Interpretation of Shape-Related Iconic Gestures in Virtual Environments

Timo Sowa and Ipke Wachsmuth

Artificial Intelligence Group, Faculty of Technology, University of Bielefeld
D-33594 Bielefeld, Germany
{tsowa,ipke}@techfak.uni-bielefeld.de

Abstract. So far, approaches towards gesture recognition focused mainly on deictic and emblematic gestures. Iconics, viewed as iconic signs in the sense of Peirce, are different from deictics and emblems, for their relation to the referent is based on similarity. In the work reported here, the breakdown of the complex notion of similarity provides the key idea towards a computational model of gesture semantics for iconic gestures. Based on an empirical study, we describe first steps towards a recognition model for shape-related iconic gestures and its implementation in a prototype gesture recognition system. Observations are focused on spatial concepts and their relation to features of iconic gestural expressions. The recognition model is based on a graph-matching method which compares the decomposed geometrical structures of gesture and object.

1 Introduction

Gesture as a communicative modality that supplements speech has widely attracted attention in the field of human-computer interaction during the past years. Applications benefit in various ways from the multimodal approach that combines different kinds of communication. The synergistic use of gesture and speech increases the bandwidth of communication and supports a more efficient style of interaction. It potentially increases naturalness and simplicity of use, since oral and gestural competences are present by nature, while WIMP[1] competence has to be learned. In virtual environments, where users may move around, it can be even impossible to use WIMP-style interfaces. Gesture and speech complement each other, so that some concepts we wish to communicate are more easily expressed in one modality than in the other. Speech, the linear and structured modality, is advantageous for abstract concepts, while gesture, as an inherently space-related modality, supports the communication of concrete and spatial content. This property of gesture makes its use in an interface particularly helpful for space-related applications.

The interpretation of iconic gestures in spatial domains is a promising idea to improve the communicative capabilities of human-computer interfaces. This paper presents an approach to utilize co-verbal iconic gestures as information

[1] Windows, Icons, Menus, and Pointing interfaces

I. Wachsmuth and T. Sowa (Eds.): GW 2001, LNAI 2298, pp. 21–33, 2002.

carriers about shape properties of objects. The application background of our research is the area of virtual design and construction, in which spatial concepts play a significant role. In particular, we focus on an intuitive interaction with the machine that (ideally) requires no learning. To approach this aim, we base the system design on the results from an empirical study about object descriptions. Before we describe this in detail, we give a brief analysis of the nature of iconic gesture and review related research.

1.1 Iconic Gestures and Shape

In the gesture typology by McNeill [11], shape-related or shape-describing hand movements are a subset of *iconic* gestures. This gesture type is used simultaneously with speech to depict the referent. Semiotically, the term *iconic* is derived from the notion of icons according to the trichotomy of signs suggested by Peirce [17]. An icon obtains meaning by *iconicity*, i.e. similarity between itself and its referent. Meaning thus becomes an inherent part of the sign.[2] The second and third type of the sign typology are index and symbol. Both require additional knowledge to be intelligible. For indices this is a shared situation and for symbols a shared social background.[3] Iconicity in gestures is multifaceted, it may occur with respect to geometric properties of referents, spatial configurations, or actions.[4] Our modelling approach is confined to the first facet, the representation of spatial properties.

The problem of gesture recognition from the viewpoint of engineering is quite often reduced to a pattern classification task. Following this paradigm, a given data stream (images in the case of video-based recognition or data from other sensor types) is classified according to a pre-defined set of categories. The class then determines the meaning of the gesture resulting in a direct mapping of gestural expression onto meaning. This engineering viewpoint is put forward, for example, by Benoit et al. [1]:

> "[The recognition of] 2D or 3D gestures is a typical pattern recognition problem. Standard pattern recognition techniques, such as template matching and feature-based recognition, are sufficient for gesture recognition." (p. 32)

The (pure) pattern recognition approach works well for emblems and application-dependent gestures which are defined in a stylized dictionary. However, the iconic gestures we consider in our approach are not part of an inventory or a standardized lexicon. Iconicity may refer to any spatial property or configuration one can think of in a given scenario. The number is potentially infinite, likewise

[2] This claim presupposes that sender and receiver share similar cognitive capabilities to perceive the sign and to mentally transfer it to its referent.

