A Model for the Representation and Processing of Shape in Coverbal Iconic Gestures

Timo Sowa (tsowa@TechFak.Uni-Bielefeld.DE)
Ipke Wachsmuth (ipke@TechFak.Uni-Bielefeld.DE)
AI Group, Faculty of Technology, Bielefeld University
33594 Bielefeld, Germany

Abstract

Verbal descriptions of object shape are frequently accompanied by iconic gestures expressing meaning via similarity to the referent. This paper presents a study examining the morphological variety of shape-related iconic gestures and the way they express aspects of object shape. It is shown that information about object extent, object boundary, and part structure is reflected both in coverbal gestures and their verbal affiliates. Based on the empirical results, a computational model for the representation and processing of multimodally communicated object shape is proposed.

Introduction

Iconic gestures are meaningful movements of the hands and arms that coincide with speech (are *coverbal*); they are semantically related to the content of speech (*co-expressive*), but convey meaning in a different way (McNeill, 1992). While speech is a symbolic modality of communication unfolding in time, iconic gestures unfold in time and space and express meaning via similarity to the referent. Though the phenomenon of iconic gestures is widely discussed in the literature, there are few concrete suggestions about the type of information they express and how their semantic content can be accessed and modeled computationally. The study as well as the modeling approach presented in this paper aim to fill this gap for the domain of shape descriptions. It is an attempt to identify and model recurring form features, units of meaning, and structures in shape-related gestures leading to a domain-specific computational model for gesture understanding.

Related Work

Shape-related coverbal gestures are not yet systematically described in the literature. However, there is related work on the morphology and semantics of iconic gestures in other domains. McNeill and Levy (1982) first examined verbal and gestural representations used to depict cartoon narrations. A comprehensive study on the use of gestures for sketchy product design 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 understanding of iconic gestures are rare. Most work has focused on symbolic gestures instead, regarding gesture understanding 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).

Study

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 subjects were asked to describe five different stimulus objects (Fig. 1). 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 rated by at least two coders were included in the analysis.

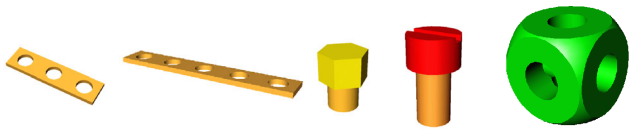


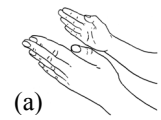
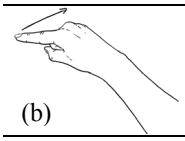
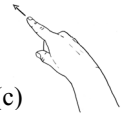
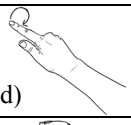
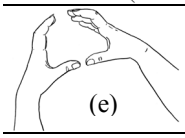
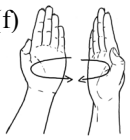
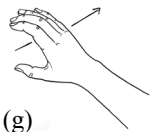
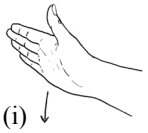
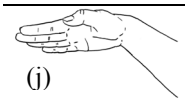
Figure 1: Stimulus objects used in the study.

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. Each gesture kind can be represented by a prototype which is an idealized realization of the form-meaning relation (Table 1). Four general gesture types can be distinguished as given below.

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 (a). The term profile refers to the course of the object's boundary and usually involves some kind of motion (b-h).

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

 (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-circluar 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

Dimensional gestures often depict "abstract" one- or two-dimensional characterizations of the 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. 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. Gesture (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 (semi-) cylindrical shape.

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) is an example of this type: The flat, moving hand indicates a particular planar side of the object without referring to the whole.

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.

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.

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" 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 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, subjects did depict the whole screw in an abstract way reducing it to its main extent.

