Real-time Indoor Scene Understanding using Bayesian Filtering with Motion Cues

Grace Tsai[†], Changhai Xu[‡], Jingen Liu[†], Benjamin Kuipers[†]
[†]Dept. of Electrical Engineering and Computer Science, University of Michigan
[‡]Dept. of Computer Science, University of Texas at Austin
[†]{gstsai,liujg,kuipers}@umich.edu [‡]changhai@cs.utexas.edu

Abstract

We present a method whereby an embodied agent using visual perception can efficiently create a model of a local indoor environment from its experience of moving within it. Our method uses motion cues to compute likelihoods of indoor structure hypotheses, based on simple, generic geometric knowledge about points, lines, planes, and motion. We present a single-image analysis, not to attempt to identify a single accurate model, but to propose a set of plausible hypotheses about the structure of the environment from an initial frame. We then use data from subsequent frames to update a Bayesian posterior probability distribution over the set of hypotheses. The likelihood function is efficiently computable by comparing the predicted location of point features on the environment model to their actual tracked locations in the image stream. Our method runs in real-time, and it avoids the need of extensive prior training and the Manhattan-world assumption, which makes it more practical and efficient for an intelligent robot to understand its surroundings compared to most previous scene understanding methods. Experimental results on a collection of indoor videos suggest that our method is capable of an unprecedented combination of accuracy and efficiency.

1. Introduction

For an embodied agent to act effectively, it must perceive its local environment. Visual perception is obviously important for biological agents, and for artificial agents it has many advantages — cost, field of view, and spatial and temporal bandwidth — over other sensors such as laser, sonar, touch or GPS. By focusing on vision as a sensor for an artificial agent, we consider the input to visual perception to be a stream of images, not simply a single image. 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. Visual processing must be done in real time, to keep up with the agent's needs.

There has been impressive recent work on scene understanding and on the derivation of depth maps from single images of indoor and outdoor scenes [9, 11, 5, 13, 19, 8, 14, 23, 1]. These methods typically depend on carefully trained prior knowledge linking local image properties to a classification of local surface orientation [9, 11, 8], to depth of surfaces in the environment [19], or to semantic labels and thence to depth [14, 23]. Using prior training knowledge with relevant domain specific examples makes these methods difficult to generalize to different environments, especially indoor environments. In addition, real-time performance may be difficult to achieve when evaluations at pixel or superpixel level are involved.

Both Structure-from-Motion [7, 17, 3, 18] and Visual SLAM methods [4, 16, 6, 12] are relatively mature methods that use a stream of visual observations to 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 carrying out actions must then be constructed from the point cloud. Methods that combine 3D point cloud and image data for semantic segmentation has been proposed (e.g. [24]). However, these methods are computationally intensive, making them difficult to apply in real time without specialized GPU hardware support.

We present an efficient method for generating and testing plausible hypotheses in the form of geometric structure models of the local environment. For this paper, we restrict our attention to indoor "floor-wall" environments as described by Delage, et al [5, 1]. These environments need not satisfy the "box" assumption [8, 23] or the Manhattan-world assumption [13]: walls are planes perpendicular to the ground plane, but not necessarily to each other. We do not assume that the ceiling is parallel with the floor, based on the fact that the ground plane is highly relevant to a mobile agent, but the ceiling is not. Indeed, our current implementation does not model the ceiling at all. We present an efficient geometric method to generate a collection of plausible ground-wall boundary hypotheses about the 3D struc-

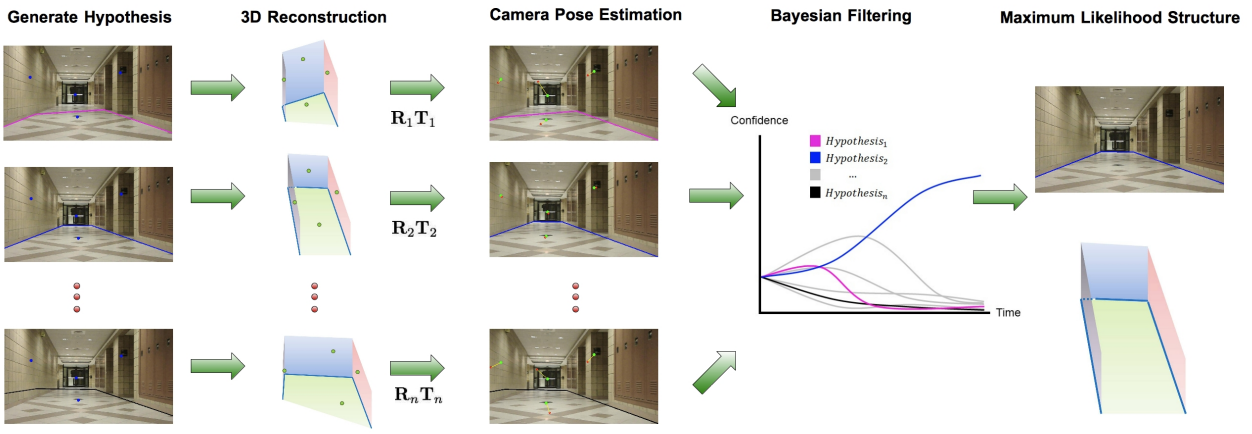


Figure 1. Our proposed framework. The first step is to generate a set of ground-wall boundary hypotheses in the first frame of the video. Given any hypothesis, a static model of the 3D planar environment is computed, and the trajectory of camera motion is determined based on the 3D static model and tracked features. Hypotheses are evaluated based on their prediction error using a Bayesian filter, and the hypothesis with the maximum posterior probability is selected.(Best viewed online, magnified, and in color.)

