A Computational Model for the Representation and Processing of Shape in Coverbal Iconic Gestures[1]

Timo Sowa and Ipke Wachsmuth

Abstract

When humans describe the shape of objects, they often use iconic gestures to depict what they want to convey to a listener. Gesturing gives them the ability to express spatial concepts directly in the spatial medium and thus provides an important communicative resource for spatial language. In order to harvest this resource in language comprehension systems, the composite signal conveyed in two different media has to be re-integrated to a common, unified meaning. In a corpus study, we examined the morphological variety of shape-related iconic gestures and the kind of shape information they express. We distinguish four sub-types of iconic gestures and show that the most frequent type, called dimensional gestures, and the lexical affiliates they co-occur with, contain information about an object's spatial extent, the course of its boundary, and the spatial relations between object parts. An analysis of the verbal utterances shows that adjectives and nouns are predominant among the lexical affiliates in our scenario. Based on the empirical results, a computational model for the representation and processing of multimodal shape descriptions is proposed.

[1] This chapter is an elaborated version of a paper first published in K. Opwis and I.-K. Penner (eds.) (2005), Proceedings of KogWis05, Basel: Schwabe Verlag. It has appeared in K. R. Coventry, T. Tenbrink, and J. A. Bateman (eds.), Spatial Language and Dialogue (Chapter 10). Oxford: Oxford University Press, 2009.

1 Introduction

When describing object shape, humans usually perform iconic gestures that coincide with speech (are coverbal). Iconic gestures are semantically related to the content of speech (co-expressive), but express meaning in a different way (McNeill, 1992, 2005). They unfold in time and space and present images, marked by a similarity between the gestural sign and the described object. In contrast, speech unfolds only in time and presents a stream of symbolic signs arbitrarily connected to meaning. Still, both modalities, the verbal and the gestural, are assumed to form an integrated system of communication (Bavelas & Chovil, 2000; Clark, 1996) with a common origin or 'idea unit' underlying the production of co-expressive speech and gesture fragments (McNeill, 1992, 2005; Kendon, 2004). Due to their inherently space-bound nature, iconic gestures may easily depict content difficult to describe using words alone. This iconic content is picked up and processed by listeners as recent studies on gestural mimicry and neuropsychological findings suggest (Kimbara, 2006; Kelly, Kravitz, & Hopkins, 2004). Though the expressive potential of iconic gestures in human-human communication is generally acknowledged, few systems make use of it in human-computer interaction. Instead, the development of comprehension systems that take the non-verbal component of natural communication into account focused much more on pointing and symbolic gestures.

Our main goal is to build computational models for the representation and processing of spatial language including shape-related gestures. We employ empirical studies on communicative behavior as the main information source for our modelling efforts. In that respect our contribution is in line with the empirically-based work of Shi & Tenbrink (this volume) on dialogue design for an instructable wheelchair and Striegnitz et al. (this volume) who focus on spatial knowledge representation for the generation of verbal and gestural route descriptions by an artificial agent. In this paper, we concentrate on the way people use iconic gestures in descriptions of object shape and we present an approach to employ such gestures in comprehension systems for spatial language. With show how to build up semantic representations for multimodal referential expressions like the noun phrase *a longish bar* + <*iconic gesture*> in which the adjective, the noun, and the gesture together form a composite signal (in the sense of Clark, 1996) specifying an object. Application areas include natural language interfaces for autonomous systems (e.g. mobile robots), virtual construction applications, and virtual design.

The paper is organized in three parts. First, we describe an empirical study on the use of iconic gestures and speech in shape descriptions. Second, we describe a formal representation for multimodal shape descriptions that captures the content of shape-related iconic gestures as well as shape-related adjectives and nouns. Finally, we sketch the application of the representation in a multimodal system for gesture and speech comprehension.

2 Gesture and speech in shape descriptions

In order to examine the morphology and the semantic aspects of shape-related coverbal gestures, an observational study was conducted which is described in more detail in (Sowa & Wachsmuth, 2003). A total of 37 participants were asked to describe five different stimulus objects (Figure 1). The objects were projected with a video beamer on a wall-size screen. It was decided to use magnified projections of the original parts since large stimuli were assumed to evoke larger and clearer gestures. The height of the green cube in Fig. 1, for instance, was about 80 centimeters on the projection screen. The original parts were not shown to the subjects.

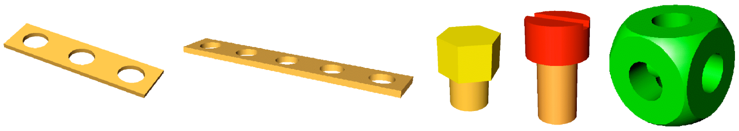


Figure 1: Stimulus objects used in the study.

The participants were told that their descriptions are videotaped and shown to another person afterwards. They were instructed to give their description in such a way that the person watching the video recording gets an idea how the objects looks like. It was mentioned that the hands can be used for the descriptions, but hand gestures were not enforced nor were the descriptions in any other way restricted. The stimulus objects are computergraphically generated parts of a toy construction kit. Two pairs of objects, the stylized screws and the bars, are quite similar. Though their basic shape is nearly identical, they differ in their sizes and proportions. That way, effects of size and proportions on iconic gestures could be examined in isolation. All gestures judged to express shape-related content were transcribed with respect to spatiotemporal features, i.e. its form, and the corresponding elements of meaning. The annotated corpus comprises 383 gestures. The analysis of verbal information in the corpus relies on the concept of *lexical affiliates* which could be single words, multiple words, or phrases to which gestures semantically relate. For each gesture transcribed, its lexical affiliate was determined independently by three coders. Only those words or parts of speech rated as lexical affiliates by at least two coders were included in the analysis.

