Automatic camera pose estimation based on a flat surface map

Yonghoon Ji, Atsushi Yamashita, Kazunori Umeda, Hajime Asama

Yonghoon Ji, Atsushi Yamashita, Kazunori Umeda, Hajime Asama, "Automatic camera pose estimation based on a flat surface map," Proc. SPIE 11172, Fourteenth International Conference on Quality Control by Artificial Vision, 111720X (16 July 2019); doi: 10.1117/12.2521780

SPIE.

Event: Fourteenth International Conference on Quality Control by Artificial Vision, 2019, Mulhouse, France

Automatic camera pose estimation based on a flat surface map

Yonghoon Ji*[a], Atsushi Yamashita[b], Kazunori Umeda[a], and Hajime Asama[b]

[a]Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Toyko, 112-8551, Japan;
[b]The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Toyko, 113-8556, Japan

ABSTRACT

This paper proposes a novel approach that performs extrinsic parameter estimation of a camera installed in a man-made environment using a single image. The problem of extrinsic parameter calibration is identical to 6DoF (six-degrees of freedom) localization problem of the camera. We take advantage of line information that is usually present in the man-made environment such as inside of the building. Our approach only requires a flat surface map for a 3D environment model which can be easily obtained from the blueprint of the artificial environment (e.g., CAD data). In order to manage the complicated 6DoF search problem, we propose a novel image descriptor defined in quantized Hough space to perform 3D-2D matching process between line features from the 3D flat surface model and the 2D single image. The proposed method can robustly estimate the complete extrinsic parameters of the camera, as we demonstrate experimentally.

Keywords: Camera calibration, global localization, intelligent space

1. INTRODUCTION

In recent years, the construction of intelligent spaces with a distributed sensor network has attracted the attention of many researchers and are becoming capable of dealing with various indoor environments.[1,2] In order to acquire reliable information from the camera network installed in the intelligent space, precise calibration work for each camera is indispensable. Here, the camera parameters which should be calibrated include intrinsic parameters such as the focal length and the distortion coefficient, and extrinsic parameters representing the 6DoF (six-degrees of freedom) position and orientation information (x, y, z, ψ, θ, φ) of a camera. Among them, we focus on a method to easily estimate the extrinsic camera parameters in this study.

Various methods for calibration of the extrinsic parameters of cameras installed in an environment have been proposed.[3,4] However, these previous approaches were able to deal with only 3DoF (x, y, φ) estimation with restrictive spatial constraints. In this respect, Ji et al. proposed a method to easily calibrate 6DoF extrinsic parameters completely using a 3D line model of the environment.[5] In this study, an image descriptor defined in Hough space to perform line-based 3D-2D matching between a real camera image and a virtual camera image which is re-projected from the 3D line model was proposed. However, there are cases where robust estimation is impossible depending on the configuration of the line information. For example, because the 3D line model does not include surface information, as shown in red lines in Fig. 1 (a), there are cases where lines to be hidden behind a surface such as a wall are incorrectly mapped to a virtual image depending on the positional relationship with the camera, which leads to the unreliable 3D-2D matching process. To remedy this problem, this study proposes a more robust extrinsic parameter calibration method that uses a 3D flat surface model instead of the 3D line model. The 3D flat surface model divides the environment into irregular flat surfaces with different colors consisting of only vertices (i.e., a set of polygons); thus, it leads to very efficient memory management compared with a general structure to represent 3D space such as a multiple voxel-based structures. As shown in Fig. 1 (b), we can solve the abovementioned problem of re-projecting the lines that should not be seen in the virtual image given that the flat surface map contains all surface information for walls, ceilings, and floors.

The remainder of this paper is organized as follows. Section 2 introduces the design of our novel image descriptor generated from image data. The particle filter-based 3D-2D matching process is presented in Section 3.

Further author information: (Send correspondence to Y. Ji)
Y. Ji: E-mail: ji@mech.chuo-u.ac.jp, Telephone: +81 (0)3 3817 1845

Fourteenth International Conference on Quality Control by Artificial Vision, edited by Christophe Cudel, Stéphane Bazeille, Nicolas Verrier, Proc. of SPIE Vol. 11172, 111720X · © 2019 SPIE · CCC code: 0277-786X/19/$21 · doi: 10.1117/12.2521780

Figure 1. Re-projected virtual images: (a) 2D image from 3D line map and (b) 2D image from 3D flat surface map.