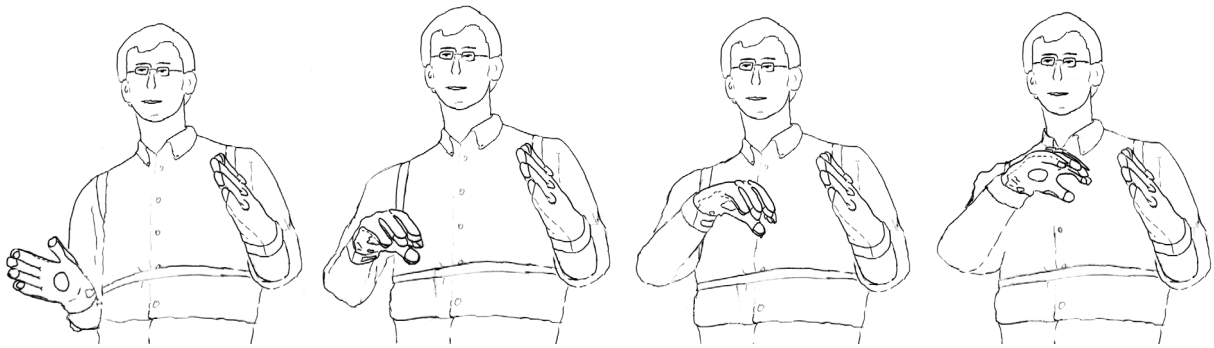


Figure 2: Explicit spatial cohesion via a two-handed gesture. Left hand is held in position.

Spatial organization

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). Examples of spatial organization can be found in the corpus data. Consider the gestures accompanying the description of the short bar (Fig. 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. 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*. 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. Fig. 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 of two or more different objects. Fig. 3b is 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. 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).

Speech

Table 2 shows the frequencies of the parts of speech among the affiliates in relation to their base frequency in the whole corpus. It is evident that adjectives and nouns are over-represented among the affiliates, while the other classes are underrepresented.

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.

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

	total (%)	relative
nouns	42.9	1.58
adjectives	29.5	4.79
verbs	4.0	0.26
prepositions	5.2	0.63
adverbs	14.2	0.65
determiners	4.2	0.27
n	478	

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.

Representation Model

Taken together, the corpus evaluation revealed three important factors to consider in a semantic representation of shape-related gestural and verbal expressions. 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. Furthermore, 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

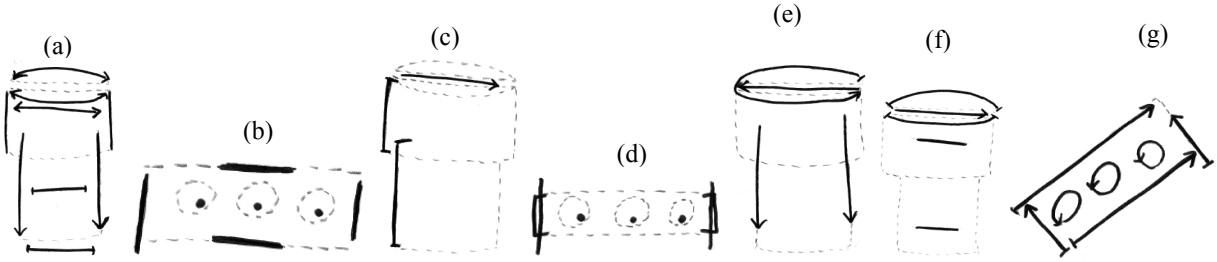


Figure 3: Implicit spatial cohesion. Solid lines indicate gesture locations (arrows stand for movement in dynamic gestures), dotted lines show the reference object.

gestures, but which has not included structured spatial organization of gesture and accompanying speech reflecting this factor (Sowa & Wachsmuth, 2002).

