2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Road Surface Estimation and Obstacle Detection Using Fisheye Stereo Camera and Monocular Depth Estimation

Hikaru Chikugo[1], Jonoshin Shiino[1], Sarthak Pathak[2], Kazunori Umeda[3]

Abstract— **In this study, we propose a method for road surface estimation and obstacle detection using a fisheye stereo camera. In obstacle detection using stereo cameras, obstacles are detected based on distance information obtained through stereo matching. However, there are regions where stereo matching cannot obtain reliable disparity. Moreover, direct obstacle detection using deep learning cannot detect obstacles that are not included in the training data. Therefore, we first detect obstacles on the road surface using the relative depth obtained from monocular depth estimation. Then, by focusing only on the obstacle regions for distance measurement, we aim to detect all obstacles. Experiments demonstrate the ability to detect only obstacles with high accuracy.**

I. INTRODUCTION

Recent years have seen remarkable advancements in autonomous driving technologies, and numerous cases have already reached practical implementation stages[1], [2]. These systems require situational awareness to detect surrounding obstacles, primarily relying on depth information. Representative sensors used to obtain such depth information include stereo cameras, LiDAR, and sonar. However, conventional sensors suffer from issues such as limited measurement range and low angular resolution, making it difficult to detect thin or laterally positioned obstacles. While some approaches combine multiple sensors or mechanisms to extend the measurable range, this often results in constraints on sensor placement, decreased maintainability, and increased costs.

In this study, we focus on fisheye stereo cameras. Fisheye cameras offer an extremely wide field of view, approximately 180°, which allows for a wide area to be measured using a single device. Furthermore, their large depth of field ensures that objects remain relatively sharp regardless of distance, which is advantageous for image recognition. Given applications such as camera sensors for autonomous parking systems or mobile robots, fisheye stereo cameras are well-suited for wide-range measurements. In previous research, Ohashi et al. proposed converting fisheye images into equirectangular images to reduce distortion and facilitate 3D reconstruction[3]. However, this approach suffers from decreased distance estimation accuracy due to increased mismatches caused by performing template matching along curved epipolar lines. Iida et al. addressed this by introducing a pseudo-bilateral filter that fuses region-based stereo matching with feature-based Structure from Motion (SfM), significantly improving accuracy by incorporating temporal image sequences[4]. Nevertheless, the approach faces limitations in processing speed due to high computational load. To overcome this, Arai et al. proposed a vertically arranged fisheye stereo camera setup, which linearizes epipolar lines and enables simplified stereo matching for improved distance estimation accuracy[5]. However, for obstacle detection purposes, it is not necessary to perform depth estimation across the entire measurement range. Sakuda et al. proposed an obstacle height estimation method that fits multiple planes to a disparity image, enabling flexible road surface estimation without strict planar constraints[6]. However, due to recursive processing involved in road surface estimation, the method struggles with real-time obstacle detection. Additionally, if the 3D reconstruction contains errors, the road surface itself may be falsely detected as an obstacle. To address this, Chikugo et al. proposed a method that incorporates intensity information to detect only obstacles while avoiding misdetection of road surfaces[7], [8]. However, since the method re-detects obstacles using high intensity values outside the previously estimated 3D obstacle region, it may falsely detect road markings or sunlit areas as obstacles[7]. Moreover, since the method uses intensity values from overlapping regions between image edges and 3D obstacle regions, it may fail in cases where road markings are misdetected or the obstacle lacks texture[8]. Therefore, we aim to develop an obstacle detection method that is robust to textureless environments by leveraging monocular depth estimation to detect obstacles on the road surface and conducting 3D measurements focusing only on these detected regions.

