Line-based Global Localization of a Spherical Camera in Manhattan Worlds

Tsubasa Goto, Sarthak Pathak, Yonghoon Ji, Hiromitsu Fujii, Atsushi Yamashita, and Hajime Asama

球形相机、室内定位

室内地图一般知。。。？

***Abstract*— Localization is an important task for mobile service robots in indoor spaces. In this research, we propose a novel technique for indoor localization using a spherical camera. Spherical cameras can obtain a complete view of the surroundings allowing the use of global environmental information. We take advantage of this in order to estimate camera position and the orientation with respect to a known 3D line map of an indoor environment, using a single image. We robustly extract 2D line information from the spherical image via spherical-gradient filtering and match it to 3D line information in the line map. Our method requires no information about the 3D-2D line correspondences. In order to avoid a complicated six degrees of freedom (6 DoF) search for position and orientation, we use a Manhattan world assumption to decompose the line information in the image. The 6 DoF localization process is divided into two phases. First, we estimate the orientation by extracting the three principle directions from the image. Then, the position is estimated by robustly matching the distribution of lines between the image and the 3D model via a spherical Hough representation. This decoupled search can robustly localize a spherical camera using a single image, as we demonstrate experimentally.**

3D线地图已知，单张图片，估计相机位姿

曼哈顿世界假设：大多数平面是三个相互正交的平面之一

先方向、后位置

I. Introduction

Recently, the use of mobile service robots in indoor spaces like offices, homes, etc. has become quite popular. Localization is a very important task for all such applications. Usual approaches like Global Positioning Systems (GPS) are difficult to use in indoor spaces and cannot provide information about six degrees of freedom (6 DoF) robot pose. Thus, many approaches have been suggested towards localization without GPS. Particularly, beacon-based localization methods can be effective in indoor spaces. Sensors or fiducial markers [1] can be placed at various positions and can help in localizing the robot. However, they involve a modification of the indoor space. Moreover, sensor-based approaches are expensive to install and maintain over time. Wi-Fi has also been shown to be effective for localization in indoor spaces, especially in office environments where many hot spots are available [2], and requires no environmental modifications. However, its accuracy is too low for most practical purposes. Furthermore, it cannot manage the orientation estimation problem.

WIFI 定位精度低

In contrast, camera-based methods can achieve high accuracy. Most approaches involving cameras use Visual Simultaneous Localization and Mapping (VSLAM) [3]. VSLAM algorithms can not only localize the robot, but also generate a 3D map of the surrounding environment. However, typically, maps of indoor environments are known and there is no need to generate them. Moreover, the maps generated by SLAM need to be registered to pre-existing maps in order to achieve global localization, which is not a trivial task [4].

All authors are with the Department of Precision Engineering, Graduate School of Engineering, The University of Tokyo, Japan {goto,pathak,ji,fujii,yamashita,asama}@robot.t.u-tokyo.ac.jp

In this research, we propose a global, camera-based indoor localization method. Towards this purpose, we use a spherical camera, which can capture information from all directions and is much more effective at self-localization as compared to a perspective camera. This is because the information obtained by the perspective camera is limited by its small field-of-view and is particularly ineffective in situations where the camera is facing a wall or an obstacle.

球形相机视野大，获得的信息更多

We focus on a map-based localization method that makes use of a 3D line map. Such line maps can be easily obtained from the architectural blueprints or construction CAD models of the environment. Since it is difficult to include accurate color information in the model, we rely on line information alone. Our proposed method estimates the position and orientation of a spherical camera by matching the line distributions in the Hough space from both arbitrary poses in the 3D line map and a real 2D spherical camera image. Our method requires no information about the 3D-2D line correspondences. In our previous work [5], we proposed a similar method that uses line information to estimate spherical camera pose. However, the extraction of lines inside the spherical image was manual. Moreover, the 6 DoF search-space was too large and a low-accuracy brute-force search was used, which can fall into a local minimum. In this work, we achieve fully automatic, global 6 DoF localization of a spherical camera using a single image via the following novel techniques:

3D线条来自CAD模型

之前的工作中线条是手工提取的，6D位姿估计搜索空间大。

- Automatic and robust detection of lines inside the spherical image is performed via a spherical Hough space. Spherical-gradient of image edge information is used to filter the votes to the line distribution. This makes the line detection and generation of the Hough line distribution robust.
- The complex 6 DoF search for position and orientation is decoupled using a Manhattan world assumption [6]. First, the orientation is estimated by decomposing the Hough line distribution into the three principle directions. Then, the position is estimated by a 3 DoF particle filter-based search that robustly compares the Hough line distributions between the image and the line map. This greatly reduces the search complexity and avoids local minima.
- The length of each line subtended on the image is used to weight it in order to reduce the influence of noise

and give more importance to longer lines that can be detected accurately. This results in highly accurate and robust position estimation.

