# Indoor RGB-D Compass from a Single Line and Plane

Pyojin Kim[1] pjinkim1215@gmail.com
Brian Coltin[2] brian.j.coltin@nasa.gov
H. Jin Kim[1] hjinkim@snu.ac.kr

[1]ASRI, Seoul National University, South Korea
[2]SGT, Inc., NASA Ames Research Center, USA

## Abstract

*We propose a novel approach to estimate the three degrees of freedom (DoF) drift-free rotational motion of an RGB-D camera from only a single line and plane in the Manhattan world (MW). Previous approaches exploit the surface normal vectors and vanishing points to achieve accurate 3-DoF rotation estimation. However, they require multiple orthogonal planes or many consistent lines to be visible throughout the entire rotation estimation process; otherwise, these approaches fail. To overcome these limitations, we present a new method that estimates absolute camera orientation from only a single line and a single plane in RANSAC, which corresponds to the theoretical minimal sampling for 3-DoF rotation estimation. Once we find an initial rotation estimate, we refine the camera orientation by minimizing the average orthogonal distance from the endpoints of the lines parallel to the MW axes. We demonstrate the effectiveness of the proposed algorithm through an extensive evaluation on a variety of RGB-D datasets and compare with other state-of-the-art methods.*

## 1. Introduction

Camera orientation estimation from a sequence of images is a fundamental problem for many applications in computer vision [23, 10] and robotics [20, 25]. Recent visual odometry (VO) and visual simultaneous localization and mapping (V-SLAM) methods [7, 9, 18] have shown promising results in estimating camera orientation from a variety of video sequences. However, these approaches cannot avoid drift error in the rotation estimate without computationally expensive SLAM techniques (loop closure, global 3D map construction).

Several studies [2, 6, 25, 14] have focused on accurate and drift-free rotation estimation in urban and indoor scenes consisting of parallel and orthogonal lines and planes, called the Manhattan world (MW) [4]. They exploit structural regularities to achieve accurate 3-DoF rotation estimation by using the distribution of surface normal vectors and points at infinity, i.e., vanishing points (VPs). The accuracy of VO has been improved dramatically in [30, 16, 15] by using the MW assumption in rotation estimation. Although they can estimate the rotational motion of the camera accurately by exploiting significant structural organization, there are still some problems: multiple orthogonal planes or many consistent lines must remain visible throughout the entire video sequence. In practice, robots often encounter harsh environments where there are insufficient structural regularities (see Fig. 1 and Fig. 7), resulting in the failure of rotation estimation or a loss of accuracy.

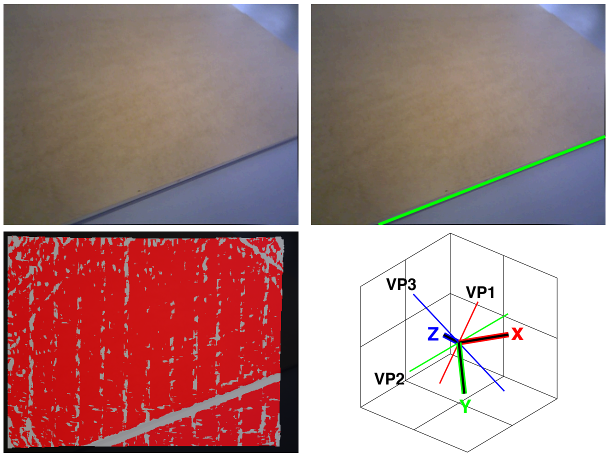


Figure 1. In an uncharacteristic scene (top left), given a single line detected in the RGB image (top right) and a single plane identified in the depth image (bottom left), the proposed method can estimate drift-free camera orientation and VPs (bottom right).

To address these issues, we propose a novel approach that estimates absolute 3-DoF camera orientation from only a single line and plane to recognize the spatial regularities of structural environments as shown in Fig. 1. We detect and track the normal vector of a plane from the depth image in order to determine two orientation angles of the 3-DoF rotation. The remaining orientation angle is computed with a line from the RGB image, which lies on the plane and is parallel to the MW axes. We incorporate the 3-DoF rotation estimation from only a single line and plane into the model estimation step of the RANSAC, which is the minimal solution for rotation estimation [2]. Furthermore, we refine the initial rotation estimate by minimizing the average orthogonal distance from the multiple lines, which are parallel to the MW axes. Our algorithm requires a plane and a line on the plane aligned with the MW to be visible, which is typically the case in most indoor environments.

Extensive evaluations show that the proposed method produces accurate and drift-free camera orientation on a variety of video sequences compared to other state-of-the-art approaches. The contributions of this work are as follows:

- We present a novel approach to estimate accurate and drift-free 3-DoF camera orientation from only a single line and plane in the RANSAC framework.
- We refine the initial rotation estimate with the parallel and orthogonal lines to obtain a more accurate 3-DoF camera orientation.
- We evaluate the proposed algorithm on the ICL-NUIM [11] and TUM [27] RGB-D datasets, showing robust, stable, and accurate performance.

