

High-accuracy Range Image Generation by Fusing Binocular and Motion Stereo Using Fisheye Stereo Camera*

Hiroataka Iida¹, Yonghoon Ji¹, *Member, IEEE*, Kazunori Umeda¹, *Member, IEEE*, Akira Ohashi², Daisuke Fukuda², Shuzo Kaneko², Junya Murayama², and Yoshitaka Uchida²

Abstract—This paper proposes a method that fuses two stereo measurements: binocular and motion stereo. The method is implemented for a fisheye stereo camera. The two stereo measurements have difficult characteristics. In case of the binocular stereo, the length of the baseline is usually small due to the installation conditions to a robot or a car. On the other hand, in the case of the motion stereo, the baseline is the distance the camera moves between two consecutive frames, and thus, the length of the baseline is usually longer than binocular stereo. In addition, the area-based approach is used for binocular stereo and the feature-based approach is used for motion stereo in the corresponding point search. In summary, the two stereo measurements have differences in the length of the baseline and the accuracy of the corresponding point search. Therefore, their fusion is studied to improve the stereo measurement. A bilateral-like filter, which is a weighted averaging method, is proposed for the fusion of the two measurements. In motion stereo, the proposed method is verified by outdoor experiments.

I. INTRODUCTION

In recent years, driving support systems for automobiles using cameras and range sensors have attracted the attention of many researchers [1], [2]. A stereo camera and a laser range finder are most widely used; however, these sensors have disadvantages, such as the narrow range of distance measurements and low measurement density respectively, which lead to the existence of undetectable objects. On the contrary, a fisheye camera is considered to be suitable for a vehicle, given that the angle of view is 180 deg or more, and its size is relatively small. There are several previous studies on the fisheye camera. Abraham et al. simplified the stereo matching process by applying stereo rectification to fisheye stereo cameras [3]. Moreau et al. exploited the fisheye stereo camera with an equisolid projection model to construct an environmental restoration method [4]. Hane et al. realized three-dimensional (3D) environmental measurements in real time using a plane-sweeping method [5]. In these studies, fisheye images are usually converted to perspective projection images in order to simplify the search process for corresponding points extracted from the images. However, since regions far away from the image center are stretched after the conversion, it becomes difficult to perform a stereo matching process in such regions. On the other hand, Schneider et al. analyzed an approach to exploiting existing

dense stereo methods with wide-angle fisheye cameras that have fields of view of more than 180 deg [6]. Ma et al. implemented a 3D reconstruction algorithm for multiple spherical images [7]. Ohashi et al. proposed a method for the fisheye stereo camera that utilizes an equirectangular image for conversion [8], [9]. The equirectangular image which is converted from the fisheye image based on the orthogonal coordinate system has no stretching in the regions far from the image center and can reduce distortion of the original fisheye image significantly so that simplification of corresponding point search can be realized. However, the issues of long range accuracy of the fisheye stereo camera due to the influence of the small baseline length and the extrinsic parameter error should be solved.

Therefore, we focus on the fusion of spatial and temporal directions to improve the accuracy of the fisheye stereo camera. Fusion of spatial and temporal directions has been considered in some studies. Hazen et al. presented a way to exploit the intrinsic spatio temporal correlations in a dynamic scene for temporal resolution enhancement of the scene [10]. However, this research is not aimed at high-accuracy 3D reconstruction. Zhu et al. presented a spatial and temporal MRF to infer high-quality dynamic depth by fusing of stereo and a TOF sensor [11]. However, energy optimization problem is not suitable for real-time operation. Based on these ideas, we focus not only on the binocular stereo that handles the same frame images but also the motion stereo that deals with time series images. In the case of the binocular stereo, area-based matching is used for corresponding points searching. Therefore, the measured density is high, but the correspondence is often erroneous. On the other hand, in the case of the motion stereo, feature-based matching is used for corresponding points searching. Therefore, the measured density is low, but there is little false correspondence. In this respect, the introduction of the motion stereo can be very useful because cameras mounted on a vehicle are usually moving. In order to take full advantages of two different stereo systems, we propose a novel filter, similar to the bilateral filter, which performs weighted averaging in the spatial and disparity directions to decrease false matching.

