Following and Interpreting Narrated Guided Tours

Sachithra Hemachandra, Thomas Kollar, Nicholas Roy and Seth Teller

***Abstract*— We describe a robotic tour-taking capability enabling a robot to acquire local knowledge of a human-occupied environment. A tour-taking robot autonomously follows a human guide through an environment, interpreting the guide's spoken utterances and the shared spatiotemporal context in order to acquire a spatially segmented and semantically labeled metrical-topological representation of the environment. The described tour-taking capability enables scalable deployment of mobile robots into human-occupied environments, and natural human-robot interaction for commanded mobility.**

Our primary contributions are an efficient, socially acceptable autonomous tour-following behavior and a tour interpretation algorithm that partitions a map into spaces labeled according to the guide's utterances. The tour-taking behavior is demonstrated in a multi-floor office building and evaluated by assessing the comfort of the tour guides, and by comparing the robot's map partitions to those produced by humans.

I. INTRODUCTION

Widespread adoption of robotic technology in human-occupied environments such as hospitals, workplaces and homes will require natural human robot interaction with unskilled people. Because robots employing standard technologies for metric mapping and navigation use environmental representations different from the spatial and topological representations used by humans [1], a robot will also need to acquire "local knowledge" of the environment (e.g. the names of places and descriptions of what people do at those locations). For example in order to understand the command "Take me to the third-floor kitchen," the robot must gain a representation of where "kitchens" are located in a particular environment.

This paper considers the problem of acquiring local knowledge by enabling a robot which has been newly introduced to an environment to acquire a semantic representation. One approach to acquire semantics about the environment is to have a human operator manually control the robot with a joystick. The operator then runs an algorithm to construct a map and manually segments and names each of the spaces. Because this approach requires a high level of technical involvement with the system, it is not feasible for large-scale deployments. An alternative approach is to have the robot autonomously explore its environment, detecting and classifying objects (through sensing and perception) and inferring generic space types (e.g., the presence of a TV might imply a living room). However, even with robust real-time object recognition, the robot would have no notion of which spaces are important to humans, or of the space names used by humans.

Hemachandra and Kollar are PhD candidates in the Computer Science and Artificial Intelligence Laboratory; Roy is faculty in the Department of Astronautics and Aeronautics; and Teller is faculty in the Department of Electrical Engineering and Computer Science, at the Massachusetts Institute of Technology, Cambridge, MA, 02139; United States. Email: {sachih, tkollar, nickroy, teller}@mit.edu

Fig. 1. Wheelchair robot in tour-guide mode

In this work, we present a scalable and flexible approach where the robot learns the environment by using human supervision. The human conducts a narrated guided tour of the new environment, describing salient locations and objects verbally, while the robot follows. The robot acquires both metrical and topological representations of the environment during the tour, and reasons about the location of semantic labels. By giving the robot a narrated guided tour in order to learn semantic information about the environment, our approach combines the advantages of the human-controlled and fully autonomous approaches above: human guidance during the tour phase provides for natural interaction and human assistance grants access to otherwise inaccessible portions of the environment. The autonomous nature of the tour-following removes the need for users to have technical knowledge of the system, or to manually encode semantic information.

There are two key challenges involved in realizing an effective tour-following and interpretation capability. First, the robot must follow the human tour guide in a socially acceptable way. Second, the robot must segment the acquired free-space map into spaces that are meaningful to humans and infer a label (name) for each space. We describe a socially acceptable person-following algorithm capable of multi-floor tour following and demonstrate a method that can bridge the gap between a robotic representation of the environment (used for the purpose of localization and navigation) and representations employed by human occupants (involving functional spaces and space names). Our method effectively adapts the mobile robot to the toured environment so that it can perform high-level speech-based navigation. We implement our solution on an autonomous wheelchair (shown in Figure 1), and demonstrate a system capable of following and interpreting a tour spanning multiple floors. A video describing the system capabilities can be found at [2].

II. RELATED WORK

Tour interpretation involves methods for simultaneous localization and mapping (SLAM), human-computer inter-

action (HCI), person following, spoken dialog management, map segmentation and human-augmented mapping.

Our tour-following capability relies on two significant components, namely person tracking and person following. Person tracking has been explored using a variety of sensors, such as vision [3] [4] and lidar [5], [6] and tracking methods based on extended Kalman filters, sample-based joint probabilistic data association filters (SJPDAF) [6], particle filters [5] and multi-hypothesis tracking (MHT) [7]. We use lidars to avoid the vulnerability of current vision-based approaches to background color, lighting conditions, and the guide's orientation; we use particle filters for tracking, similar to the implementation by Kirby et al. [5].

