Semantic Visual Understanding of Indoor Environments: from Structures to Opportunities for Action

Grace Tsai, Collin Johnson, and Benjamin Kuipers
Dept. of Electrical Engineering and Computer Science, University of Michigan
{gstsai,collinej,kuipers}@umich.edu

Abstract

We present a two-layer representation of the locally sensed 3D indoor environment. Our representation moves one step forward from capturing the geometric structure of the environment to reason about the navigation opportunities for an agent in the environment. The first layer is the Planar Semantic Model (PSM), a geometric representation in terms of meaningful planes (ground and walls). PSM captures more semantics of the indoor environment than a pure planar model because it represents a richer set of relationships among planar segments. In the second layer, we present the Action Opportunity Star (AOS), which describes the set of qualitatively distinct opportunities for robot action available in the neighborhood of the robot. Our two-layer representation is a concise representation of indoor environments, semantically meaningful to both robot and to human. It is capable of capturing incomplete knowledge of the local environment so that unknown areas can be incrementally learned as observations become available. Experimental results on a variety of indoor environments demonstrate the expressive power of our representation.

1. Introduction

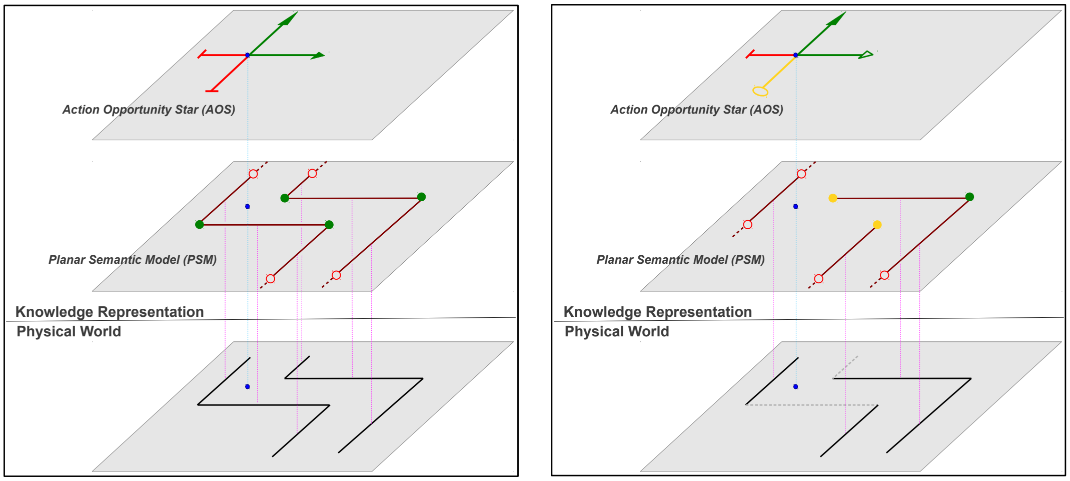
An agent must perceive its local environment to act effectively. By focusing on a vision sensor for a mobile agent, we considered the input to visual perception to be a temporally continuous stream of monocular images, not simply a single image or a collection of atemporal images. The output of visual perception must be a coherent, concise, representation of the agent's surrounding environment, at a granularity that supports the agent to make plans. Moreover, visual processing must be done on-line and real-time to keep up with the robot's needs.

A useful representation for an agent needs to concisely represent both the spatial information of the local environment and the semantic meaning of the environment in terms of the agent's action opportunities. The representation must be capable of capturing incomplete knowledge of the local environment so that unknown areas can be incrementally constructed as observations become available. Since different agents may have different action capabilities, the representation for different agents may differ. While many existing works has been proposed to represent human semantics [27, 10, 15], this paper focuses on representing the semantics of a wheeled robot that navigates in indoor environments. A useful representation for an indoor navigating robot is a concise model that captures the free-space of the environment.

There are many previous work on geometric scene understanding from a single image. A common scene representation in the image space is by labeling each pixel with a local surface orientation or a 3D depth value [13, 14, 20], where a 3D model can then be inferred. These representation are fine-grained and thus, provides no constraints on regularizing the possible 3D structure of indoor environments. A common representation specifically for indoor scenes is the image projection of a 3D planar model [4, 17, 12, 26, 21]. The planar model is concise but due to the limited field of view of a monocular camera, the scene captured in the image does not reflect the robot's immediate surrounding, so it does not provide sufficient information for the robot to make plans. A temporally coherent scene understanding result may be difficult to achieve if each frame is independently processed.

Methods such as Structure-from-Motion [11, 18, 2, 19] and Visual SLAM [3, 5, 16] 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 for navigation must then be constructed from the point cloud. Other methods [8, 6, 7] use the Manhattan-world assumption to reconstruct a planar model of an indoor environment from a collection of images. A planar model is concise and specifies free-space for navigation. However, these methods are off-line and computationally intensive, making them difficult to apply in real-time robot navigation.

