A Minimal Representation of a Map to Localize and Navigate through an Indoor Environment

1. Introduction

Localization and Navigation in an indoor environment are one of the basic problems in the area of robotics. Most of the recent SLAM or SfM based mapping systems use low level features (generally SIFT or SURF features) to infer the 3D structure of the environment and estimates it's location with respect to the inferred 3D structure. The size of the map becomes large with the increase in the size of the environment. Such heavy maps cannot be loaded into robots having a low memory constraint and therefore estimating a memory efficient map representation is interesting in these scenarios. In this paper we explore some techniques to estimate a minimal representation of a map to navigate and localize through an indoor environment without much loss of accuracy.

There are two methods to navigate from one place to another, behavior/affordance based navigation (Figure 1) and map based navigation (Figure 2). In the case of behavior/affordance based navigation, we do not create a map of the environment and do not do localization and path planning. It is based on the belief that there exist a particular behavior based solution to a particular navigation problem. For example, to navigate from room A to room B one can design a left wall follower and a room B detector [25]. The decision in this case is made on the all possible actions available in the local environment and then that action is performed. Another way to navigate is through map based navigation. Localization is a necessary building block for navigating a robot from one room to another using map based navigation. Through localization a robot can identify its current position, plan its path through the map and detect if is has reached the goal. However map based navigation requires continuous localization of a robot to identify its location with respect to the map.

We believe that a proper balance between the map based navigation and behavior based navigation can lead to a minimal representation of a map by filling the gaps using local behavior based navigation. A better behavior based navigation requires a higher spatial understanding of the environment. For example, if a robot can understand the local structure of the room and recognize the gates and corridors then each node in the map can represent a particular location and the robot can navigate through the map using the local behavior based navigation by recognizing gates and corridors to reach the destination. Using topological map to represent the map can be particularly useful where each node represents an entity (bag of features, 3D structure etc.) used to localize a robot in the global map and edge connect one node to another which is navigated using behavior based navigation.

Below we discuss the requirements of behavior/affordance based and map based navigation techniques. Holistic spatial understanding of the environment is a pre-requisite for behavior based navigation tasks. Including semantic spatial knowledge can greatly enhance the performance of robots by providing a more meaningful representation for performing complex tasks. Semantic representation of the environment leads to the formation of spatial primitives each having specific properties and which excites the desired set of behaviors from the robot. Analogously if the robot can understand higher spatial primitive then the map can be built using these spatial primitives itself. This results in memory efficient techniques for building a map where we represent landmarks using higher level spatial primitives instead of lower level features. In this paper, we explore contemporary works in the relevant area of high level scene understanding for behavior based navigation and the map representations which can encode these features. We also propose a technique to create a minimal representation for a map to localize and navigate through an indoor environment.

2. Related Work

First of all we discuss the related work in the area of scene understanding required for behavior based navigation. Holistic scene understanding using images and videos which has been studied by the computer vision community and to use that along with SLAM/SfM system which is being explored by the robotics community to produce a rich semantic map of the environment. Better understanding of the environment directly affects the map representation used to navigate through it. In the next section we discuss the possible map representations and the related work in that area.

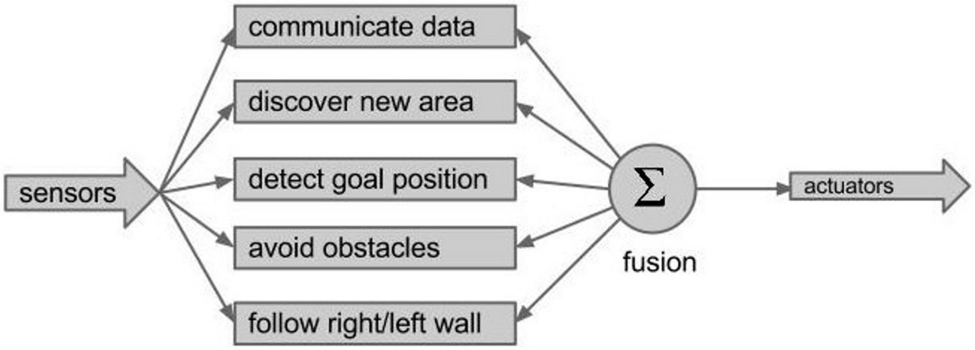

Figure 1. Architecture of Behavior based Navigation (Source: Siegwart and Nourbakhsh [25])

