Robot Navigation Using Human Cues: A robot navigation system for symbolic goal-directed exploration*

Ruth Schulz, Ben Talbot, Obadiah Lam, Feras Dayoub, Peter Corke, Ben Upcroft, and Gordon Wyeth[1]

***Abstract*— In this paper we present for the first time a complete symbolic navigation system that performs goal-directed exploration to unfamiliar environments on a physical robot. We introduce a novel construct called the abstract map to link provided symbolic spatial information with observed symbolic information and actual places in the real world. Symbolic information is observed using a text recognition system that has been developed specifically for the application of reading door labels. In the study described in this paper, the robot was provided with a floor plan and a destination. The destination was specified by a room number, used both in the floor plan and on the door to the room. The robot autonomously navigated to the destination using its text recognition, abstract map, mapping, and path planning systems. The robot used the symbolic navigation system to determine an efficient path to the destination, and reached the goal in two different real-world environments. Simulation results show that the system reduces the time required to navigate to a goal when compared to random exploration.**

I. INTRODUCTION

Humans use navigational cues—door labels, sign posts, and maps—to perform everyday navigation, particularly when visiting a location for the first time. Human navigational cues are typically symbols—text or graphics—where each symbol attributes a particular meaning to a location, or perhaps indicates a relationship to a distal location. The process of navigation using human cues is about finding meaning in the symbols located in the environment, then reasoning about those symbols to solve the navigation problem.

Robots, on the other hand, typically navigate by integrating sensor information from range sensors, cameras and odometers. Sensor information is geometric in nature, and each measurement of the environment's geometry must be made to mesh with previous measurements; the well-known problem of Simultaneous Localization and Mapping. A robot that navigates using symbolic information from human cues requires a different approach.

In this paper, we describe a robotic system that uses symbolic human cues to perform goal-directed navigation (see Fig. 1). The key contribution of the paper is the *abstract map*: a construct that links symbolic spatial information from multiple sources and robot observations to make inferences about the location of places. Sources of spatial information in the abstract map might include floor plans, campus maps, web queries, sketch maps, or even natural language statements ("Peter's office is S1104A") and directions ("Access to S1104A is via S1105"). Such symbolic spatial information is potentially very rich, but may also be redundant, inconsistent, or ambiguous. In the system described in this paper, the robot reads a graphical floor plan containing textual room names, and then uses *wild text* recognition to find door labels that can guide the robot to its goal. A simple instance of an abstract map resolves the symbolic information to infer the goal's location. We demonstrate the successful use of our symbolic navigation system to navigate to an unseen goal in two different real-world environments.

Fig. 1: Our robot after reaching its coffee delivery goal. The robot, like a delivery person, started in an unfamiliar environment and used observed symbolic information to guide itself toward the room that ordered the coffee.

II. RELATED WORK

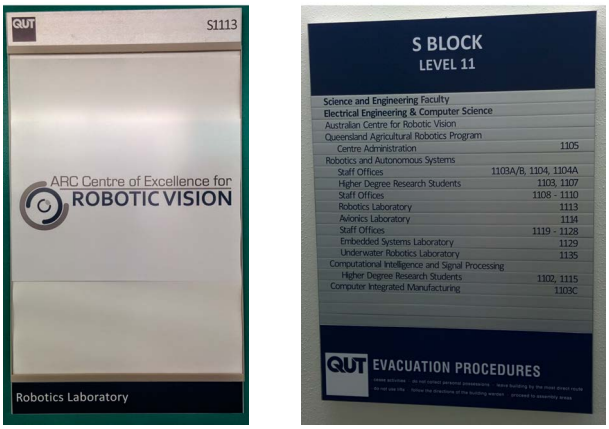
This section reviews related work on symbolic spatial information, existing robotic systems for wild text recognition, and existing symbolic navigation systems.

A. Symbolic spatial information

Symbolic spatial information comes in many forms, including natural language, route directions, gestures, signs, and pictorial representations. In this paper, we focus on information that can be independently used within buildings: floor plans and signs indicating location. Locations of interest are rooms, which are often named by numbers. Floor plans

*This research was supported under Australian Research Council's Discovery Projects funding scheme (project number DP140103216)

[1]The authors are with the Australian Centre for Robotic Vision, School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia. http://www.roboticvision.org/ email: {ruth.schulz, b.talbot, o1.lam, feras.dayoub, peter.corke, ben.upcroft, gordon.wyeth}@qut.edu.au



(a) (b)

Fig. 2: Examples of symbolic spatial information a) Door label, b) Floor directory

use the visual space in a picture to represent space between geometric features such as rooms [1]. Non-metric sketch maps have been shown to be an even more effective tool for navigation than route directions [2], using the relationship between spatial features to describe space. Room numbers are often found both in floor plans and evacuation maps as well as on the doors to the rooms and on signs providing directions to commonly used rooms (see Fig. 2).