ture of the environment from a single frame. Our method then uses information from the video stream to identify and update the Bayesian posterior probability distribution over the set of ground-wall boundary hypotheses.

Our main contribution to indoor scene understanding is a method using motion cues to compute likelihoods of hypotheses, based on simple, generic geometric knowledge about points, lines, planes, and motion. Our method relies on knowledge of indoor structure that generalizes to all indoor environment, unlike [8, 23, 13] that use the Manhattan-world assumption. Furthermore, our method avoids the need for extensive prior training with relevant domain-specific examples [9, 11, 5, 13, 19, 8, 14, 23]. In addition, our method runs in real-time since it avoids costly evaluation at pixel or superpixel level as in most single-image approaches, and it avoids costly optimization as in Structure-from-Motion. Thus, unlike other work in scene understanding, our method is practical to apply on artificial agents, which is important for developing useful computer vision tools for areas like robotics and AI.

2. 3D Model Construction and Evaluation

In this section, we describe our method for constructing a geometric model of the 3D environment under a set of hypotheses, and then finding the Bayesian posterior probability distribution over that set of hypotheses, given the observed data (Figure 1).

We assume that the environment consists of a ground plane and planar walls that are perpendicular to the ground plane but not necessarily to each other. We view the environment using a calibrated camera moving through the environment with fixed height and fixed known pitch and roll with respect to the ground plane. In this paper, we assume that the pitch and roll are known as zero. Under these assumptions, we can determine the 3D coordinates in the camera frame of any image point lying on the ground plane (Section 2.2).

We define a ground-wall boundary hypothesis as a polyline extending from the left to the right boarder of the image (similar to [5]). The initial and final segments may lie along the lower image boundary. Vertical segments correspond to occluding edges between planar walls. This paper considers only the case of three-wall hypotheses with no occluding edges. Section 2.1 describes the generation of a set of plausible ground-wall boundary hypotheses.

With the additional assumption of a specific ground-wall boundary hypothesis, we can infer the wall equations, and hence the 3D coordinates of any image point lying on a wall (Section 2.2). This allows us to construct a 3D model of the environment in the camera frame of reference, relative to a given ground-wall boundary hypothesis.

We select a set of image point features that can be tracked reliably through the frames of the video. This selection and tracking step does not depend on the ground-wall boundary hypothesis. Transforming a hypothesized 3D environment model from the camera frame to the world frame of reference, the model becomes static across the frames of the video, so the tracked points can be used to estimate camera motion from frame to frame in the video (Section 2.3).

At this point, relative to each ground-wall boundary hypothesis H_i, we have both a static model of the 3D environment and knowledge of camera motion. Using these, we can predict the motion of the image feature points over time, and compare these predictions with the observations $\mathbf{O}_t$. This comparison defines the likelihood $p(\mathbf{O}_t|H_i)$ of the observation given the hypothesis, and allows us to up-

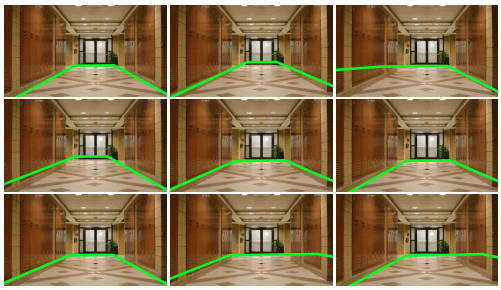


Figure 2. Examples of ground-wall boundary hypotheses.

date the Bayesian posterior probability distribution over the set of ground-wall boundary hypotheses (Section 2.4).

2.1. Ground-Wall Boundary Hypotheses

A ground-wall boundary hypothesis is defined as a polyline extending from the left to the right borders of the image, similar to [5]. The enclosed region of the polyline with the image lower border defines the ground plane. A non-vertical line segment in the polyline represents a wall plane, and vertical line segments represent occluding edges. To simplify the problem in this paper, we focus on hypotheses that consist of at most three walls (i.e. left, end, and right walls) and the ground plane with no occluding edges.

