

CERTIFICATE

It is certified that the work contained in this thesis entitled “DESIGN AND DEVELOPMENT OF DEEP LEARNING BASED APPROACHES FOR PERCEPTION AND PLANNING MODULES OF THE AUTONOMOUS VEHICLES” by “DIVYA SINGH” has been carried out under my supervision and that it has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive, Candidacy and SOTA.

Prof. RAJEEV SRIVASTAVA

Professor

Department of Computer Science and Engineering

Indian Institute of Technology

(Banaras Hindu University)

Varanasi-221005

DECLARATION BY THE CANDIDATE

I, DIVYA SINGH , certify that the work embodied in this thesis is my own bonafide work and carried out by me under the supervision of Prof. RAJEEV SRIVASTAVA from JULY-2018 to JULY-2022, at the DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, Indian Institute of Technology (BHU) Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully lifted up any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and included them in this thesis and cited as my own work.

Date :

Signature of the Student

Place : Varanasi

(DIVYA SINGH)

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my/our knowledge.

Signature of Supervisor

(Prof. RAJEEV SRIVASTAVA)
Department of Computer Science and Engineering
Indian Institute of Technology
(Banaras Hindu University)
Varanasi-221005

Signature of Head of Department

(Prof. Sanjay Kumar Singh)

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: DESIGN AND DEVELOPMENT OF DEEP LEARNING BASED APPROACHES FOR PERCEPTION AND PLANNING MODULES OF THE AUTONOMOUS VEHICLES

Name of the Student: DIVYA SINGH

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi, all rights under copyright that may exist in and for the above thesis submitted for the award of the DOCTOR OF PHILOSOPHY.

Date :

(DIVYA SINGH)

Place :

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

To
My Beloved Parents
Mrs. Sangeeta Singh
and
Mr. Santosh Kumar Singh

ACKNOWLEDGEMENTS

I want to take this opportunity to express my deep sense of gratitude to all who helped me directly or indirectly during this thesis work. Firstly, I would like to thank my supervisor, Prof. Rajeev Srivastava, for being a great mentor and the best adviser I could ever have. His advice, encouragement, and critics are a source of innovative ideas, inspiration, and causes behind the successful completion of this Thesis work. The confidence shown on me by him was the most significant source of inspiration for me. It has been a privilege working with him for several years. I am highly obliged to all the faculty members of the Computer Science and Engineering Department for their support and encouragement. I express my sincere thanks to Prof. K. K. Shukla and Dr. Pratik Chattopadhyay of the department of Computer Science and Engineering and Prof. Subir Das, Department of Mathematical Sciences IIT (BHU), for providing continuous support, encouragement, and advice. I express my sincere thanks to all the Professors, Deans, office staff, supporting staff, and Ph.D. Research Scholars of Indian Institute of Technology (BHU) Varanasi India. I express my gratitude to Director, Registrars, Deans, Heads, and Student Alumni of the Indian Institute of Technology (BHU) Varanasi.

My memory of the study period at IIT (BHU) can never be complete without mentioning my fellow research scholars. Special thanks to Dr. Alok Kumar Singh Kushwaha, Dr. Vibhav Praksh Singh, Dr. Roshan singh, Dr. Pratishta Verma, Dr. Gargi Srivastava, Dr. Ankit Jaiswal, Dr. Jani Kuntesh Ketan, Mr. Santosh Tripathy, Ms. Shweta Singh and Mr. Devendra Sharma for their great help and cooperation. I extend special thanks to the non-teaching staff in the department, particularly the late Mr. A. N. Yadav, Mr. Manoj Kumar Singh, Mr. Ravi Kumar Bharti, Mr. Prakhar Kumar, Mr. Ritesh Singh, and Mr. Shubham Pandey for their consistent support.

