The Spatial Specificity of Iconic Gestures

Stefan Kopp (skopp@techfak.uni-bielefeld.de)
Artificial Intelligence Group, Bielefeld University, P.O. 100131
D-33501 Bielefeld, GERMANY

Abstract

Humans use spontaneous gestures when communicating. But what these gestures convey is still an open question and several findings indicate that they fall short of communicating semantic information. This paper presents a study in which naïve observers had to draw images of what they saw in isolated iconic gestures. The detailed analyses of these drawings showed that observers were able to reliably extract visuospatial information from the gestures, with different hand shapes, movements, or hand orientations being differently salient and interpretable. In contrast to previous findings, these results suggest that iconic gestures can reach a level of specificity that makes them to an expedient means of conveying visuospatial information.

Introduction

When asked how to find a particular address, it is common to see a direction-giver depicting landmarks with the hands—the fork where one road joins another, the shapes of remarkable buildings, or their spatial relationship to one another. This is just one example for the use of gestures that humans frequently make to accompany their speech, especially when conveying information about objects and events in space. Just as people may draw maps or diagrams to illustrate a complex spatial layout, they may also use their hands to represent spatial content. Figure 1 shows such an example of spontaneous gesture in direction giving. While making this two-handed gesture, the speaker says "*there's a church*", and his hands impart to the utterance information about the shape of the church (left hand) and its location relative to a road (right arm, held from the previous phrase). This information was not conveyed by speech, neither in this phrase nor in previous or succeeding phrases. Yet, it seems instrumental to understanding the overall description.

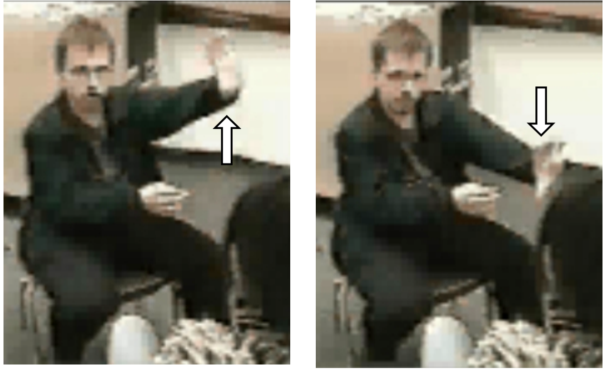


Figure 1: Coverbal gesture on "*There's a church*".

Examples of such a complementary use of speech and gestures can be found relatively easily. However, there is an ongoing discussion about whether such gestures are communicative, i.e. used by speakers to communicate information—intentionally or not—and attended to by listeners to understand this information (encoding and decoding hypotheses, cf. Bavelas et al., 2002). As discussed in the next section, considerable research has indicated that gestures fall short of communicating semantic information, i.e. that they bear a low semantic specificity. In the remainder of this paper we present a study testing the decoding hypothesis for shape-describing, iconic gestures. The results suggest that naïve observers are indeed able to extract information from isolated gestures. But, in contrast to semantic information that is apparently hard to find in gesture, this information concerns visuospatial aspects and properties of entities. We further report on an analysis showing that different features of gesture morphology, i.e., different hand shapes, movements, or hand orientations, differ in how clear and interpretably they communicate spatial information.

Semantic specificity of gesture

By gesture we generally mean all expressive movements of the hands and the arms made while speaking. Kendon (1983) suggested a continuum of coverbal hand movements, starting from so-called *adapters* that are perceived as communicatively meaningless and not related to speech. On the opposite end lie *symbolic* gestures, which have an autonomous, clear meaning and can be used like words in communication. We focus on the gestures in between these two extremes—coverbal gestures that do not have a clear cut meaning, nor are completely devoid of it (as the one shown in Fig. 1). Such gestures can vary largely in length, form, and complexity, and they appear related to the content of speech and temporally coordinated with it. Aside from *beat* gestures, small jerkily movements that do not bear any meaning, these gestures can be distinguished according to the semiotic relation that holds between their overt form and the entity they refer to. A variety of classification schemes have been proposed along those lines (e.g., in Ekman & Friesen, 1972; Feyereisen & deLannoy, 1991; McNeill, 1992). We focus here on *iconic* gestures that communicate in virtue of their resemblance with their referents, i.e. the postures and movements create an image that resembles the object, action, or event being described.

