Room Reconstruction from a Single Spherical Image by Higher-order Energy Minimization

Kosuke Fukano*, Yoshihiko Mochizuki*, Satoshi Iizuka*, Edgar Simo-Serra*, Akihiro Sugimoto†,
and Hiroshi Ishikawa*
* Waseda University, Tokyo, Japan
Email: ich38940@ruri.waseda.jp, {motchy,iizuka,esimo}@aoni.waseda.jp, hfs@waseda.jp
† National Institute of Informatics, Tokyo, Japan
Email: sugimoto@nii.ac.jp

***Abstract*—We propose a method for understanding a room from a single spherical image, i.e., reconstructing and identifying structural planes forming the ceiling, the floor, and the walls in a room. A spherical image records the light that falls onto a single viewpoint from all directions and does not require correlating geometrical information from multiple images, which facilitates robust and precise reconstruction of the room structure. In our method, we detect line segments from a given image, and classify them into two groups: segments that form the boundaries of the structural planes and those that do not. We formulate this problem as a higher-order energy minimization problem that combines the various measures of likelihood that one, two, or three line segments are part of the boundary. We minimize the energy with graph cuts to identify segments forming boundaries, from which we estimate structural the planes in 3D. Experimental results on synthetic and real images confirm the effectiveness of the proposed method.**

I. Introduction

We present a method to reconstruct (i.e., to identify the floor, the ceiling, and the walls of) a simple room in a single spherical image. An example of the input image is shown in Fig. 1. We aim to reconstruct the most basic structure of the whole room such as the walls, even in a cluttered image.

In the common 3D reconstruction technique like stereo and structure from motion, establishing a correspondence between the points in multiple images is necessary. In the cluttered indoor scenes, it is often difficult to accurately determine the correspondence due to the texture on the wall and obstructing objects like furnitures [1]. Moreover, because such methods that depend on multiview input need that each point in the scene to appear in at least two images from different angles, it is complicated to cover the whole room with the necessary number of images with adequate parallax.

In our method, we first detect line segments in the input image. The aim is to determine the borders of the floor, the ceiling, and the walls of the room among the detected line segments. Since the detected line segments include those that are not part of the borders, we classify them into the edges and other line segments using higher-order energy minimization. Finally, we infer the planes that form the room (i.e., the ceiling etc.) from the border line segments.

II. Related Work

A similar method that try to estimate the structure from a single image without using stereo or structure from motion is [2]. Similar to our method, it infers the structure of the room from the line segments detected in the input image. They however use the images from an ordinary camera, making it impossible to recover the whole room in a single image. In contrast, our method not only captures the whole room in the single image but also takes advantage of the unique geometry of the spherical image to reconstruct the walls.

Y. Zhang et al. [3] uses panorama images to find a 3D bounding boxes of a room and objects therein. However, their method is much more elaborate than ours and involves learning from a dataset of annotated panorama images. In contrast, our method does not use any training set and infer the geometry of the room solely from a single spherical image.

III. Proposed Method

In this work, we use a single spherical image taken indoors as the input, and reconstruct the shape of the room. We assume the room and the spherical image satisfy the following conditions.

- The room is roughly a cuboid, consisting of six rectangular planar walls (including the floor and the ceiling).
- All the six planes and their intersecting lines appear at least partially in the image.

Fig. 2 depicts the planes and the line segments that constitute the room.

The objective here can be thought of as dividing the image into regions corresponding to the six walls of the room. For that, we solve the problem of identifying the spherical line segments on the unit sphere that are borders of the walls. Fig. 3 illustrates the flow of the process. The method consists of the following three steps.

1) We detect line segments in the input image. Fig. 4 illustrates the process of detecting line segments in spherical images. First, we project the input image onto six planar images. Then, we detect line segments in the planar images using the method by Matas et al. [4]. Finally, we project the detected line segments back onto

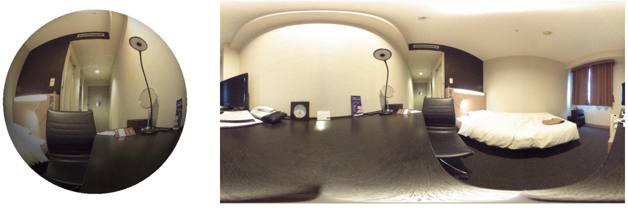


Fig. 1. An example input image. Left: A half of the spherical image. Right: the whole image projected onto a rectangle.

