Working with Robots and Objects: Revisiting Deictic Reference for Achieving Spatial Common Ground

Andrew G. Brooks
Robotic Life Group
MIT Media Laboratory
77 Massachusetts Avenue
Cambridge, MA, USA
zoz@media.mit.edu

Cynthia Breazeal
Robotic Life Group
MIT Media Laboratory
77 Massachusetts Avenue
Cambridge, MA, USA
cynthiab@media.mit.edu

ABSTRACT

Robust joint visual attention is necessary for achieving a common frame of reference between humans and robots interacting multimodally in order to work together on real-world spatial tasks involving objects. We make a comprehensive examination of one component of this process that is often otherwise implemented in an ad hoc fashion: the ability to correctly determine the object referent from deictic reference including pointing gestures and speech. We develop a modular spatial reasoning framework based around decomposition and resynthesis of speech and gesture into a language of pointing and object labeling that supports multimodal and unimodal access in both real-world and mixed-reality workspaces, accounts for the need to discriminate and sequence identical and proximate objects, assists in overcoming inherent precision limitations in deictic gesture, and assists in the extraction of those gestures. We further discuss an implementation of the framework that has been deployed on two humanoid robot platforms to date.

1. INTRODUCTION

Deictic gesture is an important non-verbal component of multimodal interaction between humans. Identified early on as a primary candidate metaphor to be transferred to human-computer interaction, deixis — in particular the "concrete" or "specific deictic", using pointing to refer to a specific object or function — is now a staple of interfaces involving 2-D spatial metaphors. In the field of human-robot interaction, deictic object reference is also an attractive communications mode. Ideal scenarios of humans and robots working together with multiple objects in the 3-D real world, however, present a more complex and unstructured problem. We have therefore revisited deictic object reference for achieving spatial common ground between the human and the robot in a scalable and robust fashion during such collaborative activities.

Human Robot Interaction '06 Salt Lake City, Utah, USA

In humans, deictic spatial gesture is a component of joint visual attention, a behavior that is theorized to be one of the developmental precursors to "theory of mind" [34, 25, 33]. Furthermore, human-human interaction studies have showed that over 50% of observed spontaneous gestures to be concrete deictics [9], and that concrete deictic gesture can effectively substitute for location descriptions in establishing joint attention [19]. We wish to support this process as part of our broader work in providing robots with the means to more deeply understand human activities, goals and intentions. The direct goal of the research reported here is to allow the human to communicate with the robot about specific objects in space — to always be "on the same page" about which object is being attended to — using the natural gestural and speech capabilities adult humans ordinarily possess: pointing and naming.

Humans' use of deictic gesture has two important characteristics which carry over to human-robot interaction. First, it is only used in cases of what Clark and Marshall refer to as "physical copresence" [5] — both participants must be able to view the referent in the situation in which the gesture occurs. The robot's pre-existing knowledge of the spatial state of the world thus should contextualize the gesture; in fact, the nature of human pointing behavior makes this a necessity. The precise point indicated by the human (the *demonstratum*) is in most cases spatially distinct from the object the human intends to indicate (the *referent*) [6] due to various factors including simple geometric error on the part of the human (e.g. parallax), the human's desire not to allow the gesture itself to occlude the robot's view, and the fact that pointing gestures in 3-D contain no inherent distance argument. Instead, the demonstratum can typically only constrain the set of potential referents. This argues for a "spatial database" approach in which deictic gestures are treated as parameterized spatial queries.

Second, like most gesture, deictic gesture is closely correlated with speech (90% of gesture occurs in conjunction with speech [22]). Natural language itself contains deictic expressions that can be disambiguated with the help of gesture, and similarly deictic gestures are usually resolved in the presence of their accompanying spoken context. Pointing gestures unaccompanied by contextualizing utterances are rare, and depend on other narrow constraints to allow them to be understood, such as specific hand configurations

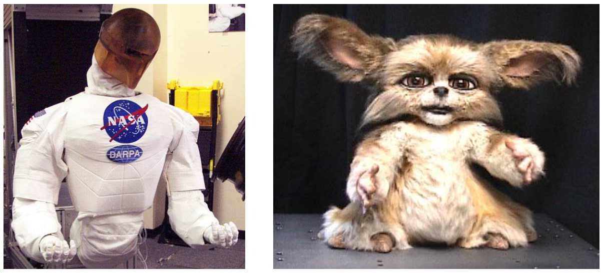


Figure 1: Humanoid robot platforms for which this deictic reference system was designed. Left: Robonaut, designed for autonomous teamwork with human astronauts in space. Right: Leonardo, designed for research into human-robot interaction.

(index finger outstretched) along with situational context (e.g. to distinguish a pointing up gesture from a symbolic reference to the numeral 1). In the typical case the robot should be able to use the accompanying speech both to assist in the spatio-temporal isolation of the gesture itself, and to constrain the demonstratum to the referent in cluttered or hierarchical situations.

