2002 IEEE International Conference on Robotics and Automation, Washington, D.C.

Using Spatial Language in a Human-Robot Dialog

Marjorie Skubic[1], Dennis Perzanowski[2], Alan Schultz[2], William Adams[2]

[1] Computer Engineering and Computer Science Department
University of Missouri-Columbia, Columbia, MO 65211
skubicm@missouri.edu

[2] Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory, Washington, DC 20375-5337
<dennisp | schultz | adams> @aic.nrl.navy.mil

Abstract

In conversation, people often use spatial relationships to describe their environment, *e.g.*, "There is a desk in front of me and a doorway behind it", and to issue directives, *e.g.*, "Go around the desk and through the doorway." In our research, we have been investigating the use of spatial relationships to establish a natural communication mechanism between people and robots, in particular, for novice users. In this paper, the work on robot spatial relationships is combined with a multi-modal robot interface developed at the Naval Research Lab. We show how linguistic spatial descriptions and other spatial information can be extracted from an evidence grid map and how this information can be used in a natural, human-robot dialog.

1. Introduction

In conversation, people often use spatial relationships to describe their environment, *e.g.*, "There is a desk in front of me and a doorway behind it", and to issue directives, *e.g.*, "Go around the desk and through the doorway". Recent cognitive models suggest that people use these types of relative spatial concepts to perform day-to-day navigation tasks and other spatial reasoning [1,2], which may explain the importance of spatial language and how it developed. In our research, we have been investigating the use of spatial relationships to establish a natural communication mechanism between people and robots, in particular, striving for an intuitive interface that will be easy for novice users to understand.

In previous work, Skubic *et al.* developed two modes of human-robot communication that utilized spatial relationships. First, using sonar sensors on a mobile robot, a model of the environment was built, and a spatial description of that environment was generated, providing linguistic communication from the robot to the user [3]. Second, a hand-drawn map was sketched on a PDA, as a means of communicating a navigation task to a robot [4]. The sketch, which represented an approximate map, was analyzed using spatial reasoning, and the navigation task was extracted as a sequence of spatial navigation states. In [5], the results of these two modes were compared for similar, but not exact environments, and found to agree.

In this paper, robot spatial reasoning is combined with a multi-modal robot interface developed at the Naval Research Laboratory (NRL) [6,7]. Spatial information is extracted from an evidence grid map, in which information from multiple sensors is accumulated over time [8]. Probabilities of occupancy are computed for grid cells and used to generate a short-term map. This map is then filtered, processed, and segmented into environment objects. Using linguistic spatial terms, a high-level spatial description is generated which describes the overall environment, and a detailed description is also generated for each object. In addition, a class of persistent objects has been created, in which objects are given locations in the map and are assigned labels provided by a user.

The robot spatial reasoning and the NRL Natural Language Processing system are combined to provide the capability of natural human-robot dialogs using spatial language. For example, a user may ask the robot, "How many objects do you see?" The robot responds, "I am sensing 5 objects." The user continues, "What objects do you see?" The robot responds, "There are objects behind me and on my left." And the dialog continues.

The paper is organized as follows. Section 2 provides an overview of the multi-modal interface. In Section 3, we discuss algorithms used to process the grid map and generate multi-level linguistic spatial descriptions. Section 4 discusses how the spatial language is used in an interactive dialog, and Section 5 provides conclusions.

2. The Multi-Modal Interface

Our research is based on previous work building a user-friendly multi-modal interface [6] for a team of dynamically autonomous [9] mobile robots (Figure 1). The interface allows the user to concentrate on the task at hand, rather than on the interaction modality. Users are allowed to choose freely and combine various modes for inputting commands and queries, including speech, gestures, and personal electronic devices. To achieve an intuitive interaction similar to communication with other humans, the interface should also support commands and spatial references such as "Follow that wall" or "Stop at the doorway on your left."

Figure 1. A Nomad 200 and an RWI ATRV

Figure 2. A Palm Pilot Personal Digital Assistant

Gestures occur frequently with natural language; some of these gestures provide crucial information to spoken commands, such as gesturing in a direction when someone says, "Go over there." Gestures are given to the robot through arm movements in a particular direction, or toward a particular location. These so-called *natural* gestures are coupled with spoken commands to clarify an otherwise ambiguous directive.

In addition to issuing verbal commands (with or without gestures), the user can also interact with the robots through a Personal Digital Assistant (PDA) (Figure 2). Discrete commands can be issued through menu buttons, or the operator can use *synthetic* gestures by pointing to coordinates on the PDA screen, or by dragging a stylus on a mapped representation of the area.