Published person-following methods fall into three main categories: following the path taken by the human; following in the direction of the human; and following side-by-side with the human. Kirby et al. [5] found that subjective evaluation favored following the person's current position. However, they did not explore which behaviors are appropriate in different situations, e.g. what action to take based on the robot and person's current configuration in the environment. Walters et al. [8] found that most people were comfortable with a robot occupying their personal zone (0.45-1.2m from the person) and social zone (1.2-3.6m) defined by normal human-human interactions while a significant minority were comfortable with a robot occupying an intimate zone (0.15-0.45m). Our method uses a variant of pure pursuit [9] to follow the guide's most recent heading, while modulating the human-robot gap based on the nature of the environment and the location of the robot with respect to the guide's spatial zones.

SLAM methods are essential for tour interpretation, as one outcome of the tour must be a map. Many SLAM methods have been proposed [10], using e.g. EKF methods [11], FastSLAM [12] and pose graphs [13]. We use a real-time implementation of iSAM [14] modified to produce multiple inter-linked planar maps from multi-floor tours.

Spatial and topological structure have been extracted from metrical maps using Voronoi Random Fields [15], where metrical maps are segmented into functional spaces such as corridors and rooms, using Ada Boost [16] and spectral clustering [17]. However, these methods rely only on spatial characteristics for segmentation, while the human notion of space is not limited to spatial features. Our method generalizes spatial segmentation methods to incorporate and exploit semantic information.

Previous work on human-augmented mapping combined human and robot spatial representations, with experiments in limited spatial settings [18], [19]. However, this work focused more on map acquisition and segmentation and did not elaborate on the person-following aspect of tour-giving. The method of Spexard et al. [19] produced exactly one cluster for each acquired label, under-segmenting when spaces are visited but not labeled. Zender et al. segmented the navigation map based solely on door traversals, under-segmenting functionally or semantically distinct spaces not separated by doorways.

III. Narrated Guided Tour Following and Interpretation

A central aspect of our method is the association of spatial and semantic information through "labeling events" that occur during a tour. The meaning of such events depends on how the guide and robot are situated with respect to one another and the space being described. Our method assumes labeling events in which the labeled space is occupied by the guide (e.g.,"I am in the kitchen"), the robot (e.g., "You are in the lounge"), or both. More complex ways in which people label the environment, such as by reference to adjoining spaces (e.g.,"On my right is the kitchen"), or distant spaces (e.g.,"The elevator is down the hall and around the corner") are outside the scope of this paper but relevant for future work. A possible approach to understanding these complex descriptions is to use Kollar et al.'s spatial description clauses [20]. The structure of the system is as follows:

- **Robot Base:**
  - Raw laser and odometry information
  - Drives the robot to a commanded rotational and translational velocity.
- **Tour Following:**
  - Person Tracking: Locates the guide using a laser rangefinder.
  - Person Following: Controls the robot to move the robot toward a guide.
- **Tour Interpretation:**
  - SLAM: Constructs real-time multi-floor maps during the guided tour.
  - Dialog Manager: Recognizes user utterances and extracts semantic labels.
  - Map Segmentation: Partitions the acquired metrical map, incorporating semantic labels.

IV. Person Following

In order for someone to successfully give a narrated guided tour to the robot, the robot must be able to follow the guide in a manner that is both efficient and socially acceptable. While people do not treat robots like humans, humans do expect certain social behaviors from robots [21]. Since one of our goals is to facilitate deployment of robots to new environments by non-technical people, behavior that conforms to their expectations is critical for success.

Our person tracker uses two lidars to extract the location of the legs of the person and a particle filter to track these leg observations. The combined field of view of the two lidars provides the robot with a 360° view of its environment. Our tracker is similar to that of Kirby et al. [5], though our initialization methods differ. To initialize the following behavior, the guide verbally indicates that they are in front of the robot. The robot then selects this person as the tour guide, picking the closest tracked target, and verbally indicating that it is tracking. The guide then starts the tour by saying "Follow me," and can pause it by saying "Stop."

