A Communicative Mediator in a Virtual Environment: Processing of Multimodal Input and Output

Timo Sowa, Stefan Kopp & Marc Erich Latoschik
Faculty of Technology, AI Group
University of Bielefeld
33501 Bielefeld, Germany
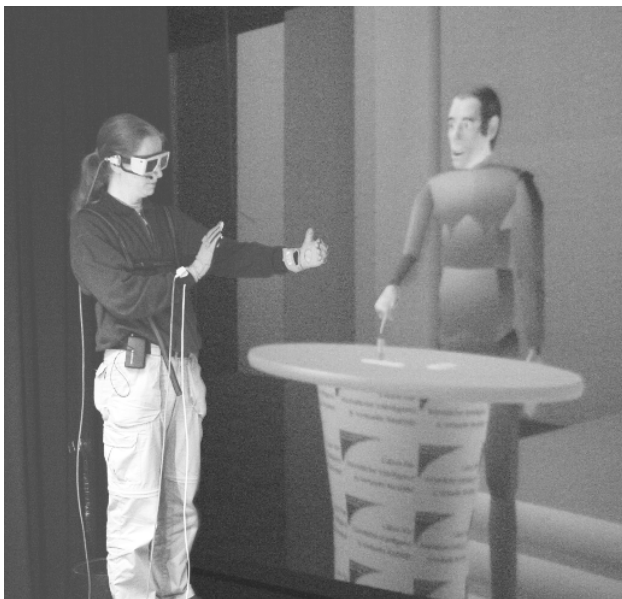+49 521 106 2921/19
{tsowa, skopp, marcl}@techfak.uni-bielefeld.de

Figure 1: Multimodal interaction with Max.

1 Abstract

This paper presents work on multimodal communication with an anthropomorphic agent. It focuses on processing of multimodal input and output employing natural language and gestures in virtual environments. On the input side, we describe our approach to recognize and interpret co-verbal gestures used for pointing, object manipulation, and object description. On the output side, we present the utterance generation module of the agent which is able to produce coordinated speech and gestures.

1.1 Keywords

multimodal communication, virtual environments, anthropomorphic agents

2 Introduction

In this paper, we give an overview of *Max* (Fig. 1), a virtual anthropomorphic agent which acts as a mediator in an immersive 3D virtual environment for simulated assembly and design tasks: The user may instruct the agent with natural multimodal utterances or may interactively manipulate the scene. The agent provides feedback combining facial and upper limb gestures with spoken utterances yielding a natural multimodal communication between the human user and the system. Max has some "expert" knowledge about construction tasks and is able to demonstrate assembly procedures to the user.

This research is carried out in a wider context of situated communication in a construction scenario [3]. It also builds on earlier work of employing an anthropomorphic agent to mediate situated language instructions which make use of varying spatial reference systems [2]. In the focus of this paper is the processing of multimodal input and output by the verbal and gestural facilities of the communicative mediator. The integrated system as conceived in Fig.2 is subject of ongoing work. Fully implemented modules for input processing and output generation as well as empirical foundations are described in the following sections.

3 Empirical Studies

A natural and intuitive multimodal interface should be designed in accordance with observable communicative behavior of humans. Besides exploiting insights from psychological and linguistic research, we conducted own empirical studies to collect data about the use of gesture and speech in our application scenario [8]. The studies concern the spatiotemporal expression as well as semantic aspects of co-verbal gestures. In the first study subjects were asked to name and to point at simple geometrical objects. The qualitative evaluation revealed information about the spatiotemporal expression of pointing gestures and possible criteria to segment gesture phases. In the second study subjects were told to describe parts from our virtual construction application. We evaluated the relation between gestural expression and gesture semantics by analyzing the way subjects employed gestural form features like hand shape or movement to express geometrical properties of the stimulus object like extent, roundness, etc. Another study using the same setting was conducted to approach the problem of gesture segmentation quantitatively, i.e. to determine the meaningful part *(stroke)*