The research described in this paper therefore contributes a framework for multimodally determining object referents from a combination of deictic gesture, speech, and spatial knowledge, in support of joint attention during physically copresent human-robot collaboration. The framework is designed to reflect and accommodate natural human deictic behavior in terms of precision and timing, and unlike other approaches to date is designed to support mutual disambiguation when referring to multiple physically identical objects. Rather than a more typical selection-based metaphor, a metadata approach of object referent labeling is chosen, to support subsequent reference in both multimodal and unimodal fashions.

An implementation of the framework is also presented, in terms of a modular system designed to interact with an untethered 3-D model-based visual tracking system, a speech recognition engine and a hierarchical spatial database. The implementation is further extended to support interactions with real robots in mixed-reality workspaces containing virtual objects. Results from usage of the implementation — in conjunction with different humanoid robot platforms (Figure 1, vision systems, speech recognition systems and spatial databases — are discussed.

2. RELATED WORK

The "Put-That-There" system of Bolt in 1980 essentially set the standard for copresent deictic gestural object reference in human-machine interaction [2]. Conceptually, the task of this system is similar to that of the human-robot collaboration task: to refer to objects by pointing and speaking. This system used a Polhemus tracker to monitor the human's arms, and the spatial data to be managed consisted of simple geometric shapes in a 2-D virtual space. Most of the subsequent advances within this task domain concern improvements to the underlying components such as tracking and speech recognition.

For a comprehensive summary of work on the visual interpretation of hand gestures, including deictic gestures, see [29]. To overcome technological limitations in natural gesture analysis, investigations into multimodal interfaces involving deixis were often restricted to constrained domains such as 2-D spaces and pen-based input (e.g. [24]). The deictic components of these efforts can be summarized as enabling this type of gestural reference in some form as part of the interface, rather than tackling problems in determining the object referent from the gesture.

Several research efforts chose to concentrate on the object referent primarily in order to use it as contextual information to assist in recognition processes. Kuniyoshi and Inoue used the object context to aid action recognition in a blocks world [17]. Moore et al. related object presence to the trajectory of the hand in order to provide action-based object recognition [26]. Strobel et al. used the spatial context of the environment (e.g. what object lies along the axis of the hand) to disambiguate the type of hand gesture being performed in order to command a domestic service robot [35]. Similarly, our first deictic reference system developed to support human tutelage of our humanoid robot Leonardo used the presence of a visually identified button object to confirm a static pointing gesture after being suggested by the hand tracking system [3]. Nagai reports using object information as a feedback signal in a system for teaching a robot to comprehend the relation between deictic gesture and attentional shift [27]. In contrast, our current work is primarily concerned with robustly connecting the demonstratum with the desired referent.

The majority of work in determining the object referent from deixis has been directed towards using gestural information to disambiguate natural language use. Kobsa et al. used gestural information in the form of a mouse pointer in a 2-D space to resolve deictic phrases [15], and more recently Huls et al. used similar mouse-based pointing as part of a system to automatically resolve deictic and anaphoric expressions [12]. Koons et al. resolved deictic expressions using visual attention in the form of both pointing and gaze direction [16]. Similar efforts are beginning to appear in the human-robot interaction literature; for example, Hanafiah et al. use onboard gaze and hand tracking to assist a robot in disambiguating inexplicit utterances, though only results for the gaze component are reported [10]. In contrast to our work, these systems concentrate on disambiguating speech as it occurs, rather than augmenting the shared spatial state with data for future object reference, such as object names.

Research concentrating primarily on determining and managing object referents from pointing is less common. The ability to use deictic gesture over a range of distances is an attractive feature in virtual environments (VEs), and Latoschik and Wachsmuth report a system that classifies pointing gestures into direction vectors that can be used to select objects, but do not tackle object discrimination [18]. A later VE system reported by Pfeiffer and Latoschik is more closely aligned with our efforts, but as above focuses more on the disambiguation of speech than gesture, and resolves multiple object reference with relative speech references rather than further gesture [32]. Hofemann et al. recently report a system dedicated to simultaneously disambiguating pointing

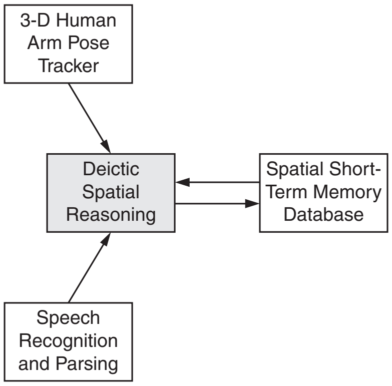

3-D Human Arm Pose Tracker
Deictic Spatial Reasoning
Spatial Short-Term Memory Database
Speech Recognition and Parsing

Figure 2: The functional units of the object reference system and their data interconnections.

gestures and determining the object referent by combining hand trajectory information with the presence of objects in a "context area", but does not deal with discrimination within homogeneous or heterogeneous collections of multiple objects. In contrast, our system is specifically designed for working with multiple, potentially visually identical objects arranged within the margin of error of human pointing and visual tracking.

