Dynamic visual understanding of the local environment for an indoor navigating robot

Grace Tsai and Benjamin Kuipers

***Abstract*— We present a method for an embodied agent with vision sensor to create a concise and useful model of the local indoor environment from its experience of moving within it. Our method generates and evaluates a set of qualitatively distinct hypotheses of the local environment and refines the parameters within each hypothesis quantitatively. Our method is a continual, incremental process that transforms current environmental-structure hypotheses into children hypotheses describing the same environment in more detail. Since our method only relies on simple geometric and probabilistic inferences, our method runs in real-time, and it avoids the need of extensive prior training and the Manhattan-world assumption, which makes it practical and efficient for a navigating robot. Experimental results on a collection of indoor videos suggests that our method is capable of modeling various structures of indoor environments.**

I. Introduction

A navigating robot must perceive its local environment. Visual perception has many advantages, such as acquiring more information for place recognition, acquiring more information at lower cost, the ability to detect drop-offs, etc. The output of visual perception must be a concise description of the agent's environment at a level of granularity that is useful to the agent in making plans or achieving a richer understanding of the environment. Visual processing must be done in real-time to keep up with the agent's needs.

Methods, such as Structure-from-Motion [9], [20], [3], [21] and Visual SLAM [4], [19], [6], [14], take a stream of visual observations and produce a model of the scene in the form of a 3D point cloud. A more concise, large-granularity model that would be useful to an agent in planning and navigation must then be constructed from the point cloud. There are methods [27], [18] that combine 3D point cloud and image data for semantic segmentation. Other methods [8], [7], use the Manhattan-world assumption to reconstruct a planar structure of an indoor environment through a collection of images. These methods are offline and computationally intensive, making them difficult to apply in real-time robot navigation.

There has been impressive recent work on visual scene understanding and on the derivation of depth maps from single images of indoor and outdoor scenes [11], [13], [5], [15], [22], [10], [16], [26], [1]. These methods typically depend on careful training with prior knowledge linking local image properties to a classification of local surface orientation [11], [13], [10], to depth of surfaces in the environment [22], or to semantic labels and then to depth [16], [26]. Dependence on prior training knowledge with relevant domain specific examples makes these methods difficult to generalize to different environments. In addition, real-time performance may be difficult to achieve when evaluations at the pixel or superpixel level are involved. Furthermore, coherent results of the 3D scene estimation may be difficult to achieve if each frame is independently processed.

In this paper, we present a concise and useful representation of an indoor environment that describes the environment by a set of meaningful planes — the ground plane G and a set of planar walls W_i that are perpendicular to the ground plane but not necessarily to each other. There is a one-to-one correspondence between this representation and a set of lines (the ground-wall boundaries) in the ground plane, represented in the same 3D world frame. We assume that the ground plane G can be unambiguously identified.

By identifying potential ground-wall boundary lines in the 2D images, we generate a set of hypotheses M^k for the 3D structure of the environment, where each

$$M^k = \{G, W_1^k, W_2^k, W_3^k, ..., W_{n_k}^k\}. \quad (1)$$

Given the camera pose, we predict how point features move in the 2D images from each hypothesis and compute the likelihood of the hypothesis by comparing the predicted and observed location of the features in each frame. These hypotheses can be efficiently tested during robot motion by Bayesian filtering [25]. However, our previous work focused only on simple corridor-like environments that consist of at most three walls where each wall intersects with its adjacent walls. Moreover, it is only capable of modeling the portion of the local environment that is visible in the first frame of the video.

In this paper, we extend our representation for wall planes to include sets of endpoints for line segments in the ground-wall boundary lines, delimiting where the wall is present and where there is an opening (e.g. at an intersection). Since such an opening may not be visible in the image from a distance, we introduce a continual, incremental process for transforming a current environmental-structure hypothesis into children hypotheses describing the same environment in more detail. Fig. 1 illustrates our approach. In addition to introducing qualitative structural improvements through new children hypotheses, we use the information in current

This work has taken place in the Intelligent Robotics Lab in the Computer Science and Engineering Division of the University of Michigan. Research of the Intelligent Robotics lab is supported in part by grants from the National Science Foundation (CPS-0931474 and IIS-1111494), and from the TEMA-Toyota Technical Center.

Grace Tsai is with Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, gstsai@umich.edu

Benjamin Kuipers is with Faculty of Computer Science and Engineering, University of Michigan, Ann Arbor, kuipers@umich.edu

