Interactive Reasoning about Spatial Concepts *

Marc Goodman, Scott Waterman, and Richard Alterman

Computer Science Department, Center for Complex Systems
Brandeis University, Waltham, MA 02254
Email: goodman@chaos.cs.brandeis.edu

Abstract

Spatial relations and spatial language form an important part of everyday reasoning. This paper describes SPATR, a system which addresses the labelling of components of objects and the interpretation of spatial relations between objects within the framework of adaptive planning. SPATR implements a model of *spatial reasoning*, which mediates among language, memory, and perception. Using a case-based approach for reasoning from past experience, SPATR makes use of spatial relationships corresponding to closed class terms, as well as a 3D, hierarchical representation of objects for retrieving relevant past experience.

Introduction

An agent using instructions to interact with a device must resolve several types of Natural Language references to spatial concepts. For example, when using an Airphone (a special type of payphone found on airplanes), an agent may be instructed to 'insert credit card, face up, with card name to the right' into the Airphone. Fully understanding the spatial aspects of this instruction requires the system to decide what is meant by "insert" (i.e., is it the kind of insert one uses when one wishes to insert a key into a lock, a pin into a voodoo doll, a bank card into an ATM, etc.), what is meant by the "face" of a credit card, what "up" means with respect to this face, where the "card name" is and whether right refers to the agent's right (a deictic reference) or some intrinsic right of the card or the Airphone.

We view many of these problems within the framework of Adaptive Planning [Alterman, 1986; Alterman, 1988]. Adaptive Planning suggests that *engagement*, or interacting with an environment, is a memory intensive task wherein an agent confronted with a 'typical' situation will behave in a manner consistent with previous experiences of that situation, and that an agent confronted with a novel situation will attempt to adapt strategies used in a previous similar experience. An implication of this approach is that experience serves to guide the direction of attention within the environment (i.e., you look for things you expect to find from past experience) and that noticing novel features of the environment guides the retrieval and adaptation of experience (e.g., noticing an airport shuttle sign when you need to get from one airport gate to another to make a connection suggests taking a shuttle rather than walking).

Within the context of the Airphone example, we view the process as follows:

- The agent approaches the Airphone for the first time. At this point in the interaction, the agent has not conducted an exhaustive visual inspection of the device, and therefore has only a skeletal representation for the Airphone.
- The agent is instructed to insert the credit card, face up with card name to the right. The 'figure' of the insert (see [Talmy, 1983] for explanations of the terms 'figure' and 'ground') is the credit card, and an appropriate geometric representation of the card is retrieved from memory. The ground of the insert is the Airphone, and the skeletal representation for the Airphone is used. The semantics of 'insert' suggest that an 'into' relationship is desired and this causes the retrieval of a likely previous experience, based on known features of the figure, and known and assumed features of the ground. For example, the agent may be reminded of inserting a bank card into an ATM.
- Inserting a bank card into an ATM consists, in part, of finding a slot on the front of the ATM, and this information is used to constrain visual search for such a slot on the front of the Airphone.
- This visual search fails and a new retrieval is performed based on different assumptions about the ground (and perhaps additional known features gained through the visual search). The agent may then be reminded of inserting a library book return

*This work was supported in part by the Defense Advanced Research Projects Agency, administered by the U.S. Air Force Office of Scientific Research under contract #F49620-88-C-0058.

card into the envelope on the front of a library book. This suggests a vertically oriented slot with an opening along its top, into which the card is inserted with a downward motion.

- Visual search succeeds for this feature and this experience of insert is chosen as the case to be adapted.
- 'Face' of a card metaphorically refers to the front of the card. We do not provide a model for linguistic metaphor, however we feel that [Martin, 1990] is more than adequate for these purposes. If the agent does not already know which side of a credit card is its front, then this decision could be handled by retrieving previous cases of labelling, and adapting these cases. For example, the agent may be reminded of a playing card where the front of the card is the side with the picture and writing on it, and label the credit card accordingly.
- Face 'up' actually means face towards the agent, given that insert requires a downward, sliding motion with the card oriented in a vertical position. This could be handled as a metaphoric meaning of up, an adaptation on the insert, or as a problem similar to the selection of the proper meaning of insert.
- We are assuming that the geometric representation of the credit card has adequate ties to semantic memory to find the 'card name' of the card.
- 'To the right' requires the agent to resolve whether 'right' refers to deictic or intrinsic right. For a full discussion of this process the reader is referred to [Alterman *et al.*, 1991].

