the results of an image sequence analysis process more easily accessible to humans

Important arguments for the application-oriented objective are the following:

1. In most cases a graphical representation of the results of an image sequence analysis process does not provide more information in less time than the original image sequence.
2. Natural language concepts guide the interpretation of what we see.
3. Natural language is a natural communication medium for humans. So computer systems which are able to communicate in natural language meet human needs much better.

Up to now, however, there have only been a few attempts at connecting vision systems with natural language access systems. Badler [Badler 75] developed a first proposal for describing object trajectories in terms of motion primitives and more complex combined motion patterns, corresponding to verbs and adverbials in natural language. Later, this framework has been adopted and improved by Tsotsos within the ALVEN system, for analyzing left ventricular heart motion [Tsotsos 85]. ALVEN does not provide natural language output, although most concepts of motion that can be recognized in the image sequence correspond to certain natural language notions of change. Theoretical aspects of linking language and perception have been considered by Waltz [Waltz 81]. He emphasizes the importance of natural language scene descriptions. LANDSCAN [Bajcsy et al. 85] is a natural language system, which answers questions about aerial images. In LANDSCAN, processing spans the entire way from sensor input to natural language output, but it only deals with static scenes. Dynamic traffic scenes have been investigated in the dialog-system HAM-ANS [Wahlster et al. 83] and in the system NAOS [Neumann 89], which generates retrospective natural language descriptions. Since a connection to the vision component could not be achieved at that time,



Figure 1: Image sequence showing an attack in a soccer game

the geometric descriptions of the analyzed time-varying scenes had to be prepared manually from the underlying image sequences.

The experience gained in HAM-ANS and in NAOS formed the starting point for our own investigations in the project VITRA (VISual TRANslator). The automatic natural language description of real world image sequences constitutes our major research goal, which has been pursued during the last ten years. In this contribution, we will motivate and describe our approach towards the integration of vision and language in a knowledge-based system. The results obtained so far will be presented and current limitations will be discussed.

2 From Visual Data to Verbal Descriptions

Automatic object recognition, i.e., the 3D-reconstruction of visible objects from 2D images, constitutes the central goal of computer vision. With respect to a natural language access to visual data, however, the verbal description of recognized objects represents just a first level of natural language scene description (see Fig. 2). In general, the generation of adequate verbal scene descriptions requires evaluation processes which lead to conceptual units of a higher level of abstraction. This comprises the explicit description of spatial configurations by means of spatial relations, the interpretation of object movements, and even the automatic recognition of presumed goals and plans of the observed agents. Such a *high-level scene analysis*, provides the crucial links between natural language processing and low-level image analysis, i.e., vision in the narrow sense.

Besides the question of which conceptual structures are to be extracted from an image sequence, it is decisive how the recognition process is realized. Low-level and high-level evaluation can rely on an *a posteriori* strategy if only retrospective descriptions of the analyzed image sequence have to be generated. A different task results if the verbal description is to be focused on what is currently happening. An *incremental* processing is required for such a simultaneous analysis and natural language description, since it often becomes necessary to talk about occurrences

Goal: Verbalization of visual perceptions	
Form: Verbal description of	Recognition of:
<ul style="list-style-type: none"> ● recognized objects: <i>“There are two cars and a bus.”</i> 	Objects
<ul style="list-style-type: none"> ● spatial positions of objects: <i>“The bus is in front of the church.”</i> 	Spatial Relations
<ul style="list-style-type: none"> ● object movements: <i>“The bus is stopping.”</i> 	Motion Events
<ul style="list-style-type: none"> ● presumed goals and plans of the observed agents: <i>“The bus is waiting in front of the traffic-light.”</i> 	Plans

Figure 2: Levels of verbal scene description

even while they are currently happening and not yet completed. Examples for this are the description of a passing maneuver just taking place in a traffic scene or the portrayal of an attack in a soccer game (see Fig. 1) 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 an evaluation strategy, which keeps pace with the progression of the scene. One might think of a robot capable of visually *‘perceiving’* its environment and compelled to be able to react immediately on the stimuli.

User modelling is another important issue in the context of automatic natural language description of time-varying scenes. A peculiarity of verbal scene descriptions is the fact, that the visual conceptualizations that an utterance elicits in the listener’s mind must be anticipated in order to generate communicatively adequate descriptions (see [Neumann 89], [Wahlster 89], and [Waltz 81]). This leads to important consequence for the integration of vision and language, since both directions of the transformation between visual data and linguistic structures have to be considered in the process of language production.