2.1 Gesture types

In order to systematize the corpus, gestures with a similar relation between form and meaning were grouped together yielding 84 different gesture kinds. The form-meaning relation was considered similar if identical spatiotemporal features had been used to express the same semantic properties. For instance, one gesture kind is marked by the distance between the tip of the thumb and another fingertip expressing object extent. Several variants of this form-meaning relation were observed. The finger opposing the thumb's tip could be the index, the middle, or even the ring finger, and the extent could be displayed horizontally or vertically. However, all of these concrete instances share the general feature that two points in space are defined by opposing fingertips, expressing the property of extent. Each gesture kind can be represented by a prototype which is an idealized realization of the form-meaning relation (see Table 1 for the most frequent gesture kinds). Four general gesture types can be distinguished as given below.

2.1.1 Dimensional gestures

The largest group is characterized by representing an object's outer dimensions via delimiting or enclosing. Such gestures may indicate spatial extent and/or the profile of intrinsic object axes. Extent refers to the stretch of space an object occupies and is often expressed by using parts of the hands or arms to indicate endpoints; cf. Table 1 (a). The term profile refers to the course of the object's boundary and usually involves some kind of motion (b-h). Dimensional gestures often depict abstract one- or two-dimensional characterizations of the three-dimensional object (*dimensional underspecification*). Gestures (a)-(c) in Table 1 are one-dimensional, i.e. depict an extent along one "line". Gesture (a) expresses extent as the space between the hands, while (b) and (c) additionally indicate the profile of this one-dimensional extent via movement.

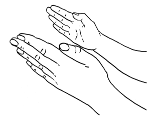
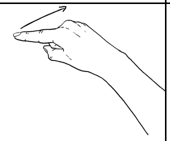
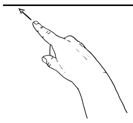
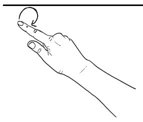
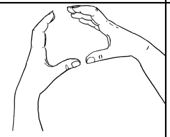
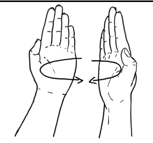
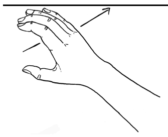
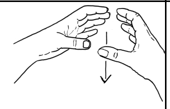
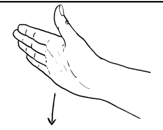
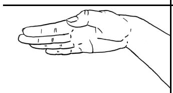
	(a) flat hands, palms facing each other; indicates extent between left and right hand
	(b) extended index finger; fingertip moving straight; orientation perpendicular to movement; indicates extent
	(c) extended index finger; hand moving along index direction which indicates the extent; used mainly to depict an *interior* path, i.e. holes
	(d) extended index finger; fingertip describes a circular trajectory; fingertip movement indicates extent and profile
	(e) rounded C-handshapes; circle open or closed; posture indicates extent and round profile
	(f) flat hands, fingers aligned; hands perform semi-circular mirrored movements, palms facing towards the center of the circle; indicates extent and round profile
	(g) hand is moving straight, perpendicular to the aperture; hand-shape indicates extent and round profile in two dimensions, movement adds another dimension
	(h) hands form an open or closed circle; hands moving downward; hand-shape indicates extent and round profile in two dimensions, movement adds another dimension
	(i) flat hand, fingers aligned; hand moves into a direction parallel to the plane of the palm; movement and hand surface indicate a face of the object
	(j) flat hand as a placeholder; indicates orientation of an object in space

Table 1: A subset of the most frequent gesture kinds represented by prototypes

Gestures (d)-(f) are two-dimensional. All of them indicate the round profile of the reference object and the extent (i.e. the diameter) either by hand-shape or by movement. Gestures (g) and (h) are three-dimensional. In both cases a two-dimensional profile created via a distinct hand-shape is extruded by a linear motion resulting in the depiction of a cylindrical shape. A detailed analysis of the usage of gesture kinds for the depiction of certain objects or parts shows that an object’s relative sizes (i.e. length : width : height ratio) partly determine the kind of gesture employed for a depiction.[2] A linear movement as in (b), (g), and (h) usually indicates a dominant extent (e.g. the length axis of the bar), whereas it is not used for an object or part without dominant axis such as the cube. In contrast, to-handed, delimiting gestures as in (a) are likewise used for dominant extents, but also for the equally-sized extents of the compact cube. Generally, hand movement and two-handed delimitation are employed for the dominant extents, while hand-shapes are typically used to display subordinate extents. Besides relative size, we also compared absolute object and part sizes (as they appear on the screen) to the size of gestures gesture (a) if this gesture kind was employed for a depiction. We found a great variance of gesturally indicated sizes for an object with constant size. From this we conclude that it is rather relative than absolute size which is reflected in a person’s dimensional gesture.

2.1.2 Surface property gestures

While dimensional gestures refer to the whole shape in terms of extent and profile, surface property gestures depict certain elements or features of an object’s surface without reference to the whole object. Prototype (i) in Table 1 is an example of this type: The flat, moving hand indicates a particular planar side of the object without referring to the whole.