of a gesture through the analysis of motion data. The evaluation is currently in progress.
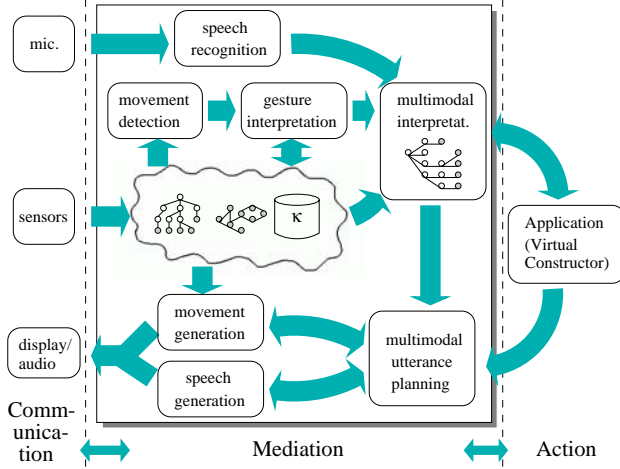

Figure 2: Overall structure of the mediator system.

The qualitative and quantitative results are used to build models on multimodal processing and to verify our assumptions about speech and gesture-based communication. Furthermore, the corpus data is available to evaluate the performance of the recognition and generation systems.

4   Input Processing

Input processing consists of four main components which access a common scene and knowledge base like illustrated in Fig. 2. The scene/knowledge base uses three different structures for the information encoding. 1) a tree expresses spatial relations of user movement and objects and enables a seamless integration into immersive VR setups. 2) a semantic net incorporates object knowledge as well as linguistic information and 3) a graph structure is used to represent shape properties of objects.

4.1   Gesture Detection

Gesture detection is closely embedded into the permanently changing scene representation. The PrOSA (Patterns On Sequences of Attributes) framework [5] consists of a sensor abstraction layer composed of so-called *actuators* that generate movement data w.r.t. a common reference frame and in a uniform rate. This data is processed in real-time by *detectors*, small calculation units that can be connected to larger *detector nets* which analytically search for specific gesture features, e.g. curvature or hand shapes.

4.2   Gesture/Speech Integration and Interpretation

Gesture interpretation is carried out using two different approaches. For the ongoing user movement in the virtual scene, *reference rays* represent significant body features, e.g., the view/pointing directions, palm normals, or reference systems. *Spacemaps* temporarily store pre-processed parameters like relative linear or angular

distance between these rays and scene objects to later reconstruct indexical integrity of deictic speech/gesture utterances.
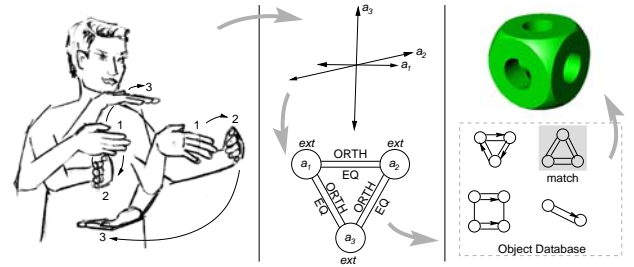

Figure 3: Interpretation of a complex iconic gesture.

The interpretation of shape-related (iconic) gestures rests upon the basic assumption that iconic gestures are *similar* to the referent they describe. In contrast to most gesture recognition approaches which directly map a gestural expression onto meaning, our model decomposes meaningful upper limb movements into shape properties. These properties represent an abstract geometrical description of the gesture that is independent from a particular realization. The property "roundness", for example, may be indicated with the thumb and index finger shaping an "O", or with the index finger tip moving on a circular trajectory. The geometrical gesture model can then be matched against a set of object models to determine the most similar object (Fig. 3). The gesture/object model is internally represented as a graph in which nodes represent shape properties and links spatial relations. The detection of similarity is performed by subgraph matching. With this approach, the decomposition of meaning is not limited to a single gesture. Properties may accumulate over a series of movements and postures as shown in the example (Fig. 3) where the idea of a cube is expressed in three gesture phases.

