Towards the Interaction of Speech, Vision, and Movement: Incremental Route Descriptions

Wolfgang Maaß
Cognitive Science Program
University of Saarbrücken
66041 Saarbrücken, Germany
E-mail address: maass@cs.uni-sb.de*

Abstract

What an agent, AG, says to a hearer, HE, about what he/she sees while moving through an environment is an interesting anchorpoint for an integrated view on visual perception, natural language and movement. This communicative act is called *incremental route description*. Human behavior such as this is investigated by different research communities, such as environmental psychology, urban planning and artificial intelligence. We briefly present a computational model called MOSES by presenting its processes and representations. In a 3-dimensional environment model, the visual perception process of MOSES starts at a high-level description of objects. In this paper, we will focus on object identification and their representation on different levels. Furthermore, we suggest a distributed planning approach which controls the input processes (visual perception and wayfinding) and the presentation processes (speech generation and gesture generation).

Introduction

When we investigate the interaction with the physical world, visual perception and natural language are the most important input and output channels. Visual perception allows us to adapt our behavior to the current environment. Natural language enables us to communicate with others about circumstances in the current environment. Hence, real intelligent behaviour of an agent in the physical world must consider both information channels. But it is not sufficient to use natural language and visual perception independently. Information presented by visual perception is fundamental when we describe spatial aspects of the environment. But also natural language can influence visual perception. For instance, when you are told to look at the right side of the street, natural language guides visual perception.

The question behind studies which investigate the connection of visual perception and natural language is: How can we discuss what we see? It is commonly assumed that information provided by visual perception is sufficient to produce natural language descriptions about spatial configurations. Using this assumption, we will discuss the interaction between visual perception, natural language, and movement. In our project VITRA (VIsual TRAnslator), we investigate the natural language access to visual data (cf. (André et al. 89)). In this context we have constructed a model for the generation of *incremental route descriptions* (cf. (Maaß 93; Maaß 94)). With incremental we mean that agent AG moves along the path towards the destination and describes actions step-by-step to a hearer HE. Therefore AG uses information obtained by visual perception.

Route descriptions are common communicative actions of everyday life which can be divided into two classes: *complete* (or *pre-trip*) route descriptions and *incremental route descriptions*. In order to give a description of the entire route, we use a complete route description. Here, a common problem for the route finders is remembering many details simultanously. En route to their destination, route finders normally cannot ask the same person for additional details. In *incremental route descriptions* given by a 'co-pilot', they receive relevant route information as it is needed. In incremental route descriptions, temporal constraints are central for both the generation and following of route descriptions. Construction of a presentation includes a minimum of four phases: determination of new information, determination of a presentation structure, transmission of the information, and consideration of time the hearer will presumably require to understand and verify the information. Furthermore, the content must be presented in accordance with strengths of each presentation mode.

Presented here is a computational model for incremental route descriptions called, MOSES. Using MOSES, we show how visually accessed information can be used for natural language descriptions. Central to our model is the determination of processes and represen-

*Current address: National Center for Geographic Information and Analysis (NCGIA), University of California, 3510 Phelps Hall, Santa Barbara, CA 93106, E-mail: maass@ncgia.ucsb.edu

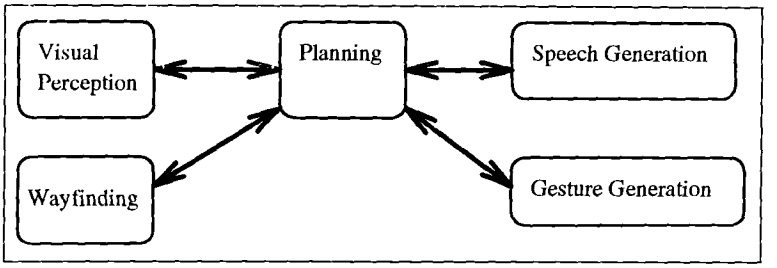

Figure 1: Process Model

tations which enable the system to behave in unknown enviroments. What we present here, can be viewed as the foundation of a system which will perform in real situations i.e., in complex and highly dynamic environments. In general, **MOSES** determines visually accessible layout information which is used to generate multimodal incremental route descriptions. **MOSES** is implemented in LISP (CLOS).