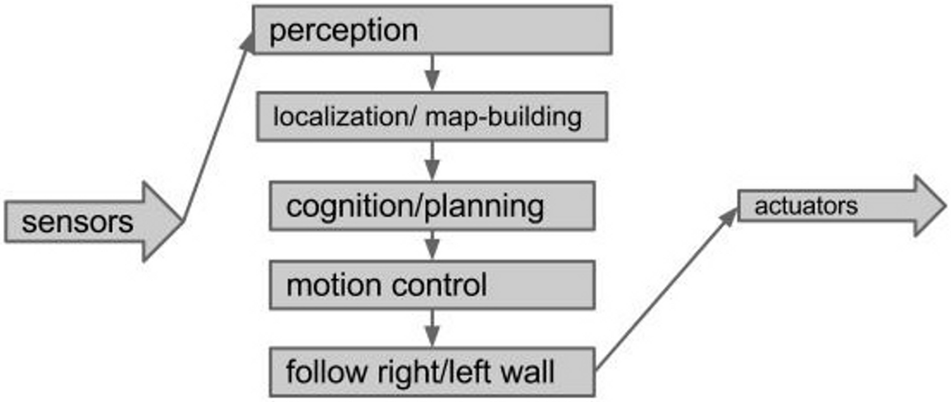

Figure 2. Architecture of Map based Navigation (Source: Siegwart and Nourbakhsh [25])

2.1. Holistic Scene Understanding

Yao et al. [34] propose an approach to holistic scene understanding that simutaneously reasons about regions, location, class, spatial extent of objects as well as the type of scene. The problem is formulated as inference in conditional random field as given in Equation 1.

$$p(\mathbf{a}) = p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{b}, \mathbf{s}) = \frac{1}{Z} \prod_{type} \prod_{\alpha} \psi_{\alpha}^{type}(\mathbf{a}_{\alpha}) \quad (1)$$

Here **x** and **y** represents the segmentation random variables, **z** represents the presence of different classes in a scene and **b** represents the set of all candidate object detections.

Schwing et al. [23] estimated the 3D scene layout of a room given a single image using a concept of integral geometry which made the structured prediction framework more efficient. Later, Schwing et al. [24] used branch and bound approach to split of space of 3D layouts and estimate the exact solution in less time than approximate inference tools. In a similar method, Xiao et al. [33] used part-based detector to model the appearance of the cuboid corners and internal edges while enforcing consistency to a 3D cuboid model. They were able to detect rectangular cuboids across many different object categories. Li et al. [17] proposed a feedback enabled cascaded classification model to maximize the joint likelihood of each of the sub-tasks like scene categorization, depth estimation, object detection and requires only a 'black-box' interface to the original classifier for each sub-task. Hedau et al. [11] estimated free space in a single image by using "box" detector to localize major furniture objects.

Hoiem and Savarese [12] did a survey of the recent work in the area of 3D scene understanding and 3D object recognition. Jianxiong Xiao [32] also gave a good overview of the related work in 3D reasoning from single image. According to Xiao, obtaining a depth map to capture a distance at each pixel is analogous to inventing a digital camera to capture the color value at each pixel. The gap between low-level depth measurements and high level shape understanding is just as large as the gap between pixel colors and high level semantic perception.

Humans as a Cue: Human activity is a very strong cue for 3D understanding of scenes. Gupta et al. [9] used the 3D scene geometry to estimate the human workspace. This method built upon the work in indoor scene understanding and the availability of motion capture data to create a joint space of human poses and scene geometry by modeling the physical interactions between the two. Recently, Fouhey et al. [7] investigated the use of human pose for scene understanding. They used the estimated human pose to extract the functional and geometrical constraints about the scene. They showed that observing people performing different actions can significantly improve estimate of 3D scene geometry.