## 2. Related Work

The use of the Manhattan world (MW) estimation for determining the orientation of a camera has been studied previously due to its importance in high-level vision applications such as 3D reconstruction and scene understanding. The approaches for understanding the structural regularities in man-made environments can be classified into either estimating the VPs from the intersection of multiple parallel lines in the image or estimating the principal normal vectors of the surface with 3D information in depth image.

A VP, which is invariant to camera translation movements, has been widely used for tracking the rotational motion of the camera accurately [1, 24, 2]. In [6, 2], Manhattan frame (MF) estimation is performed based on three lines with two of the lines parallel and orthogonal to the third in RANSAC [8], which is the minimal sampling for rotation estimation. The method in [5] finds a triplet of orthogonal vanishing points with RANSAC-based line clustering to track the camera orientation along a video sequence in real-time. [17] jointly estimates the VPs and camera orientation based on sequential Bayesian filtering without the MW assumption. These VP-based methods, however, are not robust and stable in the presence of spurious or noisy line segments. A sufficient number of parallel and orthogonal lines should exist in the image for accurate and reliable rotation estimation.

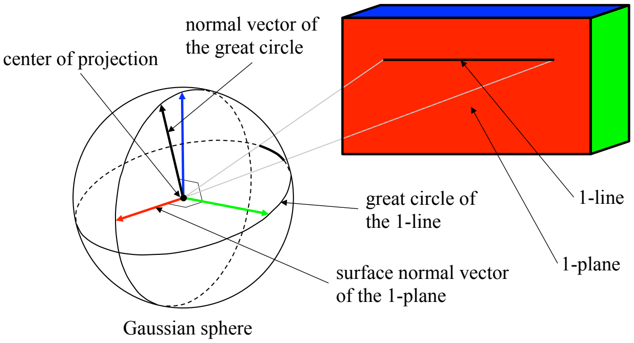


Figure 2. Geometric relationships between the line, plane, and the Gaussian sphere in the MW. The line is projected onto the Gaussian sphere as a great circle. Each orthogonal plane and its corresponding normal vector are drawn with the same color. The normal vector of the great circle (black) from the line and the normal vector of the dominant plane (red) do not have to be perpendicular.

Recent studies have utilized 3D information to estimate dominant orthogonal directions in a MW from a depth sensor like a Kinect camera. [25, 26] propose real-time maximum a posteriori (MAP) inference algorithms for estimating MW in the surface normal distribution of a scene on a GPU. The method in [30, 16] estimates drift-free camera orientation with an efficient SO(3)-constrained mean shift algorithm given the surface normal vector distribution. In [3, 14], a branch-and-bound (BnB) strategy is employed to guarantee the globally optimal Manhattan frame estimation. While these approaches based on the surface normals demonstrate more stable and accurate rotation estimation results than VP-based methods, at least two orthogonal planes must be observable in the depth images.

Prior research has used the connection between VPs in the RGB image and 3D information from depth image to perform MW estimation. Given an a priori known normal vector of the horizon plane, an estimate of the camera orientation is performed with additional line segments in RANSAC [2]. The method in [15] tracks the MW utilizing both lines and planes together, but it requires a sufficient number of lines in the image. [19] estimates global geometry of indoor MW environments by integrating RGB images with associated depth data.

## 3. Proposed Method

We propose a new method for estimating a drift-free 3-DoF rotational motion of an RGB-D camera from the RGB and depth image pairs. For each pair of RGB-D images, we perform two steps: 1) estimate the absolute camera orientation with respect to the MF using only a single line and plane in RANSAC [8]; and 2) refine this initial rotation estimate with parallel and orthogonal lines from inliers. An overview of the proposed algorithm is shown in Fig. 3.

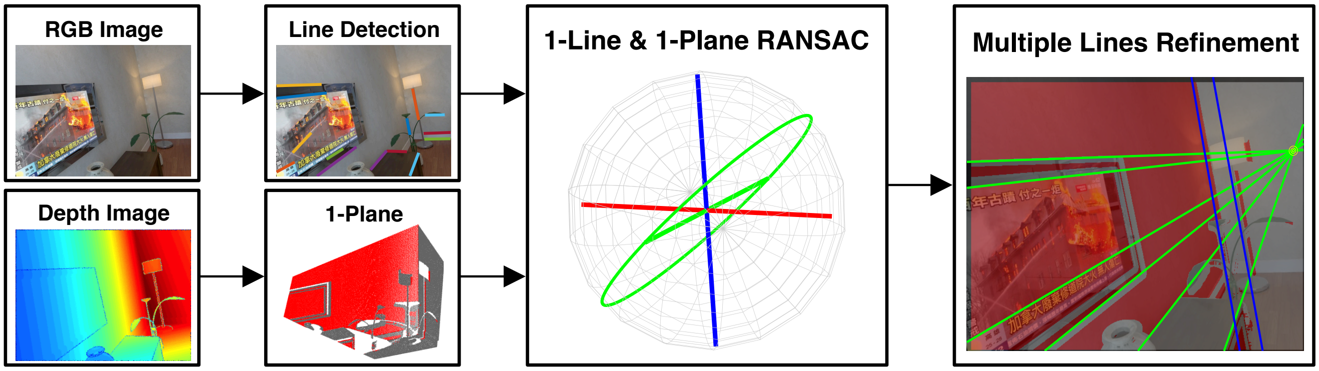


Figure 3. Overview of the proposed method. We estimate the drift-free camera orientation by using only a single line and plane in RANSAC. We refine the initial rotation estimation by minimizing the orthogonal distance from the endpoints of the parallel and orthogonal lines.

