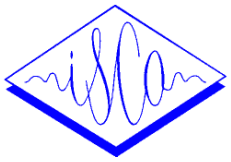
ISCA Archive
http://www.isca-speech.org/archive



ESCA Workshop on Interactive
Dialogue in Multi-Modal Systems (IDS-99)
Kloster Irsee, Germany
June 22-25, 1999

Understanding Coverbal Dimensional Gestures in a Virtual Design Environment

Timo Sowa & Ipke Wachsmuth

Faculty of Technology, AI Group
University of Bielefeld
P.O. Box 100131, 33501 Bielefeld, Germany
{tsowa, ipke}@techfak.uni-bielefeld.de

ABSTRACT

Todays multimodal systems, which allow full-body (3D) gestures and speech as input modalities, are quite restricted to easily interpretable coverbal gestures with a predefined shape and meaning. In this paper, we propose methods to abstract the concrete shape of gestures by using high-level features and to integrate them with coexpressive words using their phonological attributes. The application of this approach is discussed for a class of gestures useful in virtual design. We sketch our technical environment and first implementation approaches to build a prototype system.

1 INTRODUCTION

The use of gestures as an input modality for a multimodal system is particularly suggestive where the user wants to manipulate spatial properties of objects or spatial relations between objects. If such manipulations are performed in a virtual environment in which the user may move freely (for example in a CAVE), the use of traditional input methods – e.g. clicking and pointing with a mouse – is very obstructive. Speech and full-body (3D) gestures seem a much more natural basis for an interface to such environments [5]. Our research in this area focused mainly on object reference mediated by pointing gestures and manipulations like rotation and translation. With the current work, our aim is to enlarge the interface capabilities by adding more complex types of gestures and by investigating details about their connection to simultaneous speech input.

2 MULTIMODAL UTTERANCES

To obtain the users intention from how he or she behaves, including the analysis of all bodily movements and spoken utterances, can be regarded as the ultimate goal of building multimodal input interfaces. Although closely interwoven, the non-verbal and verbal aspects of communication show different characteristics with respect to their communicative abilities.

2.1 Verbal Communication

Verbal communication can be, roughly speaking, divided up into the questions "What is spoken?" and "How is it spoken?". The first question refers simply to the chain of words uttered by a person, whereas the second question deals with phonological issues like syllable/word stress and intonation or pausing. Words have a predefined form and meaning that can be looked up in a lexicon. Speech recognizers can use these lexical entries to determine if the "sound" from the microphone input was a word. A composition of the words by syntactic rules produces whole sentences.

Phonological attributes that modulate the word chain are used to emphasize important information in the flow of speech, for example to mark the *rheme*, the new contribution of a speaker in a discourse. The illocutionary point of an utterance, i.e. the distinction between a question, a command, or a statement, is often marked by a special intonation pattern, like a rising intonation in questions.

Recognizing and understanding verbal communication with a machine has a long tradition in pattern recognition and artificial intelligence. Despite the fact that user-independent, continuous speech recognition is now commercially available, speech understanding remains a problem far from being entirely solved. In particular, the semantic and pragmatic evaluation in consideration of application context, discourse context, and background knowledge are major factors contributing to the overall complexity of the task.

2.2 Nonverbal Communication

According to [1] bodily movements serve different functional roles, i.e. the *epistemic* function to obtain a representation of the environment by tactile feedback, the *ergodic* function to manipulate something in the physical world and, finally, the *semiotic* function to communicate meaningful information, with the latter being the most important for our purposes. Here we use the term *gesture* for movements of the upper limbs with a semiotic function.

Gestures can express nearly every aspect of an idea to be communicated [6]. In contrast to speech there is no set of predefined gestures except for emblematic gestures which share a meaning in a culture or in a social group. Since humans can astonishingly well separate gestures from other movements, there have to be cues that we use to make this distinction. Another difference to speech is the lack of a gesture syntax.

These aspects make it difficult to ascribe an intention to a single gesture apart from discourse context and concurrent speech. Nevertheless, the form of a gesture can be described by using gesture features.