2.2. Scene understanding from RGBD image

Wide availability of inexpensive depth sensors like Kinect has resulted in sudden increase in the interest on exploiting the additional depth information by the robotics and vision community [1, 14, 15, 22]. Koppula et al. [1, 14, 15] used graphical model capturing various image feature and contextual relationship to semantically label the point cloud with object classes and used that on a mobile robot for finding objects in a large cluttered room. Ren et al. [22] proposed algorithms and features to do dense scene labeling of indoor scenes using RGB-D images.

2.3. Semantic structure from motion

Bao et al. [5] proposed a framework to jointly recognize objects and at the same time discover their spatial organization in 3D. To achieve this they introduced a joint probability model to detect object and estimate 3D structure in a coherent fashion. The joint probability model is specified using Equation 2.

$$\begin{aligned} \mathbf{arg} \max_{\mathbf{Q},\mathbf{O},\mathbf{C}} \mathbf{Pr}(\mathbf{q},\mathbf{u},\mathbf{o}|\mathbf{Q},\mathbf{C},\mathbf{O}) = \\ \mathbf{arg} \max_{\mathbf{Q},\mathbf{O},\mathbf{C}} \mathbf{Pr}(\mathbf{q},\mathbf{u}|\mathbf{Q},\mathbf{C})\mathbf{Pr}(\mathbf{o}|\mathbf{O},\mathbf{C}) \end{aligned} \tag{2}$$

Here **Q,O,C** are the set of 3D points, 3D objects and camera parameters respectively. Similarly **q,u,o** are the point measurments, indicator variable and object measurments respectively.

In a more recent paper, Bao et al. [4] model the interaction between points, objects and regions and jointly estimate the location and pose of objects, regions, points and cameras in the 3D scene. The estimation problem is formulated as an energy maximization framework over points, regions and objects as showin in Equation 3.

$$\begin{aligned} \mathbf{Q},\mathbf{O},\mathbf{B},\mathbf{C} &= \mathbf{arg} \max_{\mathbf{Q},\mathbf{O},\mathbf{B},\mathbf{C}} \psi(\mathbf{Q},\mathbf{O},\mathbf{B},\mathbf{C};\mathbf{I}) \\ &= \mathbf{arg} \max_{\mathbf{Q},\mathbf{O},\mathbf{B},\mathbf{C}} \prod_{\mathbf{s}} \psi_{\mathbf{s}}^{\mathbf{CQ}} \prod_{\mathbf{t}} \psi_{\mathbf{t}}^{\mathbf{CO}} \prod_{\mathbf{r}} \psi_{\mathbf{r}}^{\mathbf{CB}} \\ &\qquad \prod_{\mathbf{t},\mathbf{s}} \psi_{\mathbf{t},\mathbf{s}}^{\mathbf{OQ}} \prod_{\mathbf{t},\mathbf{r}} \psi_{\mathbf{t},\mathbf{r}}^{\mathbf{OB}} \prod_{\mathbf{r},\mathbf{s}} \psi_{\mathbf{r},\mathbf{s}}^{\mathbf{BQ}} \end{aligned} \tag{3}$$

Here $\psi_s^{CQ}, \psi_t^{CO}, \psi_r^{CB}$ measures the consistency of 3D points, objects and regions with the image measurment respectively. $\psi_{t,s}^{OQ}, \psi_{t,r}^{OB}, \psi_{r,s}^{BQ}$ evaluates the interactions between difference scene components.

2.4. Semantic Mapping for Robots

Pronobis et al. [19, 20] proposed a complete and efficient representation of indoor spaces including semantic information. They use a multi-layered semantic mapping representation to combine information about the existence of objects in the environment with knowledge about the topology and semantic properties of space such as room size, shape and general appearance. It is used to infer semantic categories of rooms and predict existence of objects and values of other spatial properties.

