Data Miming: Inferring Spatial Object Descriptions from Human Gesture

Christian Holz[1,2] and Andrew D. Wilson[2]

[1]Hasso Plattner Institute
Potsdam, Germany
christian.holz@hpi.uni-potsdam.de

[2]Microsoft Research
Redmond, WA 98052 USA
awilson@microsoft.com

ABSTRACT

Speakers often use hand gestures when talking about or describing physical objects. Such gesture is particularly useful when the speaker is conveying distinctions of shape that are difficult to describe verbally. We present *data miming*—an approach to making sense of gestures as they are used to describe concrete physical objects. We first observe participants as they use gestures to describe real-world objects to another person. From these observations, we derive the data miming approach, which is based on a voxel representation of the space traced by the speaker's hands over the duration of the gesture. In a final proof-of-concept study, we demonstrate a prototype implementation of matching the input voxel representation to select among a database of known physical objects.

Author Keywords

Gestures, shape descriptions, 3D modeling, depth camera, object retrieval.

ACM Classification Keywords

H5.2 [Information interfaces and presentation]: User Interfaces. Input devices & strategies.

General Terms

Design, Experimentation, Human Factors.

INTRODUCTION

In conversation we sometimes resort to using hand gestures to assist in describing a shape, particularly when it would be cumbersome to describe with words alone. For example, the roofline of a new car might be conveyed by a swoop of the outstretched hand, or a particular chair style might be indicated to a shopkeeper by a series of gestures that describe the arrangement of surfaces unique to that chair. In such cases, the speaker often appears to trace the precise 3D shape of the described object. Meanwhile, the listener appears to effortlessly integrate the speaker's gestures over time to recreate the 3D shape. This exchange strikes us as a remarkably efficient and useful means of communicating the mental imagery of the speaker.

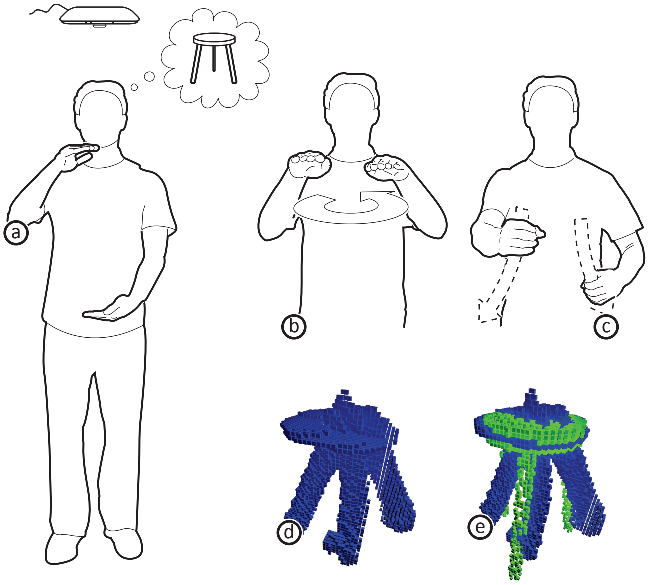

Figure 1: Data miming walkthrough. The user performs gestures in 3-space, as they might during conversations with another person, to query the database for a specific object that they have in mind (here a 3-legged stool). Users thereby visualize their mental image of the object not only by indicating the dimensions of the object (a), but more importantly the specific attributes, such as (b) the seat and (c) the legs of the chair. Our prototype system tracks the user's gestures with an overhead camera (a) and derives an internal representation of the user's intended image (d). (e) The query to the database returns the most closely matching object (green).

In this paper, we consider the use of gestures to describe physical objects. We present *data miming* as an approach to enable users to spatially describe existing 3D objects to a computer just as they would to another person.

We make two contributions in this paper. First is an *observation* of how people use gestures in a *natural way* to describe physical objects (i.e., without telling them how to use a certain gesture to specify a certain part of an object). From these observations, we derive the *data miming* approach to making sense of gestures as they are used to describe physical objects (i.e., which object was described). Our second contribution is a prototype system, which allows for walk-up use with a single overhead depth camera to sense the user's gestures. Our system follows a query-by-demonstration approach and retrieves the model in a database that most closely matches the user's descriptions.

We envision a number of practical applications for data miming. For example, shoppers in a furniture warehouse could approach a kiosk to perform gestures describing an object that they are looking for. The system would look up the closest matches in a database and present the results alongside the locations to the user. In an immersive video game (e.g., Kinect [31]), a player could summon objects by describing them spatially. The game might tailor the object to the precise dimensions indicated by the player's gestures.

DATA MIMING

In such scenarios, data miming *observes* and *makes sense* of human gesture, exploits the human sense of spatial references, and uses the richness of hand shape and motion when describing objects to *infer* the described objects (Figure 1). Similar to using gestures when talking to a human observer, data miming observes passively, thereby providing no feedback during the gesture (a-c). The user, therefore, works solely from a mental image of the described object and their gestures are used to *implicitly* create a virtual representation of the mental image (d).

The virtual representation can be used to *classify* the described object (e.g., "stool"), but also to extract the object's *specifics* in order to distinguish it from other instances in that class. For example, a user may describe *a* chair (where no further detail is required), but may also describe a *particular and existing* chair that has three legs, slanted from the center, and two feet tall (Figure 1). Without those specific details, the reference to that particular chair would be unclear.

Method

We approach this topic as follows. In an initial user study, we observe participants describing real-world objects to another person using gestures. We instruct participants to describe objects just as they would in a conversation. The purpose of this study is to determine *how* participants describe the various objects, i.e., what gestures they use, and on which parts of objects they concentrate. We analyze the patterns that recur when participants specify object dimensions and boundaries, define object faces and the shape of single parts. In addition, we observe how participants distinguish between deliberately describing meaningful parts of an object and transitioning between such parts.

From our observations, we derive the *data miming* approach, which describes how to create a virtual representation of the user's mental image from the performed gestures. We implement a prototype system that observes the user with a single depth-sensing camera and then matches the virtual representation against objects in a database.

We then run a proof-of-concept study on our prototype system to determine if the features extracted from users' actions suffice to classify and match a described object. We find that our system correctly identifies described objects in 60% of all cases from a set of 10 potential matches.