2.2.1 Gesture Features

The use of a limited feature-set to describe a gesture is one way to cope with the vast diversity of gestures. Features can be composed to describe arbitrary gestures, for example a pointing gesture description may look like: "stretched arm, long hand-body distance and a stretched index finger". Besides the concrete description we propose to abstract from the concrete shape of a gesture by using high-level features, like symmetry properties, which have the power to describe bigger classes of gestures conveying the same content. Thus, we move away from the simple shape-to-meaning mapping that is often found, but also criticised, in gesture detection systems [10].

2.2.2 Features and Concepts

Since a multimodal interface serves as a mediator between the user and the application system, the final product of the integration and interpretation stage is a command or an expression, that the application system can understand. The command language depends of course to a high degree on the application domain. Consequently, a multimodal interface includes domain-specific concepts which can be eventually expressed in a natural way by using special classes of gestures. The significant high-level features of these gesture classes have to be linked to the application concept by the interface designer.

2.3 Cross-Modal Interpretation

The information given through one modality has to be taken into account in the interpretation process for other modalities. This is especially the case for speech and gesture. Thus, the semantically coexpressive parts of speech and gesture have to be determined and integrated. Suitable means to solve this *correspondence problem* [9] have to be found for every multimodal system which allows "natural" input.

The most accentuated part of a gesture, in which the meaning is expressed, is called the gesture *stroke* [6], [4]. The stroke often coincides with the phonologically prominent syllable in speech. The detection of the gesture stroke is of great importance for a multimodal input system, but, unfortunately, it is difficult to find a formal, computable formulation of the stroke concept in general. In some gesture classes the stroke is expressed by quite easily detectable features, like a strong acceleration of the wrist in the climax (accented part) of pointing gestures. Together with the phonological information, such features could be used to detect gestures in a technical system.

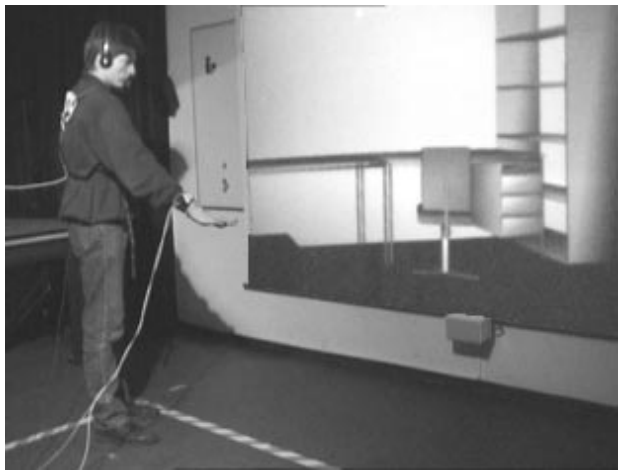


Figure 1: One-handed coverbal dimensional gesture: "Make the desk this high." (to be used in interaction with a virtual design environment)

3 SETTING AND APPROACH

Our aim is to apply the above ideas to the task of gesture and speech input integration in a virtual design environment. Our research scenario consists of a 3D scene presented on a wall-size display, which is manipulable by the user. We concentrate on size modifications of the objects in different dimensions. Figures 1 and 2 show two examples of this type of interaction. We will henceforth call such gestures *dimensional gestures*. We decided to investigate coverbal dimensional gestures because they are useful in our application scenario. As their concrete shape is multifaceted, high-level features and concepts are needed for detection and integration.

An overview of our planned prototype system is shown in figure 3. The implementation is currently in progress; dashed lines indicate the still unfinished parts of the system.

3.1 Hardware

We use three 6DOF electro-magnetic position sensors and two data-gloves for gesture detection. Two position trackers are mounted at the wrists[1], one at

[1]The mounting position is proximal to the wrist, therefore motions in the wrist do not influence the tracker.