The underlying goal of the research is to make robots capable of interacting freely with each other and with human users, at levels of autonomy that are dictated by the goals, interactions, and changing situations. To facilitate dynamically changing levels of autonomy, the status of goals is recorded in a structure known as a *context predicate* [9]. This structure allows the robots to re-assess the situation while carrying out goals, *e.g.*, re-assessing spatial relationships in the environment.

For example, a user may issue a command to the robot named Coyote, by speaking "Coyote, go over there". The command is parsed and a context predicate is constructed as a list (shown as the second element in (1)). After another command is given, "Coyote, back up this far," the second directive is stacked onto the previous element, as shown in (1). Thus, a context predicate is a stack of lists, containing semantic information obtained during the parsing of a series of commands or queries.

```
((((imper: back-direction: back)                    (1)
                (agent: system: coyote)
                (goal: direction: far)
            (goal-state: incomplete)))
    ((imper: go-direction: go
                (agent: system: coyote)
                (goal: location: there)
                (goal-state: complete))))
```

The context predicate also contains information about the status of the goals; *i.e.,* whether or not they have been achieved. Any of the participating agents in the dialog can check the context predicate to see which goals have not been completed and act on them if needed. The information in the stack is updated based on the status of goals, the existing context, and the focus of the dialog.

Given the various means of inputting commands and queries, and the creation of a context predicate, once the human user issues a command, the user can then redirect his/her attention to other matters, and the robots can go about their business without having to interrupt the user for additional information.

Since many of the commands involved spatial references, or required a priori knowledge of objects and their locations in the immediate environment, we introduced a spatial relations component. With the information of this component, we can augment the spatial information available to the various agents in the human-robot dialog and can reason about locations and objects in the environment.

3. Generating Spatial Language from Occupancy Grid Maps

The map structure used in this work is an evidence grid map [8]. The indoor environment shown in this paper is represented with a 128 x 128 x 1 cell grid, providing a two-dimensional map of the NRL lab. One cell covers approximately 11cm x 11 cm. Information from the robot sensors is accumulated over time to calculate probabilities of occupancy for each grid cell. One byte is used to store occupancy probabilities; values range from +127 (high probability of occupancy) to -127 (high probability of no occupancy), with 0 representing an unknown occupancy. For the work reported here, these maps are the sensor-fused short-term maps generated by the robot's regular localization and navigation system [10]. Examples of evidence grid maps are shown in Figures 3(a) and 4(a). For our purposes, a cell with an occupancy value ≥ +1 is considered to be occupied and is shown in black. All other cells are shown in white.

The evidence grid map is pre-processed with a sequence of operations, similar to those used for image processing, to segment the map into individual objects. First, a filter is applied through a convolution operation. A 3x3 matrix, shown below in (2), is used as the convolution kernel, K, to provide a linear filter of the map.

$$K = \begin{vmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \end{vmatrix} \qquad (2)$$

This has the effect of blurring the map, filtering single cells and filling in some disconnected regions, as shown in Figure 3(b).

An explicit fill operation is also used to further fill in vacant regions. For each unoccupied cell, if 5 or more of

its neighbors are occupied, then the cell status is changed to occupied. Eight neighbors are considered, as shown below in (3) for cell $a_{i,j}$:

$a_{i-1,j-1}$	$a_{i,j-1}$	$a_{i+1,j-1}$
$a_{i-1,j}$	$a_{i,j}$	$a_{i+1,j}$
$a_{i-1,j+1}$	$a_{i,j+1}$	$a_{i+1,j+1}$

(3)

Two passes of the fill operation are executed. Results are shown in Figure 3(c).

Finally, spurs are removed. A spur is considered to be an occupied cell with only one occupied neighbor in the four main directions (diagonal neighbors are not counted). All spurs, including those with a one-cell length, are removed. At this point, the final cell occupancy has been computed for object segmentation. Objects should be separated by at least a one-cell width.

Next, objects are labeled and loaded into a data structure for spatial reasoning. A recursive function is used to label adjacent cells. Occupied cells are initially given numeric labels for uniqueness, e.g., object #1, object #2. Once the cells are labeled, a recursive contour algorithm is used to identify the boundary of the objects. The contour is important in that it provides a representation of the environment obstacles that is used for spatial reasoning. Examples of the final segmented objects, with their identified contours, are shown in Figures 3(d) and 4(b).