The spatial knowledge is represented using four layers of abstraction. At the lowest level is the sensory layer which stores the feature level representation of the immediate environment of the robot. Above this are the place and categorical layers which contains the topological map and the pre-trained categorical models. Topmost layer is the conceptual layer which creates a unified representation relating sensed instance knowledge to general conceptual knowledge. This is a very abstract layer representing human knowledge about the environment and is generally not needed for tasks like autonomous navigation. A task like searching objects which requires higher cognitive ability will require this layer of abstraction. Chain graph probabilistic model is used to implement the conceptual layer.

In an another set of work, Tsai et al. [29] estimates a set of indoor structure hypotheses from a single image which is subsequently refined using the information from motion cues. The likelihood function in this case is computed by comparing the predicted location of point features on the environment model to their actual tracked locations in the image stream. Following that, Tsai et al. [28] extended the work to incorporate incremental changes to the hypotheses by using children hypotheses describing the same environment in more detail.

Aydemir and Jensfelt [3] used the correlation between the 3D structure and object placement to predict object locations. Aydemir et al. [2] analyzed a corpus of 567 floors, 6426 spaces with 91 room types and 8446 connections between rooms corresponding to real places and used that knowledge to predict the rest of the topology given a partial graph.

2.5. Planning to Perceive

An interesting area especially from the point of view of robotics is the ability to actively plan the robot movement to maximize the information gain or gain a reward with respect to a certain task. Göbelbecker et al. [8] presented a planning approach to active object search in unknown environments. A hierarchical model for representing object locations is used with which the planner is able to perform the indirect search. Vélez et al. [30, 31] describe an online planning framework to enable active exploration of possible object detections. They present a probabilistic approach where vantage points are identified which provide a more informative view of a potential object. However, the cost of taking a detour is then weighed against the increase in confidence regarding a particular object and the time taken to reach the actual destination.

2.6. Saliency based Mapping

Saliency plays an important role in identifying the important regions in an image and can be used to reduce the number of features which are used in mapping. Only salient features that are most likely to match under different illumination conditions, repeatable across different poses and are most informative for localizing a new image when identified across all images can be used to generate a map. Borji and Itti presented a taxonomy of different models to estimate saliency in images and videos [6]. Some of the techniques for saliency based mapping can be as described below.

- Posthoc Compression: Without any prior information about the environment, how much can we compress the optimized SLAM data with minimal loss in the accuracy. This would give us a set of points and cameras which are informative to navigation and can result in 3D saliency map of the environment. It can be done by minimizing the overall uncertainty between the 3D points selected and the cameras. Skeletal graph based compression [26] over the salient features can be used to give a minimal representation of the map without losing much accuracy

- Online Compression: Adding rules about the environment, how does the saliency map changes. It can also include rules like where to look for the salient features and some of the predefined informative salient objects in an indoor environment.

2.7. Image Retrieval based Localization

In a series of techniques popularized by Hays and Efros [10], recent large scale image based localization techniques estimates the location of the query frame by finding the nearest frame from the dataset consisting of image and location pairs. This is a purely data driven approach and is limited by the space not captured by the images in the dataset. The accuracy of image retrieval based localization improves with the space covered by the images in the dataset. Another problem with this approach is the large amount of variation in the scene appearance due to the changes in illumination, pose etc.

Overall, all of these approaches can be used together to form a structured knowledge of the environment which can be used by the robot to form a semantic representation of the surrounding and intelligently perform different tasks.

3. Related work in Map Representation

Benjamin Kuipers [16] was one of the first researchers to provide a cognitive map representation to model the knowledge a person has about the spatial structure of a large scale environment. The functions of the cognitive map are to assimilate new information about the environment, to represent the current position, and to answer route-finding and relative-position problems. The model analyzes the cognitive map in terms of symbolic descriptions of the environment and operations on those descriptions.

Sebastian Thrun [27] abstracted a topological map from an underlying metric map. He first builds a global occupancy map, and then forms a roadmap network for planning, using Voronoi diagrams. He shows that planning on the topological map leads to path lengths only a few percent greater than the grid-based paths. Zivkovic et al. [35] used a graph-cut clustering model rather than Voronoi diagrams to present the similar system. Ranganathan and Dellaert [21] proposed a probabilistic approach to infer on the space of all possible topological maps given all measurements and the posterior is updated incrementally with each successive measurement.

