3D-Lidar and RGB Camera Fusion and Semantic Segmentation

Sreehari Premkumar
ECE. Masters in Robotics
Northeastern University
Boston, United States of America

***Abstract*—This paper presents a technique for projecting 3D-lidar points onto an RGB camera image using the intrinsic parameters of the camera and extrinsic parameters while also using spherical projection. The approach involves transforming the 3D-lidar point cloud into the camera coordinate system and then projecting it onto the image plane using the camera's intrinsic parameters. The resulting 3D reconstruction of the scene is validated using real-world datasets and found to be accurate. Alongside, Semantic segmentation of the RGB image using a pre-trained model of DeeplabV3 is also performed and was compared with the projected lidar points for a more comprehensive understanding of the scene. This approach has several practical applications like autonomous driving, robotics and surveillance system.**

***Index Terms*—Autonomous Navigation, Semantic Segmentation, Lidar Camera Fusion, Intrinsic-Extrinsic Calibration**

I. Introduction

In autonomous navigation and perception, lidars and cameras are two essential sensors that work together to provide rich semantic information for safe navigation and 3D object detection. The key advantage of using cameras is their ability to capture dense and coloured representations of the environment, making it easier to identify objects such as pedestrians and traffic signs. Lidars, on the other hand, excel at extracting depth information of 3D point cloud data, which is difficult to measure using a standard RGB cameras. However, aligning the lidar points with the RGB camera image can be challenging due to the differences in their coordinate systems and calibration parameters. But by combining information from both sensors, we can overcome the limitations of each and benefit from their individual strengths. And they have some redundant information, which is always useful when fusing two different sensors.

II. Related Works

Several previous studies have explored the use of multi-modal sensor fusion for precise estimation and mapping of the features. In one of the research papers, they used RGB and Lidar based segmentation[1], where they used polar grid representation obtained after spherical projection as input to 2 different kinds of neural network architecture named SqueezeSeg and PointSeg. In this a 10% increase in the Intersection over Union (IoU) score was observed compared to only lidar. But the shortcomings were that the segmentation occured directly in the lidar Point Cloud and it is very difficult to label Point cloud data for the trainng o neural nets. Moreover there is less number of features for segmentation in point cloud data, as well as the data is sparse in nature. In one of the other notable paper PointPainting: Sequential Fusion for 3D Object Detection[2], they directly segment on the RGB image and the try to reproject the camera points into lidar domain and fuse it with the lidar data. Now this data can be used for any lidar method. Their experiments showed improvement on Point-RCNN, VoxelNet on the KITTI dataset. This paper was a major factor in my project. For the Semantic Segmentation part DeepLab: Semantic Image Segmentation with Deep Convolutional Nets was referred, where they proposed Atrous Spacial Pyramid Pooling, that helps to segment image at different levels. Their efforts set a new state-of-art at PASCAL VOC semantic segmentation task[3], which the data was trained on Cityscrapes dataset.

III. Method

In this section, the proposed method for projecting lidar points onto an RGB camera and fusing the results with semantic segmentation is discussd. The overall approach involves several steps, including intrinsic and extrinsic calibration, spherical projection of the lidar point cloud, and semantic segmentation of the RGB image. We detail each of these steps in the following subsections.

A. Dataset

For this project The KITTI Vision Benchmark Suite, was used. Specifically, Processed(synced and rectified) color stereo sequence was taken along with 3D Velodyne point clouds with around 100k points per frame in binary float matrix format. There are totally 4 different cameras in the vehicle used for data collection, of which 2 are grayscale and 2 are colored, we will be specifically using camera 2 images as shown in Fig.1. Also calibration data of camera, camera to velodyne was taken.

B. Intrinsic and Extrinsic Parameters

- P2_rect: a 3x4 projection matrix for the second camera as shown in Fig.2 . It maps 3D points in the world to 2D points in the image captured by the second camera.

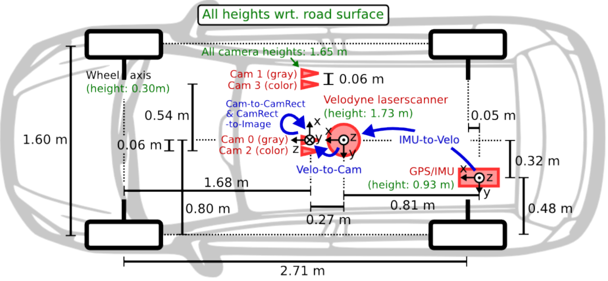


Fig. 1. camera and lidar setup in KITTI data

- R0_rect: a 3x3 rotation matrix that accounts for rectification of points in the reference camera.
- Tr_velo_to_cam: a 4x4 transformation matrix that describes the Euclidean transformation from the lidar coordinate system to the reference camera coordinate system
- Tr_imu_to_velo: a 4x4 transformation matrix that describes the Euclidean transformation from the IMU (inertial measurement unit) coordinate system to the lidar coordinate system.