My in-laws, Mrs. Vimla Singh and Mr. Hawaldar Singh, who gave me the power and brain to work out on this research and their help at every level, made me see this success. I owe thanks to my dearest husband, Dr. Ashutosh Kumar Singh for his continued and unflinching love, support and understanding during my pursuit of

Ph.D. degree that made the completion of thesis possible. You were always around at times I thought that it is impossible to continue, you helped me to keep things in perspective. I greatly value his contribution and deeply appreciate his belief in me. Words are insufficient to express my profound sense of gratitude to my friends Rakesh Yadav, Dr. Rajkumar Saini. My sisters and sister-in-law Smita singh, Rashmi Singh, Varsha Singh, Shreya Singh and Kiran Singh whose encouragement gave me physical and moral strength throughout my career as well in the present research. I extend my thanks to my uncle Rahul Singh, Satya Sheel Singh, Yashwant Singh and my brother-in-law Susheel Singh, Amit singh, brothers Bhawesh Singh, Prashant Singh, Siddhant Singh, Yash Vardhan singh who are part of my inspiration. Finally, I would like to wind up by paying my heartfelt thanks and prayers to the Almighty for their unbound love and grace.

- Divya Singh

Contents

List of Figures	xiii
List of Tables	xv
Abbreviations	xvii
Symbols	xix
Preface	xxi
1 Introduction	1
1.1 Background	1
1.2 Vision-Based Architecture of Autonomous Vehicles	3
1.2.1 Sensors	4
1.2.2 Perception System	4
1.2.2.1 Object Detection	4
1.2.2.2 Object Tracking	5
1.2.2.3 Trajectory Prediction	5
1.2.3 Mapping and Localization	6
1.2.3.1 Localization	6
1.2.3.2 Map State Estimation	6
1.2.4 Planning and Decision-Making	6
1.2.4.1 Trajectory Planning	7
1.2.4.2 Behaviour Planning	7
1.2.4.3 Motion planning	7
1.2.5 Safety Module	7
1.2.6 Control Module	8
1.3 Motivation	8
1.4 Problem statement	10
1.5 Thesis Objective	11
1.6 Contributions to the Thesis	12
1.7 Thesis Organization	14
2 Theoretical Foundation and Literature Survey	17

2.1	Introduction	18
2.2	Literature Review	18
2.2.1	Literature Review of Object Detection	19
2.2.1.1	Pioneer Methods	19
2.2.1.2	Deep-Learning Methods	20
2.2.1.3	Summary	33
2.2.2	Literature Review of Multi-Object Tracking	35
2.2.2.1	Traditional Multi-Object Tracking	36
2.2.2.2	Stereo-Vision Based MOT	37
2.2.2.3	Grid-Based MOT	38
2.2.2.4	Sensor-Fusion Based MOT	39
2.2.2.5	Deep-Learning Based MOT	40
2.2.2.6	Summary	41
2.2.3	Literature Review of Trajectory Prediction	42
2.2.3.1	Feature Encoding based TP	42
2.2.3.2	Interaction Modeling Based TP	44
2.2.3.3	Prediction Head Based TP	45
2.2.3.4	Generative Model-Based TP	46
2.2.3.5	Summary	46
2.2.4	Literature Review of Motion Planning	49
2.2.4.1	Traditional Algorithms	50
2.2.4.2	Machine-Learning and Deep-Learning Based Algorithm	52
2.2.4.3	Motion Planning using simulator	55
2.2.4.4	Summary	57
2.3	Research Gaps	59
2.3.1	Object detection	59
2.3.2	Multi-Object Tracking	60
2.3.3	Trajectory Prediction	61
2.3.4	Motion Planning	62
2.4	Benchmark Datasets and Simulator used for training and evaluation	62
2.4.1	KITTI [1]	63
2.4.2	Berkley Driving Dataset (BDD) [2]	63
2.4.3	Waymo [3]	64
2.4.4	Multi Object tracking (MOT) [4]	65
2.4.5	ARGOVERSE	66
2.4.6	APOLLOSCAPE	66
2.4.7	LYFT	66
2.4.8	CARLA Dataset	67
2.5	Evaluation Metrics	68
2.5.1	Precision	68
2.5.2	Multi-Object Tracking Accuracy	68
2.5.3	ID F1 Score	68
2.5.4	Mostly Tracked Targets	69