2.1.3 Placeholder gestures

These gestures are characterized by a body part representing the object itself. Spatial position and/or orientation properties are directly conveyed by the appropriate configuration of the body part in space. The realizations thus consist only of one-handed gestures with a distinct hand- or arm-configuration taking the approximate shape of the object. Prototype (j) is an example for a placeholder gesture. The whole hand stands for a longish, flat object and indicates its configuration in space.

[2] Cf. (Sowa, 2006a) for a discussion of the effects of object sizes and proportions on shape-related iconic gestures.

2.1.4 Spatial relation gestures

This last gesture type indicates the relative position and/or orientation of two object parts using one hand for each. Thus, spatial relation gestures are always two-handed and usually asymmetrical. They may also consist of a combination of two individual gestures from the aforementioned types.

Dimensional gestures account for 86% of all gestures, shape property gestures for 6%, placeholder and spatial relation gestures each for 2%. Given the dominance of dimensional gestures in the corpus, it seems appropriate to consider the semantic features they express, namely extent and profile, as basic features for a representation of gesture content. Dimensional underspecification further implicates to consider extents and profiles independently for each spatial dimension. A semantic representation should reflect this underspecification, i.e. it should be possible to specify just one dimension or object axis and to make no assumptions about the remaining dimensions.

2.2 Object decomposition

Some of the stimulus objects are easily decomposable into parts, for instance the screws can be composed into shank, head, and slot. Subjects usually realized this canonical object structure in their descriptions. Two object classes that apparently affect the way subjects describe the whole object can be distinguished. When the object's main body was a basic 3D geometry, like the bars and the cube, it was depicted in a gesture. For compositional objects like screws that consist of two almost equally sized parts, fewer gestures for the whole body were employed. In no case would a gesture depict the complex object shape at once, for instance, drawing an outline of the screw as T-shaped object. However, several subjects did depict the whole screw in an abstract way reducing it to its main extent.

2.3 The spatial organization of successive gestures

Gestural expressions have the potential to organize in space and to build larger structures of meaning (Emmorey, Tversky & Taylor, 2000; Enfield, 2004). They are spatially cohesive in the sense that successive gestures often employ space in a consistent way (McNeill, 1992). (For an analysis of the consistency with respect to the reference frame in verbal expressions see Vorwerg, this volume). Examples of spatial organization can be found in the corpus data. Consider the gestures accompanying the description of the short bar (Figure 2). The subject first anchors the bar in space using a two-handed symmetrical gesture indicating its longitudinal extent. The left (non-dominant) hand is held in this position, while the right (dominant) hand indicates the position and shape of the holes with three successive strokes (meaningful phases). With the initial two-handed gesture, an imagistic context introducing the main object is set up in space. The validity of the context is explicitly bound to a visible feature, namely the left hand which keeps the position and shape of the initial gesture. This kind of organization we call *explicit spatial cohesion*.

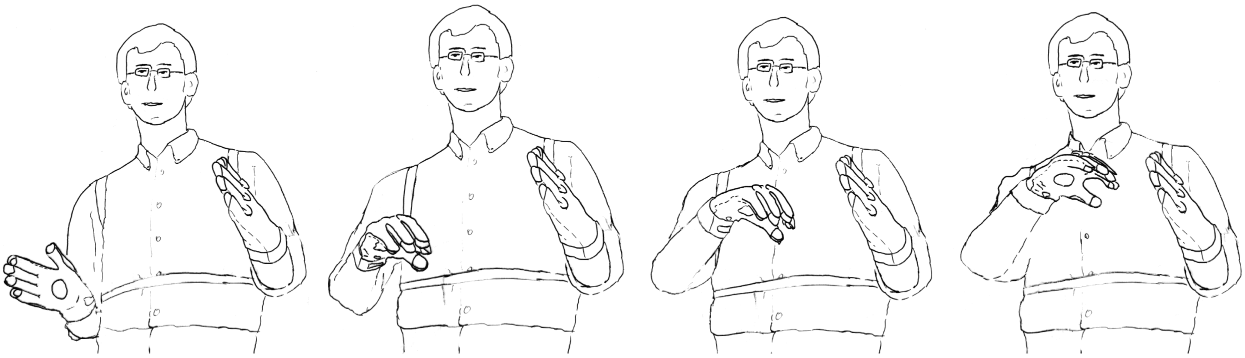


Figure 2: Explicit spatial cohesion via a two-handed gesture. Left hand is held in position.

Conversely, there is *implicit spatial cohesion* whenever the spatial relation of successive gestures reflects the relation of the reference objects, but without any visible feature indicating cohesion. Figure 3 illustrates examples in which the spatial arrangement of successive gestures coincides with the spatial relation of the objects they refer to. Spatial cohesion can bind together several semantic entities (extents, profiles) either of a single object or part, or of two or more different objects or object parts. In Fig. 3b we can see an example for the former case, called *intra-object cohesion*. The dominant dimensions of the bar (its length and width) are displayed successively with two-handed gestures (indicating parallel lines) providing a two-dimensional specification of a single object (the bar). An example for the latter case, *inter-object cohesion*, is depicted in Fig. 3c. Three cohesive gestures successively indicate different parts of the screw: the shank (lower vertical line), the head (upper vertical line), and the slot (horizontal arrow).