3. SYSTEM DESCRIPTION

Our framework for deictic object reference is a distributed, modular approach based on largely independent functional units that communicate by message passing. This enables most modules to be simultaneously available for other tasks in addition to object reference. Individual modules may be executed on different computers for performance reasons. The underlying metaphor is one of using a "deictic grammar" to assemble some combination of gestures and object names and relationships into queries to the robot's spatial database.

3.1 Functional Units

The four functional units of the framework are a vision-based human tracking system for pointing gesture extraction, a grammar-based speech understanding system, a spatial database, and the main deictic spatial reasoning system. The units and their data interconnections are shown in Figure 2. The first three components are treated in a generic fashion; they can and have been represented by different implementations that just need to satisfy the essential data requirements.

The human tracking system is a 3-D model-based tracker capable of real-time extraction and reporting of the human's arm or arms. While desirable, it is not strictly necessary for the tracker to also detect pointing "gestures" by incorporating the configuration of the hand into the model. Distinguishing pointing from non-pointing is bootstrapped with the aid of the speech module. We prefer to use an untethered vision-based tracker to maintain as much as possible the naturalness of the interaction, but this is also not a requirement. The local coordinate system of the tracker is converted into the robot's egocentric coordinate system.

The speech understanding system incorporates speech recognition and parsing according to a predefined grammar that is able to be tagged at points relating to deixis, such as direct references to objects (names and deictic expressions such as "this" and "that") and instructions indicating an object context (such as naming something). The system must thus provide access to the tags activated for a particular parse.

The spatial database primarily stores the names, locations and orientations of objects in space. It is independently updated by the robot's own perceptual and mnemonic systems. Additional levels of sophistication such as more descriptive metadata (object types and hierarchies of compound objects) and the ability to query the database by spatial partition are helpful but not required. At present we encode a basic type of an object as a prefix of its name.

The core of the system is the deictic spatial reasoning module. This unit continuously monitors the output of the human tracker, and uses the tag signals from the speech parsing to extract pointing gestures and assemble queries for potential object referents from the spatial database. Candidates returned from the spatial database are matched and confirmed by this unit, and updated result information, along with the pointing gestures themselves, are posted back to the spatial database for subsequent access by the robot's attentional mechanisms.

3.2 Gesture Extraction

Hand gestures consistently adhere to a temporal structure, or "gesture phrase" [14], comprising three phases: preparation, nucleus (peak or stroke [23]), and retraction. Visual recognition of deictic gestures is limited to the nucleus phase, by identifying the characteristic hand configuration (one finger outstretched) or making pose inference (arm outstretched and immobile). However these characteristics are frequently attenuated or absent in natural pointing actions. Conversely, as discussed earlier deictic gestures are almost always accompanied by associated utterances. We therefore choose to isolate pointing gestures temporally rather than visually.

Speech is synchronized closely with the gesture phrase, but the spoken deictic element does not directly overlap the deictic gesture's nucleus in up to 75of both spoken and gesture phrases, the overlap is substantial and predictable. Marslen-Wilson et al. observed that pointing gestures occured simultaneously with the demonstrative in the noun phrase of the utterance, or with the head of the noun phrase if no demonstrative was included, and that no deictic gestures occurred after completion of the noun phrase [21]. Kaur et al. similarly showed that during visual attention shifts, the speaker's gaze direction began to shift towards the object referent before the commencement of speech [13]. These results were matched by our own informal observations, in which subjects frequently commenced pointing prior to both the deictic and noun components of their utterances, during human subject experiments using our previous pointing system [4].

To extract pointing gestures temporally, we therefore consider "dynamic pointing", incorporating the preparation phase as well as the nucleus, in addition to the more traditional "static pointing". This also allows us to successfully recover deictic gestures in which the nucleus is not static, i.e. motioning towards an object. Because we can be confident that the pointing gesture occurred during or shortly before the user's speech, the spatial reasoning system keeps a history buffer of the arm tracking data. When a deictic phrase component is received from the speech understanding system, this buffer is analyzed to extract the best estimate of the gesture type and location. For implementation details, please see Section 4.

3.3 Object Reference

First among the ten myths of multimodal interaction identified by Oviatt is that users of a multimodal system will always tend to interact multimodally [28]. In fact, users frequently prefer unimodal interaction, particularly in the case of object reference — there should be no need for the human to point to an object each time the robot's attention is to be drawn to it. We therefore treat the traditional "point-to-select" usage as a special case of a more general "point-to-label" metaphor. Once an object has been labeled for the robot, the human can refer to it unimodally by name, as well as making partial reference by name to constrain potential referents of future deictic gestures.

Object labels are also used to allow ordering of multiple objects for tasks in which the robot must attend to them in a particular sequence. Many collaborative tasks that might be performed with robotic assistance, such as assembly of a complex object from components, require attention to otherwise identical objects in a specific order, yet this has not been widely explored in the design of tools for human-robot interaction. By incorporating a sequence number into each object label, the robot can generate appropriate behavior from future unimodal references (e.g. "first", "next", "last").