In this paper, we present a two-layer representation of


Action Opportunity Star (AOS)
Planar Semantic Model (PSM)
Knowledge Representation
Physical World
Action Opportunity Star (AOS)
Planar Semantic Model (PSM)
Knowledge Representation
Physical World

(a) Two-layer representation with complete knowledge

(b) Two-layer representation with incomplete knowledge

Figure 1. The proposed two-layer representation illustrated on the ground-plane map. (Best viewed in color). Each layer represents a different level of understanding of the local environment. The first layer models the geometric structure of the local environment. The second layer is a pure symbolic representation that describes the opportunities for robot action at a given location, based on the geometric structure determined in the first layer. Both layers are capable of representing incomplete knowledge as shown in (b). In the physical world, black solid lines represent the part of the environment that is observed by the robot, and the gray dashed lines represent the part of the environment that is not observed by the robot. The first layer is the Planar Semantic Model (PSM), which models the geometric structure of the local environment in terms of meaningful planes — the ground plane and a set of walls that are perpendicular to the ground plane but not necessarily to each other. Each wall contains a set of disjoint wall segments (red lines), delimiting where the wall is present and where is an opening. Each wall segment is represented by two endpoints, and each endpoint has its property indicating the level of understanding of the bound of the wall segment. While a *dihedral* endpoint (green dot) provides the full knowledge of the bound of its corresponding wall segment, an *occluding* endpoint (yellow dot) and an *indefinite* endpoint (red hollow dot) provide different types of incomplete knowledge of the wall intersection. The second layer is the Action Opportunity Star (AOS), describing the robot's surrounding environment by a structured set of qualitatively distinct opportunities for robot action. Each opportunity is visualized by an arrow pointing towards its associated direction, and the tip of the arrow reflects its type. The opportunity type reflects different purposes or different levels of understanding of the opportunity. A green arrow is an opportunity that is *observed* and navigable, while a red line is an *unnavigable* opportunity. A green hollow arrow is a navigable opportunity but the actual boundary of the opportunity is only *partially observed*.

the local indoor environment. Each layer represents a different level of understanding of the environment. The first layer (Section 2) models the geometric structure of the local environment in terms of planes from the image stream. The second layer (Section 3) is a pure symbolic representation that describes the opportunities for robot action (navigation), based on the geometric structure in the first layer. Figure 1 illustrates our representation. Building on top of the on-line scene understanding method [25, 22], we demonstrate an efficient method to construct the two-layer representation from a stream of images.

For the first layer, we present the Planar Semantic Model (PSM). PSM is a coarse-grained representation for the local 3D indoor environment, instead of a fine-grained representation like point clouds. PSM describes the environment in terms of a set of meaningful planes — the ground plane and a set of walls that are perpendicular to the ground but not necessarily to each other. Note that PSM is less restrictive than the Manhattan world assumption. In PSM, a wall is a set of disjoint wall segments that are co-planar in 3D. Thus, PSM is a step forward from a pure planar model because it represents richer relationships among planer segments.

For the second layer, we present the Action Opportunity Star (AOS) to describe a set of qualitatively distinctive opportunities for robot action at a given location. An *opportunity* represents a group of trajectories that can be described by the same semantic meaning of the robot's action. AOS captures where each opportunity is valid and the relationships among these opportunities. Since AOS is an abstract representation, if the surrounding PSM at two locations are similar, AOSs extracted at both locations will be the same. We present a method to extracts AOS from PSM.

Our representation is concise and useful to a navigating robot to make plans. It is able to represent incomplete knowledge of the local environment so that unknown areas can be incrementally built as observations become available. Compare to existing scene understanding work, in addition to modeling the geometric structure of the local environment, our representation takes a step forward to reason about the opportunities for robot action. Our representation supports the robot to make plans at different levels. While PSM provides information about the free-space for the robot to precisely generate a trajectory to get from one pose to another, AOS supports the robot to make plans at a higher level, such as turning right at an intersection or going forward (rather than reverse) in a corridor. Moreover, our representation supports topological mapping [1]. For example, AOS makes it easy to detect whether a robot is at a topological place, such as a hallway intersection, or on a path that links two places.

2. Planar Semantic Model

In the first layer, we present the Planar Semantic Model (PSM) to represent the locally sensed 3D indoor environment. PSM is a coarse-grained representation of an indoor environment in terms 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.

The formal definition of PSM, M is,

$$M = \{G, W_1, W_2, W_3, ..., W_n\}. \tag{1}$$

where n is the number of walls in the local environment. There is a one-to-one correspondence between this representation and a set of ground-wall boundary lines in the ground plane (the ground-plane map), represented in the same 3D coordinates. [1]

A wall W_i contains a set of disjoint wall segments that are co-planar in 3D. In the ground-plane map, the wall plane is represented by a line parametrized by (α_i, d_i). $\alpha_i \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right]$ is the orientation of the line which implies the normal direction $\mathbf{N}_i = (\cos\alpha_i, \sin\alpha_i, 0)$ of the 3D wall plane, and $d_i \in \mathbb{R}$ is the directed distance from the origin of the ground-plane map to the line. $\mathbf{N}_i$ and d_i determine the 3D equation of the wall plane.