2.5.5	Mostly Lost Targets	69
2.5.6	IDs Identity Switches	69
2.5.7	FRAG	69
2.5.8	Final Displacement Error	69
2.5.9	ADE Average Displacement Error	70
2.5.10	Infraction Management	71
2.5.11	Driving Score	71
2.5.12	Route Completion	72
2.6	Conclusion	72
3	Single-Stage Attention based Object Detection for Autonomous Vehicles	73
3.1	Introduction	73
3.2	Proposed Method and Model	76
3.2.1	Channel Spatial attention based Object Detector	77
3.2.1.1	Filter Response Normalization	78
3.2.1.2	Attention Module	79
3.3	Result Analysis and Discussion	82
3.3.1	Experimental Setup	82
3.3.2	KITTI Dataset Result	84
3.3.3	BDD Dataset Result	84
3.3.4	Ablation Study	86
3.4	Conclusion	89
4	An end-to-end Hybrid method for Multi-Object Tracking	91
4.1	Introduction	91
4.2	Proposed Method and Model	93
4.2.1	A Hybrid method for Multi-Object Tracking	93
4.2.1.1	Motion Estimation	93
4.2.1.2	Re-identification of objects	94
4.3	Results and Analysis	96
4.3.1	Experimental Setup	96
4.3.2	Loss function	98
4.3.3	Ablation study	102
4.4	Conclusion	103
5	Trajectory Prediction and Motion Planning	105
5.1	Introduction	105
5.2	Proposed Methods and Models	108
5.2.1	Graph Neural Network-based Trajectory Prediction	108
5.2.1.1	Pre-processing of Datasets	109
5.2.1.2	The architecture of the model and its working	110
5.2.1.3	Generation of the Graph	111
5.2.1.4	Spatial Sampling of Adjacency Matrix for Spatial Features	113

5.2.1.5	Convolutional Module	113
5.2.1.6	Temporal Features	114
5.2.1.7	Feature Fusion Module	114
5.2.1.8	Path Fusion	114
5.2.1.9	Trajectory Prediction Module	115
5.2.2	Result Analysis and Discussions for TP	116
5.2.2.1	Experimental setup	116
5.2.2.2	Loss Function	117
5.2.3	Motion Planning Using CARLA Simulator	124
5.2.3.1	Architecture of the model and its working	124
5.2.3.2	Algorithm for PID Controller	127
5.2.4	Result Analysis and Discussion of Motion Planning	128
5.2.4.1	Experimental Setup	128
5.2.4.2	Loss Function	128
5.2.4.3	Ablation Study	133
5.3	Conclusion	135
6	Conclusion and Future Work	137
6.1	Conclusions	137
6.2	Suggestions for Future Research	139
6.3	Future Work	139
	References	141
	List of Publications	175

List of Figures

1.1	Overview of Vision-based Architecture for Autonomous Vehicles . . .	5
1.2	The road accidents, deaths and injured people over the year in India .	9
2.1	Taxonomy of the Object Detection methods	19
2.2	Architecture of (a) Two-Stage Object Detection and (b) Single-stage Object Detection	26
2.3	Taxonomy of Multi-Object Tracking methods	35
2.4	Taxonomy of the Trajectory Prediction Models	42
2.5	Texonomy of the motion planning techniques	50
2.6	Visualization of datasets in different conditions (a)The night vision of BDD dataset (b) cloudy weather of KITTI dataset	64
3.1	Configuration of the attention modules	77
3.2	CSA-SS Architecture	78
3.3	Channel attention sub-module	80
3.4	Association of attention module with ResNet network	82
3.5	Comparison of different attention mechanism efficiency	88
4.1	The framework of the proposed model for multi-object tracking that defines the two parallel process of detected and tracked object and match these two of two consecutive frame t-1 and t, with relative scale	94
4.2	Effect of visibility (a) and size of the object (b) of the proposed model	101
4.3	Qualitative Results of the proposed model on Waymo and MOT datasets	102
5.1	Architecture of temporal graph model for Trajectory Prediction . . .	110
5.2	The architecture of each spatial-temporal block for Graph convolu- tional network	115
5.3	Basic architecture of backbone network	116
5.4	(a) GT trajectory and PT of the road-agent with agent_ID = 337 of the Apolloscape Dataset (b) GT trajectory and PT of the road-agent with agent_ID = 659 of the Apolloscape Dataset	122
5.5	(a)GT trajectory and PT of the road-agent with agent_ID = 2 of a scene with scene_ID = 127 of the Lyft Dataset (b) GT trajectory and PT of the road-agent with agent_ID = 16 of a scene with scene_ID = 141 of the Lyft Dataset	123