B. Wild text recognition

Symbolic spatial information is abundant in built environments, often as text visible to the camera of a robot exploring the world. Although systems for recognizing printed text are advanced, the application of such systems to read and interpret visible text from a robot's perspective remains a current research problem [3].

There are multiple stages in a scene text reading pipeline. First, text is detected within an image. This text is passed on to the character recognition stage, which determines individual letters and numbers in the text region. Finally, the sequence of recognized characters is collected into a word, and a dictionary of possible words may be used to correct the outputs.

A method for text detection was introduced in [4] called the stroke width transform. This uses the prior that text characters are generally the same width throughout the stroke. In [5], colour and geometrical features from detected maximally stable extremal regions [6] were used to determine whether candidate text regions contained text characters. Leading wild text recognition pipelines, such as PhotoOCR [7], treat the character recognition problem as an object recognition problem. These methods use Convolutional Neural Networks [8] as the recognition process and train the networks with very large training sets (more than 2 million images). An alternative approach is to take the output of the text detection step and pass it to a traditional document Optical Character Recognition engine such as Tesseract [9]. Previous work in [3] showed that Tesseract is unreliable due to false positives in texture images such as walls and fences.

C. Symbolic navigation systems

Symbolic navigation systems use symbolic spatial information to aid navigation. The autonomous city explorer project [10] was an early example of a symbol-based navigation system, using pointing gestures alone to guide a mobile robot to a requested destination in a city. The level of interaction has been extended in Walter's work [11] where rudimentary language was used to build a semantic graph of an environment. Fasola [12] and Kollar [13] on the other hand, used labeled spatial maps and spatial object association respectively to complete symbolic navigation directions. Elements that are required by these systems, but which are not guaranteed to be available to our system, include human supervision, qualitative feedback, symbols which are already grounded to data, or object recognition based classification, making them not directly suited for the autonomous scope of this research.

III. Approach

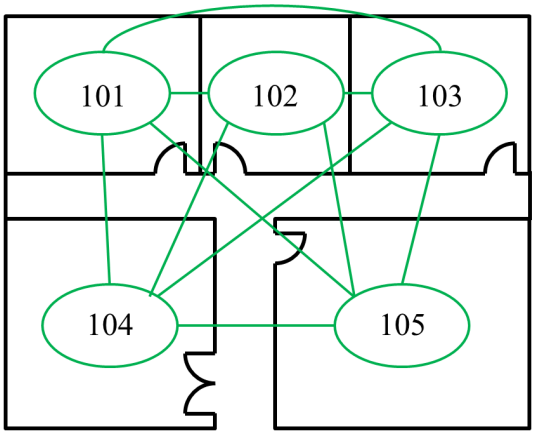
Consider a robot that is given the task of finding room '103' and provided with a floor plan that shows the positions of several nearby rooms (see Fig. 3a). One type of information that can be extracted from a floor plan is the relative location of each room, as indicated by the position of the text labels. This information, an embedded graph, can be held in an abstract map (see Fig. 3a overlay). The assumption is that the robot will be able to match the text labels of rooms on the floor plan to door labels in the world (such as Fig. 2a).

As the robot explores its world (see Fig. 3b), it builds up a representation of the world, including a map and symbol observations. Once symbolic information such as a door label has been observed and internally stored, the robot can use the information it conveys by transforming the information in the abstract map into the robot's representation of the world. The system needs symbol observations from at least two different doors to estimate the translation, scale, and rotation. Once transformed, locations in the abstract map can be used in goal-directed exploration. The robot can confirm that it has reached '103' by a symbol observation of the door label (see Fig. 3c).