II. Related Work

There are many image localization methods in literature that work by matching an image and a reference information prepared in advance. These methods can be roughly classified into two categories. The first category consists of methods that save reference images of the environment into a memory matrix and compare them to the image for which localization is desired [7], [8]. Methods using a panoramic images [7] and methods using omnidirectional cameras [8] have also been proposed earlier. However, in these methods, it is necessary to prepare images of environmental scenes in bulk, which is tedious and time-consuming. The second category consists of methods that use environmental models or maps as reference information [9]–[13]. These methods have the advantage of not needing to capture images of the environment in bulk as representations of environmental scenes can be reproduced from environmental models or maps. Ramalingam *et al.* [9] proposed a method for estimating the positions and orientations of omnidirectional camera images using skylines. Ishizuka *et al.* [10] and Cham *et al.* [11] also proposed methods using line information of the known 3D environment model. However, these methods can estimate only 3 DoF pose for a vertically oriented camera and cannot be applied for 6 DoF estimation problems. In contrast, Bleser *et al.* [12] and Ji *et al.* [13] proposed methods for estimating 6 DoF positions and orientations of perspective cameras using line information of a known 3D environment model. However, these methods can be applied to only a normal perspective projection camera and cannot be applied to a spherical camera. Moreover, the method of Ji *et al.* [13] assumed a camera attached to the wall of an indoor environment, considerably reducing the search space. There are methods using information about 3D-2D line correspondences [14]–[16]. However, these methods need line correspondences which are difficult to obtain.

In this paper, a novel method of 6 DoF estimation of any position and orientation of a camera using line information from a single image is proposed. As far as we know, a method whose problem setting is the same as the proposed method does not exist.

III. Overview of Proposed Method

An overview of our method is shown in Fig. 1. Our proposed approach takes as input a pre-constructed 3D line map of the indoor space and localizes a single spherical image clicked anywhere inside it. The line map includes doors, windows, edges, and other features that can be easily generated from an architectural blueprint or a construction model, which is usually available for indoor spaces. The processing consists of three main steps. In the first step, lines are extracted from the image using a randomized Hough transform. The Hough space is designed to be a spherical surface and also forms a 'descriptor' of the line distribution

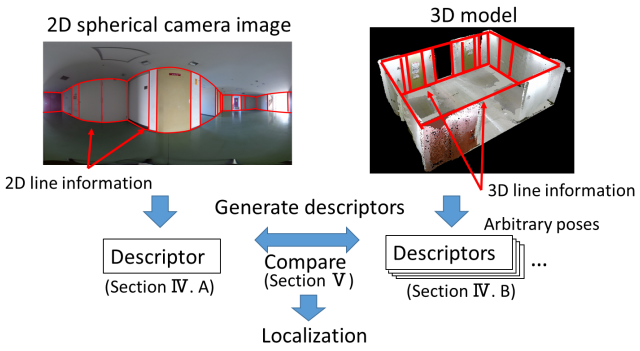


Fig. 1. Approach of the proposed method. We extract lines and generate descriptors from both a spherical image and arbitrary poses in the 3D model. These descriptors are compared to estimate the position and orientation of the spherical camera.

of the image. The same descriptor can be generated from arbitrary positions and orientations in the line map. Each line is weighted by the length that is subtended inside the image. These descriptors can be compared by searching inside the line map to localize the image. However, searching for 6 DoF pose is complicated. In order to simplify the search, orientation and position search are decoupled.

Therefore, in the next step, the descriptor is decomposed via RANdom SAmple Consensus (RANSAC) [17] to find the three principle directions simultaneously, in accordance to the Manhattan world assumption [6]. The three principle directions generated from the image are compared with the three principle directions of the line map in order to estimate the orientation of the image.

In the final step, the position of the image is estimated by robustly matching the descriptor formed from the image to descriptors generated from many arbitrary positions in the line map. The similarity between descriptors is evaluated by Earth Mover's Distance (EMD) which has the advantage of being able to evaluate the similarity of multidimensional distributions, as also used in [18]–[20]. The position at which the computed EMD is the minimum in the environment is the final position estimation result. This position is globally searched via a particle filter.

IV. Line Detection and Descriptor Generation

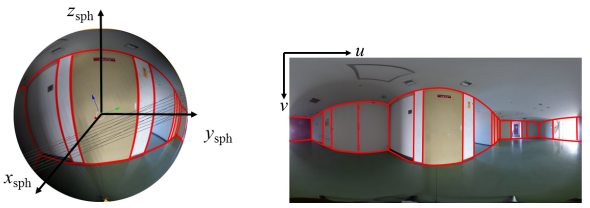
A. Line Detection in Spherical Image

The first step in our approach is to detect line information inside the spherical image. Line information is extracted by a randomized Hough transform. A spherical image is obtained as an equirectangular image as shown in Fig. 2. Lines in the environment are marked in the equirectangular image and the spherical image in red, as shown in Fig. 2.

In previous research, it has been shown that 3D lines in the environment are projected as 'great circles' in the spherical image [21]. Each great circle can be represented by a vector perpendicular to its plane from the center of the sphere, as shown in Fig. 3 (a). The direction of this vector can define a line uniquely. Thus, the Hough space is defined using a spherical surface of unit radius containing unit normal

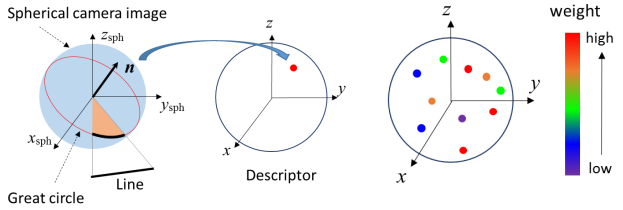
解耦位置和姿态

RANSAC+曼哈顿世界假设



(a) A spherical image (b) An equirectangular image

Fig. 2. 3D lines in the environment are represented as red lines in the equirectangular image.