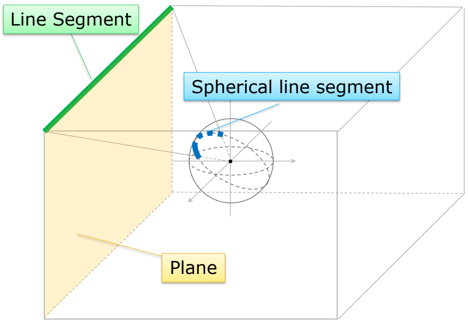


Fig. 2. Planes and line segments that constitute the room.

the unit sphere. We denote the set of the detected line segments on the sphere by $S = \{s_1, s_2, \cdots, s_n\}$.

2) We identify the subset $S^* \subset S$ of line segments corresponding to the wall borders. Since S contains line segments other than the wall borders, such as those caused by wall texture and objects present in the room, we classify the line segments into S^* and $S \backslash S^*$. We treat this as a labeling problem and solve it by minimizing a higher-order energy.
3) We infer the six walls using S^*.

We explain the second and the third steps in more detail.

A. Line Segments Extraction

To classify the line segments into S^* and $S \backslash S^*$, we treat this as a labeling problem and minimize a higher-order energy. Let n be the number of the line segments in S. We denote the i'th line segment by s_i and its length by $|s_i|$. Here, we talk about the line segments on the unit sphere, i.e., parts of great circles. So $|s_i|$ is actually the length of an arc on the unit sphere. The labeling $L = (l_i)_{i=1,\ldots,n}$ of the line segments determine which ones belong to S^*. That is, the label l_i is 1 if the line segment s_i is in S^* and 0 if it is not. We define an energy function on L such that it has smaller values when it is more likely that the labeling indicates the correct set of line segments in S^*. By finding the labeling that minimize this energy function, we infer the set S^* of line segments that constitutes the wall borders. For minimization, we use the graph-cut algorithm for higher-order energies in [5]. The energy is a weighted sum of five potential functions:

$$\begin{aligned} E(L) &= w_{\text{length}} E_{\text{length}}(L) + w_{\text{collinear}} E_{\text{collinear}}(L) \\ &+ w_{\text{cross}} E_{\text{cross}}(L) + w_{\text{corner}} E_{\text{corner}}(L) \\ &+ w_{\text{plane}} E_{\text{plane}}(L). \end{aligned} \qquad (1)$$

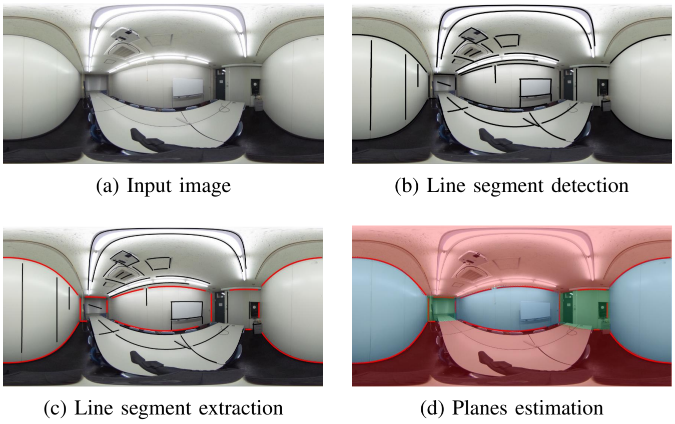

(a) Input image
(b) Line segment detection
(c) Line segment extraction
(d) Planes estimation

Fig. 3. The flow of the proposed method. (a) Input spherical image. (b) Detect line segments. The detected line segments are shown in black. (c) Classify the detected line segments into those which are borders of the walls (S^*) and all others, such as lines drawn on the wall. The line segments in S^* are shown in red. Other line segments are shown in black. (d) Using the line segments in S^*, infer the six plane walls (including the floor and the ceiling) that constitute the room. Walls facing each other are shown in the same color.