Approaches towards shape representation

Existing models for shape representation from different disciplines such as visual cognition or linguistics are briefly reviewed here with respect to their applicability to gesture representation. The 3-D model suggested by Marr & Nishihara (1978) provides some of the basic requirements described above. Their approach uses perceptual object axes as basic elements of shape. Axes are hierarchically arranged according to different levels of granularity. The disposition of a lower-level axis (part) with respect to the higher-level axis (whole) is explicitly encoded. One problem with the 3-D model is its incapacity to represent objects without a dominant axis such as coins or spheres. This problem does not appear in approaches based on volumetric primitives such as the geon model by Biederman (1987). However, this model lacks part-whole relations in the sense of different levels of abstraction. Furthermore, geons are inherently defined in 3-D and cannot be underspecified. A one-dimensional gesture specifying only one object extent could thus not be adequately represented. A suitable approach for the definition of the principal extent(s) of objects is provided by Lang (1989) within a framework for a semantic theory on dimensional adjectives. Lang defines representations called *object schemata* describing the basic gestalt properties of objects. However, the object schema approach is not hierarchical and does not distinguish between parts and wholes.

Summarizing, 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.

Modeling extent properties

For the modeling 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) dimensions.

More formally, an axis A is defined as a triple A = (i, Δ, deg) where *i* defines the integration level (1, 2, or 3), Δ a set of qualitative properties of the axis called *dimensional assignment values (DAVs)*, and *deg* a measure for the axis' numerical extent. An object schema S is then defined as a collection of one up to three object axes: $S = \{A_1, \ldots, A_n\}$. 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 than the other axes in a schema. Using different combinations of axes in an object schema, several basic objects can be represented as illustrated in Table 3. The first eight schemata in the table specify shape in three dimensions, while the last four show cases of dimensional underspecification.

Table 3: Representation of basic object types with object schemata.

Object schema	Prototype
{(1,{∅},⊥), (1,{∅},⊥), (1,{∅},⊥)}	
{(1,{*max*},⊥), (1,{∅},⊥), (1,{∅},⊥)}	
{(1,{∅},⊥), (1,{∅},⊥), (1,{*sub*},⊥)}	
{(1,{*max*},⊥), (1,{∅},⊥), (1,{*sub*},⊥)}	
{(1,{∅},⊥), (2,{∅},⊥)}	
{(1,{*max*},⊥), (2,{*sub*},⊥)}	
{(2,{∅},⊥), (1,{*sub*},⊥)}	
{(3,{∅},⊥)}	
{(1,{∅},⊥), (1,{ ∅},⊥)}	
{(1,{*max*},⊥), (1,{∅},⊥)}	
{(2,{∅},⊥)}	
{(1,{*max*},⊥)}	

Object schemata are adequate for encoding dimensional gestures and accompanying nouns or adjectives. 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. The corresponding semantic encoding would be a schema containing two axes, i.e., a one-dimensional axis representing the movement, and a two-dimensional axis representing the round shape, i.e. {(1, {max}, 40.0), (2, {sub}, 20.0)}.

Modeling 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. An example given in the following section will clarify the use of profile properties.

Modeling 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. For that purpose schemata can be arranged in a tree similar to the hierarchical structure used in the Marr & Nishihara (1978) model.

Structural aspects are represented in *imagistic descriptions*. An imagistic description *I = (C, S, a, M)* for an object consists of a set *C* of imagistic descriptions describing its parts, an object schema *S* defining its overall proportions, a spatial anchor flag *a*, and a transformation matrix *M*. The recursive definition in *C* provides a tree-like structure: The parts described in *C* are imagistic descriptions which could themselves contain further parts. The number of children is arbitrary, and if an object has no parts, C is empty. The flag *a* signals whether the description is spatially anchored in a parent coordinate system. If its value is "yes", the matrix *M* defines the position, orientation, and size of the object or part in relation to the parent description. The complete tree describing an object including all parts, parts of parts etc. is called *Imagistic Description Tree (IDT)*.

Fig. 4 shows an example of an IDT model for the screw. The part hierarchy modeled 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 modeling the parts, I_{he} for the head and I_{sh} for the shank. The head has another child node I_{sl} representing the slot.