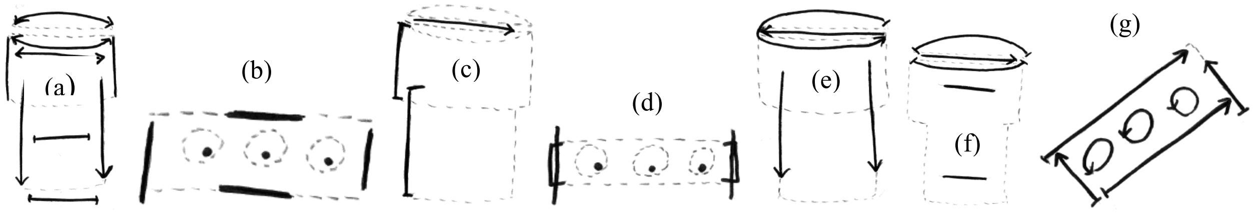


Figure 3: Implicit spatial cohesion. Solid lines indicate gesture locations (arrows stand for movement in dynamic gestures), dotted lines show the reference object.

2.4 Parts of speech associated with iconic gestures

Table 2 shows the frequency distributions of the parts of speech among the lexical affiliates (1st column, for n = 478 affiliates), their base frequencies in a representative sample of the whole corpus (2nd column), and the relative frequency of parts of speech among affiliates with respect to their base frequency (1st column divided by 2nd column). It is evident that nouns and in particular adjectives are overrepresented among the affiliates (relative frequency > 1.0), while the other classes are underrepresented (relative frequency < 1.0).

Table 2: Frequency of the word classes among the affiliates and relative to the whole corpus.

	Affiliates (%)	Corpus sample (%)	Relative (affiliates / corpus sample)
Nouns	42.9	27.2	1.58
Adjectives	29.5	6.2	4.79
Verbs	4.0	15.4	0.26
Prepositions	5.2	8.3	0.63
Adverbs	14.2	21.8	0.65
Determiners	4.2	15.6	0.27
Interjections	0.0	5.5	0.00

A semantic analysis of the affiliated nouns shows that they include references to 3-D shape such as *cylinder*, 2 or 1-D part references such as *side*, *face*, or *corner*, usually expressed after the introduction of the whole object in the discourse context, and references to object dimensions such as *length* or *diameter*. Affiliated adjectives similarly include 3-D descriptors such as *cylindrical*, 2-D expressions such as *round* or *six-sided*, and dimensional adjectives like *long* or *flat.* Furthermore, there are adjectives such as *flattened* or *dagged* describing shape properties (modifications) of base objects, and other adjectives not directly related to shape, but to object orientation and position. Most of these verbal affiliates express aspects of object extent, as in the case of dimensional adjectives, or aspects of extent combined with profile (boundary) properties as in 3-D nouns and adjectives. This shows that affiliates could refer to all spatial dimensions, or specify just two dimensions or one dimension of the object.

3 A unified shape representation for multimodal signals

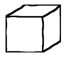
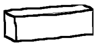
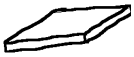
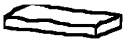
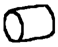
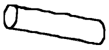
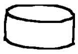
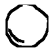
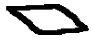
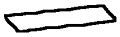
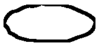
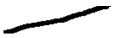
Taken together, the corpus evaluation revealed three important factors to consider in a semantic representation of shape-related gestural and verbal expressions. First, extent and profile are directly expressed in (dimensional) gestures as well as in accompanying adjectives and nouns and could be considered two basic semantic factors. Second, these elements are not expressed in isolation, but structurally organized in a spatially cohesive context. A semantic representation should thus reflect, third, the spatial arrangement of successive gestures. In the following, a shape-representation model that covers these factors is described. It extends an earlier approach which models the two factors of extent and (partly) profile information in gestures, but which has not included structured spatial organization of gesture and accompanying speech reflecting this factor (Sowa & Wachsmuth, 2002). A more detailed description of the formal structure can be found in (Sowa, 2006a, 2006b).

Models for shape representation that may inform the multimodal modelling approach can be found in different research disciplines including visual cognition, spatial reasoning, and linguistic modelling. Shape representations are usually divided into boundary- and interior-based approaches. The former primarily describe 2-D surfaces while the latter represent 3-D volumes. Interior-based approaches appear more relevant for the task, because object shape is primarily a 3-D property. A further distinction can be drawn between quantitative and qualitative representations. Purely quantitative approaches are usually employed in computer graphics and geometric modelling where precision is needed (Mortenson, 1997). Yet, as the study suggests, spatial information in gesture (and speech) is often abstract, qualitative, and underspecified. Precise approaches lack the ability of abstraction such that their applicability is limited. Cohn and Hazarika (2001) provide a summary of representations for qualitative spatial reasoning. One qualitative, interior-based method is to use volume primitives for shape approximation. This approach is exemplified by the geon model suggested by Biederman (1987) and the 3-D model by Marr and Nishihara (1978) which also introduces different levels of shape abstraction. However, geons and other volume primitives do not allow dimensional underspecification because they are inherently defined in 3-D. A one-dimensional gesture specifying a single extent could not be adequately represented. A suitable approach for the definition of the principal extent(s) of objects is provided by Lang (1989) within a semantic theory for dimensional adjectives. Lang defines representations called object schemata describing the basic gestalt properties of objects. Similarly, Clementini and Di Felice (1997) suggest properties for basic gestalt descriptions. None of these models fulfills all requirements that arise from the corpus analysis. Therefore, a new representation, called Imagistic Description Tree (IDT), is proposed in the sections to follow, which unifies the benefits of the model types above.