RELATED WORK

Data miming is related to sculpting, hand tracking through computer vision, three-dimensional shape querying, spatial memory, gesture input, and linguistic models of gesture.

Sculpting

Similar to our *data miming* approach, many sculpting systems make use of the expressiveness of human gesture. While some systems allow the user to shape and transform 3D objects using their hands (e.g., Gesture-based 3D modeling [19], *SketchMaker* [22], *Twister* [17], *Surface Drawing* [24]), others equip the user's hands with tools (e.g., *Volume sculpting* [28], heat gun and sandpaper in *Sculpting* [8]). Other systems derive 3D models from the user's sketches in the air using a stylus (e.g., *3-Draw* [23], *Spatial Sketch* [29], *HoloSketch* [5]). Users typically create and refine the 3D model iteratively based on the visual feedback provided by the sculpting application, which allows them to create high-quality 3D models. For example, different hand postures result in different brush widths in *Surface Drawing* [24] and users create high-quality virtual models with a physical brush in *CavePainting* [14].

While sculpting applications enable users to *create* new high-quality models, our *data miming* approach focuses on enabling users to specify *existing* objects. Sculpting applications are inherently *interactive*, which causes a feedback loop on the user's end; to create the intended model, a user verifies if their actions had the intended outcomes (i.e., all input needs to be rendered). The user thus directly works on the visual representation of the model. *Data miming*, in contrast, passively *observes* how the user *acts* and does not provide any feedback; the user thus works solely from their conceptual model of the object.

Spatial Memory

As it provides no feedback, *data miming* assumes that users maintain a frame of reference when performing gestures. Hinckley et al. found that the user's body becomes the spatial reference [12] in this situation; users do not rely on visual feedback when using both hands together [11]. Humans further have the ability to know where both hands are relative in space [12, 23] and can maintain a spatial anchor outside their body over a brief amount of time [3]. Baddeley and Hitch attributed this short-term visual memory for maintaining spatial relationships to the visuospatial sketchpad, a part of the working memory [1]. Gustafson et al. exploit this, such that users set a reference frame for interaction using their left hand, and draw or pick locations with their right hand without visual feedback [10].

Natural Gesture

Linguists have studied how speakers use gesture in conversation. MacNeill [18] places a series of gesture types (originally described by Kendon [15]) along a continuum according to the degree to which gesture *complements* speech. Kendon's *gesticulation* category includes those gestures that complement speech the most. MacNeill further categorizes gesticulation to include the ubiquitous *beat* gestures, which are used for emphasis, *deictic* gestures for indicating objects (pointing), *metaphoric* gestures to convey abstract meaning, and *iconic* gestures. Iconic gestures depict a concrete object or event, bear a close formal relationship to the content of the speech, and are used when trying to describe the shape or form of an object. Moving along Kendon's continuum, *emblems* include signs that carry meaning only by convention, such as the "OK" sign. *Pantomime* gestures are similar to iconic gestures in that they depict objects or actions, but do not require speech (and are distinct from gestures used in theatrical miming). Finally, sign languages stand in place of spoken languages and therefore complement speech the least.

Data miming performs matching against stored 3D models of objects. Because these models are not based on convention but rather the actual shape of real physical objects, we argue that data miming gestures are iconic or pantomime gestures, not emblems. While in the case of specifying the class by speech (e.g., "chair"), gesture and speech are clearly complementary; the dependence of one on the other does not have the complexity typical of natural iconic gestures. The gestures modeled in the present work are therefore probably best called 'pantomime' gestures.

Gesture-based systems

Many gesture-based input systems allow users to control applications. While such gestures ideally resemble corresponding real-world gestures (e.g., pushing a set of floating objects aside in *g-speak* [20], pointing in the *Perceptive Workbench* [26]), other operations do not afford a "natural" gesture and need to be defined (e.g., an input vocabulary for controlling presentations in *Charade* [2]). *Gesture Pendant* recognizes the user's hand pose and movement for the continuous control of devices [25].

Hand pose and position recognition

Many of the previously mentioned systems use gloves and motion capture devices to determine the stylus' or hand's location and pose in space. While this approach of tracking input is reliable, it requires the user to wear a device.

Alternatively, hand poses and gestures have been sensed with body-attached cameras (e.g., [10], [25]), side-view image and cast shadows (*Perceptive Workbench* [26]), and stereo images (e.g., [16]). Pavlovic et al.'s survey provides further information on modeling hands and interpretation of hand gestures [21].

Related work has often used props to distinguish between meaningful parts of gesture input from random hand motions, such as pressing a button (e.g., [14],[17],[29], aligning the thumb [24]) or pointing at a specific region to perform gestures (e.g., pointing at the screen in *Charade* [2]).

Query-by-3D model

While our prototype does not focus on the speed of matching 3D models, related work has addressed efficient 3D queries to object databases through efficient matching algorithms [7]. *Modeling-by-example* allows partial matching of 3D objects based on single parts [6]. (Tangelder and Veltkamp present a thorough review [27].)

EXPERIMENT: HOW DO USERS DESCRIBE OBJECTS?

When humans describe objects by gesture, they have a number of means to express geometry. The purpose of this user study was to find recurring patterns in those gestural descriptions. We also examined how participants distinguished between the aforementioned meaningful and random parts of gestures.

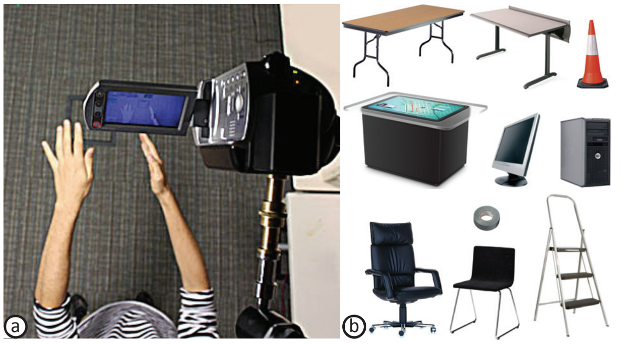

Figure 2: (a) Study setup. Participants described objects using gestures to the experimenter, recorded by an overhead camera for later analysis. (b) Participants described 10 objects.