[3] The gestural correlates of indices and symbols are, respectively, deictic gestures and emblems.

[4] This subdivision is used in [15] to define pictomimic, spatiographic, and kinemimic gestures.

are the possibilities of a gestural realization. Even identical content may be expressed differently due to personal preferences. This fundamental property of iconic gestures, and co-verbal gestures in general, is expressed by McNeill [11]:

> "The gestures I mean are the movements of the hands and arms that we see when people talk ... These gestures are the spontaneous creations of individual speakers, unique and personal. They follow general principles ... but in no sense are they elements of a fixed repertoire." (p. 1)

With McNeill, we assume that for a computerized interpretation of iconic gestures to be successful and scalable, a paradigm different from pattern classification is necessary. The detection of similarity between gesture and referent requires an analysis of their components and their components' relationships.

1.2 Related Research

There is already a multitude of recognition systems for emblems and manipulative gestures used for tasks like robot control, navigation in virtual reality, sign language recognition, or computer games. Overviews of the research in this field were compiled, for example, in [19,16]. We found only very few approaches that take up iconic gestures in human-machine communication.

One attempt to recognize and interpret co-verbal iconic hand gestures was realized in the ICONIC system [7,21]. The prototype allows a user to interact with objects in a virtual environment, for example, to move an object with a speech command "move the teapot like this" and an accompanying dynamic gesture which indicates the direction. The interpretation of a gesture is speech-driven. Whenever speech suggested the possibility of a gesture ("like this" in the example), the system looked for an appropriate gesture segment that complements the spoken utterance. Static gestures are used to indicate places of objects, whereas dynamic gestures indicated movements. Iconicity with respect to shape properties was evaluated for the hand posture. It was correlated to the shape of the object referenced verbally so that hand rotation can be applied. The authors call the process of correlating object-shape and hand-shape *Iconic Mapping*. In the following sections, we will use the same term in a more generalized sense for the mapping between gestural expression and a reference object.

Though the ICONIC approach takes iconic gestures explicitly into account, its interpretative strategy is "caught" in language-like structures. A gesture, as a unit of meaning, augments and specifies just the semantics of the spoken utterance. Thereby, the approach does not consider the very nature of the imaginal aspects of gestures. This makes it impossible to model compositional properties where subsequent gestures use space in the same way to produce a "larger" gestural image.

On the empirical side, there is a lot of research on gestural behavior, but most of it is focused on different conversational situations, for example narrative discourse, specialized on certain gesture types, like deictics, or it considers isolated, non co-verbal gestures.

In her Ph.D. thesis [4], Hummels describes a study on the use of autonomous gestures for object design. It was found that different strategies like cross-section, extrusion, contour, usage mimics, and surface design were employed to visualize a given object. The author further emphasizes that there "is no one-to-one mapping between the meaning of the gesture and structural aspects of the hands" ([4], Section 3.53). There are yet several inter-individual consistencies with respect to prevalent combinations of posture and function. These include indication of height with two flat hands or between thumb and index finger, or the cutting of material using the flat hand.

Hummels' analysis is detailed and insightful, however it remains open in how far the results from autonomous gestures used there can be transferred to co-verbal gestures we are interested in.

The lack of empirical research on the use of co-verbal gestures in our target scenario caused us to conduct a study about the properties of shape-related gestures. A short summary of the main insights is given in the following section. More details can be found in [20]. In Section 3 we present an interpretation model for iconic gestures that considers iconicity and imaginal aspects. A prototype implementation for shape-related gestures is decribed in Section 4.

2 Properties of Shape-Related Gestures: A Study

As a basis for our modeling and implementation approach we conducted an empirical study. Its aim was to analyze the gestural behavior people naturally exhibit facing the task of object description in an experimental setting that closely resembles the targeted conversational situation. A total of 37 subjects were asked to describe parts from a virtual construction application displayed on a projection screen (Fig. 1). Subjects were equipped with electromagnetic 6DOF-trackers, mounted at the wrists and the neck. For hand-shape evaluation they wore data-gloves on both hands; and speech was recorded with a microphone headset (Fig. 2, left). The descriptions were video-taped from a frontal view (Fig. 2, right shows a snapshot from the video corpus). Although the use of the hands was mentioned in the instructions, gestural explanations were not explicitly enforced. Subjects were told to explain the objects' appearance in a