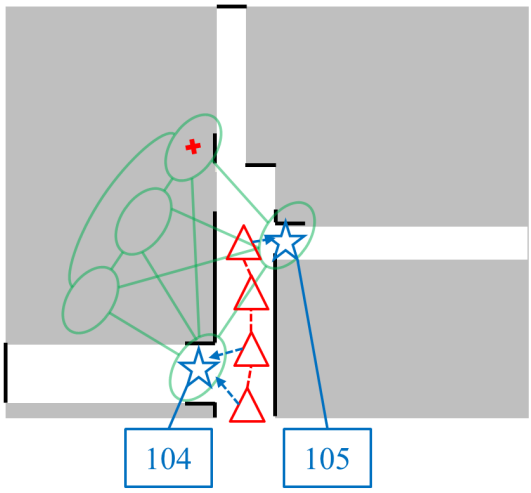
Symbolic spatial information is dealt with in the system outlined in this paper in a novel spatial data structure called the abstract map. Symbol observations are obtained from the world through the vision system. Symbolic goal-directed navigation is performed by linking the information in the abstract map with symbol observations grounded in the robot's world map, and setting goals based on the transformed information.

A. Storing symbolic information in an abstract map

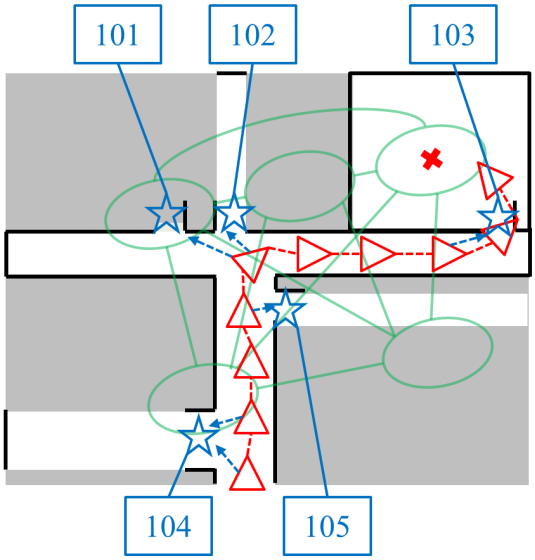
The abstract map functions as a data structure to capture pieces of symbolic spatial information into a meaningful collection, and uses this information to facilitate symbolic goal-directed exploration. In the experiment described in this paper, we illustrate the concept by constructing an abstract



(a) The robot extracts information from the floor plan (black) and constructs an abstract map (green).



(b) After two symbol observations of door labels the robot grounds its abstract map by performing an initial transform.



(c) The robot uses its grounded abstract map to plan a path to the destination, updating the transform as more door labels are observed, and reaches the goal, indicated by observing the door label 103.

Fig. 3: Exploration process, with the robot's destination set to '103'

map using just one source of symbolic spatial information—a floor plan. We show the efficacy of the approach for targeted exploration using a second source of symbolic spatial information—door labels.

The information held in the abstract map is the relative location of rooms as indicated by the pixel coordinates of room label text in the provided floor plan. For the study described here, this is a manual process; pixel locations are calculated from mouse clicks on text in the floor plan image and are stored together with the transcribed text. We are currently developing an automated system for the extraction of room locations from floor plans that are obtained either from the internet or from the robot's camera as it explores the world. All rooms that are included in the abstract map are compiled into a room-name dictionary. A vision system is required to detect symbolic spatial information in the world that matches items in the dictionary.

B. Vision system for wild text detection

In our vision system [14], designed for detecting door labels, we begin by applying a guided filter [15] to the input image from the camera as an edge enhancing technique. We use the minimally stable extremal regions detector to find potential text regions [6]. These regions are filtered using the stroke width transform [4], as well as weak geometric constraints such as minimum size and aspect ratio. The bounding box of each region is expanded horizontally to collect geometrically adjacent characters into words. We then perform character recognition. This is done using a convolutional neural network which has been trained on the computer font subset of the 74k dataset [16]. Our network (see Fig. 4) has an architecture similar to the one proposed in [17] with 4 layers, 2 of which are convolutional and 2 are fully-connected.

We trained the network for 25 epochs with small random affine distortions applied to the training set to improve generalization. We used backpropagation with an initial learning rate of 0.001, decreasing by a factor of 0.794 every 4 epochs. We then trained it for 3 additional epochs on undistorted images at a constant learning rate of 0.0002 to stabilize the weights. The network's performance on the training set is 3.3% error.

Not all text in the world corresponds to door labels. Text that does not match an item in the robot's room-name dictionary is ignored. When text corresponding to an item in the robot's room-name dictionary is detected (a symbol observation), a link between the symbol observation and the robot's map of the world is established. Each symbol observation comes with bearing and distance information calculated from the position of the text in the image from the robot's camera, the field of view of the camera, and laser range readings.