(a) Generation process (b) Generated descriptor

Fig. 3. Projection of a 3D line on a spherical image and transforming a unit normal vector into the spherical Hough space. In a spherical image, 3D lines in the environment are projected as great circles. $\boldsymbol{n}$ is a unit normal vector with respect to the plane defined by the great circle. It defines a line projected on the spherical image uniquely.

vectors that represent environmental lines. This distribution of unit normal vectors defines the arrangement of lines inside the spherical image and is unique to every position and orientation of the spherical camera inside the environment. Thus, it forms a 'descriptor' of the line information inside a spherical image.

Lines inside the spherical image are detected based on a randomized Hough transform. First, the edges inside the spherical image are generated by a Gaussian blur (to reduce the influence of noise) followed by Canny edge detection [22]. Then, a randomized voting procedure is used to generate votes to the spherical hough space, i.e. the descriptor. In order to uniquely define a line, two edge points are necessary. Thus, two edge points are selected at random from the spherical image and their cross product is taken to find the unit normal vector that defines the line between the two edge points. At the end of the voting process, all the votes close to each other are aggregated and averaged to give the final descriptor.

Thus, the number of votes corresponds to the length of the line inside the image and forms the 'weight' of each line in this descriptor. A schematic of the descriptor generation process is shown in Fig. 3 (b). The color of each point represents the number of votes, *i.e.* the weight.

The above processes are executed on an equirectangular representation of a spherical image. This results in high distortion towards the top and the bottom of the image, expanding the edges near these regions. If edge points are randomly sampled from the equirectangular image, lines towards the top and the bottom gain extra votes. Instead, it

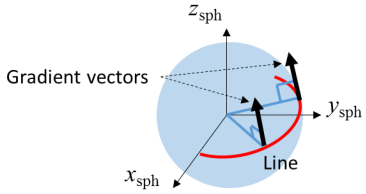


Fig. 4. The 3D directions of the spherical gradient vectors of the points that are on the same line are the same.

is necessary to randomly sample edge points uniformly from a spherical surface. In order to achieve this, we sample the vertical coordinates of the edge points in the equirectangular image from the following distribution, to account for the distortion:

$$v = \frac{r}{\pi} \arccos\left(2p - 1\right), \tag{1}$$

where $0 < p < 1$, r denotes length of vertical axis of the equirectangular image, and v denotes the selected vertical coordinate.

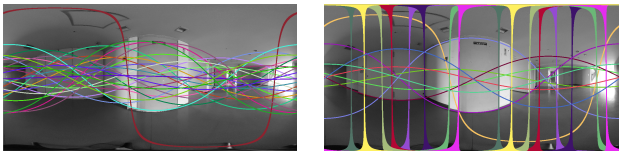
However, a naive sampling of two edge points leads to many false votes, which can reduce the accuracy and robustness of the line detection and descriptor generation. In order to reduce the number of false votes, two conditions are enforced on the randomly sampled edge points in order to make sure that they are selected from the same line. The first is a simple limitation of the distance between the two points. If two points are far from each other, the probability that they are on the same line is low.

The second is a constraint on the direction of the 'spherical gradients' at two edge points. The spherical gradient is estimated by calculating the image gradient in the equirectangular image and projecting it to the 3D spherical image surface. It is tangential to the sphere. Spherical gradient vectors of points on the same line should be oriented in the 3D same direction on a sphere as can be seen in Fig. 4, Thus, the normal vector derived from two edge points is only voted to the spherical Hough space if their spherical gradient directions are the same.

An example of line detection without these constraints is shown in Fig. 5 (a) and the same result with these constraints is shown in Fig. 5 (b). Colored curves indicate the detected lines. In the case without constraints, it can be seen that many lines are detected wrongly. Especially, vertical lines with respect to the image are missed. In comparison, in the case with constraints, the lines are detected accurately. These results show that the two constraints on sampling improve the accuracy of line detection.

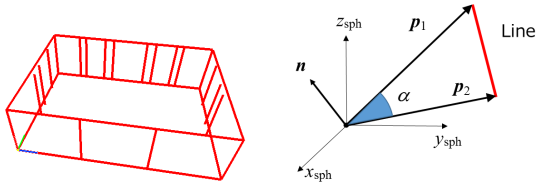
B. Descriptors from Line Map

In order to estimate the position and orientation of the spherical image in the environment, a similar descriptor needs to be generated from the line map and compared with the descriptor generated from the image. Thus, descriptors need to be generated at any 6 DoF position from the line map. An example of an indoor line map is shown in Fig. 6(a). The objective is to extract a descriptor of line information from the line map at a given camera position



(a) Without constraints (b) With constraints

Fig. 5. The result of line detection using the randomized Hough transform without and with the two proposed constraints. The colored curves show detected lines.