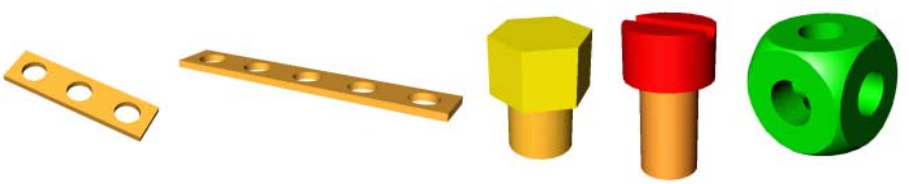

Fig. 1. The stimuli objects used in the empirical study

way that other subjects, who would watch the video afterwards, would be able to imagine the object.

As an outcome, our corpus data consists mainly of co-verbal gestures that were recorded with accompanying speech on a common time-line. The evaluation was focused on form features of the gestural expression and their relation to geometrical properties of the stimuli objects. In sum, our observations suggest that a limited set of gestural features exists to which spatial entities of the objects can be assigned. The majority of spatial entities are object extents in different dimensions, often, but not in each case, oriented like the stimulus object. The quantitative reproduction of extent is not always given, but it was comparable in relation to other extents of the object described. The significant features were linear and circular movements, the distance "vector" between the palms in two-handed gestures, the aperture of the hand in "precision-grip"-like gestures, the orientation of the palm, and a curved hand-shape. These gesture features can be combined to form a gesture like the one shown in Fig. 7 (cf. Section 4). The movement describes the dominant object's extent while the hand aperture indicates the subordinate extent. In this case, an elongated thin, or flat object could be meant.

Besides the parallel use of features in a single gesture, features may also occur sequentially in a series of gesture phrases. The cubical object (cf. Fig. 1; right), for example, was quite often described by three subsequent gestures indicating the extent in each spatial dimension. All three gesture phrases belong to a common gestural image, they form *one idea unit* [11,5]. We found that most people structurally decomposed the objects and described the parts one after another. The round-headed screw, for instance, was decomposed into the head, the slot on top, and the shaft. All parts were described independently, but their spatial relations were often retained, even beyond the scope of an idea unit. This cohe-

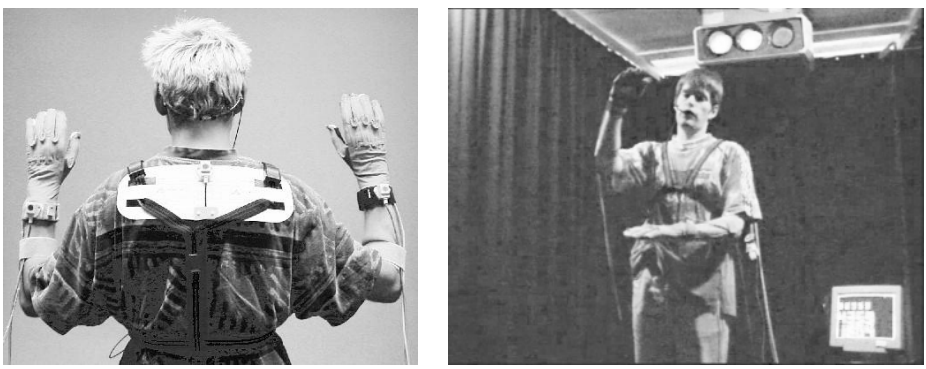

Fig. 2. Gesture and speech recording devices (left): Data gloves, motion trackers, and microphone headset. Example from the corpus (right): Showing the vertical extent of an object

sive use of space establishes a persistent gestural discourse context (also called *virtual environment* by Kita [6]), to which subsequent utterances may relate.

3 A Model of Gesture and Speech Recognition

3.1 General Framework

The prevailing psycholinguistic theories on gesture and speech production assume at least two sources of multimodal utterances. Their origin is the speakers working memory which contains, according to the nomenclature by Levelt, propositional and spatial representations [10]. Hadar and Butterworth call the corresponding mechanisms conceptual processing and visual imagery [3]. Both mechanisms influence each other in that they mutually evoke and even modify representations. It is assumed that conceptual representation is the predominant driving force of speech while visual imagery promotes gesture production (even though there is disagreement about the details of the coordination). On the surface, the production process results in co-expressive gestural and spoken utterances. The production model by McNeill assumes basically the same two origins of an utterance which he calls analytic processes and imagery [13,12]. McNeill yet refuses an information processing approach in favor of a dialectic process in which gesture and speech, viewed as materialized thinking, develop from a common source *(growth point)*.