3 The Visual Translator Conception

As it is depicted in Fig. 3, the transformation of visual data into a verbal description can roughly be subdivided into three subtasks (see [Herzog & Wazinski 94]). Starting from a sequence of digitized video frames, the processes on the sensory level concentrate on the recognition and tracking of visible objects. They provide a geometrical reconstruction of the perceived scene. Within the VITRA framework, low-level vision is carried out by our cooperation partners, the vision group at the Fraunhofer Institute IITB, Karlsruhe.

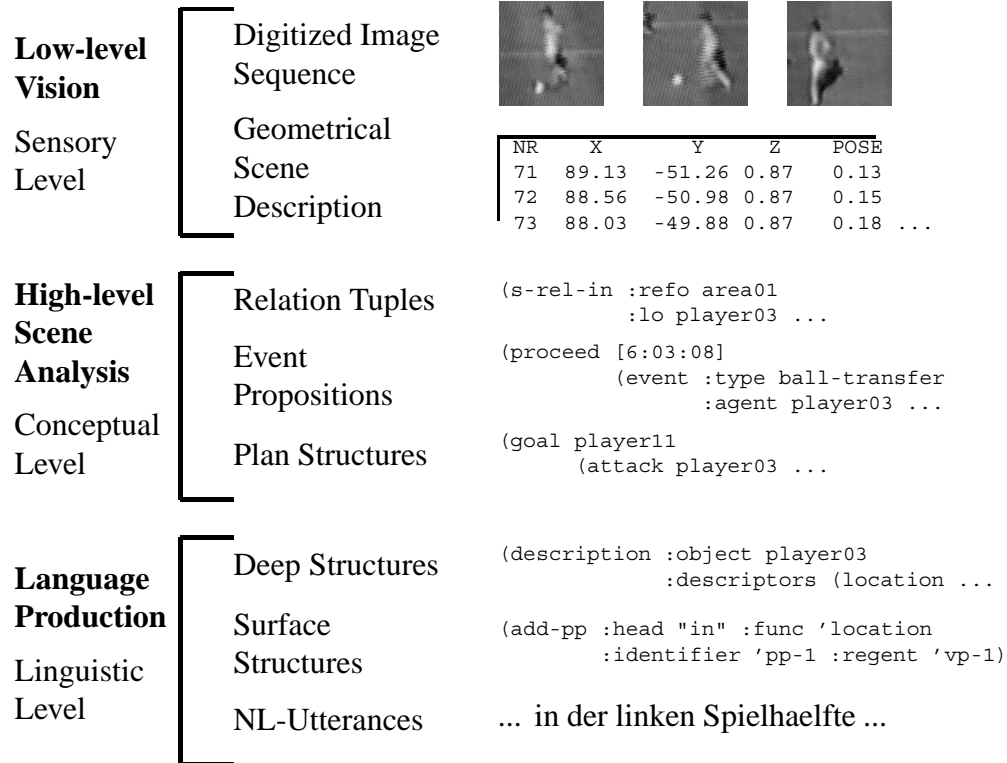


Figure 3: From visual data to a verbal description

The so-called *geometrical scene description* (GSD) has been introduced by Neumann [Neumann 89] as a representation for the intended 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)

The concept of geometrical scene description constitutes an idealized interface between low-level vision and a natural language access system. In applications, like our VITRA system, the GSD is restricted according to the practical needs. In VITRA, for example, only the trajectories of the moving objects are provided by the vision component. The stationary background of the scenes is still fed manually into the system. Fig. 4 and Fig. 5 show a short image sequence and the reconstructed GSD.

