Automatic Room Segmentation from Unstructured 3D Data of Indoor Environments

Rareş Ambruş*[1], Sebastian Claici*[2], Axel Wendt[3]

Abstract— **We present an automatic approach for the task of reconstructing a 2D floor plan from unstructured point clouds of building interiors. Our approach emphasizes accurate and robust detection of building structural elements, and unlike previous approaches does not require prior knowledge of scanning device poses. The reconstruction task is formulated as a multiclass labeling problem that we approach using energy minimization. We use intuitive priors to define the costs for the energy minimization problem, and rely on accurate wall and opening detection algorithms to ensure robustness. We provide detailed experimental evaluation results, both qualitative and quantitative, against state of the art methods and labeled ground truth data.**

I. INTRODUCTION

Robotic applications of accurate room level segmentations of indoor raw data are endless. The ability to identify walls, doors, and clutter already provides a wealth of information for various robotic systems. In addition, identifying room boundaries helps in tasks such as topological mapping, semantic mapping, automatized professional cleaning, and human-robot interaction. Within the robotics community, the development of tools such as GMapping [1] have made creating 2D grid-maps standard practice when dealing with mobile robots. Similarly, creating large scale 3D maps of various indoor environments has become increasingly easy, especially with the advent of low-cost RGB-D cameras.

However, there are several difficulties in obtaining accurate room level segmentations from point cloud data. Current scanner technology is imperfect, and noise is frequent either as registration errors, or missing data. Furthermore, interiors of residential buildings are highly cluttered with furniture and other objects. Separating clutter from permanent structures such as walls and doors is difficult as clutter can occlude permanent structures. For example, bookshelves often span the entire height of a wall, and recognizing such scenarios remains difficult.

Current approaches make a number of assumptions to make the problem tractable: knowledge of the direction of the up-vector, straight walls, alignment to a pre-defined Manhattan world frame or knowledge of the original scanning device poses from where the data was acquired. In our work we lift as many of these assumptions as possible, and present a novel method which relies only on the knowledge of the direction of the up-vector with respect to the ground. Our contributions are:

- A robust method of detecting openings in wall segments.
- An efficient method for computing a set of synthetic sensor poses / viewpoints which yield an initial labeling of the input point cloud.
- A two-step energy minimization problem: first a binary, inside-outside labeling step for noise reduction, and second a multi-label minization step which partitions the data into separate semantic entities.

We provide detailed experimental evaluation results against state of the art methods and labeled ground truth data and we show that our method outperforms related approaches.

II. RELATED WORK

Bormann et al. [2] provide a comprehensive list of methods suited for segmenting 2D maps into separate semantic entities, as well as a more detailed analysis of four selected methods. In our results section we include a comparison to the best performing method presented by Bormann et al.

Pronobis et al. [3] describe a system which learns how to combine laser and visual data to categorize semantic entities into pre-trained categories. The sensor data is processed incrementally to build a semantic map as the robot traverses the environment. Xiong et al. [4] use a region growing approach to segment out structural elements and classify them using a logistic regression model. Using a Constructive Solid Geometry model, Xiao and Furukawa [5] are able to reconstruct semantically meaningful room regions in museums, but their use of the Hough transform to detect wall segments is restricted to uncluttered scenes. Turner and Zakhor [6], [7] construct a floor plan by triangulating the 2D regions and merging them based on a graph cut approach to separate rooms.

Armeni et al. [8] describe a system for parsing and classifying 3D point clouds of entire buildings into rooms, and further into building elements (walls, doors and objects). One of the novel elements of their method is the usage of convolution operators on different axes to detect walls based on the empty spaces between them, using the assumption that the rooms are all aligned to a predefined Manhattan world

Manuscript received: September 10, 2016; Revised November 29, 2016; Accepted December 22 2016

This paper was recommended for publication by Editor Cyrill Stachniss upon evaluation of the Associate Editor and Reviewer's comments.

*Both authors contributed equally. This work was done during an internship of the first two authors at the Bosch Robotics Research group.

[1] Rareş Ambruş is with the Centre for Autonomous Systems, KTH Royal Institute of Technology, Stockholm, SE-100 44, Sweden, raambrus@kth.se

[2] Sebastian Claici is with the Computer Science and Articial Intelligence Lab, MIT, Cambridge, MA 02139, USA, sclaici@csail.mit.edu

[3] Axel Wendt is with the Bosch Research and Technology Center, Robert Bosch LLC, Palo Alto, CA 94304, USA, axel.wendt@bosch.com

Digital Object Identifier (DOI): see top of this page.