If we assume that gestures are communicative in the sense of conveying information to a recipient, then they should contribute to visual imagery and conceptual representations of the receiver. Current approaches towards speech and gesture integration just consider the conceptual representation which is built up from the fusion of oral and gestural information. However, a provision for the visuo-spatial properties of utterances, including effects like spatial cohesion, requires visual imagery. These considerations provide the basis for an abstract model of the recognition process which is shown in Fig. 3. The recipient perceives speech and gesture and analyzes them according to their structure. The

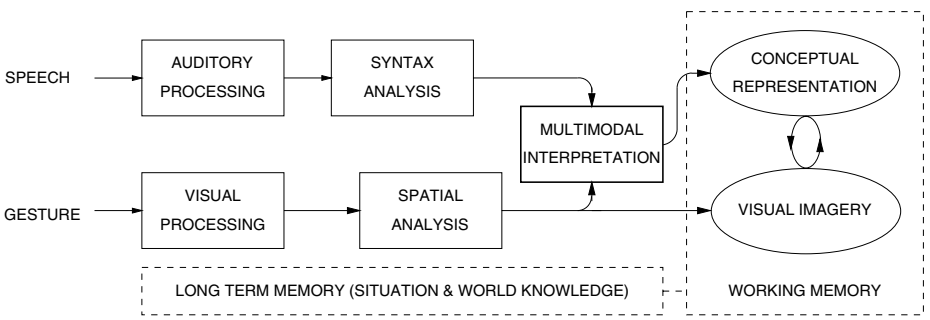

Fig. 3. An abstract processing model of gesture and speech recognition

propositional part of meaning is determined by integrating both modalities to a common conceptual representation. The multimodal interpretation step takes the categorical, language-like properties into account, in which emblems, deictics, iconic gestures, and – of course – speech, can be analyzed. At the same time a persistent image of the gestural part is built up. The property of persistence is essential for the creation of a visual discourse context that serves as a reference frame for subsequent utterances. Analogously to the production models, both representations do not exist independently from each other since conceptual processing may evoke imagery and vice versa. With this abstract recognition model in mind, we will now focus on the technical realization of the processing steps from spatial analysis to the construction of visuo-spatial imagery.

3.2 Computational Model

The application scenario of our approach towards a computational treatment of imagery is a reference identification task. The idea is to have a system that can identify a given object from among a set of objects based on a gestural description. A computational model of this process, which substantiates the spatial analysis, is shown in Fig. 4. Instead of visual information from gesture, the input consists of movement and posture data provided by data gloves and motion trackers, so that visual processing is not necessary. The spatial analysis is subdivided into three processing steps. At first, gestural features like handshape, movement, or holds are recognized. A segmentation module divides the meaningful gesture phases from subsidiary movements and meaningless postures. Following [2], we assume that it is generally possible to determine the meaningful part on the basis of movement and posture features without considering speech. The features of active phases are abstracted to basic spatial entities. Spatial entities are the building blocks of imagery, they are stored in the visual imagery component where spatial relations between different entities are determined. That way, a structured spatial representation of the gestural input is constructed.

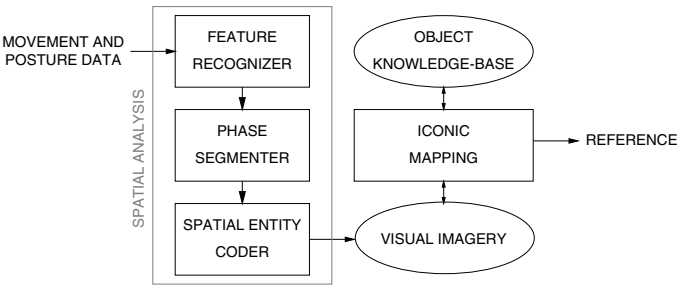

Fig. 4. A model of a recognition system for reference object recognition with shape-related gestures

A mapping mechanism compares internal object representations with the image and selects the object whose representation matches best.