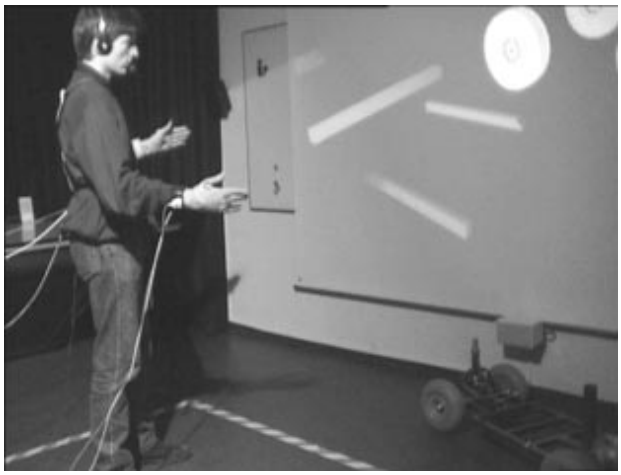


Figure 2: Two-handed coverbal dimensional gesture: "Make the rod this long."

the neck of the user. The data-gloves measure handshapes and wrist angles. Speech is captured with a microphone headset. Figure 4 shows a user wearing the sensor devices.

3.2 Body-Model

Except for very simple gestures a recognition system cannot be limited to tracking the posture and position of just one hand. The meaning of a gesture often depends on the relative position or movement with respect to the body. Deliberate gestures, for example, are often performed in a limited signing space in front of the chest. Additionally, the body provides a reference system for measurements in the virtual scene. Hence, there is a need for a model of the user's body.
A body model that meets our requirements was developed in our working group [2]. It uses the the positional and directional data from the three position sensors and solves the inverse kinematic problem via a recurrent net. The output of the body model describes the hand position and movement in symbols, using the gesture notation system *HamNoSys* [7].

3.3 Stroke Recognition

Although it is no problem for a human observer to detect the stroke in a gestural utterance, this task is generally not trivial for a machine. Currently, we investigate three cues in hand and arm movement data that indicate strokes in many gestures classes. The first two are motivated by the kinesic structure of gestures, reported in [4]: A very slow velocity of the hand that sometimes occurs after a stroke (the so-called post-stroke hold), and abrupt changes in the movement direction of the hand, which are consequences of an immediate retraction after a very short stroke phase. As a third cue we evaluate the hand tension,

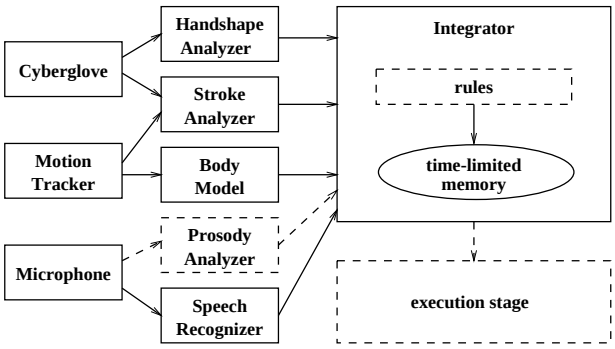


Figure 3: An overview of the planned prototype

which reaches a maximum in the expressive phase of many gestures. To compute the hand tension from the glove data we use a model proposed in [3].

3.4 Symmetry in Two-Handed Gestures

Figure 2 illustrates the use of a two-handed dimensional gesture. The intended object size is indicated by the space between the two "enclosing" hands, communicating a concept of distance or size. Besides the diversity of this gesture type (e.g. different distances, different hand orientations, etc.), there is the common property of *symmetry* as a high-level feature. Observations reveal a symmetry in hand shape and in movement which includes the co-occurrence of the stroke in both hands. Thus, the detection of symmetry is a further cue leading from motion data to a size concept.

3.5 Speech & Phonology

Currently, the speech recognition is handled by a user-independent continuous speech recognizer, a research prototype developed by the Applied Computer Science Group in our faculty. The recognizer provides the application with the detected words and their appropriate timestamps (referring to speech-onset).
In the future we plan to complement the system by prosody analyzers that detect stresses and intonation patterns.

3.6 Speech-Gesture Integration

We exploit the temporal relation between speech and gesture components of the multimodal utterance to obtain its meaning. In the interaction shown in fig.1, for example, the stroke of the dimensional gesture coincides with the word *this*. From speech context ("make ... this high") a height modification can be inferred, and the temporal relation between stroke cues and speech-onset of the word "this" indicates coexpressive gesture and speech, so that the gesture can be interpreted as the quantitative aspect of the intended height modification.