5.6	(a)GT trajectory and PT of the road-agent with agent_ID = 11 of the Argoverse Dataset (b) GT trajectory and PT of the road-agent with agent_ID = 27 of the Argoverse Dataset	123
5.7	Architecture of the Motion Planning Model	124
5.8	Ego vehicle turning at road curve	129
5.9	Variation of forward speed, steer and throttle parameters with time for the entire route of the scenario two of whose instances are shown in (a). (b) and (c). The variations of the forward speed of the vehicle and the vehicle and the values of throttle, brake and steer, with time are also shown up to the considered instance	130
5.10	Vehicle Trajectory Map for the scenario one of whose instances are shown in Figure5.8.	130
5.11	Ego vehicle overtaking another vehicle	132
5.12	Ego vehicle stopping at a STOP road-sign	132
5.13	Variation of forward speed, steer and throttle parameters with time for the entire route of the scenario two of whose instances are shown in (a). (b) and (c). The variations of the forward speed of the vehicle and the vehicle and the values of throttle, brake and steer, with time are also shown up to the considered instance	132
5.14	Vehicle Trajectory Map for the scenario one of whose instances are shown in Figure5.8.	133
5.15	Comparison of the average number of different types of infractions incurred during testing for the two models	134

List of Tables

2.1	Comprehensive details of few existing methods for object detection using Deep-learning.	34
2.2	Comparison details of few existing methods for Multi-Object Tracking using Deep-Learning Approaches	43
2.3	Comprehensive details of few existing methods for Trajectory Prediction using deep-learning methods	46
2.4	Comparison details of few existing methods for Motion Planning using Deep-Learning Approaches	57
2.5	Details of Training, Testing and validation set for KITTI, BDD, Waymo, MOT variants, Argoverse, Apolloscape, Lyft Datasets	64
3.1	Comparison of performances on KITTI validation set	85
3.2	Comparison of performances on a different backbone architectures . .	85
3.3	Comparison of performances on different backbone architectures . . .	87
3.4	Comparison of different attention mechanism efficiency in terms of number of parameters, floating point operations per second (FLOPs) and Top-1 accuracy	88
4.1	Performance Comparison between the proposed method and other latest track association MOT approaches on the 2DMOT15 Dataset. The last column has frames/second that measure the speed of the model	99
4.2	Performance Comparison between the proposed method and other latest track association MOT approaches on the MOT16 Dataset. The last column has frames/second that measure the speed of the model	100
4.3	Performance Comparison between the proposed method and other latest track association MOT approaches on the MOT17 Dataset. The last column has frames/second that measure the speed of the model	100
4.4	Performance Comparison between the proposed method and other latest track association MOT approaches on the MOT20 Dataset. The last column has frames/second that measure the speed of the model	101
4.5	Performance Comparison between the proposed method and other latest track association MOT approaches on the Waymo dataset . . .	101

4.6	Performance of the proposed model with the different backbone network through ablation study	102
5.1	Comparison of the metrics values obtained for Apolloscape and Lyft datasets using for the model with the other existing models	118
5.2	Comparison of the metrics values obtained for Argoverse dataset with prediction interval = 5 and the other existing models	119
5.3	Change in time required for one-second-ahead prediction with increasing value of N	119
5.4	Summarizes the mean values and standard deviation of the three metrics, viz. DS, RC and IM, obtained using the model and the other image-based existing models on the CARLA simulator	131