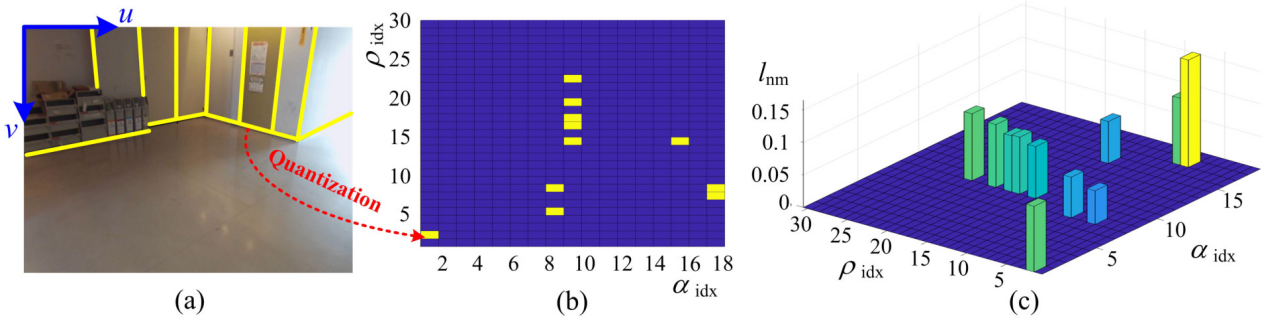

Figure 2. Image descriptor generation: (a) extracted 2D photometric line segments from image data, (b) 2D image descriptor in quantized Hough space,[5] and (c) proposed 2.5D image descriptor including normalized length of line segments in quantized Hough space.

The effectiveness of our calibration scheme is evaluated with the experiment results in Section 4. Finally, Section 5 gives conclusions of this study.

2. DESIGN OF IMAGE DESCRIPTOR

Figure 2 shows an example of the image descriptor generation. In the previous study,[5] as shown in Fig. 2 (b), our research group defined a 2D image descriptor that only the distance from the origin to each line and the slope in the image are projected on the quantized Hough space. Here, ρ_{idx} and α_{idx} respectively mean the quantized distance from the origin to the line and slope information. The image descriptor defined in Hough space expresses the distribution of lines in the captured environment from a certain camera viewpoint (i.e., a 6DoF camera pose) and it has characteristics that change sensitively depending on the viewpoint of the camera; thus, it is very useful for estimating the extrinsic parameters.[5] However, the image descriptor proposed in our previous study has a limitation that there are cases where robust estimation is not possible depending on the configuration of line information projected on the camera image. In other words, it is difficult to use in environments where there are many scenes with similar line distribution of distance ρ_{idx} and slope α_{idx}. In this respect, we further improve the image descriptor to enable a more robust estimation of the camera extrinsic parameters. As shown in Fig. 2 (c), we aim to realize more robust 3D-2D matching by extending the dimensions of the image descriptor to 2.5D from 2D. Here, l_{nm}, which is an extended component, represents normalized length information of each line segment. Extraction of line segments in the image is carried out by Canny edge detector and probabilistic Hough transform.[6] Then, after consolidating all line segments mapped to the same bean on the quantized Hough space, we can calculate the length information of each bean by finding the start and the end points among them respectively.

3. 3D-2D MATCHING

Figure 3 shows the overview of the proposed 6DoF camera pose estimation scheme based on the 3D-2D matching process using the flat surface map and the image descriptor. The image descriptors we designed for the 3D-2D matching are depicted in Fig. 2.

