UNIVERSITY of DAYTON
VISION LAB
Center of Excellence

Quinn Graehling
Jonathan Schriel
Dr. Theus Aspiras
Dr. Vijayan Asari

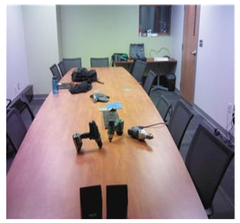
Multi-modal Data Analysis and Fusion for Robust Object Detection in 2D/3D Sensing

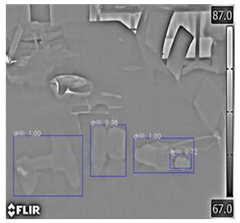
Abstract

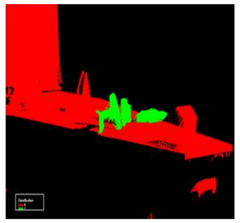
Multi-modal data is useful for complex imaging scenarios due to the exclusivity of information found in each modality, but there is a lack of meaningful comparisons of different modalities for object detection. In our work, we propose three contributions: (1) Release of a multi-modal, ground-based small object detection dataset, (2) A performance comparison of 2D and 3D imaging modalities using state-of-the-art algorithms, and (3) a multi-modal fusion framework for 2D/3D sensing.

Introduction

Multi-modal data analysis is one of the next major steps for machine learning and artificial intelligence. Previously multi-modal learning systems were scarce partly due to the lack of available datasets that provide the same environment captures in multiple modalities. This research first proposes a multi-modal dataset consisting of captures of the same environment in 2D RGB images, 2D infrared images and 3D point cloud representations. This dataset is then tested using state of the art deep learning neural networks. Finally a fusion network is proposed that is capable of learning from 2D and 3D data captures simultaneously and is capable of overcoming several of the shortcomings within the singular modality networks.







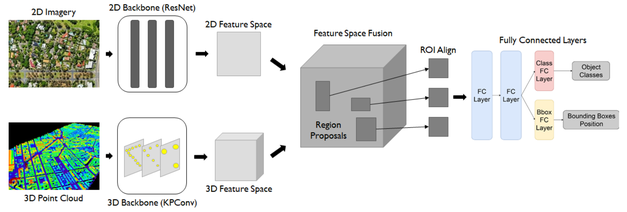
Methods

The Drills dataset is a proposed dataset consisting of natural environments captured in 2D RGB, 2D IR and 3D point cloud modalities. The dataset was captured using a FLIR infrared camera for the 2D images and a FARO LIDAR sensor for the 3D point cloud captures. The dataset consists of 100 separate labeled captures with each capture described as simple, complex and very complex depending on the number of labeled objects and the level of occlusion of these objects.

The infrared and RGB datasets were first optimized using the Multiscale Retinex image enhancement algorithm which focuses on adjusting image color saturation in accordance with relative illumination levels within the image. These enhanced 2D representations were then tested on the Mask R-CNN deep learning algorithm for object detection in 2D.

The 3D point cloud data representations were tested using the PointNet++ and KPConv deep learning networks . As opposed to the 2D data, 3D data representations consist of unordered points usually numbering in the multi millions points per capture. Therefore these networks have to perform learning with a greater number of parameters. PointNet++ solves this by performing hierarchical learning while KPConv achieves this by utilizing a deformable convolutional kernel that distributes a distance weight system..

The fusion network purposed is capable of combining the 2D Mask R-CNN algorithm and KPConv algorithm in order to perform fusion learning. It uses these existing algorithms to extract features from the data and then aligns these features using ROI align into a single feature set, effectively combining the features in the 2D and 3D representations.



Results

The Drills dataset was tested on both the Mask R-CNN with and without image enhancement and on the PointNet++ architectures to assess the benefits and shortcomings of the architectures learning in different modalities.

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

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

Conclusion

The 2D learning method was able to detect whole drills while avoiding partial detection. However, the 2D test results show several false positive detections. In contrast, the 3D learning was able to detect most drills within each capture but had a number of partial detection without successfully detecting the entire drill. Overall the 3D testing proved to more accurate at detection when measured with IoU that the 2D methods. The fusion method would be capable of detecting all of the drills within an image while avoiding the false positives detected within the 2D results. Furthermore, the 2D bounding boxes would allow for the true positives to be more fully detected by the 3D algorithm and eliminate its issue of partial detections.

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