Task and Procedure

Figure 2a shows a participant during the study, describing one of the ten objects (b). All objects were located behind the participant. For each object, the participant took a close look at the object, then turned away from the object, and described the object through gestures to the experimenter. Participants were only instructed to describe objects using gestures—not postures (e.g., assuming the posture in an office chair to augment their descriptions), but received no instructions as to which gestures to use. Participants did not use speech during the study, received no feedback, and could not see what the camera recorded. Participants finished a description of an object by lowering their arms.

As shown in Figure 2b, the objects included primitive shapes (e.g., PC = box shape, cone, tape), objects of medium complexity (table), as well as more elaborate objects (office chair, ladder). Our intention was to find the degree to which participants' descriptions agree, and the features they include when describing more complex objects.

All participants were recorded with an overhead video camera for later investigation. Overall, participants described all objects in less than six minutes. Afterwards, they filled out a questionnaire on prior experience with 3D modeling applications to determine if the tools in such applications inspired and influenced their performances.

Apparatus and Participants

The overhead video camera was mounted to the ceiling 8ft from the ground. This setup allowed the camera to capture a top-down volume of 5ft (W) × 3ft (H) × 5ft (D). We recruited 12 participants (4 female) from our institution. All participants were between 21 and 46 years old (*M*=29.5, *SD*=6.2).

Results and Observations

All participants not only maintained relative proportions of an object's parts, but also maintained relative scale across objects. For example, participants used a large fraction of their arm's length to describe the two tables, while describing the chairs smaller as appropriate. It was interesting to see participants' notion of space in each of the three dimensions (Figure 3); particularly for larger objects, participants seemed to scale objects non-uniformly to adapt sizes relative to the area covered by arm's length in each direction. This implied that, by nature of the human body, objects could always be wider than tall, and taller than deep.

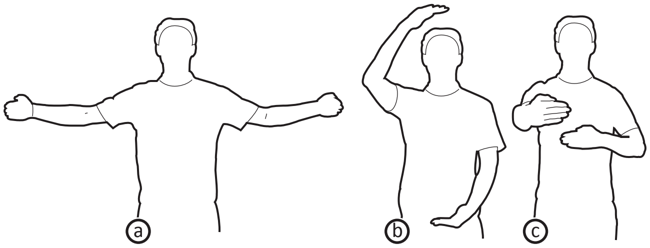

Figure 3: Arm's length is the limiting factor for specifying dimensions. When necessary, participants seemed to describe objects in non-uniform scale according to the available span for each dimension. Objects can thereby be (a) wider than (b) tall, and taller than (c) deep.

Participants mostly used a top-down approach to describe objects; after larger, more apparent surfaces they described the smaller parts. It was apparent that all participants distinguished between surfaces (e.g., flat faces of PC, monitor, Microsoft Surface, table, but also curved surfaces, such as frame of the ladder, seat and backrest of the chairs) and smaller components, such as struts and connections. We thus analyze them separately.

Surfaces and Faces of the Object

Symmetry was most apparent during participants' descriptions; all twelve used both hands in a parallel pose, facing one another to define symmetric elements of an object (e.g., PC, monitor, Microsoft Surface). Those symmetric parts did not necessarily represent the dimensions of the entire object, but would specify certain parts. Participants also used simultaneous and symmetric hand movement to describe smaller parts such as legs of chair, or frame of the ladder.

Dimensions: When the shape of objects resembled that of a box, all participants defined the dimensions of parts of objects (e.g., PC, Microsoft Surface, and monitor). Seven participants simultaneously moved both hands in a flat pose back and forth along the bounding dimensions of the object repeatedly. Three others held both hands flat in place to define those boundaries. Two participants drew wireframes of objects in box shape.

Large surfaces: All but two participants used their hands to "trace" surfaces, i.e. they moved their flat hands along surfaces, as if wiping them with their hands (e.g., tables, top of Microsoft Surface, seats of the chairs). Six of them even wiped the area within boundaries to "fill" it (Figure 4a).

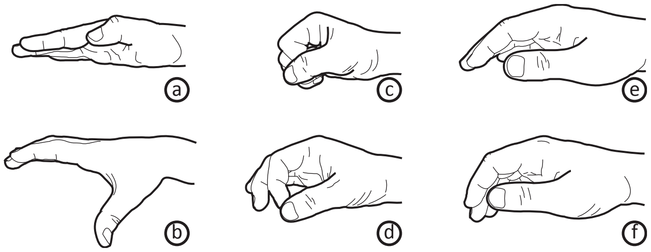

Figure 4: Hand postures used by participants to describe (a) flat and (b) curved surfaces, (c&d) struts and legs. (e&f) Hands were relaxed when not tracing a surface.

Medium surfaces: Four participants specified the outline of surfaces with their flat hands and, again, wiped the enclosed area to "fill" it (e.g., monitor, backrest of the office chair). The other eight participants abstracted those medium surfaces to a mere stroke of their flat hand, which they performed repeatedly. This was most noticeable for the two chairs. Four participants sometimes described a surface only by waving their hand repeatedly, roughly in the place of a surface (e.g., monitor, seats).

Small surfaces: All participants used their hand to "trace" smaller components of objects (e.g., steps of the ladder, outside frame of the ladder).

Curved/non-planar surfaces: All 12 participants adapted the shape of their hand to match the curved surface and "wiped" up and down the surface repeatedly. To describe the cone, for example, all participants formed a closed circle using both thumbs and index fingers and then moved their hands down, thereby driving them apart. Their fingers thereby maintained the original shape (Figure 4b).

Smaller Components

Symmetry: As with object faces, participants moved both hands symmetrically and simultaneously if the object afforded it.

Bars, struts, and legs: 6 participants used a fist and moved it along the bar to represent a straight bar (e.g., legs of the table, tripod, chair, pole of the office chair, Figure 4c). The other 6 pinched their thumb and index fingers and moved them along the bar (Figure 4d).

Interestingly, only 4 participants tried to match the actual size of a bar with their hand and only when describing objects with connected bars of varying diameters (e.g., tripod, table). The shape of their hand symbolized grabbing around bars, and they opened their hand accordingly if the bar was too big to fit inside (e.g., leg of the table).

For bigger struts, 6 participants brought their hands to a close distance, held them parallel or connected fingers and palms of both hands to enclose the space between the hands, and moved both hands to trace the shape of a component (e.g., pole and legs of the chairs, monitor stand).

