Spherical Camera Localization in Man-made Environment Using 3D-2D Matching of Line Information

Tsubasa Goto, Sarthak Pathak, Yonghoon Ji, Hiromitsu Fujii, Atsushi Yamashita, and Hajime Asama
Department of Precision Engineering
The University of Tokyo
Tokyo, Japan
Email: {goto, pathak, ji, fujii, yamashita, asama}@robot.t.u-tokyo.ac.jp

***Abstract*—In this paper, we propose a novel method for global six degree of freedoms (DoF) localization of a spherical camera in a man-made environment. Specifically, a 3D-2D matching method based on line information of a known 3D model of the environment is proposed. There are two challenging points. First is to design a unique representation of 2D line information from the image and 3D line information from the 3D environment model. Second is to evaluate similarity of the line information extracted from both a real spherical camera image taken in the environment and arbitrary 6 DoF poses in the 3D environment model in order to localize the camera. To deal with the former, a novel descriptor is designed based on a Hough space for the line information. Then, earth mover's distance (EMD) is calculated to evaluate similarity between the descriptors. We evaluated the proposed method in a real environment with its 3D model. The results demonstrated that our proposed method can effectively estimate the 6 DoF pose of a spherical camera using a single image.**

I. Introduction

Usage of mobile robots or drones has been expanding recently. For example, large infrastructures such as bridges or tunnels often need to be inspected by drones. For these uses, self-localization of a robot for inspection or monitoring of man-made structures is an important task. With regard to self-localization of robots, a Global Positioning System (GPS) can often be used [1]. However, use of GPS is difficult in places where signals are weak, e.g. under bridges, tunnels, or indoor environments. Moreover, drones cannot carry sophisticated sensors due to their loading limitations. Thus, cameras, which are lightweight and inexpensive, are a good choice. Especially, a spherical camera, which can capture in all directions, is much more effective at self-localization as compared to a perspective camera. This is because the information obtained by the perspective camera is not enough due to its limited field of view, particularly in situations where the robots moves close to a wall or an obstacle, facing it.

Approaches based on feature points extracted from an image such as visual SLAM [2] and Parallel Tracking and Mapping (PTAM) [3] can be effective to estimate the position and orientation. However, since it is necessary to track the feature points in these approach, accumulation of errors becomes problem in case of images accompanied by moving long distance. Instead, an approach that estimates the self position and orientation using the information from the whole environment can be more effective.

In this work, we focus on 3D environment models such as CAD models for construction or 3D models built by laser range finder. Since there are cases in which it is difficult to include accurate color information in the model, we propose the method to just use shape information. Especially for man-made environments, 3D line information is available. Therefore, our proposed method estimates position and orientation of a spherical camera by matching the line feature distributions from both arbitrary poses in the 3D environment model and one real spherical camera image.

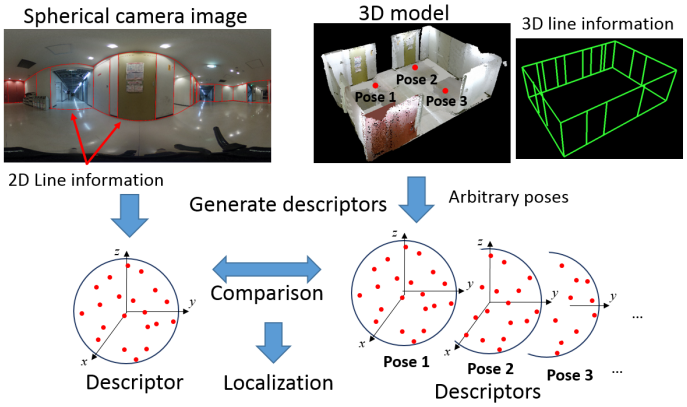

Fig. 1. Approach of the proposed method. We extract line information by generating descriptors from both a spherical camera image and arbitrary poses in the 3D model. The descriptors are compared to estimate the position and orientation of the spherical camera.