Spatial reasoning is accomplished using the histogram of forces [11], as described in previous work [3,4,5,12]. For each object, two histograms are computed (the histograms of constant forces and gravitational forces), which represent the relative spatial position between that object and the robot. Computationally, each histogram is the resultant of elementary forces in support of the proposition object #*i* is in direction θ of the robot. For fast computation, a boundary representation is used to compute the histograms. The robot contour is approximated with a rectangular bounding box. The object boundaries are taken from the contours of the segmented objects in the grid map.

The two histograms give different views of the environment; the histogram of constant forces provides a global view and the histogram of gravitational forces provides a local view. Features from the histograms are extracted and input into a system of fuzzy rules to generate a three-part linguistic spatial description: (1) a primary direction (*the object is in front*), (2) a secondary direction which acts as a linguistic hedge (*but somewhat to the right*), and (3) an assessment of the description (t*he description is satisfactory*). A fourth part describes the Euclidean distance between the object and robot (*the object is close*). In addition, a high level description is generated that describes the overall environment with respect to the robot. This is accomplished by grouping the objects into 16 (overlapping) regions located around

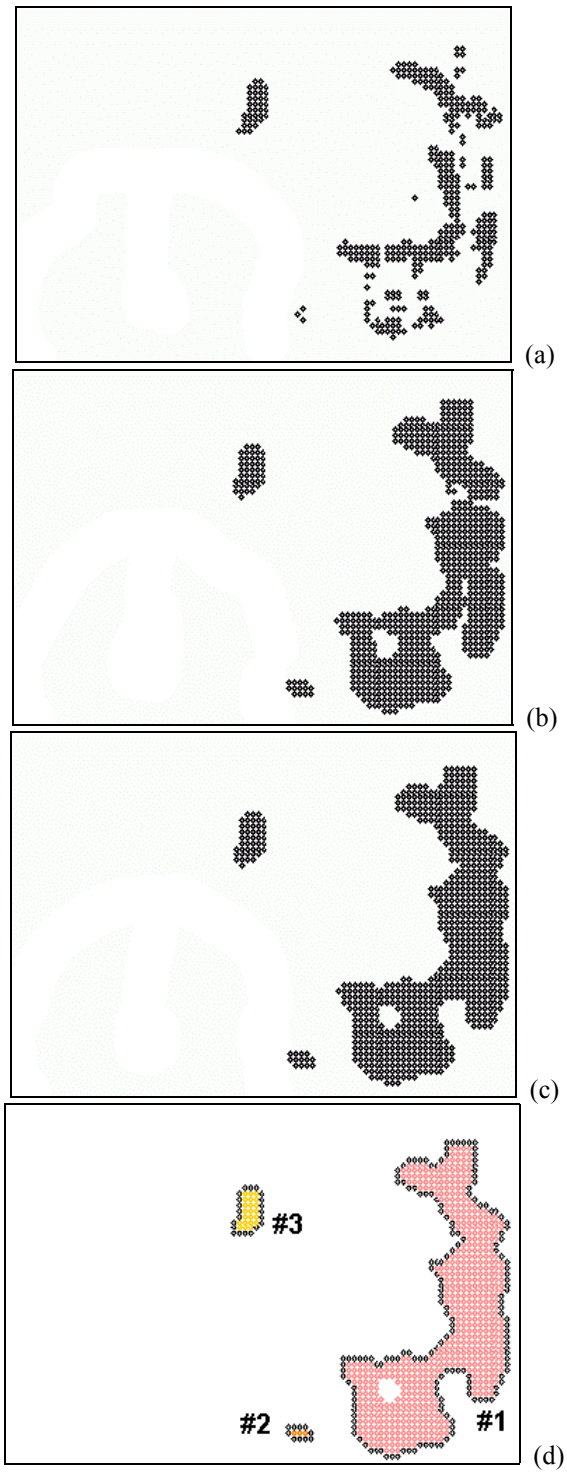
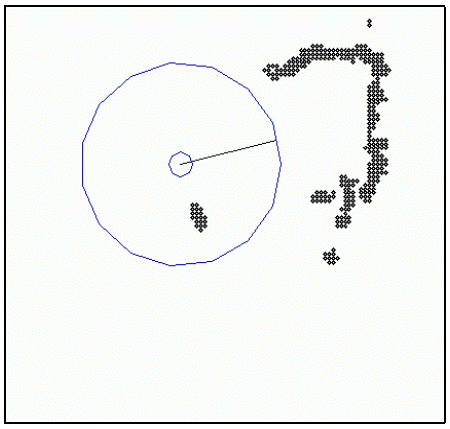
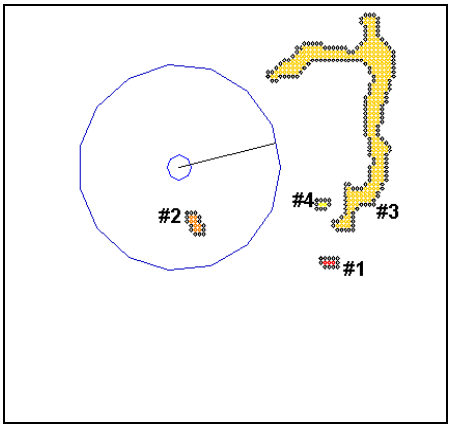