The remainder of this paper is as follows. Sections 2 and 3 describe the binocular stereo and motion stereo, respectively. Section 4 presents our proposed method that fuses each stereo result. Section 5 shows experiments we conducted. Finally, Section 6 gives the conclusion.

*This work was supported by JSPS KAKENHI Grant Number 19H04191.

¹The Course of Precision Engineering, School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, Japan
iida@sensor.mech.chuo-u.ac.jp

² Clarion, 7-2 Shintoshin, Chuo-ku, Saitama-shi, Saitama, Japan

II. BINOCULAR STEREO USING FISHEYE CAMERAS

This section briefly summarizes the binocular stereo based on an equirectangular image using the fisheye cameras [8], [9].

A. Fisheye Camera Model

Fisheye lenses often do not follow their projection models due to various factors (e.g., the deviation of the optical axis). Therefore, in this study, we use the generic omnidirectional camera model proposed by Scaramuzza et al. [12]. Intrinsic parameters of the fisheye camera are calculated using this camera model.

B. Equirectangular Image

Since fisheye images have large distortion, it is difficult to search corresponding points. To reduce these distortions of the fisheye image, the equirectangular image is used in this study [8], [9]. The equirectangular image is generated from the fisheye image through transformation using orthogonal and equidistant coordinates represented by the elevation angle λ and the azimuth angle ϕ , respectively. A conceptual diagram for converting a fisheye image to an equirectangular image is shown in Fig. 1.

C. Stereo Matching in Equirectangular images

In order to measure the distance, it is necessary to search corresponding points from two images from the left and right cameras and obtain the disparity at the corresponding point. Here, the epipolar line is generally used to reduce the processing time. In the case of the equirectangular image, the epipolar line becomes a curve because it cannot completely eliminate the distortion of the fisheye image. Assuming that the elevation angle is ϕ_0 when the azimuth angle is zero, the locus of the epipolar line on the equirectangular image is as follows:

$$\phi = \tan^{-1}(\tan \phi_0 \cos \lambda). \quad (1)$$

Therefore, the search process for corresponding points should be performed on the curve as shown in Fig. 2. Here, we apply the Sum of Absolute Difference (SAD)-based block matching which requires low computational complexity as area-based matching in the search process, because we aim to develop a real-time application. In addition, equiangular linear fitting is also performed to estimate disparity information with subpixel accuracy [13].

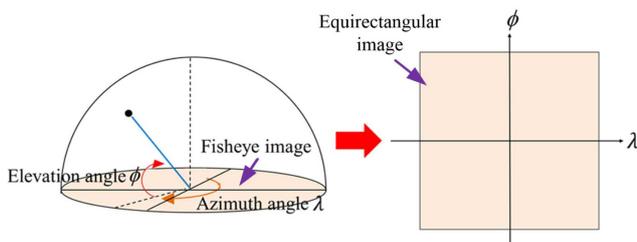


Fig. 1. Transformation from fisheye image to equirectangular image.

D. Distance Measurement using an Equirectangular Image

After searching corresponding points, the distance can be obtained as shown in Fig. 3. The measured distance D can be calculated as follows:

$$D = \frac{b \cos \lambda_l}{\sin \Delta \lambda \cos \phi_r}, \quad (2)$$

where b and $\Delta \lambda$ denote the length of the baseline and the azimuth angle of disparity between the left and right cameras, respectively. λ_l and ϕ_r are the azimuth and elevation angles from the left and right cameras to the target, respectively.

III. MOTION STEREO USING FISHEYE CAMERA

This section describes a novel scheme that applies three types of matching in the motion stereo using the fisheye cameras. This process is divided into four steps: feature point extraction, three types of matching, motion parameter calculation, and 3D reconstruction. The detailed explanation for each step is as follows.