Abbreviations

AVs	A utonomous V ehicles
NHTSA	N ational H ighway T raffic S afety A dministration
CNN	C onvolutional N eural N etwork
CARLA	C ar L earning T o A ct
C2ES	C enter F or C limate A nd E nergy S olutions
WHO	W orld H ealth O rganization
RPN	R egion P roposal N etwork
RNN	R ecurrent N eural N etwork
LSTM	L ong S hort T erm M emory
ResNet	R esidual N etwork.
TCN	T emporal C onvolutional N etwork
GRU	G ated R ecurrent U nit
RBPF	R ao B lackwellized P article F ilter
SSPF	S caling S eries P article F ilter
LBP	L ocal B inary P attern
CRR	C orrect R ecognition R ate
SVM	S upport V ector M achine
FTP	F ourier T emporal P yramid
SGD	S tochastic G radient D escent
HBRNN	H ierarchically B idirectional R ecurrent N eural N etworks
IDT	I mproved D ense T rajectory
MEI	M otion E nergy I mage

MHI	Motion History Image
ReLU	Rectified Linear Unit
FC	Fully Connected
BN	Batch Normalization
NF	Number of Feature Maps
DMMs	Depth Motion Maps
PID	Proportional Integral Derivative
DARPA	Defense Advanced Research Projects Agency
UBER	Unified Best Economical Ride
LIDAR	Light Detection And Ranging
RADAR	Radio Detection And Ranging
GPS	Global Positioning System
DARPA	Defense Advanced Research Projects Agency
DNN	Deep Neural Network
GNN	Graph Neural Network
FPN	Feature Pyramid Network
YOLO	You Only Look Once
NMS	Non Maxima suppression
SSD	Single Shot MultiBox Detector
DPM	Deformable Parts Model
BERT	Bidirectional Encoder Representation From Transformer
NLP	Natural Language Processing
MSA	Multi Headed Self Attention
DATMO	Detector And Tracker Of Moving Obstacles
RANSAC	RANdom Sample Consensus
GOTURN	Generic Object Tracking Using Regression Networks

Symbols

t_f	Last Time Step
P_j	Infraction Penalty coefficient for every j infraction instance
R_i	Percentage of the i^{th} route completed
A_c	Channel Attention Feature Map
A_s	Spatial Attention Feature Map
\otimes	Element-wise Multiplication
σ	Sigmoid function
δ	Spatial Context
\mathbf{I}	Identity Matrix
\mathbf{A}	Adjacency Matrix
μ	Threshold Parameter
\mathbf{d}	Euclidean distance
θ'_r	Chebyshev Coefficient
ξ_i^t	Encoder Vector
p^t	Positional Encoding Vector
$M_{x,y}$	Heatmap at location (x,y)
λ	Eigen Value
$\delta(M_{i,j})$	Local Spatial Context
O_t^*	Motion Vector
$ V $	Number of Vertices
$ E $	Number of Edges

Symbols

\mathbb{R}	Real Number
L_p	Laplacian of Graph
ω	Waypoint
γ	Desired Speed
α	Steering angle

PREFACE

Due to technological advancements, massive multimedia data is typically available in image and video formats. Currently, images and videos are used in many complex applications such as human-computer interaction, autonomous security systems, 3D scene understanding, sports performance analysis, etc. The autonomous vehicle is one application that uses images to analyze its surroundings. Autonomous Vehicles (AVs) have various components for execution. The Sensors are used to capture the data from the surroundings. The perception module converts this raw data into meaningful information. The mapping and localization module is used to localize the vehicles in real-world coordinates and destination locations through the Global Positioning System (GPS). The Planning module uses this meaningful information to plan their actions. The control module actuates the control through the steering, brake and accelerators.

