Car Driving Without Cameras

Jones Agwata

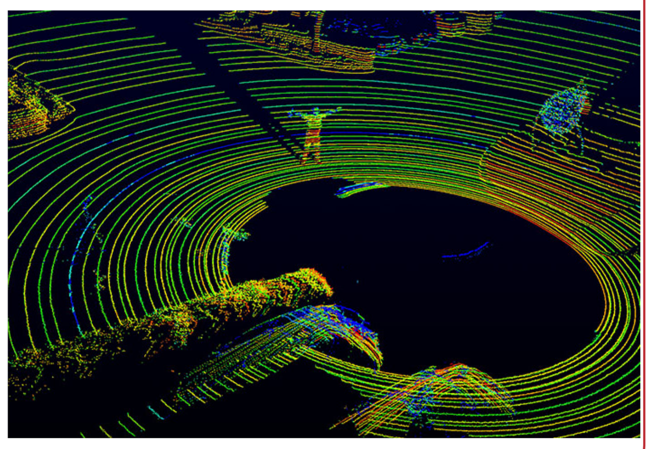
Supervisors: Dr. Luis Vaquero Gonzalez , Dr. Raul Santos-Rodriguez

University of BRISTOL

Department of Computer Science

Motivation

- Autonomous vehicles (AVs) increasingly becoming a reality.
- Multi-modal approach by fusing input from a large array of sensors including cameras and LiDARs is common.
- Approach is **expensive** and **energy inneficient.**
- Companies developing AVs seek to reduce the number of sensors.



Aim

- **Performance of different sensors under different contexts(urban or non-urban) has not been widely explored.**

1. Can we automatically detect the context of driving scenes.
2. Do sensors work better in different contexts?
3. Do some object detection models work better in different contexts?

Detecting Scene Context

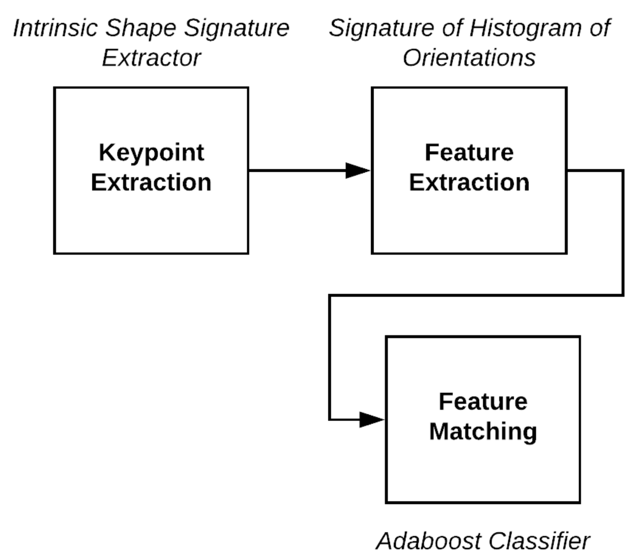
- Visually classified images with corresponding pointclouds from KITTI Dataset[1] into urban and non-urban contexts.

1. From Camera Input

- Trained linear SVM on histogram of semantic classes obtained from semantic segmentation[3].

2. From LiDAR Input

- Trained Adaboost classifier on features extracted from the keypoints of the pointcloud.