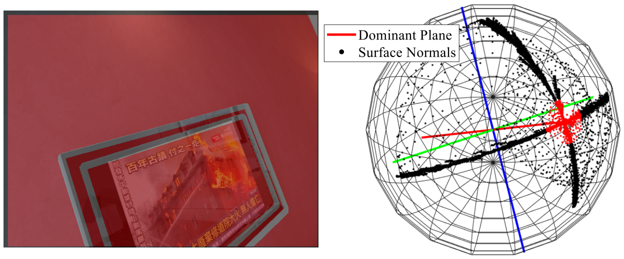


Figure 4. The normal vector (red axis) of a dominant plane (red) tracked from the distribution of the surface normals (black dots). We project the relevant surface normals inside a conic section of the red axis into the tangential plane to perform the mean shift.

### 3.1. Dominant Plane Detection and Tracking

We first detect a dominant plane in the current environment from a depth image's 3D point cloud with a RANSAC algorithm [29]. The algorithm first randomly selects three points and computes the model parameters (normal vector) of the corresponding plane. It then checks the number of inliers exceed a given threshold by calculating the distance between the 3D points and the plane. It repeats these first two steps until it finds the best (dominant) 3D plane supported by the largest number of inliers.

We track the normal vector of the dominant plane with a mean shift algorithm based on the tangent space Gaussian MF (TG-MF) model [26] given the density distribution of surface normal vectors on the Gaussian sphere $\mathbb{S}^2$ [30] in Fig. 4. The unit surface normal vector of each pixel is calculated by taking the cross product of two tangential vectors at the 3D points in the point cloud. To obtain the noiseless tangential vectors for stable surface normal vectors, we average the surrounding tangential vectors within a certain neighborhood, which can be done efficiently and quickly using integral images [13]. Unlike the previous approaches [30, 16], we only track a single normal vector of the dominant plane as shown in Fig. 4. By using the tracked (detected) normal vector from the previous frame as an initial value, we perform the mean shift algorithm in the tangent plane $\mathbb{R}^2$ of the Gaussian sphere $\mathbb{S}^2$ with a Gaussian kernel (for full details, refer to [30]). Although we could use RANSAC to discover a dominant plane for every frame, tracking is less expensive and makes a smoother estimate.

If the density distribution of the surface normal vectors around the currently tracked normal vector is too low, we reinitialize and detect a new dominant plane again with plane model-based RANSAC. We assign the normal vector of the new dominant plane to the closest axis in the MF under the assumption that the MF does not change too much between subsequent frames. There are 24 possible representations for the same MF orientation; we convert the matrices representing the MF into a unique canonical form [30] for consistent tracking.

### 3.2. One Line and One Plane RANSAC

Our approach utilizes line and plane geometric features, which provide the theoretical minimal solution for 3-DoF rotation estimation [2] as illustrated in Fig. 2. A plane provides two constraints on the two orientation angles, and the remaining orientation angle $\theta$ is constrained by a line. Once a dominant plane and a parallel line lying on the corresponding plane are found in the MW, they can determine the orientation of the Manhattan structure uniquely. By using this geometric feature, we estimate the 3-DoF drift-free rotational motion of the camera with respect to the MW in the RANSAC framework.

We detect $N$ line segments using LSD [28], and calculate their corresponding unit normal vectors of great circles on the Gaussian sphere $\mathbb{S}^2$. Each RANSAC iteration starts by randomly selecting one great circle among $N$ line segments. Given the tracked normal vector (the first VP $v_1$) of the dominant plane from the previous Section 3.1, we take cross product between the first VP and the normal vector of

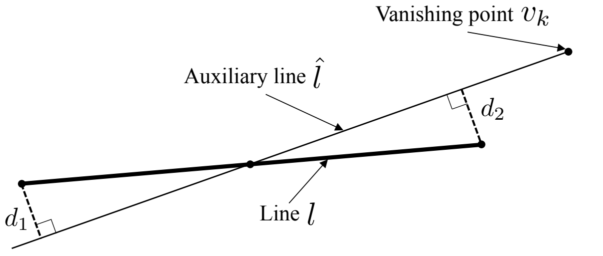


Figure 5. Orthogonal distance metric from the endpoints of the line $l$ to an auxiliary line $\hat{l}$ defined with the VP and the middle point of the line $l$ in the image space.