SPATR

This paper will describe SPATR (pronounced "spatter"), a SPATial Reasoner which models the above processes. SPATR is implemented as a distributed case-based reasoning system built on top of Cognitive Systems' CBR Shell [Goodman, 1989; Goodman, 1990]. There are three components to SPATR: the I/D Labeller, CLTER, and SPACL. These components communicate through a blackboard. Each component is geared to handle a specific type of spatial problem.

Geometric Representations of Objects

In [Marr, 1982] a model for the geometric representation of objects based on a hierarchically organized 3D model is presented. For example, the overall shape of a person is cylindrical with a generating axis running from top to bottom, and this overall shape can be decomposed into a cylinder for the torso, cylinders for the arms and legs, etc. Arms can be further decomposed into forearms and biceps, forearms can be decomposed into hands, which can be decomposed into fingers, and so on. [Biederman, 1987] advocates a similar hierarchical decomposition into 'geons' which are a fundamental set of conics. Both approaches allow the recognition and categorization of perceptual information to occur at both coarse-grained and fine-grained levels.

Figure 1: Pictorial Representations of Key, Lock, and Key into Lock.

In [Jackendoff and Landau, 1990] a set of extensions to Biederman's approach to account for certain linguistic terms is given. These extensions allow:

- The addition of secondary directing axes in addition to a main generating axis to account for characteristic orientation of objects.
- Two-dimensional as well as three-dimensional subcomponents of objects (which offers a qualitative distinction between things which are fundamentally 'sheet-like' as opposed to 'very thin').
- Hollow objects (as opposed to shapes with slightly smaller negative shapes inset).
- Texture or character for objects (such as smooth, rough, jagged).
- Surface markings and/or color on objects (for example, writing on one side of a credit card).

SPATR's representational scheme for objects is adopted from this approach. For example, SPATR's representation for an (asymmetric) key would be a frame that could be represented pictorially as shown in Figure 1. The representation specifies that the overall shape of the key is strip-like, with a directed main generating axis. For convenience, we arbitrarily label this axis as a "Front→Back" axis (we will designate directed axes from X to Y as "X→Y" and undirected axes as "X↔Y"). The shape is further decomposed into a disk for the "back" part (actually, one end of the main axis) and a strip for the "front" part. This "front" strip itself contains a secondary directed axis which we arbitrarily label the "Top→Bottom" axis. The "top" of the strip is an edge whose character is straight and the "bottom" of the strip is an edge that is jagged.

Similarly, a lock could be represented as shown in Figure 1. This representation consists of a generic (or unspecified) 3D shape which contains a negative strip (i.e. the hole where the key fits). This strip has a directed main generating axis (which we label "Front→Back" for convenience). The "front" of this strip is a negative edge which has an undirected "Top↔Bottom" axis. Note that since our representation must account for world knowledge which is not necessarily perceptual in nature (for example, that there is, in fact, a negative strip inside the lock rather than a spherical shape or a slab, when only the leading edge of that strip is visible) SPATR must either deduce or explicitly represent the distinction between what is known through direct observation (e.g. that a

lock has a negative edge) and what is known through other methods (e.g. that a lock has a negative strip).

Motion and Change in Orientation

Taking action in the world requires a representation of motion and change in orientation over time. Inserting the card, lowering the door handle, removing and inserting the handset, and removing the card all require specifying paths between objects (as in [Jackendoff and Landau, 1990] or [Talmy, 1983]). Representing these paths may require specifying starting points (as in removing the handset and credit card), one or more midpoints (as in inserting the credit card into its holder and the handset into its cradle), and/or end points of the path (as in lowering the door handle over the card and the location to which the handset is returned). An example of changing the orientation of an object is specifying the rotation of a credit card so that it is "face up with name to the right."

Actions required to place a figure into a ground are decomposed into specifications of axial alignment, rotational alignment, and movement along paths in a manner akin to [Herskovits, 1985]. For example, inserting a key into a lock is represented as a frame which specifies that the "Front→Back" axis of the key must be aligned with the "Front→Back" axis of the negative strip of the lock, the key should be rotated until its "Top→Bottom" axis is parallel to the "Top←Bottom" axis of the negative edge of the lock, and the key should be moved along the path specified by its "Front→Back" axis. This process can be represented pictorially as shown in Figure 1.