This thesis focuses on the perception and planning tasks of the autonomous vehicle using computer vision and deep learning techniques. The various tasks associated with the perception module include Object Detection, Object Tracking and Trajectory Prediction. The various tasks associated with the planning module include Trajectory Planning, Motion Planning, and Behavior planning. The significant challenges associated with these modules include highly dynamic background, gradual and abrupt illumination changes, camera jitter, shadows, reflections, and weather conditions that may cause false detection and may become a big reason for

wrong decisions in the navigation of AVs. Most of the existing methods have been reported for the different modules of AVs. This thesis addresses some of the challenges and issues arising from these sub-modules of AVs. The problem statement of the research is defined as the Study of existing methods, analyzing their merits and demerits, implementation of the algorithm and proposed new methods and models for perception and planning modules using deep learning-based approaches for providing reliable navigation of the AVs.

First, this thesis presents a detailed literature survey, including a survey of various modules and sub-modules of AVs and a study on the evaluation of modern datasets for object detection, object tracking, trajectory prediction and motion planning. Further, the hierarchy of different approaches has been discussed and research gaps have been identified. Finally, a detailed list of the dataset used for training and evaluation proposed models has been presented, followed by a discussion on performance measures used in this thesis.

The first proposed model of this thesis is for object detection, which is based on the attention mechanism. High accuracy ensures the vehicle for collision-free navigation tasks, while the faster detection speed helps make decisions quickly. In this thesis, the proposed model is a single-stage object detection that provides faster detection. The channel attention mechanism provides more fine-grained features and emphasizes that 'what' is a semantic part of a given input. Apart from the channel attention mechanisms, spatial attention emphasizes 'where' is meaningful information that is working to boost the performance of the attention block for accurate

detection. The experimental result shows that the proposed model surpasses the state-of-the-art techniques for the KITTI and BDD datasets.

Further, the research reported in this thesis has been extended for Multi-Object Tracking as a second contribution. The proposed model, An end-to-end Hybrid model for Multi-Object Tracking, involves detection-based tracking, which generally requires a scale-up of two subtasks: motion estimation and re-identification. The proposed model utilized dense-optical flow for motion estimation. The relative scale of boundary boxes is formulated to find the maximum likelihood of a couple of correct matches. The model repeats this for unmatched detection to match another trajectory (trajectories not assigned in current frames). The detection that this process cannot match is initialized as a new trajectory. The achieved state-of-the-art results of the tasks allow for high accuracy of tracking with detection and surpasses existing state-of-the-art methods by a considerable margin on MOT and Waymo publicly available datasets (Multi-Object Tracking, Waymo).

Finally, in this thesis, two methods have been proposed for trajectory prediction and motion planning. A Graph Neural Network with RNNs-based Trajectory Prediction of dynamic Agents for Autonomous Vehicles is proposed for trajectory prediction. A Semantic Supervision Guided Image-based Motion Planning of the Autonomous Vehicles method is proposed for motion planning. The trajectory prediction model extracts the spatial-temporal features using a graph neural network and predicts the long-term trajectory using LSTM (Long-short term memory). Experiments show that the proposed model effectively captures comprehensive

Spatio-temporal correlations through modeling GNN with temporal features for TP and consistently surpasses the existing state-of-the-art methods on three publicly available datasets (Lyft, Argoverse, Apolloscape) for trajectory. Compared to prior methods, The proposed model performs better for sparse datasets than for dense datasets.

The motion planning task utilized multi-view images and a CNN (Convolutional Neural Network) model for feature extraction and the number of GRUs (Gated Recurrent Units) to generate the waypoints. The ego vehicle generated a sequence of coordinates representing the waypoints in the predicted path of the vehicle for the upcoming few time steps. The model has generated the waypoints using GRUs, which are used as input for the PID (Proportional Integral Derivative) controller. The PID controller is used an inverse dynamic algorithm to derive the values of the driving parameters, like steering angle, throttle, and brake value, from the coordinates of the waypoints. The proposed model has shown an improvement in the Route Completion and Driving Score metrics that outperform state-of-the-art methods on this simulator's dataset.