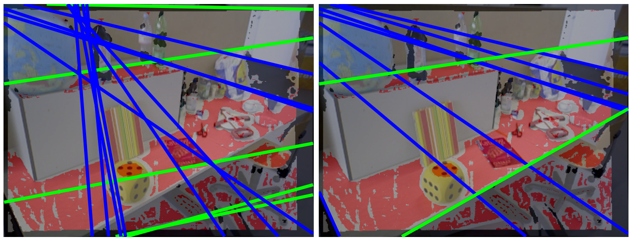


Figure 6. Improved performance to recognize the regularities of structural space by using the Eq. (2) (right) compared to the typical RANSAC algorithm (left).

the selected great circle to define the second VP $v_2$. The third VP $v_3$ (blue) is automatically determined by the cross product of the first VP (red) and the second VP (green) as shown in Fig. 2.

To evaluate the currently estimated orientation in RANSAC, we use the average orthogonal distance in the image plane as illustrated in Fig. 5 [12], which is a function of lines and camera orientation (VPs). The average orthogonal distance can be computed from the endpoints of the line $l$ to an auxiliary line $\hat{l}$, which passes through the closest VP and the middle point of the line $l$ as follows:

$$d_{i,k} = \left(d_{i1,k} + d_{i2,k}\right)/2 \tag{1}$$

$$\text{where} \quad d_{i1,k} = \frac{|A_{i,k}u_{i1} + B_{i,k}v_{i1} + C_{i,k}|}{\sqrt{A_{i,k}^2 + B_{i,k}^2}}$$

where $A_{i,k}, B_{i,k}, C_{i,k}$ are the auxiliary line parameters of the $i$-th line segment with the $k$-th VP, and $u_{i1}, v_{i1}$ is the first endpoint of the $i$-th line segment in the image plane.

Unlike the typical RANSAC algorithm in [2] which uses only the number of inliers, we find the largest consensus line set utilizing not only the average orthogonal distance $d_{i,k}$ but also the length of a line segment [21]:

$$vote\left(v_k\right) = \sum_{i=1}^{M_k} w_1 \left(1 - \frac{d_{i,k}}{t_a}\right) + w_2 \left(\frac{length(l_i)}{max(length(l))}\right) \tag{2}$$

where $M_k$, $k \in \{2, 3\}$ is the number of associated line segments for each VP $v_2$ and $v_3$, respectively. $d_{i,k}$ and $t_a$ are the average orthogonal distance of the $i$-th line segment with the $k$-th closest VP and a certain threshold defined by user (in our experiments, 1 pixel). The $i$-th line length relative to the maximum line length is also considered in the second term in Eq. (2) because the longer the lines are, the more reliable they are. The weights $w_1$ and $w_2$ denote the importance of each term, the orthogonal distance and the line length, respectively (in our experiments, 0.7 and 0.3). When we calculate the vote value in Eq. (2), we do not use the line segments parallel to the tracked normal vector of the dominant plane (the first VP $v_1$) because the plane normal tracking on the surface normal vector distribution is quite accurate [25, 16]. We find the lines and the corresponding camera orientation (VPs) leading to the highest total vote sum value from Eq. (2). It is noteworthy that the proposed one line and one plane RANSAC is computationally efficient since the number of required line sample to perform model estimation is only one similar to [22]. As the number of RANSAC iterations (computational complexity) exponentially increases depending on the number of required samples [12], using only one line sample to estimate the model makes our algorithm computationally inexpensive.

Fig. 6 shows the effectiveness of the continuous criteria in Eq. (2) compared to the standard RANSAC. When we try the standard RANSAC, it sometimes fails because it only considers whether the average orthogonal distance is smaller or larger than a certain threshold dichotomically. If there are many spurious or noisy lines, it cannot recognize the structural regularities correctly in the left of Fig. 6. We can find the correct inlier line set by using the continuous criteria written in Eq. (2) in the right of Fig. 6.

Our approach can fail when there is not any line that is parallel to the MW axes, or we cannot find any valid lines because of extreme motion blur. In other words, for our algorithm to succeed, we must have at least: 1) a plane from the depth image; and 2) a line from the RGB image, which lies on the plane and is parallel to the MW axes. While these geometric conditions may seem restrictive, our extensive experiments on multiple datasets in the Section 4 show that they often hold in most structural indoor environments, and our approach achieves better accuracy and demonstrates the effectiveness.

### 3.3. Multiple Lines Refinement

The initial rotation estimate from only a single line and plane in the previous RANSAC step can be affected by noise in the line segments, resulting in suboptimal rotation estimation. To estimate more accurate and optimal camera orientation, we further refine the initial rotation estimation from the single line and plane RANSAC by minimizing the average orthogonal distance with parallel and orthogonal lines in inliers.