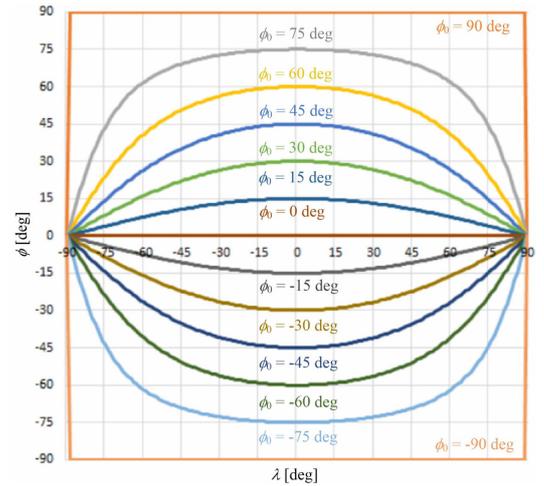


Fig. 2. Epipolar line on equirectangular image.

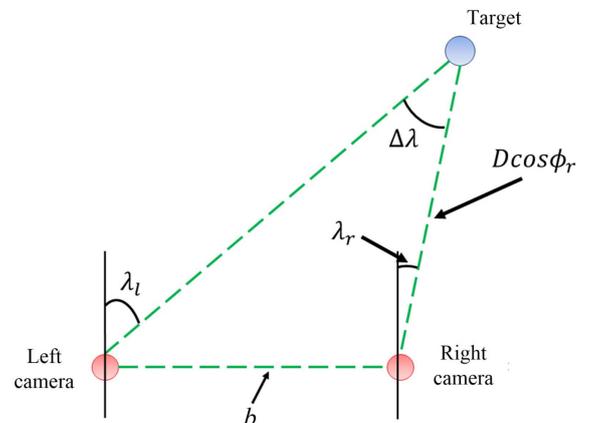


Fig. 3. Measurement of distance on equirectangular image.

A. Extraction of Feature Points

In this study, feature points on images are detected by AKAZE [14], [15]. AKAZE is capable of extracting features that are robust to changes in image blur, rotation, scale, and brightness. Furthermore, it is effective for images captured from moving cameras. Given that AKAZE searches the corresponding point on a rectangular window based on the extremum, using the equirectangular image, which has less image deformation than the fisheye image, is considered to be suitable.

B. Three Types of Matching

In general, when feature point matching is performed using only two images, the number of detected corresponding points tends to be small. To cope with this problem, we propose a novel approach that applies the matching of three kinds of patterns at the same time, as shown in Fig. 4. In other words, we use both information of the time (i.e., equirectangular images from time $t-1$ and t) and space (i.e., equirectangular images from left and right cameras). As a result, it is possible to increase the number of reliable corresponding points from the set of many feature points.

C. Motion Parameters

The motion parameters of the camera between two images are defined as the rotation matrix R and the translation vector t . These parameters are obtained using structure from motion (SfM) as follows. First, the essential matrix E is calculated from the corresponding points that are generated from the matching process in Subsection III-B. Here, the corresponding points are defined based on equirectangular coordinates (λ, ϕ) ; thus, they are converted to perspective projection coordinates (x, y, z) before finding the essential matrix E , as follows:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \tan \lambda \\ \tan \phi \\ \cos \lambda \\ 1 \end{bmatrix}. \quad (3)$$

Next, we can find motion parameters R and t through the singular value decomposition (SVD) of E .

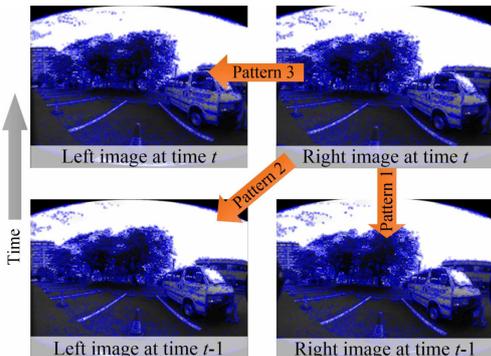


Fig. 4. Three types of matching of stereo image pairs.