II. RELATED WORK

Methods for understanding surrounding environments can be broadly categorized into two approaches: those that detect obstacles based on estimated road surfaces, and those that utilize color or appearance features. One commonly used technique in the first approach is UV-disparity-based detection [9], [10]. However, these methods rely on stereo matching to obtain disparity information. As a result, they may fail to detect textureless obstacles due to the lack of sufficient matching features. Seki et al. proposed an approach using a projection matrix to infer obstacle locations [11]. Nevertheless, their method assumes a single planar surface, which leads to issues when dealing with scenes where road inclination varies. In the second approach, methods using

[1]Precision Engineering Course, Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, Japan. {chikugo, shiino,}@sensor.mech.chuo-u.ac.jp

[2]Faculty of Engineering, Shibaura Institute of Technology, 3-7-5 Toyosu, Koto-ku, Tokyo, Japan. pathak@shibaura-it.ac.jp

[3]Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, Japan. umeda@mech.chuo-u.ac.jp

978-1-6654-5784-2/26/$31.00 ©2026 IEEE

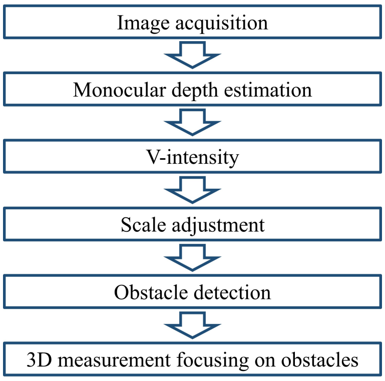


Image acquisition
Monocular depth estimation
V-intensity
Scale adjustment
Obstacle detection
3D measurement focusing on obstacles

Fig. 1: The flow of the proposed method

appearance features, deep learning-based approaches have become prominent [12], [13], [14]. While these methods achieve high performance in well-trained scenarios, they often suffer from limitations in generalization to environments different from the training data. Moreover, they tend to lack explainability, which can be a drawback in safety-critical applications.

III. PROPOSED METHOD

A. Overview

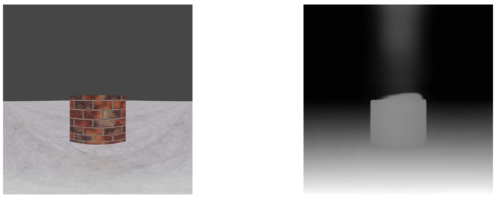
The proposed method is illustrated in Fig. 1. First, fisheye images are captured using two fisheye cameras, and they are converted into equirectangular images to reduce distortion. Next, relative depth information is estimated using a monocular depth estimation model, MiDaS [15]. Based on the obtained relative depth, a V-intensity image is generated to coarsely estimate the road surface. Here, V-intensity replaces the disparity in V-disparity [16] with brightness. Subsequently, stereo matching is applied only to points with high confidence, and the scale between MiDaS depth and stereo depth is adjusted using these points. Finally, obstacles on the road surface are detected, and depth estimation is performed only within the identified obstacle regions.

B. Image Acquisition

Fisheye images captured by fisheye cameras exhibit characteristic distortion. To reduce this distortion, the images are converted into equirectangular projections [3]. In stereo-based depth estimation using two cameras, it is generally assumed that the optical axes are perpendicular to the baseline for simplification. However, in practice, slight misalignments occur during camera installation, resulting in deviation from perfect orthogonality. Therefore, stereo rectification is performed using a checkerboard pattern to correct for these misalignments [5]. The rectification parameters for parallel alignment were obtained in advance using a checkerboard pattern.

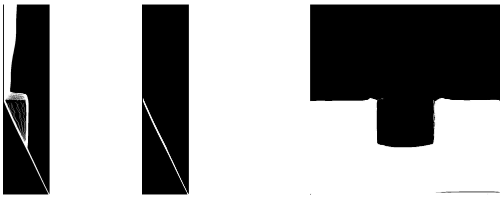
C. Monocular Depth Estimation

When using V-disparity based on disparity maps obtained from conventional stereo matching, accurate estimation of road surfaces and detection of obstacles becomes difficult in the presence of measurement errors or in textureless environments. In addition, processing speed is reduced due to the inclusion of unnecessary regions in the calculation. To address these issues, we utilize MiDaS, a monocular depth estimation model that can provide reasonably accurate environmental understanding. The result obtained using MiDaS for Fig. 2(a), is presented in Fig. 2(b). Although MiDaS performs processing across the entire image, it offers advantages over traditional stereo matching, such as faster inference and reduced sensitivity to small noise.