An issue of primary importance is whether spatial relationships should be represented as abstract generalizations over related meanings (as is practiced by [Herskovits, 1985], c.f. [Talmy, 1983], and [Chen, 1987]) or as discrete, though related instances of differing relationships (*Extended Category Structure*: [Lakoff, 1987]). [Badler *et al.*, 1990] use an abstract notion of the meaning of spatial prepositions, such as 'on,' to resolve specific instructions, such as "put the block on the table."

This approach, however, requires significant work when given an instruction like "put the picture on the wall," or "put the teakettle on the stove." When placing a picture on a wall, the supporting surface is not horizontal, and when placing a kettle on a stove, a specific sub-area of the stove (i.e. the burner) is usually meant. SPATR's approach, following Lakoff, is to treat putting a block on a table, a picture on a wall, and a teakettle on a stove as distinct senses of 'on,' and to adapt the most similar sense given a new instruction.

CLTER, (pronounced as "clutter") the CLosed-class TERm library, represents each spatial interaction between a figure and ground as a distinct case, containing geometric representations of the figure and ground, and decompositions of actions required to affect the spatial relation. A working hypothesis of CLTER is that these cases are organized primarily into less than 100 spatial categories of interaction corresponding to the closed-class terms of a language (e.g. prepositions like 'in,' 'over,' 'at,' 'into,' etc. [Talmy, 1983]). Within these spatial categories, concepts form an extended category structure [Lakoff, 1987] with the leaf nodes for a category like 'into' being cases of, for example, 'inserting key into lock' or 'inserting pen into pocket.' The position of a node in the extended category structure is defined by its spatial characteristics (i.e. features of the figure and ground which account for differences in the spatial relationships).

When given a complete spatial representation of a figure and a partial spatial representation of a ground, CLTER retrieves previously stored spatial relationships, whose grounds are used to constrain visual search for necessary components of the ground. If this visual search fails, CLTER retrieves additional cases within the given spatial category and the process repeats until the search succeeds or the possibilities are exhausted.

Orientation and Labelling of Components

Many objects have salient features which allow us to distinguish between various orientations and sides of those objects. For example, the instruction which requests that the credit card be oriented "face up with name to the right" requires us to identify the card's face and to orient it accordingly. A further example is the instruction which requests that the handset be replaced "heel first." Such instructions refer to the fronts, backs, tops, bottoms, and left and right sides of objects. SPACL, the SPAtial Component Labeller (pronounced as "spackle"), maintains a case library of object representations with associated labellings based on the encoding of 'directing axes' in those representations (e.g. the top of a bottle is generally the side with an opening, and the front of a credit card is generally the side with the picture and bank name). These cases are used to resolve the labelling of components of objects based on previous, similar labellings (e.g. The front of the Airphone is the side with the phone hanging on it, based on our knowledge that the front of a payphone is the side with the phone hanging on it). More information on SPACL may be found in [Alterman *et al.*, 1991].

The I/D (Intrinsic/Deictic) labeller distinguishes between deictic and intrinsic reference systems given a referenced object. The I/D Labeller uses a case-based implementation of an algorithm suggested by Miller and Johnson-Laird [Miller and Johnson-Laird, 1976] to determine whether spatial descriptions imply intrinsic or deictic reference systems. Semantic knowledge of the referenced object is used to decide if the object has such characteristics as inherent perceptual apparatus, a normal orientation with respect to people, a characteristic direction of motion, etc. When these criteria indicate that the referenced object has an intrinsic ref-

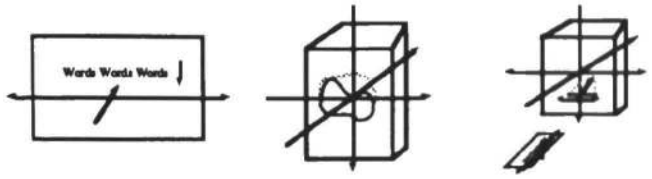

Figure 2: Pictorial Representation of Card, Airphone, and Card into ATM.

