

**AUTONOMOUS NAVIGATION OF MOBILE ROBOT IN A
DYNAMIC ENVIRONMENT USING DEEP
REINFORCEMENT LEARNING**

Dhivyadharshan Seetharaman

208814P

Master of Science in Industrial Automation

Department of Electrical Engineering
Engineering Faculty

University of Moratuwa
Sri Lanka

March 2023

**AUTONOMOUS NAVIGATION OF MOBILE ROBOT IN A
DYNAMIC ENVIRONMENT USING DEEP
REINFORCEMENT LEARNING**

Dhivyadharshan Seetharaman

208814P

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science in Industrial Automation

Department of Electrical Engineering
Engineering Faculty

University of Moratuwa
Sri Lanka

March 2023

DECLARATION

I declare that this is my own work, and this thesis does not incorporate without Acknowledgment of any material previously submitted for a Degree or Diploma in any other University or institute of higher learning, and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgment is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

The above candidate has carried out research for the Master's thesis under my supervision. I confirm that the declaration made above by the student is true and correct.

Prof. Chandima D. Pathirana

Signature of the Supervisor:

Date:

DEDICATION

I would like to dedicate this thesis to the people who aided me when I stumbled and lost my way. Especially, My parents and sister, school teachers, university professors, and lecturers.

ACKNOWLEDGMENTS

I would like to offer my sincerest gratitude and heartfelt indebtedness to my supervisor Prof.Chandima D.Pathirana, who has given me the opportunity to pursue my second-year research of MSC in Industrial Automation under his supervision and guidance. I still remember the day I joined the University of Moratuwa for my post-graduation with many knowledge gaps in my field. My supervisor, Prof.Chandima D.Pathirana, put endless efforts and deep patience into answering my numerous basic questions to fill my knowledge gaps and provided clarity on subjects by explaining the topic very deeply in this taught course. I am really obliged to him for putting additional effort into explaining countless topics in the subjects of Electrical Machines and Drives, Industrial Electronics, Control Systems Designs, and Mechatronics during my postgraduate diploma, which I missed gaining knowledge during my undergraduate period. Especially he guided me to approach the research problem in a step-by-step manner when I was trying to solve the problem with a quantum leap, and his counseling showed me the right direction when I was on the wrong way up. This research would have never been possible without his depth knowledge and proper guidance. His advice helped me approach the complex problem by dividing it into small problems in order to successfully accomplish the goal of this research within a given period.

I want to thank Prof. A.G. Buddhika P. Jayasekara, who untangled me when I was a jam at a randomization problem of python code. His advice helped me avoid getting stuck in local minima and confidently reach the global minima in my research. I would also like to thank my progress review panel members, Dr.A.M.H.S. Abeykoon and Dr.Velmanickam Logeeshan for their penetrating comments during progress review presentations which incentivized me to broaden my research and reshape it to generate valuable and sensible outputs. My special thanks go to Dr.Velmanickam Logeeshan for his friendly advice and guidance, which helped me enormously to mitigate psychological barriers during my research. I am sincerely obliged to the Department of Electrical Engineering for allowing me to access the required number of computers in the computer lab to train my Deep Reinforcement learning model for a more extended period during this research. I would like to thank all robotics lab-mates of the University of Moratuwa for their instructions and encouragement, special thanks to H.M.Ravindu T.Bandara, who has taught me significantly about the Deep Neural Networks and the deployment of the machine learning model.

Finally, no words can ever be strong enough to express my gratitude to my parents and my sister because of their great hope on me, which constantly stimulates me to take at least a few steps toward my goals. I would also like to thank God for giving me spiritual support in following my career path when I stumble.

ABSTRACT

Autonomous navigation of a mobile robot in a dynamic environment is a highly challenging application because the path to the goal frequently changes due to unpredictable movements of humans with different velocities. Deep Reinforcement completely trains a model in a simulator using a trial-and-error technique by exploring and collecting required data automatically and cheaply from the customized environment. This research develops a customized environment with robot and pedestrian models in OpenAI Gym, replicating humans' real-world measurements and motion patterns. A mathematical model was developed to encapsulate the navigation norms of humans' to teach the robot about socially compliant routes using a reward function in order to smooth the robot's navigation in a dynamic environment. A recently evolved algorithm, H-PPO, has been selected to train the model by considering the agent's hybrid action space, which consists of discrete actions parametrized by continuous values.