The person-following algorithm attempts to follow the guide in a human-friendly manner, by moving the robot toward the guide's current position subject to constraints arising from the guide's personal spatial zones [8] and nearby

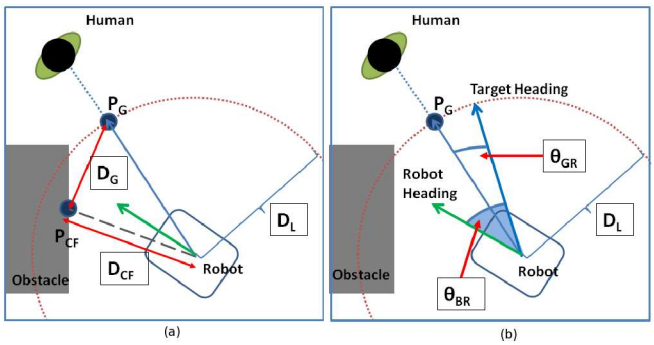

Fig. 2. Parameters used to compute desired heading: (a) parameters used for the score calculation (for an example bucket) (b) Parameters used for the velocity calculation

environmental features. The following spatial zones limit the robot behavior.

- **Intimate Zone:** In this zone (<0.45m from the guide), the robot makes no motion.
- **Personal Zone:** Between 0.45m and 1.2m from the guide, the robot either rotates in place, or exhibits full motion, depending on context.
- **Followable Zone:** When the robot is more than 1.2m from the guide moves toward the person.

These zone definitions arise from a study in which a human approached a stationary robot [8]. By observing the above zones, the algorithm aims to make the tour following behavior safe and socially acceptable. In addition to adhering to the notion of spatial zones, velocity commands are reduced when the tour guide is close to the robot or obstacles in the immediate vicinity of the robot. When the person is further away, the person following algorithm uses the location of the guide to calculate the translational and rotational velocities that move the robot toward the human.

In order to compute a rotational and translational velocities, the person following algorithm first extracts the obstacles in the environment using the lidars, and then selects a *target heading* based on the free space in each possible heading and the location of the goal (Figure 2). It then calculates the required velocities needed to move toward the target heading. The following outline in detail how this process is carried out.

1) Obstacle Extraction: The algorithm extracts the location of obstacles by registering and discretizing the field of view of the laser.

- **Registration**: Data from both lidars is transformed to the robot frame of reference.
- **Discretization of field of view**: The possible headings $\theta \in [-90°, +90°]$ are discretized into *buckets* of $2.5°$ in resolution. For each bucket, the maximum collision-free point (P_{CF}) is calculated (up to the look-ahead distance D_L) using the registered obstacle points (see Figure 2(a)).

2) Target Heading Selection: The system then computes a heading based on the environment and location of the tour guide:

- **Goal Heading and Location**: The goal point (P_G) is placed at the look-ahead distance D_L (2m) in the direction of the guide location.
- **Scoring each bucket**: Each bucket is scored based on the following formula: Where D_{CF} is the distance from the robot center to P_{CF}, and D_G is the distance from P_{CF} to the goal point P_G.

$$score[bucket] = D_{CF}/[1 + D_G^2] \tag{1}$$

- **Selection:** Once scores are calculated for the entire field of view, they are averaged using a moving window of 5 segments on each side. This averaging step significantly reduced instances of corner cutting when traveling around bends. Otherwise, moving directly toward the guide would cause the robot to travel very close to the wall, increasing the likelihood of occlusion. The heading from the bucket with the highest average score is selected as the target heading.

3) Rotational Velocity: The rotational velocity RV is computed from the difference between the target heading and the robot's current heading (θ_{BR}). The angle difference is clamped in Equation 2 to prevent rapid rotate-in-place maneuvers, where RV_M is the maximum rotational velocity:

$$RV = \frac{\text{clamp}\left(-(\pi/4), \theta_{\text{BR}}, (\pi/4)\right)}{(\pi/2)} \times RV_M \tag{2}$$

If θ_{BR} is larger than 1.4 radians, the robot turns in place until it is facing the goal heading. This behavior prevents the robot from doing large arcs when it can turn in place within a shorter time. This type of situation occurs when the guide goes around a sharp corner, or turns back and walks toward and past the robot.