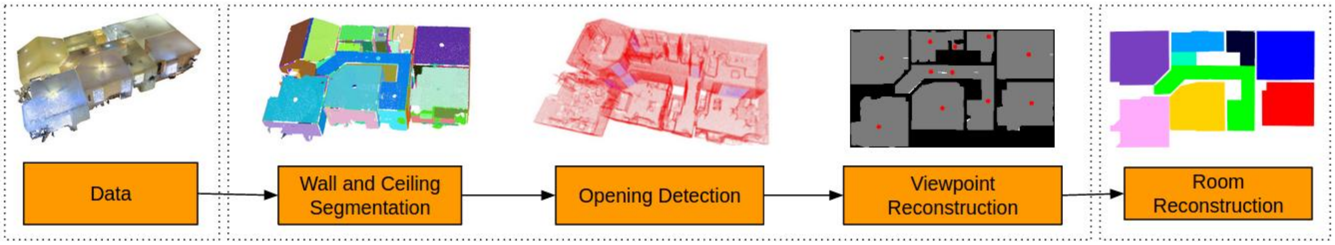

Fig. 1: System overview.

frame. Assuming a single Manhattan frame is limiting, and even simple scenarios exhibit multiple Manhattan frames [9].

Oesau et al. [10] perform an analysis along the z axis to decompose the input point cloud into horizontal and vertical structures, after which a volumetric model is extracted through a binary energy minimization process solved using Graph-cut. Closer to our work, Ochmann et al. [11] use Graph-cuts iteratively in a multi-label minimization framework. The initial set of labels is obtained from the assumption that the viewpoints from which the environment was scanned are known. In contrast, we don't use prior viewpoint information, and instead generate the initial labels automatically. A further distinction arises from detecting openings in the point cloud: while [11] employs a pre-trained SVM classifier, we perform an analysis on the projection of the wall plane on a 2D image, an operation which requires no training data. In [12] Mura et al. segment out planar patches in the 3D point cloud taking occlusions due to clutter into account. Unlike us, their method of detecting planar patches uses a smoothness-based region growing approach which is less robust in the case of noisy data. Further, the authors also use a 2D cell complex data structure to represent the patches in 2D, and define a set of weights between the faces of the complex which are propagated using diffusion maps.

In [13] Mura et al. further extend this method and explicitly encode adjacency relations between structures in a graph-based scene representation which is used to separate the walls, floor and ceiling from clutter. The data is assembled into a 3D cell complex and the optimal number of rooms is computed through a Markov Clustering step performed using visibility weights defined in the cell complex. Finally, partitioning of the cell complex space into rooms is done through a multi-label Markov Random Field label assignment. Our method is more generic as we do not rely on a-priori viewpoint information. Furthermore, we do not explicitly encode adjacencies between floor-walls-ceiling, which can be error prone especially in the case of noisy or missing data. In addition, instead of a first clustering step to obtain the number of rooms, our method first computes the energy minimization partition and subsequently merges regions by cross-checking the inferred walls with the primitives detected in an earlier step.

There is significant prior work [14], [15], [16], [17] in semantic segmentation of indoor RGB-D images using machine learning techniques. While such approaches are attractive for gaining a better semantic understanding of the rooms extracted by our algorithm, they cannot be used for the task of high level room reconstruction as they lack contextual geometric information and require RGB-D images.

III. SYSTEM OVERVIEW

Similar to prior work, we use a point cloud representation of the environment. The capturing and processing of the input point cloud is outside the scope of this work, and we assume that the data is already registered. Our method is a hierarchical parser into disjoint, semantically meaningful entities that we call rooms. We define a room as the interior of a simple closed curve in the plane that can be separated from other rooms through meaningful separators. We interpret a meaningful separator as an opening that semantically disconnects two connected regions. These can be either doors, or openings in wall segments.

We trade off between two metrics using an energy minimization formulation to obtain a maximum a posteriori labeling for each room. The energy minimization problem requires an initial labeling of the data, commonly obtained in the literature [11], [13] through the assumption that a set of viewpoints from which the point cloud was acquired is known. In our approach we lift this assumption, and instead compute a set of viewpoints from the Voronoi partition of the 2D projection of the data.

Computing the 2D projection of the data involves the detection of the ceiling, walls and openings. We define a wall as a planar patch orthogonal to the ground plane. While this does not account for all wall segments (slanted or curved walls cannot be detected), it is a common assumption in automatic reconstruction methods, and such walls account for the vast majority of structural elements. Ceiling segments are detected using a similar method, but we allow variation up to 45^o off parallel.

We detect openings using the information gained from empty spaces in wall segments. Intuitively, we adopt a heuristic approach where empty spaces are compared with a parametric model of a door for height, width, and shape.

A high level overview of our approach is shown in Fig. 1.

IV. SEGMENTATION OF STRUCTURAL ELEMENTS

We now proceed to describe in detail each section of the pipeline, starting from the coarser elements (wall and ceiling planes), down to the fine detail elements (doors and viewpoints).

A. Primitive Detection

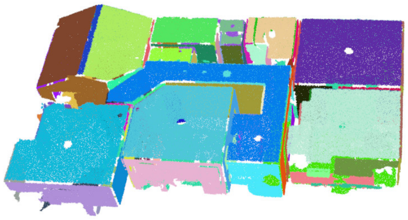

Fig. 2: Plane primitives extracted from a point cloud, arbitrarily colored.

We start from the method of Schnabel et al. [18], [19], which offers an efficient way of detecting primitive shapes of arbitrary orientations and sizes in an unordered 3D point cloud. As described in [18], each shape is described by a parametric model along with a set of supporting indices representing points in the original point cloud which are within a certain distance threshold of each other (i.e. a connected component). The primitive shapes supported are planes, cylinders and spheres. At this point we limit ourselves to plane primitives, however, we can easily extend our method to also support curved walls, as they can be parametrized by a cylindrical primitive. Fig. 2 shows the typical output of this step.