D. 3D Reconstruction

The relationship between the image coordinates (u, v) of the feature points and the 3D coordinates (X, Y, Z) in the 3D space is calculated using the perspective projection matrix P , which includes an extrinsic parameter of the camera, as follows:

$$\mathbf{u} \sim P\mathbf{X}, \quad (4)$$

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \quad (5)$$

where \sim represents equality as homogeneous coordinates. By substituting all corresponding point data obtained by feature point matching into (5), the form of the matrix-vector equation can be defined as follows:

$$B\mathbf{X} = \mathbf{b}, \quad (6)$$

$$\begin{bmatrix} p_{31}u - p_{11} & p_{32}u - p_{12} & p_{33}u - p_{13} \\ p_{31}u - p_{21} & p_{32}u - p_{22} & p_{33}u - p_{23} \\ p'_{31}u' - p'_{11} & p'_{32}u' - p'_{12} & p'_{33}u' - p'_{13} \\ p'_{31}u' - p'_{21} & p'_{32}u' - p'_{22} & p'_{33}u' - p'_{23} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} p_{14} - p_{34}u \\ p_{24} - p_{34}v \\ p'_{14} - p'_{34}u' \\ p'_{24} - p'_{34}v' \end{bmatrix}. \quad (7)$$

Here, (u, v) and (u', v') denote the corresponding points in each of two images. p and p' refer to the elements of the perspective projection matrix P for each camera state that captures each image. Consequently, 3D points $\hat{\mathbf{X}} = (\hat{X}, \hat{Y}, \hat{Z})$ are estimated with the least squares method as follows:

$$\hat{\mathbf{X}} = B^+\mathbf{b}. \quad (8)$$

However, SfM is not generally able to determine the actual scale. To solve this problem, road surface plane estimation is performed to obtain the actual scale [16]. Since the height from the plane to the center of the lens can be calculated by obtaining the plane parameters, the scale is obtained by calculating the ratio of the height to the actual camera height. Next, the 3D reconstruction results are converted to disparity values. Since the corresponding point groups obtained in Subsection III-B are in three types, this calculation is repeated three times and then integrated.

IV. FUSION OF RANGE IMAGES BY BILATERAL-LIKE FILTER

The flow of the proposed method is shown in Fig. 5. From the above processes, we can obtain two kinds of disparity information. The first information is the dense disparity obtained by the binocular stereo. The second information is the disparity for each feature point obtained by motion stereo. Since the two measurement methods differ in the matching method, their robustness against false matching is

different [17]. Therefore, we can take the fusion of two kinds of disparity information into account in order to generate more reliable range images. Hence, in this section, we propose a novel filtering method that effectively merges both information using a weighted value based on the disparity information. We call this filter as a bilateral-like filter, given that this filtering process is similar to that of the bilateral filter [18]. The updated disparity information \hat{D} of the pixel (u, v) is calculated by:

$$\hat{D} = \frac{\sum_{k=1}^n w_k D_{\text{disp}(m)k}}{\sum_{k=1}^n w_k}, \quad (9)$$

$$w_k = \exp\left(-\frac{d_k^2}{2\sigma_1^2}\right) \exp\left(-\frac{(D_{\text{disp}(b)} - D_{\text{disp}(m)k})^2}{2\sigma_2^2}\right), \quad (10)$$

$$d_k = \sqrt{(u_k - u)^2 + (v_k - v)^2}, \quad (11)$$

where σ_1 and σ_2 are the standard deviations of the Gaussian distributions representing the weights in the directions of the image space and the disparity, respectively. This filtering calculates the weighted average based on not only the distances d_k to surrounding feature points but also the difference between the disparities obtained from the binocular stereo $D_{\text{disp}(b)}$ at (u, v) and the motion stereo $D_{\text{disp}(m)}$ at surrounding feature points (u_k, v_k) . Figure 6 shows a conceptual image of weighted averaging based on the proposed bilateral-like filter. The distances, d_k , are calculated only for the surrounding feature points within a limited range, d_{th} . Thus, n means the number of the surrounding feature points within the limited range, d_{th} .

