STEREO CAMERA AND SOLID STATE LIDAR OBJECT DETECTION ON SMALL DELIVERY ROBOT

Jiarun Wei[1,†], **Ding Zhao**[2,†]

[1]Carnegie Mellon University, Pittsburgh, PA
[2]Carnegie Mellon University, Pittsburgh, PA

ABSTRACT

Stereo camera is popular in object detection due to its dense data. It works by matching the key points of the left and right images, using the geometric constraints to calculate the depth map, and then feed these into a neural network. Due to the geometric constraints, it only works for a certain range of objects and is unstable for some objects. Here we present a demo to stabilize the detection using solid state lidars, which is a very cost efficient solution especially for small and slow robots.

Keywords: Object detection, Computer Vision, Solid-state Lidar, Autonomous Delivery

1. INTRODUCTION

In the industry of autonomous delivery, it is widely proved that the perception part of the autonomous system is the bottle neck. Due to the complicated environment of the real world, it is hard for the vehicle to navigate around. The vehicle it self must get a very detailed and precise information of what the surrounding is like.

In order to achieve that, 2 types of sensors are invented, the cameras and the lidars, both of which uses the lights to percept the surrounding. However, both camera and lidars have pros and cons are always used together to get rid of the disadvantages of each other. The camera is similar to human's eyes. It captures the lights reflected from the surrounding objects projected on a 2D plane. However the reflectional lights can be confusing because the image captured by the camera and the real environment is not necessary one-to-one mapping. On contrary, the lidars launch lights it self and calculates the distance according to the time interval between the launch time and the receiving time. This is very precise in terms of the depth measurement, however lacks of semantics information due to the absence of colors.

Great efforts of the developmet of the lidars has been made to achieve a higher perception accuracy. For example, the velodyne lidar is the most important sensor. It is used to detect objects in the environment with multiple threads. However, due to the demand of high speed of real vehicles, the density of its threads are required to be very high. For small delivery robots, many researchers directly deploys the velodyne lidar, which might not be necessary given the speed and height of those small vehicles.

In this paper, a novel method to utilize a very cost efficient solid state lidar combined with stereo camera to stabilize the detection of objects on small delivery robots is proposed. We also established a prototype of the small autonomous delivery robot system and demonstrate the system's performance. This robot can navigate across within certain area within CMU campus given arbitrary initial position and target position. It is about half a person tall with omnidirectional wheels that can carry a small parcel.

2. SYSTEM DESIGN

As is shown in figure 3, the delivery robot can be divided into 6 parts: stereo camera, router, battery, solid-state lidar and the processor. The part list is shown in table 1.

Stereo camera The stereo camera is used for the detection of dynamic obstacles such as human and vehicles. Though the depth obtained by the stereo camera is not as accurate as the lidar, its object detection ability is a lot more better due to the RGB triple input channels.

Solid-state lidar The lidar is used in multiple tasks: mapping, localization and object detection. During the mapping task, the lidar scans around a certain area and reconstructs a 3D point cloud map of the surroundings. In the localization task, the lidar scans forward and locate the robot using a Normal Distribution Transform matching method, which matches the distribution of point cloud of previous map and the current scanning. In the object detection task, the lidar gets the semantics of the current scanning.

Processor The processor is used to process the data obtained by the lidar and the stereo camera. The processor uses a neural network that can process the data and generate the semantics of the current scanning.

Battery There are 2 types of batteries, one is the chassis battery which is used to power the chassis and move the whole vehicle.

[†]Joint first authors

ID	Category	Part	Qty
SAFEAI001	Joystick Controllers	Mecanum Wheel Controller	1
SAFEAI002	Joystick Controllers	Regular Wheel Controller	1
SAFEAI003	LiDARs	Solid-state Livox Mid-70 Lidars	3
SAFEAI004	LiDARs	SLAMTec Lidars	2
SAFEAI005	Cameras	ZED Camera	2
SAFEAI006	Power Supplies	Krisdonia Power Bank	1
SAFEAI007	Power Supplies	MAXOAK Power Bank	2
SAFEAI008	Computers	AGX Jetson	2
SAFEAI009	Internet communications	Router	2

TABLE 1: PART LIST

Another is the power banks which are used to power the processor, lidars, and the router. The rated power of the chassis and the power banks are different so they have to be separated.

Router The router is used to connect the lidar and the processor. It plays an important role in the synchronization process between the lidar and the processor.