The bound of each wall segment is defined by two lines that are on the wall plane and are vertical to the ground plane in 3D. By projecting these vertical lines onto the ground-plane map, the wall segment is represented by a pair of endpoints, (E^i_j, E^i_{j+1}), along the corresponding ground-wall boundary line. The formal definition of a wall W_i is,

$$W_i = \langle \alpha_i, d_i, E^i_1, E^i_2, E^i_3, ...E^i_{2m_i} \rangle \tag{2}$$

where m_i is the number of wall segments along wall i. The endpoints are ordered from the left to the right of the canonical view of the wall plane. The ordering specifies which side of the wall is free space and which side is occluded.

Each endpoint E^i_j is represented by,

$$E^i_j = \langle x^i_j, y^i_j, c^i_j \rangle \tag{3}$$

where (x^i_j, y^i_j) represent the location of the endpoint in the ground-plane map, and c^i_j specifies the type of the endpoint. There are three different types of endpoints: *dihedral*, *occluding* and *indefinite*, representing different levels of understanding of the bound of the wall segment. A *dihedral* endpoint corresponds to two visible wall segments, where the location of the endpoint is the intersection point of the two walls. An *occluding* endpoint corresponds to only one visible wall segment. An *indefinite* endpoint is the furthest observed point along its corresponding wall segment, but the actual location of the wall bound has not yet been observed due to occlusions or the end of robot's field of view. While a *dihedral* endpoint provides full knowledge of the bound of its corresponding wall segments, an *occluding* and an *indefinite* endpoint provide different types of incomplete knowledge of the wall intersection.

To extract the PSM, we implemented the method proposed in [22]. The method incrementally generates a set of qualitatively distinct hypotheses about the structure of the environment from 2D image features (e.g. points and lines), and then tests the hypotheses through a Bayesian filter based on their abilities to explain the 2D motion of a set of points tracked over a period of time.

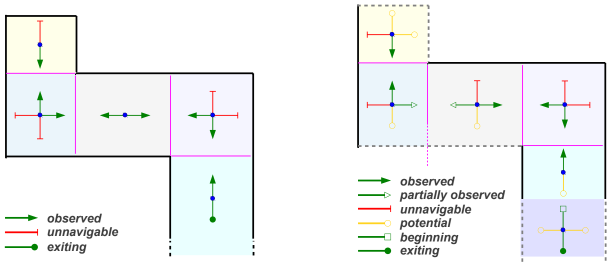
3. Action Opportunity Star

An Action Opportunity Star (AOS) is a qualitative description of the small finite set of opportunities for robot action abstracted from an infinite number of trajectories that are available within the region around the robot (the field of interest). An *opportunity* is an abstraction, representing a group of trajectories that have the same qualitative effect on the robot's state. An opportunity for action is intended to be similar to the concept of an *affordance* [9]. We define a *gateway* as a line segment on the metric map, PSM, that specifies which trajectories belong to an opportunity. All trajectories that cross a particular gateway from the side closer to the robot to the side farther from the robot belong to the same opportunity.

In addition to representing individual opportunities, AOS models the relationships among opportunities in terms of the *paths* they are on. Two opportunities that unambiguously represent opposite directions from the same field of interest are considered to be on the same path. We say that when the robot has exactly two opportunities unambiguously representing opposite directions, then the robot is *on a path*. In any other situation, the robot is *at a place*, which typically requires it to make a decision, selecting among the available opportunities [1]. For example, when the robot is at a T-intersection, it has three opportunities, associated with two paths, one of which passes through the place, while the other ends at that place.

Formally, at a given robot location, the AOS, S is defined by a list of opportunities A_i circularly ordered in the

[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.



(a) w/ complete knowledge (b) w/ incomplete knowledge

Figure 2. Examples of AOS at different locations. (Best viewed in color.) Each opportunity is visualized by an arrow and the arrow tip reflects the opportunity type. Since AOS is an abstract representation, AOSs extracted at all location within a region that has the same surrounding geometric structure are the same. The regions are shown in different colors. (b) Due to the limited field of view of a camera, the robot may have incomplete observations of the environment. Black solid lines: observed; Gray dashed-lines unobserved. AOS is capable of capturing incomplete knowledge.

counter-clockwise direction,

$$S = \{A_1, A_2, ..., A_k\} \tag{4}$$

where k is the number of opportunities around the given location. Each opportunity, A_i, is defined as,

$$A_i = \langle \pi_i, \rho_i, \tau_i, \mathbf{g}_i \rangle \tag{5}$$

where $\pi_i \in \{0, 1, ..., N_p\}$ is the path that the opportunity is on, among the N_p paths that pass through the field of interest. $\rho_i \in \{+, -\}$ is the direction along the path that the opportunity is leading onto. The path π_i and the direction ρ_i specify the relation between opportunity A_i and another opportunity A_j where $\pi_i = \pi_j$ and $\rho_i = -\rho_j$. $\mathbf{g}_i$ is the gateway associated to opportunity A_i, which is a line segment ϕ_i parameterized by two ends $(\mathbf{p}_i^1, \mathbf{p}_i^2)$ in the ground-plane map, and the qualitative traveling direction ψ_i of the opportunity is the normal direction of the gateway pointing away from the robot. τ_i specifies the type of the opportunity.