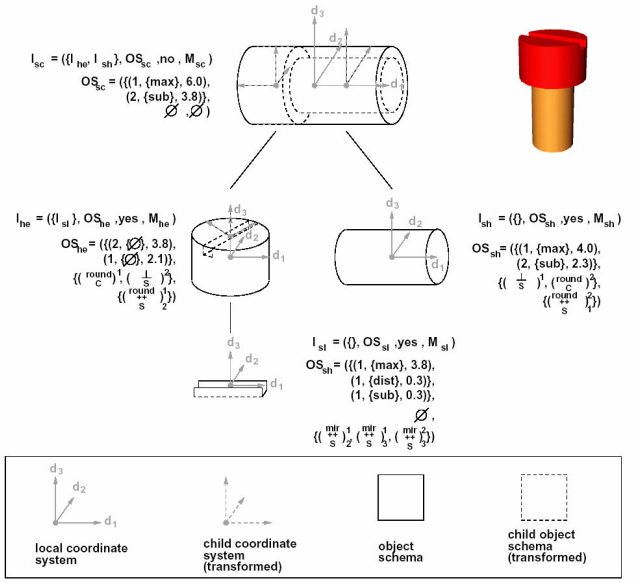


Figure 4: Example of an IDT representation for a stylized screw.

Without providing all formal details of the IDT definition, 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: 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 value and expresses perfect rotational symmetry of the axis, and *C* describes the curved boundary.

Using the IDT

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 has been tested with a prototype implementation. 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 matches to the input.

To give a rough idea, the process of interpretation is outlined in Fig. 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.

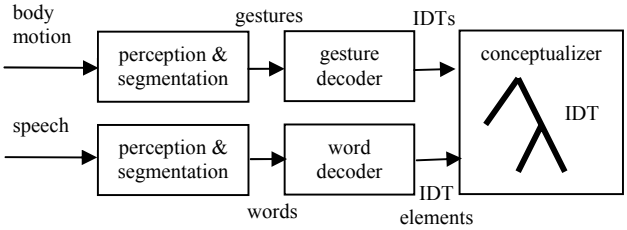


Figure 5: Interpretation process.

Two decoders, one for each modality, convert the surface descriptions into elements of an IDT representation. The word decoder looks up a lexicon to retrieve a word's semantic representation 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. Fig. 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, while the semantics of the latter is described by a 2-D object axis with additional boundary information contained in the profile vector ***p***. In both cases center ***c*** and orientation ***o*** of the axes are computed in absolute coordinates.

The subsequent processing stage, called conceptualizer in rough accordance with the speaking 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. 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.

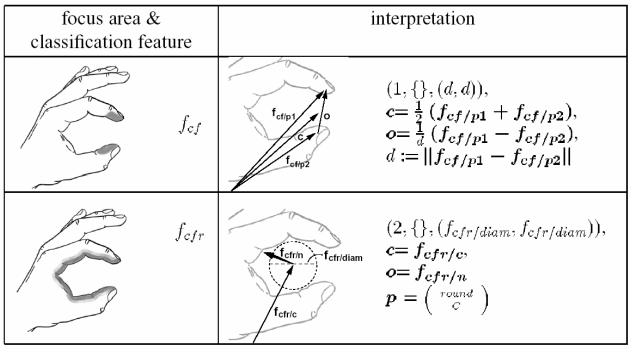


Figure 6: Two different semantic interpretations of the "C"-hand-shape in terms of IDT elements.

Conclusion

What is the meaning of shape-related iconic gestures, how do we access and model it, and how can it be unified with the semantics of shape-related verbal expressions? Based on the results of an empirical study we 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 representation and processing approach is one step towards capturing the meaning of iconic gestures in formal terms and make possible their computational treatment together with speech.

References

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

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.

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.

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: University of Chicago 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.