To generate the hypotheses, we start by extracting line segments below the camera center. Since the camera is always placed above the ground plane, all the lines within a feasible hypothesis are below the camera center. We remove vertical line segments because vertical lines imply occluding edges. Non-vertical line segments are divided into three categories (i.e. left, end and right) based on their slopes in the image. A set of hypotheses can be automatically generated by selecting and linking at most one line from each category. However, some hypotheses are infeasible in the physical world and thus, are systematically excluded from the set. Hypotheses with wall intersections outside the image borders are excluded because they violate the perspective geometry of indoor structures. In addition, a 3-wall hypothesis is excluded if its left and right walls intersect in front of the end wall. Furthermore, we define the *edge support* of a hypothesis to be the fraction of the length of the ground-wall boundary that consists of edge pixels. Hypotheses with edge support below a threshold are excluded. Examples of our ground-wall boundary hypotheses are shown in Figure 2.

2.2. Ground-Plane and Wall-Plane Points in 3D

In the camera space, the 3D location $\mathbf{P}_i = (x_i, y_i, z_i)^T$ of an image point $\mathbf{p}_i = (u_i, v_i, 1)^T$ is related by $\mathbf{P}_i = z_i \mathbf{p}_i$ for some z_i. If the point $\mathbf{P}_i$ lies on the ground plane with normal vector $\mathbf{n}_g$, the exact 3D location can be determined by the intersection of the line and the ground plane:

$$h = \mathbf{n}_g \cdot \mathbf{P}_i = z_i \mathbf{n}_g \cdot \mathbf{p}_i \tag{1}$$

where h is the distance of the optical center of the camera to the ground (camera height).

By setting the normal vector of the ground plane to $\mathbf{n}_g = (0, 1, 0)^T$, the 3D location of any point on the ground plane with image location $\mathbf{p}_i$ can be determined by

$$\mathbf{P}_i = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = h \begin{pmatrix} u_i/v_i \\ 1 \\ 1/v_i \end{pmatrix}. \tag{2}$$

The camera height is set to $h = 1$ without loss of generality.

Within a given hypothesis, a wall plane equation is determined by its corresponding ground-wall boundary line segment. We determine the normal vector of a wall plane $\mathbf{n}_w$ based on the ground-wall boundary line segment,

$$\mathbf{n}_w = \mathbf{n}_g \times \mathbf{v}_b \tag{3}$$

where $\mathbf{v}_b$ is the 3D direction vector of the boundary line segment. If a point lies on the wall plane with position $\mathbf{p}_j$ in the image space, its 3D location can be determined by

$$d_w = z_j \mathbf{n}_w \cdot \mathbf{p}_j = \mathbf{n}_w \cdot \mathbf{P}_j \tag{4}$$

where d_w can be obtained by any point on the boundary line. Notice that if a point lies on the ground-wall boundary line, both Equation 1 and 4 must be satisfied.

Based on the geometry described above, we determine the equations of the wall planes, as well as the 3D location of any given point feature on the image under a ground-wall boundary hypothesis. Once we know the 3D ground-wall structure in the camera space, we can easily transform it into the world frame of reference with the origin at $(0, h, 0)^T$ and the zero pan direction along the x-axis.

2.3. Camera Motion in the World Frame

To estimate the camera motion, we select a set of point features and track them across frames. Any method can be used but in this paper, we use KLT [20] tracking because it is more efficient than SIFT [15] and SURF [2], and it works well in our experiments.

Since the 3D locations of the feature points are static in the world coordinates, camera motion between two frames can be estimated by observing the 2D tracked points. This camera motion estimation is equivalent to estimating the rigid body transformation of the 3D point correspondences from a still camera under the camera coordinate system. In this paper, we assume that the camera is moving parallel to the ground plane so the estimated rigid body transformation of the points contains three degrees of freedom, $(\Delta x, \Delta z, \Delta\theta)$, where Δx and Δz are the translation and $\Delta\theta$ is the rotation around y-axis.

Based on the point correspondences in the image space, we reconstruct the 3D locations of the point features for both frames individually under a ground-wall boundary hypothesis. Given two corresponding 3D point sets in the camera space, $\{\mathbf{P}_i = (x_i^P, y_i^P, z_i^P)^T\}$ and $\{\mathbf{Q}_i = (x_i^Q, y_i^Q, z_i^Q)^T\}$, $i = 1...N$, in two frames, the rigid-body transformation is related by $\mathbf{Q}_i = \mathbf{R}\mathbf{P}_i + \mathbf{T}$ where $\mathbf{R}$ is the rotation matrix, and $\mathbf{T}$ is the 3D translation vector. The rotation matrix $\mathbf{R}$ has one degree of freedom, of the form

$$\mathbf{R} = \begin{pmatrix} \cos(\Delta\theta) & 0 & \sin(\Delta\theta) \\ 0 & 1 & 0 \\ -\sin(\Delta\theta) & 0 & \cos(\Delta\theta) \end{pmatrix} \tag{5}$$

and the translation vector $\mathbf{T} = (t_x, 0, t_z)^T$ has two degrees of freedom. In order to estimate the three degrees of transformation, the point correspondences are projected onto the ground plane which are denoted as $\{\mathbf{P'}_i = (x_i^P, h, z_i^P)^T\}$ and $\{\mathbf{Q'}_i = (x_i^Q, h, z_i^Q)^T\}$. The rotation matrix can be estimated by the angular difference between two corresponding vectors, $\overrightarrow{\mathbf{P'}_i\mathbf{P'}_j}$ and $\overrightarrow{\mathbf{Q'}_i\mathbf{Q'}_j}$. Our estimated $\Delta\theta$ is thus the weighted average of the angular differences,