B. Wall and Ceiling Detection

To identify ceiling plane primitives we project all the points in the point cloud onto a 2D grid aligned with the XY-plane. For each occupied cell in the grid, we find the 3D point in the point cloud with the highest Z coordinate. We use the 3D points obtained to identify relevant ceiling plane primitives, and proceed in a similar way to identify floor plane primitives.

We define walls as planar patches of points which stretch between the floor and the ceiling, with the restriction that the normal to the wall plane be perpendicular to the floor up vector. The wall candidates still contain a number of false positives corresponding to cupboards, shelves, screens, etc. which we would like to filter out. However, actual walls are most often occluded by the presence of clutter in front of the sensor, making it difficult to reason about their true height. Mura et al. [12] uses viewpoint information to reason about occlusions in the point cloud, and discard planes corresponding to false positives. However, in our setting we do not have access to the viewpoint information. Instead, we perform a per-candidate analysis and identify gaps corresponding to openings (such as doors) or due to occlusions created by clutter.

We represent each wall candidate P in Hessian normal form, $P = (\vec{n}, p)$, where $\vec{n}$ is the normal to the plane, and p is the distance to the origin. We define a 2D reference frame B in the plane, with origin aligned to the plane lower left corner. One column of B consists of the floor up vector $\vec{z}$ and the second lies in the plane and is obtained by: $\vec{z} \times \vec{n}$. Using B we project all the points of the wall candidate P to an image which we analyse for gaps and openings (see Fig. 3).

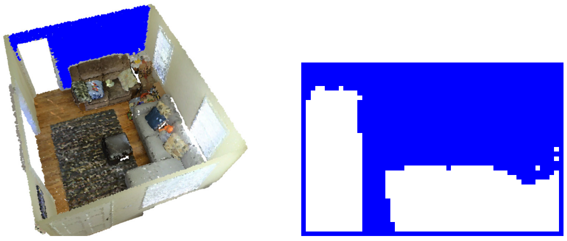

Fig. 3: Plane with projection. (a) Point cloud of a room with selected plane points colored in blue. (b) 2D projection of the wall points on a 2D plane with white pixels denoting free cells and blue pixels occupied cells.

C. Opening Detection

We define an opening as any patch of empty space contained within planar wall segments that is the required size and shape. We start with the connected components defined by each detected plane (see Fig. 3b). By analysing the 2D projection defined by a wall segment, we can detect opening candidates by looking for rectangle patches satisfying certain width and height requirements. An efficient implementation yields a solution in $O(n)$ time, where n is the number of pixels in the image. As an example, in Fig. 3b the cluster of white pixels on the left of the image corresponds to an opening, while the cluster on the right represents unobserved space due to occlusion by the couch.

D. Viewpoint Generation

We project the ceiling segmentation on the ground plane and mark all points in the bounding box of the segmentation as free space. The projections of the walls and detected openings are marked as obstacle space.

The resulting projection is noisy in areas with low ceiling point density and at the intersection between ceiling planes and wall planes. We run an energy minimization step to obtain a refined foreground/background segmentation.

This has the advantage of providing a simple and very efficient room reconstruction method. The rooms frequently cluster in connected components and a flood fill algorithm is capable of obtaining semantic labels (Fig. 4). In the results section we provide qualitative and quantitative evaluations of the semantic segmentation obtained at this point - see Table Id and Fig. 8f.

This approach however relies on almost perfect wall and opening detection, and produces jagged walls. We can leverage the simplified representation to obtain simulated viewpoints. By sampling viewpoints automatically, we do not restrict ourselves to data that encodes the original viewpoints.

The 2D projection can be described as free space (white pixels in Fig. 4b) and obstacle space (black pixels). We can compute the medial axis of each free space component following the work of [21] - Fig. 5a shows the resulting Voronoi graph. Points that will have most visibility lie along the medial axes of free space. To sample viewpoints we

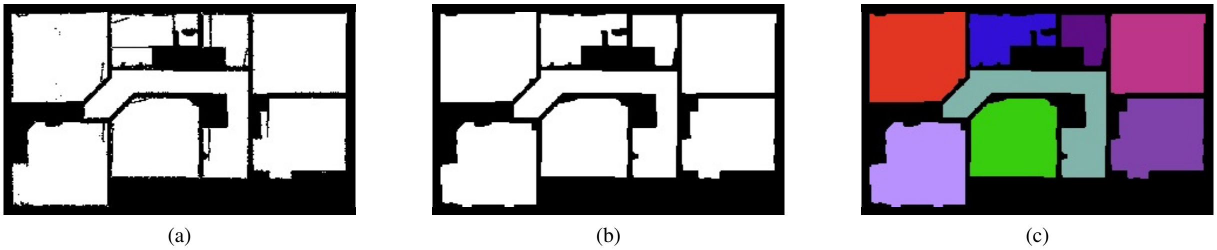
(a)
(b)
(c)

Fig. 4: Simple room reconstruction. (a) 2D projection of ceiling points and wall candidates. Notice a clear separation between rooms, but noisy interior. (b) Best foreground/background segmentation after energy minimization. (c) Rooms after flood filling.

proceed in a greedy fashion by selecting the pixel which observes the most pixels in a radius around itself. We repeat this process until all but a small fraction of the free space pixels are visible from at least one viewpoint. The radius chosen reflects the operating range of the sensor used to capture the data - this ensures that even in the case of large rooms we will need multiple viewpoints, thus ensuring we obtain an oversegmentation of the scene. In all experiments we use a radius of 3m (the typical operating range of several commercially available 3D scanners).