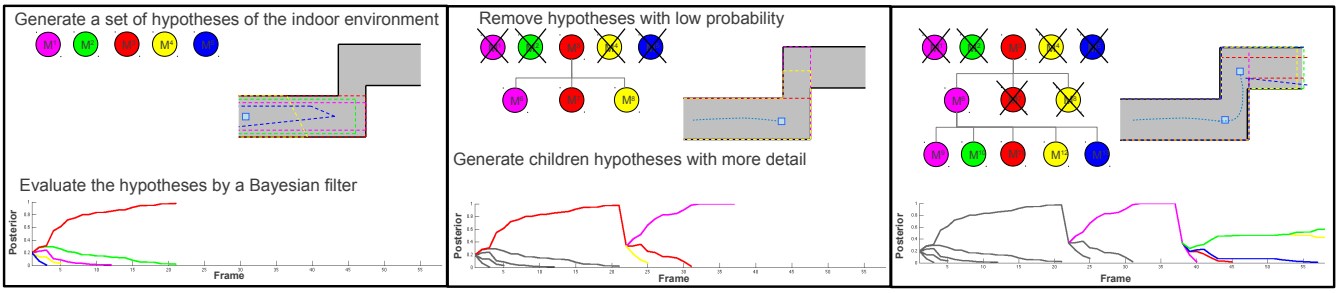


Fig. 1. Our proposed framework. (Best viewed in color.) A generate-and-test framework is proposed to build the geometric structure of the local environment. Hypotheses about the structure of the environment are generated through a continual, incremental process, and evaluated through a Bayesian filter. (left) Starting from a set of simple parent hypotheses, the Bayesian filtering framework identifies the best hypotheses and removes hypotheses with low posterior probabilities. (center and right) A set of children hypotheses are generated from the existing hypotheses to describe the same environment in more detail. Our Bayesian filtering framework continuously evaluates each new set of hypotheses.

observations to refine the quantitative precision of existing hypotheses. Since the method relies only on simple geometric and probabilistic inference, the system runs in real-time and does not rely on prior training data.

The main contribution of this paper is an online method that builds the geometric structure of the local indoor environment, without the need for prior training data or the Manhattan-world assumption. Our method generates and evaluates a set of qualitatively distinct hypotheses of the local environment while refining the parameters within each hypothesis quantitatively. Our representation is a coarse-grained description of the indoor environment in terms of meaningful planes (the ground plane and the walls), instead of a low-level fine-grained representation like point clouds. Furthermore, our representation is capable of representing partial knowledge of the local environment such that unknown areas can be incrementally built as observations become available.

II. Method

A. Representation of the Indoor Planar Structure

We represent a 3D indoor environment by a set of semantically meaningful planes, namely, a ground plane and a set of walls which are perpendicular to the ground plane but not necessarily to each other. There is a one-to-one correspondence between this representation and a set of lines (the ground-wall boundaries) in the ground plane, represented in the same 3D world frame. [1]

A wall W_i contains a set of disjoint wall segments which share the same plane equation in the 3D world coordinate. In the ground-plane map, a wall segment is represented by a pair of endpoints on the corresponding line. There are three different types of endpoints: *dihedral*, *occluding* and *indefinite*. A *dihedral* endpoint corresponds to two visible wall segments, where the location of the endpoint is the projection of the intersection of the two walls. An *occluding* endpoint corresponds to only one visible wall segment. An *indefinite* endpoint is an endpoint that is known to exist but its actual location has not yet been observed by the robot due to occlusions or the end of the robot's field of view.

[1]For a robot rolling or walking on the ground plane, the ceiling is much less relevant than the ground plane and the walls, so it can safely be omitted from the representation. An indoor flying vehicle would require us to extend this representation to include the ceiling.

B. 3D Reconstruction

In this paper, we assume that the robot moves on the ground plane and the camera is at the robot's center of rotation at a fixed height h from the ground plane. [2] We set the x-y plane of the coordinate system to be parallel to the ground plane with the origin at the initial location of the camera center. The x-axis and y-axis of the coordinate system are pointing to the front and the left of the camera. Furthermore, we define the ground-plane map as the top-down view of the world coordinate, which corresponds to the x-y plane of the world coordinate. The ground-plane map location of a 3D point $\mathbf{P} = (x, y, z)$ is $\tilde{\mathbf{p}} = (x, y)$.

In the world coordinate, the camera pose is denoted as $u = (x^c, y^c, z^c, \theta^c, \phi^c, \psi^c)$. We assume that the robot has a fixed and known tilt ϕ^c and roll ψ^c angles with respect to the ground plane, and has a fixed height $z^c = -h$ from the ground plane. Thus, the camera pose is simplified to $u = (x^c, y^c, \theta^c)$.

The 3D location $\mathbf{P}_i = (x_i, y_i, z_i)^T$ of an image point [3] $\mathbf{p}_i = (u_i, v_i, 1)^T$ that lies on the ground plane is related by

$$\mathbf{R}_{\psi^c}\mathbf{R}_{\phi^c}\mathbf{R}_c \begin{pmatrix} x_i \\ y_i \\ -h \end{pmatrix} = \lambda \begin{pmatrix} u_i \\ v_i \\ 1 \end{pmatrix} \tag{2}$$

where $\mathbf{R}_{\psi^c}$ and $\mathbf{R}_{\phi^c}$ are the rotation matrices related to the camera tilt and roll angles, respectively. In this case, the rotation matrix corresponding to the roll angle is

$$\mathbf{R}_{\psi^c} = \begin{bmatrix} \cos\psi^c & -\sin\psi^c & 0 \\ \sin\psi^c & \cos\psi^c & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

[2]Our geometry is defined under this assumption. If this is not the case, a transformation must be done so that the camera center is at the robot's center of rotation.

