



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم

بسم الله الرحمن الرحيم



MONA MAGHRABY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم



شبكة المعلومات الجامعية التوثيق الإلكتروني والميكروفيلم



MONA MAGHRABY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم

جامعة عين شمس التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغييرات



يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار



MONA MAGHRABY



3D SEMANTIC SEGMENTATION USING CAMERA-LIDAR SENSOR FUSION FOR AUTONOMOUS DRIVING

By

Khalid Mohamed Naguib Elmadawi

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communication Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2021

3D SEMANTIC SEGMENTATION USING CAMERA-LIDAR SENSOR FUSION FOR AUTONOMOUS DRIVING

By

Khalid Mohamed Naguib Elmadawi

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communication Engineering

Under the Supervision of

Prof. Dr. Hanan Ahmed Kamal

A.Prof. Dr. Omar Ahmed Nasr

.....
Professor of Control
Electronics and Communication
Engineering
Faculty of Engineering, Cairo University

.....
Associate Professor
Electronics and Communication
Engineering
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2021

3D SEMANTIC SEGMENTATION USING CAMERA-LIDAR SENSOR FUSION FOR AUTONOMOUS DRIVING

By

Khalid Mohamed Naguib Elmadawi

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communication Engineering

Approved by the Examining Committee

Prof. Dr. Hanan Ahmed Kamal,

Thesis Main Advisor

Associate Prof. Dr. Omar Ahmed Nasr,

Advisor

Prof. Dr. Mohsen Abdelrazek Rashwan,

Internal Examiner

Prof. Dr. Sherif Mahdi Abdou,

External Examiner

Professor in Faculty of Computers and Artificial Intelligence

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2021

Engineer's Name: Khalid Mohamed Naguib Elmadawi
Date of Birth: 09/06/1990
Nationality: Egyptian
E-mail: Khaled.elmadawi@gmail.com
Phone: 01149669927
Address: 15 Tiba Gardens-3B/6October First/Giza
Registration Date: 01/03/2016
Awarding Date:/..../2021
Degree: Master of Science
Department: Electronics and Communication Engineering



Supervisors:

Prof. Hanan Ahmed Kamal
Associate Prof. Omar Ahmed Nasr

Examiners:

Prof. Hanan Ahmed Kamal (Thesis main advisor)
A.Prof. Omar Ahmed Nasr (Advisor)
Prof. Mohsen Abdelrazek Rashwan (Internal examiner)
Prof. Sherif Mahdi Abdou (External examiner)

Professor in Faculty of Computers and Artificial Intelligence.

Title of Thesis:

3D SEMANTIC SEGMENTATION USING CAMERA-LIDAR SENSOR FUSION
FOR AUTONOMOUS DRIVING

Key Words:

Sensor Fusion; Environment Perception; Early Fusion; Middle Fusion; Spherical Grid Map

Summary:

This research improves the environment sensing in control systems through multi-sensor fusion. We are fusing Camera raw image data with LiDAR Point cloud resulting from having a colored point cloud, represented in a spherical grid map representation. We used the fused data in early and mid-level fusion algorithms. We evaluate our algorithms on the KITTI dataset, which provides semantic segmentation for different classes such as Cars, Pedestrians, and cyclists. We evaluated our work on two states of art architectures, namely SqueezeSeg and PointSeg, resulting in improving the mIOU score to 10% in both cases relative to the LiDAR only baseline.

Disclaimer

I hereby declare that this thesis is my own original work and that no part of it has been submitted for a degree qualification at any other university or institute.

I further declare that I have appropriately acknowledged all sources used and have cited them in the references section.

Name: khalid elmadawi

Date: .../.../.....

Signature:

Dedication

I would like to dedicate my thesis to family who supported me, and to my professors and colleagues that tough me and offered me a lot to achieve my master degree.

Acknowledgments

First, I would like to thank God for helping me and giving me the strength to overcome all the obstacles out of my hand. Then Great thanks to my colleagues in Valeo who shared my effort in my research and supporting me till we have achieved successful results. Thanks, and gratitude to Dr. Hanan Kamel and Dr. Omar Nasr for their guidance and contribution in our work through my whole research period. This work would not be carried out without the support coming from the electronics and communications department members in the Faculty of Engineering of Cairo University.