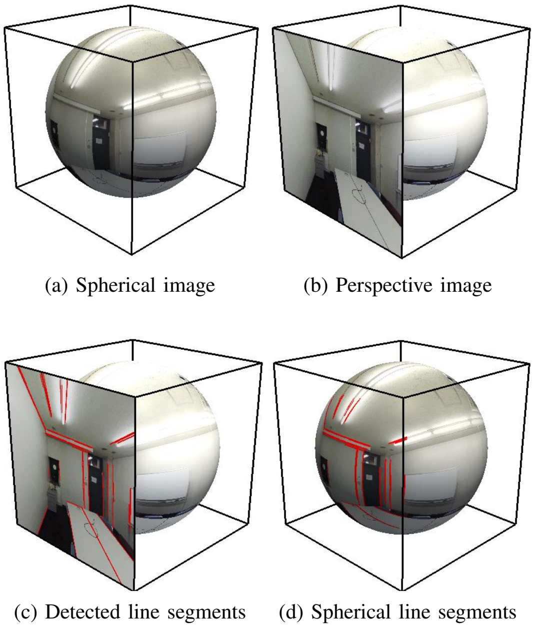

(a) Spherical image
(b) Perspective image
(c) Detected line segments
(d) Spherical line segments

Fig. 4. Detecting line segments in a spherical image. (a) A spherical image. (b) Project the input image onto six planar images. (c) Detect line segments in the planar images. (d) Project the detected segments back onto the unit sphere.

The potentials have the following meaning: E_{length} is about the length of the segments, $E_{\text{collinear}}$ is about pairs of segments on the same great circle, E_{cross} is about pairs of intersecting segments, E_{corner} is about triples of segments that form a corner, and E_{plane} is about quadruples of segments that form a rectangle. Next, we discuss each of these potentials.

1) Potential Regulating Length: Since there are various objects and textures in the scene, line segments with various lengths are detected. Many of shorter segments tend to originate from objects and textures. We can threshold the minimum length, but it is conceivable that some segments appear short even though they are actually long borders of the walls because of various reasons, such as occlusion by objects. Instead, we regulate the length of the line segment in a soft way by introducing a potential function $E_{\text{length}}(L)$ that encourages longer segments to be labeled as a part of the wall borders by giving them lower values. It is given by the following:

$$\begin{aligned} E_{\text{length}}(L) &= \sum_{i=1}^{n} f_{\text{length}}(l_i) \qquad (2) \\ f_{\text{length}}(l_i) &= \begin{cases} 1 & (l_i = 1 \wedge \frac{|s_i|}{\pi} < d_{\text{length}}) \\ 0 & (\text{otherwise}) \end{cases} \end{aligned}$$

where $0 < d_{\text{length}} < 1$ is a threshold.

2) Potential Measuring Collinearity: In detecting line segments in the spherical image, in some cases two detected segments are actually parts of one segment, because of an occlusion by objects, because of the lighting condition, or because the segment does not fit in one of the six projected planes. Let s_i and s_j be two detected line segments that are actually one segment. As such, their labels should always coincide. Thus, we introduce a potential that lowers the energy when such a pair are labeled the same. Let C be the set of pairs of indices of line segments that can be regarded to form a single segment. We define $E_{\text{collinear}}$ as follows:

$$\begin{aligned} E_{\text{collinear}}(L) &= \sum_{(i,j)\in C} f_{\text{collinear}}(l_i, l_j) \qquad (3) \\ f_{\text{collinear}}(l_i, l_j) &= \begin{cases} -e_{ij} & (l_i = l_j) \\ 0 & (\text{otherwise}) \end{cases} \end{aligned}$$

Here, e_{ij} is a positive value that becomes larger when the distance between the line segments s_i and s_j becomes smaller.

3) Potential Discouraging Segments Crossing: The line segments that are borders or boundaries of two walls should not look crossing another line segment. Examples of such crossing and non-crossing are shown in Fig. 5. Thus we introduce a potential that raise the energy value when crossing segments are labeled as borders of walls. Let F be the set of pairs of indices of line segments that cross each other. Let d_{ij} be the minimum of the distances between the crossing point and either endpoint of s_i or s_j. We define E_{cross} using F and d_{ij} as follows:

$$E_{\text{cross}}(L) = \sum_{(l_i,l_j)\in F} f_{\text{cross}}(l_i, l_j) \tag{4}$$

$$f_{\text{cross}}(l_i, l_j) = \begin{cases} d_{ij} & (l_i = l_j = 1) \\ 0 & (\text{otherwise}) \end{cases} \tag{5}$$