Currently, there is no consensus on whether iconic gestures are communicative, i.e. if speaker perform them to communicate information and if listeners draw upon them in interpreting the utterance. Two kinds of phenomena are generally interpreted in favor of the hypothesis that gestures do communicate. First, speaker gesture more frequently in face-to-face interactions with listeners (Bull, 1987; Kendon, 1983; Rime, 1982). Bavelas et al. (2002) found that speakers gesture at a higher rate when they know that they will be seen by the addressee, and they use gestures actively to compensate for problems of verbal encodibility. Second, there has been early evidence that the effectiveness of communication is enhanced by gesture (e.g., Rogers, 1978).

On the contrary, other findings seem to indicate that iconic gestures cannot communicate. Naïve observers, when viewing a gesture in isolation, are not able to reliably identify the actual lexical affiliate (Feyereisen et al., 1988; Krauss et al., 1991), even when the shaping of the gestures is related to the conceptual and semantic aspects of the affiliate (Hadar & Pinchas-Zamir, 2004). In addition, gestures are poorly recognized from visual data alone, particularly when they were originally perceived together with the accompanying speech (Krauss et al., 1991). This suggests that isolated gestures were not memorized in terms of meaning and that their meaning in verbal contexts is supposedly imputed primarily by speech. Indeed, categorization experiments showed that judgments of the semantic category of a gesture's meaning is largely a product of the verbal affiliate (Krauss et al., 1991). Finally, and contradictory to some of the aforementioned older studies, Krauss et al. (1995) reported that being able to see the speaker's gestures does not enhance the effectiveness of referential communication, when measured as the accuracy with which listeners are able to identify a described concept. All these findings seem to suggest that gestures carry relatively little semantic information. Krauss et al. (1996) even conclude that gesture does not contribute to the conveyance of the utterance's intended meaning but primarily serves intra-personal functions, e.g. to aid in the formulation of speech, and that all information it conveys is largely derivative from this. However, it is crucial to note that, by analyzing gestures for semantic information, one is putting them on par with language as a means of encoding meaning in terms of reference or predication. Results indicate that gestures fall short of communicating this kind of meaning.

Present study: Spatial specificity of iconics

The goal of our study was to investigate the communicative significance of iconic gestures. Although research to date strongly suggests that gestures do not convey any information that could be directly related to the semantics of speech, our hypothesis was that iconic gestures do have an inherent content, but that this content concerns a level of spatial meaning explicated in the imagistic content of the gesture. They communicate, then, in that this imagistic content describes meaningful geometric and spatial properties of entities. To test this hypothesis, we concentrated on the decoding problem, i.e., we addressed the question if naïve observers are able to reliably extract spatial information from isolated iconic gestures. To approach this question, we chose to let people *draw* all spatial entities or aspects they see in a gesture when not listening to the accompanying speech. By analyzing these drawings we could then test whether any spatial features are reliably perceived, and further which features of gesture morphology were used to depict them. The study was conducted on data gathered in another study on spontaneous gestures in direction giving (Kopp et al., 2004). It includes video- and audiotapes of 28 dyads (more than five hours of dialogue) in which one person explained, without any external aids such as maps, a route from point A to point B on a university campus to another person (see Fig. 2).

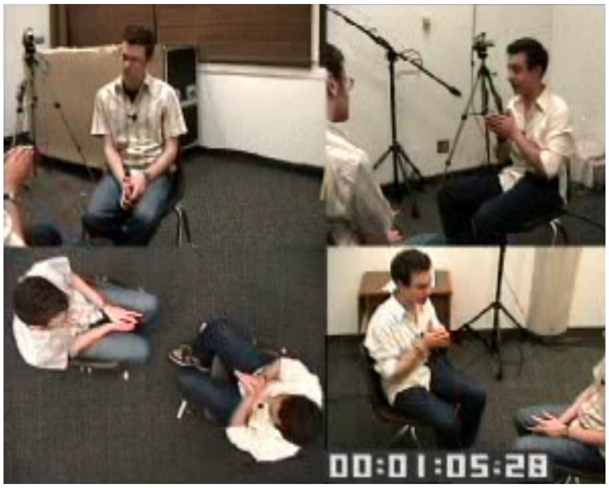


Figure 2: Example of the video data presented to subjects.