As already discussed, object reference from deictic gesture must content with a number of ambiguities. For example, the *pars-pro-toto deictic*, in which a superordinate object is indicated by pointing to an individual subcomponent, and the opposite case, the *totum-pro-parte deictic*, in which the superordinate object is used to refer attention to a specific subcomponent. Moreover, 3-D pointing gestures have no inherent distance constraint, so particular spatial layouts can present perspective ambiguities that are difficult to resolve geometrically.

We tackle these problems by allowing nested object references in which the user can first deictically indicate a parent object to constrain further references until the constraint is removed. When such a hierarchical object reference is set up, the system restricts potential matches to objects physically or conceptually dependent on the constraining object. This can occur both at the spatial database level, in terms of hierarchically constructed compound objects, and at the spatial reasoning level, in which a plane defined by the horizontal axes of the parent object is used to constrain the distance implied by the pointing gesture vector.

As mentioned above, pointing gestures themselves are also treated as virtual object references and thus can be posted to the spatial database as objects that are referred to by themselves. This provides a seamless method of providing access to pointing gestures to other attentional mechanisms (which may not be interested in object reference directly) and supports potential future extensions to reasoning about the spatial behavior of the human.

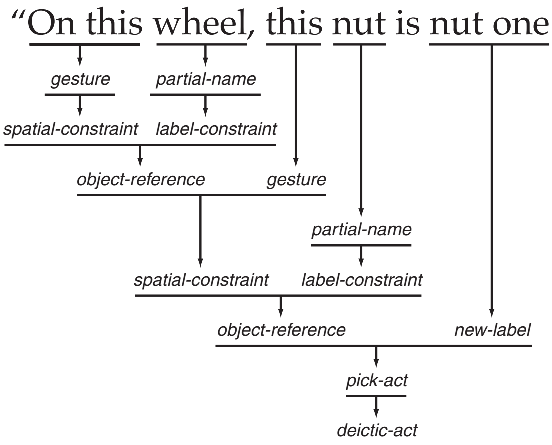


Figure 3: Parsing a sentence containing a nested object constraint and multiple deictic gestures into a single pick action intended to focus the robot's attention on an object and label it for potential unimodal reference and sequencing.

3.4 Deictic Grammar

Deictic expressions are a grammatical component of natural language, but non-verbal communication or "body language" is not strictly a language — it does not have discrete rules and grammars, but does convey coded messages that humans can interpret [20]. Our framework is based on synthesis of the spoken and gestural elements of deictic reference into a "deictic grammar", a formal language for physically and verbally referring to objects in space. Complex speech and movement are decomposed into simpler features that are sent to the deictic spatial reasoning system to be parsed according to the following ruleset, where the vertical bar '|' represents either-or selection and square brackets '[]' represent optionality:

deictic-act ⟶ **pick-act** [**go-act**] | **place-act** | **delete-act**

pick-act ⟶ **object-reference** [**new-label**] [**pick-act**]

place-act ⟶ **pick-act object-reference**

delete-act ⟶ **object-reference**

object-reference ⟶ [**spatial-constraint**] [**label-constraint**]

spatial-constraint ⟶ [**gesture**] [**object-reference**]

label-constraint ⟶ [**full-name** | **partial-name**]

In this deictic grammar, **pick-act** indicates an attention directing command with associated referent relabeling, **place-act** indicates a command to move one or more objects, **delete-act** indicates a command to remove an object from the spatial database. The **pick-act** term has been designed recursively in order to support multiple simultaneous referent resolution. In cases of multiple resolution, gestures are held in a queue and only matched against results from the

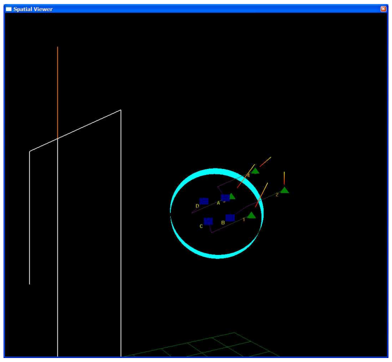


Figure 4: Example screenshot from the 3-D visualizer showing four gestures being simultaneously matched to four objects constrained by a parent object, including the actual pointing error in each case.

spatial database when the system is best equipped to do so. A special end-of-sequence marker, **go-act**, is therefore necessary to instruct the system to drain the queue of deictic references in these cases.

The **spatial-constraint** term is also recursive via **object-reference**, in order to support nested object reference as described in Section 3.3. See Figure 3 for an example parse of a typical compound deictic act, showing how a combination of multiple gestures and object references can be used to relabel a single specific object. The deictic grammar imposes some restrictions on the spoken grammar that can be used to drive it. This is acceptable for our application because the speech recognition engine itself also requires a pre-written grammar, so it just requires this framework to be taken into account when designing the space of utterances the robot can understand.