4 Implementation Approach

Based on the computational model, we implemented a first prototype. It allows the user to select an object from a set of objects displayed on the screen by iconic gestures. The setting and instrumentation is comparable to the environment that we used for the empirical study. Fig. 5 shows a user and the wall display with some differently shaped objects. The recognition system is embedded in a virtual reality framework called AVANGO, a scene-graph based software environment that allows easy access to high-level VR programming [22].

Fig. 5. The prototype system: Selecting reference objects by gestural shape description

4.1 Feature Recognizer

Hand Features The recognition system provides a module that transforms (device-dependent) posture data from the CyberGloves into a general description of the hand-shape. The module is based on a kinematic model of the human hand which includes forward-kinematics to compute hand segment positions from given joint angles [8]. Put into the scene graph, the model can be visualized, thus providing a control mechanism for the correctness of the data mapping onto the model during development and testing (Fig.6). The input data for the hand module consists of a stream of 18-dimensional vectors that represent the measurements of the bimetal bending sensors in the CyberGloves. The vector components are numbers that correspond to the joint angles in a (more or less) linear fashion. However, since the CyberGloves measurement points do not perfectly match to the joints of the model and since the hand model is underspecified

Fig. 6. Visualization of hand model (left) and real hand with CyberGlove (right)

by the glove data, some transformations and dependencies are applied to compute the missing model data. The resulting model represents the hand-shape and hand-internal movements of the user quite accurately, provided that the hand does not have contact with real objects or with the body.[5]

The hand module provides a set of fuzzy variables that express the bending of the fingers in a symbolic way, derived from HamNoSys notation [18]. The module differentiates between five basic finger shapes: angular, rounded, bent, rolled, and stretched. The set of the five corresponding variables is computed for each finger (except the thumb, which is handled separately) and for the whole hand. By this, the module can provide information like "the left hand is rounded" or "the index of the right hand is stretched" to a subsequent module. Further fuzzy information concerns the position of the thumb which may be aligned or in opposition to the other fingers. This facilitates the detection of "precision grip"-like gestures used to indicate, for example, short distances.

Movement Features Data on the location and orientation of the wrists is provided by the tracking devices. This data can be combined with the output of the hand module to compute, for example, the absolute position of the fingertips or the center of the palm. From these reference points, different movement features are detected. Currently we evaluate holds and linear movement segments. Both features may indicate meaningful gesture phases for static or dynamic gestures [20]. The detection itself is based on the PrOSA-framework, a development from our group that facilitates rapid prototyping of evaluation systems for gestural movements in VR-scenarios [9].

4.2 Phase Segmenter

Our prototype system currently supports a simple gesture segmentation based on the typology of movement phases suggested by Kendon [5]. The typology defines

[5] In this case the heuristics may fail, but iconic gestures are normally non-contact movements.

gesture units as movements that occur between two resting positions. Any unit may consist of one or more *gesture phrases*, whereas each phrase contains a gesture stroke and an optional preparation and retraction phase. The resting space in our implementation is the area below chest height. Movement segments that begin or end in the resting space are counted as preparation or retraction phases and are not considered for further analysis. A retraction to resting space also signals the end of a gesture and initializes the iconic mapping process. There are three types of gesture strokes our prototype can recognize: holds in one hand, synchronized holds in both hands, and linear movements of one hand. Transitional movements between sequential strokes can be filtered out.

4.3 Spatial Entity Coder

According to our observations, we regard the most prominent meaning categories, spatial *extent* and *diameter* (a special case of extent), as basic spatial entities. The spatial entity coder abstracts from the concrete realization of the gesture through the transformation to a set of spatial entities. This processing step thus takes the variability of gestural expressions to convey a concept into account. Spatial entities are stored in the visual imagery component and for each new entry, spatial relations to existing entries are determined. The system checks for *orthogonality*, *dominance* (of one extent over another), and *equality* of two extents. Thus, the representation of the imaginary content may be visualized as a graph with spatial entities as nodes and spatial relations as links (Fig. 7).