Figure 3. (a) The southeast part of the evidence grid map. Occupied cells are shown in black. (b) The result of the filter operation. (c) The result of the fill operation. (d) The segmented, labeled map. Physically, object #1 corresponds to a section of desks and chairs, object#2 is a file cabinet, and object #3 is a pillar.

the robot. An example of the generated descriptions is shown in Figure 4(c). See [3,4,5] and especially [12] for additional details.



(a)



(b)

DETAILED SPATIAL DESCRIPTIONS:
The #1 object is to the right-front of me
-
(the description is satisfactory)
The object is close.

The #2 object is to the right of me
-
(the description is satisfactory)
The object is very close.

The #3 object is in front of me
-
(the description is satisfactory)
The object is close.

The #4 object is mostly in front of me
but somewhat to the right
(the description is satisfactory)
The object is close.

HIGH LEVEL DESCRIPTION:
There are objects in front of me and on my right.

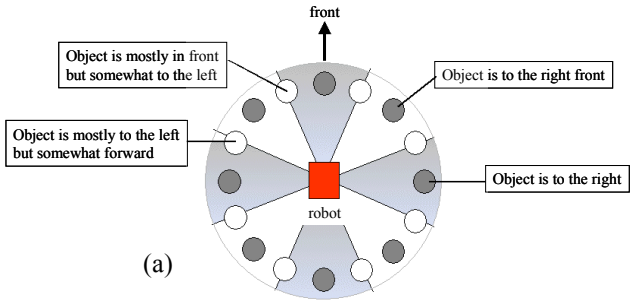
(c)

Figure 4. (a) A robot situated in the grid map. (b) The segmented, labeled map. (c) The generated descriptions. Note the robot heading. Object#2 corresponds to the same pillar in Figure 3(d).

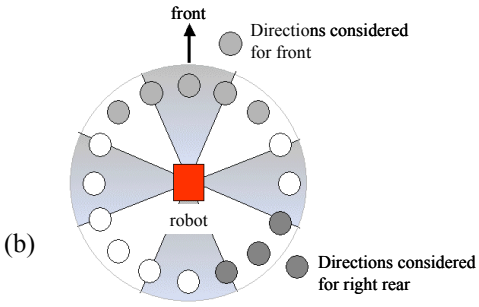
4. Integrating Spatial Language into Human-Robot Dialog

The robot control system has been implemented as a distributed system with components for path planning, map processing, localization, navigation, and handling the various interface modalities (PDA, gesture, and speech input). The spatial reasoning capabilities have been integrated into this environment in the form of a server so that any client can request the spatial description of the environment at any given time.

As the descriptions are generated, information is also stored on the relative spatial positions of the environment objects to facilitate a meaningful dialog with the robot. From the histogram computation, each object is assigned a primary direction. The 16 possible primary directions situated around the robot are illustrated in Figure 5(a), with examples of the corresponding linguistic descriptions. For each object, the primary direction is mapped to a set of 8 regions around the robot (front, rear, left, right, and the diagonals), which are used for queries. Two examples are shown in Figure 5(b). An object in any of the 5 light gray directions is considered to be in front of the robot. An object in any of the 3 dark gray directions is considered to be to the right rear. Thus, an object that is to the right front (see Figure 5(a)) would be retrieved in queries for three regions: the front, the right, and the right front of the robot.



(a)



(b)

Figure 5. Sixteen primary directions are situated around the robot (the small circles). The 8 cone-shaped sections represent the 8 basic regions (front, rear, left, right, and diagonals). (a) Examples of the corresponding linguistic descriptions. (b) Examples used for queries. An object is considered to be in front of the robot if it occupies one of the 5 light gray primary directions. Diagonal directions such as right rear comprise only 3 primary directions (dark gray).