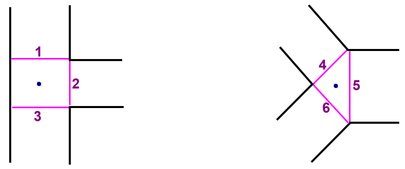
There are six different types of opportunity: *observed*, *partially observed*, *unnavigable*, *potential*, *beginning* and *exiting*, representing different purposes or different levels of understanding of the opportunity. An *observed* opportunity is navigable and leads the robot into or out of a path intersection where more than two unaligned opportunities are presented. Both ends of an *observed* gateway, $\mathbf{p}_i^1$ and $\mathbf{p}_i^2$, are fully determined. A *partially observed* opportunity plays the same role as an *observed* opportunity, except only one of the two gateway-ends is determined, leaving the actual width of the gateway undetermined. Thus, this opportunity is navigable but it contains incomplete knowledge. An *unnavigable* opportunity prohibits the robot to travel along the path due to obstacles. A *potential* opportunity exists when an opportunity leads the robot to an unobserved region, and thus, its navigability is unknown. As the robot acquires more observations around this opportunity, it can become an *observed* or an *unnavigable* opportunity. Similar to a *potential* opportunity, a *beginning* opportunity crosses the boundary between observed and unobserved regions, except a *beginning* opportunity leads the robot into the observed region. A *beginning* opportunity only occurs at the beginning of an episode (the first few frames of an image stream), where the robot only observes the environment in front of it, instead of its surrounding environment. Thus, a *beginning* opportunity is navigable but contains incomplete knowledge. The ends of the associated gateway of a *potential* or a *beginning* opportunity are specified so that the right side of the vector $\overrightarrow{\mathbf{p}_i^1\mathbf{p}_i^2}$ is the observed region while the left side of the vector is unobserved. While the above five opportunity types reflect the structure of the local environment, an *exiting* opportunity leads the robot out of the robot's field of interest. This type of opportunity usually appears when a robot is traveling on a long corridor where going forward and turning backward are the only two possible qualitative actions. Figure 2 shows examples of AOS in different situations.

3.1. Extracting Opportunities from PSM

Since a gateway links its associated opportunity to the geometric structure of the environment, we start by extracting a set of gateways within the field of interest, given a robot location. In this section, we determine only the gateway $\mathbf{g} = (\mathbf{p}^1, \mathbf{p}^2)$ and the type τ of each opportunity. In Section 3.2, we determine the other two elements, π and ρ, of the opportunities by comparing the gateways to see which opportunities are well-aligned to be on the same path.

Given the PSM, there are two major steps to collect the set of gateways within the field of interest. The first step extracts gateways that reflect the structure of the surrounding environment. All of these gateways have at least one end anchored at a PSM endpoint. The type of the PSM endpoint at which a gateway is anchored affects the type of its associated opportunity. Possible opportunity types at this stage are *observed*, *partially observed*, *potential* and *beginning*. The rest of this section describes how we extract these gateways in detail. Given the gateways from the first step, the second step extracts *exiting* opportunities from regions that are not explained by either a PSM wall segment or an existing gateway, through a circular scan around the field of interest. The gateway of an *exiting* opportunity is perpendicular to, and intersects with, a PSM wall segment or another gateway.

By using each PSM endpoint as an anchor $\mathbf{p}^1$, four gateways can be extracted with their directions $\hat{\mathbf{g}}$ parallel or perpendicular to the associated PSM wall. A gateway is valid only if it lies along the free space of PSM. A gateway must



(a) aligned opportunities (b) unaligned opportunities

Figure 3. Examples of matching opportunities to construct AOS. (Best viewed in color.) Three opportunities with their gateways (pink lines) are extracted from PSM at the robot location (blue dot). If a single, unambiguous match is found between two gateways, their associated opportunities are considered to be on the same path. (a) Gateway 1 and 3 are an unambiguous match so they are on the same path. Gateway 2 does not aligned to any gateways so it is on its own path. (b) In a Y-intersection, Gateway 4 is aligned with Gateway 5 but Gateway 5 is aligned to both Gateway 4 and 6. There are no unambiguous matches among the gateways so the three opportunities are on separate paths.

lie on the free space side of all associated walls of the PSM endpoint that it is anchoring at. Furthermore, only gateways that are within the field of interest are considered.

Given $\mathbf{p}^1$ and the direction $\hat{\mathbf{g}}$ of the gateway, we find the other gateway-end $\mathbf{p}^2$. If the gateway anchors at an *occluding* endpoint that connects a wall segment and a wall opening along the same wall, $\mathbf{p}^2$ is the other PSM endpoint that associates to the opening. Otherwise, $\mathbf{p}^2$ is the closest intersection point of a ray pointing from $\mathbf{p}^1$ in $\hat{\mathbf{g}}$ direction and a wall segment in PSM that is perpendicular to $\hat{\mathbf{g}}$. In the case where no perpendicular wall segment intersects with the ray, $\mathbf{p}^2$ is left undetermined and thus, the gateway width is also undetermined. We exclude a gateway if $\mathbf{p}^2$ is determined but it is too narrow for the robot to pass through. Finally, a gateway is removed, if its direction and gateway-ends are too similar to another gateway.