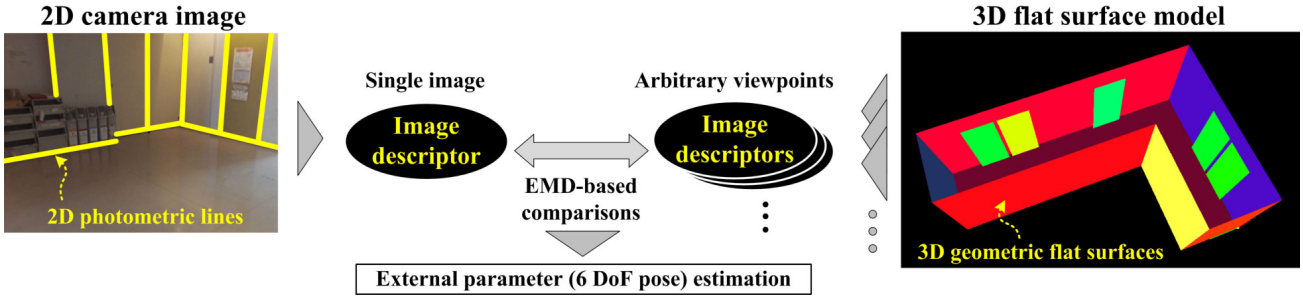

Figure 3. Conceptual image of calibration scheme of extrinsic parameters based on 3D-2D matching using flat surface map and image descriptor.

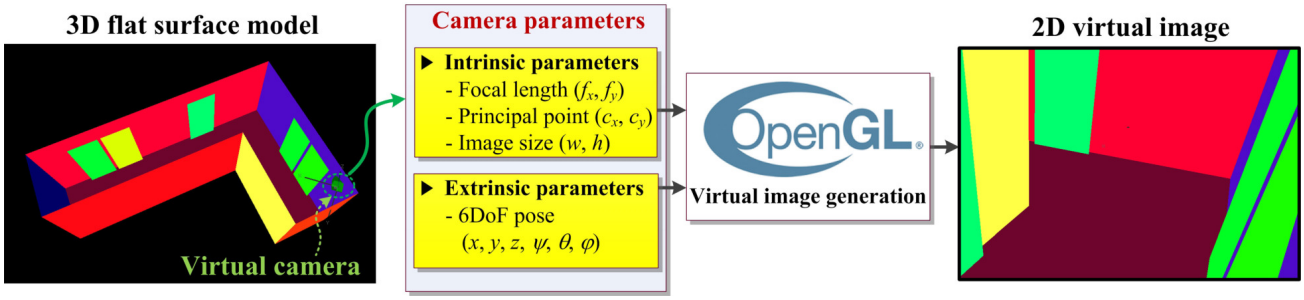

Figure 4. Generation of 2D virtual image from 3D flat surface model using OpenGL library.

First, the image descriptors are generated from a real camera image (i.e., an input image) and virtual images from arbitrary viewpoints in the flat surface map including all surface information represented in different colors. Here, the line information can be extracted easily from both the real image and the virtual images using the Canny edge detector and probabilistic Hough transform[6] as mentioned in the previous section. The basic idea that uses the flat surface map is that the line information is clearly extracted between the flat surfaces of different colors which are projected on the virtual image plane. Figure 4 represents the process of generating a virtual image from an arbitrary viewpoint. We utilize the OpenGL library to generate the virtual images from the arbitrary viewpoints on the flat surface map. The OpenGL library provides the function of the 2D projection that can generate a 2D image based on a pre-given 3D model and pre-defined camera parameters. The camera parameters are divided into two types: the intrinsic parameters (i.e., the focal length, the principal point, and the image size) and the extrinsic parameters (i.e., the viewpoint). Because the OpenGL display pipeline differs from the general pinhole camera model, we have to calculate the OpenGL projection matrix directly from the intrinsic parameters. The OpenGL projection matrix $\mathbb{K}$ to display a 2D image is defined as follows:

$$\mathbb{K} = \begin{bmatrix} 2\frac{f_x}{w} & 0 & 0 & 0 \\ 0 & 2\frac{f_y}{h} & 0 & 0 \\ \frac{1-2c_x}{w} & \frac{-1+2c_y}{h} & -\frac{z_{\text{far}}+z_{\text{near}}}{z_{\text{far}}-z_{\text{near}}} & -1 \\ 0 & 0 & -2\frac{z_{\text{far}}z_{\text{near}}}{z_{\text{far}}-z_{\text{near}}} & 0 \end{bmatrix}, \quad (1)$$

where (f_x, f_y), (c_x, c_y) and (w, h) denote the focal length, the principal point, and the image size (i.e., width and height), respectively. z_{near} and z_{far} represent the standard OpenGL near and far clipping planes, respectively. The extrinsic parameters mean the 6DoF camera pose $(x, y, z, \psi, \theta, \varphi)$.

