ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΟΜΕΑΣ ΜΕΤΑΦΟΡΩΝ ΚΑΙ ΣΥΓΚΟΙΝΩΝΙΑΚΗΣ ΥΠΟΔΟΜΗΣ

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΗΣ ΑΛΛΗΛΕΠΙΔΡΑΣΗΣ ΔΙΑΣΥΝΔΕΔΕΜΕΝΩΝ ΑΥΤΟΝΟΜΩΝ ΟΧΗΜΑΤΩΝ ΜΕ ΠΕΖΟΥΣ ΜΕ ΧΡΗΣΗ ΑΝΤΙΣΤΡΟΦΗΣ ΕΝΙΣΧΥΤΙΚΗΣ ΜΑΘΗΣΗΣ

Διπλωματική Εργασία

ΜΑΓΔΑΛΗΝΗ ΚΑΡΑΤΑΡΑΚΗ

Επιβλέπουσα: Ελένη Ι. Βλαχογιάννη, Καθηγήτρια Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά, θα ήθελα να ευχαριστήσω την κ. Ελένη Βλαχογιάννη, Καθηγήτρια της Σχολής Πολιτικών Μηχανικών Ε.Μ.Π., που μου έδωσε τη δυνατότητα να ασχοληθώ με ένα ιδιαίτερα ενδιαφέρον και σύγχρονο θέμα. Εκτιμώ την πολύτιμη καθοδήγηση που μου προσέφερε κατά την διάρκεια εκπόνησης της Διπλωματικής μου εργασίας.

Επιπλέον, θα ήθελα να ευχαριστήσω την Υποψήφια Διδάκτορα Φωτεινή Ορφανού, που μου παρείχε τα δεδομένα για την υλοποίηση της Διπλωματικής μου εργασίας και με συμβούλευσε κατά τη διάρκεια της ανάλυσης των δεδομένων και της συγγραφής της εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου, που με στήριξαν και πίστεψαν σε εμένα, τόσο κατά την διάρκεια των σπουδών μου, όσο και κατά την εκπόνηση της Διπλωματικής μου εργασίας.

Αθήνα, Οκτώβριος 2023

Μαγδαληνή Καραταράκη

Μοντελοποίηση της αλληλεπίδρασης διασυνδεδεμένων αυτόνομων οχημάτων με πεζούς με χρήση αντίστροφης ενισχυτικής μάθησης

Μαγδαληνή Καραταράκη

Επιβλέπουσα: Ελένη Ι. Βλαχογιάννη, Καθηγήτρια Ε.Μ.Π.

ΣΥΝΟΨΗ

Η παρούσα διπλωματική εργασία στοχεύει στην ανάπτυξη μιας στρατηγικής οδήγησης για τα Συνδεδεμένα Αυτοματοποιημένα Οχήματα (CAVs) κατά την διάρκεια αλληλεπιδράσεων με πεζούς. Σκοπός είναι η δημιουργία ενός αποτελεσματικού και ασφαλούς πλαισίου λήψης αποφάσεων για αυτόνομα οχήματα με τη χρήση αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας (Max-Ent IRL). Η συνάρτηση ανταμοιβής εξάγεται από τις παρατηρούμενες τροχιές των οχημάτων που λαμβάνονται μέσω ενός εικονικού πειράματος οδήγησης που διεξήχθη στην Καρλσρούη της Γερμανίας με την χρήση του προσομοιωτή CARLA. Τα αποτελέσματα της έρευνας συμβάλλουν στην ανάπτυξη των συστημάτων αυτόνομης οδήγησης, διευκολύνοντας την ενσωμάτωση των αυτόνομων οχημάτων σε αστικά περιβάλλοντα, δίνοντας παράλληλα προτεραιότητα στην ασφάλεια των πεζών. Συνιστάται περαιτέρω έρευνα για την επέκταση της εκπαίδευσης του αλγορίθμου σε ποικίλα σενάρια και διαφορετικές πιθανές καταστάσεις.

Λέξεις κλειδιά: αυτόνομα οχήματα, αντίστροφη ενισχυτική μάθηση, μέγιστη εντροπία, προσομοίωση, συμπεριφορά οχήματος, πεζός, αλληλεπίδραση, οδική ασφάλεια

Modeling the Interaction of Connected Autonomous Vehicles with Pedestrians using Inverse Reinforcement Learning

Magdalini Karataraki

Supervisor: Eleni I. Vlahogianni, Professor NTUA

ABSTRACT

This Diploma thesis aims to develop a driving strategy for Connected Automated Vehicles (CAVs) during interactions with pedestrians. The objective is to create an efficient and safe decision-making framework for autonomous vehicles using maximum entropy inverse reinforcement learning (Max-Ent IRL). The reward function is derived from observed vehicle trajectories obtained through a virtual driving experiment conducted in Karlsruhe, Germany with the use of CARLA simulator. The research outcomes contribute to the advancement of autonomous driving systems, facilitating the integration of autonomous vehicles into urban environments while prioritizing the safety of pedestrians. Further research is recommended to expand the training of the algorithm in diverse scenarios and different potential situations.