State of the art

Two topics related to the area of route descriptions which have been investigated by several researchers are action planning and spatial learning. The first computational model in the realm of spatial learning was the Tour model proposed by Kuipers (cf. (Kuipers 78)). Kuipers focuses on the relationship between what he calls *common sense knowledge of space* and the mental representation called, *cognitive maps* (cf. (Kuipers 78)). The information stored in a cognitive map includes the sequence of places and paths encountered on a route, the magnitudes of turns and distances traveled (to some low accuracy), and the observed positions of remote landmarks. Therefore he identifies four levels of descriptions of large-scale space: Sensorimotoric, procedural, topological, and metrical descriptions (cf (Kuipers & Levitt 88)). What he calls *route description* is a sequence of actions which have the effect of moving from one place to another. The main topic in Tour is the assimilation of spatial knowledge into the cognitive map. Tour does not use visual perception or maps as methods of obtaining information. It also ignores how the description is presented to the hearer.

McCalla and Schneider use a simulated taxidriver called, ELMER, in a simulated dynamic environment to present their approach of planning a route ((McCalla & Schneider 79)). Elliott and Lesk performed two experiments to investigate how people select a route (cf. (Elliott & Lesk 82)). Habel proposed a model for route descriptions (cf. (Habel 87)) based on approaches presented by Klein (cf. (Klein 83)) and Wunderlich/Reinelt (cf. (Wunderlich & Reinelt 82)). Habel presented a formalization of how objects can be represented by a representation formalism which he calls *referential network*. Leiser and Zilbershatz (cf. (Leiser & Zilbershatz 89)) determined a computational model for spatial network learning. This model, THE TRAVELLER, is based on psychological investigations by Pailhous (cf. (Pailhous 70)). The area of route descriptions is also used for investigation of the relationship between navigation and environmental learning (eg., (Piaget & Inhelder 67; Siegel & White 75)). Experiments were performed in order to test the human ability to learn about spatial enviroments. From the results of these experiments, Gopal et al. (cf. (Gopal et al. 89)) developed a psychological oriented model of environmental learning through navigation.

The model

From a phenomenological point of view, we can determine two major phases during the generation of route descriptions. First, the agent **AG** looks for an appropriate path between the starting point and destination. This is called *wayfinding*. **AG** can use different media. The most common being maps or mental representations. After the determination of the next path segment, **AG** describes the route. Usually this description is dominated by speech and spontaneous gestures but can also include sketches. While **AG** moves through the environment giving step-by-step descriptions, **AG** describes landmarks, regions, and spatial relations observed visually. **AG** also describes actions, such as turn right at the next crossing. We assume that **AG** has had no experience in the given environment. **AG** must use a map to find the path. While describing a non-trivial route, **AG** often switches between the wayfinding phase and the presentation phase. We therefore assume a central planning process which controls interaction between both phases.

We have developed a model of the *incremental route description* process ((Maaß 93; Maaß 94)). Its general architecture consists of a visual perception process, a wayfinding process, a planning process, and generation processes. The wayfinding process models the human ability to look for a route while using a map and information received visually. Interactions between visual perception, wayfinding and presenta-

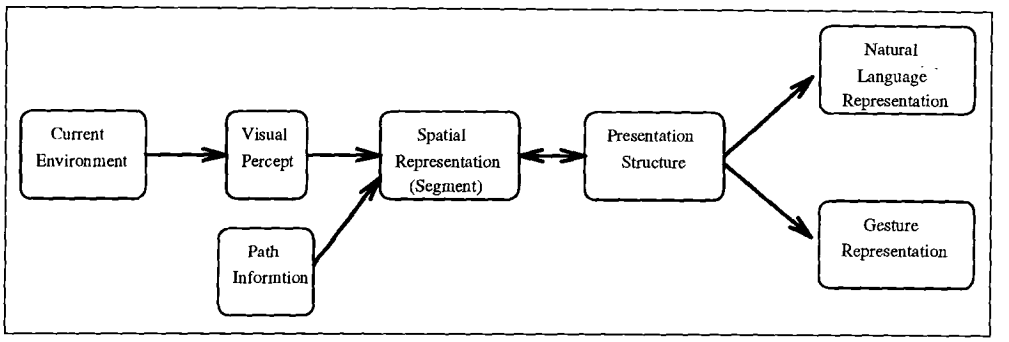

Figure 2: The representation structure model

tion processes are supervised and coordinated by a planning process (see figure 1). In general, all subprocesses depend on temporal aspects with respect to the communication coordination, on presumed properties of the questioner, and on environmental parameters such as weather, time, and brightness.