4) Translational Velocity: The translational velocity TV is calculated primarily based on the the difference in angle between the robot heading and guide direction (in Equation 3). The ratio R_{SO} (in Equation 4) reduces the translational velocity when there are nearby obstacles. The ratio R_{DO} (in Equation 5) reduces the translational velocity when there are nearby obstacles in the target heading.

$$R_{AD} = \frac{||\theta_{\text{GR}}| - \theta_{\text{BIAS}}|}{\theta_{\text{BIAS}}} \tag{3}$$

$$R_{DO} = min\left(D_{DO}/D_L, 1.0\right) \tag{4}$$

$$R_{SO} = min\left(D_{SO}/0.6, 1.0\right) \tag{5}$$

$$TV = R_{SO} \times R_{DO} \times R_{AD} \times TV_M \tag{6}$$

Here, θ_{GR} is the difference in angle between the goal and the robot heading, D_L is the look-ahead distance (2.0m), D_{DO} is the average collision-free distance in target heading, D_{SO} is the closest side gap, and TV_M is the maximum translational velocity (see Figure 2(b)).

V. Tour Interpretation

The purpose of the tour interpretation subsystem is to acquire metric, topological and semantic representations of the traversed environment during the tour. The metric representation is a set of gridmaps (one for each floor traversed) denoting an occupancy probability for each discretized (x, y) location. In this work, the iSAM [14] algorithm was used to perform multi-floor SLAM. This results in a set of metrical

maps, one for each floor visited. The topological representation is a set of nodes, one corresponding to each spatial region such that each free space gridcell in the metrical map is assigned to a particular region. Semantic information, such as the names of locations, is attached to the nodes of the topological representation of the map.

A dialog management mechanism extracts semantic labels from user utterances by continuously detecting occurrences of speech, and using a speech recognizer to extract location and object labels. We use the SUMMIT speech recognizer [22], configured with a domain-specific context-free grammar that captures the structure of language in the tour-guide domain:

```
<loc_tagging> = <perspective> <prop> <place>
<perspective> = i am [view=guide] |
                you are [view=robot] |
                we are [view=both]
<prop>        = (in | at) [pos=at] | near [pos=near]
<place>       = [a | an | the]
                (kitchen | office | lounge | ....)
```

Each labeling utterance is analyzed to determine its salient information:

- **Perspective of utterance**: A location can be labeled from the perspective of the tour guide (e.g. "I am now in the lounge") or from the perspective of the robot (e.g. "You are now near the restroom"). The metrical location of the label is extracted from the current position of the guide or the robot depending on this perspective.
- **Relative location of the labeling event**: A labeling can happen either inside the space or nearby. Spaces labeled using the keywords "in" or "at" are considered to be inside a space, and are incorporated into the segmentation process. Spaces (resp. objects) labeled using "near" do not affect the segmentation.
- **Place/object name**: The semantic label acquired from the labeling event.

After the tour ends, the map segmentation algorithm uses the metric and semantic information to create a spatial representation of the environment. The segmentation is achieved by building on the map segmentation algorithm developed by Brunskill et al. [17]. Given an occupancy grid map, the algorithm induces a graph over the sampled points by introducing edges between points that share line-of-sight visibility. It then constructs a similarity matrix from the distances between connected points. The graph is then partitioned by spectral clustering in order to maximize intra-cluster connectivity and minimize inter-cluster connectivity.

Our insight is that the addition of semantic labels to the segmentation process can enable a better partitioning of the environment. For example, under the normal spatial segmentation algorithm, an entire corridor will be segmented as a single space. However, in some environments humans will perceive the hallway as a set of separate regions (e.g. wheelchair charging stations). In such situations, these hallways should be segmented into different logical regions if they contain such semantic sub regions. In addition, large open spaces that contain complex furniture and other structures tend to get over segmented compared to the human perception of this space.

To address these issues, we create a reduced graph that takes into account the semantic information in order to perform map segmentation. This reduced graph uses ray-tracing to assign each sample to a semantic label, and only connects the sample to other samples that were assigned to the same label (instead of all of the close neighbors in space). When no semantic label is available, the graph is generated as before, connecting samples by line of sight visibility. By formulating the graph in this way, the reduced graph tends to create stronger connections to nearby semantic labels, thereby creating a more intuitive segmentation of the environment. After creating a similarity matrix (where points with shorter distances have higher values), the graph is segmented using spectral clustering [17]. The result is a set of points (representing the entire map) segmented into a small number of regions.

VI. Performance

In order to validate that the system provides a socially acceptable and efficient tour, can be used for multi-floor guided tours, and that it segments metrical maps more robustly than prior approaches, we carried out a series of experiments on a speech-commandable autonomous wheelchair platform (Figure 1).

