Fusing LiDAR and Camera Data for Advanced Context Recognition in Autonomous Navigation Sensory Systems Through Multidimensional Deep Neural Network Architectures

Zhang, Emily (School: Cherry Creek High School)

Current robotic and vehicular systems for navigation, such as SLAM, rely heavily on sensory input to reconstruct and navigate environments but are unable to detect the structural integrity of objects. For example, Boston Dynamics' BigDog and Spot are equipped with both visual cameras and Light Detection and Ranging (LiDAR) sensors, but they have largely separate uses, and each cannot perform context recognition fast enough alone. LiDAR data cannot be used to accurately classify specific objects, while visual data cannot be used to find the position of individual objects other than by computationally expensive region generation methods, which fail regardless in low-light situations. Thus, this project proposes a method to fuse both LiDAR and camera data by using two different network architectures. First, LiDAR point cloud data is fed through a VoxelNet architecture to locate 3D bounding boxes around objects — this effectively performs object detection. Then, the 3D bounds are projected onto the 2D image space, and these regions of interest are ran through a separate convolutional neural network (CNN) based on the ResNet-101 architecture to classify the objects. Both networks were trained on the KITTI dataset, a widely-used autonomous driving dataset of both point cloud and camera information, so the CNN classifies objects into cars, pedestrians, cyclists, trucks, and vans. The VoxelNet reached an average precision (AP) of 75.52, 54.73, and 53.92 respectively for the easy, moderate, and hard occlusion levels of the KITTI dataset. The CNN attained an accuracy of 93.60% and an AP of 93.62, which outperforms previous metrics.

Awards Won:
Association for Computing Machinery: First Award of $4,000