4) Potential Assessing Three Segments Forming A Corner: Consider three planes P_1, P_2, and P_3 in the 3D space such that each pair among them are perpendicular to each other. Let T_1 be the intersection of P_2 and P_3, T_2 the intersection of P_1 and P_3, and T_3 the intersection of P_1 and P_2. The segments T_1, T_2, and T_3 intersect perpendicularly at one point, called the perpendicular vertex, in the 3D space [6]. Let s_1, s_2, and s_3 be the projection of T_1, T_2, and T_3 onto the unit sphere. Then s_1, s_2, and s_3 are called the line segments that constitute a corner. Fig. 6 shows a perpendicular vertex and its line segments that constitute a corner. When s_1, s_2, and s_3 are line segments that constitute a corner, they are likely to be in S^*. Thus we add a potential that encourage such triple to have label 1. First, let q_1, q_2, and q_3 be the intersections (points on the unit sphere) of the three pairs of line segments out of s_1, s_2, and s_3. Let g be the barycenter of the three points and d_{ijk} the average distance between g and q_1, q_2, and q_3. Then the potential $E_{\text{corner}}(L)$ is defined as follows:

$$\begin{aligned} E_{\text{corner}}(L) &= \sum_{(i,j,k)\in A} f_{\text{corner}}(l_i, l_j, l_k) \qquad (6) \\ f_{\text{corner}}(l_i, l_j, l_k) &= \begin{cases} -\cos(d_{ijk}) & (l_i = l_j = l_k = 1) \\ 0 & (\text{otherwise}) \end{cases} \end{aligned}$$

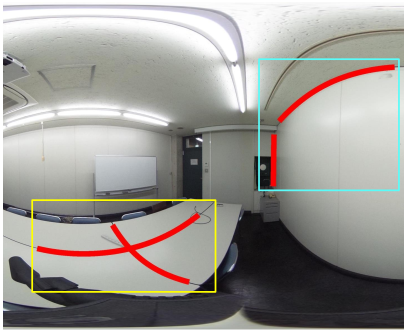


Fig. 5. Crossing segments and non-crossing segments that are boundary of the walls. The two line segments in the yellow box on the left cross each other. The ones in the cyan box on the right do not, and are borders of walls.

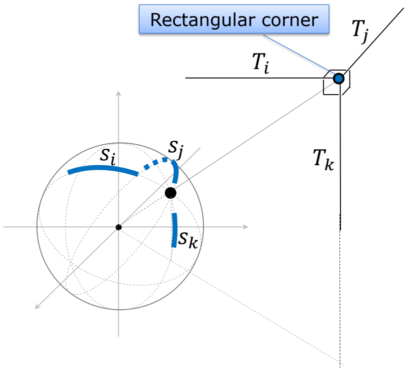


Fig. 6. A perpendicular vertex and its line segments that constitute a corner.

5) Potential Estimating Planes: The potential $E_{\text{plane}}(L)$ estimates how likely sets of four line segments form planes. It lowers the energy if the four are likely to form a rectangle and they are all labeled 1. It is defined as follows:

$$\begin{aligned} E_{\text{plane}}(L) &= \sum_{(h,i,j,k)\in V} f_{\text{plane}}(l_h, l_i, l_j, l_k) \qquad (7) \\ f_{\text{plane}}(l_h, l_i, l_j, l_k) &= \begin{cases} -r_{\text{plane}} & (l_h = l_i = l_j = l_k = 1) \\ 0 & (\text{otherwise}) \end{cases} \end{aligned}$$

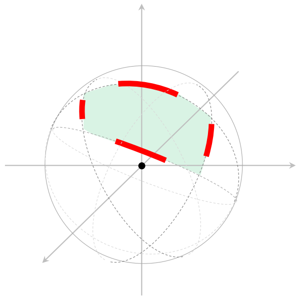


Fig. 7. Four line segments that can be part of a rectangle (left) and that cannot (right).