C. Symbolic navigation with an abstract map

In order for the robot to use the symbolic spatial information from the floor plan when navigating, a link needs to be established between this information and the world

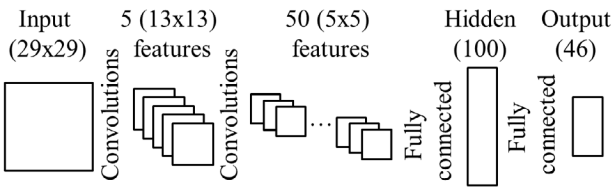


Fig. 4: Network architecture for character recognition

representation used by the robot via symbols observed in the world. The robot operates in its own coordinate frame and the symbolic spatial information from the abstract map is in the pixel coordinate frame of the floor plan. A scaled 2D coordinate transform is required to link between these two frames.

A 2D coordinate transform T is comprised of horizontal and vertical scaling, horizontal and vertical translation, and rotation (t_{s_x}, t_{s_y}, t_{t_x}, t_{t_y}, and t_θ respectively). Under the assumption that floor plans are scaled equally in both directions, as is the robot's coordinate system, the scaling can be combined into one parameter (t_s). The order was defined as rotate, scale, then translate.

For a piece of symbolic spatial information of the form 'a is at x, y in floor plan F' (denoted by F_{x_a}, F_{y_a}), the coordinate transform T to convert this to location a in the robot's coordinate frame R (denoted by R_{x_a}, R_{y_a}) is defined as:

$$\begin{bmatrix} R_{x_a} \\ R_{y_a} \\ 1 \end{bmatrix} = T(t_\theta \rightarrow t_s \rightarrow t_x, t_y) \begin{bmatrix} F_{x_a} \\ F_{y_a} \\ 1 \end{bmatrix} \tag{1}$$

As a homogeneous coordinate transform matrix, the transform operation can be expressed as:

$$\begin{aligned} \begin{bmatrix} R_{x_a} \\ R_{y_a} \\ 1 \end{bmatrix} &= \begin{bmatrix} t_s\cos(t_\theta) & -t_s\sin(t_\theta) & t_x \\ t_s\sin(t_\theta) & t_s\cos(t_\theta) & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} F_{x_a} \\ F_{y_a} \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} t_1 & -t_2 & t_3 \\ t_2 & t_1 & t_4 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} F_{x_a} \\ F_{y_a} \\ 1 \end{bmatrix} \end{aligned} \tag{2}$$

As the robot gathers more symbolic spatial observations, the system of linear equations relating places in the floor plan to those in the robot's map grows as follows:

$$\begin{bmatrix} R_{x_a} \\ R_{y_a} \\ \vdots \\ R_{x_N} \\ R_{y_N} \end{bmatrix} = \begin{bmatrix} F_{x_a} & -F_{y_a} & 1 & 0 \\ F_{y_a} & F_{x_a} & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ F_{x_N} & -F_{y_N} & 1 & 0 \\ F_{y_N} & F_{x_N} & 0 & 1 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{bmatrix} \tag{3}$$

This linear system of equations is solved using an ordinary least squares estimator to find the linear transformation. The estimates for these parameters are used to transform the symbolic spatial information from the floor plan to the robot's map. An estimate for these parameters can only be obtained once the robot has observed the door labels for two distinct locations (four equations for four parameters). Not all doors in the real world have labels that are used in the floor plan. However, the system works with small numbers of observations, with every additional door label observation refining the robot's estimate of the transform.

Once the abstract map has been transformed into the coordinate frame of the robot's world map, the coordinates for the destination can be set as the robot's goal. Standard methods for map building and path planning are used that allow a goal to be set beyond the edges of the current world map.

IV. Experimental Setup

A. Robot platform

The robot used in this paper to demonstrate and evaluate the system is a GuiaBot from MobileRobots, shown in Fig. 1. This robot is equipped with a spherical video camera, Ladybug2, which provides the robot with six 0.75M pixel $1/3''$ CCD sensors and a FireWire interface. The robot has four on-board computers all running Robotic Operating System, ROS.

B. Environments

Experiments were performed in two environments, both in simulation and in the real world. The simulated environments were created from cleaned versions of the floor plans, with no noise in mapping or localization, and a simulated door label detector providing symbol observations. The environments were Levels 11 and 4 of S Block at QUT Gardens Point Campus (see Fig. 5). Printed door labels were added to the environment to enable successful OCR using the Ladybug2 cameras.