With this spatial information and linguistic descriptions, we can now establish a dialog using spatial language. Queries can be made with the 8 basic regions illustrated in Figure 5. Consider the scene in Figure 4 for the dialog below. Note that the user has all of the interface modalities available and can view the graphical display of the scene while talking to the robot. The robot responds using synthesized speech output [13].

User: "How many objects do you see?"
Robot: "I am sensing 4 objects."
User: "What objects do you see?"
Robot: "There are objects in front of me and on my right."
User: "Where is the nearest object on your right front?"
Robot: "The #4 object is mostly in front of me but somewhat to the right. The object is close."
User: "Where is the nearest object in any direction?"
Robot: "The #2 object is to the right of me. The object is very close."
User: "Where is the nearest object on your left?"
Robot: "I am not sensing any objects in that direction."

Motion directives can also be given using spatial language, as shown in the dialog below:

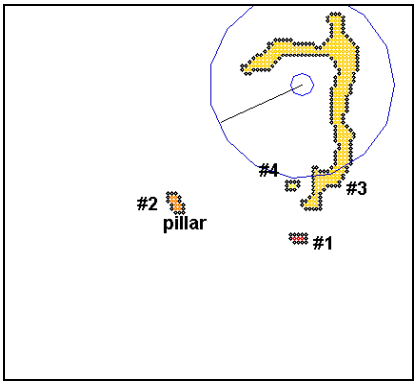
User: "Go to the nearest object in front of you."
Robot: "Looking for the nearest object in front of me."
"I found it. Going to the object."

We have also defined a class of persistent objects that have a designated location and are assigned a label by the user. Persistent objects are created from the objects identified in the grid map, through a dialog with the robot. For example, consider the scene in Figure 6.

User: "Where is the nearest object in front of you?"
Robot: "The #2 object is mostly in front of me but somewhat to the left. The object is close."
User: "The #2 object is a pillar."
Robot: "I now know that the #2 object is a pillar. The pillar is mostly in front of me but somewhat to the left. The object is close."

As the robot moves around the environment, it remembers where the pillar is and will continue to generate spatial information about the pillar. For the scene in Figure 7, three objects have been identified from the grid map; the cells occupied by the pillar are also shown in the figure. The user can now establish a dialog using the defined object.

User: "Where is the pillar?"
Robot: "The pillar is mostly in front of me but somewhat to the right. The object is very close."



(a)

DETAILED SPATIAL DESCRIPTIONS:
The #1 object is mostly to the left of me
but somewhat forward
(the description is satisfactory)
The object is close.

The #2 object is mostly in front of me
but somewhat to the left
(the description is satisfactory)
The object is close.

I am surrounded from the rear
(surrounded by the #3 object).
The object is very close.

The #4 object is mostly to the left of me
but somewhat forward
(the description is satisfactory)
The object is close.

HIGH LEVEL DESCRIPTION:
There are objects on my front left.
I am surrounded from the rear.

(b)

Figure 6. Creating persistent objects. (a) The robot situated in the segmented map. (b) The generated descriptions before defining the "pillar".

Physically, the pillar and object #3 are the same although they do not occupy exactly the same grid cells. Note, however, that the description of both object #3 and the pillar, as shown in Figure 7(b), are exactly the same. In future work, we will explore algorithms for connecting persistent objects with those identified dynamically from the grid map.

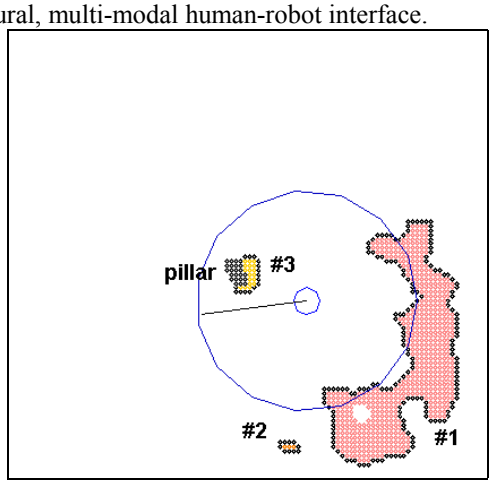
Figure 6 also shows an example of the surrounded relation, which provides capabilities of high level spatial reasoning. In [12], we introduce 3 levels of surrounded based on the width of the force histograms, *e.g.*, (1) *I am surrounded on the right*, (2) *I am surrounded with an opening on the left*, and (3) *I am surrounded*.