(a) An example of line map (b) Descriptor generation

Fig. 6. Schematic of transforming a 3D line in the line map to the spherical Hough space. A cross product of two position vectors $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ of end points of a line derive a normal vector $\boldsymbol{n}$. The angle α is the weight for computing EMD.

and orientation. All calculations are done with respect to this position and orientation. As shown in Fig. 6, the descriptor can be generated by directly transforming each line present in the line map into the spherical Hough space. If $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are the two end points of a line, the unit normal vector $\boldsymbol{n}$ representing it can be derived by a cross product of the position vectors of these two points from the camera position.

Each unit normal vector is weighted with the angle α subtended by its end points. This makes sure that the weight of the line in the descriptor generated from the line map corresponds to the number of votes obtained from the same line if a spherical image was clicked at the same position. The use of these weights raises the accuracy and robustness while comparing descriptors generated from the model and the image, as will be shown later via experiments.

V. Decoupled Position and Orientation Estimation

Once the descriptor is generated from the spherical image, it can be compared to descriptors generated at arbitrary positions and orientations from the line map in order to localize the image. In our previous work [5], this was done as a 6 DoF brute force search. Moreover, the weight system mentioned in the previous section was not used. In such a scenario, the probability of the search leading to a local minimum is quite high and the accuracy is quite low. Global search methods like particle filters [23], [24] can be used, but can suffer from the curse of dimensionality and lead to a prohibitively high search time. Instead, we propose the use of the Manhattan world assumption [6] in order to decouple the search for position and orientation.

According to the Manhattan world assumption [6], most lines in an indoor space lie alone three principle directions.

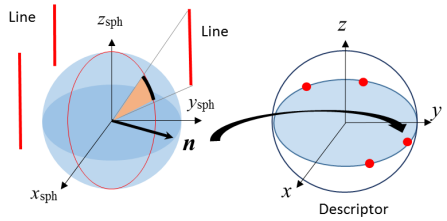


Fig. 7. Lines whose directions are the same are transformed on the same plane in the descriptor. For example, lines along with z axis are transformed on the xy plane in the descriptor.

These principle directions can be estimated from the line information extracted from the image and compared to the principle directions of lines in the line map. Thus, the orientation can be easily estimated, leaving a simpler 3 DoF search for the position.

While constructing the line map, we place the origin in a manner such that the axes lie in the principle directions of the lines. Thus, it is only necessary to estimate the principle directions in the image. This is done via a 3-line RANSAC [17] that simultaneously estimates all three directions, similar to the method followed in [25].

A. Extracting Principle Directions from Spherical Image

In order to extract the three principle directions from the image, we make use of the fact that the unit normal vectors obtained from all lines in the same principle direction lie on the same plane, as shown in Fig. 7. For example, lines which lie along with z axis induce unit normal vectors on the xy plane. Under the Manhattan world assumption [6], there should be three such orthogonal planes. Thus, the descriptor generated from the image should have orthogonal three planes. Extracting these three planes can directly estimate the orientation of the image. This is because they depend on the camera orientation with respect to the line map, which fits a Manhattan world.

The three orthogonal planes are estimated by a RANSAC [17] approach, which is commonly used to estimate parameters of a pre-defined model. In this case, the 'model' consists of three planes represented by their normal vectors. In order to estimate three planes, at least three lines are required. Thus, three unit normal vectors representing their respective lines are randomly sampled from the image descriptor and three planes are generated from them as follows. Vector 1, which is normal to Plane 1 is obtained by a cross product of any two of them. Vector 2, which is normal to Plane 2 is obtained by a cross product of remaining unit vector and Vector 1. Vector 3, which is normal to Plane 3 is computed by a cross product of Vector 1 and Vector 2. This generates three orthogonal planes that represent the principle directions in accordance to the sampled lines.

The next step is to check for the how many other lines in the descriptor satisfy these three planes, i.e., the number of inliers. This is done by estimating whether the unit vector of each line lies in any one of the planes. If the angle from the closest plane is less than a pre-defined threshold, it is counted to be an inlier. While calculating the number of inliers as

well, the weight system described in the previous section is used in order to decrease the influence of noise and small lines. The sampling is repeated a large number of times and the three planes with the largest number of weighted inliers are chosen as the final estimated orthogonal three vector. The three planes estimated by this method are automatically constructed to satisfy the orthogonality of the three principle directions in the same manner as used in [25]. Note that our proposed method is performed under the Manhattan world assumption which assumes orthogonal principal directions. Thus, in the case that an environment does not follow the assumption, RANSAC may not be able to estimate the three principal directions. In addition to RANSAC, a Mixture of Gaussians approach [26] can be used to consider noise in line detection process, which we will consider for future work.

Once the three planes have been estimated, the three principle directions of the image are known. The three principle directions in the line map are set in advance to correspond to the axes of the origin. However, correspondence of the three directions in the image to the three axes needs to be known in order to uniquely estimate the image orientation. This problem is simplified by the assumption that most lines in the image are vertical, as can be seen from Fig. 6(a). Hence, the direction with the highest number of lines corresponds to the vertical direction. The ambiguity of the other two directions (and their negatives) is resolved using the descriptor generated in the previous section and the same distance metric that is used for position estimation, as described in the next section. If the environment does not follow this assumption, e.g., in case of stairs, it is also possible to estimate the correspondence of the three directions at the same time by using the distance metric. Once the three axes in the image are uniquely known, the orientation of the image is uniquely estimated, similar to the method in [25].