(a) Input image (b) Normalized depth

Fig. 2: Results of MiDaS



(a) (b) (c)

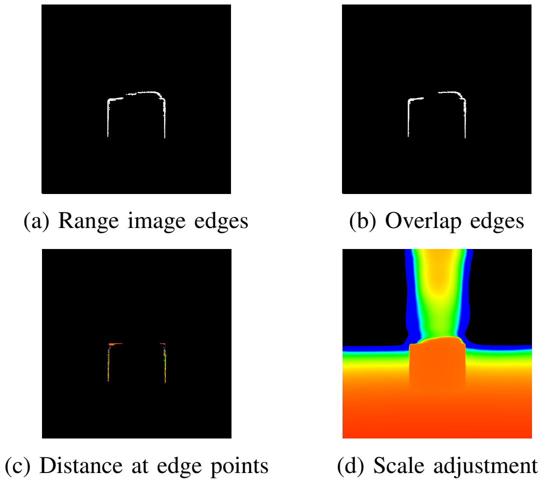
Fig. 3: V-intensity corresponding to Fig. 2. (a) V-intensity image, (b) Road region extracted from the V-intensity image, (c) Road region (shown in white) overlaid on the input image.

D. V-intensity

The relative depth values obtained using MiDaS are normalized to a range of 0 to 255. Therefore, the relative depth can be treated as image intensity. Based on this, we generate a V-intensity image, which represents the frequency of intensity values along the vertical axis of the relative range image, as shown in Fig. 3(a). In Fig. 3(a), the diagonal region corresponds to the road surface, while the vertical structures represent obstacles: the sky and distant background objects such as buildings. To extract the road surface region, we focus on the bottom of the diagonal region and define the area within a certain threshold from the bottom as the road surface. If the value exceeds the threshold, the corresponding region is assumed to be of similar size as neighboring road regions. The resulting extracted road surface is shown in Fig. 3(b). Fig. 3(c) shows the extracted road surface region projected back onto the input image.

E. Scale Adjustment via Stereo Matching

Since the depth value output by MiDaS is relative, it is not possible to directly obtain the absolute distance from the camera to the obstacle. Therefore, we perform scale adjustment by aligning the relative depth obtained from



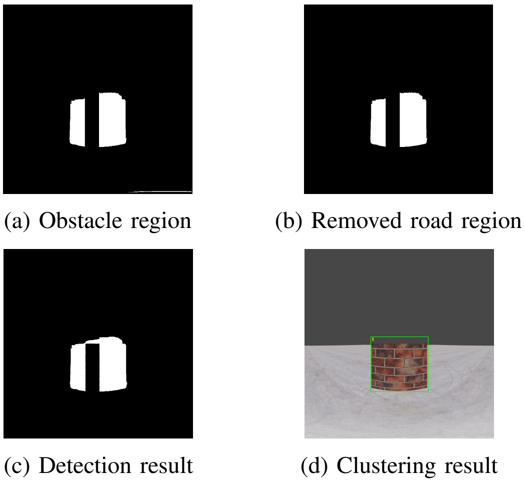
(a) Range image edges

(b) Overlap edges

(c) Distance at edge points

(d) Scale adjustment

Fig. 4: Scale adjustment between stereo and MiDaS



(a) Obstacle region

(b) Removed road region

(c) Detection result

(d) Clustering result

Fig. 5: Obstacle detection

MiDaS with the reliable absolute depth values obtained through stereo matching, focusing particularly on edge regions. If stereo matching is applied to the entire image, noise in texture-less regions can interfere with accurate scale estimation. Additionally, full-image processing can significantly reduce the processing speed. To address these issues, stereo matching is performed only at the points where edges in the relative range image and the input image overlap. Obstacles are expected to generate edges in the relative range image. Furthermore, edge-adjacent regions are likely to yield reliable depth through stereo matching. For this reason, we use edge information from the relative range image. Specifically, only edge pixels that are in contact with the estimated road surface are used for scale adjustment.