Level of detail: 3 participants abstracted the complex shape of the office chair's foot base to a single disc. All other participants ignored the base and indicated the stand only.

Questionnaire: While nine of the 12 participants reported prior usage of 3D modeling applications (e.g., AutoCAD, Sketchup and the like), only two participants stated that those tools influenced their descriptions. They reported trying to perform extrusion/push and pull operations by moving their hands, and other times they performed sweeping and revolving.

Discussion and Analysis of Patterns

All participants began describing an object spatially in a top-down fashion immediately after looking at it. While participants received no instructions as to the detail of their description, all participants followed a similar approach. They abstracted the form of the object, often specified large components and faces first, and finally described some of the characteristic, but smaller components. For instance, while all participants indicated the armrests, the pole and foot of the office chair, few described the support of the armrests or the bars connecting the backrest to the seat. Similarly, participants described the ladder by indicating all three steps and then highlighting the outer frame.

Participants often described those parts first that most clearly represented the function of the object (e.g., backrest, seat of the chairs, top of MS Surface, steps of the ladder). They then described the parts that hold the object together.

Participants made use of symmetric appearances whenever possible; they used both hands with mirrored gestures to describe the shape. Likewise, participants used both hands to specify dimensions, either by defining constraining planes or "drawing" the bounding box. The actual dimensions of medium- and small-sized surfaces seemed to be unimportant to participants, as only a few times did a participant constrain the dimensions of such objects.

The majority of participants adapted the shape of their hand to that of the described object or component, stretching (to describe a planar surface, Figure 4a) or curling and bringing together fingers (for a round surface, Figure 4b) as necessary. In contrast, when participants moved their hands to the next part of an object, they would relax their hands and allow them to assume their natural posture (Figure 4e&f).

For smaller components of an object, such as bars and stands, participants had similar conceptual models for their description. They either formed a fist or pinched their thumb and index finger to indicate both round and square bars, along whose shape they then moved the hand. The majority of participants thereby ignored the actual diameter of those bars, using hand motion to indicate the shape of such bars (Figure 4c&d).

Expressiveness of the user's hand: hand postures

While participants mostly varied hand yaw and roll, they typically varied hand pitch only when indicate parallel parts by a vertical pose (Figure 3c). We assume this to be due to the limited range of angles for hand pitch. When the hand is vertical, however, moving the elbow can extend this range. In contrast, hand roll and yaw cover a larger range; elbow movement also supports the range of hand yaw.

We observed that participants focused on using their hands to describe shapes, and not their arms. This coincides with participants' statements.

Delimiters: meaningful vs. random parts of gestures

In addition to stretching the hand to indicate activity as mentioned above, all participants generally deliberately described parts of the object more slowly, while moving their hands faster when transitioning to another object part. For smaller surfaces, participants dwelled in one position for a brief amount of time. For lager surfaces, participants repeatedly described the surface and often more carefully than when moving their hands to another part.

Whenever two components were closely collocated, participants did not dwell between components, but rather treated them as a compound part and changed hand orientation while they were moving their hands (e.g., connected backrest and seat of a chair). Participants often repeatedly indicated this compound component through gestures.

THE DATA MIMING APPROACH

The previous observations allow us to design *data miming* as an approach to translate the observed gestures, as they occur, into implications for a virtual representation that seeks to reproduce the user's mental image. In particular, we argue that we need not rely on predefined gestures that manifest themselves as a particular part of an object upon recognition.

Basic Approach

The analysis of participants' gestures in the previous study showed that they often traced an object's surfaces and structural elements, thereby essentially recreating the object based on their spatial memory. That they often repeatedly traced those identifying parts suggests that the virtual representation of the user's description should also build up over time. Those parts that the user has spent more time describing should be weighted more strongly than parts they have covered only briefly.

Since participants mostly described surfaces of different sizes by waving their hand in the respective area, the user's hand should create a trace in the virtual representation. Since the actual path of the gesture is less important, the position and orientation of the user's hands are essential to translate motions correctly. In conjunction with the time-aware sensing of gestures, such traces become more meaningful to the virtual representation as the user repeatedly or more slowly covers a certain part of the object. The analysis of gestures thereby entirely focuses on the user's hands and neglects position and posture of the user's arms and body.

Meaningful vs. Random Motions

Ideally, our approach recognizes and translates only the meaningful parts of a user's gesture, while ignoring motions that only serve to transition the hands to the next part of the object.

Our first study showed that participants briefly relax their muscles while moving the hands to another part of the object, whereas they typically align or stretch their fingers, or flex their muscles to signal a meaningful hand pose. While it would be desirable to capture this distinction, changes in finger postures and curvature are fairly subtle (Figure 5).

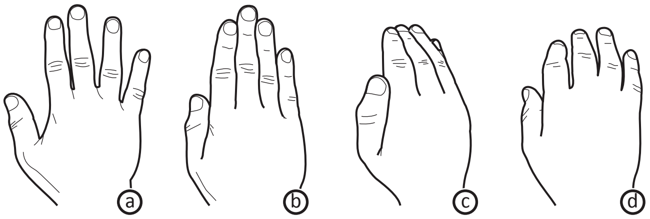

Figure 5: Hand poses and corresponding meanings. (a) Stretched hand, used to indicate a shape (e.g., flat surface), (b) stretched hand, fingers together, which also shows intention to indicate a shape, (c) curved hand shape and fingers together, suggesting that this motion is meaningful, and (d) relaxed pose when transitioning (cf. Figure 4e&f).

Considering the difficulty of sensing muscle relaxation with a camera, *data miming* forgoes the interpretation of finger curvature to derive the meaning of the current gesture. Instead, the user's hands constantly leave a footprint in the virtual representation whose position and orientation corresponds to those of the user's hands in the real world. That is, the orientation and posture of the hand at all times determines the volume of the component added to the virtual representation (i.e., a flat, tilted hand makes a flat, slanted small-sized impact on the virtual representation).

By simply replicating the volumes of the user's hands and representing them in the virtual space, our approach allows for sensing flat and curved hand postures (e.g., flat surfaces, surface of a sphere) and also accounts for smaller elements when users form a fist or pinch their fingers (e.g., legs of a chair). It also allows us to consider both hands separately.