This direction giving task demanded the speaker to communicate complex spatial and visual information and all direction givers made frequent and spontaneous use of coverbal iconic gestures. Each videotape shows four synchronized camera views (Fig. 2), three of them recording the speaker from different perspectives (left-front, right-front, top). Using the TASX Annotator software (Milde & Gut, 2002), 10 dyads were segmented for the direction-giver's gestures and the expressive, meaning-bearing phases of each gesture was marked. To ensure rigor, all coding was carried out by a minimum of two coders, with any disagreements resolved by discussion. Annotation resulted in a total of 2424 gestures out of which 20 gestures that referred to concrete objects (landmarks or parts of landmarks, no actions) where selected for this study.

Method

Subjects. 11 graduate students or research assistants, who were not involved in any part of this study, volunteered as subjects.

Materials. Paper sheets were prepared for the subjects to draw on. To easy maintenance of perspective and to allow subjects to revert to easier two-dimensional drawing, every sheet contained four separate fields. Three fields corresponded to the perspectives presented in the video data, indicated by schematic pictures of the gesturer as seen from that perspective (cf. Fig. 4). The last field was empty and could be used freely. Additionally, each sheet contained checkboxes for indicating an absolute size (small, medium, or large).

Procedure. The subjects were provided with the paper sheets and seated in front of a computer monitor where isolated gestures were presented, i.e., short video clips of single gestures were shown without audio (see Fig. 2 for a screenshot of the presented video material). Subjects were instructed to carefully watch each gesture as often as they needed to find out the "image" the gesture seems to depict, and to redraw it in the fields of the sheet. Specifically, they were told not to wonder *what* is being depicted but only to concentrate on the presented spatial aspects, like shapes, lines, planes, bodies, locations, directions, extents, or configurations. They did so for all 20 gestures. To give an example, Fig. 4 shows three different drawings that we obtained for the gesture shown in Fig 3. While the first two participants saw a three-dimensional, box-like shape, the third subject perceived only two parallel upright planes.

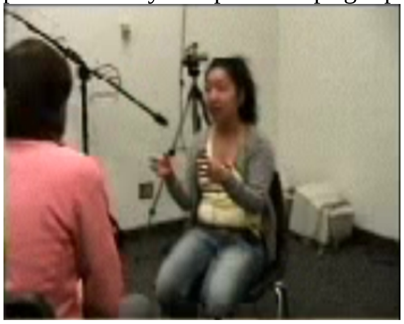


Figure 3: Stimulus gesture.

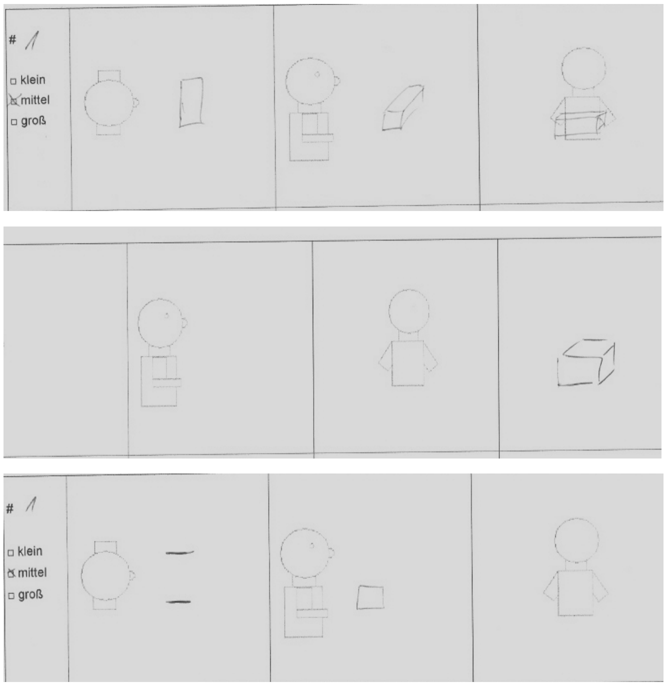


Figure 4: Drawings made by three subjects from the gesture shown in Fig. 3 (each row one subject).