The edge image of the relative depth that contact with the road and the edge image used for stereo matching are shown in Fig. 4(a) and Fig. 4(b), respectively. Stereo matching is performed using block matching. In addition, texture filters are applied to reduce noise, and mismatches are handled appropriately. Mismatch correction is based on the disparity between corresponding pixels in the stereo image pair. Stereo matching is applied only to the overlapping edge points where the disparity between the relative depth edge and the input image edge is below a certain threshold. The result of edge-focused stereo matching is shown in Fig. 4(c). To estimate the scale, the average ratio between the stereo-matched absolute depth and the corresponding MiDaS-relative depth is computed across all matched pixels. This scale factor is then multiplied with the relative depth to convert it into absolute distance. The result of the scale-adjusted depth is shown in Fig. 4(d). In both Fig. 4(c) and Fig. 4(d), closer distances are shown in red, while farther distances appear in blue. In Fig. 4(d), some regions near the top of the image exhibit incorrect distance estimates, which are likely caused by the limited reliability of stereo matching in textureless or distant areas, resulting in inaccurate scale adjustment.

F. Obstacle Detection

To estimate the obstacle regions, Fig. 3(c) and Fig. 4(b) are used. If the number of pixels in Fig. 4(b) exceeds a threshold, the region between the edge and the road surface is considered an obstacle region. If the number of pixels is below the threshold, a similar process is performed using Fig. 3(c) and Fig. 4(a). Furthermore, if the number of pixels in Fig. 4(a) is also below the threshold, the non-road region in Fig. 3(c) that falls within the area in Fig. 4(d) where the distance is determined (indicated in color) is considered the obstacle region. The detected obstacle region from this process is shown in Fig. 5(a). However, in the lower right of Fig. 5(a), a road area is mistakenly detected as an obstacle region. Therefore, noise removal is applied to each detected obstacle region using the neighboring road parameters. These road parameters are estimated using the least squares method based on the scaled MiDaS depth. By applying the road plane estimated from the obtained road parameters, the road surface around each obstacle region is determined, and regions that are farther than a threshold from this road plane are classified as obstacles, while others are considered part of the road and removed from the obstacle regions. This approach enables robustness to sloped surfaces. The result after this processing is shown in Fig. 5(b), where the road region mistakenly detected as an obstacle in the lower right is successfully removed. After noise removal, the scaled MiDaS depth and the brightness of the input image near the obstacle region are analyzed, and if they are similar, those points are added to the obstacle region. The result of this process is shown in Fig. 5(c). Finally, clustering is performed on the obstacle regions by considering the connectivity in the binary image, as shown in Fig. 5(d).

G. 3D Measurement Focused on Obstacles

For each clustered obstacle, the distance from the camera and the height of the obstacle are estimated. To estimate the distance from the camera, the distances obtained from stereo matching and MiDaS are compared. Then, the height is estimated using the distance from the camera and the road surface parameters. For the stereo-based distance, Fig. 4(c) is used. From Fig. 4(c), some points, five in the following experiments, with the shortest distances are selected, and their average is taken as the stereo-based distance from the camera. For the MiDaS-based distance, Fig. 4(b) and Fig. 4(d) are used. From the edge areas shown in Fig. 4(b), some points with the shortest distances are selected, and their average is taken as the MiDaS-based distance. When comparing the two distances, if the difference between them is below a threshold, it is considered that both have high reliability. In such cases, considering safety, the smaller of the two distances is selected. This is based on the idea that selecting the closer distance allows for faster response in obstacle avoidance. If the difference between the two distances exceeds the threshold, it is assumed that at least one of them has low reliability. Since the stereo-based distance is calculated as the average of some points, it may contain noise due to mismatches, leading to a large difference from the MiDaS-based distance. On the other hand, since the MiDaS-based distance uses edge regions from Fig. 4(b) and tends to contain less noise, its reliability is considered higher when mismatched points are included in the stereo matching. Therefore, if the difference between the two distances exceeds the threshold and the stereo-based distance is shorter than the MiDaS-based distance, the MiDaS-based distance is adopted as the final distance from the camera.