Modeling the Scene

Data miming creates a virtual representation of the user's description in a discretized 3D volume consisting of $l \times m \times n$ voxels. This *voxel space* represents the "memory" of the system. Each voxel is either active or inactive. A scene begins with only inactive voxels and over the course of observing the user's gestures, voxels are activated as appropriate. Voxels also have a certain weight, which is increased as the user repeatedly activates the voxel. This allows us to capture how users trace the parts of the object: slower and more careful tracing indicates a more meaningful part and thus increased weight, while a faster (random) motion indicates a less meaningful part of the description (as in [13]). The set of voxels below a certain weight are thus ignored, leaving only meaningful parts of the gestures.

The 3D-scene approach we use is world-anchored, such that its location and orientation does not adapt to the user's position or orientation. While the center of the scene is always in front of the user (i.e., world anchored), users are able to maintain this spatial anchor [3], as object descriptions take only a few seconds in our case.

Retrieving matching objects from the database

To retrieve the identity of the user-described object, our approach relies on a database of candidate objects in voxel representation. Data miming selects the most closely matching object from the database as follows. For each candidate object, the user-created model is aligned with the database model for comparison and measurement of similarity. As byproduct we obtain the scale and rotation difference from the user's creation.

Account for the user's under-specification

As objects are mostly assembled from characteristic components, humans describe such characteristic parts separately. People also seem to make implicit assumptions about their audience; they do not describe less significant parts, parts that seem implicitly necessary (e.g., connecting parts between surfaces, such as backrest and seat in a chair) or features that do not serve to uniquely identify the object.

We reflect this *fragmentary* modeling on the user's part in the matching process by allowing the user to omit any part of the object, trusting that the user will specify enough detail given some familiarity with the class of objects under consideration and the variability of shape within that class. In the study of discourse, Grice's maxims of co-operation describe the tendency for speakers to lead the listener to a correct understanding. In particular, Grice's Maxim of Quantity holds that the speaker will make their contribution as informative as required, and no more [9].

IMPLEMENTATION OF THE DATA MIMING SYSTEM

We implemented a prototype system based on the *data miming* approach described in the previous section. This prototype works on an end-user system and requires only a single depth-sensing camera that is mounted above the user.

Hardware and Processing

Our prototype system uses a Microsoft Kinect camera which provides depth images at 30Hz and a resolution of 640×480 (Figure 6a). The camera has a diagonal field-of-view of 70°. The prototype processes each camera frame in less than 15ms, thus providing real-time processing of the user's gestures and translation into *voxel* representation.

Our prototype system first transforms every pixel in the input image into world coordinates and then crops coordinates (i.e., pixels) outside a volume of 3ftW × 2ftH × 2.5ftD (Figure 6b). This removes the floor, walls, and potential other objects from the depth image (e.g., the chair and table in Figure 6a). The prototype then identifies the user's arms in the image, distinguishing between contiguous regions with only gradually changing depth values to account for overlapping arms, and extracts the user's hands from those regions (c). The prototype system thereby assumes that the user's arms enter from outside and reach into the volume (as in [26]). Our prototype then finds the most-distant point of the hand, measuring distance as the length of a path *within* the shape of the arm (i.e., not Euclidean distance), to account for bent elbows and wrists. To extract the user's hands, we assume a constant hand length (depending on the distance to the camera), which has proven to work well in our tests. Our prototype also provides an optional calibration for the user's particular hand size.

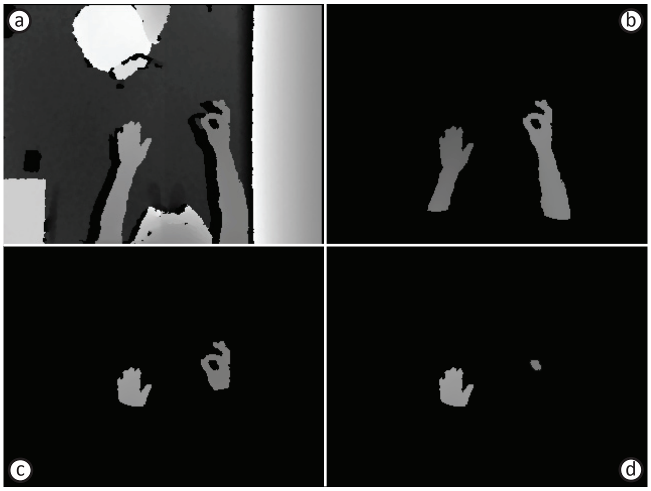

Figure 6: Our prototype processes the raw image (a) as follows: (b) crop to world coordinates (background removal), (c) extract hands, and (d) check for enclosed regions. Detected regions and hands add to the *voxel space* (cf. Figure 1d).

Our prototype calculates the orientation and volume of both hands by tracking the visible area of the hand over time; the system calculates the roll and pitch angle of each hand from the changes in depth values across the visible area. If the visible area is too small, such as for vertical hand roll (only thumb and index finger are visible top-down), our prototype estimates based on prior observations how much of the hand must be occluded, and determines the hand orientation accordingly. Calculating the yaw angle of the hand is straightforward considering the camera is mounted above the user's head. From the observations over time, our system reconstructs the posture of each hand in 3-space, as well as its precise extent in the *z*-axis (i.e., the axis of line-of-sight of the camera).

Having calculated the orientation of the hands, our prototype then directly translates the position and orientation of the hand into locations of voxels in the *voxel space*. As mentioned before, this comprises activating all voxels in an area that has the same depth, location, and orientation as the user's hand. To account for fast hand motions, the system additionally considers the direction of hand movement and activates voxels between the two positions of the same hand in consecutive camera frames.

Our prototype system detects users' intentions to create finer elements by pinching their fingers and thumb together or moving both hands together. As shown in Figure 6d, as soon as our prototype detects such an enclosed region, it processes this region as opposed to the hand (as in [30]). It samples the depth values for this region from the surrounding area (i.e., the hand). Voxels become active if they share a location with this enclosed region; the prototype dismisses the actual shape of the hand if it encloses a region. This allows users to indicate thinner elements, such as table legs or tripod struts. The same applies if the user connects both thumbs and index fingers, thereby enclosing a bigger area.

Modeling the time-aware scene (*voxel space*)

Our prototype implements the *voxel space* as a three-dimensional array of positive numbers, effectively a 3D histogram. Each voxel has a constant width, height, and depth (10mm in our implementation). We placed the center of the *voxel space* directly in front of the user, roughly at torso level (Figure 6a).

