Multimodal Object Detection and Ranging Based on Camera and Lidar Sensor Fusion for Autonomous Driving

Danish Khan[1], Minjin Baek[1], Min Young Kim[2] and Dong Seog Han[2,*]
[1]Center for ICT and Automotive Convergence, Kyungpook National University, Daegu, South Korea
[2]School of Electronics Engineering, Kyungpook National University, Daegu, South Korea
danish@knu.ac.kr, mbaek@knu.ac.kr, minykim@knu.ac.kr, dshan@knu.ac.kr

***Abstract*—A robust perception system is critical in autonomous driving. It is responsible for object detection, classification, and ranging under challenging circumstances. Camera and lidar sensors provide complementary information, and by combining these two modalities, we can increase the robustness and accuracy of the overall perception system. This paper presents the implementation of sensor fusion based perception using camera images and lidar point clouds for object detection and ranging in a real-time driving environment. The experiment results obtained with our test vehicle demonstrate that the perception of vehicle surroundings can be more effectively achieved by means of camera-lidar sensor fusion compared with using a single type of sensor.**

***Index Terms*—camera, lidar, sensor fusion, perception, object detection, ranging, autonomous driving**

I. Introduction

Object detection by a single type of sensor is not sufficient to ensure safety for automated driving systems. Every sensor has its advantages and shortcomings. Identification of color, shape, and type of obstacles is relatively easy with cameras compared with other sensors. But generally, cameras cannot capture depth or range information. Methods are available to recover 3D information from camera images. However, this is the fundamental problem with the cameras. In contrast, lidars can depict the 3D surroundings of the vehicle more accurately, often with a much larger field of view.

The lidar is less prone to weather-related conditions and is not affected by ambient light variations compared with the camera. Lidar data contain range information, but lidar-based detection of objects is not straightforward. First, the 3D lidar points that correspond to objects are not as dense as how the objects appear in the camera image. Second, the variation of geometric shapes is much higher. The accuracy of the geometric shape determined based on lidar measurements can largely vary for objects located at different distances because the density of the 3D points decreases as the distance to the object increases. 3D object detection is very challenging in such cases. Considering aforementioned advantages and disadvantages of each type of sensor, it is highly beneficial to use both sensors together for the high level of perception accuracy and robustness that is required for autonomous driving.

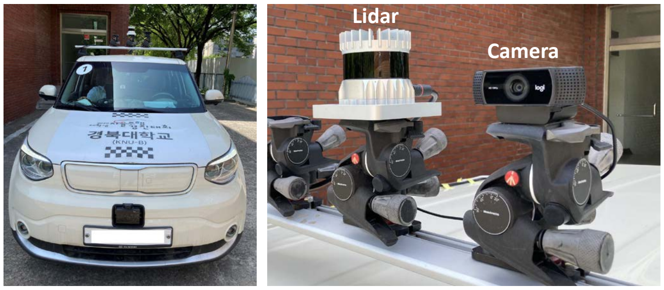

Fig. 1. Sensor configuration for the multimodal object detection and ranging.

In this paper, the implementation of multimodal object detection and ranging in a real-time driving environment is presented. The camera-lidar calibration process as well as the sensor fusion based approach for perception are described.

II. Camera-Lidar Calibration

The object detection and ranging setup consists of a 128-channel lidar (Ouster OS1-128) and a camera (Logitech C922). The camera and lidar sensors are mounted on the roof rack of the test vehicle as shown in Fig. 1. The camera-lidar calibration is the process of estimating the relative position and orientation of one sensor with respect to the other, which yields calibration parameters that can be used to transform the sensor measurement data into one unified coordinate system.

To estimate the rotation and translation between the camera and lidar, optical and geometrical characteristics of the camera such as focal length, principal point, and distortion coefficient are required. These characteristics are attributed to camera intrinsics and can be estimated by using a camera intrinsic calibration method. We used Zhang's method [1] in which a checkerboard target is used to estimate a set of feature points. These feature points are then related at different view points to estimate the intrinsic matrix. Once the intrinsic matrix is obtained, the extrinsic parameters of the camera and lidar are acquired by using the perspective-n-point (PnP) algorithm. The PnP algorithm minimizes the reprojection error between the 3D lidar points and their corresponding points in the 2D

[*]Corresponding author: Dong Seog Han.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A1A03043144).