4. Related work in Localization

Dervish: office navigating robot by Illah Nourbakhsh [18] was one the first robots to succesfully navigate using a topological map using its own sensors to navigate from a chosen starting position to a target room. It also won the 1994 AAAI National Robot Contest. Dervish employed a probabilistic Markov localization and used a multiple-hypothesis belief state over a topological environmental representation.

5. Related work in Navigation

Konolige et al. [13] recently proposed an approach for navigation in hybrid maps consisting of a topological graph overlaid with local occupancy grids. The topological graph is built on top of a graph SLAM system, which can be efficiently optimized even for very large environments. The navigation is done locally using metric maps which are connected using a topological map.

6. Proposed Approach

The three major components involved in this work are Map Representation, Localization and Navigation.

6.1. Map Representation

Map representation is one of the most important component of the whole pipeline. The map representation directly affects the techniques used to localize in it and navigate through it. We believe that a *topological map* representation consisting of nodes and edges is a generic model to represent a map. It is analogous to Kuipers PATH-PLACE model. According to Siegwart and Nourbakhsh [25] there are three important fundamental relationships while choosing a map representation:

- The precision of the map must appropriately match the precision with which the robot needs to achieve its goal.
- The precision of the map and the type of features represented must match the precision and data types returned by the robot's sensors.
- The complexity of the map representation has direct impact on the computational complexity of reasoning about mapping, localization and navigation.

Therefore its important to design the map representation according to the capabilities of the localization and navigation. For each entity in the graph which include nodes and edges, we have to define how can we know if a robot has reached this node or edge (localization) and the actions which can be performed in this node or edge (included navigation). Each node in a low level topological map can represent a 3D point or a camera pose which are connected by the edges if the camera sees the 3D point and in a higher level topological map each node can represent a concept like a room or a collection of rooms and an edge can represent a path connecting those nodes. One important characteristics to think about while deciding between what to represent as nodes and what to represent as edges is to find the ease to localize in a certain node and the ease to navigate in a certain edge. If a place in the 3D scene has higher distinctiveness than its surrounding then we can create a node representing that location and the edges should connect all such places using easily navigable paths. Table 1 summarizes the characteristics of nodes and edges in a topological map.

We propose a hybrid representation to specify each node and edge. Accordingly each node and edge can be represented either using a metric map or a semantic map depending upon the robot's capability to understand it. The decision to choose between higher and lower level representations for each node and edge can vary across different regions. For example, it is better to have a apperance based representation (bag of visual words) for a cluttered room than to fit a 3D layout which can be more difficult. However it might be easier to use a line based representation for particularly feature less regions like corridors connecting different rooms. The size of the map can depend on the level of abstraction we use to represent a particular node. This directly translates to robot's cognition ability to localize and navigate through that node. Figure 3 depicts the increase in the size of the map with the level of abstraction in the map and the corresponding decrease in the cognition ability to localize in it. Higher cognition ability requires a better behavior based navigation strategy to better understand the local environment and navigate from one node to the other. In regions where it is tough to localize in the global map, we rely on local behavior based navigation to move from one place to another. In another way, intentionally not localizing in certain regions and relying on behavior based navigation can help reduce the size of the map.