Multimodal integration and interpretation is carried out using an *enhanced ATN* for a combined syntactic/semantic analysis. It uses additional state transition timing information to allow continuous processing of parallel occurring time-stamped utterances. Scene and application context related information is incorporated by appending query functions as state transition constraints. The integration results are passed to the application and to the utterance planning module.

5   Utterance Generation

Output generation starts from a concise specification of the desired multimodal utterance. The overall process comprises two interacting main stages: Generating verbal/nonverbal parts and coordinating them (see Fig. 2).

5.1   Utterance Specification

We developed a XML-based specification language which provides flexible means of expressing multimodal utterances in a given context. Such descriptions are

centered around the verbal utterance (in German) which can be augmented by several behaviors. To this end, certain points in time during the spoken utterance are defined by marking up the textual output. Timing of each explicitly represented behavior is then specified w.r.t. the appropriate time points such that behaviors can occur subsequently or overlap in time. In our current system, we adopt the empirically suggested assumption of an one-to-one correspondence between a single gesture and some sort of intonational unit (cf. [6]). The verbal part of a complex multimodal utterance that comprises multiple gestures must therefore be divided in *chunks* by annotating the corresponding time tags.

Gestural behaviors can be defined by the specification of a required communicative function sufficient for the agent to choose an appropriate behavior from its lexicon of XML-compliant gesture representations. Alternatively, the desired gesture can be explicitly defined in terms of its main spatiotemporal features using our gesture markup language.

The following example utterance comprises two deictic gestures. The first one is defined explicitly by hand shape and finger orientation, which is required to point to a target location (bound to a certain object's position). The second gesture is specified in terms of a communicative function ('refer_to_loc').

```
<definiton>
  <parameter name="target_location_1" default="bolt-1"/>
  <parameter name="target_location_2" default="bar-1"/>

  <utterance>
    <specification>
    And now take <time id="t1"/> this bolt <time id="t2"
    chunkborder="true"/> and put it into <time id="t3"/> this
    bar. <time id="t4"/>
    </specification>
    <behaviorspec id="gesture_1">
      <gesture>
        <constraints><parallel>
          <static slot="HandShape" value="BSifinger"/>
          <static slot="ExtFingerOrientation" value=
            "$target_location_1" mode="pointTo"/>
        </parallel></constraints>
      </gesture>
      <timing><onset id="t1"/><end id="t2"/></timing>
    </behaviorspec>
    <behaviorspec id="gesture_2">
      <gesture>
        <function name="refer_to_loc">
          <param name="refloc" value="$target_location_2"/>
        </function>
      </gesture>
      <timing><onset id="t3"/><end id="t4"/></timing>
```

```
    </behaviorspec>
  </utterance></definition>
```

Additional behaviors include arbitrary bodily movements, defined as parametric keyframe animations combined with ease in/out, and facial animations given as sequences of face muscle values.

5.2   Gesture/Speech Generation

Our work on synthesizing multimodal output so far focussed on generating gestural and verbal behaviors with particular emphasis on how to achieve temporal coordination. The demand for naturalness of the agent's movements has most often led to creating behaviors beforehand which are either captured from real humans or manually predefined. Yet, the employed techniques do not provide a satisfactorily high degree of adaptability to varying movement constraints as found in co-verbal gesture (cf. [1]). Therefore, we developed a model for creating gesture animations on-the-fly that reproduces major characteristics of human movement and provides sufficient flexibility. Our approach combines movement planning (described in [4]) with execution: A motor program applies a certain number of *local motor programs (LMP)* simultaneously which have been instantiated, prepared, and arranged during planning. Each LMP is able to activate and complete itself at run-time ensuring continuous motion. In addition, LMPs are able to pass control between each other. LMPs control submovements over a designated period of time employing suited motion generation methods, most of them well-known in computer animation. For forming arm trajectories a new method based on non-uniform B-Splines was developed targeting at the simulation of kinematic properties of human gestural movements like characteristics of ballistically executed phases. In particular, the dynamics of the stroke can be adjusted by way of timing the velocity peak which helps to create synchronized accentuation of verbal and nonverbal output as found in multimodal utterances.