Next, the 6DoF pose estimation process is performed by evaluating similarities between the image descriptors from the real camera image and the virtual images. The similarity between the image descriptors is evaluated using the earth movers distance (EMD).[7] In order to compute EMD between two image descriptors, each of

image descriptors are converted to sets of clusters, $\mathbb{C}^{(\mathrm{I})}$ and $\mathbb{C}^{(\mathrm{V})}$, as follows:

$$\mathbb{C}^{(\mathrm{I})} = \left[(\boldsymbol{c}_i^{(\mathrm{I})}, \varpi_i^{(\mathrm{I})}) \quad | \quad 1 \leqq i \leqq N^{(\mathrm{I})} \right], \tag{2}$$

$$\mathbb{C}^{(\mathrm{V})} = \left[(\boldsymbol{c}_j^{(\mathrm{V})}, \varpi_j^{(\mathrm{V})}) \quad | \quad 1 \leqq j \leqq N^{(\mathrm{V})} \right], \tag{3}$$

where $\boldsymbol{c}$ and ϖ denote the cluster representative which means the coordinates of the image descriptor (ρ_{idx}, α_{idx}) and weight value l_{nm} that corresponds to the cluster, respectively. The size of each cluster N is equal to the number of beans that are weighted. In other words, it means the number of elements in the image descriptor that contain length information l_{nm} (e.g., $N = 11$ in the case of Fig. 2). EMD calculation defined in this study is as follows:

$$EMD(\mathbb{C}^{(\mathrm{I})}, \mathbb{C}^{(\mathrm{V})}) = \frac{\sum_{i=1}^{N^{(\mathrm{I})}} \sum_{j=1}^{N^{(\mathrm{V})}} f_{ij} d_{ij}}{\sum_{i=1}^{N^{(\mathrm{I})}} \sum_{j=1}^{N^{(\mathrm{V})}} f_{ij}}, \tag{4}$$

$$d_{ij} = \sqrt{\tau_\rho (\Delta\rho_{\mathrm{idx}})^2 + \tau_\alpha (\Delta\alpha_{\mathrm{idx}})^2}, \tag{5}$$

$$\Delta\rho_{\mathrm{idx}} = \rho_j^{(\mathrm{V})} - \rho_i^{(\mathrm{I})}, \tag{6}$$

$$\Delta\alpha_{\mathrm{idx}} = \arctan\left(\frac{\tan\alpha_j^{(\mathrm{V})} - \tan\alpha_i^{(\mathrm{I})}}{1 + \tan\alpha_j^{(\mathrm{V})} \tan\alpha_i^{(\mathrm{I})}} \right). \tag{7}$$

Here, d_{ij} in Eq. (5) denotes the user-defined distance, which is the distance between clusters $\boldsymbol{c}_i^{(\mathrm{I})}$ and $\boldsymbol{c}_i^{(\mathrm{V})}$. τ_ρ and τ_α respectively mean the weight parameters that adjust the importance of the distance distribution and slope distribution. f_{ij} in Eq. (4) denotes a flow element between $\boldsymbol{c}_i^{(\mathrm{I})}$ and $\boldsymbol{c}_j^{(\mathrm{V})}$ which can be derived by solving the well-known transportation problem taking the corresponding weights $\varpi_i^{(\mathrm{I})}$ and $\varpi_j^{(\mathrm{V})}$ into consideration. EMD can measure the similarity between two multi-dimensional distributions robustly even if there are some little errors in the descriptors. More details on EMD can be found in.[7]

Consequently, the 6DoF camera pose at which the evaluation value of EMD is the minimum can be found as the optimal extrinsic camera parameters. We use a particle filter that does not require initial values as an extrinsic parameter to carry out the abovementioned 3D-2D matching process. In addition, the particle filter can easily take constraints on the solution space (i.e., the searching space for the 6DoF camera poses) into consideration. In general, because of space limitations, the cameras are installed on the occupied region, such as interior walls; thus, the 3D flat surface model of the environment can provide very useful information which can be divided into two types of constraints: a constraint on camera position (x, y, z) and a constraint on camera orientation (ψ, θ, φ). Therefore, these useful constraints from the flat surface map can be applied to the particle filter to reduce the solution space dramatically.