The resulting set of viewpoints is displayed in Fig. 5b.

To recover viewpoints in the original pointcloud, we project the 2D viewpoints back into the original 3D space and place them at mean height.

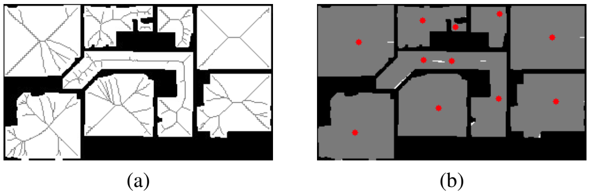
(a)
(b)

Fig. 5: Initial viewpoint labeling. (a) Voronoi graph. (b) Viewpoints in 2D shown in red. Points visible from at least one viewpoint shown in gray, while unobserved points are colored white.

V. ROOM RECONSTRUCTION

To obtain semantically meaningful rooms from the extracted primitives, we make use of a cell complex data structure. This has the advantage of producing a planar graph that encodes spatial relationships between regions defined by wall segments. The graph structure allows us to define unary and binary potentials for use with an energy minimization algorithm for semantic labeling.

A. Cell Complex

A cell complex or arrangement of lines [22], [23] is a geometric data structure describing how the intersection of a set of lines partitions space. To construct a cell complex, we project the points associated with each wall segment to the ground plane and find a line of best fit by solving the associated least-squares problem. To deal with curved wall surfaces, we can find piecewise linear approximations of the projected curve and insert each individual line segment into the cell complex.

The cell complex induces a planar graph in Euclidean space with vertices representing intersections between line segments, and edges naturally induced by the segments. Every planar graph induces a dual graph with vertices representing faces in the primal and edges between adjacent faces (see Fig. 6). We work with the dual to obtain a per face labeling. We define an energy minimization problem on the dual graph by associating unary potentials with the vertices (representing the faces in the cell complex) and binary potentials with the edges.

Formally, we propose to solve the following

$$\min_{\mathbf{l}} \sum_{v \in V} U_v(l_v) + \sum_{v,w \in E} B_{v,w}(l_v, l_w) \tag{1}$$

where $\mathbf{l} \in \mathcal{L}^{|V|}$ is a per vertex label vector drawn from a finite labeling set, $U_v : \mathcal{L} \to [0,1]$ is the unary potential function associated with vertex v, and $B_{v,w} : \mathcal{L} \times \mathcal{L} \to [0,1]$ is the binary potential function associated with the edge (v,w). It is important for our purposes that the true number of rooms be at most equal to $|\mathcal{L}|$ (i.e. the initial segmentation has to be an oversegmentation).

The unary potentials describe the likelihood that regions have an associated (coarse) labeling obtained from an oversegmentation. The unary potentials are a guess on the true partition of the rooms. To define easy to compute unary potentials, we make a simplifying assumption that the rooms are close to convex. Using the synthetic viewpoints defined

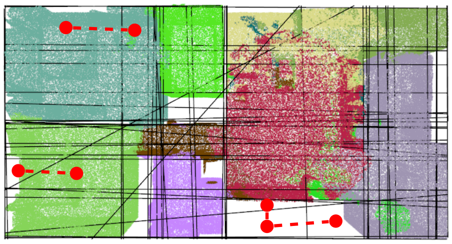

Fig. 6: Cell complex example. This figure shows the main components of a cell complex data structure. We show the full primal graph as black edges. A few examples of dual vertices and dual edges are shown. Dual vertices (corresponding to primal faces) are shown as red circles, and dual edges are shown as dashed red lines.

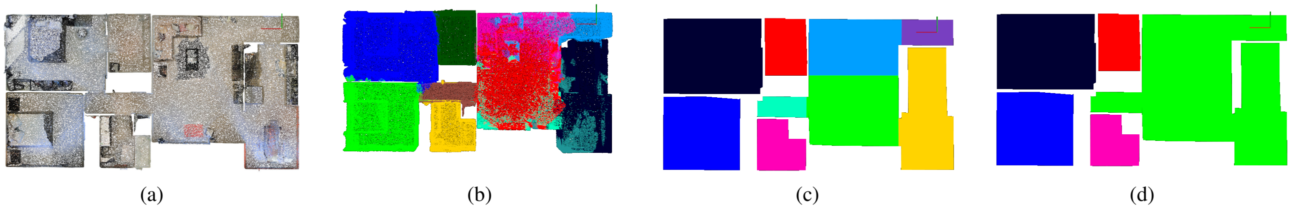

Fig. 7: Room reconstruction end-to-end: (a) Initial point cloud. (b) Initial viewpoint labeling of point cloud - each color represents points associated with a different viewpoint. (c) Initial room segmentation. (d) Final room segmentation, after merging.

previously, we can label points within a fixed radius from each viewpoint using a viewpoint specific color. This captures the intuition that points that can be seen from the same viewpoint are more likely to be part of the same room. An example coloring of cloud points based on viewpoint visibility is shown in Fig. 7b.