B. Position Estimation

After estimating the orientation of a camera, the position can be estimated as a 3 DoF search by evaluating the similarity of the descriptors generated from the spherical image and from the line map. In this method, EMD is used as an evaluation function to compute the similarity. In order to globally search for the position of highest similarity, a particle filter is applied.

1) Similarity Evaluation of Descriptors: In the proposed approach, no correspondence of lines between the image and the line map is estimated. Moreover, the detected lines may have noise, mistakes, etc. Therefore, the evaluation function needs to be robust. Hence, we adopt EMD as a metric that can compute the similarity of the descriptors. EMD is a measure of the distance between two multi-dimensional distributions. It requires no correspondence and it measures the 'amount of work needed to convert one distribution into another'. A simple example with a one-dimensional distribution is shown in Fig. 8. Unlike the L_2 norm, it can take partial matches into account in a natural way. For example, if two distributions are slightly displaced from each other, the L_2 norm will result in a high error. However, the EMD between them will remain small. In other words, it can manage the noise and mistakes present in line detection.

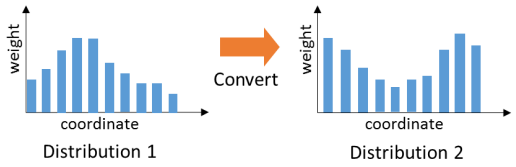


Fig. 8. Schematic of EMD computation. EMD measures the amount of minimum work needed to convert distribution 1 to distribution 2.

In our case, the descriptors can be treated as spherical distributions in order to evaluate their similarity using EMD. The descriptor obtained from the spherical image is treated as distribution 1 in Fig. 8 and descriptors obtained from arbitrary positions in the line map are treated as distribution 2 in Fig. 8. In order to compute EMD between them, each descriptor is converted to a set of clusters $\boldsymbol{Q}$:

$$\boldsymbol{Q}^{(1)} = \left[(\boldsymbol{n}_i^{(1)}, w_i^{(1)}) \mid 1 \leqq i \leqq N^{(1)}\right], \tag{2}$$

$$\boldsymbol{Q}^{(2)} = \left[(\boldsymbol{n}_j^{(2)}, w_j^{(2)}) \mid 1 \leqq j \leqq N^{(2)}\right], \tag{3}$$

where $\boldsymbol{n}$ and w denote the unit normal vectors in the descriptor and the weights that belongs to the clusters, respectively. As explained earlier, the weight w for the descriptor from the spherical image is the number of votes in the randomized Hough transform, and for the descriptor from the line map, the weight w is the angle between $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ as seen from the camera center. This weighing scheme is effective because lines closer to the image are projected to be longer and are more important since they can be detected with greater accuracy. The size of each cluster N is equal to number of the unit normal vectors transformed into the spherical Hough space. EMD is defined as follows:

$$EMD(\boldsymbol{Q}^{(1)}, \boldsymbol{Q}^{(2)}) = \frac{\Sigma_{i=1}^{N^{(1)}} \Sigma_{j=1}^{N^{(2)}} f_{ij} d_{ij}}{\Sigma_{i=1}^{N^{(1)}} \Sigma_{j=1}^{N^{(2)}} f_{ij}}, \tag{4}$$

where d_{ij} denotes a user-defined ground distance. The distance d_{ij} is the angle between two unit normal vectors computed as follows:

$$d_{ij} = \arccos(\boldsymbol{n}_i^{(1)} \cdot \boldsymbol{n}_j^{(2)}). \tag{5}$$

The variable f_{ij} denotes a 'flow element' that is derived by solving the transportation problem using the weight w. Additional details on EMD can be found in [18].

The minimum value of EMD is 0 and its value becomes larger as similarity of the descriptors is lower. The position of a camera is estimated as the pose at which the EMD between the descriptor obtained from the spherical image becomes the minimum with respect to the descriptor obtained from a particular position in the line map. This position is searched for using a particle filter [23], [24]. Particles are generated at the orientation extracted in the previous step at many random positions inside the line map, and propagated. The final position is estimated by calculating a weighted sum of all particles after convergence.

VI. Experiment

Three kind of experiments were conducted to demonstrate the performance of our method.

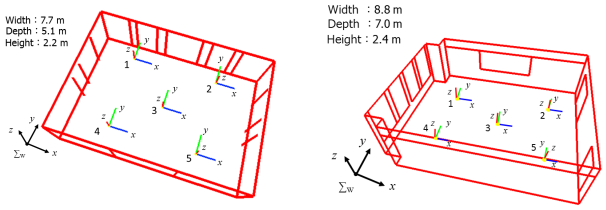
In experiment 1, spherical images captured at five different positions in two indoor environments were localized in order to evaluate robustness and accuracy. In addition, many spherical images generated at randomly selected poses in a simulation environment were localized.