A. Socially Acceptable Behavior

1) Procedure: In order to understand the tour guide's perception of the overall behavior of the robot, its ability to keep up, his/her comfort levels regarding the different behaviors, and perception of the tour-guide behavior, we performed three trials with five subjects per trial. Trial 1 (T1) used our notion of adhering to different personal boundaries. Trial 2 (T2) kept a small (0.3m) fixed standoff from the person, and trial 3 (T3) kept a large standoff (2.0m) from the person. The subjects were not informed about the different behaviors. Subjective feedback was obtained by scoring the above described criteria.

2) Results: The results of the subjective evaluations are given in Table I. The standard deviations are given in parenthesis.

Criterion	T1	T2	T3
Overall Preference (1-worst 10-best)	7.2 ($\pm$1.1)	7.0 ($\pm$2.5)	3.6 ($\pm$2.5)
Ability to keep up (1-poor 10-perfect)	8.0 ($\pm$1.9)	8.6 ($\pm$1.7)	2.6 ($\pm$1.5)
Comfort w/ closeness (1-aggressive 10-timid)	5.5 ($\pm$1.7)	2.5 ($\pm$1.3)	8.5 ($\pm$0.6)
Starting Distance (1-close 10-far)	5.6 ($\pm$0.9)	4.6 ($\pm$0.5)	8.6 ($\pm$0.5)

TABLE I

Subjective responses to person-following behaviors

Overall our method (T1) was rated slightly higher (mean 7.2$\pm$1.1) than T2 (mean 7.0.$\pm$2.5), and significantly higher than T3 (mean 3.6$\pm$2.5). Regarding comfort with closeness, our method was rated close to the best, while subjects in

T2 found the robot to be too aggressive and T3 found the robot to be too timid. Both T1 (mean 8.0±1.9) and T2 (mean 8.6±1.7) scored highly in the ability to keep up. Subjects felt that both methods T1 and T2 were close to the ideal starting distance while T3 was felt to start following when the person was too far away.

As a part of the studies that we performed, we found that the ideal gap between the person and the robot should not be static and that subjects' comfort levels revolved around a combination of the speed of the robot (relative to the person) and its distance from them. For example, while the subjects felt uncomfortable with the robot closely approaching them at higher speeds, slow approach was always tolerated. In addition, subjects' comfort levels change based on their confidence in the the system. Most subjects remarked that even during T2 (closest approach trial), they were comfortable with the approach once they were confident of its ability to stop. However, all subjects were familiar with robots and therefore might have higher comfort levels with such systems. Since in our method (T1), the robot started to decelerate earlier than T2, the participants felt more comfortable the system. Subjects' preference regarding the starting distance of the robot appears to be different than the stopping distance. Since in the second trial (T2), the robot started to move as soon as the person was farther than 0.3m away, it was best at keeping up with the guide (and therefore ranked highest in that category). This was also a contributing factor to most subjects' evaluation of overall performance.

B. Efficiency in Conducting Multi-floor Tours

1) Procedure: Since our system is aimed at acquiring environment information based on guided tours, four subjects conducted multi-floor tours around the Stata building on the MIT campus. Two types of tours were conducted: one in which the wheelchair followed the tour guide autonomously, and the other in which a person drove the wheelchair to follow the guide.

To ensure consistency, a predetermined route was taken by the tour guide, and the same set of spaces were described by the subjects. The route spanned two floors, which included two elevator transits (beginning and concluding on the same floor). The tour guides were instructed regarding the typical grammar recognized by the system. All modules other than the map partitioner were run on-line. We measured the time taken in both tours, and the average speeds of the wheelchair (when moving) under the two methods.

2) Results: Table II summarizes the results of the multi-floor tour following trials. During all multi-floor trials, the SLAM module constructed consistent multi-floor maps. While all tagged locations outlined at the start of the trial were in the speech recognizer vocabulary, some out-of-grammar utterances were not recognized.

On average the autonomous person following took 1.2 mins or 7% longer to complete the same route. Noticeable delays in the autonomous tours were caused by occasional person tracking failures and mis-recognized "stop" commands. The robot lost track of the guide in one instance when exiting the elevator (out of 8 elevator exits in total). In all tracking failures, the tour guide was able to restart the tracking by standing in front of the wheelchair and communicating that fact. During the autonomous tours, the robot lost track of the guide on average 2.25 times per tour (283m was traveled on average between tracking failures). Compared to this, the person tracker performed much better when the wheelchair was manually driven behind the guide (on average 0.5 tracking failures per tour). All tours were completed autonomously and there were no wheelchair-guide collisions.