To estimate the height of an obstacle, the heights of the upper edge points obtained from Fig. 4(b) are calculated, and their average is taken as the height of the obstacle. If the equation of the road surface plane is given as

$$z = ax + by + c, \tag{1}$$

then the height H of a point (x_0, y_0, z_0) from the road surface can be computed by the following equation:

$$H = \frac{|ax_0 + by_0 + c - z_0|}{\sqrt{a^2 + b^2 + 1}}. \tag{2}$$

Here, a, b, and c represent the road surface parameters. These parameters are obtained using the least squares method based on the distances in the vicinity of each obstacle region.

IV. Accuracy Evaluation Experiment

A. Experimental Conditions

In this experiment, a virtual environment was created using the 3DCG software Blender. We verified whether the proposed method can detect obstacles without misidentifying the road surface. Additionally, we evaluated the error in the estimated distance from the camera to the obstacle, the error in obstacle height, and the processing speed. In Blender, the Cycles render engine was used, and the lens was set to an equirectangular panoramic projection. The resolution of the camera was set to 1048×1048 pixels, with both horizontal and vertical fields of view set to 180°. The baseline length between the two cameras was set to 0.072 m. Since the images obtained from Blender are ideal, both the internal and external camera parameters can be assumed to be error-free. The experimental environment is shown in Fig. 6 and Fig. 7. Fig. 6 shows the overview of the experimental setup, while Fig. 7 presents the experimental conditions. The elevation angle in the image was fixed at 0°, and square block was placed at azimuth angles of $-60°$, 0°, and 60° as measurement positions. The block had a side length of 1 m, and the distance from the camera to the block was measured at five locations: 1 m, 3 m, 5 m, 7 m, and 9 m.

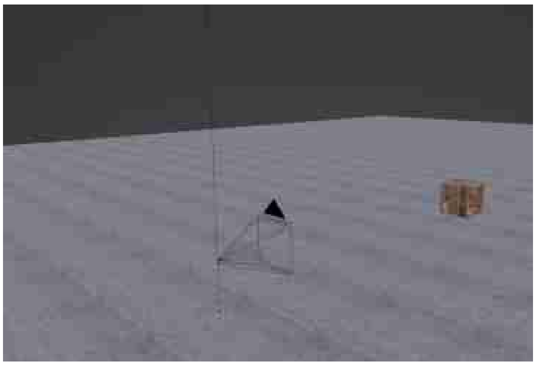


Fig. 6: Overview of the experimental setup

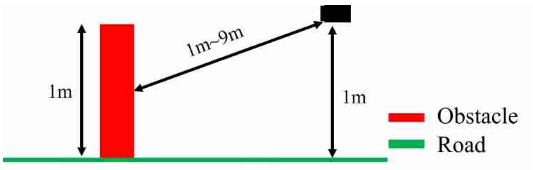


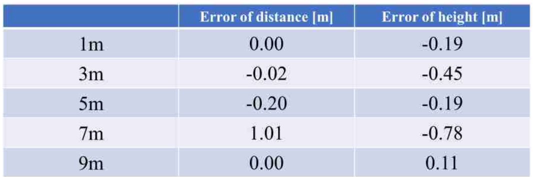
Fig. 7: Experimental conditions

B. Experimental Results

The experimental results are shown in Tables I to IV. Tables I to III present the errors in the estimated distance from the camera to the obstacle and in obstacle height at azimuth angles of $-60°$, 0°, and 60°, respectively. Table IV shows the average processing time for each processing step and the overall average processing speed.

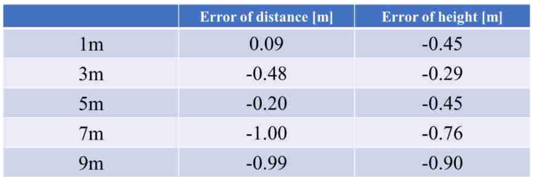
From the results, it can be confirmed that in all conditions, the proposed method was able to detect the target block without misidentifying the road surface. The error in the estimated distance from the camera to the obstacle ranged from a minimum of 0 m to a maximum of 5.90 m. For example, in cases such as the 1 m distance at 0° azimuth and the 9 m distance at 0° azimuth, the estimation error was 0 m, indicating high accuracy. However, in some cases, the error exceeded 1 m, showing that the method was not robust under all conditions. This is likely due to the distance estimation method described in Section III-G. The method calculates the average of the five closest distance values obtained by stereo matching. Therefore, if even one of those points has a large error, the final average error will also be large. Furthermore, since MiDaS was trained on perspective projection images rather than equirectangular images, there is a possibility of increased error when the distance from the camera is estimated using MiDaS.