Since the tracked normal vector of the dominant plane on the surface normal vector distribution is relatively accurate [25, 16], the cost function, which is the average orthogonal distance written in Eq. (1), is only a function of the remaining one orientation angle $\theta$ constrained by multiple inlier lines. We express the 3-DoF camera orientation (VPs) as the axis-angle representation where the direction of an axis of rotation is the tracked unit normal vector of the dominant plane, and the magnitude of the rotation about the axis is the remaining orientation angle $\theta$. The optimal drift-free camera orientation, which minimizes the orthogonal distance of all parallel and orthogonal inlier lines found in the RANSAC, can be obtained by solving the following optimization problem:

$$\theta^{*} = \arg\min_{\theta} \sum_{k=2}^{3} \sum_{i=1}^{M_k} \left(d_{i,k}\left(\theta\right)\right)^2 \tag{3}$$

where $M_k$, $k \in \{2, 3\}$ is the number of parallel or orthogonal lines related to the $k$-th VP counted in the RANSAC as inliers. $d_{i,k}(\theta)$ denotes the orthogonal distance of the $i$-th line segment with the $k$-th VP in the image space. We use the Levenberg–Marquardt (LM) algorithm for solving Eq. (3). By additionally constraining the remaining orientation angle $\theta$ from the parallel and orthogonal lines found in RANSAC, we can estimate more accurate and consistent rotational motion compared to the initial rotation estimate directly from the RANSAC process.

Note that the first RANSAC step (Section 3.2) and the second (Section 3.3) optimization of the algorithm seem to be redundant as both estimate the rotational motion of the camera. The additional refinement step, however, makes the estimated camera orientation more accurate and consistent by utilizing multiple lines. We validate the effect of the refinement in the next evaluation section.

# 4. Evaluation

We evaluate the proposed approach on a variety of RGB-D video sequences in man-made structural environments:

- *ICL-NUIM* [11] is a synthetic dataset consisting of a collection of RGB and depth images at 30 Hz captured in a living room and office with ground-truth camera orientation. The synthesized RGB and depth images are corrupted by the modeled sensor noise to simulate typically observed real-world artifacts. It is challenging to estimate the accurate 3-DoF camera rotation throughout the entire video sequences due to very low texture and a single plane as shown in Fig. 7.

- *TUM RGB-D* [27] is a famous dataset for VO/V-SLAM evaluation, containing RGB-D images from a Microsoft Kinect RGB-D camera in various indoor environments as shown in Fig. 7. It is recorded in room-scale environments with ground-truth camera trajectories provided by a motion capture system.

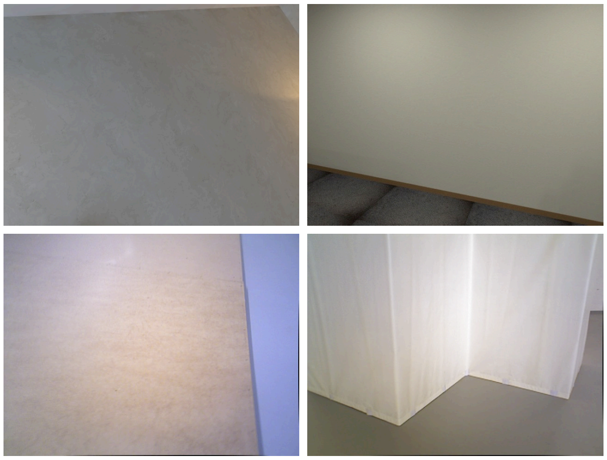


Figure 7. Example images of the Manhattan world from the ICL-NUIM [11] (top) and TUM RGB-D [27] (bottom) datasets, which are captured in a very uncharacteristic scene.

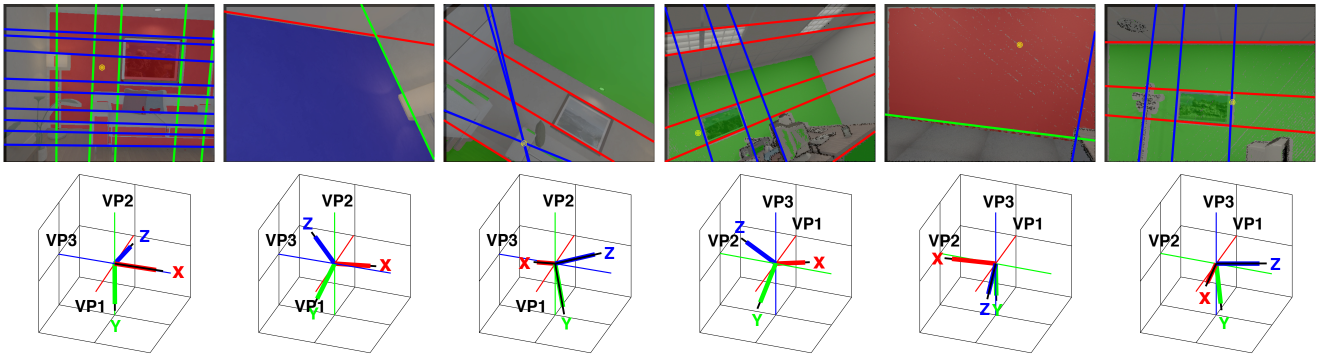
| Experiment | Proposed | GOME | OLRE | OPRE | ROVE | # of frame |
|---|---|---|---|---|---|---|
| Living Room 0 | **0.31** | × | × | × | × | 1507 |
| Living Room 1 | **0.38** | 8.56 | 3.72 | 0.97 | 26.74 | 965 |
| Living Room 2 | **0.34** | 8.15 | 4.21 | 0.49 | 39.71 | 880 |
| Living Room 3 | **0.35** | × | × | 1.34 | × | 1240 |
| Office Room 0 | 0.37 | 5.12 | 6.71 | **0.18** | 29.11 | 1507 |
| Office Room 1 | 0.37 | × | × | **0.32** | 34.98 | 965 |
| Office Room 2 | 0.38 | 6.67 | 10.91 | **0.33** | 60.54 | 880 |
| Office Room 3 | 0.38 | 5.57 | 3.41 | **0.21** | 10.67 | 1240 |