The top level commands were chosen in order to primarily support object reference in the real world, with extensions to support mixed-reality workspaces (i.e. where objects can be moved by the spatial database directly, and where object deletion makes sense). We chose not to specifically incorporate object creation, as this system is designed for the spatial reference of objects that the robot knows to be physically present. None of the current module interconnections specifically deals with objects the robot knows about but which are not present, so the spatial reasoning system can not post arbitrary objects to the spatial database. This functionality could be added by having another module read posted gestures from the spatial database and comminicate with the speech system directly.

4. IMPLEMENTATION DETAILS

The spatial reasoning system is written in C++ and runs on an IBM workstation running Windows 2000. It incorporates a 3-D visualizer of objects and pointing gestures, written using OpenGL (Figure 4).

The functionality of the other modules has been provided by a number of other systems. During our involvement with the Robonaut project, visual tracking of the human was performed by the system developed by Huber [11]. Speech recognition was accomplished using ViaVoice and the natural language understanding system Nautilus developed at the Naval Research laboratories [30]. The spatial database used was the Sensory Egosphere (SES) developed at Vanderbilt University [31]. Intermodule communication used Network Data Delivery System by Real-Time Innovations, Inc.

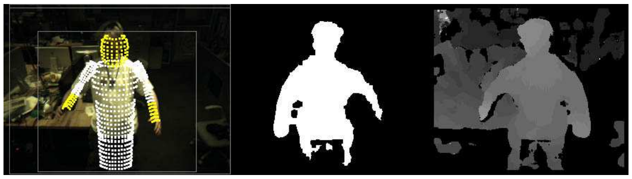


Figure 5: The robot currently has untethered real-time perception of both human upper arms and forearms, using the VTracker articulated body tracker.

Currently, for our humanoid robot Leonardo, we have replaced each of these modules. Human body tracking is performed by the VTracker articulated body tracker, developed and kindly provided for our use by the Vision Interfaces Group at the MIT Computer Science and Artificial Intelligence Laboratory, which uses an interval constraint propagation algorithm on stereo range data to provide a real-time position estimate of the human's torso and left and right shoulders, elbows and wrists [8] (Figure 5). Speech recognition and parsing is performed by CMU Sphinx 4, using grammars developed in our laboratory [7]. We have developed our own basic implementations for the spatial database and intermodule communications, though compatibility with the SES has been retained.

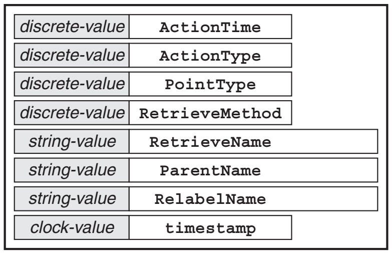
4.1 Gesture Extraction

Extraction of static nucleus deictic gestures is straightforward but varies with the visual tracking system. The Robonaut vision system tracks only the human forearm, so the principal axis of the forearm from elbow to wrist is used as the pointing vector. The VTracker system tracks both forearm and upper arm, so we use the vector between the shoulder and the wrist, which actually provides a better estimate of the distant pointing location than the forearm alone for typical human pointing gestures.

To extract the preparation phase of deictic gestures, or those with a dynamic nucleus, the time series of the position of the end effector is used instead. Principal component analysis is performed on a moving buffer of position values prior to the speech trigger. The first principal component is used as the orientation vector of the gesture. This vector is then linearly best fit to the collection of position values to determine its translational position and direction. This method has a lower accuracy than the static case, but is typically used in situations when the problem is well constrained by the spatial arrangement of the objects or the hierarchical context, as users tend to point more carefully when the situation is ambiguous.

4.2 Object Matching

The spatial reasoning system performs geometric matching between extracted deictic gestures and lists of candidate ob-



discrete-value	`ActionTime`
discrete-value	`ActionType`
discrete-value	`PointType`
discrete-value	`RetrieveMethod`
string-value	`RetrieveName`
string-value	`ParentName`
string-value	`RelabelName`
clock-value	`timestamp`

Figure 6: Contents of the speech message data structure sent to the spatial reasoning system to control and coordinate its activities.

jects returned by the spatial database. The spatial database stores an object by its name and six floating point numbers representing its position and orientation in the robot's coordinate frame of reference. Currently the spatial database returns all objects satisfying a query by full or partial object name, although we have designed for future support of query by object hierarchy or by spatial partition based on the set of gestures involved in the particular circumstance.

The returned set of candidate objects is compared against the set of gestures using a linear best fit. Presently the size of the gesture queue is limited to 12 gestures, which allows an exhaustive search of the matching space to be performed to avoid encountering local minima. Clearly the accuracy of the system increases as the number of simultaneous matches is performed, but we do not envisage any tasks for our robot that require simultaneous sequencing of more than this number of identical objects.

Due to limitations in the spatial databases used so far, which do not store extent information, objects are currently treated as point targets and gestures as vectors. When a parent object is used as a spatial constraint, the horizontal axes of the parent object, rotated to match the pose of the object, are extrapolated to an intersection plane that reduces the gesture vectors to points from which linear distances can be calculated. When no parent object has been defined, the orthogonal distance from the candidate objects to the gesture vectors is used.