Activating a voxel in the system increases its count in the histogram. This implies that voxels through which the user passes repeatedly or more slowly (i.e., meaningful parts of the object description) will have a higher count than voxels the user passes through when moving the arms to the next, meaningful location. Simple thresholding across all voxels in the space leaves the meaningful and relevant parts.

Matching

We implemented two different techniques to match objects represented in *voxel spaces*. While an interactive system would benefit from real-time matching, our proof-of-concept implementation is not focused on execution speed.

Iterative alignment uses the iterative closest point (ICP) algorithm to register two models [32]. The prototype runs ICP after pre-aligning both models (by translating and rotating to match the principal components). This preparation additionally adapts the scale of both models uniformly.

Brute force tests four levels of quarter-rotation around the *z*-axis (vertical) and any combination of translations within 16cm×16cm×16cm. We neglected rotations around *x* and *y* (horizontal), as users often maintain an object's orientation around those axes, while they tend to "turn" objects towards them when describing (i.e., they rotate about the *z* axis). The number of *z* rotations for this algorithm ideally corresponds to the number of vertical faces in the object, often four. This algorithm also pre-aligns both models and adapts their scale uniformly.

While ICP is computationally expensive and takes around 8s to compare two models in our prototype, brute force takes less than 1s, because it operates in the discrete *voxel space* (looking up voxels is fast). However, ICP is more flexible in that it rotates the objects around all three axes to find the best match.

PROOF-OF-CONCEPT EXPERIMENT

The purpose of this study was to verify the relevance of our observations during the first user study and determine if our prototype implementation is capable of recognizing users' descriptions. In addition, we were also curious to see how our prototype system performs during walk-up use.

Task

As in the first study, participants' task was to describe a 3D object through gestures to the experimenter (Figure 7). Each trial started with a projection of the object onto the wall *behind* the experimenter, such that only the participant could see it. After the participant had studied the object, the projection was turned off and the participant described the object based on their mental representation. The experimenter then tried to guess the object and noted it down.

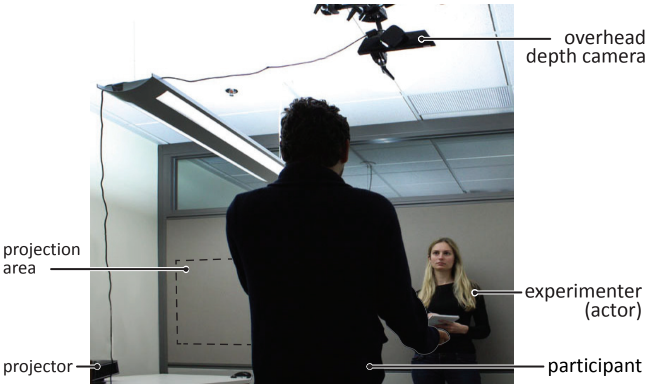

Figure 7: A participant describes an object to the experimenter (here portrayed by an actor). The object had been previously shown to the participant in the projection area, which the experimenter could not see.

In order to encourage participants to provide a sufficiently detailed description of each object, the experimenter never saw the object projected on the wall. This required the participant to perform careful and understandable gestures.

As in the first study, participants were instructed to describe objects using gestures and not complement their gestures with speech. They received no feedback from the system or the experimenter, as this might have impacted how they perform descriptions.

Procedure

Participants described objects from four different categories (Figure 8), each of which contained ten objects. The choice of categories allowed us to separately examine the suitability of our prototype system for simplistic 3D objects, as well as more complex, real-world objects.

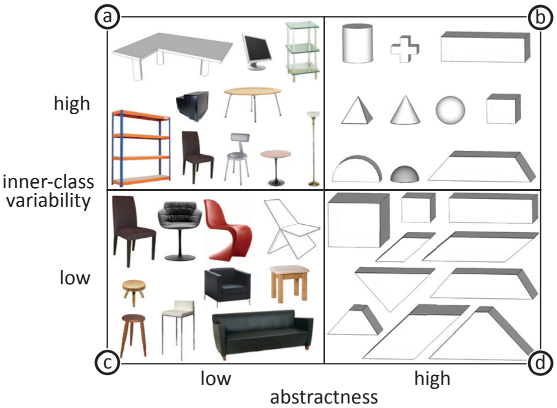

Figure 8: The objects used in the study were split into 2x2 categories: (a) office furniture, (b) primitive solids, (c) chairs, and (d) parallelepipeds.

Before the study, the experimenter presented all categories and objects to the participant, such that the participant could become familiar with the set of objects and learn about the level of detail necessary in the description to uniquely determine the object. (Some objects were in more than one category; depending on the other objects in each category, however, different levels of detail might be necessary to provide an unambiguous description.) The experimenter announced the category at the beginning of each trial and held up a printed sheet illustrating all contained objects to remind the participant.

The study consisted of two sessions: *walk-up* and *instructed*. Participants always began with the walk-up session, in which they described two objects from each category without any instructions. After completing the first session, the experimenter explained to the participant that the system derives faces and legs/struts from the participant's gestures and that single postures did not add to object recognition. The experimenter additionally encouraged the participant to describe each object's elements carefully in the second session. Overall, participants described three objects from each category during the second session.

Participants were instructed to stand on an X on the ground and face the experimenter when gesturing. While this did not preclude participants from gesturing outside the capture volume or occluding their hands, it ensured that the camera would capture most of the gestures.

To avoid priming participants, we made no reference to "miming" during instructions. The absence of feedback for participants during the study ensured that we could test the gestures participants would use when describing objects to a person as opposed to a camera. The use of an experimenter thus precluded participants from falling into an unnaturally detailed and slow demonstration mode.

Categories as well as object selection from the categories were counterbalanced across participants. Overall, each participant completed all 20 trials (4 conditions × (2 objects in the first and 3 objects in the second session)) in less than 20 minutes. Participants filled out a questionnaire on experience with 3D modeling tools after the study.

Apparatus and Participants

Our system captured all interaction with a Microsoft Kinect camera. During the study, the system recorded a video of participants' gestures at 640×480 pixels resolution with depth information at 30Hz. We processed and evaluated all study videos post-hoc with our prototype. The system was running Windows 7 Ultimate, powered by an Intel Core2Duo 2.13 GHz processor and 6GB of RAM. We used an off-the-shelf projector to present objects to participants.