Shape feature coding

The procedure yielded 220 sheets, six of which did not comply with the instructions (for example, subjects wrote down a possible referent like "*dough*" but did not draw anything). The remaining 214 drawings were analyzed for the spatial features of the shapes people have sketched. We adopted and slightly modified a representation used in our lab for iconic gesture analysis (Sowa & Wachsmuth, 2003). It is based on the component-based perceptual representations proposed by Marr & Nishihara (1978) and Biederman (1987), as well as on work on spatial language (Lang, 1989; Jackendoff, 1991). It describes a geometrical shape in terms of its differentiable extents, defined as axes of different size, type, and form. Depending on rotational symmetry, an axis may cover up to three spatial dimensions (*integration level*, from 1D-integrated to 2D- or 3D-intergrated as in a circle or a sphere). Additionally, each axis has a form and direction in space (linear or curved), has a qualitative measure of relative and absolute size, and is marked as either bounded or unbounded. We refer to these discernable shape properties as *gesture image features* (GIFs). Directions are defined relative to the base axes of the gesturer's frame of reference (left-right, back-front, up-down) or combinations of half-axes (e.g., "L+F" for left+front).

Table 1: Shape descriptions examples.

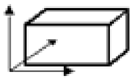
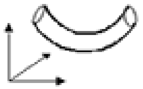
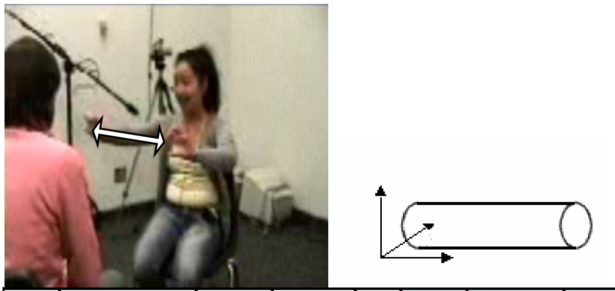
	Int. lvl	Form, direct.	Rel. size	Abs. size	Boun.
	1D	L-R	Max	Med	+
	1D	B-F	Min	Small	+
	1D	U-D	Min	Small	+
	1D	ARC R+D> R+U	Max	Med	+
	2D	B-F/R-L	Min	Small	+

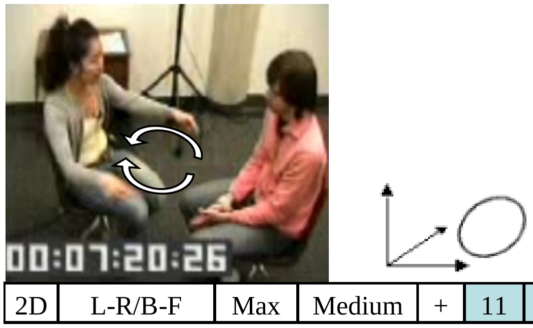
Table 1 shows descriptions of two example shapes. While one can find three different axes in the first shape, there are only two discernable GIFs in the second shape. Its first main axis is arced ("ARC") and starts out in the right+down direction ("R+D"), ending right+up ("R+U"). The second axis integrates two dimensions of space, notably, the plane spanned by the back-front and right-left axes ("B-F/R-L"). Both shapes have one major axis, whose size relative to the other axes is "Max" and whose absolute size is normal ("Med"). All axes are bounded, i.e. have clearly defined, finite extents.

Results

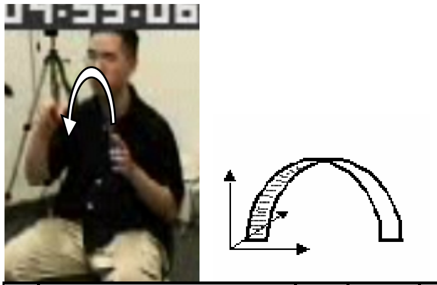
After annotating all 214 drawings, we could determine which GIF was perceived in a gesture and by how many observes. For every gesture, there were up to 33 axes (maximal 3 axes per shape times 11 subjects), with possibly different levels of integration, form, directions, sizes, or boundedness. Despite this high number of possible features, the observers did not perceive more than three to five different GIFs in a gesture (average 4.25, S.D.=1.71). Ordered by number of sightings, the most salient GIFs were seen on average by 9.25 subjects (S.D.=1.94), 6.3 subjects (S.D.=3.13), and 2.95 subjects (S.D.=1.99), respectively. For example, in the gesture shown in Fig. 3, 72.7% (Percentage Agreement=0.51) of the subjects saw a linear left-right extent, 54.4% (P.A.=0.27) spotted a linear back-front axis, and the same number of people an up-down axis. All agreed on the absolute and relative sizes of the axes, as well as that the shape is bounded in each of its axes.