Keywords: autonomous vehicles, inverse reinforcement learning, maximum entropy, simulation, vehicle behavior, pedestrian, interaction, road safety

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια τα αυτόνομα οχήματα έχουν αρχίσει να εισάγονται σταδιακά στην κυκλοφορία. Βασικό κίνητρο της έρευνας πάνω στα αυτόνομα οχήματα είναι η αύξηση της ασφάλειας στο δρόμο και η αναπαραγωγή της ανθρώπινης συμπεριφοράς από τα AV για την ευκολότερη ενσωμάτωσή τους στην κυκλοφορία. Στόχος της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη μιας στρατηγικής οδήγησης για διασυνδεδεμένα αυτόνομα οχήματα σε αλληλεπίδραση με πεζούς με την χρήση της αντίστροφης ενισχυτικής μάθησης. Πιο συγκεκριμένα, μέσα από τις παρατηρούμενες τροχιές των αυτόνομων οχημάτων με την παρουσία πεζού, το μοντέλο προσπαθεί να εξάγει την συνάρτηση ανταμοιβής που περιγράφει βέλτιστα την συμπεριφορά των AV, με τη χρήση του αλγόριθμου αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας.

Στο πλαίσιο της βιβλιογραφικής ανασκόπησης αναδείχθηκαν τα πλεονεκτήματα και οι προκλήσεις της χρήσης αντίστροφης ενισχυτικής μάθησης για την ανάπτυξη στρατηγικών οδήγησης που βασίζονται σε επιδείξεις ειδικών, καθώς και η περιορισμένη έρευνα πάνω στη χρήση IRL σε σενάρια με αυτόνομα οχήματα υπό την παρουσία πεζών. Τα βασικά πλεονεκτήματα που προέκυψαν είναι η αποτύπωση των ανθρώπινων προτιμήσεων μέσα από τις παρατηρούμενες τροχιές και ο αυτόματος σχεδιασμός της συνάρτησης ανταμοιβής. Επιπλέον, οι αλγόριθμοι IRL οδηγούν σε ρεαλιστικές απεικονίσεις, διότι βασίζονται σε πραγματικά δεδομένα, και άρα σε ταχύτερη εκπαίδευση του μοντέλου. Τέλος, το μοντέλο αναπτύσσει μια ασφαλή στρατηγική αλληλεπίδρασης με τους χρήστες του δρόμου οδηγώντας στην ευκολότερη αποδοχή του.

Η συλλογή των δεδομένων έγινε μέσω ενός πειράματος εικονικής πραγματικότητας που πραγματοποιήθηκε από τον FZI στην Καρλσρούη της Γερμανίας με σκοπό την παρατήρηση της συμπεριφοράς του αυτόνομου οχήματος, όταν ένας πεζός επιχειρεί να διασχίσει τον δρόμο σε διαφορετικά σενάρια και καταστάσεις. Το πείραμα πραγματοποιήθηκε με την χρήση του προσομοιωτή CARLA και η εξαγωγή των δεδομένων έγινε απευθείας από το περιβάλλον. Οι τροχιές των οχημάτων και τα υπόλοιπα κυκλοφοριακά μεγέθη που υπολογίστηκαν, χρησιμοποιήθηκαν στη συνέχεια για την διαμόρφωση του τελικού αλγόριθμου. Από έναν χάρτη υψηλής ευκρίνειας (HD) του οδικού δικτύου σε μορφή Lanelet, υπολογίστηκαν τιμές, όπως η επιτάχυνση, η χρονική απόσταση, το κενό από το προπορευόμενο όχημα, η πλευρική μετατόπιση προς το κέντρο της λωρίδας και η απόσταση από πλευρικά αντικείμενα εκτός της λωρίδας. Έπειτα, καθορίστηκαν οι καταστάσεις με βάσεις τις ταχύτητες και τις αποστάσεις AV και πεζού. Στην ανάπτυξη του μοντέλου χρησιμοποιήθηκαν συνολικά 16 τροχιές για την εκπαίδευση του αλγορίθμου.