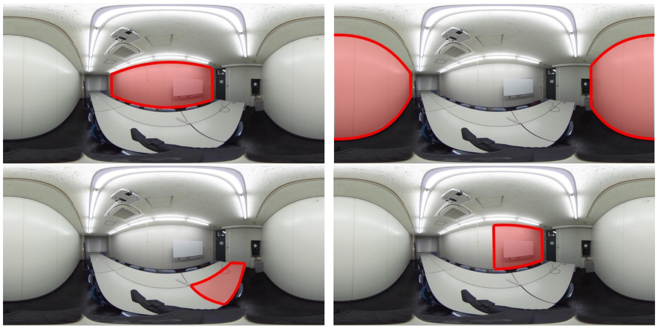


Fig. 8. Examples of planes that constitute the room (upper row) and those that do not (lower row).

Here, V is the set of quadruples of line segments that can possibly form a rectangle in the 3D space. This is determined by examining necessary conditions on spatial relationships between the segments. We refer to [7] and omit the details here because of the lack of space. Fig. 7 illustrates one such condition. Also, r_{plane} is the ratio of the sum of the lengths of the four segments to the perimeter of the rectangle that can be formed, projected on the unit sphere.

B. Inferring the Walls

Using the set S^* of line segments likely to form the borders of the walls, we infer the planes that constitute the ceiling, the floor, and the walls in the following steps:

1) Generate the set $\mathcal{Q}$ of candidate quadruples of line segments that can form planes.
2) Eliminate the candidates that cannot constitute the room, i.e., those that cannot be the four borders of a floor, a ceiling, or a wall of the room.

To generate the candidates, we use the same necessary conditions [7] on spatial relationships between the segments we used in §III-A5. The set $\mathcal{Q}$ includes both planes that constitute the room (i.e., are either the ceiling, the floor, or the walls) and those do not. Examples of both cases are shown in Fig. 8. Here, planes that constitute the room have common borders with other such planes. Therefore, if any of the borders of Q is not also an border of another plane in $\mathcal{Q}$, Q can be excluded as the candidate. We eliminate the planes that do not constitute the room as follows:

1) For $Q \in \mathcal{Q}$, determine if a plane in $\mathcal{Q}$ with a common border with Q can be found for each of Q's borders.
2) If they cannot be found, remove Q from $\mathcal{Q}$.
3) Repeat 1 and 2 for all $Q \in \mathcal{Q}$.
4) Repeat 1 through 3 Until the number of the elements of $\mathcal{Q}$ no longer decreases.

We take the set $\mathcal{Q}$ after this process as the set of the planes constituting the room. Finally, we cluster the normal direction of the remaining planes with k-means with $k = 6$.

C. Reconstructing the Room

To obtain the whole room shape as a cuboid, we can utilize the method in Y. Zhang et al. [3]. The algorithm uses the segments corresponding to each edge of a cuboid to construct a set of constraints for a least-square problem. Since we know the correspondence between $\mathcal{Q}$ and the faces of a cuboid, we can supply the necessary information to the algorithm. The method also needs the axes of vanishing point directions to which the cuboid should be aligned. Thus we need to estimate the vanishing points corresponding to the input image by some existing method [8].

IV. Experimental Results

Here, we describe the experiments and their results we performed to evaluate our method. Fig. 9 shows the two input images with resolution 3584×1792 pixels (left) and 2048×1024 pixels (right), respectively. Fig. 10 shows the detected line segments in blue and the selected lines by energy minimization in red.

The weights were automatically chosen by trying different combinations and then choosing the one that resulted in the fewest line segments, out of the combinations that the final wall inference does not fail (by ending up with empty $\mathcal{Q}$).

For the left image, 314 segments were detected, which contain all the wall borders. The automatically chosen weights for the minimization were $w_{\text{length}} = 1$, $w_{\text{collinear}} = 5$, $w_{\text{cross}} = 100$, and $w_{\text{corner}} = 6$. After the energy minimization, 128 segments were selected as possible borders. Then, 29792 plane candidates were generated, and 1691 planes out of them remained after the wall inference. The planes are clustered into six directions as shown in Fig. 11.

The right image is the more complicated case. The 551 detected segments include the edges of the table, a woman, and her hand. The automatically chosen weights were $w_{\text{length}} = 3$, $w_{\text{collinear}} = 1$, $w_{\text{corner}} = 74$, and $w_{\text{corner}} = 3$. After the energy minimization, 219 segments were selected. Then, 42721 plane candidates were generated, and 135 planes out of them remained after the wall inference. Fig. 12 shows the eliminated candidates.