First, the model failed to learn due over fit in a simple environment, and then it learned when the task was randomized. The various approaches have been investigated to enhance the model's generalizability as much as possible in the simulator. Finally, the agent is trained in each environment separately. Despite this research has not considered the complex scenario as randomizing the whole environment initially, the developed model was able to scrutinize the performance of the recently evolved algorithm H-PPO in obstacle avoidance applications, and the developed model can learn obstacle avoidance in a dynamic environment by respecting social norms in the long-range motion for laboratory application. However, the success rate of the model trained later in a fully randomized 3-pedestrians environment was 86.67% out of 30 episodes of testing, which is higher than the previous research [1]. Further investigation has to be carried out in future work by adding memory ability to the model in order to enhance the performance, reduce the training time and mitigate the performance collapse during the learning phase.

Keywords: Intelligent Mobile Robot, Navigation, OpenAI Gym, Deep Neural Network, Deep Reinforcement Learning, H-PPO algorithm, Machine Learning, Parametrized Action Space.

Table of Contents

DECLARATION	i
DEDICATION	ii
ACKNOWLEDGMENTS	iii
ABSTRACT	iv
LIST OF FIGURES	ix
LIST OF TABLES	xi
LIST OF ABBREVIATIONS	xii
NOMENCLATURE.....	xiii
LIST OF APPENDICES	xiv
CHAPTER 1	1
INTRODUCTION	1
1.1 Background and Rationale	1
1.2 Motivation.....	2
1.3 Problem Identification.....	2
1.4 The Aim and the objectives.....	3
1.5 Limitations of the research.....	3
1.6 Thesis summary	4
1.7 Thesis outline	4
CHAPTER 2	6
LITERATURE REVIEW.....	6
2.1 Selection of machine learning tool.....	6
2.2 Deep Reinforcement learning	7
2.2.1 The parallelization technique in Deep Reinforcement Learning	8
2.3 Recent Researches in the DRL.....	10
2.3.1 Socially attentive network based robot navigation	10
2.3.2 The robot navigation in a cluttered environment.....	11
2.3.3 Socially aware robot.....	12
2.3.4 A3C-based robot navigation in a simulator	13

2.3.5 Graph Convolutional Network based robot navigation	14
2.4 The Mechanism to respect pedestrian-norms.....	15
2.5 Selection of algorithm.....	16
2.6 Detailed Description of the selected algorithm.....	18
2.7 The architecture of the deep reinforcement learning	22
CHAPTER 3	24
OPTIMAL DESIGN	24
3.1 Conceptual Designs.....	24
3.1.1 Conceptual Design 1	24
3.1.2 Conceptual Design 2	25
3.1.3 Conceptual Design 3	26
3.2 Selection of the Conceptual Design	27
3.2 Investigation of each conceptual design from the perspective of each objective.28	
3.2.1 Design	28
3.2.2 Cost	29
3.2.3 Effectiveness	30
3.3 Novelties of the research.....	31
CHAPTER 4	33
OVERALL SYSTEM AND MODELLING.....	33
4.1 System of systems and block diagram	33
4.2 Functional flow chart of the overall system.....	34
4.3 Mathematical model.....	36
4.3.1 Rendering of the robot and pedestrians in the simulator.....	36
4.3.2 Reinforcement Learning.....	39
4.3.3 Definition of collision avoidance	41
4.3.4 Reward Function (The Expert Knowledge)	42
4.3.5 Mathematical constraints to respect pedestrian-norms	46
CHAPTER 5	49
SOFTWARE IMPLEMENTATION	49
5.1 Customizing Simulation Environment.....	49