[3]The camera needs to be calibrated so that the image point is on the normalized image plane (focal length $f = 1$).

and the rotation matrix corresponding to the tilt angle is

$$\mathbf{R}_{\phi^c} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi^c & -\sin\phi^c \\ 0 & \sin\phi^c & \cos\phi^c \end{bmatrix}. \quad (4)$$

$\mathbf{R}_c$ is the matrix that transforms the location of a 3D point from the image coordinate to the world coordinate:

$$\mathbf{R}_c = \begin{bmatrix} 0 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}. \quad (5)$$

Solving (2) gives us the 3D location of the ground plane point in the world coordinate,

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \begin{pmatrix} \lambda_i(\cos\phi^c - v_i \sin\phi^c) \\ -\lambda_i(\sin\psi^c \sin\phi^c + u_i \cos\psi^c + v_i \cos\phi^c \sin\psi^c) \\ -h \end{pmatrix} \quad (6)$$

where

$$\lambda_i = \frac{h}{\cos\psi^c \sin\phi^c - u_i \sin\psi^c + v_i \cos\psi^c \cos\phi^c}.$$

A wall plane W_j in the 3D world coordinate corresponds to a line parametrized by (α_j, d_j) in the ground-plane map. $\alpha_j \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right]$ is the orientation of the line which implies the normal direction of the wall plane in the 3D coordinate, and $d_j \in \mathbb{R}$ is the directed distance from the origin of the ground-plane map to the line. Since the walls are perpendicular to the ground plane, the normal vector of the wall $\mathbf{N}_j$ in the world coordinate is $\mathbf{N}_j = (\cos\alpha_j, \sin\alpha_j, 0)$, and the directed distance from the origin of the world coordinate to the plane is d_j. $\mathbf{N}_j$ and d_j determine the equation of the wall in the 3D world coordinate.

We start by selecting any two points along the line and reconstruct their locations in the ground-plane map, $\tilde{\mathbf{p}}_1 = (x_1, y_1)^T$ and $\tilde{\mathbf{p}}_2 = (x_2, y_2)^T$, using the geometry of ground plane points. Given the two points, α_j can be determined by,

$$\alpha_j = -\arctan\frac{x_1 - x_2}{y_1 - y_2}. \quad (7)$$

and thus, the normal vector of the corresponding line in the ground-plane map is $\mathbf{n}_j = (\cos\alpha_j, \sin\alpha_j)^T$. The directed distance d_j from the origin to the line can be determined by,

$$d_j = \mathbf{n}_j \cdot \tilde{\mathbf{p}}_1 = \mathbf{n}_j \cdot \tilde{\mathbf{p}}_2 \quad (8)$$

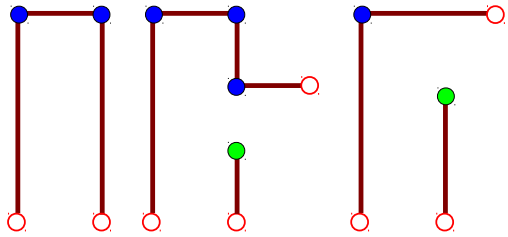
If a point lies on the wall plane with position $\mathbf{p}_i$ in the image, its 3D location in the world coordinate is related by

$$d_j = \mathbf{N}_j \cdot (\lambda_j \mathbf{R}_c^{-1} \mathbf{R}_{\phi_t}^{-1} \mathbf{R}_{\phi_r}^{-1} \mathbf{p}_i) = \mathbf{N}_j \cdot \mathbf{P}_i. \quad (9)$$

Solving λ_j in (9) gives us the 3D location of the point $\mathbf{P}_j$.

C. Hypotheses Generation

Given the camera pose, the image projection of the ground-wall boundary of an indoor planar structure is a polyline extending from the left to the right borders of the image, where the initial and final segments may lie along the lower image border. A non-vertical line segment corresponds to a wall segment in the indoor planar structure and vertical segments correspond to occluding edges between planar wall segments. Thus, a set of hypotheses of the 3D local environment can be generated from 2D image features.



(a) parent (b) child (type 1) (c) child (type 2)

Fig. 2. Types of child hypotheses (Best viewed in color.) On the ground-map space, the walls are represented by (red) lines and a set of endpoints delimiting where the wall is present. *Dihedral* endpoints are marked as blue and *occluding* endpoints are marked as green. *Indefinite* endpoints are marked as red hollow points. Given the parent hypothesis (a), two types of hypothesis can be generated to describe the environment in more detail. (b) adds two endpoints to an existing wall segment to form an opening, and adds new walls that are visible through the openings. Note that in this case, the representation captures the fact that the two wall segments are parts of the same wall plane. (c) creates an opening between two walls that are intersecting in the parent hypothesis.

We demonstrated an efficient method to generate a set of simple hypotheses in corridor-like environments that consist of at most three walls where each wall intersects with its adjacent walls in [25]. The method generates a set of hypotheses by connecting lines with certain constraints. To obtain the lines, we extract line segments by edge linking and then merge line segments to form a set of long straight lines [24]. In this paper, in addition, we transform a current hypothesis to a set of children hypotheses describing the same environment in more detail.

Two types of child hypothesis can be generated from a current hypothesis. The first type of child hypothesis adds openings along walls. These children hypotheses add endpoints to the existing wall segments in the parent hypothesis to create the openings, and add new walls that are visible through the openings (Fig. 2(b)). For each visible wall, a set of candidate openings are generated. A candidate opening consists of two endpoints belonging to two adjacent segments of the same wall to create the gap. We start by extracting image corner features along the projected ground-wall boundary line, and collect a set of corner pairs that are wide enough in 3D for the robot to pass through. For each corner pair, we combine line segments between the two corners to form a set of candidate openings. A child hypothesis is generated by selecting at most one candidate opening from each visible wall.