$$\mathbf{P}_{rect}^{(i)} = \begin{pmatrix} f_u^{(i)} & 0 & c_u^{(i)} & -f_u^{(i)} b_x^{(i)} \\ 0 & f_v^{(i)} & c_v^{(i)} & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Fig. 2. Projection matrix, which converts rectified co-ords to image plane

The projection matrix P2_rect, along with the rectification matrix R0_rect and the transformation matrix Tr_velo_to_cam, can be used to project 3D lidar points into the second camera's image plane. Conversely, the inverse of these matrices can be used to project 2D points in the image plane into 3D points in the lidar coordinate system.

C. Projection

There are few major Projections :

- First the Lidar Points (Point Cloud - PC) have to be converted to Reference camera frame

$$PC\_cam\_ref = Tr\_velo\_to\_cam * PC \tag{1}$$

- Next rotational Rectification of the projected data

$$PC\_cam\_ref\_rectified = R0\_rect * PC\_cam\_ref \tag{2}$$

Now that the Extrinsic calibration has been performed, the lidar data will be available with reference to the cameras coordinate axis. To bring the points into X-Y image plane of the camera, intrinsic calibration was performed.

- Transform to Image Plane :

$$PC\_in\_Cam2\_image = P2\_rect * PC\_cam\_ref\_rectified \tag{3}$$

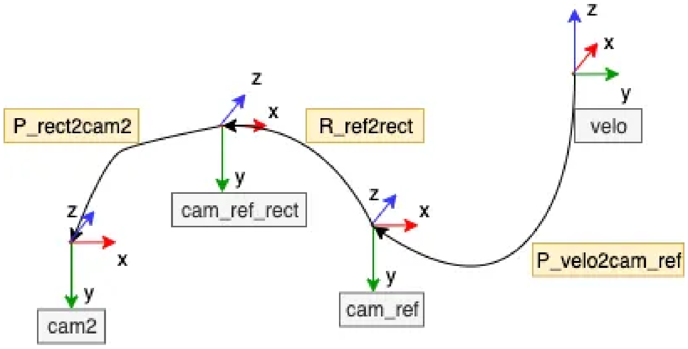


Fig. 3. Transformation Graph

D. Spherical Projection

A possible extension is to use a Spherical projection before performing intrinsic calibration. Which will give us a cylindrical panorama view around the Cam_ref frame. Now we can unwrap this cylinder to get an image of full 360°and use it as an extended image input for a Neural Network. This image is also referred to as a Polar Grid Map(PGM).

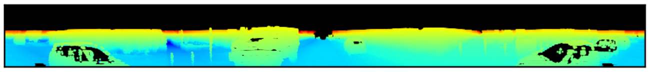


Fig. 4. 360°Polar Grid Map formed from spherical projection

E. Semantic Segmentation

Semantic segmentation is a computer vision task that involves dividing an image into multiple segments, each corresponding to a different object or region in the image. The aim of semantic segmentation is to label each pixel in the image with its corresponding class. This task is more challenging than traditional object detection, as it requires pixel-level accuracy and the ability to differentiate between objects of similar appearance.