Thus, the objective of this research is 6 DoF localization by 3D-2D matching using the known 3D environment model and a single spherical image.

II. Related Work

A number of approaches have been developed in 3D-2D matching for the position and orientation estimation for mobile robots or computer vision applications. Ramalingam *et al.* [4] proposed a method for estimating the positions and orientations of omnidirectional camera images using skylines. Ishizuka *et al.* [5] and Cham *et al.* [6] also proposed methods using the line information of the known 3D environment model. However, these methods can estimate only 3 DoF pose on a plane for a camera in a vertical orientation and cannot be applied to complete 6 DoF estimation problems. For 6 DoF localization, Ji *et al.* [7] and Bleser *et al.* [8] proposed methods for estimating the position and orientation of a perspective camera using the line information of a known 3D environment model. However, these methods can be applied to only a normal perspective projection camera and distortion of lines must be considered for a spherical camera.

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

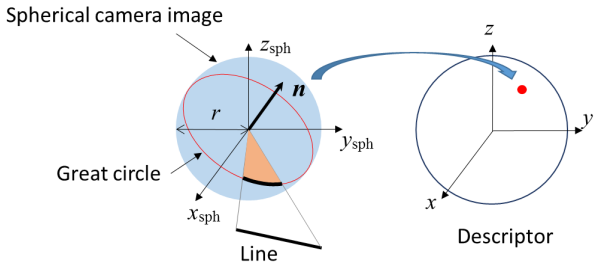

Fig. 3. Schematic of projection of a line and transforming a unit normal vector into spherical Hough space. In a spherical image, 3D lines in the environment are projected as great circles. $\boldsymbol{n}$ is a unit normal vector with respect to the plane defined by the great circle.

In this paper, a novel method of 6 DoF estimation of the position and orientation using the line information from a spherical image is proposed.

III. Proposed Method

A. Overview

Figure 1 shows the approach of our method. Line information is extracted in order to generate descriptors from both a spherical camera image and arbitrary poses in the 3D model. The descriptor in this method represents the distributions of 3D straight lines in the scene of the environment. Lines in the environment are projected as great circles on the spherical image. A great circle is a circle on the surface of a sphere that passes through its center and divides the sphere into two halves. These lines can be defined by the normal vectors defined by these circles. Consequently, the distribution of these normal vectors generate the descriptors. The descriptors correspond to the positions and orientations of the view points in the environment. Therefore, it is possible to estimate position and orientation by evaluating similarity of the descriptors from both a spherical camera image and arbitrary poses in the 3D model. The similarity of them is evaluated by earth mover's distance (EMD) [9] which has an advantage of being able to evaluate similarity of multidimensional distributions. The pose where computed EMD is the minimum in the environment is the final estimation result.

B. Generation of Descriptors

The camera pose is estimated by comparing the descriptors from both a 2D spherical image and the 3D environmental model as mentioned above. The descriptors of the line information from 2D spherical image can be obtained from the lines within the image. Meanwhile, the descriptors from arbitrary poses in the 3D model can be obtained directly from the line information contained in the model. Since these descriptors depend on the position and orientation of a camera, they can describe not only the line information but also the information of the position and orientation of the camera.

1) A descriptor from a spherical image: The line information is extracted by a randomized Hough transform to generate descriptor. Lines in the environment are represented as red lines in an equirectangular image as shown in Fig. 2. In a spherical image, 3D lines in the environment are projected as great circles as shown in Fig. 3. A vector $\boldsymbol{n}$ denotes a unit normal vector with respect to the plane defined by the great circle. This great circle can define a line uniquely. Accordingly, the Hough space is defined as a unit sphere that takes unit normal vectors as parameters that define environmental line information. The line information is extracted by transforming the unit normal vector into this spherical Hough space. This expression of distribution of lines on the spherical Hough space can describe the line information and the pose of a spherical camera, i.e. the spherical Hough space is the descriptor from a spherical image.