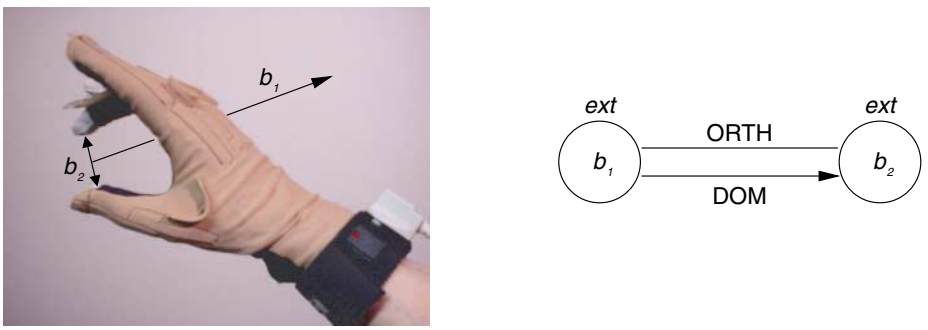

Fig. 7. Spatial representation of a dynamic gesture (hand moves right along b_1 direction): Spatial extents b_1 and b_2 are recognized and checked for spatial relations. b_1 is orthogonal and dominant to b_2

4.4 Iconic Mapping

The iconic mapping module determines the most appropriate object from the object knowledge base that matches the content of visual imagery. The output of

the iconic mapping may be used as a reference object in an application command. The mapping is based on a comparison of the gesture representation graph with the pre-defined object graphs by means of subgraph matching [14]. The mapping process evaluates for each object graph how much cost is expended to transform the gesture graph into a true subgraph of the model. The cost functions for insertion, deletion, and modification of nodes and links are defined individually. Our prototype computes the cost function according to some simple heuristics: Generally, it is assumed that the object model is complete, meaning that the gesture graph usually does not contain additional nodes and links. Therefore, a deletion of nodes and links is more expensive than an insertion. Node types can be `extent` and `diameter`, where extent is a superclass of diameter. The transformation from extent nodes to diameter nodes is therefore cheaper than vice versa. The transformation cost for each node and link total to a sum that represents the mismatch between gesture and object. Fig. 8 exemplifies the recognition process for a cubical object.

From the three-phase gesture unit indicating the cube's extent in all dimensions a gesture model graph with three nodes for the spatial entities is built. Entity relations `ORTH` for orthogonality and `EQ` for equality are determined. The graph is then matched against the object database. Note that a complete match is not necessarily needed to recognize the object. Even if only two of the three object extents are indicated by gesture, the cube interpretation would match the gesture model better than, for example, the bar.

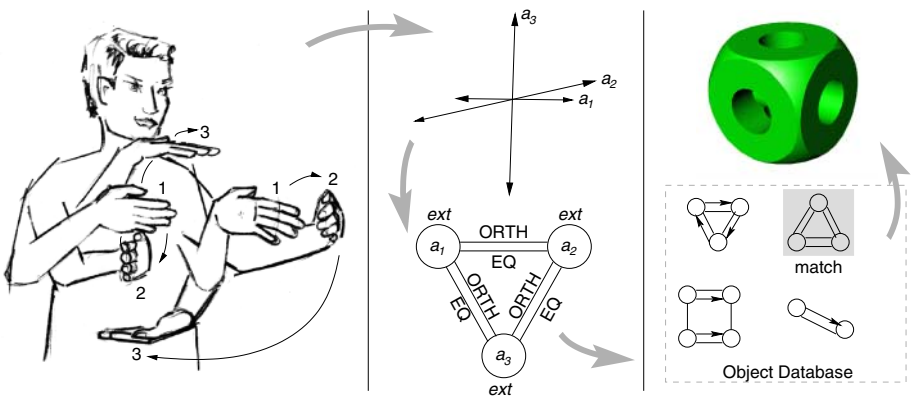

Fig. 8. The recognition process: A three-phase gesture unit (left) is mapped on a static model of imagery features (middle) and compared to object models (right)

5 Conclusion and Further Work

We presented an approach towards the interpretation of shape-describing gestures that is applicable in virtual design and virtual construction tasks. The model goes beyond a simple gesture-to-meaning mapping by way of decomposing the gestural utterance into spatial entities and their relations. That way, a *static* spatial representation of the *dynamic* gesture is built up. Our approach emphasizes the importance of imagery in addition to conceptual structures to construct a visuo-spatial discourse context. AI techniques from image understanding, i.e. graph-based representations and image recognition by means of subgraph isomorphism, are used for realization of a running system.