4.3 Deictic Grammar Parsing

The deictic grammar given in Section 3.4 represents a system model, of which our implementation is one realization. In our implementation the input to the grammar parser consists of a sequence of activated speech tags produced by the speech recognition system as a result of its parse of the human's speech. The structure of the tag is shown in Figure 6.

There are four discrete-valued fields. The *ActionTime* field indicates whether the tag represents a tag to be enqueued or executed immediately. The *ActionType* field defines the action context of the tag (and its corresponding gesture), such as a deictic pick or place action. The *PointType* field can be used to force the gesture extraction mechanism to limit itself to a static or dynamic interpretation of the arm state, if desired. The *RetrieveMethod* field informs the spatial reasoning system which parameters to use in its retrieval of object candidates from the spatial database. The string fields *RetrieveName*, *ParentName* and *RelabelName* contain the values for spatial database retrieval and object referent reposting, if applicable. Finally, a timestamp assists in maintaining synchronizations between the various modules, which do not necessarily share a common clock.

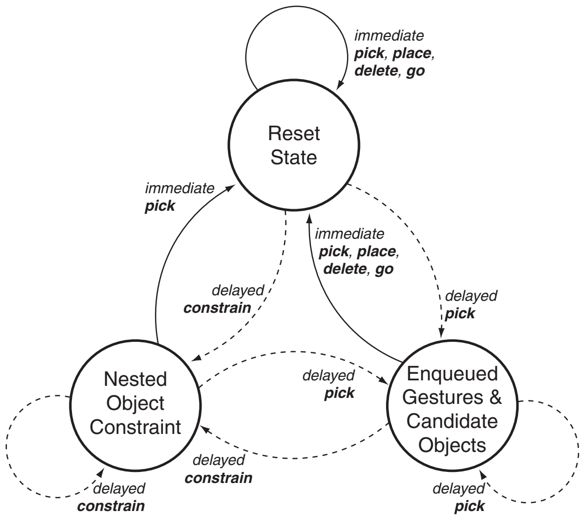
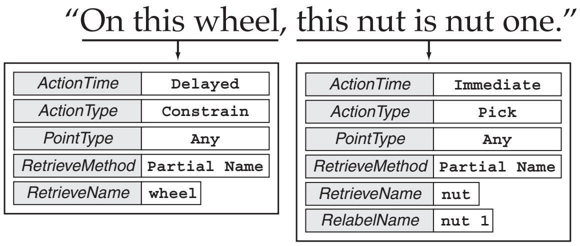


Figure 7: General state model of the spatial reasoning system. Edge transitions are caused by incoming speech messages, and may have communications with the spatial database associated with them. Solid edges represent transitions in which new object information may be posted to the spatial database (e.g. object relabeling or coordinate transformations). Dashed edges represent transitions in which object information may be read but not posted.

These incoming speech tags are processed by using their contents to trigger changes in the state of the spatial reasoning system that simulate the behavior of the deictic grammar model. A state diagram of the reasoning system is shown in Figure 7, and an example tag sequence corresponding to the example parse of Figure 3 is shown in Figure 8. The tag structure is easily extended to more complex grammars and modules (e.g. affirmation/negation of pointing results, more complex database lookups) with corresponding alterations to the state model. We presently assume properly formed input from the speech system and simply discard tag sequences resulting in error conditions.

5. RESULTS

The initial version of this system was used as part of NASA's humanoid robot Robonaut (Figure 1 Left) for its Demo 2 for the DARPA MARS program, a report on which can be found in [1]. In this task, the human instructed the robot to tighten the four nuts on an ordinary automobile wheel in a specific sequence by using deictic gesture and speech to constrain the robot's attention to the wheel object and then similarly pointing and labeling the nuts in order (Figure 9). The discrimination distance between the nuts was roughly on the order of the size of the nuts themselves (approximately 3cm).



ActionTime	Delayed
ActionType	Constrain
PointType	Any
RetrieveMethod	Partial Name
RetrieveName	wheel

ActionTime	Immediate
ActionType	Pick
PointType	Any
RetrieveMethod	Partial Name
RetrieveName	nut
RelabelName	nut 1

Figure 8: The example utterance as parsed into speech tag messages, each of which triggers an object reference incorporating a deictic gesture. The first determines the nested object constraint and the second matches the ultimate object referent and attaches its new label.

Using multiple simultaneous gesture-object matching, the deictic spatial reference system was able to robustly resolve the object referents correctly in the presence of tracking and human pointing error. Using the additional constraint that in the event that the spatial database returned the same number of candidate object referents as gestures the system was to assume a unique one-to-one match, the system was also able to detect single duplicate labeling errors.

The current version of the system, extended for mixed-reality workspace support, is intended for use in collaborative tasks with our humanoid robot Leonardo (Figure 1 Right), replacing a less sophisticated earlier deictic object reference system that operated in two spatial dimensions. We expect this system to support significantly more complex natural human-robot interactions involving objects in the world.