We recruited 15 participants (5 female) from our institution. Participants were between 26 and 47 years old (*M*=33.6, *SD*=7.6) and received a small gratuity for their time.

Results

Outliers and limitations

We experienced two types of outliers during the study. The first relates to occlusion and the limits of the capture volume; while we experienced only five such cases in the walk-up session, because participants leaned into the capture volume with their head, we removed ten more trials from the instructed session (of 300 overall). In some of these cases, participants held one hand up, though inactive, and thereby obstructed the camera view onto their other hand, which was gesturing. In other cases, they moved their hands outside the capture volume, or leaned back also gesturing outside. (We removed no more than two trials from one participant, and never more than one per category.)

The second type of outlier relates to variations in human behavior. In some cases, participants performed gestures too fast to be captured by the camera, finishing a trial in less than two seconds. Two participants, for example, held their fingers and arms together to indicate the cross (Figure 8b). While data miming does not account for recognizing shapes from a single posture, we chose not to remove these "outliers" for the analysis of this study.

We take a two-fold approach in measuring the recognition rate of our prototype. We check if the most-closely matching model from the database is the model the participant had to describe ("*top-match*"), as well as if the described object is among the three most-closely matching models ("*closest-three*"). All numbers compare against chance = 10% for the first approach and 34% for the second.

As shown in Figure 9a, our prototype system retrieved the participant's described object in 60% of all trials (i.e., both sessions and all categories). In 87% of all cases, our prototype retrieved the intended object amongst the three most-closely matching models. We now break out these numbers into the different factors.

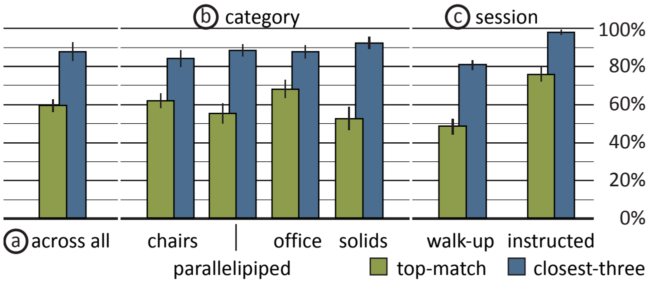

Figure 9: Matching results for participants' performances, showing success rates for *top-match* (green; the matching returned the described object) and *closest-three* (blue; the described object is among the three most-closely matched objects). (a) Accuracy across all trials, (b) broken down to the four categories, and (c) the two sessions. Error bars encode standard error of the mean.

Categories

A two-way ANOVA on recognition rate with participant as the random variable did not find an effect of category on *top-match* performance (p=0.18) or *closest-three* matches (p=0.3). The numbers, however, exhibit interesting patterns (Figure 9b). While the matched object for descriptions of primitive solids was only correct in about half of all cases, this number increased substantially for the *closest-three* approach (92%). The accuracy of matching parallelepipeds exhibits a similar pattern.

Session

A paired-samples t-test comparing accuracy of matches between the walk-up session and the instructed session of our study found a significant increase of correct matches in the instructed part for both *top-match* (t_{14}=4.276, p<0.002) and *closest-three* (t_{14}=6.035, p<0.001). As shown in Figure 9c, our prototype correctly matched participants' descriptions in ¾ of all cases during the instructed part (98% for *closest-three*).

A two-way ANOVA on category and session did not find a significant interaction between the two for *top-match* (p=0.126) nor *closest-three* (p=0.7).

Questionnaire

While four participants stated to have prior experience with 3D modeling tools, none reported to have thought about operations in such tools during the study.

Discussion

The results of the study support our *data miming* approach, which performed particularly well after participants had been given a few instructions (increasing success in 50% more of the cases to obtain the described object as the top match) or when returning several most-closely matching objects. Surprisingly, however, the recognition rate of the primitive solids and parallelepipeds was lower than that of the other two categories when considering the most closely matching object. The considerable difference to *closest-three*, however, suggests that this low performance is due to the similarity of objects in those two categories in general.

Our study also supports findings in the related work. Working solely on their mental image and receiving no feedback, participants' gestures resulted in 3D representations that could mostly be recognized ([1, 10], no-feedback condition in [29]). Participants used fluid and directed gestures to describe object parts, such that straight lines (e.g., shelf, struts) and faces (e.g., seats) looked proper and not jagged [29].

Our results also confirm that participants were able to maintain a spatial reference while performing gestures using both hands [3, 11, 12]. While some object parts were displayed (see video figure), most have only slight offsets (e.g. legs in Figure 1). Participants maintained relative scale and references when gesturing due to their visuospatial memory [1], which also reflects prior insights [10, 11, 12, 23].

We suspect that our study differs from the real world in that users might be more familiar with the object they are describing. On the other hand, our study design matches the case where the user only roughly knows what they want and "sketches" the desired object. A system for such a task would benefit from the *closest-three* approach; a user specifies an object through gestures, upon which the system returns the three most-closely matching objects. The user then picks one, starts over, or continues to gesture more detail, because it has become apparent that they underspecified the object.

Alternatively, the *closest-three* results might be consumed by a larger system that models the context of the interaction, such as the spoken dialogue (e.g., [4]). This extra information might easily disambiguate the user's input and narrow the class of objects under consideration (and, conversely, gesture might disambiguate other aspects of the interaction, such as speech). For example, the user could say "chair" and then specify a particular chair by gesturing to indicate a uniquely identifying feature (or set of features) of the chair's shape. Interactive systems can thus use data miming for input as users describe objects focusing on their mental image.

CONCLUSIONS AND FUTURE WORK

We have presented *data miming* as an approach to inferring spatial objects from the user's gestures when describing physical objects. Our approach results from observations in a user study, in which participants described real-world objects to another person using gestures. We implemented a proof-of-concept prototype system, which passively observes the user's gestures with a depth camera and models the trace of the user's hands with a discrete and time-aware voxel representation. Data miming proved useful in a final study and recognized participants' iconic or pantomime gestures while describing physical objects. That our implementation performs as well as it does suggests that people do carry 3D mental images of objects, and that they readily replicate this 3D imagery in gesture with surprising fidelity.