V. EXPERIMENTS

A. Experimental Conditions

We conducted experiments to evaluate the accuracy of the long range measurement. The cameras used in the experiment were FLIR Flea3 equipped with a fisheye lens, SPACE TV1634M. The intrinsic parameters of the fisheye lens were

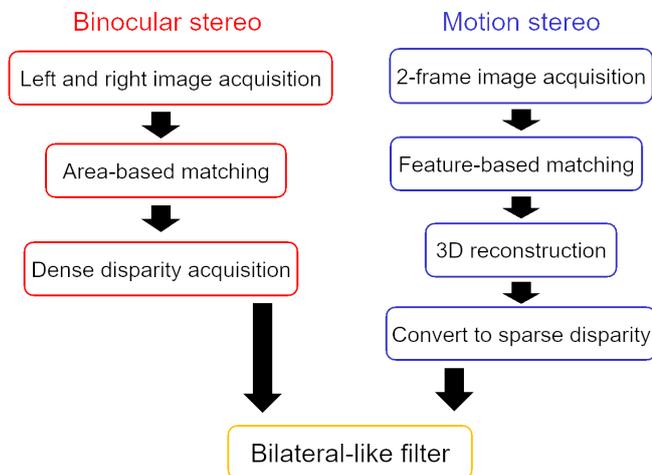


Fig. 5. Flow chart of the proposed method: bilateral-like filter.

estimated using the OcamCalib Toolbox for MATLAB [12]. The resolution of camera was $1,328 \times 1,048$ pixels, and its baseline was 52 mm. The angle of view was 165 deg in the horizontal direction and 132 deg in the vertical direction. Figure 7 shows the fisheye stereo camera used for our experiments. For the binocular stereo, the template size was 7×7 pixels, and the disparity search range was 48 pixels. For the motion stereo, images taken by the fisheye stereo camera moving 0.15 m in the direction of the optical axis between two frames were used. Figure 8 shows a diagram in which the experimental environment and measurement points are shown by color. Measurement distances were set to 7 m and 10 m, and the measurement target was a rhombus made of black and white paper. True value was measured by a laser range finder (BOSCH Inc.). The measurement values of three points in total, i.e., the point of interest and its upper and lower points, were taken as one measurement, and two measurements were performed. The mean and the standard deviation of the measurement errors were obtained from totally six points. Table I summarizes the parameters appearing in bilateral-like filter for experimental condition. To evaluate the matching of three kinds of patterns in motion stereo, we compared experiments using three patterns and pattern 1 only in Fig. 4 when applying the bilateral-like filter.

B. Experimental Results

Range images at a measurement distance of 7 m when the target is at the center of the image are shown Fig. 9. Since no

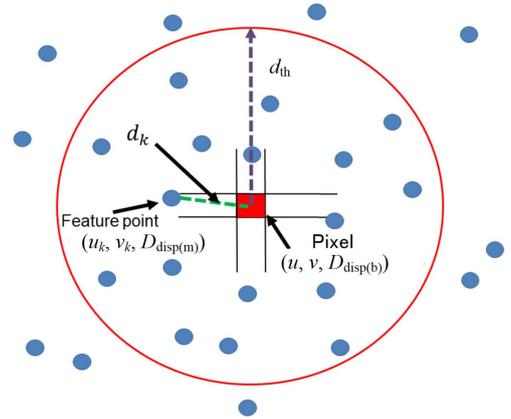


Fig. 6. Conceptual image of update disparity information using bilateral-like filter.