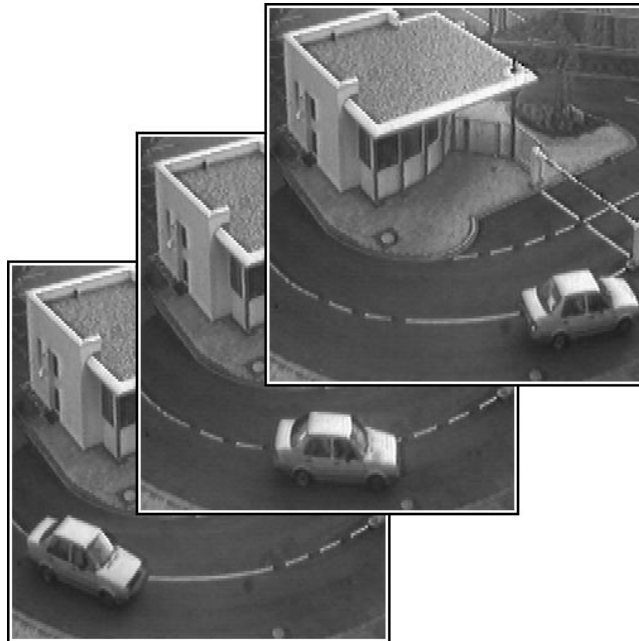


Figure 4: A short traffic scene

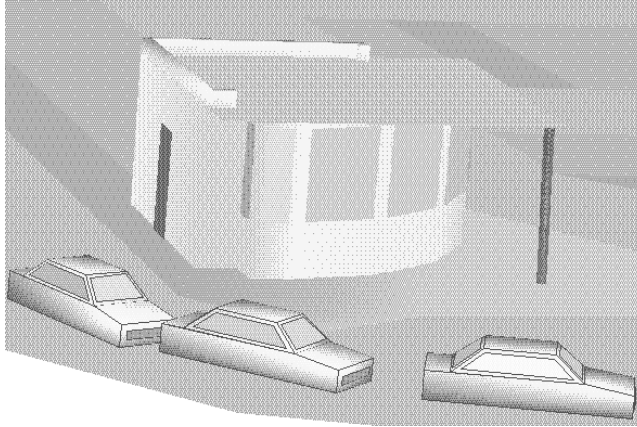


Figure 5: Visualization of the reconstructed GSD

The contents of the GSD, which is constructed incrementally, as new visual data arrive, are further interpreted during high-level scene analysis. In VITRA, incremental high-level scene analysis comprises modules for the evaluation of spatial relations (see [André et al. 87], [Herzog 95], [Gapp 94]) and the recognition of interesting motion events (see [Herzog 92]), as well as presumed intentions, plans, and plan interactions of the observed agents (see [Retz-Schmidt 91]). The conceptual structures on the cognitive level bridge the gap between visual data and natural language concepts, such as spatial prepositions, motion or action verbs and temporal adverbs.

Language production in VITRA includes processes that handle the selection, linearization and encoding of propositions (see [André et al. 88], [Herzog & Wazinski 94]). To meet the requirements of simultaneous scene description, an incremental generation strategy is employed in our approach. The listener model provides an imagination component (see [Blocher & Schirra 95], [Schirra & Stopp 93]), in order to anticipate the listener's visual conceptualizations of the described scene. Since we do not assign simple truth values to spatial predications, but instead have introduced a measure of degrees of applicability that expresses the extent to which a spatial relation is applicable, the construction of a plausible visual imagination, as well as the selection of relevant propositions is facilitated.

4 Experimental Results

During the last years several image sequences from different domains of discourse have been evaluated for our experimental investigations within the VITRA project. As a first step, trajectories of the center of gravity of object candidates in the image

plane could be provided and utilized to answer simple questions about observations in a short traffic scene [Schirra et al. 87]. Moving object candidates are segmented from the stationary ones by computing and analysing displacement vector fields.

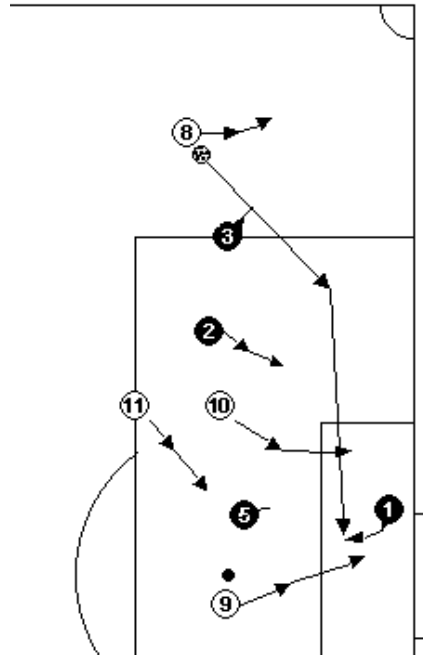


Figure 6: 2D representation in the soccer domain

The same image analysis method, which considers object candidates to be essentially rigid, has been applied to short sections of an image sequence recorded from a soccer game (see Fig. 1). The calibration of the camera allowed for the transformation of trajectories from the image plane into world coordinates. The as yet partial trajectories delivered by the vision component were used to synthesize interactively a realistic GSD, with object candidates assigned to previously known players and the ball. Together with an instantiated model of the static background (see Fig. 6), this information served as input for the VITRA system, which generates a running report for the scene under consideration [Herzog et al. 89].

