Jonathan Schierl
Quinn Graehling
Theus Aspiras
Vijayan Asari

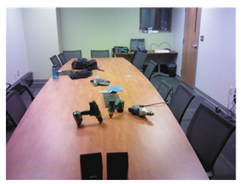
Multi-modal Data Analysis and Fusion for Robust Object Detection in 2D/3D Sensing

Abstract

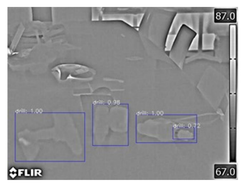
Currently, there is a lack of comparison between different modalities in object detection. We propose a dataset specifically captured for the comparison of 2D infrared and 3D point cloud data. We then analyze these modalities separately, using state of the art deep learning architectures. Finally, we will create a fusion network that uses both modalities, to increase the overall accuracy of detection

Dataset

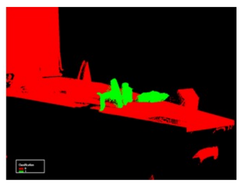
This dataset was captured specifically for comparison of 3D LiDAR point-cloud and 2D infrared data. The FARO 3D LIDAR Sensor and FLIR t650sc Infrared Camera were used, respectively. The sensors were placed next to each other for each capture, to retain a similar resolution. Visual spectrum images were also captured with the FLIR camera. The RGB, infrared, and point cloud sample images can be seen below:



RGB



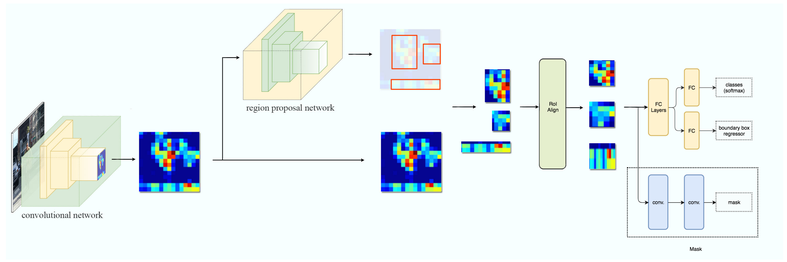
Infrared



Point Cloud

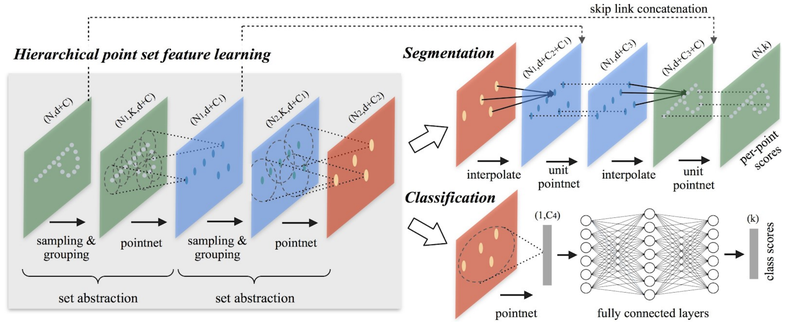
Method

These modalities were trained separately using state of the art deep learning architectures. For 2D analysis, the Mask R-CNN network was used. Included in this architecture was a pretrained backbone used for pretraining.



Mask R-CNN Architecture

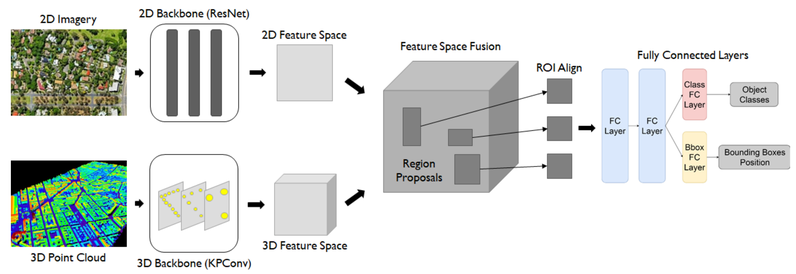
The 2D infrared images, however, lacked contrast. To better utilize the backbone of this architecture, a image-enhancement algorithm, called Retinex, was used. For 3D analysis, the PointNet++ architecture was used.



PointNet++ Architecture

Fusion Network

A fusion network will be created that utilizes the strengths of each modality, to increase the accuracy of detection. The 2D portion of this network will process the infrared images using a ResNet backbone, as done in the Mask R-CNN network, to extract a 2D feature space. Similarly, the 3D Point clouds will use KPConv backbone to extract a similar 3D feature space. These feature spaces can be combined, and fed through the rest of the Mask R-CNN network, for detection.



Fusion Network Architecture

Results

Model	Epochs	Time per Epoch (min)	Drill Class Accuracy	Overall Accuracy
Standard PointNet++	30	51:44	80.83	99.95
PointNet++ Costum 3 Layer	30	46:32	84.42	99.97
PointNet++ Costum 4 Layer	30	46:52	79.52	99.97
PointNet++ Costum 5 Layer	30	60:32	80.77	99.96
PointNet++ Costum 3 Layer	400	46:32	89.10	99.96

Model	Epochs	Preprocessing	Backbone	Drill Class Accuracy
Maskrcnn benchmark	15000	None	ResNet-50	0.5198
Maskrcnn benchmark	15000	None	ResNet-101	0.6077
Maskrcnn benchmark	15000	None	ResNet-152	0.6106
Maskrcnn benchmark	15000	Retinex	ResNet-50	0.6189
Maskrcnn benchmark	15000	Retinex	ResNet-101	0.6237
Maskrcnn benchmark	15000	Retinex	ResNet-152	0.5935

Conclusion

In general, the 3D network performed better, where it correctly identified all of the drills, whereas the 2D network did not. The drop in accuracy was due to noise points. The 3D network performed best at 89.1% accuracy, and the 2D network performed best with the Retinex preprocessing at 62.37% accuracy.

UNIVERSITY of DAYTON
Vision Lab
Kettering Laboratories
300 College Park
Dayton, OH 45469-0232

Center of Excellence for Computer Vision and Wide Area Surveillance Research
visionlab.udayton.edu

Dr. Vijayan K. Asari
Director
(937) 229-4504
vasari1@udayton.edu