We use the colored points to obtain a potential associated to each face (recall that faces correspond to vertices in the dual graph). For each face in the cell complex, we calculate the density of points of each color whose projection to the plane falls within that face.

We define a unary potential associated with each face and color as the ratio of points colored with that specific color over all points that fall within a face. Formally, let c_{ij} be the number of points associated with viewpoint j that fall in face i. For each face i we define a potential $\theta^i_j = \frac{c_{ij}}{\sum_j c_{ij}}$.

To detect faces in the cell complex that correspond to empty space, we perform a density threshold check. As the subsampling grid size is given as 0.05m, we can compute an average density for each meter squared. If a cell complex face does not contain a density of points that is within a given factor of the average density, we mark it as empty. We create a fictitious viewpoint labeled 0 for the empty spaces, and to each empty region we associate a potential $\theta^i_j = 1$ if $j \neq 0$ (i.e. j is any other label), and $\theta^i_j = 0$ if $j = 0$ (i.e. j is the empty label).

Binary potentials are obtained from the wall information. Each edge in the cell complex is within the linear span of a wall segment. If we write $\mathbf{e}$ for the cell complex edge, and $\mathbf{w}$ for the original wall segment, we obtain a binary potential between the two faces separated by edge $\mathbf{e}$ as

$$B_{u,v}(l_u, l_v) = \begin{cases} 0 & \text{if} \quad l_u = l_v \\ 1 - \frac{|\mathbf{e} \cap \mathbf{w}|}{|\mathbf{e}|} & \text{otherwise.} \end{cases}$$

Here $|\mathbf{e}|$ denotes the length of the segment $\mathbf{e}$, and $\mathbf{e} \cap \mathbf{w}$ is the segment intersection of $\mathbf{e}$ and $\mathbf{w}$. This can be either a segment, a point, or empty.

This potential intuitively describes the likelihood that two regions are contained within the same room. If there is a large wall segment that separates the two regions, then the regions are unlikely to cluster together. Note that we must enforce $B_{u,v}(l, l) = 0$ to maintain semi-metricness of the B function. This is required to apply the graph-cut algorithm of Boykov and Kolmogorov [24].

B. Room Segmentation

We solve the minimization problem (1) using the alpha expansion algorithm of Boykov and Kolmogorov [24].

The output of the energy minimization can be seen in 7c. The room segmentation algorithm can lead to an oversegmentation in cases where imaginary walls are inferred. This is the case for long corridors, and such regions can be merged in a post-processing step.

To merge two regions, we detect the overlap between the edge separating two regions and the wall segments that were detected by the RANSAC algorithm. If the overlap is small (less than 20% of the wall overlaps the inferred separator), we choose to merge the two regions into one. We show the output of this step in Fig. 7d.

VI. EVALUATION AND RESULTS

To test our implementation as well as compare with related work we have created a dataset consisting of 10 large scale point clouds, with ground truth segmentation labels[1] (see Fig. 9a and b). All point clouds were subsampled in a voxel grid at a resolution of 0.05m. We mention that our data was collected using cheap RGBD sensors and presents certain challenges which invalidate some of the assumptions made by the related work such as increased noise, occasional registration errors, slanted ceilings, up-vector not perfectly aligned with (0,0,1).

For evaluation we perform standard intersection over union (IoU) and precision/recall tests and compare with other state of the art methods. The intersection over union metric is defined as the ratio between the area of the intersection between the results and the ground truth and the union of the same. We compare the 2D segmentation results of the previous sections with hand labeled data. The IoU is taken from a best match with the labeled data (see Fig. 8). Additionally, we also perform a coarse evaluation by measuring precision and recall of detected rooms. We compute true positives as regions whose intersection with the ground truth covers at least half of the room surface area. If a room covers multiple ground truth rooms, we only account for one (see e.g. room 10 in Fig. 9). False positives are rooms not present in the ground truth labeling, while false negatives are given either by a room that is not detected by the method, or by a room that is incorrectly connected to a different room. The results

[1] The data is available upon request.

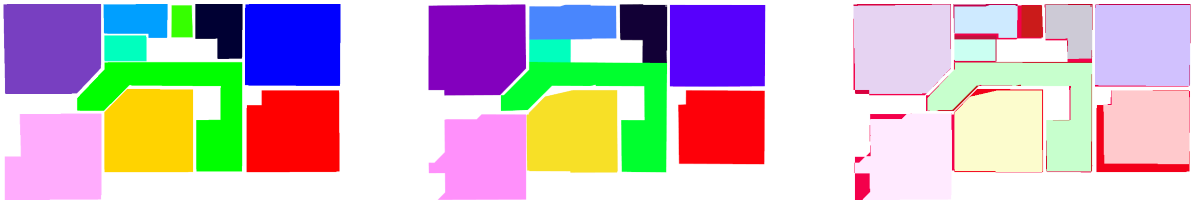
(a)
(b)
(c)

Fig. 8: Intersection over union operation. (a) Ground truth labeled data set. (b) Our results on the same data set. (c) Intersection over union visualization; regions colored red are incorrectly labeled.

