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. 117

Visual Grounding of Route Descriptions in Dynamic Environments

Wolfgang Maaß, Jörg Baus, Joachim Paul

Juli 1995

ISSN 0944-7814

117

Visual Grounding of Route Descriptions in Dynamic Environments

Wolfgang Maaß, Jörg Baus, Joachim Paul
University of Saarbrücken
66041 Saarbrücken, Germany
Email address: maass@cs.uni-sb.de

Abstract. A software agent who gives incremental, i.e. step-by-step, route descriptions while moving through an environment is an interesting starting-point for an integrated view on visual perception and natural language generation. We present a computational model, called MOSES. In particular we show how visual data is transformed into visuo-spatial representations. An object selection process based on visual features starts at a high-level description of objects in a synthetic three-dimensional environment. We found by experiments that incremental route descriptions can be classified by a small set of syntactic and semantic structures. By consideration of temporal constraints, visuo-spatial structures, path-related intentions, as well as rhetorical abilities of the speaker, a selection process extracts description schemata as input for the language generation process. These schemata are modeled by a modified subset of Jackendoff's conceptual semantics formalism.

Keywords: visual object selection, route descriptions, spatial knowledge representation

1 Introduction

The model presented here is part of a larger project, called VITRA (VIsual TRAnslator), where we investigate aspects of the interaction of vision and language ([Herzog et al. 89]). A particular interest lies on the information flow from analysis of visual data to language generation. We focus on how visual information can be used for grounding descriptions in the environment. In cooperation with the visual perception group of the IIFB at the Fraunhofer Institute at the University of Karlsruhe, we have shown how real-world visual data in dynamic environments can be used in natural language descriptions ([Herzog et al. 89, Schirra et al. 87, Huang et al. 94]). A model-based approach is used for automatically generating 3D-representations of the environment. This approach has been examined in a soccer domain ([André et al. 89]) and a dynamic traffic scene domain ([Schirra et al. 87]). The model presented here depends on experiences gained in these domains. Our current work is related to problems which occur if an agent moves through real or synthetic environments.

The agent's task during its movement is to incrementally describe a route from a starting point to a destination by refering to visually obtained objects. In general, the whole complexity of AI research is involved, e.g., control laws for movement, early-vision processing, high-level vision, naive physics, temporal and spatial reasoning, knowledge representation, planning, and language processing. In a first approach, we have implemented a software agent[1], called `MOSES`, who describes a path in a synthetic 3D-environment. `MOSES` can only refer to visually obtained objects (landmarks) in the current situation. Information about the path is extracted from a map by using an incremental path-finding procedure (for more details see [Maaß 94]). In this paper, we will sketch the process architecture and associated representations.

2 Related work

It has not changed so much since Waltz mentioned that the interaction between researchers in visual perception and language processing is quite small ([Waltz 81]). Although there are recently more activities in this area ([McKevitt 94b, McKevitt 94a]), we are still lacking a complete theory. In contrast to this problem, this topic is in psychology, as well as in philosophy subject of long-standing controversies. Although there are evidences to believe that the integration of vision and language is quite complex the human cognitive system has found an efficient way to integrate these two complex modules. Unwrapping the functional architecture of the interaction of vision and language will provide important insights about cognitive processes and representations in general. On the other hand, related investigations may help to find efficient and useful computational architectures in general. If we better understand how humans refer to visual information in language we might be able to build more cooperative systems.