Fig. 7. Fisheye stereo camera.

feature points were extracted in the regions without texture in the image, the distance was not measured. Thus, such regions are represented in black. The mean and the standard deviation of the measurement errors are shown Figs. 10 and 11. Here, the color of the bar graph corresponds to the measurement point of Fig. 8 (b). NULL represents the case where the distance is not measured due to the weak texture. From Fig. 9, thanks to the bilateral-like filter, measurement errors were significantly reduced for buildings in distant locations because feature-based matching is robust against various environments. From Figs. 10 and 11, the distance accuracy was improved by the proposed method. This is due to the baseline of motion stereo that is nearly three times of that of binocular stereo. Moreover, since the scale reconstruction by road surface estimation was performed very accurately, the accuracy of the distance was significantly improved. In addition, it is shown that the distance accuracy is better when three types of matching was applied, which verifies the effectiveness of the proposed method. In particular, when

one type of matching (i.e., pattern 1) was applied, the distance accuracy at the center of the image was extremely poor. This is because in the 3D reconstruction of pattern 1 in Fig. 4, the baseline is in the optical axis direction. Therefore, it is considered that the weighting by position of the image according to the type of matching is effective. As for calculation cost, short processing time of 0.7 s was realized by introducing parallel operation using GPU. Note that the same processing by the CPU takes more than 40 s.

VI. CONCLUSIONS

In this study, we realized accurate range image generation using the fisheye stereo camera by fusing the dense disparity information obtained from the binocular stereo and the disparity of feature points obtained from the motion stereo. The accuracy of the distance measurement using the fisheye stereo camera based on the equirectangular image is significantly improved by introducing a bilateral-like filter that calculates the weighted average according to the Gaussian distribution in the directions of the image space and the disparity.

As a future work, we will introduce weighting based on the position of the image according to the type of matching.

REFERENCES

- [1] J. W. Perng, P. Y. Liu, K. Q. Zhong, and Y. W. Hsu, "Front object recognition system for vehicles based on sensor fusion using stereo vision and laser range finder," in *Proceedings of the 2017 IEEE International Conference on Consumer Electronics*, pp. 261-262, 2017.
- [2] N. Sasaki, N. Iijima, and D. Uchiyama, "Development of ranging method for inter-vehicle distance using visible light communication and image processing," in *Proceedings of the 2015 15th International Conference on Control Automation and Systems*, pp. 666-670, 2015.
- [3] S. Abraham and W. Forstner, "Fish-eye stereo calibration and epipolar rectification," *Journal of Photogrammetry and Remote Sensing*, vol. 59, no. 5, pp. 278-288, 2005.
- [4] J. Moreau, S. Ambellouis, and Y. Ruichek, "Equisolid fisheye stereo- vision calibration and point cloud computation," in *Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 167-172, 2013.
- [5] C. Hane, L. Heng, G. H. Lee, A. Sizov, and M. Pellefeys, "Real-time direct dense matching on fisheye images using plane-sweeping stereo," in *Proceedings of the 2014 2nd International Conference on 3D Vision*, pp. 57-64, 2014.
- [6] J. Schneider, C. Stachniss, and W. Forstner, "On the accuracy of dense fisheye stereo," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 227-234, 2016.
- [7] C. Ma, L. Shi, H. Huang, and M. Yan, "3D reconstruction from full-view fisheye camera," *arXiv*, 2015.
- [8] A. Ohashi, Y. Tanaka, G. Masuyama, K. Umeda, D. Fukuda, T. Ogata, T. Narita, S. Kaneko, Y. Uchida, and K. Irie "Fisheye stereo camera using equirectangular images," in *Proceedings of the 2016 11th France-Japan Congress on Mechatronics 9th Europe-Asia Congress on Mechatronics 17th International Conference on Research and Education in Mechatronics*, pp. 284-289, 2016.
- [9] A. Ohashi, F. Yamano, G. Masuyama, K. Umeda, D. Fukuda, K. Irie, S. Kaneko, J. Murayama, and Y. Uchida, "Stereo rectification for equirectangular images," in *Proceedings of the 2017 IEEE/SICE International Symposium on System Integration*, 2017.
- [10] D. Hazen, R. Puri, and K. Ramchandran, "Multi-camera Video Resolution Enhancement by Fusion of Spatial Disparity and Temporal Motion fields," in *Proceedings of the Fourth IEEE International Conference on Computer Vision Systems (ICVS 2006)*, 2006.
- [11] J. Zhu, L. Wang, J. Gao, and R. Yang, "Spatial-Temporal Fusion for High Accuracy Depth Maps Using Dynamic MRFs," in *Proceedings of the IEEE Transactions on Pattern Analysis and Machine*, vol. 32, no. 5, pp. 899-909, 2010.