In its current state, Max is able to produce complex multimodal utterances with exactly synchronized verbal and gestural parts. The assignment of a co-verbal gesture in the XML specification affects the intonation of the verbal utterance due to the fact that a recognizable contrastive stress in speech serves as a synchronization point for the gesture's timing [7]. To this end, a set of SABLE[1] tags is utilized to tag words or syllables to be emphasized in speech. A text-to-speech system (described in [9]) was developed to control prosodic parameters (e.g., speech rate and intonation) in order to pre-plan the timing of stressed syllables. Timing information up to phoneme-level is utilized to compose lip-synchronous speech animations as well as to complete the time definition of the

---

[1] SABLE is an international standard for marking up text input to speech synthesizers.

accompanying gesture which is then created dynamically. Within a single chunk we apply an *absolute-time-based* scheduling as proposed by Cassell et al. [1] and set the onset of the gesture stroke to precede the onset of the corresponding linguistic element by approximately one syllable's duration (0.2 ms). The stroke is set to span the whole phrase framed by the corresponding time tags before retraction starts. Once a chunk has been completed (verbal part and gesture *stroke* fully performed), the execution of the next chunk is started by completing and activating the corresponding LMPs. Thus, co-articulation effects like fluent gesture transitions emerge from activation of the subsequent gesture, resp. its LMPs, before the preceding one has been fully retracted. Additional synchronization mechanisms, including the integration of explicit holds in the gesture, are to be evaluated in ongoing work.

REFERENCES

1. Cassell, J., Vilhjálmsson, H., and Bickmore, T. BEAT: the Behavior Expression Animation Toolkit. In *Proceedings of the 2001 conference on Computer Graphics* (*SIGGRAPH)*, pp. 477-486, 2001.

2. Jörding, T., and Wachsmuth, I. An anthropomorphic agent for the use of spatial language. In Coventry, K., and Olivier, P., editors, *Spatial Language: Cognitive and Computational Aspects,* chapter 4. Kluwer, Dordrecht, 2001. in press.

3. Jung, B., Latoschik, M.E., and Wachsmuth, I. Knowledge-based assembly simulation for virtual prototype modeling. In *Proc. of the 24th Annual Conf. of the IEEE Industrial Electronics Society – IECON '98,* 1998, pp. 2152-2157.

4. Kopp, S., and Wachsmuth, I. A knowledge-based approach for lifelike gesture animation. In Horn, W., editor, *Proc. of the 14th European Conf. on Artificial Intelligence – ECAI 2000,* pp. 661-667, Amsterdam, 2000. IOS Press.

5. Latoschik, M.E. A gesture processing framework for multimodal interaction in virtual reality. In *AFRIGRAPH 2001 conference proceedings, 2001.* To be published.

6. McNeill, D. *Hand and mind: What gestures reveal about thought*, Univ. of Chicago Press, 1992.

7. de Ruiter, J.-P. *Gesture and Speech Production.* Ph.D. thesis, Katholic University Nijmegen, 1998.

8. Sowa, T., and Wachsmuth, I. Coverbal iconic gestures for object descriptions in virtual environments: An empirical study. *Technical Report 2001/03, Collaborative Research Center "Situated Artificial Communicators" (SFB 360)*, University of Bielefeld, 2001.

9. Wachsmuth, I., and Kopp, S. Lifelike gesture synthesis and timing for conversational agents. *Proc. of Gesture Wokshop 2001,* to be published.