The interaction between vision and language has been a main question in cognitive psychology during the last decades. But the general question has been decomposed into many virtually smaller ones. Most popular has been the *imagery debate* in which it is discussed whether representations of visual information is coded in a propositional or dipictional format ([Pylyshyn 81, Kosslyn et al. 90]. Experiments in this area are mostly situated in semantic-free small-scale environments and do not consider realistic scenarios. Real-world situations are investigated by experiments about how large-scale affects human anticipation of the environment (cf. [Downs & Stea 73, Downs & Stea 77]). Large scale space environments include several sources of non-visual and non-spatial information. A viewer always has a set of goals towards her environment which direct her behavior. If she walks along a route she intends to reach a destination. She uses her environment in a goal-directed habit. This influences her perception of the

[1] The complete model is implemented in LISP (CLOS and CLIM) on different hardware platforms. We would like to thank Gerd Herzog, Jochen Müller, Eva Stopp and Anselm Blocher for contributing important software modules and valuable information. This work is partly supported by the cognitive science program 'Kognitionswissenschaft' at the University of Saarbrücken.

environment and in particular her focus of attention. But an important question is what happens in-between vision and language. It is confirmed by neurophysiologial and psychophysical studies that one of the main task of vision is to select appropriate information. Visual attention is the process which focuses on particular parts of a given scene and leads visual perception. A common view is that information selected by visual processes is stored in 3-dimensional representations (e.g., [Marr 82]). But what happens after objects are represented is more or less speculative. There are several approaches to harmonize different kinds of structures into a general framework ([Johnson-Laird 83, Bryant 92, Landau & Jackendoff 93]). But beside models which try to find a link between vision and language, several studies have been conducted to determine an independent representation format of spatial knowledge of large-scale space, generaly called *cognitive maps* (cf. [Tolman 48, Downs & Stea 77, Gärling et al. 84]). Beside spatial information, cognitive maps depend on information about temporal constraints, viewer properties, and other conceptual factors. How we memorize and retrieve spatial knowledge strongly depends on situation in which we experienced the environment. But experiments showed that we must assume different kinds of formats for identical spatial knowledge. We talk about new environments by refering to perspective views (route knowledge) and to familiar environments by refering to bird's-eye perspectives (map knowledge) (see [Siegel & White 75]).

The question now is what kind of processes and representations are involved if we talk by refering to visual information. On one hand, a general framework must be able to deal with any kind of environment, whether we directly perceive it or whether we refer to memorized representations. It must also include contextual constraints and viewer/speaker and listener properties. As already mentioned, we are far from having such a theory. Therefore we restrict our proposed model to a minimal set of processes and representations which allows the system to produce descriptions similar to those given by humans in comparable situations. There are only a few computational models in which visual and linguistic structures are related to one another. In some approaches it is investigated how to analyse time-varying environments from a static view-point. The goal in LandScan is to guide low-level visual processes by textual input and to provide descriptions of visual information ([Bajcsy et al. 85]). NAOS ([Neumann & Novak 83]) and a model proposed by Howarth and Buxton ([Howarth & Buxton 93]) are able to describe time-varying visual data in traffic scenes. Other models do not depend on visual information but on spatial configurations as domains for textual queries and descriptions ([Winograd 72, Waltz 81, Wahlster et al. 83, Novak & Bulko 90]). Associated to route descriptions is navigation. Computational models of navigation provide important insights about how visual data is transformed into spatial representations and how it can be used for language processes (e.g., [Kuipers 78, Gopal et al. 89, Leiser & Zilbershatz 89, McCalla & Schneider 79, Vere & Bickmore 89]). In particular, systems developed in robotics focus on the interaction of physical navigation and on low-level vision processes (e.g., [Brooks & Maes 94]). But those systems are usually lacking language abilities.

3 The proposed model

In static environments a system has usually no deadlines for its behavior patterns. This is different for systems in dynamic environments. But also synthetic environments provide strong constraints on agents. In MOSES we focus on how to select visuo-spatial information units which are salient in a given situation. Descriptions given by MOSES are compared with descriptions given by test persons in comparable real-world environments.