6. CONCLUSIONS & FUTURE WORK

By formally incorporating deictic gestures into the language of object reference, the framework we have developed provides tighter integration of natural multimodal interaction between humans and robots. The framework further supports natural human behavior by explicitly concentrating on persistent object disambiguation for unimodal speech reference through the use of user-defined labels. The deictic grammar construct with support for nested object reference and multiple simultaneous object referent resolution has been demonstrated to be robust to the imprecision of contemporary motion tracking and human pointing actions themselves.

The implementation of the spatial reasoning module that we have developed is a comprehensive and flexible system for achieving human-robot joint attention with real and virtual objects. It has been deployed as part of the computational resources of two different complex humanoid robots, each of which has different implementations of the various other subcomponents such as vision and speech, and has successfully been used as part of a demonstration of a real-world cooperative assembly task.

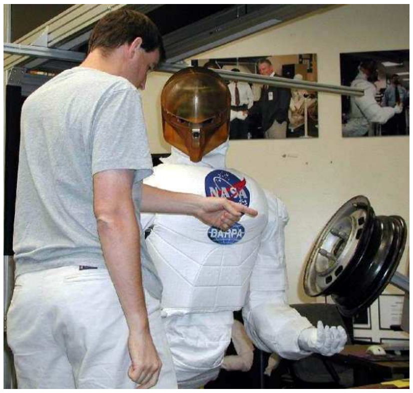


Figure 9: Execution of the pointing and labeling component of the Robonaut nut-tightening task. The four nuts are indicated by the human in order to label them with their sequence numbers, and the gestures then matched to the correct object referents simultaneously.

As this system currently meets our basic requirements, we do not envisage significant future work on the core elements of the deictic grammar and spatial reasoning process. The principal element in need of increased sophistication is the spatial database implementation. Ultimately it would be useful for the robot’s spatial short term memory to be arranged more along the lines of a virtual environment scene graph, incorporating details such as object shape and volumetric extents to allow more accurate gesture matching and proper geometric querying. It might also be useful to merge our current work on social referencing into the spatial database, providing the potential for emotional parameterizations to spatial queries (e.g., pointing at a collection of objects and asking for one’s favorite).

7. ACKNOWLEDGEMENTS

The authors are grateful for the past and ongoing collaborations with the following individuals and groups. Dr David Demirdjian and Dr Trevor Darrell of the Vision Interfaces Group at the MIT Computer Science and Artificial Intelligence Laboratory for the VTracker software. Dr Robert Ambrose, Dr Bill Bluethmann and the Dextrous Robotics Laboratory at NASA Johnson Space Center for our opportunity to be involved in the Robonaut project. Stan Winston Studio provided the physical Leonardo robot. Leonardo’s cognitive architecture is built on top of the C5M code base of Dr Bruce Blumberg’s Synthetic Characters Group at the MIT Media Lab. This work was funded in part by the *Digital Life* and *Things That Think* consortia and the DARPA MARS program.

8. REFERENCES

[1] Bluethmann, W., Ambrose, R., Diftler, M., Huber, E., Fagg, A., Rosenstein, M., Platt, R., Grupen, R., Breazeal, C., Brooks, A.G., Lockerd, A., Peters, R.A. II, Jenkins, O.C., Matarić, M., and Bugajska, M. Building an autonomous humanoid tool user. In *Proc. IEEE-RAS/RSJ Int’l Conf. on Humanoid Robots (Humanoids ’04)*, Los Angeles, California, November 2004.

[2] Bolt, R.A. Put-That-There: Voice and gesture at the graphics interface. *ACM Computer Graphics*, 14(3):262–270, 1980.

[3] Breazeal, C., Brooks, A.G., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., and Chilongo, D. Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robots*, 1(2):315–348, 2004.

[4] Breazeal, C., Kidd, C.D., Lockerd Thomaz, A., Hoffman, G., and Berlin, M. Effects of nonverbal communication on

efficiency and robustness in human-robot teamwork. In *Proc. International Conference on Intelligent Robots and Systems*, 2004.

[5] Clark, H.H. and Marshall, C.R. Definite reference and mutual knowledge. In Joshi, A.K., Webber, B.L., and Sag, I.A., editors, *Elements of Discourse Understanding*. Cambridge University Press, Cambridge, 1981.

[6] Clark, H.H., Schreuder, R., and Buttrick, S. Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22:245–258, 1983.

[7] CMU Sphinx Group. Open Source Speech Recognition Engines. `http://cmusphinx.sourceforge.net/`.

[8] Demirdjian, D., Ko, T., and Darrell, T. Constraining human body tracking. In *Proc. International Conference on Computer Vision*, Nice, France, October 2003.

[9] Gullberg, M. Gestures in spatial descriptions. In *Working Papers 47*, pages 87–97. Lund University, Department of Linguistics, 1999.

[10] Hanafiah, Z.M., Yamazaki, C., Nakamura, A., and Kuno, Y. Human-robot speech interface understanding inexplicit utterances using vision. In *Late Breaking Results of the 2004 Conference on Human Factors and Computing Systems (CHI'04)*, pages 1321–1324. ACM Press, April 24–29 2004.