3. METHODS

Stereo R-CNN based 3D Object Detection [1] is deployed in the system. It can be decomposed into several parts: feature extraction, reigion proposal, reigion of intersection alignment, keypoint prediction and 3D bounding box estimation, which involves deep neural networks such as ResNet and binocular geometric constraints. In the following experiment, the lidar point clouds are used to verify the detection.

Feature extraction The ResNet [2] is one of the most commonly used feature extraction methods. Its architecture is shown in figure 1. Unlike the regular sequential CNNs, the ResNet adds the forward propagation from the input of each blocks directly to their output.

Reigion proposal The Faster R-CNN network[3] is used to propose the regions of interest. It can be divided into 3 steps: anchor generation, foreground-background classification and bounding box regression. The anchors are generated in the feature space of the output of the feature extraction backbone. They are mapped back to the image regions according to the aspect ratios and scales, which are all trainable parameters.

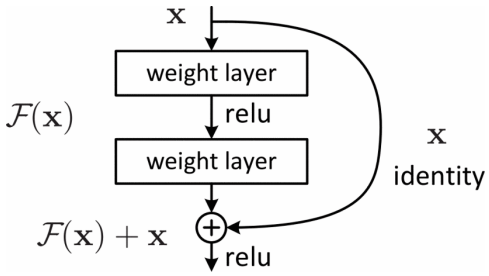

FIGURE 1: RESNET BLOCK

Region of intersection alignment Given the ground truth bounding boxes, the candidate boxes are further classified into foreground or background using the overlapping ratio with the ground truth as criterion. The ground truth is a heavy work which requirs a lot of human resources.

Keypoint prediction and 3D bounding box estimation After the region of intersection is calculated, the network can do a regression to estimate the position and size of the bounding boxes using backpropagation.

4. EXPERIMENT

The over all work flow is shown in figure 2. Before the actual experiment begins, the stereo camera is calibrated using the intrinsic matrix K and the distortion coefficients DC, which are defined as:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

Where f_x and f_y are the focal length along x and y axis. c_x and c_y are the offsets of the principal axis.

$$DC = \begin{bmatrix} k_1 & k_2 & p_1 & p_2 & k_3 \end{bmatrix} \tag{2}$$

Such that the radio and tangential distortion are given as:

$$\begin{aligned} x_r &= x\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) \\ y_r &= y\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) \\ x_t &= x + \left[2p_1 xy + p_2\left(r^2 + 2x^2\right)\right] \\ y_t &= y + \left[p_1\left(r^2 + 2y^2\right) + 2p_2 xy\right] \end{aligned} \tag{3}$$

The corresponding parameters' values are shown in table 2.

Perception Besides the Solid-state lidar and the stereo camera, which is introduced in the previous section, there are 2 more sensors: the IMU and the motor encoder. The IMU is used to measure the dynamic properties of the robot. For example the acceleration and the angular acceleration. The veolcity and the estimated position can be obtained by integrating the acceleration. It should be noticed that the position obtained here can not be used directly because there is drifts due to the intrinsic noise of the IMU, hence the point cloud matching localization of the lidar is necessary. The motor encoder is used to measure the rotational speed of the wheels so that the commanded speed can be maintained using a PID control policy.

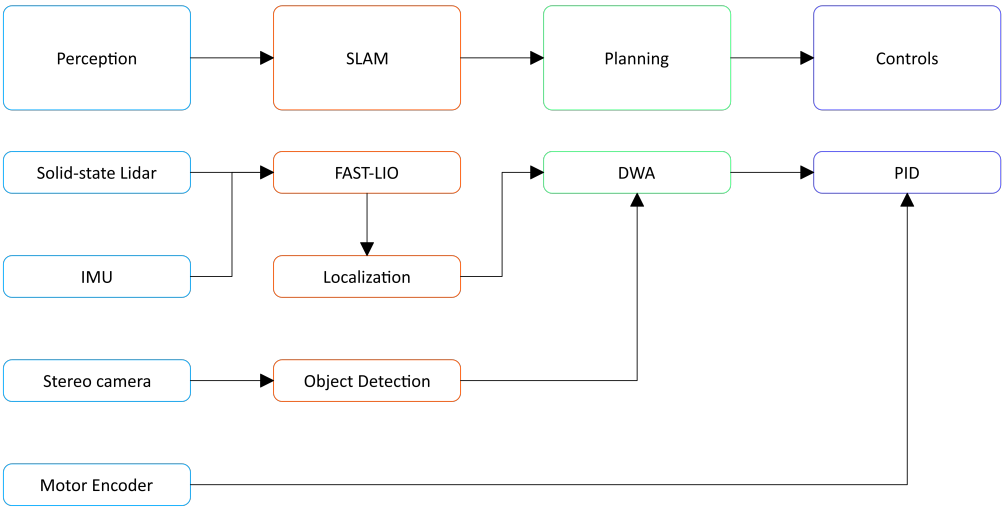

FIGURE 2: FLOW CHART

	Left lence	Right lence
f_x	525.575	528.400
f_y	525.1050	527.8600
c_x	355.7350	627.9750
c_y	350.3825	370.8590
k_1	-0.0408	-0.0453
k_2	0.0081	0.0140
p_1	0.0000	0.0000
p_2	0.0000	0.0000
k_3	-0.0041	-0.0041