In experiment 2, a single spherical image from experiment 1 was rotated 10 deg. at a time from -30 deg. to 30 deg. around the x axis and the position and orientation were estimated at each step. Rotation of spherical images causes large, non-linear changes in image content, making it difficult to detect the same lines in each case. The purpose of this experiment was to evaluate robustness for different camera orientations.

In experiment 3, in order to compare with our previous work [5], the results of our proposed method were compared with results obtained using a non-weighted descriptor, and results obtained from a 6 DoF brute-force search in a coarse-to-fine scheme. 7 different trials were performed. For the brute-force, coarse-to-fine search, the initial point was set randomly in each trial. The minimum step size of the search was 0.1 m for the position, and 1 deg in each axis for the orientation. For the particle filter-based searches (with and without the weights), 10,000 particles were set randomly in each trial. The purpose of this experiment was to confirm the robustness of global localization when using the proposed weighing scheme and the particle filter-based search. In each case, if the estimation error was within 0.3 m for the position and within 5 deg in each axis for the orientation, the estimation was defined to be successful. The success rate and maximum error in all trial were evaluated.

A. Experimental Setup

The Ricoh Theta S spherical camera was used to obtain spherical images for experiments. It directly captures an equirectangular image and requires no external calibration. Equirectangular images of resolution 1200×600 pixels were used. In this experiment, we assume that the intrinsic calibration and spherical image generation implemented by the camera manufacturer is accurate enough to generate a spherical image. A section of a building corridor and a conference room were used as experimental environments. A 3D environment models of the experimental environments were generated by Simultaneous Localization and Mapping (SLAM) using a mobile robot equipped with RGB-D sensor in advance [13]. The line maps of the experimental environments are shown in Figs. 9(a) and 9(b) were generated by manually selecting the coordinates of the end points of each line. They can also be obtained from an architectural blueprints and are quite easy to generate. In experiment 1, spherical images were captured at five points in both environments at the same orientations. Spherical images captured at point 3 in both environments are shown in Fig. 10. In addition, a simulation environment representing an elevator hall is shown in Fig. 11(a). The camera poses were localized



(a) The section of the corridor (b) The conference room

Fig. 9. The line maps of the experimental environments. Spherical images were taken at points 1 to 5.

(a) Spherical image of the section of the corridor (b) Spherical image of the conference room

Fig. 10. Spherical images captured at point 3 in both environments.

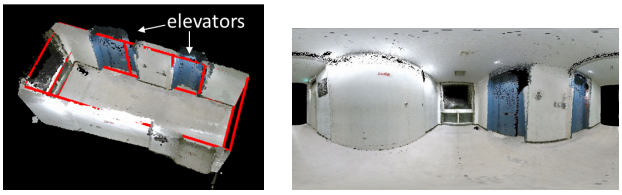
using spherical images generated at randomly selected points and compared to the groundtruth. An example of generated spherical image is shown in Fig. 11(b). Localization was done 100 times in the simulation environment. Lines and principle directions were extracted in each image. The threshold for RANSAC-based principle direction extraction described in Section V-A was $\sin^{-1}(0.05)$ radians.

B. Result and Discussion

An example of the intermediate result of line detection and principle direction extraction is shown in Fig. 12. The colorful lines denote detected lines and the pairs of red, blue, and green points denote the three principle directions (i.e. the vanishing points). The camera orientation was estimated using these principle directions and the camera position was searched for using a particle filter.

1) Experiment 1: The results of the experiments in the section of the corridor and the conference room are shown in Table I and II respectively. The values show the errors of the estimated positions and orientations. The maximum error obtained was 0.26 m for position and 3.6 deg for orientation in one axis. If the model is very big, detected lines may be shorter. Consequently, inaccuracy in line detection leads to larger errors of position and orientation.In the simulation environment, successful cases were defined as those having maximum errors within 0.3 m for translation and 5 degrees for rotation. Localization succeeded 95 out of 100 times. Our proposed method succeeded in accurately estimating the position and orientation of a spherical camera at different positions.

It can be noticed that the estimation errors at point 3 were slightly larger than those at other points. This is because the camera position at point 3 was more distant from every wall in the environment as compared to other points.



(a) The 3D model of the elevator hall with line map

(b) An example of a generated spherical image

Fig. 11. The simulation environment and a spherical image.

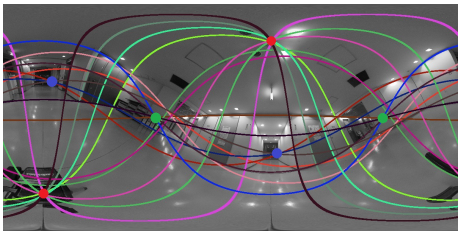


Fig. 12. An example of the intermediate result of line detection and principle direction extraction. The colored lines denote detected lines and the pairs of red, blue, and green points denote the three principle directions.

Consequently, the detected lines were shorter, making line detection inaccurate. This also corroborates the importance of using the length of the line subtended as a weight for the estimation (revisited in experiment 3).

Regarding the processing time, it totally takes about 30 minutes to estimate position and orientation of a single image. The processing time changes depending on the number of edge points or the size of the experimental environment.