TABLE I: Distance and height errors at 0° azimuth



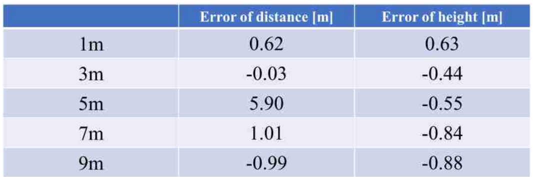
	Error of distance [m]	Error of height [m]
1m	0.00	-0.19
3m	-0.02	-0.45
5m	-0.20	-0.19
7m	1.01	-0.78
9m	0.00	0.11

TABLE II: Distance and height errors at $-60°$ azimuth



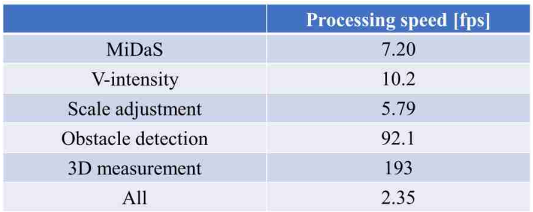
	Error of distance [m]	Error of height [m]
1m	0.09	-0.45
3m	-0.48	-0.29
5m	-0.20	-0.45
7m	-1.00	-0.76
9m	-0.99	-0.90

TABLE III: Distance and height errors at 60° azimuth



	Error of distance [m]	Error of height [m]
1m	0.62	0.63
3m	-0.03	-0.44
5m	5.90	-0.55
7m	1.01	-0.84
9m	-0.99	-0.88

TABLE IV: Processing speed



	Processing speed [fps]
MiDaS	7.20
V-intensity	10.2
Scale adjustment	5.79
Obstacle detection	92.1
3D measurement	193
All	2.35

The error in obstacle height ranged from a minimum of 0.11 m to a maximum of -0.90 m. While the method showed good accuracy in some cases, in most cases the errors were relatively large. This is likely because the road surface parameters near the obstacle were estimated using least squares based on the distance values at the lower part of the image. In detection results like the one shown in Fig. 8, the distance values within the obstacle region are also used to estimate the road surface, which prevents accurate estimation. As a result, the estimated obstacle heights tend to be lower than the actual heights in most situations.

According to Table IV, the overall processing speed was 2.35 fps, which is insufficient for autonomous driving applications. The most time-consuming process was scale adjustment, particularly the stereo matching step within that process.

V. Conclusion

In this paper, we proposed an obstacle detection method focusing on obstacles on the road surface using a fisheye stereo camera and the monocular depth estimation model MiDaS. The experiments demonstrated that obstacles could be detected without misidentifying the road surface. However, sufficient accuracy in measurement and processing speed were not achieved.

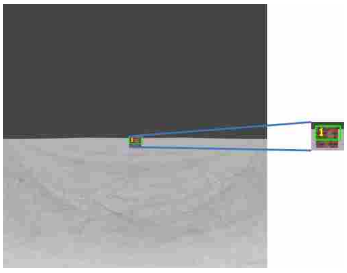


Fig. 8: Detection result at 7m and 0°

As future work, we plan to conduct experiments in textureless environments and real-world scenarios. We will also perform comparative experiments with conventional methods. Furthermore, by incorporating feature points into stereo matching, we aim to improve both measurement accuracy and processing speed.

References

[1] Jau-Woei Perng, Pei-Yu Liu, Kai-Quan Zhon and Ya-Wen Hsu, “Front object recognition system for vehicles based on sensor fusion using stereo vision and laser range finder,” *2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*, pp.261–262, 2017.

[2] Nobuo Sasaki, Naoyuki Iijima and Daiki Uchiyama, “Development of ranging method for inter-vehicle distance using visible light communication and image processing,” *2015 15th International Conference on Control, Automation and Systems (ICCAS)*, pp.666–670, 2015.