The second type of child hypothesis creates an opening between two walls that are intersecting in the parent hypothesis (Fig. 2(c)). From each *dihedral* endpoint that corresponds to a concave wall intersection, a set of candidate openings can be generated by transforming the *dihedral* endpoint into an *occluding* endpoint for one wall segment and an *indefinite* endpoint for the other wall segment. Thus, we search for image corner features along both associated wall segments of the *dihedral* endpoint to become a candidate *occluding* endpoint. A candidate opening is generated by a candidate

occluding endpoint that provides a feasible gap for the robot to pass through. A child hypothesis is generated by selecting at most one candidate opening from each concave *dihedral* endpoint.

In addition to the above two transformations, we apply a constrained version of the simple hypothesis generation process [25] to generate children hypotheses. Certain elements of the child hypotheses are constrained to have values from their parent hypothesis, but we generate many hypotheses for the previously unobserved portions of the environment.

D. Refining Planar Structure

We use the information in current observations to refine the quantitative precision of each existing hypothesis. The generic Extended Kalman Filter (EKF) is used to estimate the parameters of each wall plane and the ground plane map location of each *occluding* endpoint. The estimate update is carried out in two stages, prediction and correction.

Prediction:

$$\begin{aligned} \hat{\mu}_t &= g(\mu_{t-1}) \\ \hat{\Sigma}_t &= G_t \Sigma_{t-1} G_t^T + Q_t, \end{aligned} \tag{10}$$

where $G_t = \frac{\partial g(\mu)}{\partial \mu}\big|_{\mu=\mu_{t-1}}$ and Q_t is the prediction noise.

Correction:

$$\begin{aligned} \mu_t &= \hat{\mu}_t + K_t(z_t - h(\hat{\mu}_t)) \\ \Sigma_t &= (I - K_t H_t)\hat{\Sigma}_t \\ K_t &= \hat{\Sigma}_t H_t^T (H_t \hat{\Sigma}_t H_t^T + R_t)^{-1}, \end{aligned} \tag{11}$$

where $H_t = \frac{\partial h(\mu)}{\partial \mu}\big|_{\mu=\hat{\mu}_t}$ and R_t is the measurement noise.

For each wall w that is visible, the parameters of the plane at frame t are $\mu_t^w = (\alpha_t, d_t)^T$. Since the walls are static in the world coordinate, the state prediction function is,

$$g(\mu_{t-1}^w) = \mu_{t-1}^w. \tag{12}$$

To obtain a 3D plane measurement z_t^w of the wall, we project the predicted ground-wall boundary to the image space and find the best match between the boundary line and a set of lines under the camera center. Using the 3D reconstruction method described in Section II-B, the measurement z_t^w can be parametrized as a 3D wall, $z_t^w = (z_\alpha, z_d)^T$. Given the camera pose $u_t = (x_t^c, y_t^c, \theta_t^c)^T$ at frame t, the predicted measurement $\hat{z}_t^w = h(\hat{\mu}_t^w, u_t)$ is obtained by

$$\begin{aligned} \hat{z}_t^w &= \begin{bmatrix} \hat{\alpha}_t - \theta_t^c \\ \hat{d}_t - \cos\hat{\alpha}_t x_t - \sin\hat{\alpha}_t y_t \end{bmatrix} \\ &\textit{or} \\ \hat{z}_t^w &= \begin{bmatrix} \hat{\alpha}_t - \theta_t^c + \pi \\ -\hat{d}_t + \cos\hat{\alpha}_t x_t + \sin\hat{\alpha}_t y_t \end{bmatrix}. \end{aligned} \tag{13}$$

Once the parameters of the walls are refined, we refine the location of each *occluding* endpoint that is visible in the current frame. The ground-plane map location of the endpoint at frame t is represented as $\mu_t^e = (x_t, y_t)^T$. Given the refined parameters of its associated wall with $\mu_t^w = (\alpha_t, d_t)$ and uncertainty Σ_t^w, the state prediction function for the endpoint location $\hat{\mu}_t^e = g(\mu_{t-1}^e, \mu_t^w)$ projects the point μ_{t-1}^e onto the wall and the process noise Q_t is,

$$Q_t = F_t \Sigma_t^{wall} F_t^T \tag{14}$$

where $F_t = \frac{\partial g(\mu^e, \mu^w)}{\partial \mu^w}\big|_{\mu^w=\mu_t^w}$. By projecting the endpoint onto the image space and matching with point features extracted along the ground wall boundary, we collect a measurement of the endpoint which is represented in the ground map space $z_t^e = (z_x, z_y)^T$. The predicted measurement of the endpoint is computed by,

$$\begin{aligned} \hat{z}_t &= h(\hat{\mu}_t^e, , u_t) \\ &= \begin{bmatrix} (x_t - x_t^c)\cos\theta_t^c - (y_t - y_t^c)\sin\theta_t^c \\ (x_t - x_t^c)\sin\theta_t^c + (y_t - y_t^c)\cos\theta_t^c \end{bmatrix}. \end{aligned} \tag{15}$$

The location of a *dihedral* endpoint is updated by finding the intersection of the two associated walls with their refined parameters. For an *indefinite* endpoint, we update its location based on the robot pose and its field of view.