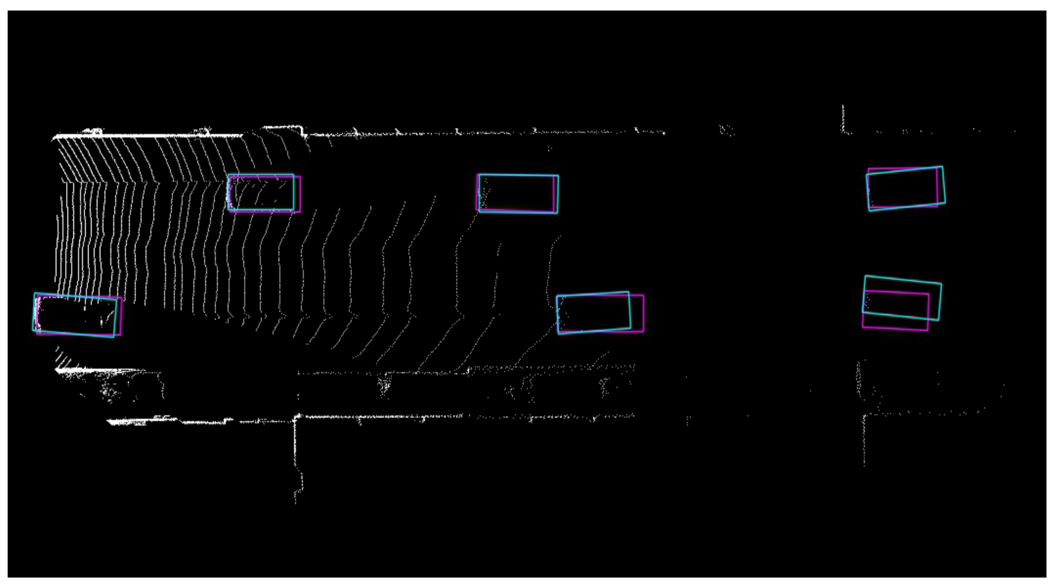


Model Types

Two models were evaluated on the divided context dataset.

1. VoxelNet[4]- LiDAR Only

- Point cloud object detection neural networks.
- Extracts and trains on features from voxels containing pointclouds using a Region Proposal Network.



2. Aggregated View Object Detection[2]- Image & LiDAR

- Multimodal object detection model.
- Fuses features from image and pointclouds in a Region Proposal Network.

Evaluation

1. Context Detection

Context Detection using PointCloud Feature Matching

	precision	recall	f1-score	support
non-urban	**0.52**	0.45	0.48	206
urban	0.51	**0.58**	**0.55**	206
avg / total	0.51	0.51	0.51	412

Context Detection using Image Segmentation

	precision	recall	f1-score	support
non-urban	0.81	**0.9**	**0.85**	193
urban	**0.9**	0.81	**0.85**	218
avg / total	0.86	0.85	0.85	411

2. Model Performance

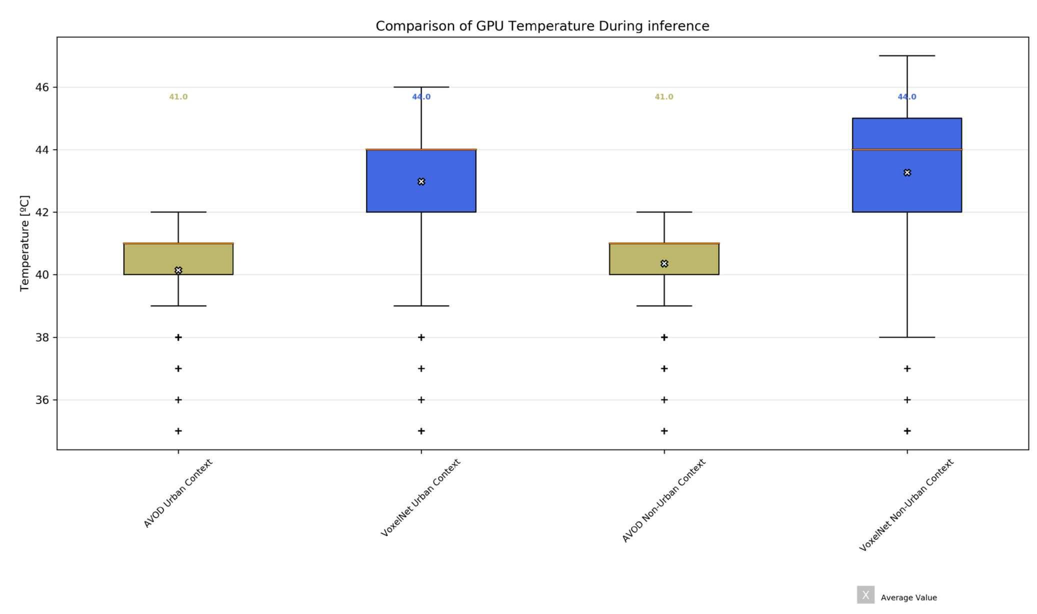
Average Precision

Context		Urban			Non-urban		
Difficulty		Easy	Med	Hard	Easy	Med	Hard
AVOD	Car 2D BB	**86.98**	**77.10**	**68.01**	**89.18**	**79.75**	**78.55**
	Car BEV BB	86.00	74.35	65.62	87.11	**76.85**	**75.72**
	Car 3D BB	**75.44**	**63.74**	**54.33**	**75.43**	**64.19**	**62.90**
Voxel Net	Car 2D BB	69.59	65.98	59.28	77.52	67.73	62.28
	Car BEV BB	**86.34**	**76.18**	**68.10**	**88.63**	75.36	69.50
	Car 3D BB	73.63	58.47	50.74	68.62	49.45	45.80