C. Task

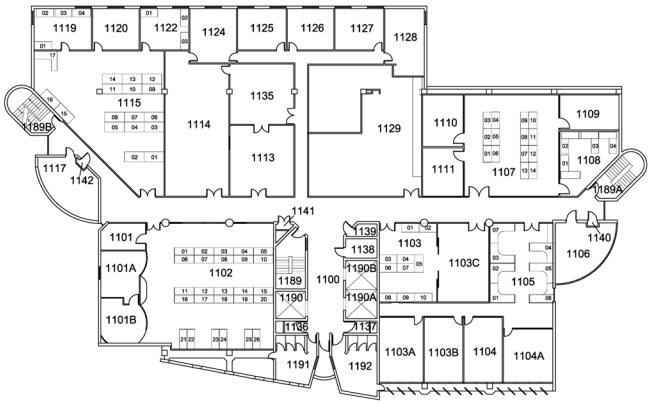
The task for the robot was to reach a destination, specified by a number, for example, '1105'. The number matched the room name on both the floor plan and the physical door to the room. Successful task completion was defined as the robot stopping outside the door to the room.

D. Experiment

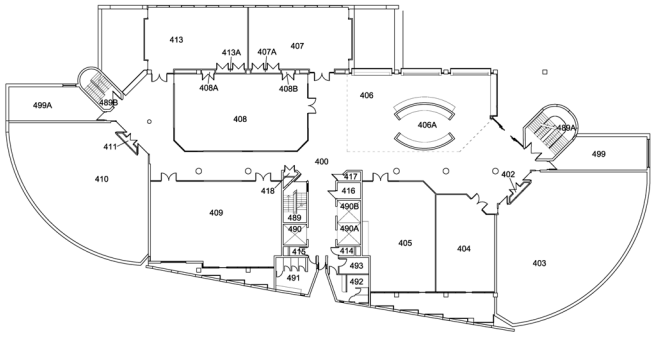
The system was tested in 10 trials in both simulated environments, and the results were confirmed by a single trial in both real-world environments. An abstract map and associated room-name dictionary constructed from a floor plan of the environment was provided to the robot prior to each trial. The robot was started at the same initial position and assigned a destination from the room-name dictionary. In the simulated environments, we compared our system with a left wall-follower and a right wall-follower using random exploration to attempt to reach the destination.

V. Results

In each of the ten trials in both of the simulated environments, our system successfully reached the destination, often after backtracking several times to find a path. In both trials in the real-world environments, our system successfully reached the goal location, using an efficient path after two door labels were detected.

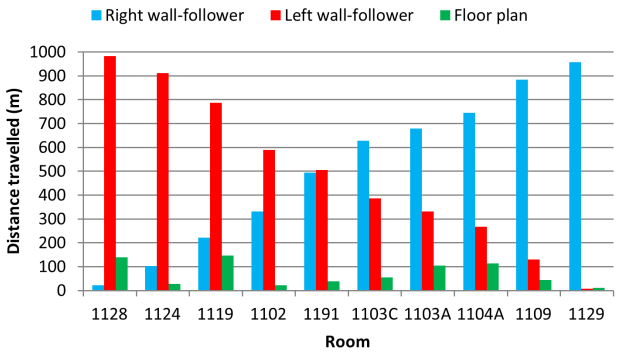


(a) QUT Gardens Point S Block Level 11

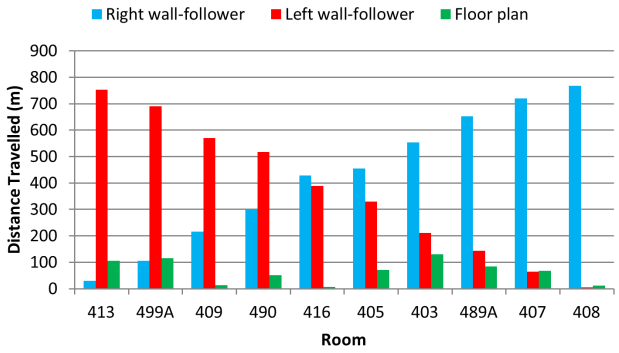


(b) QUT Gardens Point S Block Level 4

Fig. 5: Floor plans



(a) QUT Gardens Point S Block Level 11



(b) QUT Gardens Point S Block Level 4

Fig. 6: Total distance traveled to reach the goal for 10 trials in the simulated environments