$$\cos(\Delta\theta) = \frac{1}{\sum\omega_{ij}} \sum_{i\neq j} \omega_{ij} \frac{\overrightarrow{\mathbf{P'}_i\mathbf{P'}_j} \cdot \overrightarrow{\mathbf{Q'}_i\mathbf{Q'}_j}}{\|\overrightarrow{\mathbf{P'}_i\mathbf{P'}_j}\|\|\overrightarrow{\mathbf{Q'}_i\mathbf{Q'}_j}\|} \tag{6}$$

where ω_{ij} is defined as

$$\omega_{ij} = \frac{(1/z_i^P + 1/z_i^Q)}{2} \frac{(1/z_j^P + 1/z_j^Q)}{2}. \tag{7}$$

Since the reconstructed 3D positions of distant points are less accurate, they are given lower weights.

The translation vector $\mathbf{T}$ is then estimated by the weighted average of the differences between $\mathbf{R}\mathbf{P'}_i$ and $\mathbf{Q'}_i$,

$$\mathbf{T} = \frac{1}{\sum\omega_i} \sum_{i=1}^{N} \omega_i(\mathbf{Q'}_i - \mathbf{R}\mathbf{P'}_i) \tag{8}$$

where $\omega_i = (1/z_i^P + 1/z_i^Q)/2$.

The pose change of the camera in the world frame of reference can be determined based on the estimated rotation $\mathbf{R}$ and translation $\mathbf{T}$ of the 3D features points locations under the camera coordinate. If we set the camera location of one frame to $(0, 0, h)^T$ with zero pan along the x-axis, the camera pose, $(x_c, z_c, \theta_c)^T$, of that frame is $(0, 0, 0)^T$. Then, the pose of the camera in the other frame becomes $(t_x, t_z, \Delta\theta)^T$, where t_x and t_z are the x and z components of $\mathbf{T}$ in the camera coordinates.

2.4. Evaluating Hypotheses

At this point, relative to each ground-wall boundary hypothesis, we have both a static model of the 3D environment and knowledge of camera motion. Using these, we can predict the motion of the image feature points over time, and compare these predictions with the tracked features.

Ground-wall boundary hypotheses are evaluated by Bayesian filtering. We define the first frame of the video as our reference frame and compare it with each of the other frames. For each frame, the likelihood for each hypothesis with respect to the reference frame is computed. Using a Bayesian filter allows us to accumulate the likelihoods from all the frames in order to select the hypothesis with the maximum posterior probability at the end of the video.

Given a set of hypotheses, $\{H_i\}$, $i = 1...N$, the posterior distribution over the hypotheses at time step m can be expressed by Bayes rule,

$$p(H_i|\mathbf{O}_1, \mathbf{O}_2, ..., \mathbf{O}_m) \propto p(H_i) \prod_{t=1...m} p(\mathbf{O}_t|H_i) \tag{9}$$

where $\mathbf{O}_t$ is the set of features whose tracked and predicted locations are compared at time step t. If we have no information about the correctness of the hypotheses from the reference frame, the prior probability $p(H_i)$ in Equation 9 is uniformly distributed over all the hypotheses.