In both cases, we correctly identified all the six planes. Finally, we applied the method [3] to estimate the cuboid for each set of candidates as shown in Fig. 13.

A. Analysis of the weight for E_{cross}

Changing the weight w_{cross} of E_{cross}, we computed the Precision and Recall for the labeling for the right image. Other

Fig. 9. Two input images.

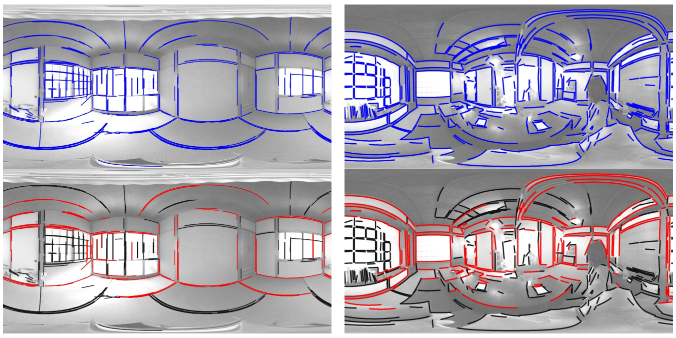


Fig. 10. Detected line segments (top) and classification results by energy minimization (bottom).

weights used were: $w_{\text{length}} = 1, w_{\text{collinear}} = 100, w_{\text{corner}} = 6$, and $w_{\text{plane}} = 0$.

Table I shows the classification accuracy by energy minimization with varying w_{cross}. The ground truth for the accuracy calculation is manually annotated as shown in Fig. 14. Fig. 15 shows the corresponding results.

Discussion. From Table I and Fig. 15, it can be seen that when w_{cross} is small most segments are classified to be wall borders, while the accuracy is the highest when $w_{\text{cross}} = 1500$. This indicates that E_{cross} is effective. On the other hand the Recall suffers when w_{cross} is too big.

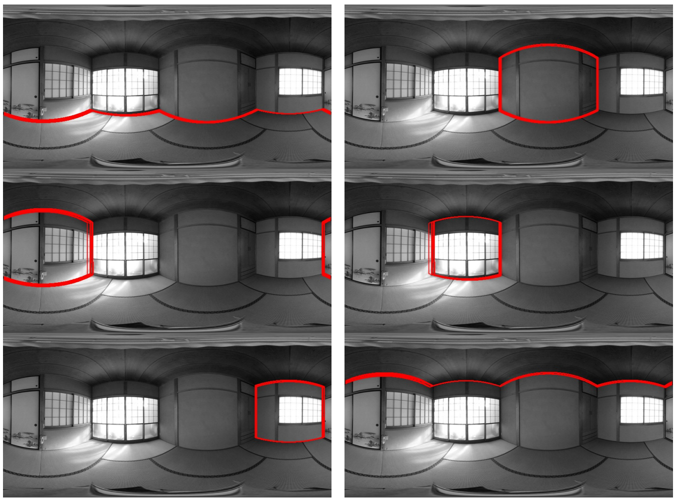


Fig. 11. Eliminated candidates of the ceiling, floor, and walls.

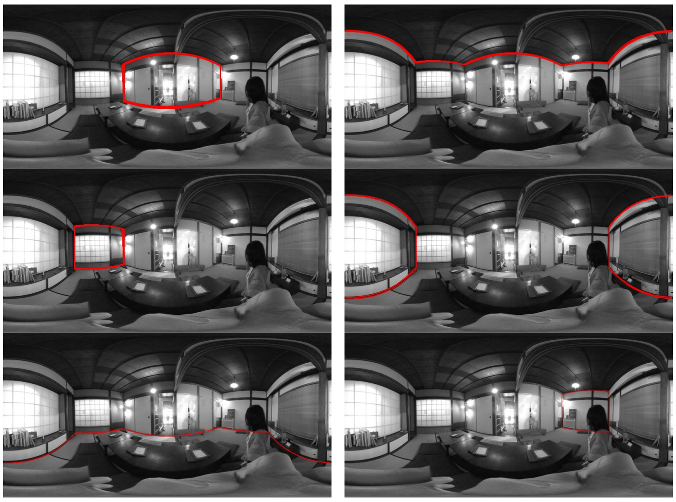


Fig. 12. Eliminated candidates of the ceiling, floor, and walls.