The more recent model-based approach described in [Koller et al. 92] accomplishes the automatic 3D-reconstruction of vehicles in traffic scenes and provides more reliable trajectories for our natural language access system. In addition to object recognition, a coarse classification of recognized vehicles (e.g. *'platform van'*, *'estate car'*) is attempted as well.

The vision group has even been concerned with the model-based recognition of non-rigid mobile objects. Research described in [Rohr 94] concentrates on the 3D-

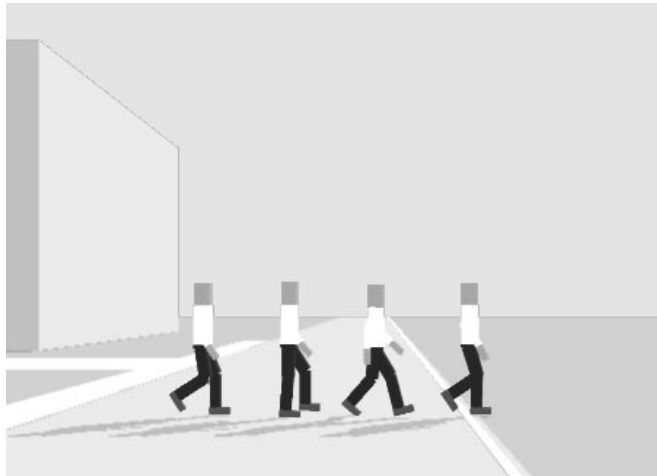


Figure 7: Automatic recognition of a walking person

reconstruction of the movements of articulated bodies. A cylindric representation and a kinematic model of human walking, which is based on medical data, is utilized for the incremental recognition of pedestrians. The algorithm determines the 3D positions as well as the postures of moving persons. As yet, one of the analyzed image sequences has been further investigated within the VITRA system (see [Herzog & Rohr 95]).

Fig. 7 shows a complete frame of this time-varying scene and relevant details comprising the entire sequence. A visualization of the reconstructed GSD is given as well. The displayed trajectory has been derived from the real world image sequence without manual interaction. Our experimental studies of real world image sequences are summarized in Table 1.

Domain	Multilane street intersection	Soccer	Parking area	Multilane street intersection	Pedestrian
Viewpoint	fixed	fixed	fixed	fixed	fixed
Monocular	yes	yes	yes	yes	yes
Number of frames (Duration)	130 (5.2s)	681 (27.2s)	81 (3.2s) & 400 (16s)	50 (2s)	80 (3.2s)
Frame size	512 * 512 8 bit no color	512 * 512 8 bit no color	512 * 512 8 bit no color	512 * 512 8 bit no color	256 * 256 8 bit no color
Number of mobile objects	10	13	1 & 2	9	1
Trajectory data format	2D image plane	2D world coordinates	3D	3D	3D
Point density (in VITRA)	0.8s	0.1s	0.04s	0.04s	0.04s

Table 1: Investigated real world image sequences

5 Discussion and Outlook

VITRA provides an operational form of referential semantics that reaches down to the sensoric level. The approach has been applied to real world image sequences

and an integration of vision and language processing could be achieved. These are first promising results, but we are still far from a universally applicable AI system capable of describing an arbitrary sequence of images. From the point of view of natural language generation, low-level vision is still restricted to rather short and relatively simple image sequences.

The role of the geometrical scene description, as it has been presented here, constitutes a crucial limitation of our current approach, since it excludes the bidirectional interleaving of image processing and scene analysis. Interaction between low-level and high-level analysis is required if VITRA is to become robust for insufficiencies in low-level vision. In more advanced applications active sensor control and focussing techniques are necessary to compensate the computational complexity of the evaluation. Resource-adaptive behaviour of the algorithms in general becomes an important issue if real-time performance has to be attempted seriously.

The simultaneous evaluation and natural language description of real world image sequences provides a basis for further investigations on the integration of vision and language. Current research in VITRA concentrates on transferring the developed methodology to application areas like (1) intelligent multimedia systems (see [André et al. 94]), (2) driver support systems in road vehicles (see [Maaß 94; Maaß et al. 93]), and (3) autonomous mobile robot systems (see [Längle et al. 95; Stopp et al. 94]).