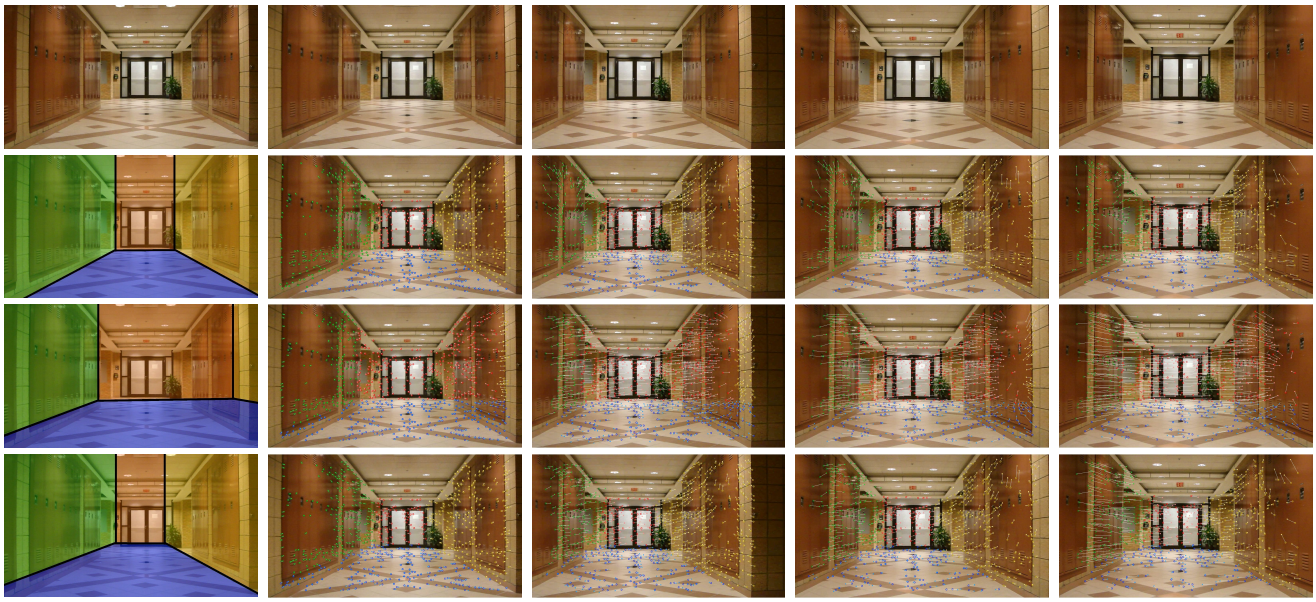
For each time step, the observation $\mathbf{O}_t$ is a set of evidence from the feature points, $\{o_1^t, o_2^t, ...o_n^t\}$, observed in frame t. The likelihood of an individual point o_j^t at image location $\mathbf{L}(o_j^t)$ is modeled by a normal distribution with mean at the predicted image location $\hat{\mathbf{L}}(o_j^t)$ in frame m. $\hat{\mathbf{L}}(o_j^t)$ and $\mathbf{L}(o_j^t)$ are related by the rotation matrix $\mathbf{R}$ and translation vector $\mathbf{T}$ as described in Section 2.3. Since the likelihood is only depending on the distances between $\mathbf{L}(o_j^t)$ and $\hat{\mathbf{L}}(o_j^t)$, the individual likelihood is equivalent to modeling the prediction error between the two with a zero mean normal distribution with variance σ. By combining the likelihoods from individual points, the likelihood of hypothesis H_i at time step t is,

$$p(\mathbf{O}_t|H_i) \propto \prod_{j=0}^{n} \exp \frac{-||\hat{\mathbf{L}}(o_j^t) - \mathbf{L}(o_j^t)||^2}{2\sigma^2}. \tag{10}$$

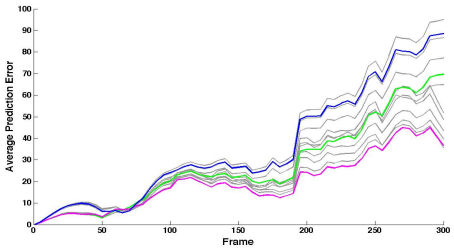
3. Results

We tested our approach on 11 videos with resolution 1280×720 in various indoor environments (Figure 5). We included corridors that violate the Manhattan-world assumption, corridors with glass walls, reflections and partially occluded edges, and rooms with various sizes. The number of frames in each video ranges from 300 to 380 and the hypotheses are evaluated every 5 frames. The overall motion in each video is about 3 meters moving forward with slight direction changes. Frames from one video in our dataset are shown in the top row of Figure 3(a).

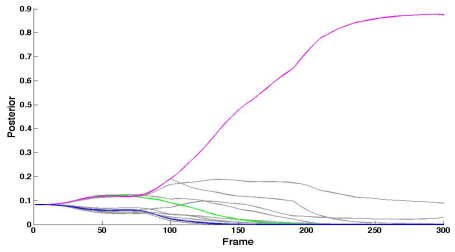
The efficiency of our method is shown in Table 1. The computational time is related to the number of feature points



(a)



(b) Prediction Error



(c) Posterior Probability Distribution after each Frame

Figure 3. (a) The top row are frames 1, 50, 100, 150 and 200 from one of our datasets. The bottom rows are examples of our hypothesis generated in the first frame (first column) and the predicted locations (crosses) and tracked locations (circles) of the feature points on each frame. The discrepancy of the locations are the white lines. (b) The overall trend of the prediction error increases with the motion due to feature tracking quality. The number of existing tracked features decreases as the motion increases and these features are mostly from the distant area in the first frame which has low resolution. The abrupt increase at frame 185 is because of the sudden camera movement which reduces the tracking quality. (c) All hypotheses are equally likely in the first frame. Hypotheses with low accuracy drop significantly in the first few frames, while the one with the highest accuracy gradually stands out among the rest. The most accurate hypothesis need not to be the one with minimum prediction error all the time in order to get the maximum posterior probability in the end. (Best viewed in color)