Το περιβάλλον χαρακτηρίζεται από 12 καταστάσεις με βάση την ταχύτητα του οχήματος, την χωρική απόσταση και την διαφορά ταχύτητας μεταξύ οχήματος και πεζού. Για τον χωρισμό των χαρακτηριστικών σε επιμέρους επίπεδα χρησιμοποιήθηκε η μέθοδος ομαδοποίησης k-means. Οι διαθέσιμες ενέργειες περιλαμβάνουν τις ενέργειες που μπορεί να κάνει το αυτόνομο όχημα και διακρίνονται σε επιτάχυνση δύο επιπέδων και επιβράδυνση τριών επιπέδων- ομαλή, μέση και απότομη. Τα κρίσιμα όρια για την επιτάχυνση και την επιβράδυνση ορίστηκαν με βάση την ανάλυση K-means και λαμβάνοντας υπόψη τα όρια που περιγράφονται στην βιβλιογραφία.

Για την διαμόρφωση του τελικού αλγορίθμου αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας έγιναν δοκιμές για την έρευση των παραμέτρων που οδηγούν σε σύγκλιση. Στον τελικό αλγόριθμο, χρησιμοποιήθηκε ο εκπτωτικός παράγοντας 0.9, ο αριθμός των εποχών είναι 35 και ο ρυθμός εκμάθησης ισούται με 0.02. Με βάση αυτές τις τιμές προέκυψε το τελικό διάγραμμα των βαρών της συνάρτησης ανταμοιβής. Ο αλγόριθμος εκπαιδεύτηκε ικανοποιητικά και μπορεί να χρησιμοποιηθεί για τον σχεδιασμό μιας πιο ασφαλής στρατηγικής οδήγησης για αυτόνομα οχήματα σε περιβάλλοντα αλληλεπίδρασης με πεζούς.

Πιο συγκεκριμένα, ο αλγόριθμος μπορεί να προβλέψει σωστά τις συμπεριφορές των οχημάτων σε συνάρτηση με τους πεζούς, όπως φαίνεται από τα βάρη της συνάρτησης ανταμοιβής. Επιλέγει μικρότερες ταχύτητες υπό την παρουσία πεζών, προτιμά μεγαλύτερα χωρικά κενά με τους πεζούς, όπως και μεγαλύτερη διαφορά ταχύτητας. Επιπλέον, δίνει τις ανάλογες ανταμοιβές σε κάθε κατάσταση με βάση την σημασία των χαρακτηριστικών που τις αποτελούν, όπως φαίνεται από το Διαγρ.7. Όλα αυτά οδηγούν στη διαμόρφωση μιας ασφαλής στρατηγικής οδήγησης που ανταποκρίνεται στις παρεχόμενες τροχιές των AV.

Ο αλγόριθμος αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας (Max-Ent IRL) που αναπτύχθηκε μπορεί να εκπαιδευθεί περαιτέρω με την χρήση περισσότερων δεδομένων και να αξιολογηθεί σε πραγματικά δεδομένα. Με τον τρόπο αυτό θα είναι σε θέση να αντιμετωπίζει περισσότερες καταστάσεις και να προβεί σε περισσότερες ενέργειες. Επίσης, μπορούν να προστεθούν στο μοντέλο περισσότερες πληροφορίες για το περιβάλλον, ώστε να μπορεί να λάβει καλύτερες αποφάσεις με βάση αυτές πχ καιρικές συνθήκες, προφίλ πεζών. Οι αλγόριθμοι αντίστροφης ενισχυτικής μάθησης έχουν την δυνατότητα περαιτέρω μάθησης και μετά το πέρας της εκπαίδευση τους, το οποίο τους καθιστά ιδανικούς για χρήση σε αυτόνομα οχήματα για την εκμάθηση των προτιμήσεων των οδηγών και για την προσαρμογή τους σε καινούρια δεδομένα που θα παρουσιαστούν.

Η παρούσα Διπλωματική εργασία αποτελεί μία από τις πρώτες έρευνες που χρησιμοποιούν την μέθοδο μέγιστης εντροπίας για την πλοήγηση αυτόνομων οχημάτων με την παρουσία πεζών. Η μέθοδος Max-Ent IRL έχει σημαντικές δυνατότητες για την ανάπτυξη στρατηγικών οδήγησης για AV με την παρουσία πεζών. Αξιοποιώντας τροχιές ειδικών και μεγιστοποιώντας την εντροπία, το Max-Ent IRL μπορεί να εξάγει αμερόληπτες συναρτήσεις ανταμοιβής, επιτρέποντας στα αυτόνομα οχήματα να μαθαίνουν ασφαλείς και κοινωνικά αποδεκτές συμπεριφορές οδήγησης γύρω από τους πεζούς. Αυτό οδηγεί σε βελτιωμένη ασφάλεια και ευρύτερη αποδοχή των αυτόνομων οχημάτων σε πραγματικά σενάρια κυκλοφορίας.