978-1-6654-9927-9/22/$31.00 ©2022 IEEE

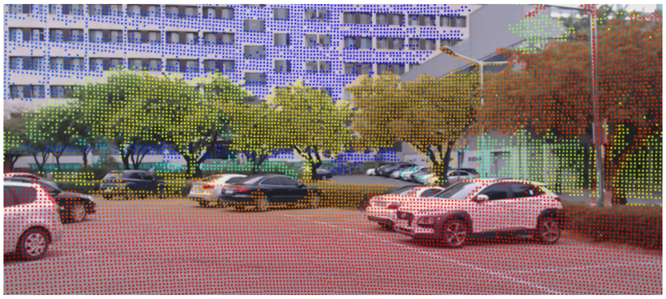

Fig. 2. Projection of the 3D lidar points on the 2D camera image.

camera image to estimate the pose. The 2D-3D corresponding points were carefully selected based on the reflectivity map of the lidar measurements from multiple planer checkerboard targets. We used multiple targets at different locations because higher calibration accuracy can be achieved compared with using a single plane [2]. Fig. 2 shows the 3D lidar points mapped onto their corresponding 2D image pixels using the transformation matrix acquired by the aforementioned camera-lidar calibration method.

III. Object Detection and Ranging

The camera and lidar fusion steps for the multimodal object detection and ranging approach described in this paper are illustrated in Fig. 3. For object detection based on camera images, we used the YOLOv3 object detector [3], which provides high real-time performance because it does not require a region proposal network and directly regresses to detect objects. For 3D object detection based on lidar measurements, we used PointPillars [4], which is based on the PointNet [5] encoder and extracts local and global features from the 3D point clouds. PointPillars converts the 3D points into pillar representation hence there is no need to perform vertical binning that is required in other representations such as voxels. Using the calibration parameters, we transformed the data obtained from two different sensors into one coordinate system by projecting the point clouds onto the images.

The Robot Operating System (ROS) and Autoware [6] were utilized for the implementation of the multimodal object detection and ranging system. For our experiment conducted in real-time driving conditions, we used a computer equipped with Intel Core i9-9900K CPU, Nvidia Titan RTX, and 32 GB of memory. For the purpose of the preliminary performance evaluation, the weights pretrained with the Microsoft COCO dataset [7] were used for YOLOv3 (with Darknet-53 as the backbone network), and the weights pretrained on the KITTI dataset [8] were used for PointPillars. The experiment result as shown in Fig. 4 illustrates successful detection and ranging based on the sensor fusion approach described in this paper.

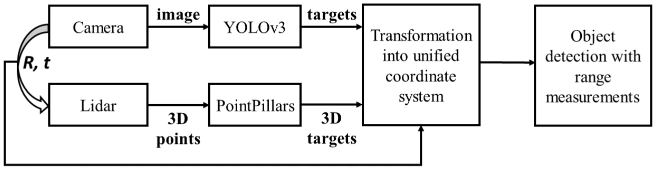
Camera
2D image
YOLOv3
2D targets
R, t
Lidar
3D points
PointPillars
3D targets
Transformation into unified coordinate system
Object detection with range measurements

Fig. 3. Flow diagram of the multimodal object detection and ranging

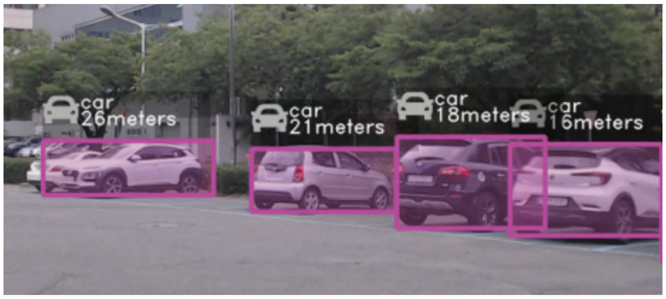

Fig. 4. Experiment result of object detection and ranging.

IV. Conclusion

In this paper, we presented the implementation of the multimodal object detection and ranging system for real-time perception of the driving environment. The camera and lidar sensors were carefully calibrated so that the 3D lidar points can be accurately projected on the 2D camera image. The results from the preliminary driving tests demonstrated that the camera-lidar fusion approach enables accurate and reliable detection and ranging and that the two sensors effectively complement one another. Our future work will include optimization of the detection models and training on larger datasets. We also plan to use early fusion techniques to reconstruct the pixel-wise depth from sparse lidar points and dense 2D images. Additionally, we will extend this study to cooperative perception via vehicular wireless communications.

References

[1] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[2] F. Youyang, W. Qing, Y. Yuan, and Y. Chao, "Robust improvement solution to perspective-n-point problem," *International Journal of Advanced Robotic Systems*, vol. 16, no. 6, pp. 1–15, 2019.

[3] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[4] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[5] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 652–660.

[6] S. Kato, E. Takeuchi, Y. Ishiguro, Y. Ninomiya, K. Takeda, and T. Hamada, "An open approach to autonomous vehicles," *IEEE Micro*, vol. 35, no. 6, pp. 60–68, 2015.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[8] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.