1D	L-R	Max	Large	+	11	100%	1
2D	B-F/U-D	Min	Small	+	6	54.5%	0.27



2D	L-R/B-F	Max	Medium	+	11	100%	1



1D	ARC U>U+R>D+R>D	Max	Small	+	11	100%	1
1D	Back-Front	Min	Small	+	3	27.3%	0.05
1D	Left-Right	Min	Small	+	1	9.09%	0

Figure 5: Most frequent GIFs and the corresponding image prototypes for different gestures (see text for explanation of tables).

From these numbers we could determine an *image prototype* for each gesture. This prototype is the shape defined by the GIFs with the highest agreement values, which—in this combination—cover up as many spatial dimensions as possible. Figure 5 shows the image prototypes for three stimulus gestures that are relatively representative for the variety in gesture morphology captured by our study (one- vs. two-handed, different hand shapes, linear vs. curved trajectories, dynamic vs. static features). For each gesture, the derived image prototype is shown along with a table of the corresponding, most frequently recognized GIFs it consists of (one GIF per row, described as in Table 1). The shaded columns state how often each GIF was seen (from left to right, absolute numbers, percentages, and percentage agreement). Interestingly, there were great variations of agreement upon different GIFs. Each of three gestures of Fig. 5 depicted one spatial feature so saliently that it was reported by all subjects (P.A.=1). Across all gestures, 14 GIFs reached this perfect level of agreement, while four GIFs still achieved a level of P.A.=0.8 (reported by nine subjects). On the other hand, some GIFs were perceived with an extremely low reliability, like the width and thickness of the arc in Fig. 5 bottom. To find out if these differences were due to mere inter-subject variability or if they resulted from the very postures and movements used by the speakers to depict the GIFs, we conducted a second analysis of the correspondence between features of gesture morphology and features of spatial meaning, as far as captured by our representation.

Spatial specificity and gesture morphology

Morphology coding

In another pass of video data annotating, the morphology of each gesture was coded, using a scheme based on the McNeill Coding Manual (McNeill, 1992), refined for the purpose of our study. *Hand shape* was denoted in terms of ASL (American Sign Language) shape symbols, optionally modified with the terms “loose”, “bent”, “open”, or “spread”. *Hand orientation* was coded in terms of the direction the palm is facing, and the direction the fingers would point in if they were extended. Both were coded in terms of six speaker-centric, base- or half-axes. *Hand location* was described relative to a zoning of the space in front of the gesturer, augmented with a symbol to denote the distance between the hand and the body. Finally, movement in any of the three features was described using symbols to denote its shape (line, arc, circle, chop, or wiggle), its direction, and its extent. In addition, two-handed configurations (e.g., palms together) as well as movements of one hand relative to the other (e.g., hands move as mirror images) were explicitly denoted. Accuracy of this morphology coding, which was done by only one coder, was assessed by asking four members of our lab, who had never seen any of the gestures before, to reproduce 75 randomly chosen gestures from the dataset, solely on the basis of the codes. Each person was videotaped while reproducing the test gestures. These video recordings were then compared with the original subject data to rate similarity between the original gestures and the recreated ones (from 1=identical to 4=completely different). For each gesture, similarity was judged independently and separately for hand shape, orientation, and location. Then, the arithmetic means were calculated across all features and gestures. The resulting average value for all gesture ratings was 1.54 (SD=0.44), with static gestures being reproduced more accurately than gestures including movement. This indicated, first, that the codes captured almost all of the information needed to recreate a gesture and, second, that the number of errors in our form coding was well within acceptable limits.