E. Hypothesis Evaluation

Given a set of hypotheses, $\mathbf{M} = \{M^1, M^2, ..., M^N\}$, the posterior distribution over the hypotheses at frame t can be expressed by Bayes rule,

$$\begin{aligned} p\left(M^i|\mathbf{O}^1, \mathbf{O}^2, ..., \mathbf{O}^t\right) &\propto p\left(M^i|\mathbf{O}^1, ..., \mathbf{O}^{t-1}\right) p\left(\mathbf{O}^t|M^i\right) \\ &\propto p\left(M^i\right) \prod_{j=1...t} p\left(\mathbf{O}^j|M^i\right) \end{aligned} \tag{16}$$

where $\mathbf{O}^t$ is a set of feature correspondences that the robot observed at frame t. In order to keep the number of hypotheses N tractable, hypotheses with posterior probabilities lower than a threshold ($\frac{0.1}{N}$) are removed.

The likelihood function $p\left(\mathbf{O}^t|M^i\right)$ is defined by the ability of hypothesis M^i to explain the motion of a set of observable features $\mathbf{O}^t$ at frame t. To compute the likelihood, we extract a set of point feature correspondences $\mathbf{O}^t = \{o_1^t, o_2^t, ..., o_{n_t}^t\}$ between frame t and frame $t - t_w$ in the image space, where t_w is automatically adjusted to ensure that the number of the point features exceeds a threshold. Any feature matching or tracking methods can be used but in this paper, KLT [23] tracking is used because it is more efficient than SIFT [17] and SURF [2]. Since point features may lose track or go out of sight from time to time, new point features are extracted as they become visible in order to preserve enough observations.

Given the camera pose and hypothesis M^i, we predict the location $\hat{\mathbf{L}}^i(o_j^t)$ of a previously observed feature point o_j^t to the current frame by reconstructing its 3D location in the world coordinate (Section II-B), and then project the feature point onto frame t. The likelihood of an individual point correspondence o_j^t at image location $\mathbf{L}^i(o_j^t)$ is modeled by a normal distribution with mean at the predicted image location $\hat{\mathbf{L}}^i(o_j^t)$ in frame t. Since the likelihood only depends on the distances between $\mathbf{L}^i(o_j^t)$ and $\hat{\mathbf{L}}^i(o_j^t)$, the individual likelihood is equivalent to modeling the prediction error between the two with a zero mean normal distribution with

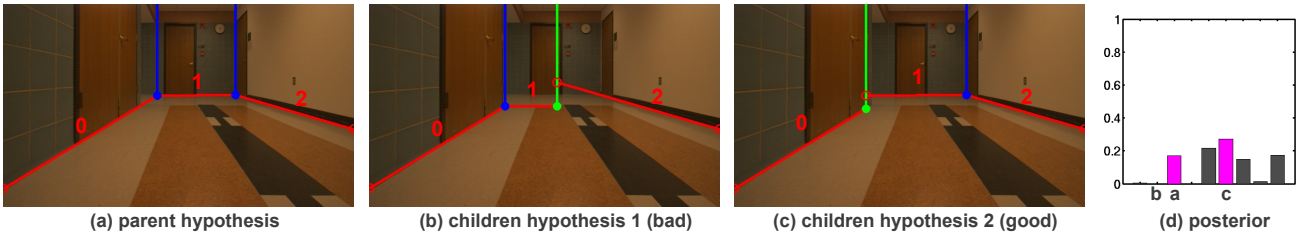


Fig. 3. Examples of the hypotheses generated by our method. (Best viewed in color.) The ground wall boundaries are plotted as red. Color dots represent *dihedral* endpoints (blue) and *occluding* endpoints (green). (b) and (c) are children hypotheses generated from (a). (d) is their posterior probability at the current time frame.

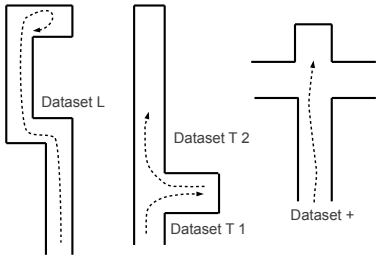


Fig. 4. Our datasets. Dataset L contains three L-intersections. In this dataset, the robot traveled through two long corridors connected by two adjacent L-intersections and finally turned at the last intersection. In dataset Dataset T 1, the robot traveled from the major corridor and turned at a T-intersection, whereas in Dataset T 2, the robot traveled from the minor corridor and turned at a T-intersection. Dataset + has one +-intersection, and the robot traveled through the intersection without turning.

variance σ ($\sigma = 20$ in our experiments). By combining the likelihoods from individual points, the likelihood of hypothesis M^i at time step t is,

$$p\left(\mathbf{O}^t|M^i\right) \propto \prod_{j=0}^{n} \exp \frac{-||\hat{\mathbf{L}^i}(o_j^t) - \mathbf{L}^i(o_j^t)||^2}{2\sigma^2}. \qquad (17)$$

Since we have no motion information to evaluate about the correctness of the hypotheses generated in the initial frame, their prior probability $p\left(M^i\right)$ in (16) is uniformly distributed over all the hypotheses. A child hypothesis that is added to the set at frame t has a prior probability equal to the posterior probability of its parent hypothesis at frame $t-1$.