	(a) **Bormann et al. [2]**			(b) **Ochmann et al. [11]**			(c) **Mura et al. [13]**			(d) **Sec. IV**			(e) **Sec. V**		
	Prec.	Rec.	IoU	Prec.	Rec.	IoU	Prec.	Rec.	IoU	Prec.	Rec.	IoU	Prec.	Rec.	IoU
1	1	0.9	0.71	0.8	0.8	0.74	0.9	1	0.75	1	0.8	0.92	1	0.9	**0.93**
2	0.75	1	0.77	0.6	1	0.7	1	1	0.9	0.83	0.83	0.81	1	1	**0.95**
3	1	0.82	0.77	0.8	0.73	0.83	1	1	0.81	1	1	0.87	1	1	**0.95**
4	0.8	1	0.83	0.57	1	0.62	1	0.75	0.88	1	1	**0.91**	1	0.75	0.89
5	0.72	1	0.88	0.63	1	0.73	1	1	**0.95**	1	1	0.92	1	1	0.94
6	1	0.75	0.65	1	1	0.88	1	0.92	0.9	1	1	0.91	1	1	**0.93**
7	1	0.5	0.67	1	1	0.86	1	1	0.86	0.91	0.91	**0.90**	1	0.8	0.82
8	1	1	0.8	0.67	0.8	0.78	1	1	0.9	1	1	**0.94**	1	1	0.91
9	0.75	0.75	0.85	0.89	1	0.79	1	1	**0.95**	1	0.75	0.87	0.89	1	0.92
10	0.47	0.9	0.61	0.69	0.9	0.68	1	0.1	0.44	0.7	0.7	**0.7**	0.64	0.9	**0.7**
Mean	0.85	0.86	0.75	0.77	0.92	0.76	**0.99**	0.88	0.83	0.94	0.90	0.87	0.95	**0.94**	**0.89**

TABLE I: Results and comparison with state of the art. We measure precision and recall based on the number of detected rooms against the ground truth labeling. To obtain a measure for the error in area, we compute the intersection over union score of the best labeling pairing between each method and the ground truth. (The results shown for Bormann et al. [2] correspond to the Voronoi segmentation method.)

are summarized in Table I. Qualitative results are displayed in Fig. 9.

We perform the same evaluation on three of the methods described in [2]: morphological, distance and Voronoi methods. In Table Ia we show the results of the Voronoi segmentation method, which performed the best. We note that the methods described in [2] were applied after the initial 2D projection was computed, and thus after walls and doors were segmented out and all other clutter removed. In keeping with the original intent of [2] we did not mark the doors as impassable areas. In some cases [2] yields good results (see Fig. 9d, rows 6 and 9), however, overall these methods don't generalize well and perform poorly when applied to a more varied dataset. We compare here with the segmentation resulting after the steps described in Section IV of our method - see Table Id, and in Fig. 9f.

We notice a much better segmentation as compared to [2] which we attribute to our energy minimization step which smooths out some of the 2D projection noise, as well as the marking of doors as obstacles. Notice however that when the door detection fails the flood-fill operation joins rooms together, as is the case in e.g. Fig. 9f rows 2 and 9.

Even though precision and recall for room detection are high in these methods, the area of the rooms is more prone to errors, thus making these approaches less attractive for analysis of indoor reconstructions. A second reason to prefer cell complex approaches is that they yield much more intuitive room boundaries composed of few straight line segments. Notably, we compare with the work of Mura et al. [13] and Ochmann et al. [11] whose methods share many similarities to our own. Since [11], [13] require viewpoint information, we supply the viewpoint positions computed by our method for the purpose of the comparison.

The method of Ochmann et al. [11] - see Table Ib and Fig. 9c, yields good results in terms of detecting the true walls of the environment, and the resulting segments follow closely the outline of the ground truth segmentation. However, we notice that the method is prone to over-segmenting the environment. [11] includes a method for identifying ”false” walls induced by the energy minimization, based on a supervised machine learning approach, which appears to yield false results on our data thus leading to the over-segmentation.

The method of Mura et al. [13] also uses a cell complex to define an energy minimization problem, however the reasoning is done in 3D. The results, reported in Table Ic and Fig. 9e, show that the method performs quite well and with a few exceptions is able to segment environment correctly. [13] has one real failure case on our data (see Fig. 9e, row 10), where the method fails to segment out the environment. We mention here that this instance of our dataset is quite challenging, as it consists of an atypical layout: two very large rooms with very high ceilings, and connected to a number of smaller rooms; this instance is also much bigger in terms of real world volume than the others. We attribute the poor performance of [13] to (i) failure to encode the environment primitives of this challenging instance into one of the 6 types of structural paths described in the method, and (ii) failure of the Markov clustering algorithm to compute the correct number of rooms.