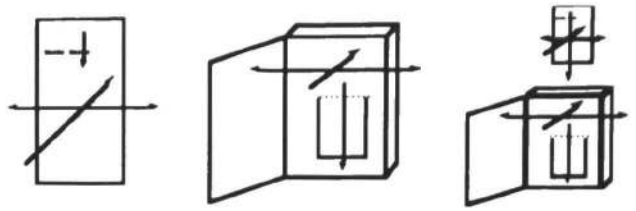

Figure 3: Pictorial Representation of Return Card, Library Book, and Insertion.

erent for the relation specified, the intrinsic system is used. A deictic system of reference is used when no intrinsic referent exists, or when insufficient semantic knowledge is present. Additional information on the I/D Labeller is also in [Alterman *et al.*, 1991].

FLOABN

SPATR has been applied to spatial problems within the context of the FLOABN (For Lack Of A Better Name) project (*Instruction Usage*: [Alterman and Zito-Wolf, 1990], *Adaptive Planning and Learning*: [Zito-Wolf and Alterman, 1990; Alterman *et al.*, 1991]). The FLOABN project attempts to adapt previously devised plans for the use of devices to new types of devices. In doing so, it may need to make use of instructions as well as generic semantic knowledge of device use.

SPATR will be invoked either through a failure in the adaptive planner (e.g. the planner tries to insert the credit card in a certain way, fails, and posts its failure to the blackboard causing SPATR to suggest different types of insertion), through text inferencing on instructions (e.g. the inferencer posts a partial interpretation of "Face up with card name to the right" and SPATR activates to resolve references to the card's face, and the orientation of the card [Carpenter and Alterman, 1991]), or through external reference resolution (e.g. the text inferencer interprets "door handle" to refer to a specific door handle, causing the reference resolver to look for a door handle, causing a request to provide a schematic description of a door handle, which activates SPATR).

SPATR Example

Given a new device, such as the Airphone, and an instruction to "insert credit card face up, with card name to right," into it, CLTER starts with a full representation for the credit card and a partial representation for the Airphone as shown in Figure 2. This representation states that the Airphone is overall slab-shaped, with directed axis from top to bottom, front to back and left to right (the directed axis follows from the device being mounted against a wall, in a particular orientation). The representation also states the expectation (from semantic memory) that the Airphone will have a negative space of undefined shape somewhere on its front where the card may be inserted.

Indices are traversed and the most similar previous case of insert is retrieved. The indices used for retrieval can be interpreted as:

- The Airphone has a main or secondary axis from top to bottom. This is true since the Airphone is mounted in a particular orientation.
- The Airphone is three dimensional. This is true from a cursory examination of the Airphone.
- There is a subcomponent of the Airphone on its front or back. This is true from our assumption that there will be a place on the front of the Airphone in which to insert the credit card.
- There is not a subcomponent of the Airphone with a top or bottom that has further subcomponents. This index serves to discriminate between cases where there is a slot or hole on the top of the object (e.g. a piggy bank or trash can) and other cases. Since we can't know whether the Airphone has or lacks this feature, we assume the Airphone doesn't have it. Note that our strategy will be to jump to conclusions to get a reminding, use the specific information off of the reminding to test its validity, and reverse assumptions if we fail.

The case retrieved is that of inserting a bank card into an ATM. The actions represented by this insertion are shown pictorially in Figure 2. SPATR posts a request to the blackboard for the (unimplemented) perceptual system to find a slot similar to the ATM's, which fails. Based on this failure we reverse our assumption about slots or holes on top, and re-retrieve. The result of this new retrieval is inserting a return card into a library book. A pictorial representation for this insertion is shown in Figure 3. Based on this retrieval, SPATR asks the perceptual system to find a slot oriented vertically on the front of the Airphone, with an opening along its top. This request succeeds and this meaning of 'into' is chosen as the appropriate precedent for adaptation.

Next, SPATR must identify the front of the credit card. When given a schematic representation of a credit card, as in Figure 2, SPACL retrieves a previous, labelled representation for a similar object (in this case a library book return card, shown in Figure 3). The return card is selected since both objects are sheets (i.e. flat rectangles) and both objects have words on a side referenced by a directed axis (i.e. there is a differentiation between a side which has words and the opposite

side). The return card specifies that the front of the return card is the side with the words (return dates) on it. SPACL uses this case to identify that the front of the credit card is the side with the words on it.

Finally, SPATR must decide whether "name to the right" refers to deictic or intrinsic right. The I/D labeller [Alterman *et al.*, 1991] decides that right is deictic (i.e. the viewer's) right, by analogy with mirrors, pictures, and playing cards, all of which are sheet-like objects with images and/or words on one side.