III. Evaluation

We tested our approach on four video datasets [4] with resolution 965×400 in various indoor environments, such as L-intersections, T-intersections and +-intersections (Fig. 4). The videos were collected by a camera that was mounted on a wheeled robot with near zero tilt and roll angle with respect to the ground plane. For all datasets, the robot pose at each frame is provided. [5]

Examples of the hypotheses generated from our system are shown in Fig. 3. Fig. 5 demonstrates our method for transforming a current environmental-structure hypothesis into children hypotheses and shows how the Bayesian filter converges to the best hypothesis. We have presented our results for different structures of the environment in Fig. 6. Sometimes due to the lack of features and motion, the Bayesian filter might converge to a bad hypothesis as shown in Fig. 7, but our system is still able to use the information

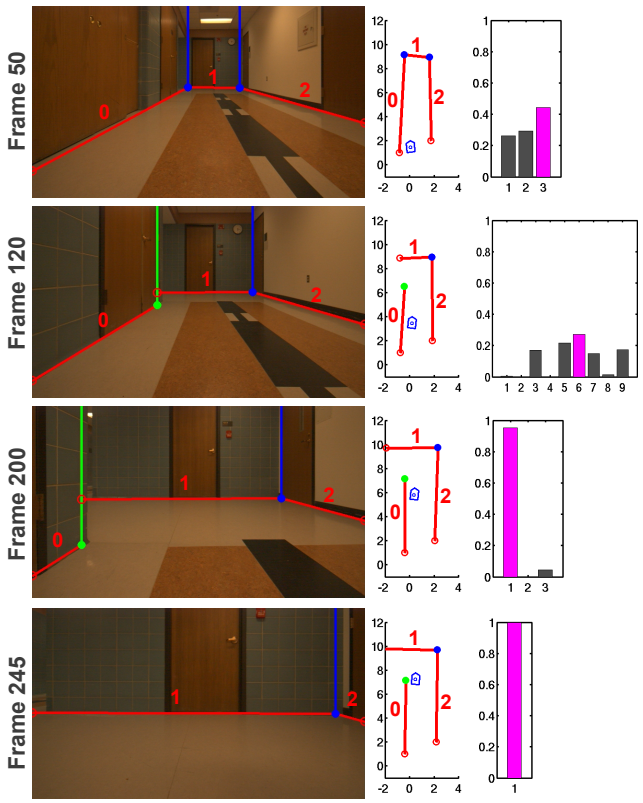


Fig. 5. Examples of our continual, incremental generate-and-test framework on L-intersection. (Best viewed in color). A set of simple hypotheses (where each wall intersects with its adjacent walls) were generated from the first frame of the video. (Frame 50) The Bayesian filter converged to three simple hypotheses. The other two hypotheses have the same side walls but a different planar equation for the end wall from the one shown. (Frame 120) Hypotheses with low probability were removed and good hypotheses generated children hypotheses to describe the scene in more detail. (Frame 200) and (Frame 245) Our Bayesian filter continues to evaluate all the hypotheses and gradually converges to the best hypothesis. At the end, the robot has a single current model of the surrounding environment even though much of it is not in view. Theses four frames are marked on the posterior probability time evaluation plot in Fig. 8.

[4]Datasets available at `http://www.eecs.umich.edu/~gstsai/release/Umich_indoor_corridor_2012_dataset.html`.

[5]We use an occupancy grid mapping algorithm with a laser range finder to obtain the robot pose. In fact, any method can be used to provide the robot pose. For example, wheel odometry or visual odometry.

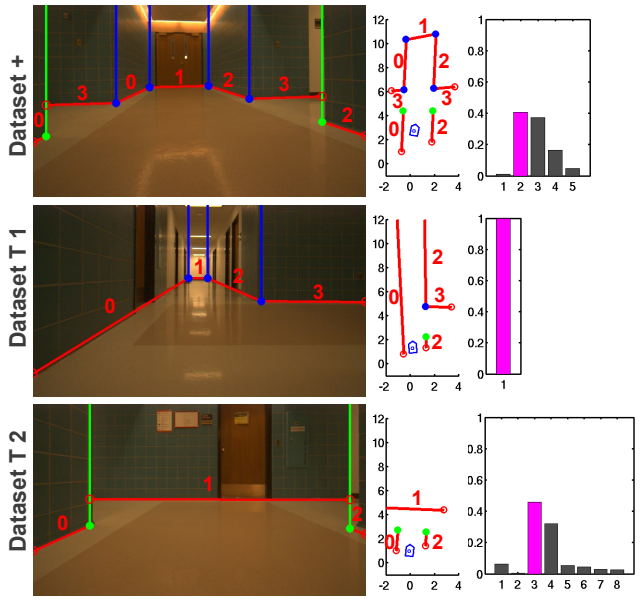


Fig. 6. Examples of our framework in different environments. (Best viewed in color). Notice that in Dataset +, we not only identify the planes but we model them as three walls where each wall has an opening. The second best hypothesis in Dataset + has a similar structure except with a different equation for the wall (wall 3) that is visible from the openings. In Dataset T 1, our method models the right side as one wall with two endpoints that creates the opening.