Analysis and results

To study the correspondence between GIFs and features of gesture morphology (MF, henceforth), we counted for each GIF how often it was seen by subjects in a gesture with a certain MF (or combination of MFs), and how often it was not. Looking at the MFs, we could then test for a causation of GIFs by them, i.e. how probable it is that an observer will perceive a certain GIF when the gesture comprises that MF. Table 2 shows the results for the five GIFs we have tested: linear extents in the three main directions, arced 1D-integrated extent, and 2D-integrated extents. For each GIF, the MFs are shown along with the number of gestures they were comprised by (#G), the number of sightings of the GIF in these gestures (#seen) and the respective percentages (%).

Table 2: Causation of GIFs by MFs.

MF	**#G**	**#seen**	**%**	
GIF = 1D-integrated, linear, left-right				
Linear movement L-R	5	51/55	92,7	p<0.004
Palms facing, finger F, ASL loose_5, B	3	24/32	75	
other	12	34/127	26.8	
GIF = 1D-integrated, linear, up-down				
Linear movement U-D	4	33/44	75	p<0.005
Fing. U, ASL loose_5, B	1	11/11	100	
Fing. F, palm R/L, ASL loose_5, B, C	2	18/21	85.7	
Other	13	41/138	29.7	
GIF = 1D-integrated, linear, back-front				
Fing. F, ASL loose_5, B	6	45/64	70.3	p<0.001
Other	14	21/150	14	
GIF = 2D-integrated				
Arced movement	1	11/11	100	p<0.001
Fing. in GIF plane, ASL C	3	18/33	54.6	
Fing perp. to GIF plane, ASL bent_5, open_C	2	9/20	45	
Other	14	18/150	12	
GIF = 1D-integrated, arced				
Arced movement, large, in frontal plane	3	31/33	93.9	p<0.001
Other	17	12/181	6.63	

Statistical tests (ANOVA) indicated that the influence of morphological features on the spatial interpretation by observers is significant, and that this influence can even be traced down to the level of single MFs and single GIFs. Some MFs thereby reach surprisingly high levels of significance. For example, movement along a linear or arced trajectory seems to be a very reliable and salient way to depict a linear or curved extent or, in other words, to covey this particular piece of spatial information about, say, an object or event. Likewise, finger direction combined with a straightened hand shape (e.g., fingers pointing forward and spread as in ASL B) seems to be a good “depictor” of linear extent. This means also that, for some GIFs, there are several ways of conveying these particular features of spatial information.

Conclusion and general discussion

The primary question of interest here is what is the inherent content of an iconic gesture? Considerable previous research has indicated that conversational gestures do not convey semantic information and Krauss et al. (1991) have even stated that gestures are not an effective guide to the intended interpretation of the original utterance. In fact, we observed in our study that people strive to interpret a gesture in terms of conceptual reference rather than to conceptualize the abstract visuospatial properties, and we saw that a successful interpretation of an isolated gesture in this sense is virtually impossible. Nevertheless, the participants in our study were able to extract visuospatial features from the isolated gestures, yet with remarkable differences in agreement. The morphological features seem to differ with respect to how salient they depict certain spatial information and, consequently, how good they can be interpreted for this information. Still, some MFs did reach a perfect level of agreement and we can conclude that, if an iconic gesture in isolation may not be semantically specific, it can very well possess sufficient specificity to be communicative of spatial information as far as the decoding problem is regarded.

Generalizing this conclusion to the encoding hypothesis, one could assume that speakers also take advantage of this specificity by using gestures to depict visuospatial properties in support of conveying their communicative intent. The study described here was carried out in a larger research context of work on computer simulations of natural language and gesture generation with an embodied agent (Kopp et al., 2004). One hypothesis that underlies our generation of iconic gestures is that MFs can convey distinct visuospatial properties, and that these associations can be used to derive a gesture directly from the agent’s spatial knowledge about a referent. Our findings on decodability of iconic gestures encourage us in this and the spatial specificity we have found in them amounts to distinguishable visuospatial features of, e.g., objects, actions, or events. We believe that such information can be part of the speaker’s overall communicative intent, e.g. as distinguishing features in referring to an object, and examples like in Fig. 1 demonstrate that gestures can play even an exclusive role in encoding and conveying this information. The simulation also allows us to test another hypothesis, namely that MFs may be combined in a single gesture such that their spatial aspects are conflated into a more complex shape. This seems possible given that single MFs or combinations of a few MFs, which do not make for complete gesture morphologies yet, are already indicative of GIFs. Our results here do not disprove this hypothesis, but they do not provide positive evidence for it either as we have not analyzed for the contributions of single MFs to image prototypes. Another, more extensive study addressing these encoding problems is underway (Kopp et al., in press). Finally, we must point out that our study did not employ a very large number of subjects (11) and gestures (20), and that bigger data sets are needed to enhance statistical significance. Also, a more fine-grained study on gestures that differ only in single MFs would allow for a more systematic investigation of how single properties of gesture morphology influence the specificity of the gesture and thus the possible interpretation by an observer.