We outperform the state of the art for quantitative measures of area which are often desirable when gathering statistics on building layouts - see Table Ie and Fig. 9g. Our results are better even though we have presented a more generic approach, independent of viewpoint information. We note that instance 10 of our dataset is a partial failure case for our method as well, due to the fact that in parts of the environment the ceiling is not visible / has not been scanned, which results in our method marking those areas as empty. Comparing our final results with the segmentation resulting after Section IV we note that we are able to improve the overall segmentation of the environment into rooms, while at the same time the resulting segments have much clearer boundaries composed of a few straight-line segments. The runtime of our method varies between 30 - 120 seconds, depending on the complexity (i.e. number of points) of the input point cloud.

Our approach suffers from the typical issues present in cell complex reconstruction approaches. First, as we rely on a parametric plane detection method to discover walls, atypical candidate wall planes are not accurately detected (e.g. in the case of buildings where the walls are not perpendicular to the ground plane). A volumetric cell complex approach can sidestep this issue, but it is less robust to noise and computationally more expensive due to the increase in dimensionality. Second, we rely extensively on the presence of ceiling points to detect regions that are outside, with the disadvantage that regions with partial / unobserved ceilings are not detected accurately.

VII. CONCLUSIONS

We have presented an automatic method for reconstructing room information from raw point cloud data using only 3D point information. Our method is the first such method to not rely on viewpoint information without relying on a Manhattan frame assumption. We have given extensive quantitative results that prove our method outperforms the state of the art on both fine grain and coarse grain segmentation tasks.

There are several avenues for future research. One appealing direction is to leverage the power of object recognition systems to aid in a truly semantic manner. Knowledge of, for instance, which room is the kitchen versus which room is a bedroom would be invaluable for robotic applications. A different avenue for research involves the use of volumetric primitives for full 3D model reconstructions.

ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge Sebastian Ochmann [11] and Claudio Mura [12], [13] for their help regarding the comparison to their work, as well as Patric Jensfelt and John Folkesson for their guidance and comments which greatly improved this manuscript.

REFERENCES

[1] G. Grisetti, C. Stachniss, and W. Burgard, "Improved techniques for grid mapping with rao-blackwellized particle filters," *IEEE transactions on Robotics*, vol. 23, no. 1, pp. 34–46, 2007.

[2] R. Bormann, F. Jordan, W. Li, J. Hampp, *et al.*, "Room segmentation: Survey, implementation, and analysis," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1019–1026, IEEE, 2016.

[3] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *The International Journal of Robotics Research*, 2009.

[4] X. Xiong, A. Adan, B. Akinci, and D. Huber, "Automatic creation of semantically rich 3d building models from laser scanner data," *Automation in Construction*, vol. 31, pp. 325–337, 2013.

[5] J. Xiao and Y. Furukawa, "Reconstructing the world's museums," *International Journal of Computer Vision*, vol. 110, no. 3, pp. 243–258, 2014.

[6] E. Turner, P. Cheng, and A. Zakhor, "Fast, automated, scalable generation of textured 3d models of indoor environments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 3, pp. 409–421, 2015.

[7] E. Turner and A. Zakhor, "Floor plan generation and room labeling of indoor environments from laser range data," in *Computer Graphics Theory and Applications (GRAPP), 2014 International Conference on*, pp. 1–12, IEEE, 2014.

[8] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3d semantic parsing of large-scale indoor spaces," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR, 2016.

[9] J. Straub, G. Rosman, O. Freifeld, J. J. Leonard, and J. W. Fisher, "A mixture of manhattan frames: Beyond the manhattan world," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3770–3777, IEEE, 2014.

[10] S. Oesau, F. Lafarge, and P. Alliez, "Indoor scene reconstruction using feature sensitive primitive extraction and graph-cut," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 90, pp. 68–82, 2014.

[11] S. Ochmann, R. Vock, R. Wessel, and R. Klein, "Automatic reconstruction of parametric building models from indoor point clouds," *Computers & Graphics*, vol. 54, pp. 94–103, 2016.

[12] C. Mura, O. Mattausch, A. J. Villanueva, E. Gobbetti, and R. Pajarola, "Automatic room detection and reconstruction in cluttered indoor environments with complex room layouts," *Computers & Graphics*, vol. 44, pp. 20–32, 2014.

[13] C. Mura, O. Mattausch, and R. Pajarola, "Piecewise-planar reconstruction of multi-room interiors with arbitrary wall arrangements," *Computer Graphics Forum*, 2016.

[14] A. Hermans, G. Floros, and B. Leibe, "Dense 3d semantic mapping of indoor scenes from rgb-d images," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2631–2638, IEEE, 2014.

[15] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[16] S. Song and J. Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," *arXiv preprint arXiv:1511.02300*, 2015.

[17] X. Ren, L. Bo, and D. Fox, "Rgb-(d) scene labeling: Features and algorithms," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2759–2766, IEEE, 2012.

[18] R. Schnabel, R. Wahl, and R. Klein, "Efficient ransac for point-cloud shape detection," in *Computer graphics forum*, vol. 26, pp. 214–226, Wiley Online Library, 2007.