	Nodes	Edges
Localization	Easy to localize	Don't Care
Navigation	Don't Care	Easy to Navigate

Table 1. Characteristics of nodes and edges in a topological map

6.2. Localization

Robot's localization and the reference system according to Benjamin Kuiper [16] can have the following characteristics:

- Position is often given as a relation between two places and a reference system, not an absolute property of a single place.
- There are many reference sytems with respect to which positions may be defined.
- The relation between different reference systems may be unknown.
- Many people orient themselves with respect to conspicuous landmarks.
- A reference system need not be tied to a large geographical structure like a mountain or a set of grid-structured streets. It may be created to represent the locations of a small set of nearby but mutually invisible places, whose positions are computed by dead-reckoning along short routes

Initially we'll consider the case of global localization where we continuously localize against a global map us-

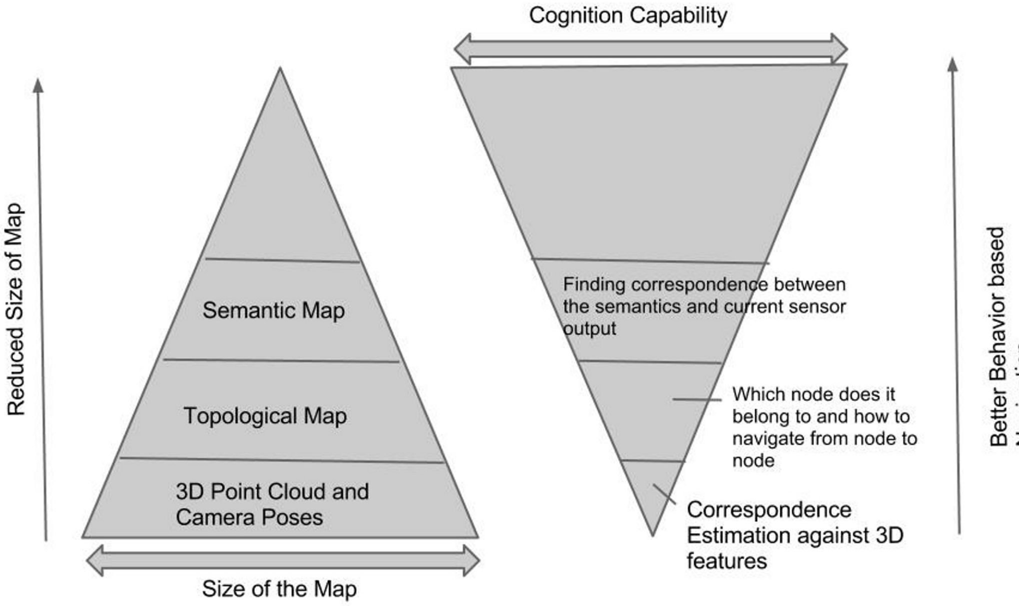
Cognition Capability
Reduced Size of Map
Semantic Map
Topological Map
3D Point Cloud and Camera Poses
Size of the Map
Finding correspondence between the semantics and current sensor output
Which node does it belong to and how to navigate from node to node
Correspondence Estimation against 3D features
Better Behavior based Navigation

Figure 3. Map Representation: Increase in the size of the map corresponds to decrease in the cognition capability of the robot and vice versa

ing landmarks. This can be followed by fusing information from dead-reckoning and

6.3. Navigation

To find a proper balance between Behavior/Affordance based navigation and Map based navigation can lead to a minimal representation for a robot. Initially we can start without using any behavior/affordance based navigation and using only map based navigation to find how much can we reduce the map without losing the accuracy of map based navigation. After this we can further adaptively reduce the map in some regions where behavior based navigation works fine and use map based navigation for the rest of the regions.

6.4. Plan

The plan for this problem can be something like this.

- **First stage:** Find minimal representation of a map to continuously localize in the environment and navigate using map based navigation
- **Second stage:** Find minimal representation to localize in the nodes using map based navigation and navigate along the edges using behavior based navigation.

References

[1] A. Anand, H. S. Koppula, T. Joachims, and A. Saxena. Contextually guided semantic labeling and search for 3d point clouds. *CoRR*, abs/1111.5358, 2011.

[2] A. Aydemir, E. Järleberg, S. Prentice, and P. Jensfelt. Predicting what lies ahead in the topology of indoor environments. In *Spatial Cognition*, pages 1–16, 2012.

[3] A. Aydemir and P. Jensfelt. Exploiting and modeling local 3d structure for predicting object locations. In *IROS*, pages 3885–3892, 2012.