MOSES receives information about the environment by a visual input process and by searching paths on street maps (see figure 1). A main problem for the visual perception process is to select interesting objects. An object is selected if corresponding visual features are salient. Features are basic properties of objects, such as color, width, height, and direction of movement. In our current implementation, MOSES considers three different feature types: Color, height, and width (c.f. [Maaß 95]). Each feature is projected onto a type-specific, e.g, a color feature is integrated in the color feature map. A feature map is a two-dimensional projection of the scene and preserves its topographical structure. For each feature, MOSES computes a *salience value* which is modeled by a potential field representation. The intensity of a feature defines the third dimension of a feature map (vertical cuts through feature maps are shown in figure 2 and 3). MOSES determines a mean value for each feature map. For instance, the mean value of a color feature map is the arithmetic mean of all RGB values. The distance between a mean value and a value of a particular feature represents the salience value of this feature, i.e. the maximum of the corresponding potential. The extension of a potential field is the visible enlargement of that feature (see figure 2 and 3). Feature maps are joined in a *global feature map*. Therefore all feature maps are projected onto one another. A problem is how to combine feature maps of different types. We have asked test persons to rate feature types in different scenes. According to these ratings we determined factors for the combination of feature maps. As we will describe next, MOSES is mainly lead by his[2] intentions. If he wants to turn left at the next decision point he focuses his attention to an area on the left. In our experiments we found that objects on the opposite side are neglected in descriptions even if they have been quite salient. According to this phenomenon, MOSES has a *spatial attention window* which is used to increase salience values of features which lie in the focus area and decrease those which lie outside of this area (see figure 3). The location of the maximum value in the global feature map corresponds to the most salient location in a given scene. The object selection process now determines objects which corresponds to salient locations and passes them to the visuo-spatial buffer (see figure 1).

Descriptions are always accompanied by intentions. MOSES' main intention is to follow a path from a starting point to a destination. This global goal is subdivided into simpler goals as he moves along the path. His intentions are mainly path-related. By refering to the path-finding process, he knows where to go at

[2] For historical reasons, we refer to MOSES by masculine pronouns.