3.1 Modelling extent properties

For the modelling of extent properties we adopt the idea of an object schema as proposed by Lang (1989). Each object is described by a collection of up to three axes which represent the object's extents. An axis may cover one, two, or three spatial dimensions. A schema for a cylinder, for instance, would contain two axes. The first axis describes its height and is associated with one dimension. The second axis is associated with the remaining two (indistinguishable - due to rotational symmetry) dimensions. The "object schema" representation is appropriate as a model for object descriptions with dimensional gestures and adjectives/nouns, because it singles out the axes and their relations and thus allows dimensional underspecification.

Table 3: Representation of basic object types with object schemata

Object schema	Prototype
$\{(1,\{\emptyset\},\bot), (1,\{\emptyset\},\bot), (1,\{\emptyset\},\bot)\}$	
$\{(1,\{\mathit{max}\},\bot), (1,\{\emptyset\},\bot), (1,\{\emptyset\},\bot)\}$	
$\{(1,\{\emptyset\},\bot), (1,\{\emptyset\},\bot), (1,\{\mathit{sub}\},\bot)\}$	
$\{(1,\{\mathit{max}\},\bot), (1,\{\emptyset\},\bot), (1,\{\mathit{sub}\},\bot)\}$	
$\{(1,\{\emptyset\},\bot), (2,\{\emptyset\},\bot)\}$	
$\{(1,\{\mathit{max}\},\bot), (2,\{\mathit{sub}\},\bot)\}$	
$\{(2,\{\emptyset\},\bot), (1,\{\mathit{sub}\},\bot)\}$	
$\{(3,\{\emptyset\},\bot)\}$	
$\{(1,\{\emptyset\},\bot), (1,\{ \emptyset\},\bot)\}$	
$\{(1,\{\mathit{max}\},\bot), (1,\{\emptyset\},\bot)\}$	
$\{(2,\{\emptyset\},\bot)\}$	
$\{(1,\{\mathit{max}\},\bot)\}$	

Using different combinations of axes in an object schema, several basic object types with certain spatial characteristics can be represented as illustrated in Table 3. The first schema describes an object with three discernable axes of almost equal size. The typical instance, or prototype, for an object with these characteristics would be a cube. The first eight schemata in the table specify shape in three dimensions (illustrated with 3D-prototypes), while the last four show cases of dimensional underspecification. In Table 3 and in the following we use this formal notation: Curly brackets {…} delimit an *object schema*. The bracketed elements describe the *object axes*. An object axis has three properties, (1) a number (1, 2, or 3) describing how many canonical dimensions it is associated with, (2) a set of qualitative properties called *dimensional assignment values (DAVs)*, and (3) a measure for the axis' numerical extent (e.g., in centimetres). We use triplets in round brackets (…) to indicate these properties. If a particular axis within a schema is labeled with the DAV *max*, it is the one with the largest numerical extent which corresponds to the length of the object. The DAV *sub* stands for substance and expresses minimality of the extent as compared to the other axes. It corresponds to object thickness. The unspecified DAV ∅ stands for an axis which is not significantly different in extent from the other axes in a schema. We use the symbol ⊥ to indicate that a numerical extent of an axis is not specified.

Here are two examples of how the spatial properties inherent in dimensional adjectives, nouns, and gestures can be represented using this model. Consider the adjective *longish*: its conceptualization in terms of an object schema would be {(1, {max}, ⊥)}. This means that a longish object is characterized by an object schema containing at least one axis which covers a single dimension and which is quantitatively most extended. Similarly, dimensional gestures can be semantically encoded using object schemata. Consider gesture prototype (h) in Table 1. The hands symmetrically form a round shape which is combined with a downward motion. Assume further that the extent of the motion is 40 cm, and the extent (diameter) of the circle formed by both hands is 20 cm. Both the movement component and the hand-shape are assumed to indicate spatial extent on the highest level of abstraction. The corresponding semantic encoding would thus be a schema containing two axes, i.e., a one-dimensional axis representing the movement extent, and a two-dimensional axis representing the extent of the hand-shape, i.e. {(1, {max}, 40.0), (2, {sub}, 20.0)}.

3.2 Modelling profile properties

While extent properties refer to the basic proportions of an object, profile features provide additional information on the object's boundary. We adopt three general properties (symmetry, size, and edge) from the geon model here, with some modifications. The *symmetry property* expresses regularities of the boundary with respect to one axis or a symmetric relation between two axes. The *size property* reflects the change of an axis' extent when moving along another axis. The *edge property* determines whether an object's boundary consists of straight segments that form sharp corners, or of curvy, smooth edges. Profile properties are defined by a profile vector containing symmetry, size, and edge properties for each object axis or pair of axes. Considering gesture prototype (h) in Table 1 again, it is possible to infer profile properties of the object expressed in the gesture in addition to the basic extent information encoded in {(1, {max}, 40.0), (2, {sub}, 20.0)}. The fact that the hand-shape does not change during the downward movement indicates a constant extent (diameter) along the movement axis.

This can be captured with a profile vector containing a size entry for “constancy” of the second axis when moving along the first. A combination of two static gestures, e.g. (e) + (a) in Table 1, would not provide such profile information. Another example for the use of profile properties is given in the following section.