Table 1. Comparison of the average value of the absolute rotation error (degrees) on ICL-NUIM benchmark [11].

We compare the proposed algorithm against other state-of-the-art 3-DoF camera orientation estimation methods using lines and planes, namely GOME [14], OLRE [2], OPRE [30], and ROVE [17]. GOME and OPRE estimate the drift-free rotational motion of the camera by tracking the distribution of the surface normal vectors from the depth images, while OLRE and ROVE utilize many consistent line features from the RGB images to estimate the camera orientation. The proposed method, GOME, OLRE, and OPRE rely on the MW assumption whereas ROVE does not require the MW in the scene.

## 4.1. ICL-NUIM Dataset

We measure the mean value of the absolute rotation error (ARE) [30] in degrees, and present the evaluation results in Table 1. The smallest rotation error for each dataset is bolded. Other methods using only multiple lines or planes sometimes fail to track the camera orientation (marked as



Frame 196 Frame 931 Frame 1478 Frame 160 Frame 530 Frame 918

(a) Living Room 0 (b) Office Room 1

Figure 8. Results of the proposed method in the ICL-NUIM dataset (a) 'Living Room 0' and (b) 'Office Room 1'. Clustered lines and segmented plane with inferred MW are overlaid on the RGB images (top). Colored thick, thin lines denote the estimated 3-DoF camera orientation and the MW (VPs), and the black lines represent the true pose of the camera (bottom). Each of the colored lines and plane in the top images corresponds to the VP of the same color in the bottom.

× in Table 1) due to multiple lines or orthogonal planes not always being visible throughout the entire video sequences. In 'Living Room 0', at one point the camera sees only a single plane with very low texture, leading to failure of other approaches. The proposed method can continue tracking the absolute camera orientation stably and accurately as shown in Fig. 8. The tracked normal vector of the dominant plane is changed depending on the current situation.

Our approach outperforms the other methods for most cases. In 'Office Room' environments, OPRE performs slightly better thanks to sufficient surface normals distribution throughout the estimation period, but the proposed algorithm performs nearly as well. The average ARE of the proposed method is 0.36 degrees, while GOME, OLRE, OPRE, and ROVE are 6.82, 5.79, 0.55, and 33.63 degrees respectively. Since ROVE does not utilize the MW assumption, ROVE cannot estimate the drift-free camera orientation, resulting in accumulation of ARE over time. The main reason for the improved performance is that the proposed method can stably track the absolute rotations even when the camera sees only a planar surface with little texture by exploiting the minimal sampling (one line and one plane) to recognize structural regularities.

The advantage of the additional refinement step in the proposed method described in Section 3.3 becomes clear when plotting the ARE statistics from the dataset 'Living Room 1' in Fig. 9. We can observe that there are some large ARE (marked as a red cross) from the proposed method when the refinement step is not performed. The optimization with parallel and orthogonal lines found in the RANSAC as inliers enables to estimate the drift-free camera rotation more consistently and accurately.