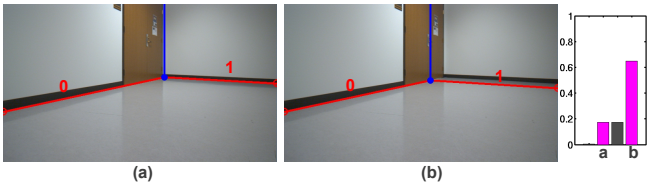


Fig. 7. An example when the Bayesian filter failed to identify the best hypothesis. (Best viewed in color.) This scene corresponds to frame 734 to frame 837 in Fig. 8. Due to the lack of observed point features and motion, our Bayesian filter might converge to a bad hypothesis (frame 760). In fact, in this dataset, the robot did not accumulate enough information to identify the best hypothesis before the end wall went out of view. However, our system is still able to use the information of the current scene (the left wall) to generate a set of new hypotheses to describe the local environment in the next scene (frame 843).

of the current scene to generate a set of new hypotheses to describe the local environment in the next scene. However, if the ground-wall boundaries are blocked by objects, our method might fail because the best hypothesis might not be generated in the first place. Thus, our future work is to extend this framework to handle cluttered environments.

For each test video, we manually labeled the ground truth classification of the planes (i.e. the walls, ground and ceiling plane) for all pixels every 10 frames in order to evaluate our results quantitatively. In each frame, we define the accuracy of a hypothesis being the percentage of the pixels that have the correct classification in the frame. Since the ceiling plane is not included in our hypotheses, we skipped the ceiling pixels in our evaluation. Two types of quantitative accuracy are reported for each dataset. *MAP hypothesis accuracy* is the accuracy of the hypothesis of the maximum posterior probability at each frame. *Weighted accuracy* is the weighted average accuracy of all the existing hypotheses at each evaluated frame where the weight of each hypothesis is equal to its posterior probability. Our quantitative results are reported in Fig. 9. Fig. 8 shows the performance of our Bayesian filtering framework at each time frame.

IV. Conclusion

We present a useful representation of an indoor environment that describes the environment by a set of meaningful planes — the ground plane and a set of planar walls that are perpendicular to the ground plane but not necessarily to each other. By analyzing 2D image features, a set of hypotheses about the 3D structure of the local environment can be generated. We use a Bayesian filtering framework to evaluate the set of hypotheses using information accumulated through motion and remove hypotheses with low posterior probability. Since many details of the environment, such as openings and intersections, are not visible in the image from a distance, we propose a continual, incremental process for transforming a current environmental-structure hypothesis into children hypotheses describing the same environment in more detail. In addition to introducing qualitative structural improvements through new children hypotheses, we use the information in current observations to refine the quantitative precision of existing hypotheses. Our experimental results suggest that our method is capable of accurately modeling a variety of indoor environments, including L-intersections, T-intersections and +-intersections. Since our method only relies on simple geometric and probabilistic inferences, our method runs in real-time, and it avoids the need for extensive prior training or the Manhattan-world assumption, which makes it practical and efficient for a navigating robot.

References

[1] O. Barinova, V. Konushin, A. Yakubenko, K. Lee, H. Lim, and A. Konushin. Fast automatic single-view 3-d reconstruction of urban scenes. *ECCV*, pages II: 100–113, 2008.
[2] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *CVIU*, 110:346–359, 2008.
[3] N. Cornelis, K. Cornelis, and L. V. Gool. Fast compact city modeling for navigation pre-visualization. *CVPR*, 2006.
[4] A. J. Davison. Real-time simultaneous localization and mapping with a single camera. *ICCV*, 2003.
[5] E. Delage, H. Lee, and A. Y. Ng. A dynamic Bayesian network model for autonomous 3d reconstruction from a single indoor image. *CVPR*, pages 2418–2428, 2006.
[6] A. Flint, C. Mei, D. Murray, and I. Reid. Growing semantically meaningful models for visual slam. *CVPR*, 2010.
[7] A. Flint, D. Murray, and I. Reid. Manhattan scene understanding using monocular, stereo, and 3d features. *ICCV*, 2011.
[8] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Reconstructing building interiors from images. *ICCV*, 2009.
[9] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.
[10] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. *ICCV*, 2009.
[11] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. *ICCV*, 2005.
[12] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, October 2007.
[13] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *CVPR*, 2:2137–2144, 2006.
[14] G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. *ISMAR*, 2007.