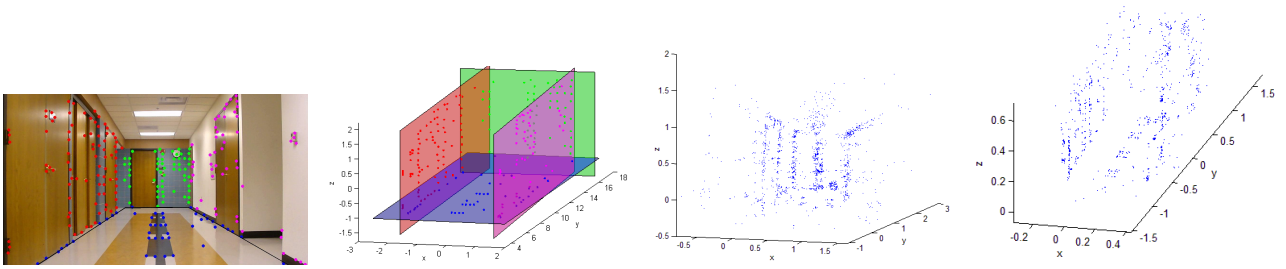
and the number of hypotheses as shown in the table. Our algorithm runs in real-time and the computational time (in C/C++ using an Intel Core 2 Quad CPU 2.33GHz) is less than the video length.

For each test video, we manually labeled the ground truth classification of the planes (i.e. three walls, ground plane and ceiling plane) for all pixels in the first frame in order to evaluate our results quantitatively. We define the accuracy of a hypothesis being the percentage of the pixels that have the correct classification in the first frame. Since the ceiling plane is not included in our hypotheses, we omitted the ceiling pixels from our evaluation.

Figure 3 shows our results in hypothesis generation, motion prediction and Bayesian filtering. [1] Even though the overall error increases with motion due to the quality of feature tracking, hypotheses that are closer to the actual indoor structure have relatively low errors compared to others since the hypotheses are evaluated based on the same set of feature points. Figure 5 shows our performances in various indoor environments in which we demonstrated our capability to deal with non-Manhattan world structures, as well as noisy feature points.

To compare our approach to state-of-the-art methods, we apply the indoor classifier in [10] and the box layout estima-

[1] Since our ground-wall boundary hypotheses do not model the ceiling plane, feature points from the ceiling plane will be misleading in the evaluation. These points are excluded using essentially the technique used to identify the ground plane in Section 2.1.



(a) Input video (b) Our method (c) Bundler [21] (d) PTAM [12]

Figure 4. Comparison between our approximate 3D planar model with other multi-image reconstruction methods. (Best viewed in color) (a) First frame of our input video and the hypothesis with maximum posterior probability (black line) determined by our framework using frame 1 to 300. Based on the hypothesis, feature points are classified into ground plane (blue) and left (red), front (green) and right (pink) walls. (b) 3D planar model and 3D locations reconstructed using the geometry described in Section 2.2 given the image location of the hypothesis. (c) 3D point cloud reconstructed by Bundler [21] using frame 5, 10, ..., 300 in the video. Only the distant area of the corridor are reconstructed because Bundler only considered SIFT feature points that frequently appear among the image set. (d) 3D point cloud reconstructed by PTAM [12] using frame 1 to 300 where the first 10 frames are used for initialization.

Datasets	EECS Building	Library 1	Locker Room	Non-parallel 1	Non-parallel 2	Basement
Our Method	**85.78%**	**91.13%**	**97.91%**	**89.61%**	84.89%	**99.71%**
[10]	85.49%	88.13%	71.49%	86.23%	**85.44%**	72.82%
[10]+MRF	84.57%	83.10%	83.18%	86.47%	60.16%	77.99%
[8]	79.15%	83.32%	87.72%	53.10%	66.11%	89.79%
Datasets	Study Room	Library 2	Glass Wall	Object	Two Walls	Average
Our Method	**95.59%**	**88.39%**	85.56%	**94.73%**	**97.71%**	**92.09%**
[10]	76.43%	88.38%	58.87%	94.45%	92.02%	82.07%
[10]+MRF	72.90%	84.78%	64.39%	88.62%	91.80%	79.62%
[8]	89.23%	78.42%	**87.55%**	88.16%	95.63%	81.96%

Table 2. Classification accuracy. We compared our results quantitatively with [8] and [10], and we further extend [10] to incorporate temporal information in order to make a fair comparison (see text for more detail). Note that while evaluating [10], the most likely label is assigned to each pixel and pixels with most likely label "sky", "porous" or "solid" are excluded in the evaluation. While evaluating [8], pixels with label "ceiling" are excluded in the evaluation.

Datasets	NF	NH	VL	CT
EECS Building	256	15	10 s	5.77 s
Library 1	233	12	10 s	6.18 s
Locker Room	245	16	10 s	6.70 s
Non-parallel 1	286	13	10 s	6.49 s
Non-parallel 2	268	8	11.67 s	7.60 s
Basement	282	9	10 s	7.27 s
Study Room	248	15	10 s	8.18 s
Library 2	250	20	12.67 s	7.84 s
Glass Wall	403	10	10 s	7.34 s
Object	336	15	10 s	7.50 s
Two Walls	242	10	10 s	5.37 s

Table 1. Computational time analysis. (NF: number of features; NH: number of hypotheses; VL: video length; CT: computational time)

tor in [8] to the first frame of each video. Furthermore, we extended the method in [10] by applying it to the same subset of frames that our method used (e.g. 60 frames out of 300), and combined the labels across frames using a spatial-temporal Markov Random Field linking superpixels within and across the frames, similar to [24]. We refer to these results as "[10]+MRF". Notice that adding temporal information to [10] does not necessarily improve the result in the first frame because incorrect labels in later frames affect the label in the first frame.