4. EXPERIMENTAL RESULTS

In order to evaluate our 6DoF pose estimation framework that uses the flat surface map and the improved image descriptor, a comparative experiment in a simulation environment was conducted with a virtual camera. The size of the simulation environment shown in Fig. 5 (a) was 5 m × 8 m × 3 m, including various line features located on the sides of the walls, doors, and windows. We assume that the environment is fully expressible by a set of multiple rectangular planes in this experiment. However, of course, using more complex polygonal planes, we can deal with a more complex environment as well. Here, the real 6DoF pose of the virtual camera that should be estimated are represented in purple color and corresponding simulated image the area behind the wall is not projected is shown in Fig. 5 (b). The coordinate system adopted in this study (i.e., the relationship between the camera coordinate frame and the world coordinate frame in the 3D space) is also appeared in Fig. 5 (a). In this study, the optical axis of the camera is defined as x-axis in the camera coordinate frame. The real 6DoF pose of the virtual camera installed on the wall was (3.7 m, 8.0 m, 1.9 m, 0.0 deg, 16.0 deg, -92.0 deg). In this simulation experiment, intrinsic parameters (f_x, f_y) and (c_x, c_y) of the virtual camera were set to (930, 930), and (640, 400),

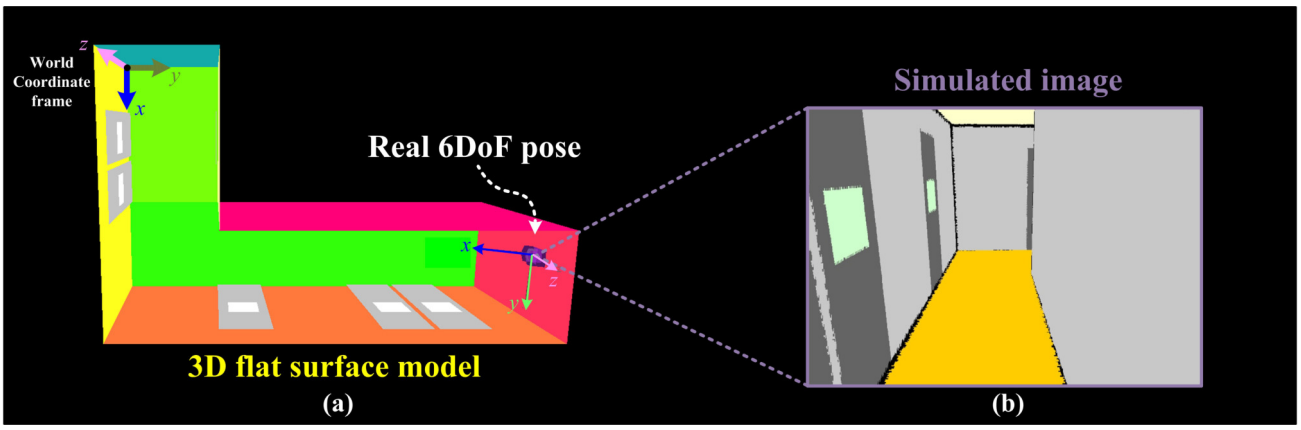

Figure 5. Simulation environment with virtual camera: (a) flat surface map of simulation environment and (b) simulated image from virtual camera.