6 Technical Notes

The VITRA system is written in Common Lisp and CLOS, with the graphical user interface implemented in CLIM. In addition, the Geomview package is employed as a display engine for the visualization and animation of geometrical scene descriptions. The VITRA workbench has been developed on Symbolics UX1200S Lisp Coprocessors, and on HP 9720 as well as on SUN Workstations.

Acknowledgements

The work described in this paper was supported by the Special Collaborative Program on AI and Knowledge-based Systems (SFB 314), project N2: VITRA, of the German Science Foundation (DFG). The support of the Geometry Center, Minneapolis, MN, which provided the Geomview software, is gratefully acknowledged.

References

- [AI Review 94] Artificial Intelligence Review Journal, 8, Special Volume on the Integration of Natural Language and Vision Processing, 1994.
- [Amsili et al. 95] P. Amsili, M. Borillo und L. Vieu (Hrsg.). *Proc. of the 5th Toulouse International Workshop "Time, Space, and Movement: Meaning and Knowledge in the Sensible World"*, Château de Bonas, France, 1995. Groupe "Langue, Raisonnement, Calcul", Toulouse.
- [André et al. 87] E. André, G. Bosch, G. Herzog und T. Rist. Coping with the Intrinsic and the Deictic Uses of Spatial Prepositions. In: K. Jorrand und L. Sgurev (Hrsg.), *Artificial Intelligence II: Methodology, Systems, Applications*, pp. 375–382. Amsterdam: North-Holland, 1987.
- [André et al. 88] E. André, G. Herzog und T. Rist. On the Simultaneous Interpretation of Real World Image Sequences and their Natural Language Description: The System SOCCER. In: *Proc. of the 8th ECAI*, pp. 449–454, Munich, 1988.
- [André et al. 94] E. André, G. Herzog und T. Rist. Multimedia Presentation of Interpreted Visual Data. In: P. McKeivitt (Hrsg.), *Proc. of AAAI-94 Workshop on "Integration of Natural Language and Vision Processing"*, pp. 74–82, Seattle, WA, 1994.
- [Badler 75] N. I. Badler. Temporal Scene Analysis: Conceptual Description of Object Movements. Technical Report 80, Computer Science Department, Univ. of Toronto, 1975.
- [Bajcsy et al. 85] R. Bajcsy, A. Joshi, E. Krotkov und 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.
- [Blocher & Schirra 95] A. Blocher und J. R. J. Schirra. Optional Deep Case Filling and Focus Control with Mental Images: ANTLIMA-KOREF. In: *Proc. of the 14th IJCAI*, Montreal, Canada, 1995. to appear.
- [Gapp 94] K.-P. Gapp. Basic Meanings of Spatial Relations: Computation and Evaluation in 3D Space. In: *Proc. of AAAI-94*, pp. 1393–1398, Seattle, WA, 1994.
- [Herzog & Rohr 95] G. Herzog und K. Rohr. Integrating Vision and Language: Towards Automatic Description of Human Movements. In: C.-R. Rollinger und W. Brauer (Hrsg.), *KI-95: Advances in Artificial Intelligence*. Berlin, Heidelberg: Springer, 1995. to appear.