Although promising, our results show that next versions of data miming need to incorporate multiple cameras to mitigate occlusion, and also recognize compound hand postures. While easy to detect for humans, future prototypes need to recognize distorted and out-of-shape objects. While the findings reported in this paper are limited to the tested objects, we believe they extend to others that consist of struts and faces.

Of course, linguists have observed many gesture strategies that do not follow a 3D model. For example, a participant might have indicated the red 'S'-shaped chair (Figure 8c) with a "wavy" motion of the hand. While the form of this gesture might not match the actual 3D shape of the chair's back, it instead conveys an abstract quality of the shape. A more feature-based or machine-learning approach driven by training examples might capture some non-literal aspects of gesture behavior. Extending data miming to handle such metaphoric gestures is an interesting avenue of future work.

ACKNOWLEDGMENTS

We thank Hrvoje Benko, Ken Hinckley, and Johnny Lee for valuable feedback and discussions, as well as all participants of our studies for their time.

REFERENCES

1. Baddeley, A.D., and Hitch, G. 1974. Working memory. In G.H. Bower (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory 8*, 47-89.
2. Baudel, T. and Beaudouin-Lafon, M. Charade: remote control of objects using free-hand gestures. *CACM* 36, 7(1993), 28-35.
3. Billinghurst M., Bowskill J., Dyer N., and Morphett, J. Spatial information displays on a wearable Computer. *IEEE CG&A* 18, (1998), 24-30.
4. Bohus, D., Horvitz, E. 2009. Dialog in the Open World: Platform and Applications. *Proc. ICMI '09*, 31-38.
5. Deering, M. F. HoloSketch: a virtual reality sketching/animation tool. *ACM Trans. Comput.-Hum. Interact.* 2, 3(1995), 220-238.
6. Funkhouser, T., Kazhdan, M., Shilane, P., Min, P., Kiefer, W., Tal, A., Rusinkiewicz, S., and Dobkin, D. Modeling by example. *ACM Trans. Graph.* 23, 3(2004), 652-663.
7. Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D., and Jacobs, D. A search engine for 3D models. *ACM Trans. Graph.* 22, 1(2003), 83-105.
8. Galyean, T.A., Hughes, J.F. Sculpting: an interactive volumetric modeling technique. *SIGGRAPH Comput. Graph.* 25, 4(2004), 267-274.
9. Grice, P. Studies in the Way of Words. Harvard University Press: Cambridge. 1989.
10. Gustafson, S., Bierwirth, D., and Baudisch, P. Imaginary Interfaces: Spatial Interaction with Empty Hands and Without Visual Feedback. *Proc. UIST'10.*
11. Hinckley, K., Pausch, R., and Proffitt, D. Attention and visual feedback: the bimanual frame of reference. *Proc. I3D '97*, 121-ff.
12. Hinckley, K., Pausch, R., Goble, J. C., and Kassell, N. F. A survey of design issues in spatial input. *Proc. UIST '94*, 213-222.
13. Holz, C. and Feiner, S. Relaxed Selection Techniques for Querying Time-Series Graphs. *Proc. UIST '09*, 213-222.
14. Keefe, D. F., Feliz, D. A., Moscovich, T., Laidlaw, D. H., and LaViola, J. J. CavePainting: a fully immersive 3D artistic medium and interactive experience. *Proc. I3D '01*, 85-93.
15. Kendon, A. How gesture can become like words. In F. Payatos (ed.), *Cross-Cultural Perspectives in Nonverbal Communication,* 131-141. Toronto: Hogrefe.
16. Lee, J. and Kunii, T. L. Model-Based Analysis of Hand Posture. *IEEE Comput. Graph. Appl.* 15, 5(1995), 77-86.
17. Llamas, I., Kim, B., Gargus, J., Rossignac, J., and Shaw, C. D. Twister: a space-warp operator for the two-handed editing of 3D shapes. *ACM Trans. Graph* 22, 3(2003), 663-668.
18. McNeill, D. Hand and Mind: What Gesture Reveal about Thought. Chicago: University of Chicago Press. 1992.
19. Nishino, H., Utsumiya, K., and Korida, K. 3D object modeling using spatial and pictographic gestures. *Proc. VRST '98*, 51-58.
20. oblong g-speak. http://www.oblong.com
21. Pavlovic, V.I., Sharma, R., Huang, T.S. Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Trans. Pattern Analysis and Machine Intelligence* 19, 7(1997), 677-695.
22. Pratini, E. New Approaches to 3D Gestural Modeling – the 3D SketchMaker Project. *Proc. eCAADe '01*, 466-471.
23. Sachs, E., Roberts, A., and Stoops, D. 3-Draw: A Tool for Designing 3D Shapes. *IEEE Comput. Graph* 11, 6(1991), 18-26.
24. Schkolne, S., Pruett, M., and Schröder, P. Surface drawing: creating organic 3D shapes with the hand and tangible tools. *Proc. CHI '01*, 261-268.
25. Starner, T., Auxier, J., Ashbrook, D., and Gandy, M. The Gesture Pendant: A Self-illuminating, Wearable, Infrared Computer Vision System for Home Automation Control and Medical Monitoring. *Proc. ISWC '00*, 87-94.
26. Starner, T., Leibe, B., Minnen, D., Westeyn, T.L., Hurst, A., Weeks, J. The perceptive workbench: Computer-vision-based gesture tracking, object tracking, and 3D reconstruction for augmented desks. *Machine Vision and Appl.* 14, 1(2003), 59-71.
27. Tangelder, J. W. and Veltkamp, R. C. A survey of content based 3D shape retrieval methods. *Multimedia Tools Appl.* 39, 3(2008), 441-471.
28. Wang, S. W. and Kaufman, A. E. Volume sculpting. *Proc. I3D '95*, 151-156.
29. Willis, K. D., Lin, J., Mitani, J., and Igarashi, T. Spatial sketch: bridging between movement & fabrication. *Proc. TEI '10*, 5-12.
30. Wilson, A.D. Robust computer vision-based detection of pinching for one and two-handed gesture input. *Proc. UIST '06*, 255-258.
31. Xbox Kinect. http://www.xbox.com/kinect.
32. Zhang, Z. Iterative point matching for registration of free-form curves and surfaces. *Int. J. Comput. Vision* 13, 2(1994), 119-152.