3.3 Modelling structure by an IDT

Object schemata are the building blocks of the IDT. They provide a description of an object’s overall proportions and its major profile properties, but do not model structure and spatial relations. To this end, schemata can be arranged in a tree similar to the hierarchical structure used in the Marr and Nishihara (1978) model.

Structural aspects are represented in *imagistic descriptions*. An imagistic description for an object consists of a set of imagistic descriptions describing its parts, an object schema defining its overall proportions, a spatial anchor flag and a transformation matrix. This recursive definition provides a tree-like structure: The parts described are imagistic descriptions which could themselves contain further parts. The number of children is arbitrary. The spatial anchor flag signals whether the description is spatially anchored in a parent coordinate system. If its value is *yes*, the transformation matrix defines the position, orientation, and size of the object or part in relation to the parent description. An imagistic description of a perceived gesture, for instance, is spatially anchored because the gesture is performed in space and can be assigned spatial coordinates, while an imagistic description of an adjective is not spatially anchored. The complete tree describing an object including all parts, parts of parts etc. is called an *Imagistic Description Tree (IDT)*.

Figure 4 shows an example of an IDT model for the screw. The part hierarchy modelled by the three layers of the tree follows its perceptually salient decomposition. The top-level node I_{sc} represents the whole screw and has two child nodes modelling the parts, I_{he} for the head and I_{sh} for the shank. The head has another child node I_{sl} representing the slot.

Without providing all formal details of the IDT definition (Sowa, 2006a, 2006b), a closer look at node I_{he} representing the head will suffice to illustrate the model. The imagistic description I_{he} defines the slot representation I_{sl} as the only part. OS_{he} is the object schema that defines the basic proportions (axes) of the head. It contains two axes:

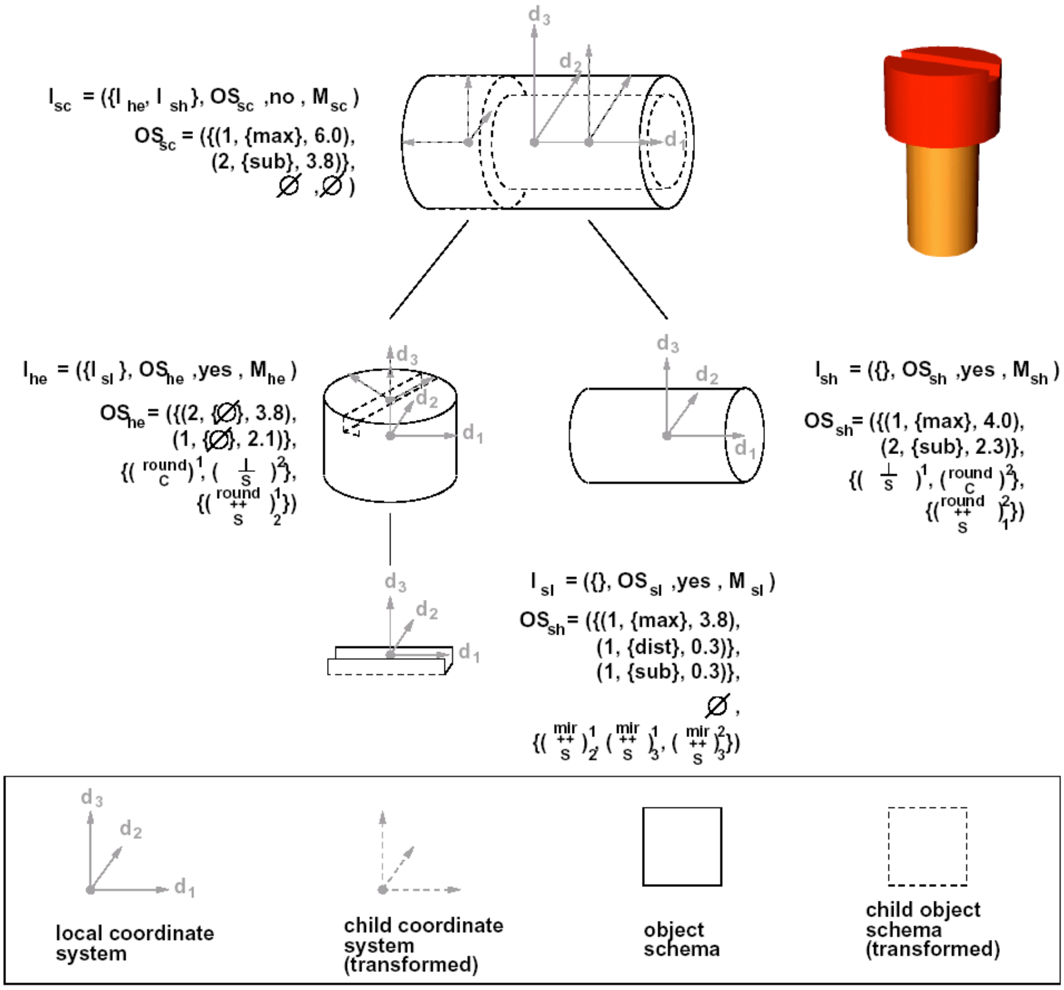

I_sc = ({I_he, I_sh}, OS_sc ,no , M_sc)
OS_sc = ({(1, {max}, 6.0),
(2, {sub}, 3.8)},
∅ ,∅)
I_he = ({I_sl }, OS_he ,yes , M_he)
OS_he = ({(2, {∅}, 3.8),
(1, {∅}, 2.1)},
{(round C)1, (⊥ S)2},
{(round ++ S)1 2})
I_sh = ({}, OS_sh ,yes , M_sh)
OS_sh = ({(1, {max}, 4.0),
(2, {sub}, 2.3)},
{(⊥ S)1, (round C)2},
{(round ++ S)2 1})
I_sl = ({}, OS_sl ,yes , M_sl)
OS_sh = ({(1, {max}, 3.8),
(1, {dist}, 0.3)},
(1, {sub}, 0.3)},
∅,
{(mir ++ S)1 2, (mir ++ S)1 3, (mir ++ S)2 3})
d1
d2
d3
local coordinate system
child coordinate system (transformed)
object schema
child object schema (transformed)

Figure 4: Example of an IDT representation for a stylized screw.

The first covers two dimensions *(d1, d2)* and represents the “diameter” with a numerical extent of 3.8 units. The second axis covers one dimension *(d3)* and represents the “height” of the cylinder which is 2.1 units. Since there is neither a perceptually dominant axis corresponding to “length”, nor a subordinated one corresponding to “thickness”, both axes are qualitatively described by the unspecified DAV ∅. The object schema definition is further augmented by profile vectors. It contains, for instance, the entry *(round, C)* for the first axis, where *round* is a symmetry property and expresses rotational symmetry of the axis, and *C* describes the curved boundary.

4 Using the IDT in a prototype system

The IDT model forms the conceptual basis to represent shape-related information acquired via gesture and speech for usage in an operational gesture understanding system. The applicability of the IDT representation and a gesture and speech processing model have been tested with a prototype system. Gesture (motion) data is captured via data-gloves and motion trackers. The system is able to recognize and to conceptualize shape-related gestures and verbal expressions and to determine target objects which most closely match the input. To give a rough idea, the process of interpretation is outlined in Figure 5. Gesture and speech are perceived and segmented. The result of the segmentation process are uninterpreted surface descriptions of single words and gestures. For gestures, this surface description consists of a collection of spatiotemporal features.

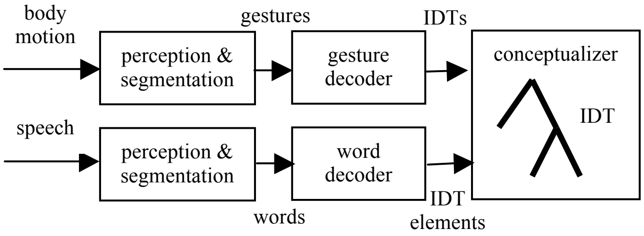


Figure 5: Interpretation process.

Two decoders, one for each modality, convert the surface descriptions into elements of an IDT representation. The word decoder retrieves an adjective's or noun's semantic representation from a lexicon in terms of a complete IDT. The gesture decoder analyzes the spatiotemporal features and transforms them into a set of object axis descriptions according to the form-meaning relations observed in the study. Since gesture and speech can be ambiguous, both decoders may output a set of alternative interpretations.

Figure 6 illustrates the decoding of a C-shape hand gesture. Subjects used it in two different ways (hand regions marked grey): to indicate extent between the thumb's and index finger's tip and to depict a round profile with the curvature of the fingers. The former interpretation is represented by a 1-D object axis *(1, {}, d)*, while the semantics of the latter is described by a 2-D object axis *(2, {}, dia)* with additional boundary information *(round, C)*. Which one of the two interpretations is correct cannot be determined at this stage without further contextual information. Thus, both of them are forwarded to the next processing stage. The subsequent processing stage, called conceptualizer in rough accordance with the speech production and comprehension model suggested by Levelt (1989), maintains a spatial context model in form of a dedicated IDT. This model can be considered the system's "spatial imagination". In the conceptualizer, incoming interpretations from the decoders are unified with the current model.

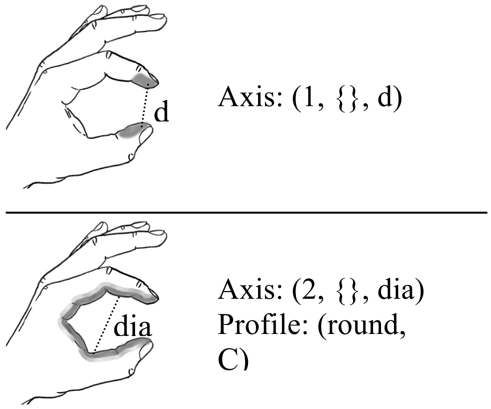


Figure 6: Two different semantic interpretations of the "C"-hand-shape in terms of IDT elements.

Integration of IDTs from verbal information is formally accomplished via a unification procedure that merges two compatible IDTs into a single one. Object axes resulting from gesture interpretation are inserted into the existing IDT. That way, successive gestures and words are integrated step-by-step to result in a unified spatial representation of an object description. Alternative interpretations may be ruled out during unification due to incompatibilities. Eventually, only one unified interpretation remains.