5.1.1	Required tools to create a custom environment	49
5.1.2	Description of the environment.....	52
5.2	The selected algorithm	53
5.2.1	Algorithm: On-Policy PPO with clipping	54
5.2.2	The explanation steps of the algorithm	55
CHAPTER 6	57
TRAINING AND DISCUSSIONS	57
6.1	Training in a simple environment	57
6.1.1	Graph analysis	58
6.1.2	Parameter tuning	60
6.2	Randomizing the task.....	64
6.3	Poor generalizability of the model	66
6.3.1	Increasing the stop value	67
6.3.2	Modification of collision penalization	67
6.3.3	Altering the pedestrian-norm-inducing weight	68
6.3.4	Parallelization technique	70
6.4	Final training and discussions	75
6.4.1	Training the agent in the environment with one pedestrian.....	75
6.4.2	Training the agent in the environment with two pedestrians	76
6.4.3	Training the agent in the environment with three pedestrians	78
6.4.4	Training the agent in the environment with four pedestrians	80
6.4.5	Training the agent in a fully randomized environment with three pedestrians	84
6.5	Testing and Comparison	85
6.5.1	Testing.....	85
6.5.2	Comparison	86
CHAPTER 7	92
CONCLUSIONS AND FUTURE WORKS	92
7.1	Conclusions.....	92
7.2	Future works	93

APPENDIX..... 95
REFERENCES..... 96

LIST OF FIGURES

Figure 2.1: The Reinforcement Learning Model	8
Figure 2.2: Overview of the LM-SARL. Adapted from [9].....	10
Figure 2.3: The network architecture of the SOADRL. Adapted from [10].....	11
Figure 2.4: Social-norm inducing mechanism (The top and bottom rows depict the left-handed and the right-handed rules, respectively). Adapted from [11].....	12
Figure 2.5: Motion learning using CNN. Adapted from [14]	13
Figure 2.6: The interactions between humans and robot as a graph. Adapted from [15] ..	14
Figure 2.7: The H-PPO algorithm outweighing other threes. Adapted from [38]	17
Figure 2.8: Hybrid actor-critic architecture for parameterized action space. Adapted from [36]	18
Figure 2.9: The architecture of the actor-critic reinforcement learning. Adapted from [47]	23
Figure 2.10: The operation of the Proximal Policy Optimization (PPO) algorithm based on a clipped objective. Adapted from [48]	23
Figure 3.1: Overall view of the conceptual design 1	24
Figure 3.2: Overall view of the conceptual design 2	25
Figure 3.3: Overall view of the conceptual design 3	26
Figure 4.1: The overall block diagram to depict where the trained model is going to be installed	34
Figure 4.2: The robot model created in OpenAI gym.....	36
Figure 4.3: The pedestrian model created in OpenAI gym according to the measurements of real-world's pedestrians. Adapted from [24], [46].....	37
Figure 4.4: Scaled-down of models in the simulator	38
Figure 4.5: The 2D depiction of the environment.....	38
Figure 4.6: Collision distance	41
Figure 4.7 : The outline of the collision circle	41
Figure 4.8: The robot's angles for the direction related reward function	43
Figure 4.9: Encapsulating the passing behavior.....	46
Figure 4.10: Encapsulating the crossing behavior	47
Figure 4.11: Encapsulating the overtaking behavior.....	48
Figure 5.1: Customized Environment in the OpenAI gym	50
Figure 5.2: The parameterized action space (Hybrid action space)	53
Figure 6.1: The reward mean graph of the agent's training in the environment with two pedestrians.....	58
Figure 6.2: The average entropy loss of the agent's training in the environment with two pedestrians.....	59
Figure 6.3: The maximum reward graph for six different maximum number of steps....	60
Figure 6.4: Test results of the agent in the environment with one pedestrian	76
Figure 6.5: Test results of the agent in the environment with two pedestrians.....	78
Figure 6.6: Test results of the agent in the environment with three pedestrians.....	80
Figure 6.7: Test results of the agent in the environment with four pedestrians	82

Figure 6.8: The comparison of two mean reward graphs of trainings with stop values 1.64 and 1.65 in the environment with four pedestrians83

Figure 6.9: The comparison of two average entropy loss graphs of trainings with stop values 1.64 and 1.65 in the environment with four pedestrians83

Figure 6.10: The randomization setup of the environment85

Figure 6.11: Comparing the mean rewards of the best-trained models in each environment87