Acknowledgments

The author is grateful to Jens Heemeyer for his invaluable help with data annotation and analysis. Justine Cassell and Paul Tepper are collaborators in the overarching research effort and gave a lot of thoughtful comments.

References

Bavelas, J.B., Kenwood, C., Johnson, T., & Phillips, B. (2002). An experimental study of when and how speakers use gestures to communicate. *Gesture*, 2, 1-17.

Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94, 115-147.

Bull, P. (1987). *Gesture and Posture*. Oxford, Engl.: Pergamon Press.

Ekman, P. & Friesen, W.V. (1972). Hand movements. *Journal of Communication*, 22, 353-374.

Feyereisen, P. & deLannoy, J.-D. (1991). *Gesture and speech: Psychological investigations*. Cambridge: Cambridge University Press.

Feyereisen, P., Van de Wiele, M. & Dubois, F. (1988). The meaning of gestures: What can be understood without speech? *Cahiers de Psychologie Cognitive*, 8, 3-25.

Hadar, U. & Pinchas-Zamir, L. (2004). The semantic specificity of gesture. *Journal of Language and Social Psychology*, 23. 204-214.

Jackendoff, R. (1991). Parts and Boundaries, *Cognition*, 41, 9-45.

Kendon, A. (1983). Gesture and speech: How they interact. In J.M. Weinmann & R.P. Harrison (Eds.), *Nonverbal interaction*. Beverly Hills, CA: Sage.

Kopp, S., Tepper, P. & Cassell, J. (2004). Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output. *Proceedings of ICMI'04*, 97-104.

Kopp, S., Tepper, P., Ferriman, K. & Cassell, J. (in press). Trading Spaces: How Humans and Humanoids use Speech and Gesture to Give Directions. *Spatial Cognition and Computation*.

Krauss, R.M., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In M. Zanna (Ed.), *Advances in experimental social psychology*, 389-450, San Diego, CA: Academic Press.

Krauss, R.M., Dushay, R.A., Chen, Y., & Rauscher, F. (1995). The communicative value of conversational hand gestures. *Journal of Experimental Social Psychology* , 31, 533-552.

Krauss, R.M., Morrel-Samuels, P. & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, 61, 743-754.

Lang, E. (1989). The semantics of dimensional designation of spatial objects. In M. Bierwisch & E. Lang (Eds.), *Dimensional adjectives*. Berlin: Springer.

Marr, D. & Nishihara, H.K. (1978). Representation and recognition of the spatial organization of three dimensional structure. In *Proceedings of the Royal Society of London*, 200, 269-294.

McNeill, D. (1992). *Hand and mind*. Chicago: The University of Chicago Press.

Milde, J.-T. & Gut, U. (2002). The TASX-environment: an XML-based toolset for time aligned speech corpora. *Int. Conf. On Language Resources & Evaluation*, Las Palmas.

Rime, B. (1982). The elimination of visible behaviour from social interactions: Effects on verbal, nonverbal and interpersonal behavior. *European Journal of Social Psychology*, 12, 113-129.

Rogers, W.T. (1978). The contribution of kinesic illustrators toward the comprehension of verbal behaviors within utterances. *Human Communication Research*, 5, 54-62.

Sowa, T. & Wachsmuth, I. (2003). Coverbal Iconic Gestures for Object Descriptions in Virtual Environments: An Empirical Study. In M. Rector, I. Poggi & N. Trigo (Eds.). Proceedings of "*Gestures. Meaning and Use*", 365-376.