From each remaining gateway, we form an opportunity and determine its type τ by: 1) the type of the anchoring PSM endpoint; 2) whether $\mathbf{p}^2$ is determined; and 3) the robot's location. A gateway that anchors at a *dihedral* or an *occluding* endpoint forms an *observed* opportunity if $\mathbf{p}^2$ is determined, and forms a *partially observed* opportunity otherwise. A gateway that anchors at an *indefinite* endpoint is a boundary line between an observed and an unobserved region in the PSM, and thus forms a *potential* or a *beginning* opportunity. We arrange the order of $(\mathbf{p}^1, \mathbf{p}^2)$ so that the observed region is on the right side of vector $\overrightarrow{\mathbf{p}_i^1\mathbf{p}_i^2}$ and the unobserved region is on the left. A *potential* opportunity is formed if the robot is located on the observed side of the vector, and a *beginning* opportunity is formed otherwise.

3.2. Extracting AOS from Opportunities

Given a set of opportunities, each provided with only the gateway $\mathbf{g}$ and the opportunity type τ, this section determines the other two elements $\langle \pi, \rho \rangle$ of each opportunity and the ordering among the opportunities to construct the complete AOS. Since $\langle \pi, \rho \rangle$ of the opportunities captures the relationships among them, the complete AOS is extracted by pairing up opportunities if their gateways are well-aligned to form a path. Thus, AOS is extracted by determining the number of paths N_p passing through the field of interest.

First, we define a *bounding box* to represent the smallest bounding box enclosing all gateways. Second, for each pair of opportunities, their gateways $(\mathbf{g}_i, \mathbf{g}_j)$ are compared using the similarity measurement,

$$sim(\mathbf{g}_i, \mathbf{g}_j) = -\cos(\psi_i - \psi_j)\max(0, \frac{l_{\mathbf{g}_i,\mathbf{g}_j}}{l_{\mathbf{g}_i}}) \qquad (6)$$

where ψ_i is the normal direction of gateway $\mathbf{g}_i$ pointing away from the robot. $l_{\mathbf{g}_i}$ is the length of the *bounding box* edge that intersects by a line in the opposite direction of ψ_i, and $l_{\mathbf{g}_i,\mathbf{g}_j}$ is the shortest distance from the gateway line ϕ_i to the center of gateway $\mathbf{g}_j$. Note that this quantity is not symmetric, $sim(\mathbf{g}_i, \mathbf{g}_j) \neq sim(\mathbf{g}_j, \mathbf{g}_i)$. The similarity measurement is designed to account for two factors. The first metric considers how similar the gateway directions are. Orthogonal gateways are not on the same path, while gateways with ψ_g pointing in opposite directions may be on the same path. The second metric considers the amount of overlap between the gateways relative to the size of the *bounding box* enclosing all gateways. Two gateways with more overlap are more likely to be on the same path. If there is no overlap, the gateways are not on the same path.

Starting from an empty set of paths Π that pass through the field of interest, we carry out an exhaustive search among the opportunities to find unambiguous matches using the similarity measurement. If a single, unambiguous match is found between two opportunities, they are considered to be on the same path, and thus the path is added to the set Π. If an opportunity belongs to no paths or to more than one path in the existing path set Π, a separate path is created for the opportunity. Figure 3 shows examples for aligned and unaligned gateways. After the search is done, if a path in Π is associated to only one opportunity A_i, an *unnavigable* opportunity A_j is generated with $\pi_j = \pi_i$ and $\rho_j = -\rho_i$ to describe the opposite side of the path. Finally, the complete AOS is formed by ordering the opportunities so that the normal directions of their gateways are sorted in the counter-clockwise direction.

4. Results

We tested our approach on The Michigan Indoor Corridor 2012 Video Dataset [22]. The dataset has four video sequences with resolution 965×400 in various indoor environments, (i.e. +, T, and L intersections). The field of view of the camera is about $82\,^{\circ}$. For all sequences, the robot pose at each frame is provided.

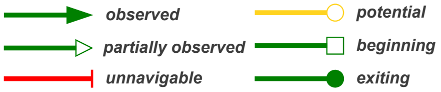


Figure 4. Visualization of different opportunity types. (Best viewed in color.) Filled arrow: full knowledge; Hollow shaped arrow: incomplete knowledge; Green: navigable; Yellow: potentially navigable; Red: unnavigable.

For each frame t, we select the maximum *a posteriori* PSM hypothesis at the current frame and extract the AOS from the PSM at the current robot location. For each example (Figure 5,6,7,8), the first column is the image projection of the PSM ground-wall boundaries. The second column visualizes the PSM in the ground-plane map with the robot pose plotted in blue. In the PSM, a green dot represents a *dihedral* endpoint, a yellow dot represents an *occluding* endpoint, and a red hollow dot represents an *indefinite* endpoint. Each wall has an index automatically assigned by the implemented system, and all the wall segments contained in that wall are marked by the same index. The third column is the AOS at the current robot location. Each opportunity A_i is shown directed along its associated path with an arrow reflecting its type (Figure 4), and a label for its path index and its direction along the path $\langle \pi_i, \rho_i \rangle$.