Our quantitative results are reported in Table 2. Applying all four methods to our datasets, we obtained a mean accuracy of 92.09% for our method, a mean accuracy of 82.06% for [10] in its original single-image mode, a mean accuracy of 79.62% for [10]+MRF and a mean accuracy of 81.96% in [8]. One reason for this substantial difference is that [10] and [8] depend strongly on training data, which is likely to be specific to certain environments. By contrast, our method applies a very general geometric likelihood criterion to semantically meaningful planar hypotheses. In addition, [8] uses the "box" assumption, while our methods does not require the walls to be perpendicular to each other.

Even though our focus is on scene understanding, we compared our 3D planar model with multiple image reconstruction approaches, Bundler [21] and PTAM [12], as

shown in Figure 4. Bundler [21] has trouble with simple forward motion because it only considered SIFT points that frequently appear among the image set for 3D reconstructions and camera pose estimation. Thus, only the far end of the corridor was reconstructed. Our approximate 3D reconstruction is comparable with [12], but in addition to point clouds, our model provides semantic information about the indoor environments (e.g. walls and ground plane). We also used J-linkage [22] to fit planes to the 3D point clouds from [21] and [12]. These results do not contribute meaningful information for indoor scene understanding, because the plane-fitting process is easily misled by accidental groupings within the point cloud. Our hypothesis-generation process focuses on semantically plausible hypotheses for indoor environments.

4. Conclusion

We have demonstrated a new method for efficiently generating and testing models of indoor environments. We apply single-image geometric methods to an initial frame of a video to propose a set of plausible ground-wall boundary hypotheses for explaining the 3D structure of the local environment. Within the context of each hypothesis, our method estimates camera motion, then uses the 3D structural hypothesis plus camera motion to predict the image-space motion of point features on the walls. A likelihood function for the observed data, given each hypothesis, can then be computed from the stream of data in subsequent frames. The Bayesian posterior probability distribution is updated using these likelihoods from each subsequently analyzed frame in the stream, almost always leading to rapid identification of the best hypothesis. We demonstrate qualitative and quantitative results on videos collected of motion in a variety of indoor environments, including non-Manhattan-world environments and ones with glass walls and windows, shadows, and other difficult image features. Our experimental results suggest that our method is capable of an unprecedented combination of accuracy and efficiency. Our goal is to enable an embodied agent with visual perception to understand its environment well enough to act effectively within it.

Acknowledgements

The authors thank Chia-Wei Luo and Bangxin Hu for help on experiments and data collection, and Silvio Savarese for useful feedbacks. This work has taken place in the Intelligent Robotics Lab 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 from the TEMA-Toyota Technical Center. Coauthor CX is also supported in part by a grant from the National Science Foundation (IIS-0713150) to the University of Texas at Austin.

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. 1

[2] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, 110:346–359, 2008. 3

[3] N. Cornelis, K. Cornelis, and L. V. Gool. Fast compact city modeling for navigation pre-visualization. *CVPR*, 2006. 1

[4] A. J. Davison. Real-time simultaneous localization and mapping with a single camera. *ICCV*, 2003. 1

[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. 1, 2, 3

[6] A. Flint, C. Mei, D. Murray, and I. Reid. Growing semantically meaningful models for visual slam. *CVPR*, 2010. 1

[7] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. 1

[8] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. *ICCV*, 2009. 1, 2, 6, 8

[9] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. *ICCV*, 2005. 1, 2

[10] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, October 2007. 5, 6, 8

[11] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *CVPR*, 2:2137–2144, 2006. 1, 2

[12] G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. *ISMAR*, 2007. 1, 6, 7

[13] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. *CVPR*, 2009. 1, 2

[14] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. *CVPR*, 2010. 1, 2

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 3

[16] P. Newman, D. Cole, and K. Ho. Outdoor SLAM using visual appearance and laser ranging. *ICRA*, 2006. 1

[17] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Trans. PAMI*, 26(6):756–770, 2004. 1

[18] 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. 1

[19] A. Saxena, M. Sun, and A. Ng. Make3d: learning 3d scene structure from a single still image. *IEEE Trans. PAMI*, 30:824–840, 2009. 1, 2

[20] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593 – 600, 1994. 3

[21] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring image collections in 3d. *SIGGRAPH*, 2006. 6, 7

[22] R. Toldo and A. Fusiello. Robust multiple structure estimation with J-linkage. *ECCV*, 2008. 7

[23] H. Wang, S. Gould, and D. Koller. Discriminative learning with latent variables for cluttered indoor scene understanding. *ECCV*, 2010. 1, 2