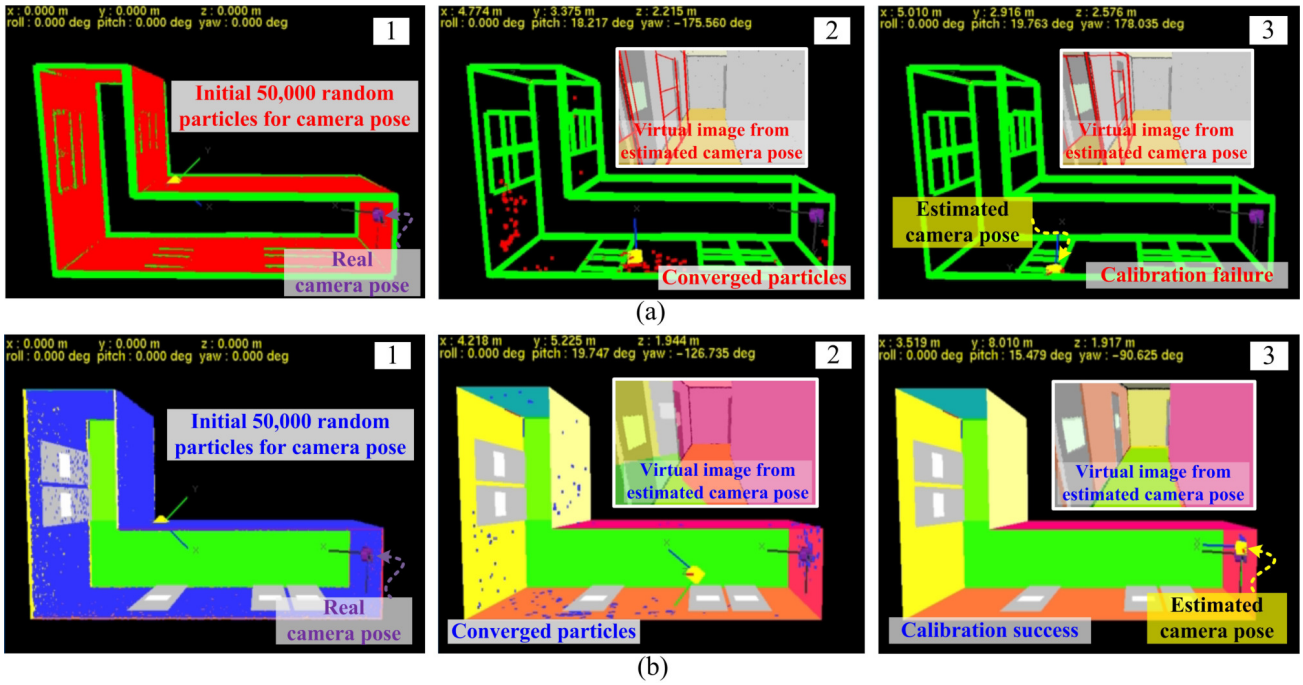

Figure 6. Simulation experimental results for several stages of particle filter in case of using: (a) 3D line model[5] and (b) using 3D flat surface model.

respectively. The image size (w, h) was (1280, 800). Since the roll angle rotation (i.e., the rotation on the optical axis) has a slight influence on the scope of the camera observation, we fixed it at zero deg and estimation was not carried out. In this simulation, a maximum of 50,000 particles which have camera position and orientation information (i.e., the 6DoF pose) were used on the wall of the whole environment and the number of particles is adjusted according to the distribution of the particles. In the particle filter process, it is necessary to generate virtual images corresponding to the number of particles at every step. Thus, considering the processing speed, the size of the virtual images generated by the OpenGL library was reduced to (128, 80) (i.e., 1/10 size) and the extracted line parameters were multiplied by ten times matching up the scale of them with the input image.

The several stages of the particle filter iterations and convergence process for the camera pose are illustrated in Fig. 6. The particles are initialized globally based not on the initial conditions, but on the constraints of the position and the orientation as mentioned in section 3. Fig. 6 (a) and Fig. 7 (a) shows the estimation results of