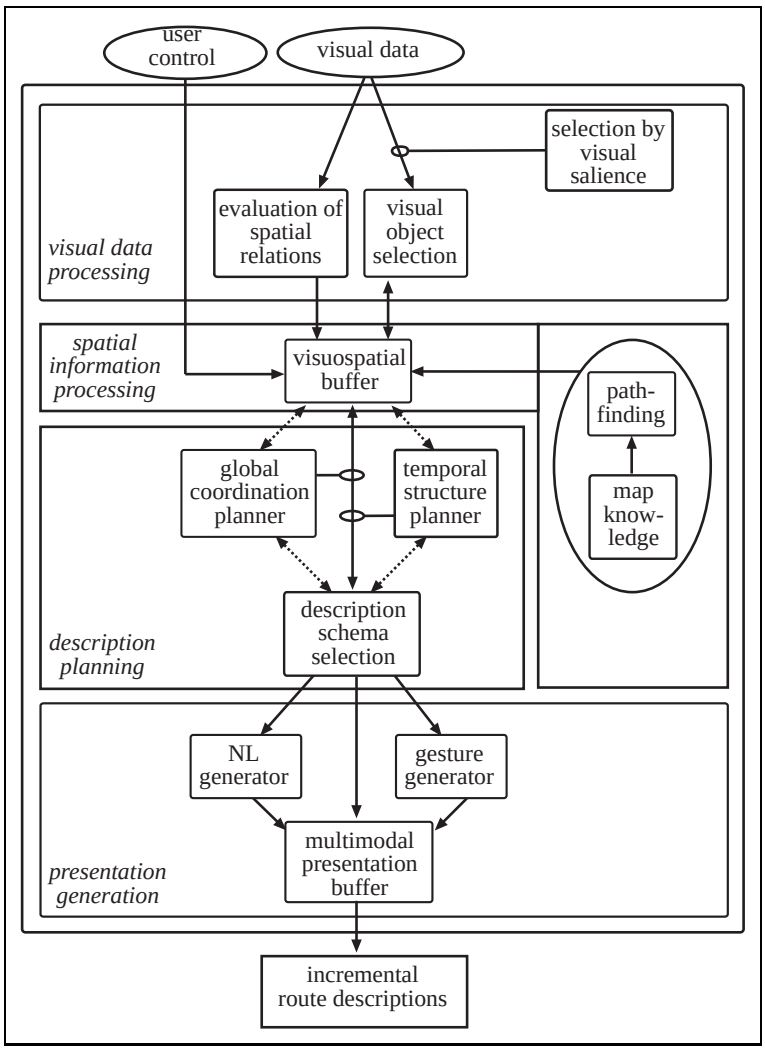

Fig. 1. Process model of MOES

each decision point. As already mentioned, path-related intentions control the focus of attention. But these intentions are also important for the language generation process. For instance, test persons have had severe problems to generate appropriate descriptions of a turn-left action when they were forced to refer to an object on the right side. Beside path-related intentions, MOSES also has the intention to describe the route incrementally at appropriate time-points.

Now that salient objects (L1 is the first building on the left side of the crossing

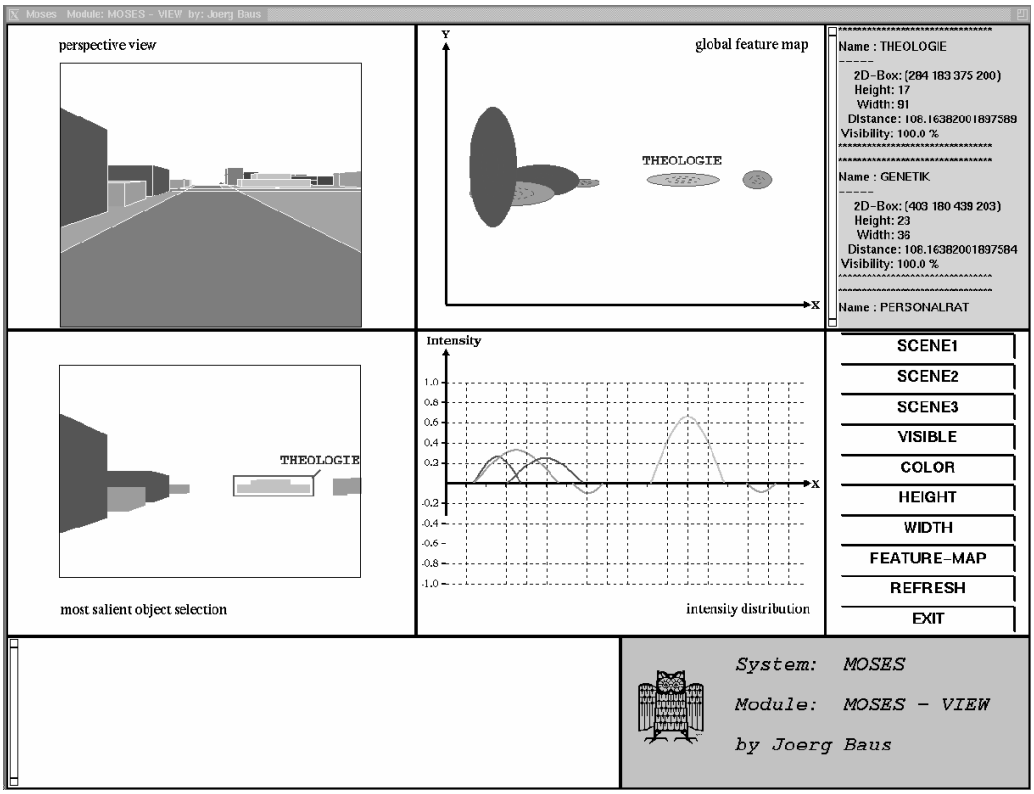

Fig. 2. Object selection by reference to the global feature map in a complex scene

scenario indicated in figure 3) are selected and sent to the visuo-spatial buffer, they are interrelated to MOSES' current position (CP) and street items (S1, S2, and C) (crossings and street segments) by topographical spatial relations (see figure 4). This representation is called *configuration description*. A configuration description provides explicit information about the spatial structure of a situation. Route descriptions mainly depend on the spatial structure represented by configuration descriptions. MOSES considers the given configuration description, intentions, the temporal structure of the situation, and his linguistic abilities to select an appropriate description schema. The temporal structure of a situation is constrained by the speed of MOSES and the distance to the next decision point. MOSES makes assumptions about how long it will probably take to reach the next decision point. According to this time interval, those schemata are selected which can be used to generate a description right in time. The next filter selects from