- [Herzog & Wazinski 94] G. Herzog und P. Wazinski. VISual TRANslator: Linking Perceptions and Natural Language Descriptions. *Artificial Intelligence Review*, 8(2/3):175–187, 1994.
- [Herzog et al. 89] G. Herzog, C.-K. Sung, E. André, W. Enkelmann, H.-H. Nagel, T. Rist, W. Wahlster und G. Zimmermann. Incremental Natural Language Description of Dynamic Imagery. In: C. Freksa und W. Brauer (Hrsg.), *Wissensbasierte Systeme. 3. Int. GI-Kongreß*, pp. 153–162. Berlin, Heidelberg: Springer, 1989.
- [Herzog 92] G. Herzog. Utilizing Interval-Based Event Representations for Incremental High-Level Scene Analysis. In: M. Aurnague, A. Borillo, M. Borillo und M. Bras (Hrsg.), *Proc. of the 4th International Workshop on Semantics of Time, Space, and Movement and Spatio-Temporal Reasoning*, pp. 425–435, Château de Bonas, France, 1992. Groupe “Langue, Raisonnement, Calcul”, Toulouse.
- [Herzog 95] G. Herzog. Coping with Static and Dynamic Spatial Relations. In: P. Amsili, M. Borillo und L. Vieu (Hrsg.), *Proc. of TSM’95, Time, Space, and Movement: Meaning and Knowledge in the Sensible World*, pp. C 47–59, Château de Bonas, France, 1995. Groupe “Langue, Raisonnement, Calcul”, Toulouse.
- [Koller et al. 92] D. Koller, K. Daniilidis, T. Thórhallson und H.-H. Nagel. Model-based Object Tracking in Traffic Scenes. In: G. Sandini (Hrsg.), *Proc. of Second European Conf. on Computer Vision*, pp. 437–452. Berlin, Heidelberg: Springer, 1992.
- [Längle et al. 95] T. Längle, T. C. Lüth, G. Herzog und E. Stopp. KANTRA - A Natural Language Interface for Intelligent Robots. In: U. Rembold, R. Dillman, L. O. Hertzberger und T. Kanade (Hrsg.), *Intelligent Autonomous Systems (IAS 4)*, pp. 357–364. Amsterdam: IOS, 1995.
- [Maaß et al. 93] W. Maaß, P. Wazinski und G. Herzog. VITRA GUIDE: Multimodal Route Descriptions for Computer Assisted Vehicle Navigation. In: *Proc. of the Sixth Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems IEA/AIE-93*, pp. 144–147, Edinburgh, Scotland, 1993.
- [Maaß 94] W. Maaß. From Visual Perception to Multimodal Communication: Incremental Route Descriptions. *Artificial Intelligence Review*, 8(2/3):159–174, 1994.
- [McKevitt 94] P. McKevitt (Hrsg.). *Proc. of AAAI-94 Workshop on Integration of Natural Language and Vision Processing*, Seattle, WA, 1994.

- [Neumann 89] B. Neumann. Natural Language Description of Time-Varying Scenes. In: D. L. Waltz (Hrsg.), *Semantic Structures: Advances in Natural Language Processing*, pp. 167–207. Hillsdale, NJ: Lawrence Erlbaum, 1989.
- [Retz-Schmidt 91] G. Retz-Schmidt. Recognizing Intentions, Interactions, and Causes of Plan Failures. *User Modeling and User-Adapted Interaction*, 1:173–202, 1991.
- [Rohr 94] K. Rohr. Towards Model-based Recognition of Human Movements in Image Sequences. *Computer Vision, Graphics, and Image Processing (CVGIP): Image Understanding*, 59(1):94–115, 1994.
- [Schirra & Stopp 93] J. R. J. Schirra und E. Stopp. ANTLIMA — A Listener Model with Mental Images. In: *Proc. of the 13th IJCAI*, pp. 175–180, Chambery, France, 1993.
- [Schirra et al. 87] J. R. J. Schirra, G. Bosch, C.-K. Sung und G. Zimmermann. From Image Sequences to Natural Language: A First Step Towards Automatic Perception and Description of Motions. *Applied Artificial Intelligence*, 1:287–305, 1987.
- [Stopp et al. 94] E. Stopp, K.-P. Gapp, G. Herzog, T. Längle und T. C. Lüth. Utilizing Spatial Relations for Natural Language Access to an Autonomous Mobile Robot. In: B. Nebel und L. Dreschler-Fischer (Hrsg.), *KI-94: Advances in Artificial Intelligence*, pp. 39–50. Berlin, Heidelberg: Springer, 1994.
- [Tsotsos 85] J. K. Tsotsos. Knowledge Organization and its Role in Representation and Interpretation for Time-Varying Data: the ALVEN System. *Computational Intelligence*, 1:16–32, 1985.
- [Wahlster et al. 83] W. Wahlster, H. Marburger, A. Jameson und S. Busemann. Over-answering Yes-No Questions: Extended Responses in a NL Interface to a Vision System. In: *Proc. of the 8th IJCAI*, pp. 643–646, Karlsruhe, FRG, 1983.
- [Wahlster 89] W. Wahlster. Natural Language Systems: Some Research Trends. In: H. Schnell und N. O. Bernsen (Hrsg.), *Logic and Linguistics: Research Directions in Cognitive Science - European Perspectives, Vol. 2*, pp. 171–183. Hillsdale, NJ: Lawrence Erlbaum, 1989.
- [Waltz 81] D. L. Waltz. Understanding and Generating Scene Descriptions. In: A. Joshi, B. L. Webber und I.A. Sag (Hrsg.), *Elements of Discourse Understanding*, pp. 266–281. Cambridge, London: Cambridge University Press, 1981.