5. Concluding Remarks

In this paper, we showed how an evidence grid map is processed so that linguistic expressions can be generated to describe the environment with respect to a robot. Also, we showed how spatial language can be integrated into a multi-modal robot interface to provide

capabilities for a natural, human-robot dialog. The work thus far illustrates further questions that need to be addressed, *e.g.*, what is the most useful spatial language needed for a dialog, and what is the best frame of reference for different types of tasks.

In the future, we intend to address these problems. We also want to explore the use of spatial information in robot behaviors, to facilitate directives with respect to objects in the environment, *e.g.,* "Move forward until the pillar is behind you". We will also explore the concept of unoccupied spatial regions, which can be used in commands such as "Go to the left of the pillar" or "Go around the pillar". This continued work in supporting and developing spatial language contributes to the natural, multi-modal human-robot interface.



(a)

The #1 object is behind me
but extends to the left relative to me
(the description is satisfactory)
The object is very close.

The #2 object is mostly to the left of me
but somewhat forward
(the description is satisfactory)
The object is close.

The #3 object is mostly in front of me
but somewhat to the right
(the description is satisfactory)
The object is very close.

The pillar is mostly in front of me
but somewhat to the right
(the description is satisfactory)
The object is very close.

(b)

Figure 7. (a) The scene from Figure 3 with the persistent object "pillar". (b) The generated detailed descriptions.

Acknowledgements

This research has been supported by ONR and the Naval Research Laboratory. The authors would also like to acknowledge the help of Magda Bugajska at NRL and Dr. Pascal Matsakis and Dr. Jim Keller at UMC.

References

[1] F.H. Previc, "The Neuropsychology of 3-D Space", Psychological Review, 1998, vol. 124, no. 2, pp. 123-164.

[2] C. Schunn, T. Harrison. Personal communication. 2001.

[3] M. Skubic, G. Chronis, P. Matsakis and J. Keller, "Generating Linguistic Spatial Descriptions from Sonar Readings Using the Histogram of Forces", in *Proc. of the 2001 IEEE Intl. Conf. on Robotics and Automation*, May, 2001, Seoul, Korea, pp. 485-490.

[4] M. Skubic, P. Matsakis, B. Forrester and G. Chronis, "Extracting Navigation States from a Hand-Drawn Map", in *Proc. of the 2001 IEEE Intl. Conf. on Robotics and Automation*, May, 2001, Seoul, Korea, pp. 259-264.

[5] M. Skubic, G. Chronis, P. Matsakis and J. Keller. "Spatial Relations for Tactical Robot Navigation", in *Proc. of the SPIE, Unmanned Ground Vehicle Technology III*, April, 2001, Orlando, FL.

[6] D. Perzanowski, A.C. Schultz, W. Adams, E. Marsh, M. Bugajska, "Building a Multimodal Human-Robot Interface", *IEEE Intelligent Systems*, Jan./Feb, 2001, pp. 16-20.

[7] W. Adams, D. Perzanowski, A.C. Schultz, "Learning, Storage and Use of Spatial Information in a Robotics Domain", *Proc. of the ICML 2000 Workshop on Machine Learning of Spatial Language*, Stanford Univ.: AAAI Press, pp. 23-27.

[8] M.C. Martin, H.P. Moravec, "Robot Evidence Grids", Technical Report #CMU-RI-TR-96-06, Carnegie Mellon University, Mar., 1996.

[9] D. Perzanowski, A. Schultz, W. Adams, and E. Marsh, "Goal Tracking in a Natural Language Interface: Towards Achieving Adjustable Autonomy," *Proc. of the 1999 IEEE Intl. Symp. on Computational Intelligence in Robotics and Automation,* Monterey, CA, Nov,1999, pp.208-213.

[10] A. Schultz, W. Adams and B. Yamauchi, "Integrating Exploration, Localization, Navigation and Planning with a Common Representation," *Autonomous Robots*, vol.6, no.3, May, 1999.

[11] P. Matsakis and L. Wendling, "A New Way to Represent the Relative Position between Areal Objects", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 7, pp. 634-643, 1999.

[12] M. Skubic, P. Matsakis, G. Chronis, and J. Keller, "Generating Multi-Level Linguistic Spatial Descriptions from Range Sensor Readings Using the Histogram of Forces," Submitted to *Autonomous Robots*

[13] D. Perzanowski *et al.*, "Multi-Modal Navigation of Robots Using Spatial Relations: A Videotaped Demonstration", submitted to ICRA 2002