5 Discussion

While codified gestures with a fixed, culture-dependent meaning and pointing gestures have been described to some extent in the literature, there is not much work on the morphology and semantics of iconic gestures in specific domains. Exceptions are the early work by McNeill and Levy (1982) who first examined verbal and gestural representations used to depict cartoon narrations, and, more recently, Kopp, Tepper, and Cassell (2004), who examined iconic gestures in route descriptions for gesture synthesis in embodied conversational agents. A comprehensive study on the use of gestures for product design processes including shape description was conducted by Hummels (2000). In contrast to the work presented here, her study focuses not on coverbal, but autonomous gestures performed independently of speech.

Computational models for the comprehension of iconic gestures and semantic fusion with speech are rare. Most work has focused on pointing gestures or symbolic gestures instead, regarding gesture comprehension as a mere pattern classification problem. Seminal work on the understanding of iconic gestures for object placement and movement descriptions was done by Koons, Sparrell, and Thorisson (1993). Yet, their approach is focused on applying spatial transformations depicted with iconic gestures ("place the box here", "move it like this") to objects, but not on the integration of verbal and gestural information.

Our work addresses the lack of a semantic foundation for the integration of iconic gestures and speech in a specific domain. Extending the approach to multimodal dialogue systems, the IDT model could serve as a part of the spatial discourse context shared between a human and a computer system embodied in a virtual agent or robot. This would enable both interlocutors to fill space with meaning.

6 Conclusion

This chapter has addressed the meaning of shape-related iconic gestures. It has considered how such meanings can be accessed and modelled, as well as how they can be unified with the semantics of shape-related verbal expressions. Based on a comprehensive corpus of speech-gesture shape descriptions acquired from an empirical study, we have proposed the Imagistic Description Tree (IDT) as a representation for the semantics of multimodal shape-related expressions, and outlined its application in a gesture understanding system. The IDT models object extent, profile, and structure as the salient semantic elements contained in gesture and speech. The IDT representation is an important step towards capturing the meaning of iconic gestures in formal terms and making possible their computational treatment together with speech. An application has been outlined that can algorithmically generate interpretive operational shape descriptions from gesture and speech input modalities.

References

Bavelas, J.B., & Chovil, N. (2000). Visible acts of meaning. An integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology*, *19*(2), 163-194.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Clementini, E., & Di Felice, P. (1997). A global framework for qualitative shape description. *GeoInformatics, 1,* 11-27.

Cohn, A., & Hazarika, S. (2001). Qualitative spatial representation and reasoning. *Fundamenta Informaticae, 46*, 1-29.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Emmorey, K., Tversky, B., & Taylor, H. A. (2000). Using space to describe space: Perspective in speech, sign, and gesture. *Spatial Cognition and Computation*, *2*, 157–180.

Enfield, N.J. (2004). On linear segmentation and combinatorics in co-speech gesture. *Semiotica, 149-1/4*, 57-123.

Hummels, C. (2000). *Gestural design tools: prototypes, experiments and scenarios.* Doctoral dissertation, Technische Universiteit Delft.

Kelly, S.D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89*(1), 253-260.

Kendon, A. (2004). *Gesture. Visible action as utterance.* Cambridge, UK: Cambridge University Press.

Kimbara, I. (2006). On gestural mimicry. *Gesture, 6*(1), 39-61.

Koons, D. B., Sparrell, C. J., & Thorisson, K. R. (1993). Integrating simultaneous input from speech, gaze and hand gestures. In M. T. Maybury (Ed.), *Intelligent multimedia interfaces*. Cambridge (MA): MIT Press.

Kopp, S., Tepper, P., & Cassell, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (pp. 97-104). New York: ACM Press.

Lang, E. (1989). The semantics of dimensional designation of spatial objects. In M. Bierwisch & E. Lang (Eds.), *Dimensional adjectives: Grammatical structure and conceptual interpretation*. Berlin, Heidelberg, New York: Springer.

Levelt, W. (1989). *Speaking.* Cambridge, Massachusetts: MIT Press.

Marr, D., & Nishihara, H. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society, Series B*, *200*, 269–294.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago: The University of Chicago Press.

McNeill, D. (2005). *Gesture & thought*. Chicago: The University of Chicago Press.

McNeill, D., & Levy, E. (1982). Conceptual representations in language activity and gesture. In R. Jarvella & W. Klein (Eds.), *Speech, place, and action* (pp. 271–295). Chichester: John Wiley & Sons.

Mortenson, M. E. (1997). *Geometric modeling*. New York: Wiley.

Shi, H., & Tenbrink, T. (this volume). Telling Rolland where to go: HRI dialogues on route navigation.

Sowa, T. (2006a). *Understanding coverbal iconic gestures in shape descriptions.* Berlin: Akademische Verlagsgesellschaft Aka.

Sowa, T. (2006b). Towards the integration of shape-related information in 3-D gestures and speech. In *Proceedings of the Eighth International Conference on Multimodal Interfaces* (pp. 92-99). New York: ACM Press.

Sowa, T., & Wachsmuth, I. (2002). Interpretation of shape-related iconic gestures in virtual environments. In I. Wachsmuth & T. Sowa (Eds.), *Gesture and sign language in human-computer interaction*. Berlin: Springer.

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.), *Gestures: Meaning and use* (pp. 365–376). Porto, Portugal: Edições Universidade Fernando Pessoa.

Striegnitz, K., Tepper, P., Lovett, A., & Cassell, J. (this volume). Knowledge representation for generating locating gestures in route directions.

Vorwerg, C. (this volume). Consistency in successive spatial utterances.