A. *Simulation-World Experiments*

The results for the three different systems (left wall-follower, right wall-follower, and our floor-plan system) across the ten trials for both simulated environments (Level 11 and Level 4 of S Block) can be seen in Fig. 6. In the simulated Level 11 environment, the wall-following system took, on average, more than 6 times as long and the robot traveled approximately 7 times as far as with our system. The only cases where our system was outperformed by the wall-following system was when the goal location was the first room on its path. In the simulated Level 4 environment, the wall-following system took, on average, more than 6 times as long and the robot traveled approximately 6 times as far as with our system. In this environment, our system was outperformed by the wall-following system when the goal location was the first or second room on its path.

B. *Real-World Experiments*

Here we provide a detailed description of the trial on Level 11 of S Block, shown also in the accompanying video [1]. The robot was provided with an abstract map (see Fig. 7a) constructed from the floor plan of S Block Level 11 (see Fig. 5a). The robot was initially located in the lobby next to the data room '1138' and the destination was set to '1105'. The first door label, '1138', was detected as the robot spun around to an exploration goal behind its starting position. The robot continued to explore towards the bathrooms, '1191' and '1192', and detected the label '1136'. The robot then estimated an initial transform between the world map and abstract map, set a goal for the destination, and planned a path to the goal (see Fig. 7b).

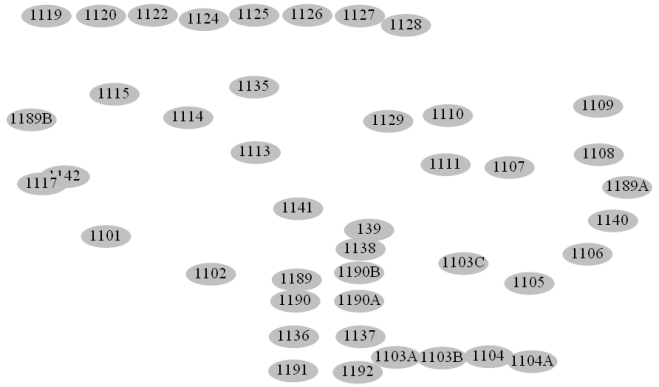
As the robot followed its path to the goal, it received more evidence in the form of door labels '1139' and '1107', updating its transform each time. The goal set by the robot was inside Room 1105 rather than at the door, due to the location of the label '1105' on the floor plan. The robot initially attempted to enter the door to room '1105' before detecting the door label and stopping as visual text recognition indicated that the goal had been reached (see Fig. 7c).

In the second environment, Level 4 of S Block, the robot was initially located in the lobby next to the data room '416', and the destination was set to '410'. The robot set an initial transform after detecting door labels for '416' and '417', and followed the path plan to '410' (see Fig. 8).

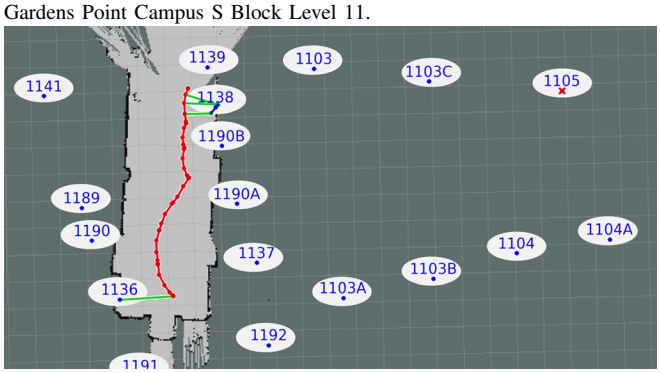
VI. Conclusions and Future Work

The system presented in this paper provides an architecture for thinking about how to use symbolic spatial information to aid navigation. These studies demonstrated the utility of a symbolic goal-directed exploration system in two real-world environments. Our system can use an embedded graph created from information in a metric floor plan to plan