Figure 5 demonstrates our incremental method of building the PSM of the observed environment from a temporally continuous stream of images, and the AOSs extracted in various locations within the PSM. Figure 6 demonstrates that PSM is a step forward from a pure planar model because it represents a richer set of relationships among planar segments. Figure 7 compares results from two sequences acquired around the same intersection with different trajectories. In all of these examples, due to the limited field of view of the monocular camera, it is impossible for the robot to realize that it is at an intersection solely from the current image. Thus, a temporally continuous stream of images is essential for coherent visual scene understanding.

The maximum *a posteriori* PSM hypothesis is correct [2], 92.18% of the time. The main reason that our method fails to select the correct hypothesis is lack of feature. One can overcome this problem by applying methods that maintains a set of informative features to discriminate the hypotheses [24]. Nevertheless, among the frames with an incorrect PSM, our method is still able to extract the correct AOS, 73.72% of the time. This happens because the incorrect PSM hypothesis has the same structure layout of the correct one, except the actual locations of the walls are off. Figure 8 is an example of this situation.

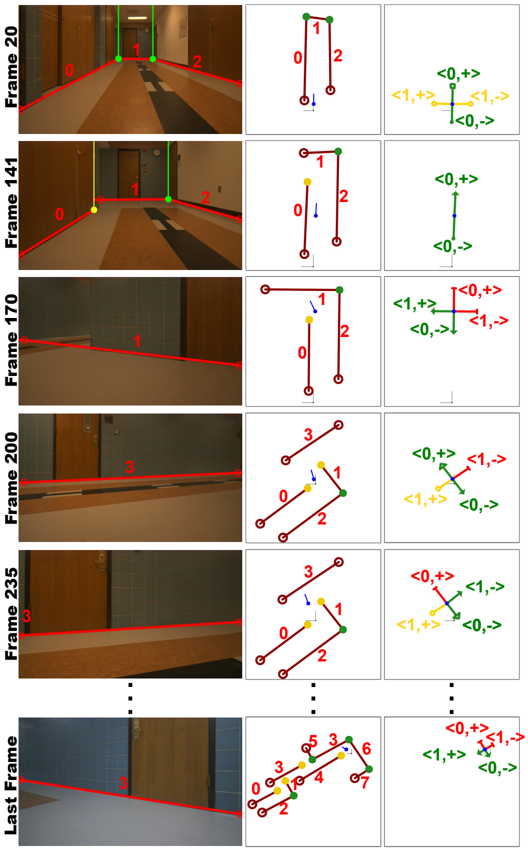


Figure 5. Examples of the two-layer representation on Dataset L that demonstrates the on-line incremental process of building the PSM [23] and the AOS extracted at various locations. (Best viewed in color.) This sequence contains three L-intersections. The robot traveled through two long corridors connected by two adjacent L-intersections and finally made a U-turn at the last L-intersection. **Frame 20:** Due to the field of view of the monocular camera, no information of the PSM around the robot's immediate surrounding is available when the process begins. Only the environment in front of the robot is observed. Thus, the *beginning* opportunity leads the robot into the region that has been modeled. **Frame 141:** As more observations become available, PSM with more detail (the first L-intersection) of the environment are incrementally built. Although there is still incomplete knowledge in the PSM in the distance, the robot is now in a long corridor with full knowledge of its current surrounding. Thus, in the AOS, the *observed* opportunity leads the robot towards the L-intersection, while the *exiting* opportunity leads the robot out of its field of interest. **Frame 170:** The robot is at the first L-intersection and has full knowledge of the opportunities available at the intersection. **Frame 200:** More details of the environment is captured with the PSM. The robot has incomplete knowledge of its surrounding. The wall on the left side of the robot is unobserved, so the *potential* opportunity leads the robot towards the unobserved region. **Frame 235:** The robot is in the second L-intersection but unlike the first one, the robot has incomplete knowledge of the intersection. **Last Frame:** The final PSM is constructed from the sequence. At this point, the system cannot yet conclude that the endpoint of Wall 4 is an *dihedral* endpoint that connects to Wall 1.

[2] We consider a PSM hypothesis correct if the geometric structure within 4 meters of the vision cone is correctly modeled. Thus, for a given frame, it is possible to have more than one correct PSM hypothesis, if the differences are further than 4 meters away.