[24] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. *ICCV*, 2009. 1, 6, 8

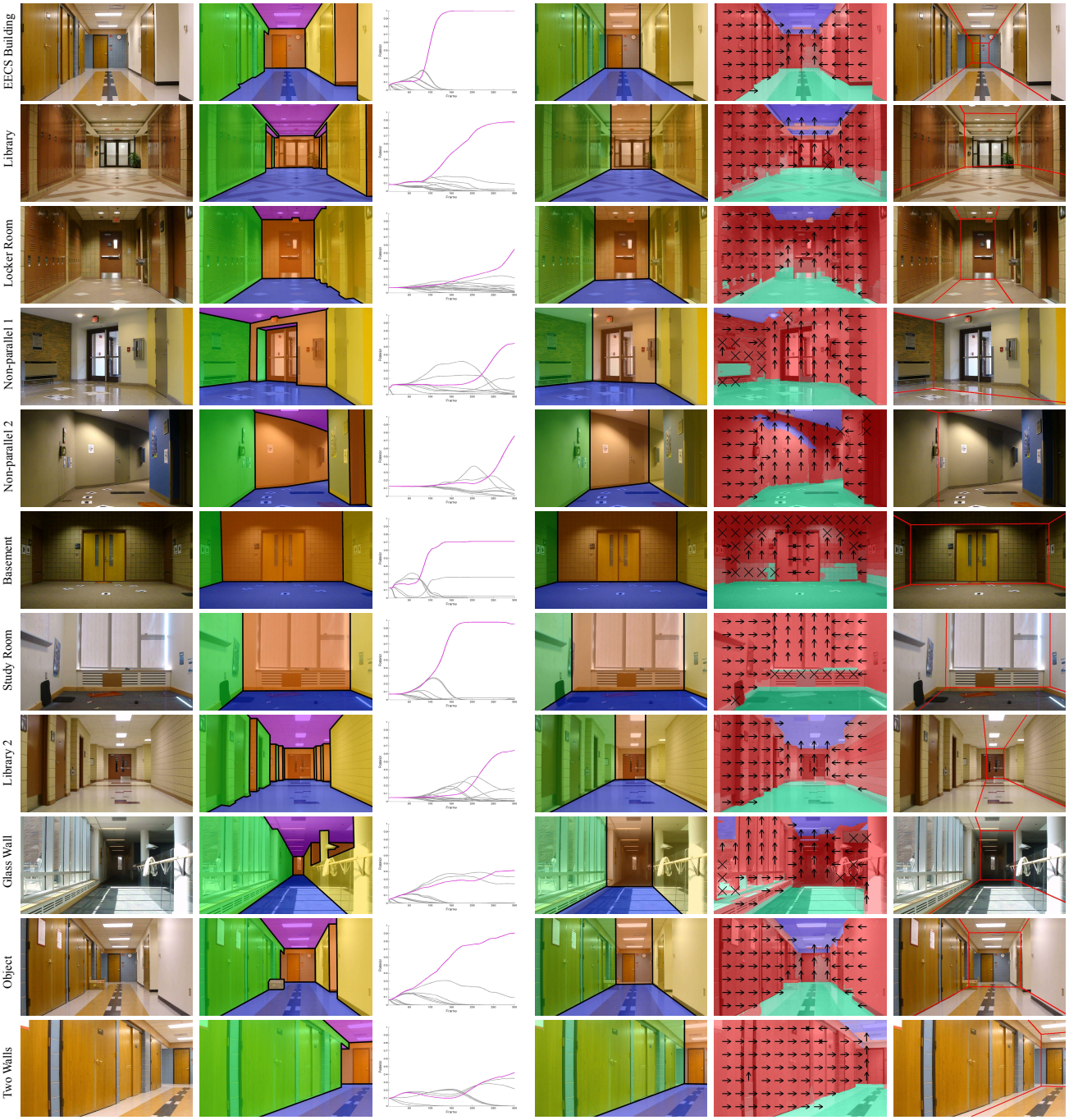


Figure 5. Examples (Best viewed in color). The first column is the first frame of each video and the second column is our manually labeled ground truth. Column three shows the posterior distribution over the hypotheses. The hypothesis with maximum posterior probability at the end of the video is shown in pink. The fourth column shows the hypothesis with maximum posterior probability. The fifth column is the results from [10] on the first frame of the video using their classifiers trained for indoor images and the last column is the results of box layout estimation from [8] on the first frame of the video. Non-parallel 1 and Non-parallel 2 demonstrate our capability to identify non-Manhattan world structures, unlike [8]. Furthermore, our simple ground-wall models enable us to ignore objects that stick out of the wall as in Locker Room, Non-Parallel 1 and Object. Object also demonstrates our capability to deal with a partially occluded ground-wall boundary. Our method is a generalized framework that can deal with any number of walls by generalizing the hypothesis generation (Two Walls). Our method works fairly well even with noisy feature due to reflections (Glass Wall). Compared to [10] and [8], our method generalizes across different environments since we do not rely on any training from the image properties, which can be scene sensitive. We further applied [10] to multiple frames of the videos and built a MRF to combine temporal informations similar to [24]. See Table 2 for quantitative comparisons.