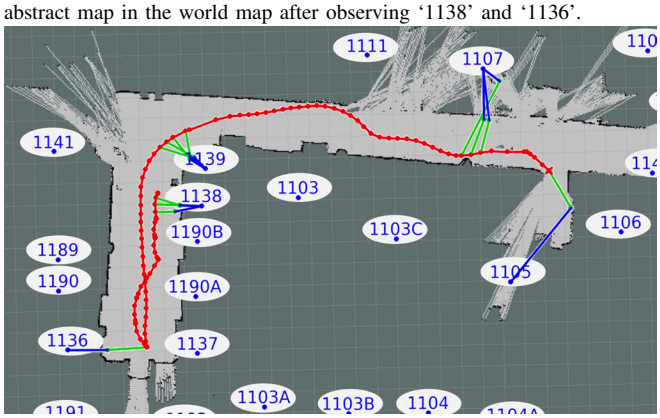
[1]The video is available at http://tinyurl.com/HumanCues



(a) Visualization of the abstract map constructed from a floor plan of QUT Gardens Point Campus S Block Level 11.



(b) The robot explored the world and estimated the transform to ground the abstract map in the world map after observing '1138' and '1136'.



(c) After more evidence ('1139' and '1107'), the robot updated the transform. As the robot neared the goal coordinate estimated from the abstract map transform, it observed '1105' and stopped at the destination.

Fig. 7: Real-World Experiment in S Block Level 11

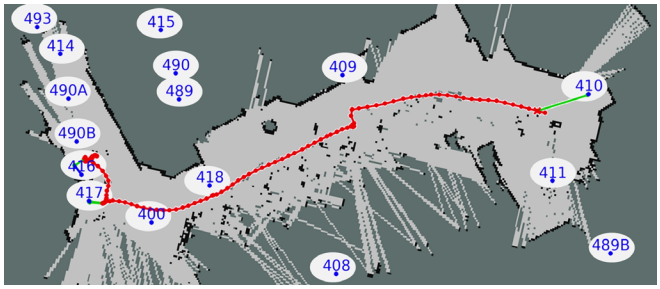


Fig. 8: Real-World Experiment in S Block Level 4: The robot started near the lifts, set an initial transform after detecting '416' and '417', and successfully reached '410'

efficient paths to the rooms shown in the floor plan.

The proof-of-concept system presented in this paper will become more useful as we extend it to different types of potentially conflicting information that humans typically use for navigating the world, including non-metric information; different types of potentially noisy symbol observations from the robot's cameras; and different methods of navigation.

References

[1] B. Tversky, "Structures of mental spaces: How people think about space," *Environment and behavior*, vol. 35, pp. 66–80, 2003.

[2] J. Wang and R. Li, "An empirical study on pertinent aspects of sketch maps for navigation," in *11th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE, 2012, pp. 130–139.

[3] I. Posner, P. Corke, and P. Newman, "Using text-spotting to query the world," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2010, pp. 3181–3186.

[4] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 2963–2970.

[5] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3538–3545.

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[7] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "PhotoOCR: Reading text in uncontrolled conditions," in *International Conference on Computer Vision (ICCV)*. IEEE, 2013, pp. 785–792.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] R. Smith, "An overview of the Tesseract OCR engine." in *9th International Conference on Document Analysis and Recognition (ICDAR)*, vol. 7, 2007, pp. 629–633.

[10] A. Bauer, K. Klasing, G. Lidoris, Q. Mühlbauer, F. Rohrmüller, S. Sosnowski, T. Xu, K. Kühnlenz, D. Wollherr, and M. Buss, "The autonomous city explorer: Towards natural human-robot interaction in urban environments," *International Journal of Social Robotics*, vol. 1, no. 2, pp. 127–140, 2009.

[11] M. R. Walter, S. Hemachandra, B. Homberg, S. Tellex, and S. Teller, "Learning semantic maps from natural language descriptions," in *Proceedings of Robotics: Science and Systems*, Berlin, Germany, 2013.

[12] J. Fasola and M. J. Mataric, "Using semantic fields to model dynamic spatial relations in a robot architecture for natural language instruction of service robots," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2013, pp. 143–150.

[13] T. T. F. Kollar, "Learning to understand spatial language for robotic navigation and mobile manipulation," Ph.D. dissertation, Massachusetts Institute of Technology, 2011.

[14] O. Lam, F. Dayoub, R. Schulz, and P. Corke, "Text recognition approaches for indoor robotics: a comparison," in *Australasian Conference on Robotics and Automation*. ARAA, 2014.

[15] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Computer Vision–ECCV*. Springer, 2010, pp. 1–14.

[16] T. de Campos, B. R. Babu, and M. Varma, "Character recognition in natural images," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2009.

[17] P. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR*, 2003, pp. 958–963.