Scenario	Time	Distance	Speed
Autonomous	17.9 min	424 m	0.58 m/s
Manual	16.7 min	436 m	0.64 m/s

TABLE II
GUIDED TOUR RESULTS

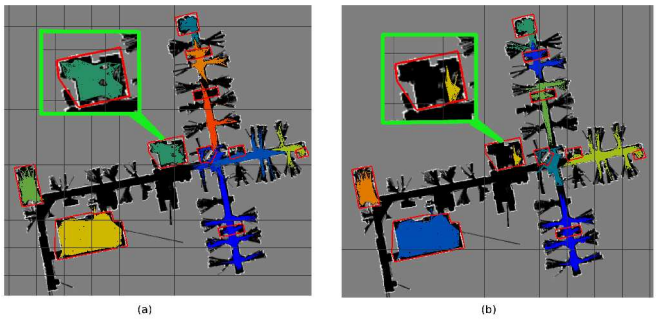

Fig. 3. Map Segmentation TBH First Floor (ground truth polygons are shown in red): (a) Semantic segmentation result (b) basic segmentation results.

C. Map Segmentation Performance

1) Procedure: We asked one of the tour guides to segment the locations that she had tagged during the tour on a map by outlining the bounding boxes for each region. We compared this with the segmented maps created by both the basic method [17] and our semantic segmentation method. In all, we compared the different results from three floors from two different buildings. The comparison was done by calculating the overlapping regions (from the respective algorithm and the human segmentation) and dividing by the combined region covered by the two methods, which assigns a value of 100% only when both segments are in agreement, but less than that otherwise. The region area was approximated using the number of sample points generated by the segmentation algorithm (from the RRG).

$$Accuracy = \frac{\text{Ground truth Region} \cap \text{Assigned Region}}{\text{Ground truth Region} \cup \text{Assigned Region}}$$

In addition, to evaluate the sensitivity of the map segmentation to the location of the tagging event, we generated four random samples each, from within the ground truth space for each labeled location. We used these to simulate different tagging locations and ran the segmentation algorithm on these points.

2) Results: Table III summarizes the results of the semantic segmentation evaluation using labels acquired during actual guided tours. We have also compared the performance

of the two algorithms on tagged locations acquired during actual tours. The semantic segmentation method appears to improve the results in all floors. Floors such as Stata-3 had fewer labeled locations, which might indicate the reason for lower gains. Overall, the ground truth accuracy tends to vary depending on the type of space considered, for example locations with high spatial structure tend to have very high accuracies with the ground truth, where as spaces such as charging stations which are less spatially pronounced tend to score lower.

Floor	Accuracy [17]	Accuracy (Semantic)
Stata-3	62.8 %	68.9%
TBH-1	39.7 %	47.7 %
TBH-2	54.4 %	60.3%

TABLE III
SEGMENTATION RESULTS (ACTUAL TOUR)

In Table IV, we have compared prior work to our algorithm by synthesizing label locations within annotated regions to simulate different areas that users might use to label a region. We show between 5% and 10% improvement over previous work.

Floor	Accuracy [17]	Accuracy (Semantic)
TBH-1	39.7 %	45.1 %
TBH-2	54.4 %	64.2 %

TABLE IV
SEGMENTATION RESULTS (PERTURBED LABELS)

Figure 3 shows a comparison of the basic and the semantic segmentation, and highlights a situation where a space that was over segmented in the basic algorithm is correctly segmented using semantic information. The map segmentation performance shows that even though our semantic based segmentation manages to segment all labeled locations, the boundaries in certain spaces do not conform greatly with the human definitions. This is most prominent for spaces lacking in spatial structure (e.g. wheelchair charging stations along a corridor at TBH).

VII. CONCLUSION

We have demonstrated an end-to-end solution on a robotic wheelchair that facilitates scalable, user-friendly deployment of a robot system into new environments. Our robotic platform is a speech-capable system with the capacity to interpret and implement simple high-level navigation commands. In addition, the wheelchair is able to follow a narrated guided tour from a human and learns a metric, topological and semantic map of the environment. By creating flexible user-friendly systems, we hope to improve the acceptance of complex robotics systems in everyday human environments, and improve the quality of life of people in assisted-living environments.