Associated with each process are unique representation structures. Visual perception receives information from the current environment. We follow Neisser's proposal that the visual perception is led by schemata which enable us to select information even in complex and dynamic environments ((Neisser 76)). We use the term *visual percept* to denote an object-oriented representation structure (see figure 2). Therefore it seems to be reasonable to consider only approximated shape properties of an object when computing spatial relations (cf. (Landau & Jackendoff 93)). In most cases, it is sufficient to approximate the object to be localized with its center of gravity, because its position is the only necessary information that counts for the applicability of the spatial relation. In our system we are using the following simplifications at the moment:

(1) *Center of gravity.* (2) *2D representation*: The base of each object (Necessary when perceiving objects from a bird eye's view, e.g., in maps). (3) *Smallest circumscribing rectangle* (4) *Smallest circumscribing right parallelepiped* (5) *3D representation*: The complete description of an object. Beside a *geometric representation*, the representation of an object integrates a *conceptual representation* which includes size, color, and intrinsic properties of the object (cf. (G... p 94; Gapp & Maaß 94)).

The wayfinding process extracts path information from a map. We also use a dualistic representation. On one hand, AG has survey information for the whole path necessary to keep track of a global orientation and to provide abstract information about distance to the destination. On the other hand, AG matches this general information into an exact topographical partition of the path which leads AG's movement and description in the current time interval.

Spatial information from maps and observation is integrated by the planning process into a central spatial representation structure called, *segment*. Because of the incrementality of the whole process, this structure is extended step-by-step. Using this representation, the planning process extracts object information and relates relevant objects with one another to produce a presentation structure. This presentation structure determines the content of the verbal and gestic description.

Conclusion

For the integration of visual perception and natural language we have presented a computational approach in the area of incremental route descriptions. Although this is only a step towards integration of both fields, we have shortly presented how visually accessible knowledge of landmarks and knowledge of routes can be related to one another to act in the current environment. Specifically, we have focused on those processes and representations involved when generating incremental route descriptions: visual perception, wayfinding, planning, and presentation. The multi-agent architecture of our model is composed of two types of processes, processes which run continuously and those which are activated by demand. Within the spatial representation landmarks and routes are interrelated, and spatial relations are established. In order to give an adequate description, the presentation planner selects part of the obtained spatial information. The planner determines the contents of the description. The content is used by the activated speech and gesture generator for construction of the description.

The general visual identification process, specifically the determination of salience criteria, will be focused in the future. In addition, we will investigate how changes in the environment and movement of the speaker influence his/her behaviour in a physical world.

References

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.

R. J. **Elliott** and M. E. **Lesk**. *Route Finding in Street Maps by Computers and People.* In: Proc. of AAAI-82, pp. 258–261, Pittsburgh, PA, 1982.

K.-P. **Gapp** and W. **Maaß**. *Spatial Layout Identification and Incremental Descriptions.* submitted, 1994.

K.-P. **Gapp**. *On the Basic Meanings of Spatial Relations: Computation and Evaluation in 3D Space.* In: To appear in: Proc. of the 12 th AAAI-94, Seattle, WA, 1994.

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.

Ch. **Habel**. *Prozedurale Aspekte der Wegplanung und Wegbeschreibung.* LILOG-Report 17, IBM, Stuttgart, 1987.

W. **Klein**. *Deixis and Spatial Orientation in Route Directions.* In: H. L. Pick and L. P. Acredolo (eds.), Spatial Orientation: Theory, Research, and Application, pp. 283–311. New York, London: Plenum, 1983.

B.J. **Kuipers** and T.S. **Levitt**. *Navigation and Mapping in Large-Scale Spaces.* AI Magazine, pp. 25–43, 1988.

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

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

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

W. **Maaß**. *A Cognitive Model for the Process of Multimodal, Incremental Route Description.* In: Proc. of the European Conference on Spatial Information Theory. Springer, 1993.

W. **Maaß**. *From Visual Perception to Multimodal Communication: Incremental Route Descriptions.* AI Review Journal, 1994. Special Volume (Issues 1,2,3): Integration of Natural Language and Vision Processing, forthcoming.

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

U. **Neisser**. *Cognition and Reality.* San Francisco: Freeman, 1976.

J. **Pailhous**. *La Representation de l'Espace Urbain - L'example du Chauffeur de Taxi.* Presses Universitaires de France, 1970.

J. **Piaget** and B. **Inhelder**. *The childs conception of space.* New York: Norton, 1967. Originaly published in French, 1948.

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.

D. **Wunderlich** and R. **Reinelt**. *How to Get There From Here.* In: R. J. Jarvella and W. Klein (eds.), Speech, Place, and Action, pp. 183–201. Chichester: Wiley, 1982.