TABLE 2: CAMERA INTRINSIC PARAMETERS

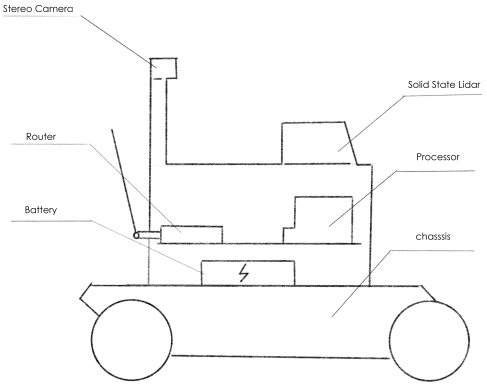

FIGURE 3: SYSTEM DESIGN

Simultaneously Localization and Mapping There are 2 sensors in charge of the localization process: the IMU and the lidar. The IMU is stable as long as there is no strong magnetic field around. It will never be influenced by the surrounding obstacles. However, it will have large drift upon small errors of the accelerometer. On the other hand, the lidar point cloud matching method will correct it self when more points are detected. However, the lidar is very sensitive with respect to the envirnonment. Given the pros and cons of these 2 sensors. The final localization is done by merging these 2 sensors' data using the Kalmen filter. Figure 5 shows the cost map of the route between CMU's Hunt library and gym.

Planning The planning is divided into global planning and local planning. The global planning is done by using the A* algorithm and the local planning is done by DWA algorithm. The A* algorithm is a breath first search based method with weights added to each steps of the exploration. The Dynamic Window Approach is a local planning method that searches the robot control space u, where:

$$u = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \end{bmatrix} \tag{4}$$

Control The control is done by using the PID control policy. The PID control policy is a feedback control policy that is used to maintain the speed of the robot.

5. RESULTS

Figure 6 shows 2 views of the stereo camera object detection result of a person. The results are shown in blue bounding box with green vertices. The strips in the figures are the lidar point cloud. It is shown that the person in the bounding box is closely surrounded by the strips, which shows that the lidar point cloud overlaps with the camera object detection. Figure 4 shows the planning result of the vehicle. The green line is the global planning path, the red line is the local planning path and the blue line is the localization.

According to the figures, the stereo camera's object detection

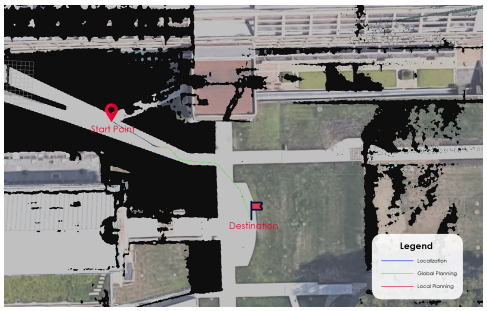

FIGURE 4: COST MAP

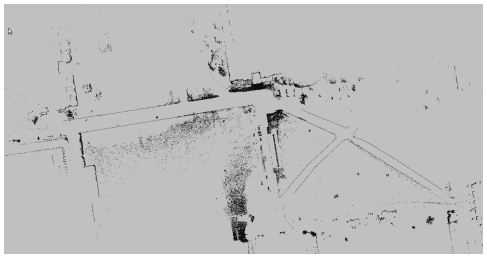

FIGURE 5: COST MAP FOR ANOTHER ROUTE

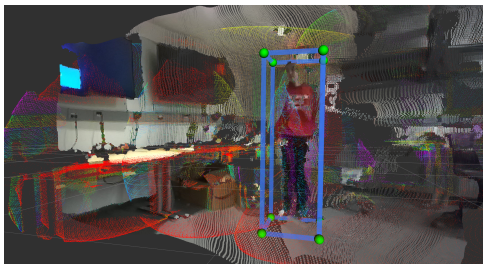

FIGURE 6: OBJECT DETECTION VERIFIED BY LIDAR

is relatively accurate even under indoor envirnonment. The lidar's point cloud is also shown to have a very good alignment with the depth map given by the stereo camera.