In particular, as is often the case with dynamic environment, occlusions can be present in the image. In order to evaluate such situations, experiments were conducted using images occluded by obstacles. The camera pose was estimated using an image with a person who was standing at about 1.5 m from the camera as shown in Fig. 13(a). The maximum errors were 0.1 m for translation and 2 degrees for rotation. The localization succeeded with small errors. In addition, we estimated the camera pose using an image with circles as occlusions, for example, as shown in Fig. 13(b). The circles were positioned at random and the size of each circle is 2 % of the area in the image. The number of circles was set as 2, 4, 6, 8 and 10, and five trials were conducted for each case. If the maximum errors were within 0.3 m for translation and 5 degrees for rotation in any axis, the trial was considered successful. The results are shown in Table III.

(a) Image with a person as an occlusion

(b) An example image with circles as occlusions

Fig. 13. Images with a person and circles as occlusions.

TABLE I

ERRORS OF ESTIMATION IN THE SECTION OF THE CORRIDOR

Pose	x [m]	y [m]	z [m]	ϕ [deg]	θ [deg]	ψ [deg]
1	0.02	0.06	0.01	0.3	1.4	1.3
2	0.04	0.06	0.01	0.4	1.1	0.7
3	0.24	0.07	0.09	2.9	0.2	3.0
4	0.06	0.01	0.01	0.9	1.7	2.1
5	0.09	0.02	0.05	0.2	0.9	1.3

TABLE II

ERRORS OF ESTIMATION IN THE CONFERENCE ROOM

Pose	x [m]	y [m]	z [m]	ϕ [deg]	θ [deg]	ψ [deg]
1	0.02	0.08	0.05	0.6	2.1	1.7
2	0.06	0.07	0.03	3.6	3.2	0.8
3	0.26	0.2	0.13	0.3	0.2	0.0
4	0.05	0.05	0.06	1.2	2.3	0.6
5	0.11	0.05	0.04	0.8	0.7	1.4

These results show the robustness of the proposed method to occlusions.

2) Experiment 2: The results of experiment 2 are shown in Table IV. The values show the errors of the estimated positions and orientations. The maximum error obtained was 0.12 m for position and 2.5 deg for orientation in one axis. These results indicate that estimation accuracy does not depend on the camera orientation and is stable. This is possible due to the sampling scheme adopted during the randomized Hough voting in subsection IV-A. The errors were stable for each orientation and thus, the robustness for change of the orientation can be confirmed.

3) Experiment 3: The results of experiment 3 are shown in Table V. They show the success rate and maximum error of all the trials. The success rate of orientation estimation using the descriptor with no weights was very low. In the position search using particle filter, the orientation of the particles was set as the ground truth. Even then, the particles did not converge and we could not get the final result. This shows that it is essential to weigh each line corresponding to its length inside the image.

Meanwhile, the success rate of the brute-force, coarse-to-fine search for 6 DoF localization was very low. This is because the search can easily fall into a local minima depending on the starting point. Meanwhile, in our proposed method, the position and orientation estimation is decoupled and the position is searched using a particle filter, which can globally estimate the position. Falling into a local minima can be avoided and the position and orientation can be robustly estimated.

VII. CONCLUSION

In this research, a novel method for 6 DoF localization of a spherical camera within a known 3D model i.e. line map of an indoor environment was proposed. A novel descriptor was designed based on a spherical Hough space for representation of line information from both a 2D spherical image and a 3D line map without requiring information about the 3D-2D line correspondences. The length of a line subtended inside a spherical image was used as a weight. The orientation of a spherical camera was estimated under a Manhattan world

TABLE III

NUMBER OF SUCCESSES USING IMAGES WITH OCCLUDING CIRCLES

Number of circles	2	4	6	8	10
Number of successes	5	4	4	4	3

TABLE IV

ERRORS OF ESTIMATION IN EXPERIMENT 2

Rotation	x [m]	y [m]	z [m]	ϕ [deg]	θ [deg]	ψ [deg]
-30 deg	0.07	0.04	0.01	1.0	1.2	1.8
-20 deg	0.04	0.04	0.04	0.9	1.1	0.3
-10 deg	0.03	0.02	0.03	0.5	0.6	0.1
10 deg	0.07	0.04	0.05	0.0	0.7	1.8
20 deg	0.12	0.07	0.01	0.9	0.3	2.0
30 deg	0.03	0.06	0.00	1.5	2.5	2.3

assumption. Then, in order to search for the position, EMD was used to effectively and robustly compute the similarity between these descriptors and a particle filter-based search was used. This decoupled 6 DoF estimation was able to avoid falling a local minima effectively. Experiments were conducted in a real environment and the results demonstrated that our proposed method could effectively estimate the 6 DoF pose of a spherical camera up to about 0.26 m and 3.6 deg in each axis within a 3D model using a single spherical image. This research demonstrated that the weighting of each line is crucial for an adequate matching. Hence, in order to estimate the camera pose more robustly, we will develop a weighting scheme using surrounding color information. In this study, we focused on an indoor environment. Future work will also extend the developed method to outdoor environments.

ACKNOWLEDGEMENT

We wish to thank Dr. Angela Faragasso for her cooperation in improving this paper.

This work was in part supported by the Council for Science, Technology and Innovation, "Cross-ministerial Strategic Innovation Promotion Program (SIP), Infrastructure Maintenance, Renovation, and Management" (funding agency: NEDO).