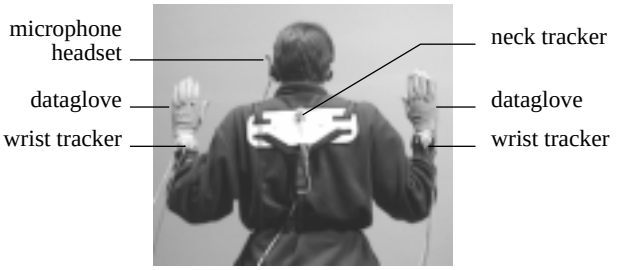


Figure 4: Sensor devices for gesture and speech detection

A system for multi-level temporal integration of symbolic data is used to perform the speech-gesture integration task [8]. The mechanism rests on the assumption that the relevant information on a particular level of integration is limited to a temporal window. Hence, the integrator module needs only watch the content of the integration window. To control the temporal chunks, either a fixed-size window can be used or segmenting signals can be given by external events indicating the beginning of a new temporal window, e.g. intonation, stroke, holds, etc.

Knowledge about the integration is represented by explicit rules in our integrator mechanism. The precondition of each rule requires the existence of base symbols (which represent input information) and a temporal relation between them. If the precondition if satisfied (e.g., symmetric handshape and position + stroke + word "high" are present concurrently), a new symbol (e.g., height modification) is produced.

4 CONCLUSION

In this paper, we argued in favor of an abstract view on coverbal gestures that is not bound to a concrete form. With this approach more complex classes of gestures can be recognized by the interface system. Furthermore, we proposed to make use of phonological information from speech to solve the speech-gesture correspondence problem.

The completion of our prototype system as it is shown in figure 3 is our next aim. This includes the implementation of a prosody analyzer, the elaboration of a rule set for speech-gesture integration and a connection to our graphical output system.

On the empirical side we plan experiments in our environment to gain more exact and meaningful data about the types of gestures used in our scenario, and their temporal relation to speech. Based on these data we can refine the integration method with the establishment of new – or more differentiated – integration rules. Further work has to be done to detect similarities in the shape of dimensional gestures, i.e. how different users express the "distance concept" with gestures. Regarding two-handed gestures, the detection of symmetry seems a good starting point.

References

[1] J. L. Crowley and J. Coutaz. Vision for man machine interaction. In *Proceedings of Engeneering Human Computer Interaction*, pages 28–45, Grand Targhee, USA, Aug. 1995. EHCI'95, Chapman and Hall, Ltd.

[2] M. Fröhlich and I. Wachsmuth. Gesture recognition of the upper limbs - from signal to symbol. In *Gesture and Sign Language in Human-Computer Interaction: Proceedings of Gesture Workshop '97*, pages 173–184, 1997.

[3] P. A. Harling and A. Edwards. Hand tension as a gesture segmentation cue. In *Progress in Gestural Interaction: Proceedings of Gesture Workshop '96*, pages 75–87, 1996.

[4] A. Kendon. Current issues in the study of gestures. In Nespoulous, Perron, and Lecours, editors, *The Biological Foundations of Gestures: Motor and Semiotic Aspects*. Lawrence Erlbaum Associates, Hillsday N.J., 1986.

[5] M. Latoschik, M. Fröhlich, B. Jung, and I. Wachsmuth. Utilize speech and gestures to realize natural interaction in a virtual environment. In *IECON'98 - Proceedings of the 24th Annual Conference of the IEEE Industrial Electronics Society*, volume 4, pages 2028–2033. IEEE, 1998.

[6] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.

[7] S. Prillwitz, R. Leven, H. Zienert, T. Hamke, and J. Henning. *HamNoSys Version 2.0: Hamburg Notation System for Sign Languages: An Introductory Guide*, volume 5 of *International Studies on Sign Language and Communication of the Deaf*. Signum Press, Hamburg, Germany, 1989.

[8] T. Sowa, M. Fröhlich, and M. Latoschik. Temporal symbolic integration applied to a multimodal system using gestures and speech. To appear in the Proceedings of the Gesture Workshop '99, March 17-19th, Gif-sur-Yvette, France.

[9] R. K. Srihari. Computational models for integrating linguistic and visual information: A survey. *Artificial Intelligence Review*, 8:349–369, 1994.

[10] A. D. Wexelblat. An approach to natural gesture in virtual environments. *acm Transactions on Computer-Human Interaction*, 2(3):179–200, 1995.