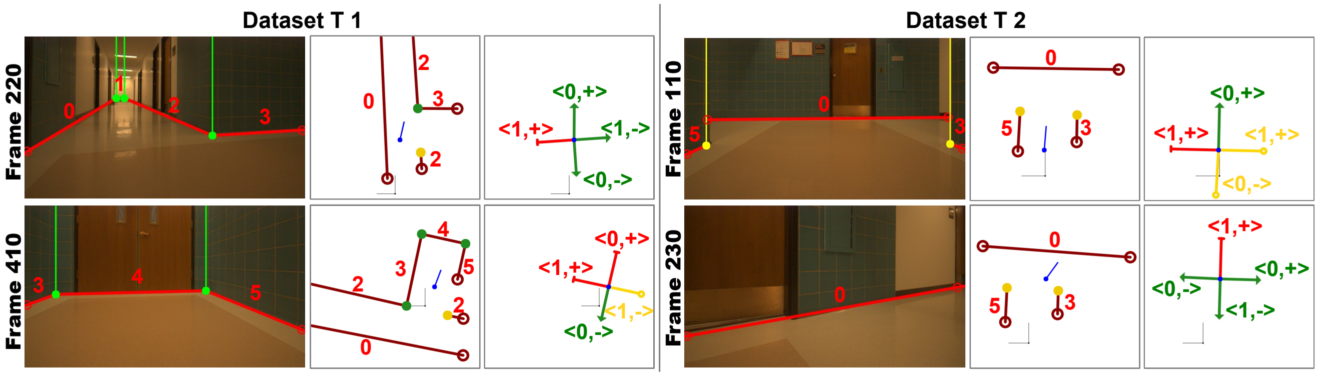


Figure 7. Examples of the two-layer representation on Dataset T 1 and Dataset T 2. (Best viewed in color.) In the two video sequences, the robot traveled around the same T-intersection in the physical world. In Dataset T 1, the robot traveled from the major corridor and made a right turn onto the minor corridor at the intersection, whereas, in Dataset T 2, the robot traveled from the minor corridor to the major corridor. We process each sequence independently and compare the results from the two. To clarify the comparison, we aligned the wall indices so that the same wall in the physical world has the same index. **(T1)Frame 200:** PSM models the T-intersection by three walls (Wall 0, 2 and 3). Wall 2 contains two disjoint wall segments, and the gap between the wall segments is the opening of the T-intersection. The AOS captures the opportunities for actions at the T-intersection with full knowledge. **(T1)Frame 410:** PSM continues to model the dead-end at the minor corridor. Due to lack of observations, the video sequence contains no clue for Wall 5 to intersect with Wall 2. Thus, in AOS, the *potential* opportunity captures the incomplete knowledge of the missing information between Wall 5 and Wall 2. **(T2)Frame 110:** The robot is at approximately the same location as the robot in (T1)Frame 410 but in the opposite direction. Since the observations of the two sequences of the same environment are different, PSMs from the two sequences captures different forms of partial knowledges of the environment. Consequently, the AOSs extracted from the two sequences captures partial knowledge of different part of the robot's surrounding. The two AOSs contains no conflicting opportunities. Note that since the robot was facing in opposite direction in the two sequences, one of the AOSs needs to be rotated at about 180 ° in order to match the other one. Thus, by acquiring more observations around a *potential* opportunity, it can become an *observed* or an *unnavigable* opportunity. **(T2)Frame 230:** The robot is at the T intersection and has full knowledge of intersection. Since the robot is at the same T-intersection as the (T1)Frame 200, the AOSs in both situations are the same. In fact, any location within the T-intersection will have the same AOS. Moreover, if the structure and the knowledge of the robot's surrounding of two locations are similar, AOSs extracted in both locations will be the same.

5. Conclusion

We presented a two-layer representation of the locally sensed 3D indoor environment. Each layer represents a different level of understanding of the environment. The first layer, Planar Semantic Model (PSM), is a coarse-grained geometric representation of the indoor environment in terms of ground plane and walls. A wall consists of a set of disjoint wall segments that are co-planar in 3D. Thus, PSM is a step forward of a pure planar model because it represents a richer set of relationships among planar segments. The second layer, Action Opportunity Star (AOS), describes a structured set of qualitatively distinct opportunities for robot action at a given location. An opportunity is an abstraction of a group of trajectories that have the same semantic meaning in terms of robot action. We demonstrated an algorithm to extract AOS from PSM.

Our representation is a concise and semantically meaningful representation of an indoor environment to both human and indoor navigating robots. It is able to represent incomplete knowledge of the local environment so that missing information can be incrementally built as observations become available. Unlike existing scene understanding works, in addition to modeling the geometric structure of the environment, our representation takes a step forward to reason about the robot's action opportunities. Furthermore, our representation supports topological mapping [1] because the robot can detect whether it is at a topological place (e.g. hallway intersection) or not by checking the number of paths in the AOS. Experimental results on a variety of indoor environments demonstrated the expressive power of our representation.

Our future work includes two directions. One direction is to model cluttered environment with the proposed two-layer representation. The main challenge is to model the regions that are explained by the PSM and identify regions that are not explained by the PSM. Once the PSM is extracted, the AOS can be extracted by our proposed method. The second direction is to apply our representation for a robot to explore the unknown parts of the environment. Our representation allows the robot to make plans in different levels. The robot will select a navigable opportunity that has incomplete knowledge from the AOS to choose a target pose, and find a trajectory to get from its current pose to that target within the free-space of PSM.