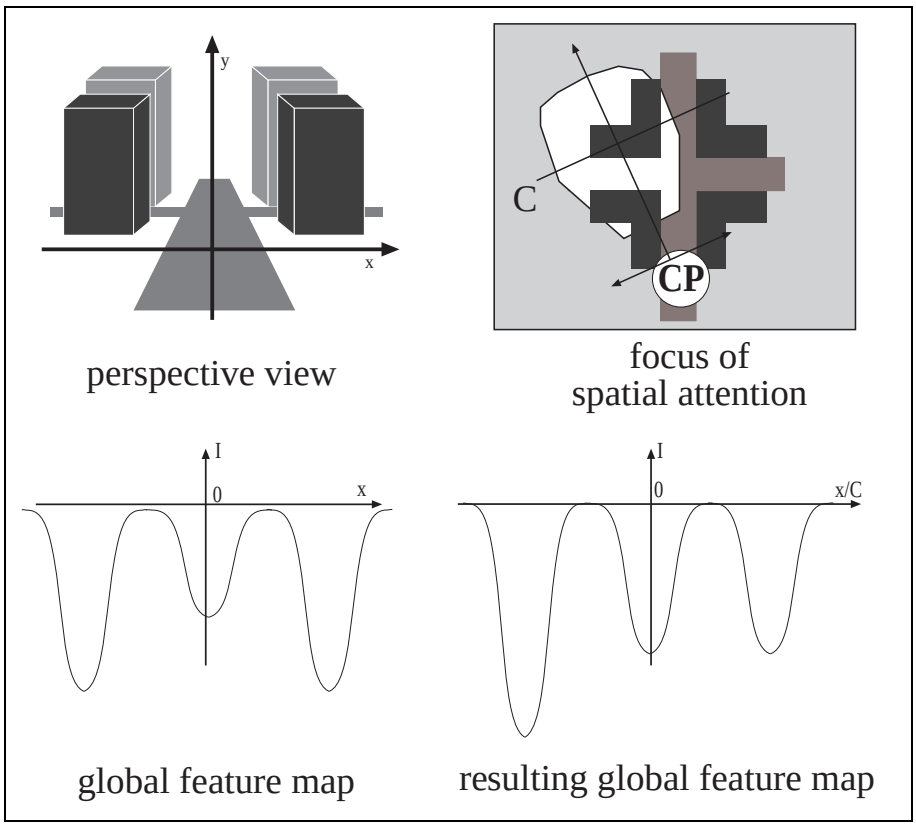

Fig. 3. Modifying potential field representations by a focus of spatial attention in a crossing scenario

these schemata those which correspond to the intended action at the next decision point. During the next selection step, those schemata are extracted which assume a similar spatial structure as given by the configuration description. If there are objects selected by the object selection process those schemata are used which include a reference to salient objects at appropriate places. Incremental route descriptions are generally ill-formed sentences. There are differences between speakers. Some are able to smoothly include objects in their descriptions. Others describe the action first followed by an indication of the location. `MOSES` can be used in two different rhetoric modes: poor and normal.