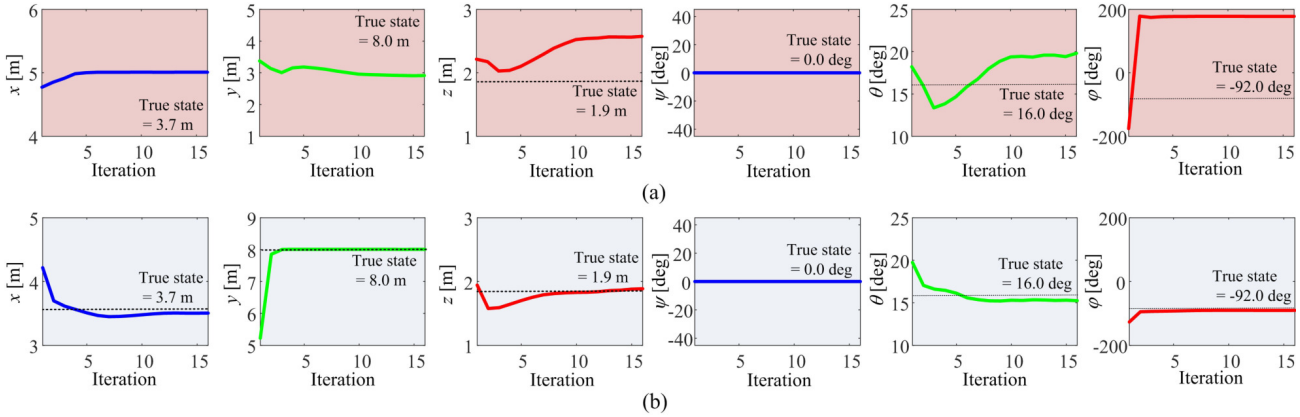

Figure 7. Convergence process for camera parameters $(x, y, z, \psi, \theta, \varphi)$ in case of using: (a) 3D line model[5] and (b) 3D flat surface model.

the camera pose using the 3D line model proposed in the previous study.[5] As mentioned in the introduction, due to the problem of re-projecting the lines that should not be seen in the virtual image, particles converge to the wrong position, and estimation of the camera pose (5.01 m, 2.92 m, 2.58 m, 0.00 deg, 19.76 deg, 178.04 deg) failed. On the other hand, when using the 3D flat surface model, the particles accurately converged on the real pose as shown in Fig. 6 (b) and Fig. 7 (b), and estimation of the camera pose (3.52 m, 8.01 m, 1.92 m, 0.00 deg, 15.48 deg, -90.63 deg) succeed. Here, the images illustrated in Fig. 6 are alpha blended images which include both the simulated image (i.e., the input image) and the generated virtual images from estimated pose at each iteration. In case of using the 3D line model, the projected lines on the image plane did not match with the simulated image even after the particles had converged. On the other hand, in case of using the 3D flat surface model, we can initially easily recognize large differences between the simulated image and the generated virtual image, but after convergence, these images are almost identically matched with small errors.

5. CONCLUSION

This paper proposed a novel method to easily calibrate the extrinsic camera parameters (i.e., the 6DoF camera pose) of the camera installed on the environment. We used the 3D flat surface model instead of the 3D line model in order to solve the limitation of the previous study[5] on some cases of re-projecting the lines that should not be seen in the virtual image. In conclusion, it was possible to realize 3D-2D matching that does not depend on the spatial limitations of the environment, and dramatically improves the robustness of extrinsic parameter estimation.

REFERENCES

[1] Lee, J. H. and Hashimoto, H., "Intelligent spaceconcept and contents," *Advanced Robotics* **16(3)**, 265–280 (2002).

[2] Sato, T., Nishida, Y., and Mizoguchi, H., "Robotic room: symbiosis with human through behavior media," *Robotic and Autonomous Systems* **18**, 185–194 (1996).

[3] Rahimi, A., Dunagan, B., and Darrell, T., "Simultaneous calibration and tracking with an of non-overlapping sensors," *Proc. IEEE Conference on Computer Vision and Pattern Recognition* **1**, 187–194 (2004).

[4] Funiak, S., Guestrin, C., Paskin, M., and R. Sukthankar, R., "Distributed localization of networked cameras," *Proc. 5th International Conference on Information Processing in Sensor Networks*, 34–42 (2006).

[5] Ji, Y., Yamashita, A., and Asama, H., "Automatic calibration of camera sensor network based on 3d texture map information," *Robotic and Autonomous Systems* **87**, 313–328 (2017).

[6] Kiryati, N., Eldar, Y., and Bruckstein, A. M., "A probabilistic hough transform," *Pattern Recognition* **24(4)**, 303–316 (1991).

[7] Rubner, Y., Tomasi, C., and Guibas, L. J., "A metric for distributions with applications to image databases," *Proc. 6th IEEE International Conference on Computer Vision*, 59–66 (1998).