In order to obtain these unit vectors, line detection in the spherical image is necessary. The lines are detected based on a randomized Hough transform. For Hough transform, edge detection is necessary. First, a real spherical image is blurred by a Gaussian operator for reduction of influences of noise. Next, the edge image is generated from the blurred image using the Canny edge detection operator. Here, we refer to points on the edge as edge points. Randomized Hough transform implements a voting procedure repeatedly for a certain number of times. The voting procedure is as follows:

1) Select two edge points from the edge image randomly (because two points can completely define a line).
2) Derive a unit normal vector by a cross product of position vectors of the selected points.
3) Update the spherical Hough space with voting the derived unit normal vector.

If a newly voted point is close to previously voted ones, they are averaged and a new point is created. In order to make sure that the two points selected are from the same line, a constraint is applied to limit the distance between the two points. In this way, the spherical Hough space voted repeatedly is the descriptor from a real spherical image.

2) Descriptors from 3D environmental model: To extract line information from the 3D environmental model, a line map that contains only line information of the 3D environment model is prepared. For example, the lines of Fig. 4(a) are drawn as Fig. 4(b). In order to extract the descriptor of line information at an arbitrary camera pose, the camera pose is set as the origin and orientation of the 3D line map. All calculations are done with respect to the camera pose. In the same way as the descriptor from a spherical image, the lines in the line map are transformed directly into the spherical Hough space. A unit normal vector $\boldsymbol{n}$ is derived by a cross product of two position vectors $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ of the start and end points of the line. These unit normal vectors are inserted into a spherical hough space which forms the descriptor from a 3D

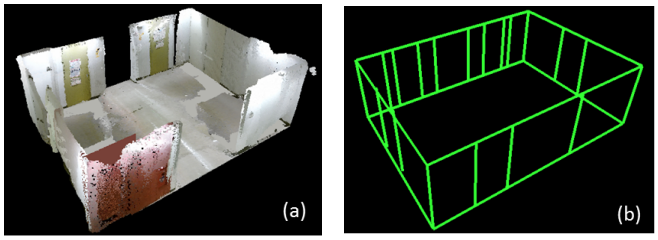

Fig. 4. 3D environment model and line map: (a) 3D environment model, (b) line map in which only lines are drawn

environmental model with respect to the arbitrarily given pose. Thus, the descriptor for any arbitrary camera pose in the 3D model can be generated.

C. Similarity Evaluation of the Descriptors

The similarity of the descriptors mentioned above is calculated as below. For line based matching, it is often the case that lines are not detected perfectly. Wrong lines can be detected and lines can be missed or slightly displaced. Therefore, the evaluation function should be robust. Then, we adopt earth mover's distance (EMD) as the evaluation function which computes the similarity of the descriptors. EMD is a measure of the distance between two multi-dimensional distributions. It measures the amount of work needed to convert one distribution into another. Due to this, 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 line detection errors. That is why EMD is qualified to compute the similarity of the descriptors.

In our case, the descriptors can be treated as spherical distributions in order to evaluate their similarity using EMD. In order to compute EMD between two descriptors, each descriptor is convert to a set of clusters $\boldsymbol{Q}$.

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

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

where $\boldsymbol{q}$ and w denote the coordinates of the descriptor and the weight values that belongs to that cluster, respectively. The weight w for the descriptor from a spherical image is the number of votes in the randomized Hough transform. As for the descriptor from a 3D model, the weight w is the angle between $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ as seen from the camera center. This is because lines that are closer to the image are projected as longer and are more important since they are easy to be detected from a real spherical image. The size of cluster N is equal to number of the unit normal vectors transformed into the spherical Hough space. EMD is defined as follows:

$$EMD(\boldsymbol{Q}^{(1)}, \boldsymbol{Q}^{(2)}) = \frac{\Sigma_{i=1}^{N^{(1)}} \Sigma_{j=1}^{N^{(2)}} f_{ij} d_{ij}}{\Sigma_{i=1}^{N^{(1)}} \Sigma_{j=1}^{N^{(2)}} f_{ij}}, \tag{3}$$

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

$$d_{ij} = \arccos(\boldsymbol{n}_i^{(1)} \cdot \boldsymbol{n}_j^{(2)}), \tag{4}$$

where $\boldsymbol{n}^{(1)}$ and $\boldsymbol{n}^{(2)}$ denote the unit normal vector of the descriptors from a real spherical image and 3D environmental model, respectively. 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 [9].

