Artificial Intelligence Review **8**: 159–174, 1994.

From Vision to Multimodal Communication: Incremental Route Descriptions*

WOLFGANG MAAß

Cognitive Science Program, Universität des Saarlandes, D-66041 Saarbrücken, E-mail: maass@cs.uni-sb.de

Abstract. In the last few years, within cognitive science, there has been a growing interest in the connection between vision and natural language. The question of interest is: How can we discuss what we see. With this question in mind, we will look at the area of *incremental route descriptions*. Here, a speaker step-by-step presents the relevant route information in a 3D-environment. The speaker must adjust his/her descriptions to the currently visible objects. Two major questions arise in this context: 1. How is visually obtained information used in natural language generation? and 2. How are these modalities coordinated? We will present a computational framework for the interaction of vision and natural language descriptions which integrates several processes and representations. Specifically discussed is the interaction between the spatial representation and the presentation representation used for natural language descriptions. We have implemented a prototypical version of the proposed model, called MOSES.

Key words: spatial cognition, wayfinding, multimodal presentation, object representation.

1. INTRODUCTION

Recently in cognitive science there has been a growing interest in the connection between vision and natural language. The reductionist approach to separate each of these topics tends to ignore the fact that human beings integrate both abilities to act and communicate in their environment.

In the project VITRA (Visual Translator), we are investigating the connection between vision and multimodal communication in dynamic environments from a computational point of view (cf. Herzog and Wazinski, this volume). In a subproject, we are specialising on the generation of multimodal incremental route descriptions which combine natural language and spontaneous gestures[1] (cf. Maaß 1993). The main question for this kind of communication is how to lead a person to his/her destination by describing the route during the trip itself. Then persons were asked to give incremental route descriptions always used different modalities, e.g. speech and spontaneous gestures, to describe the route. Thus, two of the interesting subquestions are: (1) How is visually obtained information used in natural language production? and (2) How are these modalities coordinated? Our model, called MOSES, is led by the psychological results

of (cf. Allen and Kautz 1985). This report states that humans mentally partition the continuous and complex spatial environment into segments of information. These segments represent a partial view of the actual environment. We propose two selection steps: the visual selection and the presentation selection. These selections reduce the complexity of information and the complexity of computation, making it possible to describe the route efficiently.

In Section 2, we suggest some of the major concerns which are central for incremental route descriptions, followed by a discussion of related research. For the process of incremental route descriptions, we have designed a computational model which is presented in Section 3. Two separate input processes determine the behaviour of the entire process: the visual recognition (Section 3.1) and the determination of path information obtained from maps (Section 3.2). The interaction between the spatial representation and the linguistic representation, underlying the presentation processes is presented in Section 3.3. In Section 4.1 we demonstrate how to plan the presentation structures in order to effectively communicate the obtained information. The modespecific generators for natural language and gestures are briefly mentioned in Section 4.2 and, in Section 5, we give a conclusion and a projection on some open questions.

2. MOTIVATION

Route descriptions are common communicative actions in 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 whole route we use complete route descriptions. Here, a well-known problem for the route finders is remembering many details at one time. En route to their destination, they normally cannot ask the same person for more details. In *incremental route descriptions*, e.g., descriptions given by a co-driver, the route finders receive relevant route information as it is needed. This reduces the cognitive load. Central to incremental route descriptions are temporal constraints on both the generation and following of route descriptions. The construction of a presentation involves, a minimum of the following phases: determination of new information, determination of a presentation structure, transmission of the information, and consideration of the length of time the hearer will presumably require to understand and verify the information. Furthermore, the information must be presented in accordance with the strengths of each presentation mode, while taking into account the information to be presented and the current environment. In this work, we address the first two phases. In the scenario considered here, the speaker, SP, and hearer, H, travel by car in an urban environment. The task for SP is to give adequate information to H. With the term *adequate* we mean that SP must determine what is relevant to the hearer.

In this project we are not concerned with the use of long-term mental representations of spatial information, often called *cognitive maps*, but rather with the phenomena which arise when the speaker uses both a map and visible information of the current environment to determine and describe a route in unknown