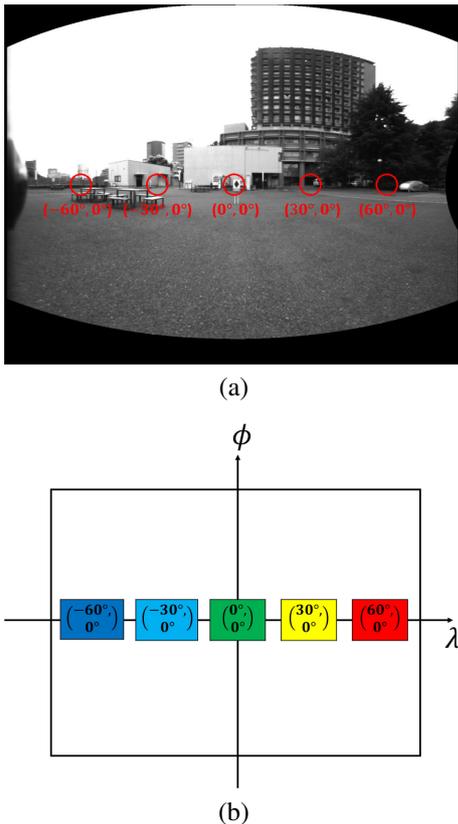


Fig. 8. Experimental conditions: (a) outdoor environment and (b) measurement points.

TABLE I
PARAMETERS FOR BILATERAL-LIKE FILTER

Parameter	
Distance threshold d_{th} [pixel]	12
Standard deviation σ_1 [pixel]	6
Standard deviation σ_2 [pixel]	5

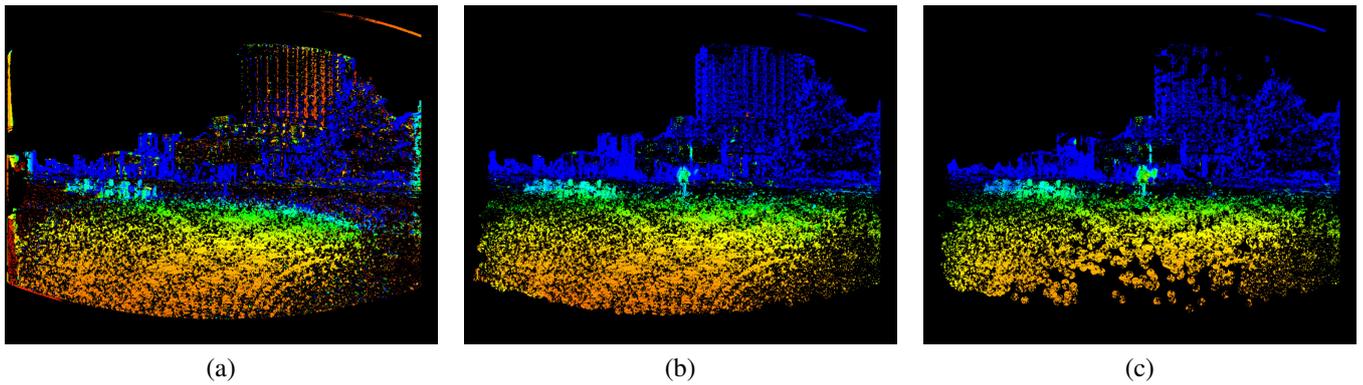


Fig. 9. Comparison of range images: (a) without bilateral-like filter, (b) with bilateral-like filter using three types of matching for motion stereo, and (c) with bilateral-like filter using one type of matching for motion stereo. Color represents distance values from 0 m (red) to 10 m and above (blue).