The minimum value of EMD is 0 and value of EMD becomes larger as similarity of the descriptors is lower. The position and orientation of the camera is estimated as that pose whose EMD from the descriptor obtained by the spherical image becomes the minimum.

Fig. 5. Experimental setup showing the RICOH THETA S spherical camera placed in the environment.

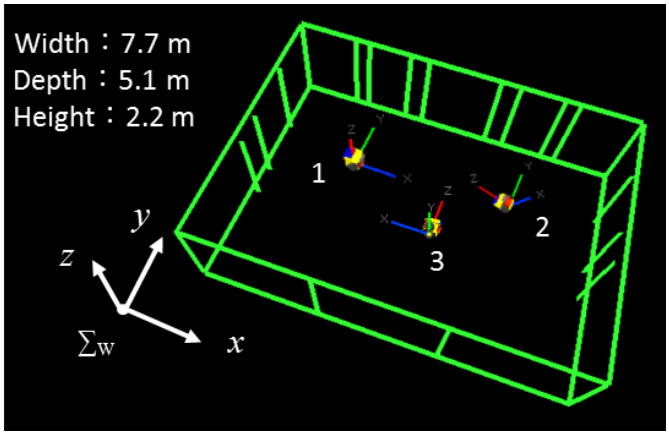

Fig. 6. The line map of the experimental environment. Spherical images were taken at points 1 to 3 at different orientations as shown in Table I.

TABLE I. THE POSITIONS AND ORIENTATIONS OF THE POINTS 1 TO 3.

Pose	x [m]	y [m]	z [m]	ϕ [deg]	θ [deg]	ψ [deg]
1	2.0	4.1	0.9	0	0	0
2	5.6	4.1	0.9	0	-45	0
3	3.8	2.8	0.9	90	0	180

IV. EXPERIMENT

A. Experimental Setup

A spherical camera used for this experiment was RICOH THETA S shown in Fig. 5. The 3D environment model of a corridor was prepared in advance. Figure 6 shows the line map of the experimental environment. Spherical images were taken at points 1 to 3 with respective orientations. The position and orientation of each point are shown in Table I. This line map was created by manually determined start points and end points of the lines from the 3D environment map. The descriptors from the 3D environmental model with respect to

arbitrary poses were generated by using this line map. The descriptors from the captured spherical images were generated by manually detecting the lines in each image.

In order to estimate the position and orientation, a full coarse-to-fine search in the environment was conducted as a 3 step approach. In the first step, EMD was calculated at every 0.5 m in the direction of each axis and every 45 deg rotation around each axis and the pose whose EMD was at the minimum was decided. In the second step, with reference to the result of the first step, EMD was calculated at every 0.1 m in the direction of each axis within $\pm$0.5 m and every 15 deg rotation around each axis within $\pm$45 deg. In the third step, EMD was calculated at every 1 deg rotation around each axis with reference to the result of second step.

The pose whose EMD was minimum in the third step was regarded as the final estimation result. In short, estimation of the position and orientation was performed with a least count of 0.1 m and 1 deg.

B. Experimental Result

The estimation errors for the 3 poses are shown in Table II. Our proposed method succeeded in estimating the position and orientation of the spherical camera up to 0.3 m and 1 deg for each axis.

TABLE II. THE ERRORS OF THE ESTIMATION RESULT.