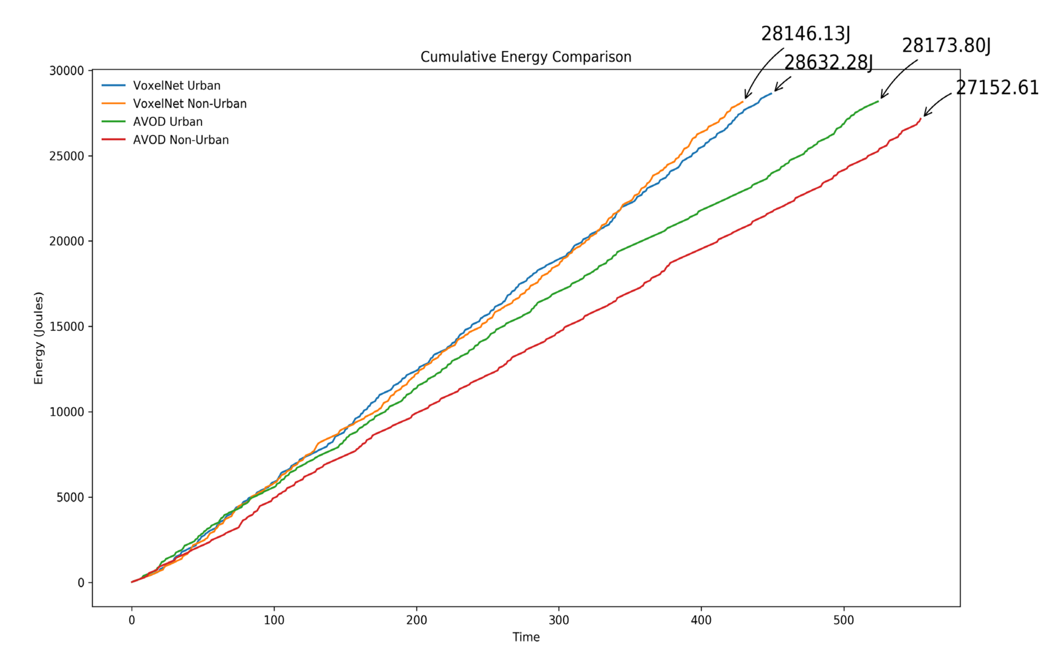
Inference Time

	VoxelNet			AVOD		
	min	max	mean	min	max	mean
Urban	0.113	2.224	0.127	**0.096**	**2.506**	**0.113**
Non urban	0.113	**3.813**	0.129	**0.095**	2.457	**0.112**

Temperature



Energy



Discussion

- Context detection using semantic histograms of images is fairly accurate.
- Context detection using point cloud feature matching was quite poor. This could be as a result of the sparse nature of the point clouds thus affecting feature matching due to varying point cloud sampling.
- Some images were difficult to visually categorise into urban or non-urban images due to lack of temporal and spatial information thus affecting performance of both context detectors.

- AVOD proved to be better than VoxelNet in many performance metrics. However VoxelNet was better in Bird's Eye View detection in all urban difficulties and the easiest non-urban difficulty.
- Computation-wise, VoxelNet generated a higher average temperature on a NVIDIA P100 GPU as compared to AVOD. However the energy consumed was similar.

References

[1] Vision meets robotics: The KITTI dataset. International Journal of Robotics Research (IJRR), 2013.

[2] Joint 3d proposal generation and object detection from view aggregation. arXiv preprint arXiv:1712.02294, 2017.

[3] Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.

[4] Voxelnet: End-to-end learning for point cloud based 3d object detection. arXiv preprint arXiv:1711.06396, 2017.