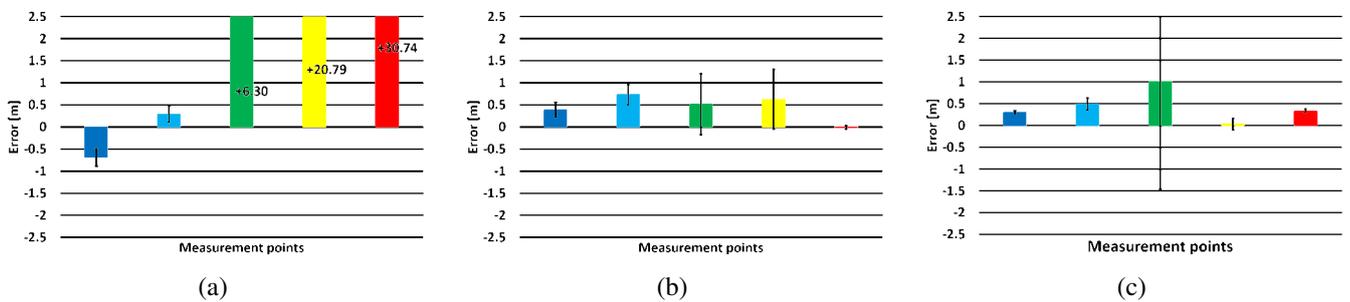


Fig. 10. Mean and standard deviation of errors at distance 7 m: (a) without bilateral-like filter, (b) with bilateral-like filter including three types of matching, and (c) with bilateral-like filter including one type of matching.

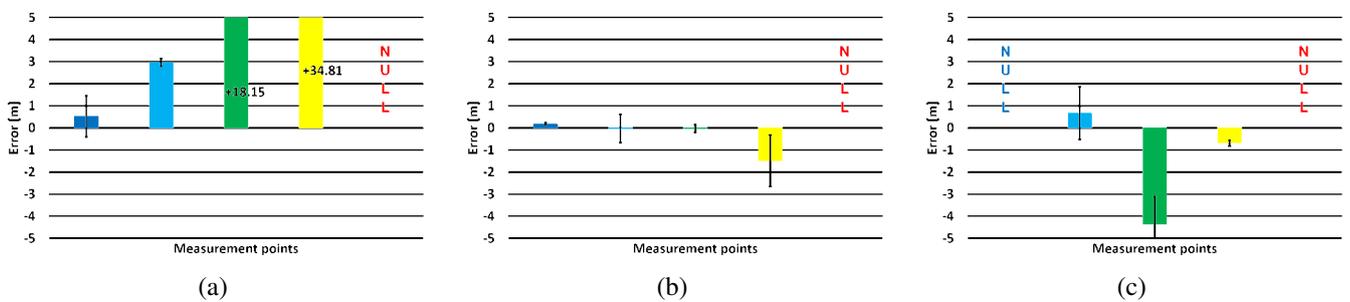


Fig. 11. Mean and standard deviation of errors at distance 10 m: (a) without bilateral-like filter, (b) with bilateral-like filter using three types of matching for motion stereo, and (c) with bilateral-like filter using one type of matching for motion stereo.

[12] D. Scaramuzza, A. Martinelli, and R. Slegwart, "A toolbox for easily calibrating omnidirectional cameras," in *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5695-5701, 2006.

[13] M. Shimizu and M. Okutomi, "Sub-pixel estimation error cancellation on area-based matching," *International Journal of Computer Vision*, vol. 63, no. 3, pp.207-224, 2005.

[14] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE Features," *Computer Vision—ECCV 2012. Lecture Notes in Computer Science*, vol. 7577, pp. 214-227, 2012.

[15] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in *Proceedings of the British Machine Vision Conference*, 2013.

[16] K. Yamaguchi, T. Kato, and Y. Ninomiya, "Moving Obstacle Detection using Monocular Vision," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, 2006.

[17] N. Jayanthi and S. Indu, "Comparison of image matching techniques," *International Journal of Latest Trends in Engineering and Technology*, vol. 7, issue 3, pp. 396-401, 2016.

[18] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the 1998 IEEE 6th International Conference on Computer Vision*, pp. 839-846, 1998.