Pose	x [m]	y [m]	z [m]	ϕ [deg]	θ [deg]	ψ [deg]
1	0.1	0.0	0.0	1	0	1
2	0.0	0.0	0.0	1	1	1
3	0.3	0.1	0.1	1	1	0

It is noted that the estimation errors at the point 3 are slightly larger than those of the point 1 and 2. This is because point 3 is far from every wall in the environment as compared to points 1 and 2. In the computation of EMD, the weight w is designed to become larger as the lines come closer to the camera because they can be easily detected. In the case of point 3, the camera position is almost at the center of the experimental environment, far away from all lines as Fig. 7, thus making line detection difficult. Moreover, the effect of noise is also enhanced in such cases. Nevertheless, the accuracy of this result is adequate. Thus, the effectiveness of our method is confirmed.

V. CONCLUSION

A novel method for 6 DoF localization of a spherical camera within a known 3D model of a man-made environment was proposed in this study. A new descriptor was designed based on a spherical Hough space for representation of line information from both a 2D spherical image and a 3D environmental model. EMD was used to effectively and robustly compute the similarity between these descriptors. Finally, an experiment was conducted in a real environment and the results demonstrate that our proposed method can effectively estimate the 6 DoF pose of a spherical camera up to 0.3 m and 1 deg in each axis within a 3D model using a single spherical image.

As future work, we plan to improve the automatic line detection from a real spherical image to make it robust and accurate as well as to develop a method to find the pose without resorting to a full search.

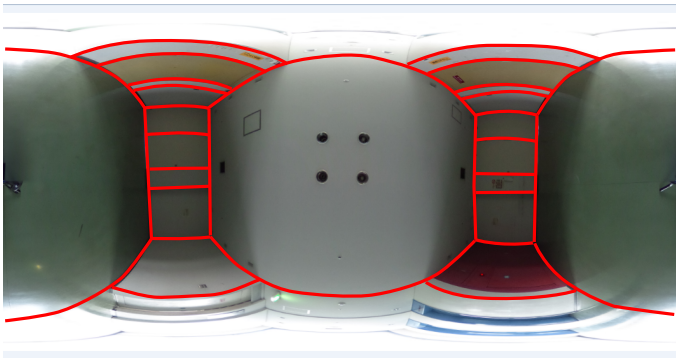

Fig. 7. The spherical image taken at the point 3. Its position is far from every wall and there are no lines that are very close to the camera.

ACKNOWLEDGEMENTS

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] S. Kim, C. Roh, S. Kang, and M. Park: "Outdoor Navigation of a Mobile Robot Using Differential GPS and Curb Detection", *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, pp. 3414–3419, 2007.

[2] D. Caruso, J. Engel, and D. Cremers: "Large-Scale Direct SLAM for Omnidirectional Cameras", *Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 141–148, 2015.

[3] G. Klein and D. Murray: "Parallel Tracking and Mapping for Small AR Workspaces", *Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 225–234, 2007.

[4] S. Ramalingam, S. Bouaziz, P. Sturm, and M. Brand: "Geolocalization Using Skylines from Omni-Image", *Proceeding of the 2011 IEEE International Conference on Computer Vision Workshop*, pp. 23–30, 2011.

[5] D. Ishizuka, A. Yamashita, R. Kawanishi, T. Kaneko, and H. Asama: "Self-localizaion of Mobile Robot Equipped with Omnidirectional Camera Using Image Matching and 3D-2D Edge Matching", *Proceedings of the 2011 IEEE International Conference on Computer Vision Workshop*, pp. 272–279, 2011.

[6] 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", *Proceedings of the 2010 IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 366–373, 2010.

[7] Y. Ji, A. Yamashita, and H. Asama: "Automatic Calibration of Camera Sensor Networks Based on 3D Texture Map Information", *Robotics and Autonomous Systems*, 2016, doi:10.1016/j.robot.2016.09.015.

[8] G. Bleser, H. Wuest, and D. Stricker:"Online Camera Pose Estimation in Partially Known and Dynamic Scenes", *Proceedings of the 2006 IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 56–65, 2006.

[9] Y. Rubner, C. Tomasi, and L. J. Guibas: "A Metric for Distributions with Application to Image Databases", *Proceedings of the 1998 IEEE International Conference on Computer Vision*, pp. 59–66, 1998.