Table of Contents

DISCLAIMER	I
DEDICATION	II
ACKNOWLEDGMENTS	III
TABLE OF CONTENTS	IV
LIST OF TABLES	VI
LIST OF FIGURES	VII
NOMENCLATURE	XI
ABSTRACT	XII
CHAPTER 1 : INTRODUCTION	1
1.1. MOTIVATION	2
1.2. PROBLEM DEFINITION	7
1.3. THESIS CONTRIBUTION	9
1.4. ORGANIZATION OF THESIS	10
CHAPTER 2 : LITERATURE REVIEW	11
2.1. INTRODUCTION	11
2.2. NEURAL NETWORKS	13
2.2.1. Fully connected Neural Networks	15
2.2.2. Convolution Neural Network	16
2.3. CONVOLUTION AUTO-ENCODERS	18
2.4. FULLY CONVOLUTION NETWORKS	19
2.5. U-NET CONVOLUTION NETWORKS	19
2.6. CAMERA SEMANTIC SEGMENTATION SENSOR MODALITY RELATED WORK	20
2.7. CAMERA DETECTION RELATED WORK	21
2.7.1. Sliding Window	23
2.7.2. Convolution Implementation of Sliding Window	25
2.7.3. Intersection Over Union	27
2.7.4. Nonmax Suppression	27
2.7.5. Anchor Boxes	29
2.7.6. Yolo Algorithm.....	30
2.8. LiDAR SEMANTIC SEGMENTATION SENSOR MODALITY RELATED WORK	30
2.8.1. Top-View Scaled LiDAR Representation	31
2.8.2. Voxel-Net	33
2.8.3. PointNet	34
2.8.4. Spherical Grid Maps	37
2.9. SENSOR MODALITIES FUSION	38
2.10. DATASETS	40

2.10.1.	KITTI	41
2.10.2.	KAIST.....	42
2.10.3.	H3D.....	43
2.10.4.	nuScenes.....	44
2.10.5.	Argoverse	45
2.10.6.	Lyft L5	46
2.10.7.	Waymo Open	47
2.10.8.	Dataset comparison	47
2.11.	SUMMARY	49
CHAPTER 3 : RGB AND LIDAR FUSION SEMANTIC SEGMENTATION.....		50
3.1.	UNI-MODAL ARCHITECTURES.....	50
3.1.1.	Camera Semantic Segmentation	50
3.1.2.	LiDAR Semantic Segmentation	53
3.2.	NO FUSION ARCHITECTURE.....	53
3.3.	EARLY FUSION ARCHITECTURE	55
3.3.1.	Sensor Calibration	55
3.3.2.	Intrinsic Transformation.....	57
3.3.3.	Extrinsic Transformation.....	59
3.3.4.	Early fusion architecture.....	61
3.4.	MID-FUSION ARCHITECTURE.....	62
3.5.	SUMMARY	64
CHAPTER 4 : EXPERIMENTS AND RESULTS.....		66
4.1.	EXPERIMENTAL SETUP	66
4.2.	RESULTS	67
4.3.	SUMMARY	71
CHAPTER 5 : CONCLUSION AND FUTURE WORK.....		72

List of Tables

Table 1: Different sensor modalities performance Vs different features.	12
Table 2: Different combination of multi-sensor modalities performance Vs different features.	12
Table 3: A comparison table was published by nuscenets [46] representing variety keys for different datasets.	48
Table 4: Quantitative evaluation on KITTI Raw dataset using SqueezeSeg architecture.	67
Table 5: Quantitative evaluation on KITTI Raw dataset using PointSeg architecture...	67