ACKNOWLEDGMENTS

Thanks to Professor Ding and all members in the team of Safe AI Scout at Safe AI Lab, the system of our prototype is developed in a very smooth way. I am very grateful to the members of the team not only for their hard work and dedication but also for their suggestions and feedbacks.

The time at Safe AI Lab is very valuable. Not only did I established the software and hard ware system for an autonomous delivery robot, but also I got the opportunity to learn the techniques in order for a real world system to work. For example the software environment and the hardware environment are very important for the software development. Sometimes we spent days on the compatibility between software packages. I also learned that the cooling system is very important for the performance of the robot.

I also want to say thank you to my mentors for their high level guidance for what the whole architecture of an autonomous robot will be like. This gives us a very clear idea of how to build a robot at the very beginning of our development.

REFERENCES

[1] Li, Peiliang, Chen, Xiaozhi and Shen, Shaojie. "Stereo R-CNN based 3D Object Detection for Autonomous Driving." *CoRR* Vol. abs/1902.09738. URL 1902.09738, URL http://arxiv.org/abs/1902.09738.

[2] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing and Sun, Jian. "Deep Residual Learning for Image Recognition." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: pp. 770–778. 2016. DOI 10.1109/CVPR.2016.90.

[3] Ren, Shaoqing, He, Kaiming, Girshick, Ross B. and Sun, Jian. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *CoRR* Vol. abs/1506.01497. URL 1506.01497, URL http://arxiv.org/abs/1506.01497.

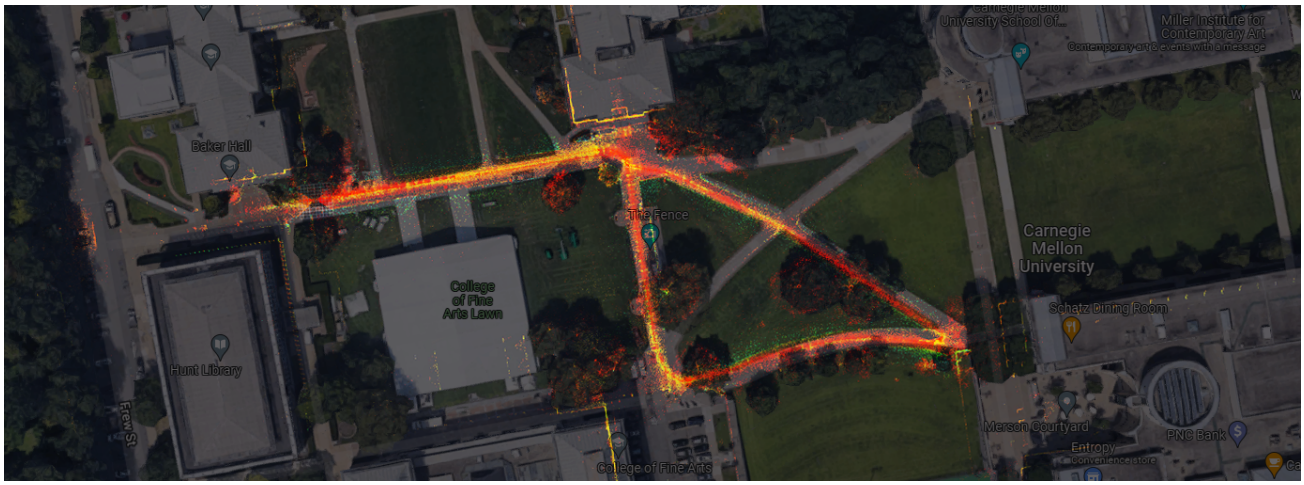

FIGURE 7: COMPARISON WITH GOOGLE MAP

APPENDIX A. CODES FOR CAMERA LIDAR CALIBRATION

```
import pyzed.sl as sl # Import the stereo camera's library
import cv2 # Import OpenCV
init_params = sl.InitParameters() # Create a sl.InitParameters object
init_params.camera_resolution = sl.RESOLUTION.HD2K # Use HD2K resolution
zed = sl.Camera() # Create a camera object
zed.open(init_params) # Open the camera
if not zed.is_opened(): # Check if camera opened successfully
  exit()
print("Opening ZED Camera...") # Remind the user that the camera is opening
zed.grab(init_params) # Grab a new image
image = sl.Mat() # Create a sl.Mat object

zed.retrieve_image(image) # Retrieve the left image
cv2.imwrite("test.png", image.get_data()) # Save the image
```