The final interpretation for "insert card face up, with name to the right" is, therefore, orient the credit card with the side which contains words towards the agent, with the edge of the card that contains the name toward the agent's right, and insert the card into a vertically positioned slot of the airphone with an opening towards the top of the slot with a downwards, sliding motion.

Conclusions

We have identified several problems of spatial reference and provided a framework for their solution which is consistent with Adaptive Planning. We have implemented systems to label components of objects, resolve references to intrinsic or deictic coordinate systems, and choose appropriate instances of actions and spatial relationships to constrain visual search and drive action.

References

[Alterman and Zito-Wolf, 1990] Richard Alterman and Roland Zito-Wolf. Planning and Understanding: Revisited. In *Proc. AAAI 1990 Spring Symposium*, Computer Science Department, Brandeis University, 1990.

[Alterman *et al.*, 1991] Richard Alterman, Roland Zito-Wolf, Tamitha Carpenter, Marc Goodman, and Scott Waterman. Readings from the FLOABN Project, Vol. 1. Draft CS-91-157, Brandeis University, Computer Science Department, Brandeis University, Waltham Massachusetts 02254-9110, 1991.

[Alterman, 1986] Richard Alterman. An adaptive planner. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 65–69, Philadelphia, PA, August 1986.

[Alterman, 1988] Richard Alterman. Adaptive planning. *Cognitive Science*, 12:393–421, 1988.

[Badler *et al.*, 1990] Norman Badler, Bonnie Webber, Jugal Kalita, and Jeffry Esakov. Animation from instructions. In N. Badler, B. Barsky, and D. Zeltzer, editors, *Making Them Move: Mechanics, Control and Animation of Articulated Figures*, chapter 3, pages 51–93. Morgan Kaufmann Publishers, Los Altos, CA, 1990.

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

[Carpenter and Alterman, 1991] Tamitha Carpenter and Richard Alterman. Explanation-based learning and summarization. To Appear in *Proceedings of the Thirteenth Cognitive Science Conference*, 1991.

[Chen, 1987] Su-Shing Chen. A geometric approach to multisensor fusion and spatial reasoning. In Avi Kak and Su shing Chen, editors, *Proceedings of the 1987 Workshop on Spatial Reasoning and Multi-Sensor Fusion*, pages 201–210. Morgan Kaufmann Publishers, Inc., 1987.

[Goodman, 1989] Marc Goodman. CBR In Battle Planning. In *Second Proceedings of a Workshop on Case Based Reasoning*, pages 312–326, 1989.

[Goodman, 1990] Marc Goodman. Prism: A case-based telex classifier. In Alain Rappaport and Reid Smith, editor, *Proceedings of the Second Conference on Innovative Applications of Artificial Intelligence*, pages 86–90, 1990.

[Herskovits, 1985] Annette Herskovits. Semantics and pragmatics of locative expressions. *Cognitive Science*, 9:341–378, 1985.

[Jackendoff and Landau, 1990] Ray Jackendoff and Barbara Landau. Spatial Language and Spatial Cognition: A Swarthmore FestSchrift for Lila Gleitman. In J. Kegl D. J. Napoli, editor, *Bridges Between Psychology and Linguistics*. Lawrence Erlbaum Associates, Inc., 1990.

[Lakoff, 1987] George Lakoff. *Women, Fire, and Dangerous Things*. The University of Chicago Press, 1987.

[Marr, 1982] D. Marr. *Vision*. W. H. Freeman, San Francisco, California, 1982.

[Martin, 1990] James H. Martin. Computer understanding of conventional metaphoric language. Technical Report CU-CS-473-90, University of Colorado, Boulder, Computer Science Department and Institute of Cognitive Science, 1990.

[Miller and Johnson-Laird, 1976] George A. Miller and Phillip N. Johnson-Laird. *Language and Perception*. The Belknap Press of Harvard University Press, Campbridge, Massachusetts, 1976.

[Talmy, 1983] Leonard Talmy. How language structures space. In Herbert Pick and Linda Acredolo, editors, *Spatial Orientation: Theory, Research, and Application*, pages 225–282. Plenum Press, 1983.

[Zito-Wolf and Alterman, 1990] Roland Zito-Wolf and Richard Alterman. Ad-Hoc, Fail-Safe Plan Learning. In *Proceedings of the Twelfth Cognitive Science Conference*, Computer Science Department, Brandeis University, 1990. Lawrence Erlbaum Associates.