Most of all, `MOSES` descriptions depend on his type of movement. When he moves with average car speed, intervals between decision points are sometimes quite short. In those situations, he mainly refers to route knowledge. If he moves with walking speed, he has more time and can refer to objects. For instance, if

a salient object is on the left side and his intention is to turn left he gives the description in two parts: "Please, turn left behind the red building on the left side." .... "Now, please." First he gives a complete description of the intended action by refering to objects. Then, just before the action must be performed, he gives a short hint. This selection process extracts and instantiates one or more description schemata. If there are more than one schemata, MOSES uses the first one. It is clear that a more sophisticated conflict resolution procedure would be helpful. But in our domain we found that our simple strategy serves quite well.

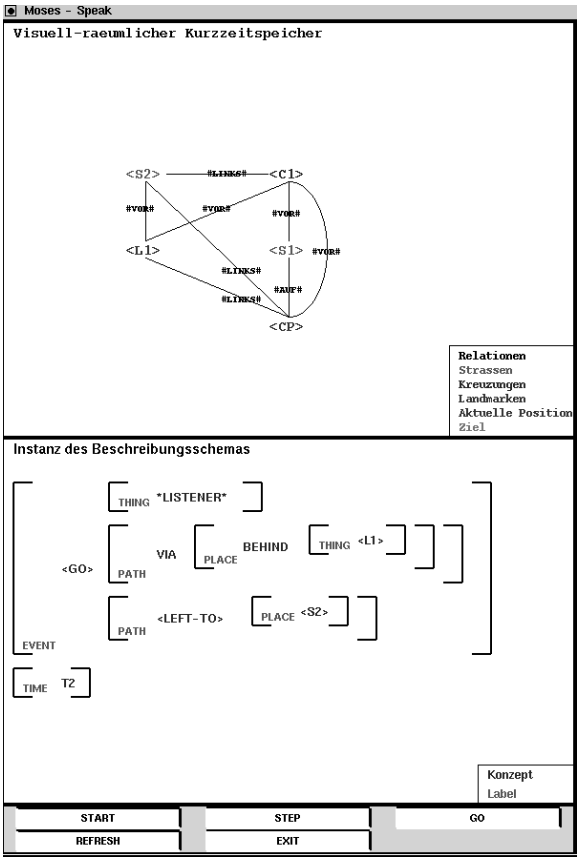

Fig. 4. Example configuration description of a turn-left action by refering to a landmark

A description schema provides the semantic of a route description. The structure of a schema is based on Jackendoff's *conceptual semantics* (see [Jackendoff 83]).

Because Jackendoff only presented his framework of conceptual semantics by refering to simple utterances we carefully extended his formalism (see figure 4). MOSES has a repertoire of almost 60 description schemata. Basic constituents of a description schema are *things* (persons), *locations* (places), and *paths*. They are used in higher-order structures, such as *events* and *states*. The general structure of an event consists of the instance of the speaker MOSES followed by a path and a place. Hence we can represent utterances such as: "Please, turn left behind the building on the left side." Figure 4 shows the conceptual structure of this description. In the last step, a description schema is transformed into input structures for the language generator. We have two generators, one based on the tree-adjoining-grammar approach ([Finkler & Schauder 92]) and another one which is simpler but faster.

4 Summary and future work

MOSES is a model for generating incremental route descriptions. This strategy is based on experiences we gained in other domains in cooperation with a vision processing group. Nevertheless it is obvious that the range MOSES is dealing with is far beyond the capabilities of nowadays low-level vision processing systems. We have sketched the process model of MOSES. It consists of a object selection process based on visual features, an incremental path-finding process on street maps, a visuo-spatial buffer which determines configuration descriptions, and a selection process for description schemata. Finally these description schemata are transformed into input structures for a language generation process.

MOSES provides an approach for integrating several complex processes of vision processing and natural language processing. Therefore it is clear that almost every process can be refined by future work. But nonetheless, MOSES provides a robust experimental environment for gaining more experience about the relation of vision and language in dynamic environments. In the near future, we will integrate capabilities for tracking and describing external events, such as a car which turns left at a crossing or a persons crossing the street ([Herzog & Rohr 95]). Another important goal for the future is to combine this software agent with a physical agent equipped with motoric abilities, a vision unit and speech output.