[19] N. Bore, P. Jensfelt, and J. Folkesson, "Querying 3d data by adjacency graphs," in *International Conference on Computer Vision Systems*, pp. 243–252, Springer, 2015.

[20] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 271–279, 1989.

[21] S. Thrun and A. Bücken, "Integrating grid-based and topological maps for mobile robot navigation," in *Proceedings of the National Conference on Artificial Intelligence*, pp. 944–951, 1996.

[22] B. Chazelle and H. Edelsbrunner, "An optimal algorithm for intersecting line segments in the plane," *Journal of the ACM (JACM)*, vol. 39, no. 1, pp. 1–54, 1992.

[23] H. Edelsbrunner, J. O'Rourke, and R. Seidel, "Constructing arrangements of lines and hyperplanes with applications," *SIAM Journal on Computing*, vol. 15, no. 2, pp. 341–363, 1986.

[24] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

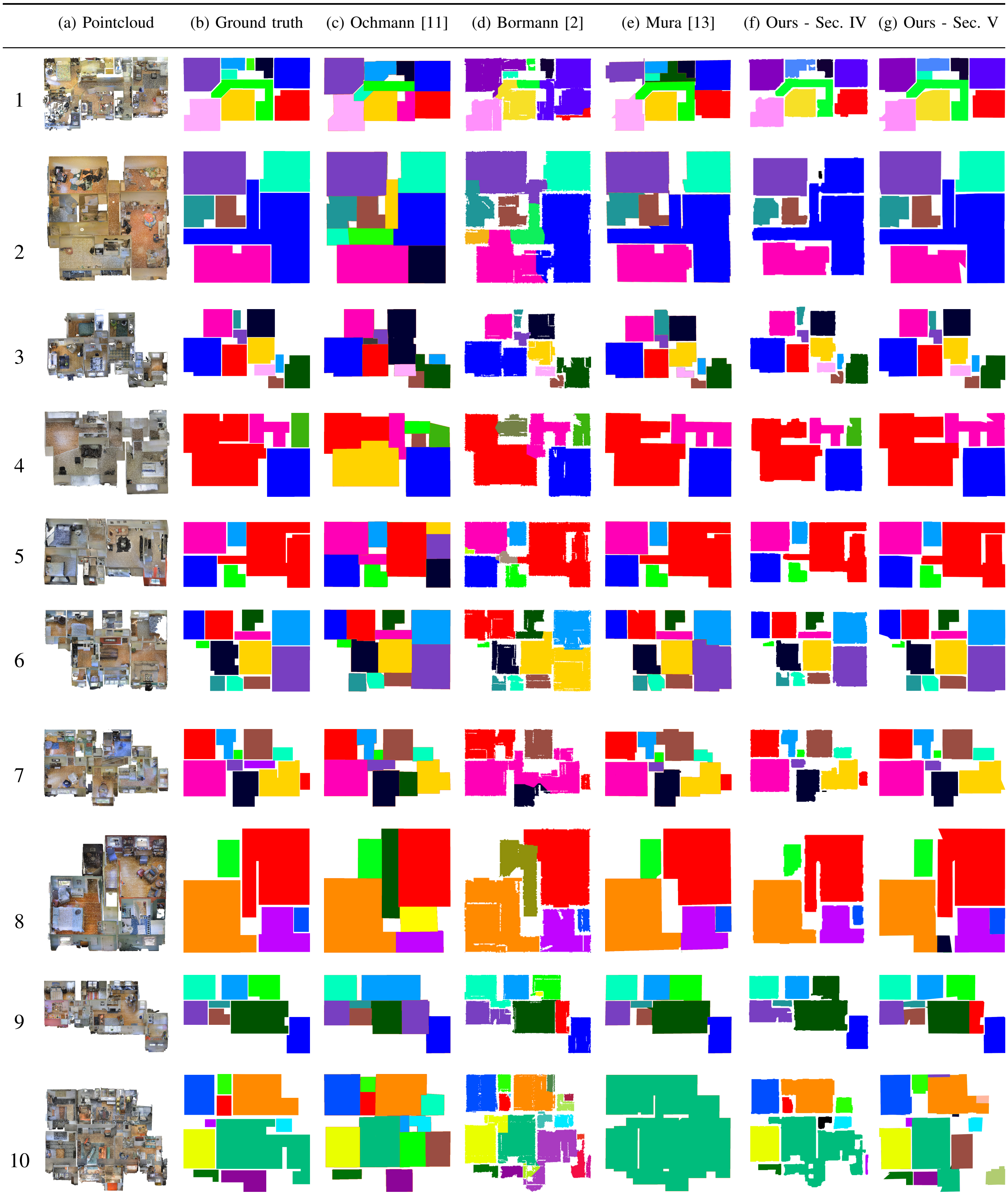

Fig. 9: Qualitative results of semantic segmentations: (a) shows the original data (ceilings removed for clarity) and (b) the ground truth labelling. We compare with the methods from [11] [2] [13], displayed in (c),(d) and (e). We also show our results in (f) - simple segmentation as defined in Section IV and (g) - segmentation after the energy minimization step described in Section V. All images show a top-down view of an orthographic projection of the data.