VIII. ACKNOWLEDGMENTS

We are grateful to Abraham Bachrach and Dr. Bryan Reimer at MIT, and to Don Fredette at The Boston Home.

REFERENCES

[1] T. P. McNamara, "Mental representations of spatial relations," *Cognitive Psychology*, 1986.
[2] "Narrated Guided Tour Demonstration." `http://rvsn.csail.mit.edu/wheelchair/tourguide.mov`.
[3] Z. Chen and S. Birchfield, "Person following with a mobile robot using binocular feature-based tracking," in *IROS*, pp. 815–820, Oct. 2007.
[4] C. Schlegel, J. Illmann, K. Jaberg, M. Schuster, and R. Wotz, "Vision based person tracking with a mobile robot," in *Proceedings of the Ninth British Machine Vision Conference (BMVC)*, pp. 418–427, 1998.
[5] R. Kirby, J. Forlizzi, and R. Simmons, "Natural Person-Following Behavior for Social Robots," in *HRI*, pp. 17–24, March 2007.
[6] E. Topp and H. Christensen, "Tracking for following and passing persons," in *IROS*, pp. 2321–2327, Aug. 2005.
[7] M. Luber, G. D. Tipaldi, and K. O. Arras, "Place-dependent people tracking," in *ISRR*, 2009.
[8] M. L. Walters, K. Dautenhahn, R. te Boekhorst, K. L. Koay, C. Kaouri, S. Woods, C. Nehaniv, D. Lee, and I. Werry, "The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment," in *HRI*, pp. 347–352, 2005.
[9] R. C. Coulter, "Implementation of the Pure Pursuit Path Tracking Algorithm," Tech. Rep. CMU-RI-TR-92-01, Robotics Institute, Pittsburgh, PA, January 1992.
[10] H. Durrant-Whyte and T. Bailey, " Simultaneous Localisation and Mapping (SLAM): Part I The Essential Algorithms," *IEEE Robotics and Automation Magazine*, 2006.
[11] R. Smith and P. Cheeseman, "On the Representation and Estimation of Spatial Uncertainty," in *IJRR*, vol. 5, pp. 56–68, 1986.
[12] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM 2.0: An Improved Particle Filtering Algorithm for Simultaneous Localization and Mapping that Provably Converges," in *IJCAI*, pp. 1151–1156, 2003.
[13] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental Smoothing and Mapping," *Robotics, IEEE Transactions on*, vol. 24, no. 6, pp. 1365–1378, 2008.
[14] B. Kim, M. Kaess, L. Fletcher, J. Leonard, A. Bachrach, N. Roy, and S. Teller, "Multiple relative pose graphs for robust cooperative mapping," in *ICRA*, (Anchorage, AK), pp. 3185 –3192, May 2010.
[15] S. Friedman, H. Pasula, and D. Fox, "Voronoi random fields: Extracting the topological structure of indoor environments via place labeling," in *IJCAI*, 2007.
[16] O. M. Mozos, C. Stachniss, and W. Burgard, "Supervised Learning of Places from Range Data using AdaBoost," in *ICRA*, pp. 1730–1735, Apr. 2005.
[17] E. Brunskill, T. Kollar, and N. Roy, "Topological Mapping Using Spectral Clustering and Classification," in *IROS*, pp. 3491–3496, 2007.
[18] H. Zender, P. Jensfelt, O. M. Mozos, G.-J. M. Kruijff, and W. Burgard, "An integrated robotic system for spatial understanding and situated interaction in indoor environments," in *AAAI*, pp. 1584–1589, 2007.
[19] T. Spexard, S. Li, B. Wrede, J. Fritsch, G. Sagerer, O. Booij, Z. Zivkovic, B. Terwijn, and B. Krose, "BIRON, where are you? Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization," in *IROS*, pp. 934–940, Oct. 2006.
[20] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *HRI*, pp. 259–266, 2010.
[21] E. Pacchierotti, H. Christensen, and P. Jensfelt, "Human-robot embodied interaction in hallway settings: a pilot user study," in *ROMAN*, pp. 164–171, Aug. 2005.
[22] J. Glass, "A Probabilistic Framework for Segment-Based Speech Recognition," in *Computer Speech and Language 17*, pp. 137–152, 2003.