References

[André et al. 89] E. **André**, G. **Herzog**, and T. **Rist**. *Natural Language Access to Visual Data: Dealing with Space and Movement.* In: F. Nef and M. Borillo (eds.), Logical Semantics of Time, Space and Movement in Natural Language. Proc. of 1^{st} Workshop. Hermès, 1989.

[Bajcsy et al. 85] R. **Bajcsy**, A. **Joshi**, E. **Krotkov**, and A. **Zwarico**. *LandScan: A Natural Language and Computer Vision System for Analyzing Aerial Images.* In: Proc. of the 9^{th} IJCAI, pp. 919–921, Los Angeles, CA, 1985.

[Brooks & Maes 94] Rodney A. **Brooks** and Pattie **Maes** (eds.). *ARTIFICIAL LIFE IV, Proceedings of the fourth International Workshop on the Synthesis and Sim-*

ulation of Living Systems. MIT Press, 1994. 6-8th July 1994, MIT, Cambridge, MA, USA.

[Bryant 92] D. J. **Bryant**. *A Spatial Representation System in Humans.* Journal of Memory and Language, 31:74–98, 1992.

[Downs & Stea 73] R. M. **Downs** and D. **Stea**. *Image and Environment: Cognitive Mapping and Spatial Behaviour.* Chicago: Aldine, 1973.

[Downs & Stea 77] R. M. **Downs** and D. **Stea**. *Maps in Mind: Reflections on Cognitive Mapping.* New York: Harper & Row, 1977.

[Finkler & Schauder 92] W. **Finkler** and A. **Schauder**. *Effects of Incremental Output on Incremental Natural Language Generation.* In: Proc. of the 10 th ECAI, pp. 505–507, Vienna, 1992.

[Gärling et al. 84] T. **Gärling**, T. **Böök**, and E. **Lindberg**. *Cognitive Mapping of Large-Scale Environments.* Environment and Behavior, 16(1):3–34, 1984.

[Gopal et al. 89] S. **Gopal**, R. **Klatzky**, and T. **Smith**. *NAVIGATOR: A Psychologically Based Model of Environmental Learning Through Navigation.* Journal of Environmental Psychology, 9:309–331, 1989.

[Herzog & Rohr 95] G. **Herzog** and K. **Rohr**. *Integrating Vision and Language: Towards Automatic Description of Human Movements.* to appear, 1995.

[Herzog et al. 89] G. **Herzog**, C.-K. **Sung**, E. **André**, W. **Enkelmann**, H.-H. **Nagel**, T. **Rist**, W. **Wahlster**, and G. **Zimmermann**. *Incremental Natural Language Description of Dynamic Imagery.* In: Ch. Freksa and W. Brauer (eds.), Wissensbasierte Systeme. 3. Internationaler GI-Kongreß, pp. 153–162. Berlin, Heidelberg: Springer, 1989.

[Howarth & Buxton 93] R. **Howarth** and H. **Buxton**. *Selective attention in dynamic vision.* In: Proc. of the 13 th IJCAI, pp. 1579–1584, Chambery, France, 1993.

[Huang et al. 94] D. **Huang**, D. **Koller**, J. **Malik**, B. **Ogasawara**, B. **Rao**, S. **Russell**, and J. **Weber**. *Automatic Symbolic Traffic Scene Analysis Using Belief Networks.* In: Proc. of AAAI-94, pp. 966–972, Seattle, WA, 1994. MIT Press.

[Jackendoff 83] R. **Jackendoff**. *Semantics and Cognition.* Cambridge, MA: MIT Press, 1983.

[Johnson-Laird 83] P. N. **Johnson-Laird**. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* Cambridge University Press, 1983.

[Kosslyn et al. 90] S. **Kosslyn**, R. **Flynn**, J. **Amsterdam**, and G. **Wang**. *Components of high-level vision: A cognitive neuroscience analysis and accounts of neurological syndromes.* Cognition, 34:203–277, 1990.