[11] Huber, E. and Baker, K. Using a hybrid of silhouette and range templates for real-time pose estimation. In *Proc. International Conference on Robotics and Automation*, pages 1652–1657, New Orleans, Louisiana, 2004. IEEE.

[12] Huls, C., Bos, E., and Claassen, W. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics*, 21(1):59–79, 1995.

[13] Kaur, M., Tremaine, M., Huang, N., Wilder, J., and Gacovski, Z. Where is 'it'? event synchronization in gaze-speech input systems. In *Proc. 5th International Conference on Multimodal Interfaces (ICMI'03)*, pages 151–158, November 2003.

[14] Kendon, A. Current issues in the study of gesture. In Nespoulous, J.-L., Perron, P., and Lecours, A.R., editors, *The Biological Foundations of Gestures*, pages 23–47. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[15] Kobsa, A., Allgayer, J., Reddig, C., Reithinger. N., Schmauks, D., Harbusch, K., and Wahlster, W. Combining deictic gestures and natural language for referent identification. In *Proc. 11th Conference on Computational Linguistics*, pages 356–361, Bonn, Germany, 1986.

[16] Koons, D.B., Sparrell, C.J., and Thorisson, K.R. Integrating simultaneous input from speech, gaze, and hand gestures. In Maybury, M., editor, *Intelligent Multimedia Interfaces*, pages 257–276. MIT Press, Menlo Park, CA, 1993.

[17] Kuniyoshi, Y. and Inoue, H. Qualitative recognition of ongoing human action sequences. In *Proc. International Joint Conference on Artificial Intelligence*, pages 1600–1609, 1993.

[18] Latoschik, M.E. and Wachsmuth, I. Exploiting distant pointing gestures for object selection in a virtual environment. In Wachsmuth, I. and Fröhlich, M., editors, *Gesture and Sign Language in Human-Computer Interaction*, volume 1371 of *Lecture Notes in Artificial Intelligence*, pages 185–196. Springer-Verlag, 1998.

[19] Louwerse, M.M. and Bangerter, A. Focusing attention with deictic gestures and linguistic expressions. In *Proc. XXVII Annual Conference of the Cognitive Science Society (CogSci 2005)*, Stresa, Italy, July 21–23 2005.

[20] Machotka, P. and Spiegel, J. *The Articulate Body*. Irvington, 1982.

[21] Marslen-Wilson, W., Levy, E., and Tyler, L.K. Producing interpretable discourse: The establishment and maintenance of reference. In Jarvella, R.J. and Klein, W., editors, *Speech, Place and Action: Studies in Deixis and Related Topics*. Wiley, 1982.

[22] McNeill, D. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, IL, 1992.

[23] McNeill, D. and Levy, E. Conceptual representations in language activity and gesture. In Jarvella, R.J. and Klein, W., editors, *Speech, Place and Action: Studies in Deixis and Related Topics*. Wiley, 1982.

[24] Milota, A.D. and Blattner, M.M. Multimodal interfaces with voice and gesture input. In *Proc. International Conference on Systems, Man and Cybernetics*, pages 2760–2765, Vancouver, Canada, October 1995. IEEE.

[25] Moore, C. and Dunham, P.J., editors. *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum Associates, 1995.

[26] Moore, D., Essa, I., and Hayes, M. Exploiting human actions and object context for recognition tasks. In *Proc. International Conference on Computer Vision*, Corfu, Greece, 1999.

[27] Nagai, Y. Learning to comprehend deictic gestures in robots and human infants. In *Proc. 14th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN'05)*, pages 217–222, Nashville, TN, August 2005.

[28] Oviatt, S. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.

[29] Pavlovic, V.I., Sharma, R., and Huang, T.S. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):677–695, July 1997.

[30] Perzanowski, D., Schultz, A.C., Adams, W., Marsh, E., and Bugajska, M. Building a multimodal human-robot interface. *IEEE Intelligent Systems*, 16(1):16–21, 2001.

[31] Peters, R.A.II, Hambuchen, K.E., Kawamura, K., and Wilkes, D.M. The sensory ego-sphere as a short-term memory for humanoids. In *Proc. IEEE-RAS/RSJ Int'l Conf. on Humanoid Robots (Humanoids '01)*, pages 451–459, Tokyo, Japan, 2001.

[32] Pfeiffer, T. and Latoschik, M.E. Resolving object references in multimodal dialogues for immersive virtual environments. In *Proc. IEEE Virtual Reality Conference (VR'04)*, Chicago, IL, March 27–31 2004.

[33] Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

[34] Scaife, M. and Bruner, J.S. The capacity for joint visual attention in the infant. *Nature*, 253:265–266, 1975.

[35] Strobel, M., Illmann, J., Kluge, B., and Marrone, F. Using spatial context knowledge in gesture recognition for commanding a domestic service robot. In *Proc. 11th IEEE Workshop on Robot and Human Interactive Communication (RO-MAN'02)*, pages 468–473, Berlin, Germany, September 25–27 2002.