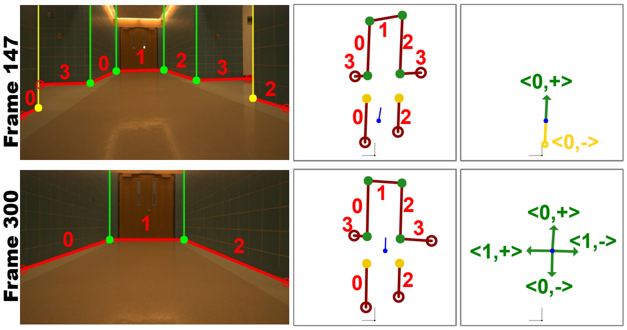
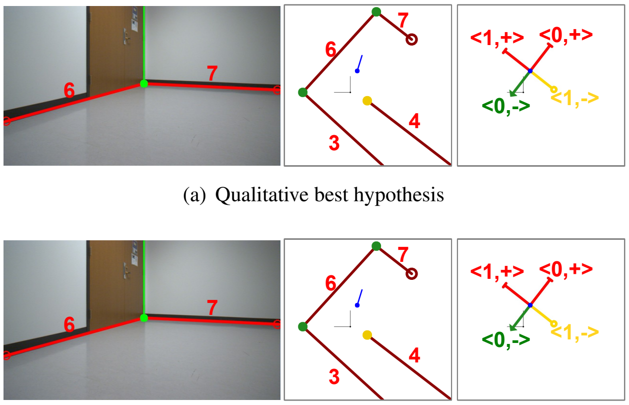


Figure 6. Examples of the two-layer representation on Dataset + that demonstrates the expressive power of the PSM. (Best viewed in color.) Dataset + has one +-intersection along a long corridor, and the robot traveled from one end of the corridor to the intersection without making any turn. PSM models the +-intersection by three walls (Wall 0, 2, and 3), each with two disjoint wall segments. Thus, PSM is a step forward from a pure planar model because it represents a richer set of relationships among planar segments. The robot is at the +-intersection, and has full knowledge of the intersection. Notice that it is impossible for the robot to realize that it is at a +-intersection solely from the current image. Thus, a temporally continuous stream of images is essential for coherent visual scene understanding.



(a) Qualitative best hypothesis

(b) Maximum posterior hypothesis

Figure 8. An example when the maximum *a posteriori* hypothesis is not a correct PSM hypothesis but the AOS is correct. (Best viewed in color.) Due to lack of feature points, our method may fail to identify the correct hypothesis. In this case, the actual location of Wall 7 is not correctly identified. However, if the incorrect PSM has the correct structure layout, the extracted AOS will still be the same as the correct PSM hypothesis.

Acknowledgment

This work has taken place in the Intelligent Robotics Lab in the Computer Science and Engineering Division of the University of Michigan. Research in the lab is supported in part by grants from the National Science Foundation (CPS-0931474, IIS-1111494, and IIS-1252987).

References

[1] P. Beeson, J. Modayil, and B. Kuipers. Factoring the mapping problem: Mobile robot map-building in the Hybrid Spatial Semantic Hierarchy. *IJRR*, 29(4):428–459, 2010. 3, 7

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

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

[4] 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

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

[6] A. Flint, D. Murray, and I. Reid. Manhattan scene understanding using monocular, stereo, and 3d features. *ICCV*, 2011. 1

[7] A. Furlan, S. Miller, D. G. Sorrenti, L. Fei-Fei, and S. Savarese. Free your camera: 3d indoor scene understanding from arbitrary camera motion. *BMVC*, 2013. 1

[8] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Reconstructing building interiors from images. *ICCV*, 2009. 1

[9] J. J. Gibson. *The Senses Considered as Perceptual Systems*. Houghton Mifflin, Boston, 1966. 3

[10] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. *CVPR*, 2011. 1

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

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

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

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

[15] Y. Jiang and A. Saxena. Infinite latent conditional random fields for modeling environments through humans. *RSS*, 2013. 1

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

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

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

[19] 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

[20] 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

[21] C. J. Taylor and A. Cowley. Parsing indoor scenes using rgb-d imagery. *RSS*, 2012. 1

[22] G. Tsai and B. Kuipers. Dynamic visual understanding of the local environment for an indoor navigating robot. *IROS*, 2012. Dataset: www.eecs.umich.edu/~gstsai/release/Umich_indoor_corridor_2012_dataset.html. 2, 3, 5

[23] G. Tsai and B. Kuipers. Dynamic visual understanding of the local environment for an indoor navigating robot. *IROS*, 2012. 6

[24] G. Tsai and B. Kuipers. Focusing attention on visual features that matter. *BMVC*, 2013. 6

[25] G. Tsai, C. Xu, J. Liu, and B. Kuipers. Real-time indoor scene understanding using Bayesian filtering with motion cues. *ICCV*, 2011. 2

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

[27] Y. Zhao and S.-C. Zhu. Scene parsing by integrating function, geometry and appearance models. *CVPR*, 2013. 1