[Kuipers 78] B. **Kuipers**. *Modelling Spatial Knowledge.* Cognitive Science, 2:129–153, 1978.

[Landau & Jackendoff 93] B. **Landau** and R. **Jackendoff**. *"What" and "where" in spatial language and spatial cognition.* Behavioral and Brain Sciences, 16:217–265, 1993.

[Leiser & Zilbershatz 89] D. **Leiser** and A. **Zilbershatz**. *THE TRAVELLER: A Computational Model of Spatial Network Learning.* Environmental and Behaviour, 21(4):435–463, 1989.

[Maaß 94] W. **Maaß**. *From Visual Perception to Multimodal Communication: Incremental Route Descriptions.* Artificial Intelligence Review Journal, 8(5/6), December 1994. Special Volume on Integration of Natural Language and Vision Processing.

[Maaß 95] W. **Maaß**. *How Spatial Information Connects Visual Perception and Natural Language Generation in Dynamic Environments: Towards a Computational Model.* submitted, 1995.

[Marr 82] D. **Marr**. *Vision: a computational investigation into the human representation and processing of visual information.* San Francisco: Freemann, 1982.

[McCalla & Schneider 79] G. **McCalla** and P. **Schneider**. *The Execution of Plans in an Independent Dynamic Microworld.* Proc. of the 6 th IJCAI, pp. 553–555, 1979.

[McKevitt 94a] P. **McKevitt** (ed.). *Integration of Natural Language and Vision Processing.* AAAI-94 Workshop. Seattle, WA, 1994.

[McKevitt 94b] P. **McKevitt** (ed.). *Special Volume on the Integration of Natural Language and Vision Processing*, volume 8: Artificial Intelligence Review Journal. Dordrecht: Kluwer, 1994.

[Neumann & Novak 83] B. **Neumann** and H.-J. **Novak**. *Natural-Language Oriented Event Models for Image Sequences Interpretation: The Issues.* Technical Note 34, CSRG, University of Toronto, 1983.

[Novak & Bulko 90] G. S. **Novak** and W. C. **Bulko**. *Understanding Natural Language with Diagrams.* In: Proc. of AAAI-90, pp. 465–470, St. Paul, MN, 1990.

[Pylyshyn 81] Z. W. **Pylyshyn**. *The Imagery Debate: Analogue Media versus Tacit Knowledge.* Psychological Review, 87:16–45, 1981.

[Schirra et al. 87] J. R. J. **Schirra**, G. **Bosch**, C.-K. **Sung**, and 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.

[Siegel & White 75] A. W. **Siegel** and S. H. **White**. *The Development of Spatial Representation of Large-Scale Environments.* In: W. Reese (ed.), Advances in Child Developement and Behaviour, volume 10, pp. 9–55. New York: Academic Press, 1975.

[Tolman 48] E.C. **Tolman**. *Cognitive Maps in Rats and Men.* Psychological Review, 55:189–208, 1948.

[Vere & Bickmore 89] S. **Vere** and T. **Bickmore**. *A Basic Agent.* Report, Lockheed AI Center, Palo Alto, CA, 1989.

[Wahlster et al. 83] W. **Wahlster**, H. **Marburger**, A. **Jameson**, and S. **Busemann**. *Over-answering y-No Questions: Extended Responses in a NL Interface to a Vision System.* In: Proc. of the 8 th IJCAI, pp. 643–646, Karlsruhe, FRG, 1983.

[Waltz 81] D. L. **Waltz**. *Understanding and Generating Scene Descriptions.* In: A. Joshi, B. L. Webber, and I.A. Sag (eds.), Elements of Discourse Understanding, pp. 266–281. Cambridge, London: Cambridge University Press, 1981.

[Winograd 72] Terry **Winograd**. *Understanding Natural Language.* New York: Academic Press, 1972.

This article was processed using the LaTeX macro package with LLNCS style