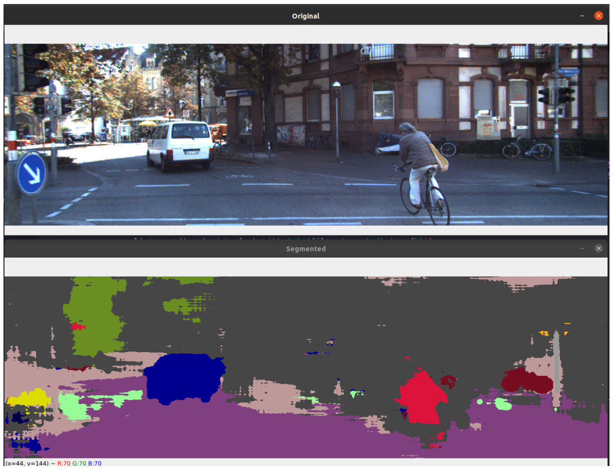


Fig. 5. Segmentation comparison on continuos data with 0.5 FPS

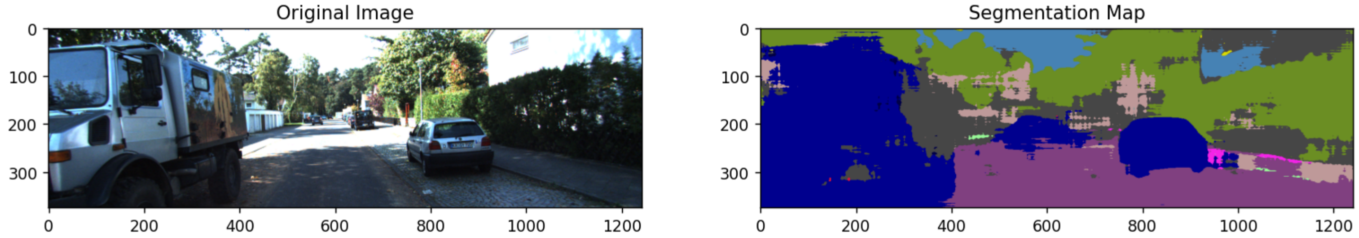


Fig. 6. Semantic Segmentation using DeeplabV3

Deeplabv3 is a popular deep learning architecture for image segmentation that has been trained on large-scale datasets such as COCO and Pascal VOC. It uses a deep neural network to learn a mapping between input images and their corresponding segmentation maps. The pre-trained Deeplabv3 model that is used in this context has already been trained on a large dataset of images and their corresponding semantic segmentation maps. This means that it has already learned to recognize various objects and their semantic labels, such as cars, pedestrians, and buildings. By using this pre-trained model, the task of semantic segmentation can be performed much faster and with higher accuracy than training a new model from scratch.

F. Segmentation of Lidar data

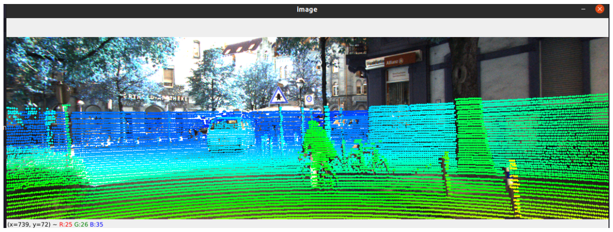
There is correspondence 1 to 1 mapping between the lidar points in the image plane and the image pixels itself. There is also 1 to 1 mapping between image and segmented image. Hence we can directly map the segmented image mask with the lidar points and use this information to segment the Lidar data.

IV. Experiments and Results

When this was performed on the KITTI dataset, The projection of lidar points on the image had less than 10% error, and the projection can be run in real time. The semantic Segmentation using DeeplabV3, although once a state of art, works well, but has constraints over usage of GPU and computation power. Specifically, I was having a speed around 0.3 - 0.4 Frames per second, which caused issue with syncing both segmentation and projection to finally segment the lidar data. The results obtained were still better than using neural nets to train on segmenting directly on lidar data and some of the other such methods.

V. Discussion and Summary

In summary, the project involves using a combination of LiDAR and camera data to perform object detection and semantic segmentation for autonomous driving. The LiDAR data is used to generate a point cloud, which is then projected onto the camera image using various calibration matrices. The resulting image is used for object detection, and the RGB image is used for semantic segmentation using a pre-trained deep learning model. Then using correspondence of the segmented image and point cloud in image plane, segmentation of 3D lidar data is possible. The project also explores the use of spherical projection to obtain a panoramic view of the environment, which can be used as an extended input for the neural network. Overall, the project showcases the potential of combining LiDAR and camera data for advanced perception tasks in autonomous driving.



(x=739, y=72) ~ R:25 G:26 B:35

Fig. 7. Lidar projected on image

References

[1] Khaled El Madawy, Hazem Rashed, Ahmad El Sallab, Omar Nasr, Hanan Kamel, & Senthil Kumar Yogamani (2019). RGB and LiDAR fusion based 3D Semantic Segmentation for Autonomous Driving. CoRR, abs/1906.00208.

[2] Sourabh Vora, Alex H. Lang, Bassam Helou, & Oscar Beijbom (2019). PointPainting: Sequential Fusion for 3D Object Detection. CoRR, abs/1911.10150.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, & Alan L. Yuille (2016). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. CoRR, abs/1606.00915.

[4] Pierre Biasutti, Aurélie Bugeau, Jean-François Aujol, & Mathieu Bredif (2019). RIU-Net: Embarrassingly simple semantic segmentation of 3D LiDAR point cloud. CoRR, abs/1905.08748.