REFERENCES

[1] E. Olson, "Apriltag: A robust and flexible visual fiducial system," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2011, pp. 3400–3407.

[2] R. Miyagusuku, A. Yamashita, and H. Asama, "Improving gaussian processes based mapping of wireless signals using path loss models," in *Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 4610–4615.

[3] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Proceedings of the IEEE/ACM International Symposium on Mixed and Augmented Reality*, November 2007, pp. 225–234.

[4] R. Yoshimura, H. Date, S. Kanai, R. Honma, K. Oda, and T. Ikeda, "Automatic registration of mls point clouds and sfm meshes of urban area," *Geo-spatial Information Science*, vol. 19, no. 3, pp. 171–181, 2016.

[5] T. Goto, S. Pathak, Y. Ji, H. Fujii, A. Yamashita, and H. Asama, "Spherical camera localization in man-made environment using 3d-2d matching of line information," in *Proceedings of the International Workshop on Advanced Image Technology*, 2017.

[6] J. M. Coughlan and A. L. Yuille, "The manhattan world assumption: Regularities in scene statistics which enable bayesian inference," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, 2000, pp. 809–815.

TABLE V

RESULTS OF EXPERIMENT 3

	No weight		C-t-F search		Proposed	
	Pos.	Ori.	Pos.	Ori.	Pos.	Ori.
Success rate	0 %	14.3 %	57.1%	57.1 %	100 %	100 %
Max error	-	44.9 deg	2.50 m	180 deg	0.14 m	3.5 deg

[7] A. Torii, Y. Dong, M. Okutomi, J. Sivic, and T. Pajdla:, "Efficient localization of panoramic images using tiled image descriptors," *Information and Media Technologies*, vol. 9, no. 3, pp. 351–355, 2014.

[8] N. Aihara, H. Iwasa, N. Yokoya, and H. Takemura:, "Memory-based self-localization using omnidirectional images," in *Proceedings of the 14th IEEE International Conference on Pattern Recognition*, 1998, pp. 1799–1803.

[9] S. Ramalingam, S. Bouaziz, P. Sturm, and M. Brand:, "Geolocalization using skylines from omni-image," in *Proceeding of the 2011 IEEE International Conference on Computer Vision Workshop*, 2011, pp. 23–30.

[10] D. Ishizuka, A. Yamashita, R. Kawanishi, T. Kaneko, and H. Asama:, "Proceedings of the 2011 ieee international conference on computer vision workshop," 2011, pp. 272–279.

[11] T. Cham, A. Ciptadi, W. Tan, M. Pham, and L. Chia:, "Estimating camera pose from a single urban ground-view omnidirectional image and a 2d building outline map," in *Proceedings of the 2010 IEEE International Conference on Computer Vision and Pattern Recognition*, 2010, pp. 366–373.

[12] G. Bleser, H. Wuest, and D. Stricker:"Online, "Camera pose estimation in partially known and dynamic scenes"," in *Proceedings of the 2006 IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2006, pp. 56–65.

[13] Y. Ji, A. Yamashita, and H. Asama:, "Automatic calibration of camera sensor networks based on 3d texture map information," *Robotics and Autonomous Systems*, vol. 87, pp. 313–328, 2017.

[14] C. Xu, L. Zhang, L. Cheng, and R. Koch, "Pose estimation from line correspondences: A complete analysis and a series of solutions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1209–1222, 2017.

[15] F. M. Mirzaei and S. I. Roumeliotis, "Globally optimal pose estimation from line correspondences," in *proceedings of the 2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 5581–5588.

[16] A. Vakhitov, J. Funke, and F. Moreno-Noguer, "Accurate and linear time pose estimation from points and lines," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 583–599.

[17] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[18] Y. Rubner, C. Tomasi, and L. J. Guibas:, "A metric for distributions with application to image databases," in *Proceedings of the 1998 IEEE International Conference on Computer Vision*, 1998, pp. 59–66.

[19] K. Grauman and T. Darrell, "Fast contour matching using approximate earth mover's distance," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 1, 2004, pp. I–I.

[20] J. Rabin, J. Delon, and Y. Gousseau, "Circular earth mover's distance for the comparison of local features," in *proceedings of the 19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[21] A. Torii and A. Imiya, "The randomized-hough-transform-based method for great-circle detection on sphere," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1186–1192, 2007.

[22] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 679–698, 1986.

[23] N. Gordon, B. Ristic, and S. Arulampalam, "Beyond the kalman filter: Particle filters for tracking applications," *Artech House*, 2004.

[24] A. Smith, A. Doucet, N. de Freitas, and N. Gordon, "Sequential monte carlo methods in practice," *Springer Science & Business Media*, 2013.

[25] J. C. Bazin and M. Pollefeys, "3-line ransac for orthogonal vanishing point detection," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 4282–4287.

[26] A. Minagawa, N. Tagawa, T. Moriya, and T. Gotoh, "Vanishing point and vanishing line estimation with line clustering," *IEICE TRANSACTIONS on Information and Systems*, vol. 83, no. 7, pp. 1574–1582, 2000.