The inclusion of speech in our prototype system remains an open issue for further research. To achieve it, a cross-link between imagistic and conceptual memory is necessary. Another point for improvement is gesture segmentation. A detailed analysis of gestural features that indicate gesture strokes is still in progress. Likewise an evaluation of the overall system performance is subject to future work.

Acknowledgment

This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center "Situated Artificial Communicators" (SFB 360).

References

1. Christian Benoit, Jean-Claude Martin, Catherine Pelachaud, Lambert Schomaker, and Bernhard Suhm. Audio-visual and multimodal speech systems. In D. Gibbon, editor, *Handbook of Standards and Resources for Spoken Language Systems - Supplement Volume D.* to appear. 22
2. P. Feyereisen, M. Van de Wiele, and F. Dubois. The meaning of gestures: What can be understood without speech? *European Bulletin of Cognitive Psychology*, 8:3–25, 1988. 27
3. U. Hadar and B. Butterworth. Iconic gestures, imagery, and word retrieval in speech. *Semiotica*, 115(1/2):147–172, 1997. 26
4. Caroline Hummels. *Gestural design tools: prototypes, experiments and scenarios.* PhD thesis, Technische Universiteit Delft, 2000. 24
5. A. Kendon. Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key, editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. Mouton, The Hague, 1980. 25, 29
6. Sotaro Kita. How representational gestures help speaking. In McNeill [13], chapter 8, pages 162–185. 26
7. David B. Koons, Sparrell Carlton J., and Thorisson Kristinn R. *Intelligent Multimedia Interfaces*, chapter 11. MIT Press, Cambridge, Mass., USA, 1993. 23
8. Stefan Kopp and Ipke Wachsmuth. A knowledge-based approach for lifelike gesture animation. In W. Horn, editor, *ECAI 2000 - Proceedings of the 14th European Conference on Artificial Intelligence*, pages 663–667, Amsterdam, 2000. IOS Press. 28

9. Marc Erich Latoschik. A general framework for multimodal interaction in virtual reality systems: PrOSA. In *VR2001 workshop proceedings: The Future of VR and AR Interfaces: Multi-modal, Humanoid, Adaptive and Intelligent*, 2001. in press. 29
10. W. J. Levelt. *Speaking*. MIT press, Cambridge, Massachusetts, 1989. 26
11. D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992. 22, 23, 25
12. David McNeill. Catchments and contexts: Non-modular factors in speech and gesture production. In McNeill [13], chapter 15, pages 312–328. 26
13. David McNeill, editor. *Language and Gesture*. Language, Culture and Cognition. Cambridge University Press, Cambridge, 2000. 26, 32, 33
14. Bruno T. Messmer. *Efficient Graph Matching Algorithms for Preprocessed Model Graphs*. PhD thesis, University of Bern, Switzerland, 1996. 31
15. Jean-Luc Nespoulous and Andre Roch Lecours. Gestures: Nature and function. In Nespoulous, Perron, and Lecours, editors, *The Biological Foundations of Gestures: Motor and Semiotic Aspects*. Lawrence Erlbaum Associates, Hillsday N. J., 1986. 22
16. Vladimir I. Pavlovic, Rajeev Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):677–695, July 1997. 23
17. Charles Sanders Peirce. *Collected Papers of Charles Sanders Peirce*. The Belknap Press of Harvard University Press, Cambridge, 1965. 22
18. Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, and Jan 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, 1989. 29
19. Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E. McCullough, and Rashid Ansari. Gesture and speech multimodal conversational interaction in monocular video. Course Notes of the Interdisciplinary College "Cognitive and Neurosciences", Günne, Germany, March 2001. 23
20. Timo Sowa and Ipke Wachsmuth. 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. 24, 29
21. Carlton J. Sparrell and David B. Koons. Interpretation of coverbal depictive gestures. In *AAAI Spring Symposium Series: Intelligent Multi-Media Multi-Modal Systems*, pages 8–12. Stanford University, March 1994. 23
22. Henrik Tramberend. Avocado: A distributed virtual reality framework. In *Proceedings of the IEEE Virtual Reality*, pages 14–21, 1999. 28