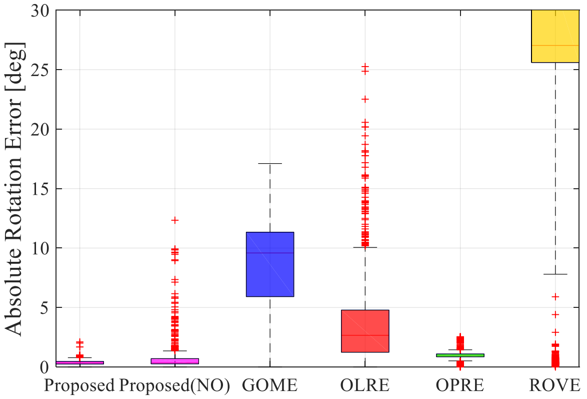


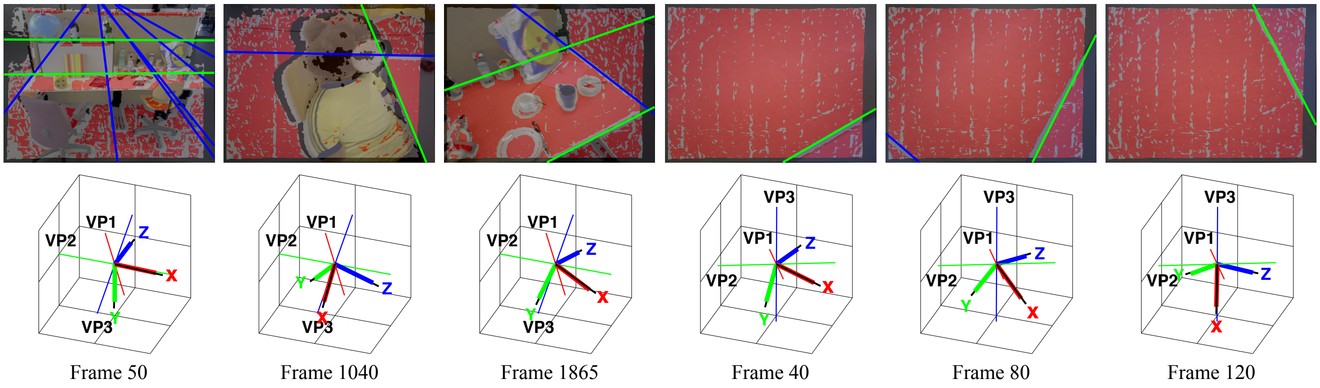
Figure 9. Comparison of the proposed approach with and without refinement step (NO) versus the other algorithms. We use the absolute rotation error from the 'Living Room 1' to obtain error statistics.

| Experiment | Proposed | GOME | OLRE | OPRE | ROVE | # of frame |
|---|---|---|---|---|---|---|
| fr3_longoffice | **1.75** | × | × | 4.99 | × | 2488 |
| fr3_nostruc_notex | **1.51** | × | × | × | × | 239 |
| fr3_nostruc_tex | **2.15** | × | 46.18 | × | 16.45 | 1639 |
| fr3_struc_notex | **1.96** | 4.07 | 11.22 | 3.01 | × | 794 |
| fr3_struc_tex | **2.92** | 4.71 | 8.21 | 3.81 | 13.73 | 907 |
| fr3_cabinet | 2.48 | 2.59 | × | **2.42** | × | 1112 |
| fr3_large_cabinet | **2.04** | 3.74 | 38.12 | 36.34 | 28.41 | 984 |

Table 2. Comparison of the average value of the absolute rotation error (degrees) on TUM RGB-D dataset [27].

### 4.2. TUM RGB-D Dataset

We evaluate the proposed and other algorithms on the video sequences of the TUM RGB-D dataset, which contain structural regularities (lines or planes) in the observed



(a) fr3_longoffice

(b) fr3_nostruc_notex

Figure 10. Inferred MW (VPs) orientation, the clustered lines and plane with the proposed method are overlaid on top of the RGB images in the TUM RGB-D dataset (a) 'fr3_longoffice' and (b) 'fr3_nostruc_notex'. The thick blue line (Z) in the bottom represents the head of the camera. The proposed method accurately estimates the absolute 3-DoF camera orientation and the VPs from a single line and plane.

scenes. We also investigate the effect of the existence of structure and texture components in the scenes on the camera rotation estimation. Table 2 compares the average ARE results of the proposed and other methods. Our method can track accurate and drift-free camera rotational motion even in insufficient (imperfect) structural environments like 'fr3_longoffice' or 'fr3_nostruc_notex' as shown in Fig. 10. However, other approaches require at least two orthogonal planes (GOME, OPRE) or many consistent line segments (OLRE, ROVE) throughout the entire motion estimation process. While other methods are significantly affected by the presence or absence of the structure and texture components in the scenes, the proposed method shows accurate MW estimation not only in abundant but also in very low structure and texture environments with the help of the minimal solution (one line and one plane).

We can also observe the effect of the refinement step in the proposed method by drawing the boxplot of the ARE from the dataset 'fr3_struc_tex' in Fig. 11. Outliers marked as red cross are removed, and the average ARE of the proposed method decreases thanks to the proposed additional refinement step.

Please refer to the video clips submitted with this paper showing more details about the experiments.[1]

## 5. Conclusion

We propose a new method that is able to perform accurate and drift-free camera orientation estimation under insufficient structural environments by exploiting a single line and plane in RANSAC, which are the minimal solution for 3-DoF rotation estimation. We refine the initial rotation estimate by minimizing the average orthogonal distance from the endpoints of the parallel and orthogonal lines found in the RANSAC as inliers. The proposed algorithm is tested thoroughly with a large number of RGB-D datasets on the video sequences, and shows accurate and drift-free rotation estimation results in the environments where the structural regularities are challenging to find. Our method is currently tested with an RGB-D camera in indoor environments. In the future, we will try to implement the proposed algorithm with a stereo camera and possibly extend to outdoor urban environments.