Figure 6.12: Comparing the average entropy losses of the best-trained models in each environment.....87

Figure 6.13: The Failure rates of each models without considering the fully randomized 3- pedestrians’ environment.....88

Figure 6.14: Success rate.....89

LIST OF TABLES

Table 3.1: Analyzing each conceptual design from the Design point of view	28
Table 3.2 : Analyzing each conceptual design from the Cost point of view	29
Table 3.3: Analyzing each conceptual design from the Effectiveness point of view	30
Table 3.4: Novelties of the research.....	31
Table 5.1: On-Policy PPO Algorithm with Clipping. Adapted from [27]	54
Table 6.1: Manually searching the learning rate	63
Table 6.2: The recorded training results of the model using a secondly defined reward function	65
Table 6.3: Analyzing the test results by modifying the collision penalization weight	67
Table 6.4: The test results of the model by gradually increasing the stop value for a norm weight of 0.02	69
Table 6.5: The test results of the model by modifying the maximum turning angle from 20° to 10°	70
Table 6.6: The recorded test results while increasing the parallel environment.....	71
Table 6.7: The test results of the re-trained model for a fixed parallel environment.....	72
Table 6.8: The test results of the re-trained model in the fixed parallel environment with constant stop value and zero pedestrian-norm weight	73
Table 6.9: The finalized parallel environment	74
Table 6.10: The selected parameters to train the model in the environment with 1-ped.	75
Table 6.11: Gradually increasing the stop value in the environment with two pedestrians	77
Table 6.12: Gradually increasing the stop value in the environment with three pedestrians	79
Table 6.13: Gradually increasing the stop value in the environment with four pedestrians	81
Table 6.14: Testing the higher number of pedestrians' models in the lower number of pedestrians' environments.....	86

LIST OF ABBREVIATIONS

API	Application Programming Interface
A3C	Asynchronous Advantage Actor-Critic Learning
CNN	Convolutional Neural Network
CADRL	Collision Avoidance with Deep Reinforcement Learning
DRL	Deep Reinforcement Learning
DNN	Deep Neural Network
DDPG	Deep Deterministic Policy Gradient
GNRON	Goal Not Reachable when Obstacles are Nearby
GCN	Graph Convolutional Network
GOSELO	Goal-Directed Obstacle and Self-Location Map for Robot Navigation using Reactive Neural Networks
GRU	Gated Recurrent Unit
H-PPO	Hybrid Proximal Policy Optimization
HyAR	Hybrid Action Representation
LiDAR	Light Detection and Ranging
LSTM	Long Short-Term Memory
MP-DQN	Multi-Pass Deep Q-Networks
MLP	Multi-Layer Perceptron
PPO	Proximal Policy Optimization
P-DQN	Parameterized Deep Q-Networks
PAMDPs	Parametrized Action Markov Decision Processes
RL	Reinforcement Learning
SLAM	Simultaneous Localization and Mapping
SOADRL	Social Obstacle Avoidance using Deep Reinforcement Learning
SA-CADRL	Socially Aware Collision Avoidance with Deep Reinforcement
TD	Temporal Difference

NOMENCLATURE

$R(\tau)$	Total Discounted Reward
T	Horizon (Episode Length)
θ	Heading angle of the robot
x_r, y_r	Co-ordinates of the robot
v_r	Speed of the robot
r	The outer radius of the robot
d_G	The distance between the center point of the robot and the goal
p_{ix}, p_{iy}	Co-ordinates of pedestrians
v_p	Constant speed of pedestrians
θ_i	Heading angle of pedestrians
d_{ci}	Collision distance between the robot and pedestrians
θ_{Rn}	Robot's norm-inducing angle
θ_{Pin}	Pedestrians' norm-inducing angle
a_t	The complete action
a	The chosen discrete action
x_a	The chosen continuous parameters for discrete action a
$\pi_{\theta_{old}}$	The old policy
$A_t^{\pi_{\theta_{old}}}$	Advantage function calculated by the old policy
ϵ	The clip ratio

LIST OF APPENDICES

Appendix – A	CD/DVD.....	94
--------------------	-------------	----