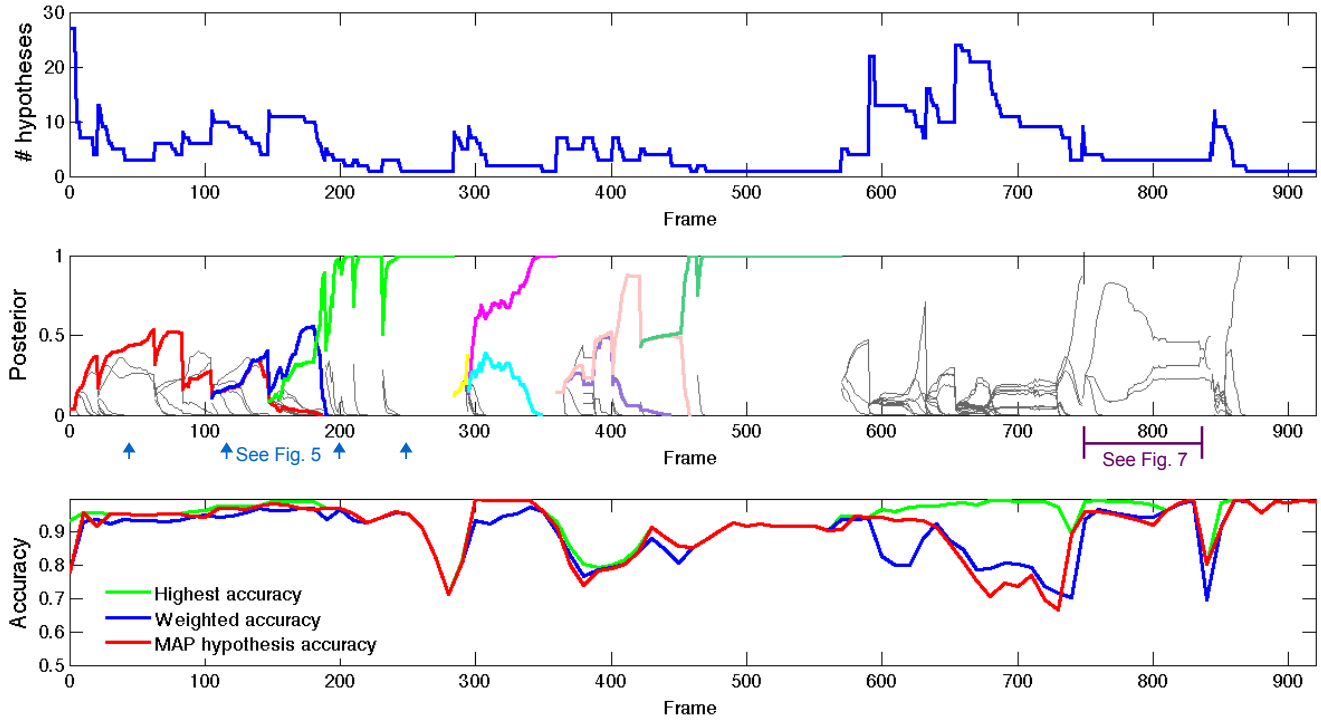


Fig. 8. Bayesian filtering results of Dataset L. (Best viewed in color). The horizontal axis of the graphs is the time frame of the video. **The top graph** shows the number of hypotheses active in each frame. Notice that our method efficiently focuses on a small set of plausible hypotheses (at most 27 hypotheses in this case) at each frame. In our experiments, a set of children hypotheses are generated every 20 frames or when less than 70% of the current image is explained by existing hypotheses. **The second graph** shows the posterior probability of the hypotheses at each frame. Every time when a set of children hypotheses are generated, the posterior probability of the best hypothesis drops. We highlighted several hypotheses to illustrate our method. The red hypothesis is a simple three-wall hypothesis where each wall intersects with its adjacent walls. Both the blue and the green hypotheses (the correct hypothesis) are children hypotheses of the red one with the correct opening structure with different width generated at frame 105 and frame 147 respectively. Thus, the posterior probability of the red hypothesis started to decrease and eventually it was removed from the set, while both the blue and green hypotheses had a higher and higher posterior probability. As the robot approached the opening, more point features around the opening were observed and thus, the probability of the blue hypothesis dropped and was removed from the set at frame 190. Similar phenomena occur several times throughout the video. **The third graph** plots the accuracy of the system at each time frame (the accuracy is evaluated every 10 frames). The green line shows the maximum accuracy among the existing hypotheses. The blue line shows the *weighted accuracy* among of the existing hypotheses, where the weight of each hypothesis is equal to its posterior probability. The red line shows the accuracy of the hypothesis with the maximum posterior probability. The Bayesian filter identifies the best hypothesis for most of the time. Notice a good children hypothesis was generated at frame 654 with a low prior, but as the robot approaches the opening and accumulates more evidences along the travel, the correct hypothesis becomes the winner at frame 737

Dataset	Dataset L	Dataset +	Dataset T 1	Dataset T 2	Overall
MAP hypothesis accuracy	91.00%	94.23%	92.71%	96.38%	93.76%
weighted accuracy	90.04%	92.77%	92.21%	95.37%	92.80%
number of frames	900	300	410	360	1970

Fig. 9. Accuracy. To the best of our knowledge, there is no directly comparable work but in [25], quantitative comparison with [12] and [10] demonstrates our approach has higher accuracies on most environments. Accuracies of both [12] and [10] are about 80% on similar datasets.

[15] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. *CVPR*, 2009.

[16] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. *CVPR*, 2010.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[18] D. Munoz, J. A. Bagnell, N. Vandapel, and M. Hebert. Contextual classification with functional max-margin markov networks. *CVPR*, 2009.

[19] P. Newman, D. Cole, and K. Ho. Outdoor SLAM using visual appearance and laser ranging. *ICRA*, 2006.

[20] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Trans. PAMI*, 26(6):756–770, 2004.

[21] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3D reconstruction from video. *IJCV*, 78(2-3):143–167, 2008.

[22] A. Saxena, M. Sun, and A. Ng. Make3d: learning 3d scene structure from a single still image. *IEEE Trans. PAMI*, 30:824–840, 2009.

[23] J. Shi and C. Tomasi. Good features to track. *CVPR*, 1994.

[24] J. Tavares and A. Padilha. A new approach for merging edge line segments. *In 7 Congresso Portuguls de Reconhecimento de Padres, Aveiro,*, 1995.

[25] G. Tsai, C. Xu, J. Liu, and B. Kuipers. Real-time indoor scene understanding using Bayesian filtering with motion cues. *ICCV*, 2011.

[26] H. Wang, S. Gould, and D. Koller. Discriminative learning with latent variables for cluttered indoor scene understanding. *ECCV*, 2010.

[27] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. *ICCV*, 2009.