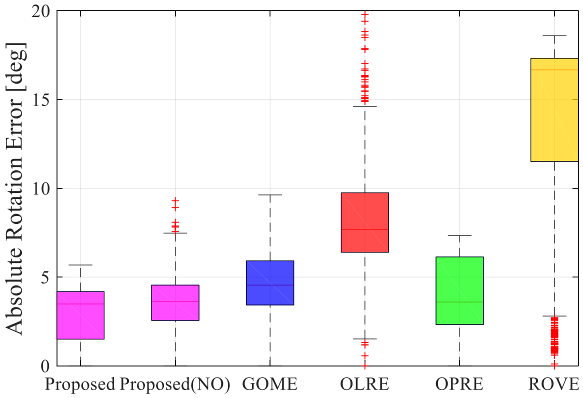


Figure 11. The statistical distribution of the absolute rotation error from the 'fr3_struc_tex' for each method. 'Proposed(NO)' denotes the performance of the proposed approach without refinement step with multiple lines.

## Acknowledgements

This work was supported by the Samsung Smart Campus Research Center at Seoul National University (0115-20170013) and the Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Ministry of Trade,

[1] Video available at `https://youtu.be/qusvgMequqM`

Industry & Energy (MOTIE, Korea) (No. 10080301).

## References

[1] J. C. Bazin, C. Demonceaux, P. Vasseur, and I. Kweon. Motion estimation by decoupling rotation and translation in catadioptric vision. *Computer Vision and Image Understanding*, 2010.

[2] J.-C. Bazin and M. Pollefeys. 3-line RANSAC for orthogonal vanishing point detection. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*.

[3] J.-C. Bazin, Y. Seo, C. Demonceaux, P. Vasseur, K. Ikeuchi, I. Kweon, and M. Pollefeys. Globally optimal line clustering and vanishing point estimation in Manhattan world. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*.

[4] J. M. Coughlan and A. L. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*.

[5] W. Elloumi, S. Treuillet, and R. Leconge. Real-time camera orientation estimation based on vanishing point tracking under Manhattan world assumption. *Journal of Real-Time Image Processing*, 2017.

[6] A. Elqursh and A. Elgammal. Line-based relative pose estimation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*.

[7] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In *Readings in computer vision*. Elsevier, 1987.

[9] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza. SVO: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 2017.

[10] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from RGB-D images. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*.

[11] A. Handa, T. Whelan, J. McDonald, and A. J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *Robotics and automation (ICRA), 2014 IEEE international conference on*.

[12] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[13] D. Holz, S. Holzer, R. B. Rusu, and S. Behnke. Real-time plane segmentation using RGB-D cameras. In *Robot Soccer World Cup*. Springer, 2011.

[14] K. Joo, T.-H. Oh, J. Kim, and I. So Kweon. Globally optimal Manhattan frame estimation in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[15] P. Kim, B. Coltin, and H. J. Kim. Low-drift visual odometry in structured environments by decoupling rotational and translational motion. In *Robotics and automation (ICRA), 2018 IEEE international conference on*.

[16] P. Kim, B. Coltin, and H. J. Kim. Visual odometry with drift-free rotation estimation using indoor scene regularities. In *2017 British Machine Vision Conference*.

[17] J.-K. Lee, K.-J. Yoon, et al. Real-time joint estimation of camera orientation and vanishing points. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*.

[18] R. Mur-Artal and J. D. Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 2017.

[19] N. Neverova, D. Muselet, and A. Trémeau. 2 1/2 D scene reconstruction of indoor scenes from single RGB-D images. In *Computational Color Imaging*. Springer, 2013.

[20] C. F. Olson, L. H. Matthies, M. Schoppers, and M. W. Maimone. Stereo ego-motion improvements for robust rover navigation. In *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*.

[21] C. Rother. A new approach to vanishing point detection in architectural environments. *Image and Vision Computing*, 2002.

[22] D. Scaramuzza and F. Fraundorfer. Visual odometry [tutorial]. *IEEE robotics & automation magazine*, 2011.

[23] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*. Springer, 2012.

[24] S. N. Sinha, D. Steedly, and R. Szeliski. A multi-stage linear approach to structure from motion. In *European Conference on Computer Vision*. Springer, 2010.

[25] J. Straub, N. Bhandari, J. J. Leonard, and J. W. Fisher. Real-time Manhattan world rotation estimation in 3D. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*.

[26] J. Straub, O. Freifeld, G. Rosman, J. J. Leonard, and J. W. Fisher. The Manhattan frame model–Manhattan world inference in the space of surface normals. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[27] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*.

[28] R. G. Von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. LSD: A fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence*, 2010.

[29] M. Y. Yang and W. Förstner. Plane detection in point cloud data. In *Proceedings of the 2nd int conf on machine control guidance, Bonn*, 2010.

[30] Y. Zhou, L. Kneip, C. Rodriguez, and H. Li. Divide and conquer: Efficient density-based tracking of 3D sensors in Manhattan worlds. In *Asian Conference on Computer Vision*. Springer, 2016.