[4] S. Y. Bao, M. Bagra, Y.-W. Chao, and S. Savarese. Semantic structure from motion with points, regions, and objects. In *CVPR*, 2012.

[5] S. Y. Bao and S. Savarese. Semantic structure from motion. In *CVPR*, 2011.

[6] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):185–207, 2013.

[7] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single view geometry. In *ECCV (5)*, pages 732–745, 2012.

[8] M. Göbelbecker, A. Aydemir, A. Pronobis, K. Sjöö, and P. Jensfelt. A planning approach to active visual search in large environments. In *Automated Action Planning for Autonomous Mobile Robots*, 2011.

[9] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR*, pages 1961–1968, 2011.

[10] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[11] V. Hedau, D. Hoiem, and D. A. Forsyth. Recovering free space of indoor scenes from a single image. In *CVPR*, pages 2807–2814, 2012.

[12] D. Hoiem and S. Savarese. *Representations and Techniques for 3D Object Recognition and Scene Interpretation*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.

[13] K. Konolige, E. Marder-Eppstein, and B. Marthi. Navigation in hybrid metric-topological maps. In *ICRA*, pages 3041–3047, 2011.

[14] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena. Labeling 3d scenes for personal assistant robots. *CoRR*, abs/1106.5551, 2011.

[15] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *NIPS*, pages 244–252, 2011.

[16] B. J. Kuipers. Representing knowledge of large-scale space, 1977.

[17] C. Li, A. Kowdle, A. Saxena, and T. Chen. Toward holistic scene understanding: Feedback enabled cascaded classification models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7):1394–1408, 2012.

[18] I. R. Nourbakhsh, R. Powers, and S. Birchfield. Dervish - an office-navigating robot. *AI Magazine*, 16(2):53–60, 1995.

[19] A. Pronobis. *Semantic Mapping with Mobile Robots*. PhD thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2011.

[20] A. Pronobis and P. Jensfelt. Large-scale semantic mapping and reasoning with heterogeneous modalities. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3515 –3522, may 2012.

[21] A. Ranganathan and F. Dellaert. Online probabilistic topological mapping. *I. J. Robotic Res.*, 30(6):755–771, 2011.

[22] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, pages 2759–2766, 2012.

[23] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient structured prediction for 3d indoor scene understanding. In *CVPR*, pages 2815–2822, 2012.

[24] A. G. Schwing and R. Urtasun. Efficient exact inference for 3d indoor scene understanding. In *ECCV (6)*, pages 299–313, 2012.

[25] R. Siegwart and I. R. Nourbakhsh. *Introduction to Autonomous Mobile Robots*. Bradford Company, Scituate, MA, USA, 2004.

[26] N. Snavely, S. M. Seitz, and R. Szeliski. Skeletal graphs for efficient structure from motion. In *CVPR*, 2008.

[27] S. Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence*, 99(1):21–71, 1998.

[28] G. Tsai and B. Kuipers. Dynamic visual understanding of the local environment for an indoor navigating robot. In *IROS*, pages 4695–4701, 2012.

[29] G. Tsai, C. Xu, J. Liu, and B. Kuipers. Real-time indoor scene understanding using bayesian filtering with motion cues. In *ICCV*, pages 121–128, 2011.

[30] J. Vélez, G. Hemann, A. S. Huang, I. Posner, and N. Roy. Active exploration for robust object detection. In *IJCAI*, pages 2752–2757, 2011.

[31] J. Vélez, G. Hemann, A. S. Huang, I. Posner, and N. Roy. Planning to perceive: Exploiting mobility for robust object detection. In *ICAPS*, 2011.

[32] J. Xiao. 3d reconstruction is not just a low-level task: retrospect and survey. In *MIT-TR*, 2012.

[33] J. Xiao, B. C. Russel, and A. Torralba. Localizing 3d cuboids in single-view images. In *ECCV*, 2012.

[34] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, pages 702 –709, june 2012.

[35] Z. Zivkovic, B. Bakker, and B. Krose. Hierarchical map building and planning based on graph partitioning. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 803 –809, may 2006.