[3] Akira Ohashi, Fumito Yamano, Gakuto Masuyama, Kazunori Umeda, Daisuke Fukuda, Kota Irie, Shuzo Kaneko, Junya Murayama and Yoshitaka Uchida, “Development of ranging method for inter-vehicle distance using visible light communication and image processing,” *Stereo rectification for equirectangular images*, pp.535–540, 2017.

[4] Hirotaka Iida, Yonghoon Ji, Kazunori Umeda, Akira Ohashi, Daisuke Fukuda, Shuzo Kaneko, Junya Murayama and Yoshitaka Uchida, “High-accuracy Range Image Generation by Fusing Binocular and Motion Stereo Using Fisheye Stereo Camera,” *2020 IEEE/SICE International Symposium on System Integration (SII)*, pp.343–348, 2020.

[5] Hikaru Chikugo, Kento Arai, Sarthak Pathak, and Kazunori Umeda, “Fisheye Stereo Camera Using Fisheye Vertical Stereo Method,” in Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2024, pp. 3512–3518.

[6] Tomoyu Sakuda, Hikaru Chikugo, Kento Arai, Sarthak Pathak and Kazunori Umeda, “Estimation of Road Surface Plane and Object Height Focusing on the Division Scale in Disparity Image Using Fisheye Stereo Camera,” *Journal of Robotics and Mechatronics*, vol.35-5, pp.1354–1365, 2023.

[7] Hikaru Chikugo, Kento Arai, Sarthak Pathak, and Kazunori Umeda, “Detection and Height Estimation of Obstacles Using Disparity and Brightness Information with Fisheye Stereo Camera,” *Proceedings of the Robotics and Mechatronics Conference*, 2023, pp.2A1–D07.2023. (in Japanese)

[8] Hikaru Chikugo, Tomoyu Sakuda, Sarthak Pathak and Kazunori Umeda, “Obstacle Detection and Height Estimation Using Fisheye Stereo Camera Considering Intensity Information,” *2024 IEEE/SICE International Symposium on System Integration (SII)*, pp.147–152, 2024.

[9] Mingguo Liu, Chenxing Shan, Haofeng Zhang and Qingyuan Xia, “Stereo Vision Based Road Free Space Detection,” *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, pp.272–276, 2016.

[10] Wenjie Song, Yi Yang, Mengyin Fu, Fan Qiu and Meiling Wang, “Real-Time Obstacles Detection and Status Classification for Collision Warning in a Vehicle Active Safety System,” *IEEE Transactions on Intelligent Transportation Systems*, vol.19-3, pp.758–773, 2018.

[11] Mingguo Liu, Chenxing Shan, Haofeng Zhang and Qingyuan Xia, “Robust obstacle detection in general road environment based on road extraction and pose estimation,” *2006 IEEE Intelligent Vehicles Symp*, pp.437–444, 2006.

[12] Peiliang Li, Xiaozhi Chen and Shaojie Shen, “Stereo R-CNN based 3D Object Detection for Autonomous Driving,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.7636–7644, 2019.

[13] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng and Xiaogang Wang, “GS3D: An Efficient 3D Object Detection Framework for Autonomous Driving,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1019–1028, 2019.

[14] Changxin Zhou, Yazhou Liu, Quansen Sun and Pongsak Lasang, “Vehicle Detection and Disparity Estimation Using Blended Stereo Images,” *IEEE Transactions on Intelligent Vehicles*, vol.6-4, pp.690–698, 2021.

[15] Reiner Birkl and Diana Wofk and Matthias Müller, “MiDaS v3.1 – A Model Zoo for Robust Monocular Relative Depth Estimation,” arXiv preprint arXiv:2307.14460, 2023.

[16] Raphael Labayrade, DidieI Aubert, and Jean-Philippe Tarel, “Real time obstacle detection in stereovision on non flat road geometry through ”v-disparity” representation,” *Intelligent Vehicle Symposium, 2002. IEEE*, vol.2, pp.646–651, 2002.