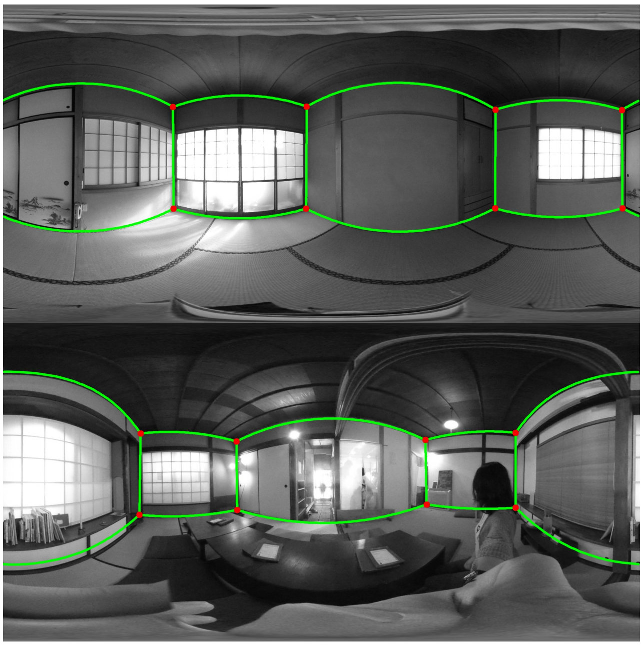


Fig. 13. Final results.

B. Degenerate case

For some cases, the wall inference may fail, i.e., the set of planes $\mathcal{Q}$ becomes empty. This is because insufficient segments are detected to reconstruct the room or the classification result is not accurate enough.

Fig. 16 is the classification result of the detected segments

TABLE I
CLASSIFICATION ACCURACY WITH VARYING w_{cross}.

w_{cross}	0	100	500	1000	1500	2000	2500
Precision	0.28	0.31	0.54	0.69	0.94	0.91	0.20
Recall	1.00	1.00	0.98	0.98	0.98	0.69	0.01

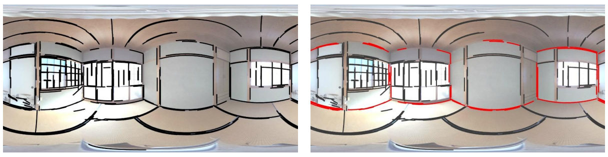
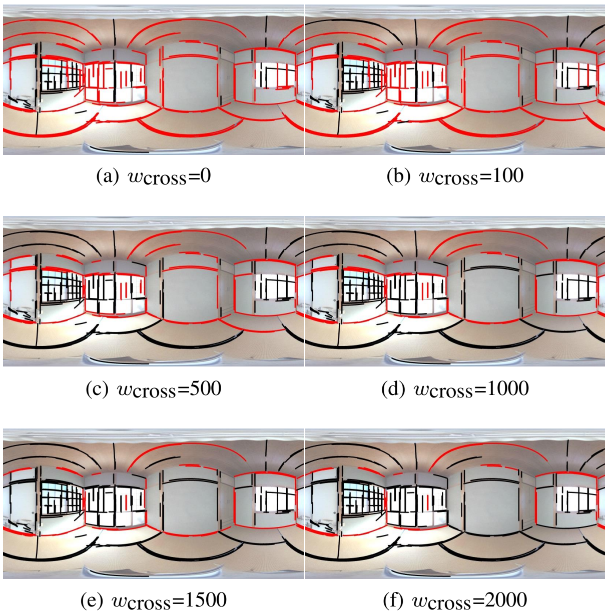


Fig. 14. Ground truth for the accuracy evaluation. Left: Detected line segments. Right: Those that are actually part of the wall borders are shown in red.



(a) w_{cross}=0

(b) w_{cross}=100

(c) w_{cross}=500

(d) w_{cross}=1000

(e) w_{cross}=1500

(f) w_{cross}=2000

Fig. 15. Classification results with varying w_{cross}. Those segments that were labeled 1 (i.e., as wall borders) are shown in red.

for an image. In this case, the inference failed. However, an intermediate $\mathcal{Q}$ roughly forms a cuboid without the most of outliers as shown in Fig. 17. We can also apply the cuboid estimation method [3] to get the approximate result (Fig. 18).