List of Figures

Figure 1: Urban sprawl and the use of land, degrading the environment [1].	1
Figure 2: Traffic jam problem due the wide use of vehicles [2].	1
Figure 3: Autonomous Driving Revolution [3].	2
Figure 4: connectivity between vehicles, and the surrounding environment [4].	2
Figure 5: Autonomous driving provides safety not only for the vehicle, also for everyone around it [5].	3
Figure 6: Autonomous driving saves time for the drivers [6].	3
Figure 7: Autonomous driving decrease the emission of CO2 and helps us to save the environment [7].	4
Figure 8: Autonomous driving creates new job opportunities [8].	4
Figure 9: Five levels that defines autonomous driving vehicles.	6
Figure 10: Autonomous driving main modules [9].	7
Figure 11: Waymo sensor suite for its autonomous driving vehicles [10].	8
Figure 12: Tesla sensor suite for its autonomous driving vehicles [11].	9
Figure 13: Neural network main components, neural network input, neural network weights, and neural network output [14].	13
Figure 14: Explanation of how the input would be represented for the neural network, how the network would take a decision through forward propagation, and how would it learn from the error between the prediction and ground truth through back propagation [14].	14
Figure 15: Extracted features of fully connected neural networks, that makes the neural network build its decision which class it should select [15].	15
Figure 16: convolution process on an image for a single kernel [16].	16
Figure 17: Same convolution process can be done through multiple convolution filter creating multiple feature maps and extracting more features from the image [16].	16
Figure 18: Input image of the convolution neural network, of size of 28x28, with one channel that contains value of 0 or 1 [18].	17
Figure 19: Convolution neural networks kernels after training, representing what the features that makes the neural network build its decision upon it [18].	17
Figure 20: feature maps resulted from the convolution of the input layer, from the above kernels, resulting to a set of extracted features that help the neural network to build its decision [18].	18
Figure 21: The full convolution process on a single kernel, generating a single feature map [18].	18
Figure 22: Convolution auto-encoder that removes the noise from an image and helps it to generate a better less noisy image [19].	19
Figure 23: Fully convolution Neural Network that represents the skip connection approach [20].	19
Figure 24: U-net architecture for automotive application [22].	20
Figure 25: 2D image semantic segmentation, where fig(a,e) are the input for the DNN, fig(b,f) area the ground truth of the predicted classes, fig(c,g) are the output of FCN architecture, and fig(d,h) are the output of the presented in [23].	21
Figure 26: From left to right successively image classification Vs image classification with localization, Vs multiple object detection [27].	22
Figure 27: Convolutional neural network that outputs class of the detected output with the position of the bounding box [27].	22

Figure 28: Bounding box representation with its coordinates [27] .	23
Figure 29: Training set example for sliding window algorithm [28].	24
Figure 30: Different sizes of sliding window [28].	24
Figure 31: Fully connected neural network representation [29].	25
Figure 32: Fully convolutional neural network representation for a fully connected neural network [29].	25
Figure 33: mapping of the convolution filters applied on the input image to each output grid pixel [29].	26
Figure 34: Sliding window example showing box represented by each output grid pixel, and the red bounding box that represents the detected object in the middle grid pixel [29].	26
Figure 35: If intersection over union between the detected object and the ground truth is greater than 0.5, then it is considered as a correct detection [30].	27
Figure 36: Output grid representation of the input image, where the result of the output grid would be 19x19 pixels, each pixel can represent an object bounding box [31].	28
Figure 37: an example for multiple detection per object, where the right vehicle is detected three times, with three probabilities (0.9,0.7,0.6) and the non max suppression algorithm will discard the low probability bounding boxes and will select the highest bounding box, same for the left vehicle was detected three times, with probabilities of (0.8,0.65,0.6) will select the bounding box with the highest probability of detection, and will discard the other two bounding boxes [31].	28
Figure 38: Representation of how multiple anchors can be represented to detect multiple object per grid cell in the output detection grid cell [32].	29
Figure 39: YOLO detection algorithm [33].	30
Figure 40: 2D sub space representation of the Cartesian coordinates of the sensor FOV [34], each cell contains information of the scan points in this subspace.	31
Figure 41: 2D top view of the LiDAR perception representing the subspace representation of each cell to a certain class, in top image represents the ground truth, middle image how the 2D grid cell looks like after projecting the LiDAR scan points to it, low image represents the output of the DNN presented in [35].	32
Figure 42: 2D top view scaled LiDAR representation provides the ability to apply object detection either by Yolo algorithm as presented in [34] or by the output of the semantic segmentation as presented in [35].	32
Figure 43: 3D voxelization in presenting 3D pointcloud, each voxel contains one point or more of the point cloud, and can be presented in Binary voxels or density voxles as presented in [36].	33
Figure 44: PointNet approach different DNN classification, part segmentation, and semantic segmentation output [37].	34
Figure 45: PointNet semantic segmentation for multiple classes per scan point [chairs, ceiling, tables, ground, walls, doors, white board, windows] [37].	34
Figure 46: PointNet robustness with the decreased number of points, depending on the critical points acting as a global feature to achieve its required task, for either classification, part segmentation, or semantic segmentation [37].	35
Figure 47: PointNet semantic segmentation output Vs ground turth, presenting the output of PointNet right, and Groundtruth output left as presented in [38].	36
Figure 48: PointNet++ method for local and global feature extraction in the 3D point clouds [39].	36
Figure 49: Super Point point cloud semantic segmentation, where image(a) presents the point cloud in (X, Y, Z, R, G, B), image(b) represents geometric partitioning, image(c)	