V. Conclusion

We present a method to reconstruct a simple room from a single spherical image. We first detect line segments, which we classify into borders and others using higher-order energy minimization. Then we infer the planes that constitute the room from the boundary line segments. We correctly determined the structure of the rooms in experiments with real images. Future work would include loosening the assumption that we imposed here.

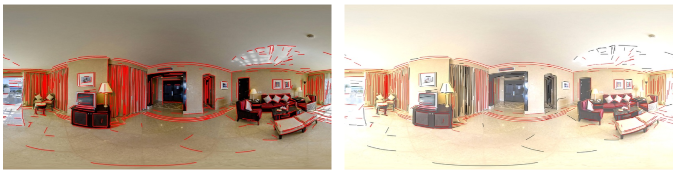


Fig. 16. Detected segments and the classification result by energy minimization.

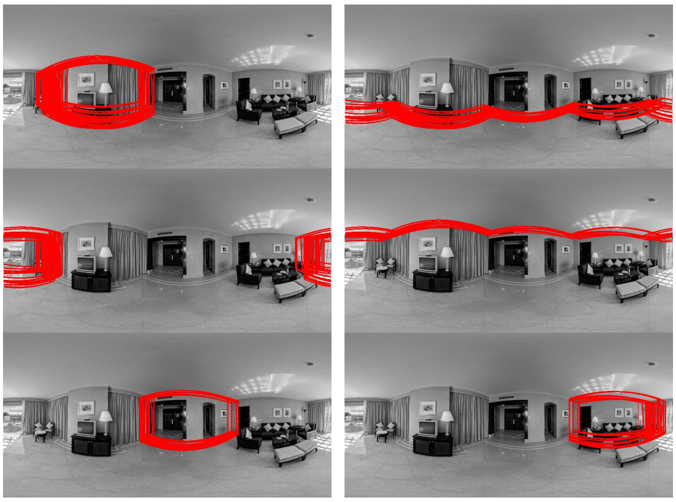


Fig. 17. Eliminated candidates at one step before the algorithm failed.

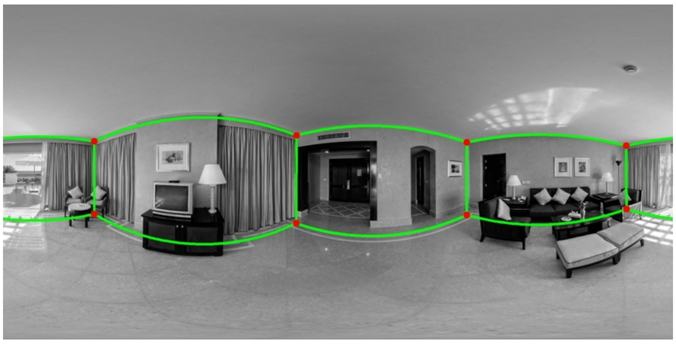


Fig. 18. The estimated cuboid.

Acknowledgment

This work was partially supported by JST CREST.

References

[1] R. Cabral and Y. Furukawa, “Piecewise planar and compact floorplan reconstruction from images,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014.

[2] D. C. Lee, M. Hebert, and T. Kanade, “Geometric reasoning for single image structure recovery,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2009.

[3] Y. Zhang, S. Song, P. Tan, and J. Xiao, “PanoContext: A whole-room 3d context model for panoramic scene understanding,” in *The 13th European Conference on Computer Vision (ECCV)*, Sep. 2014.

[4] J. Matas, C. Galambos, and J. Kittler, “Robust detection of lines using the progressive probabilistic Hough transform,” *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 119 – 137, 2000.

[5] H. Ishikawa, “Transformation of general binary MRF minimization to the first-order case,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1234–1249, 2011.

[6] K. Kanatani, *Group-Theoretical Methods in Image Understanding*, ser. Springer Series in Information Science. Springer Berlin Heidelberg, 1990, vol. 20.

[7] H. Kato and M. Billinghurst, “Marker tracking and HMD calibration for a video-based augmented reality conferencing system,” in *the 2nd IEEE and ACM International Workshop on Augmented Reality*, 1999, pp. 85–.

[8] C. Rother, “A new approach for vanishing point detection in architectural environments,” in *The British Machine Vision Conference (BMVC)*, 2000.