Η χρήση αντίστροφης ενισχυτικής μάθησης για την διαμόρφωση στρατηγικών πλοήγησης για AV είναι καθοριστική, διότι μπορεί να εξάγει ανθρώπινες αξίες και προτιμήσεις από παρατηρούμενες τροχιές οχημάτων και πεζών, οδηγώντας σε μεγαλύτερη ασφάλεια και αποδοχή στον δρόμο. Η έρευνα πάνω στην αντίστροφη ενισχυτική μάθηση είναι σχετικά πρόσφατη, καθώς το πεδίο αναπτύχθηκε τα τελευταία χρόνια και οι πλήρεις δυνατότητες της δεν έχουν αξιοποιηθεί ακόμα στο έπακρον.

Με την συλλογή περισσότερων δεδομένων από πραγματικά σενάρια αλληλεπίδρασης AV με πεζούς και την αξιοποίηση των δυνατοτήτων της αντίστροφης ενισχυτικής μάθησης, θα καταστεί

δυνατό να σχεδιαστούν αυτόνομα οχήματα που θα αυξήσουν κατακόρυφα την ασφάλεια και θα εμπνεύσουν αξιοπιστία στους χρήστες του οδικού δικτύου.

ΠΕΡΙΕΧΟΜΕΝΑ

Ευρετήριο Πινάκων

Ευρετήριο Εικόνων

Ευρετήριο Διαγραμμάτων

Λεξιλόγιο

AI: τεχνητή νοημοσύνη

AV: αυτόνομο όχημα

BDM: λήψη αποφάσεων συμπεριφοράς

CAV: διασυνδεδεμένα αυτόνομα οχήματα

CNN: συνελικτικό νευρωνικό δίκτυο

DNN: βαθύ νευρωνικό δίκτυο

DQN: βαθύ Q- δίκτυο (είδος νευρωνικού δικτύου)

HDV: όχημα που οδηγείται από ανθρώπους

IRL: αντίστροφη ενισχυτική μάθηση

IVE: εμβυθιστικό εικονικό περιβάλλον

LSTM: μακροπρόθεσμη βραχυπρόθεσμη μνήμη

MDP: διαδικασία λήψεως αποφάσεων Markov

ML: μηχανική εκμάθηση

MORL: ενισχυτική μάθηση πολλαπλών στόχων

MOMDP: διαδικασία λήψεως αποφάσεων Markov πολλαπλών στόχων

Max-Ent IRL: αντίστροφη ενισχυτική μάθηση μέγιστης εντροπίας

MaaS: υπηρεσία μετακίνησης (Mobility as a Service)

MEP: αρχή μέγιστης εντροπίας

RL: ενισχυτική μάθηση

SAE: Σύλλογος Μηχανικών Αυτοκινήτων (Society of Automotive Engineers)

SORL: ενισχυτική μάθηση ενιαίου στόχου

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

1.1 ΓΕΝΙΚΑ

Η σημερινή εποχή χαρακτηρίζεται από την 4η βιομηχανική επανάσταση που περιλαμβάνει την αυτοματοποίηση και την ανταλλαγή δεδομένων στις τεχνολογίες παραγωγής. Η εξέλιξη της τεχνολογίας οδήγησε στην ανάπτυξη της Τεχνητής Νοημοσύνης (Artificial Intelligence), της Μηχανικής Μάθησης (Machine Learning) και της Επιστήμης των Δεδομένων (Data Science).

Σχεδόν το 95% των τροχαίων ατυχημάτων στην ΕΕ οφείλεται σε κάποιο βαθμό σε ανθρώπινο σφάλμα (Αυτόνομα αυτοκίνητα στην ΕΕ: από επιστημονική φαντασία...σε απτή πραγματικότητα, 2019). Τη σήμερον ημέρα υπάρχει πληθώρα επιλογών μετακίνησης και σταδιακή εισαγωγή στοιχείων έξυπνης αστικής κινητικότητας στις μετακινήσεις. Η εισαγωγή των αυτόνομων οχημάτων έχει ήδη ξεκινήσει και θα φέρει δραστικές αλλαγές στις μεταφορές στο πλαίσιο των έξυπνων πόλεων. Τα μη επανδρωμένα οχήματα θα αυξήσουν την οδική ασφάλεια και θα περιορίσουν το περιβαλλοντικό αντίκτυπο των μεταφορών, θα μειώσουν την κυκλοφοριακή συμφόρηση και θα αυξήσουν την προσβασιμότητα. Πριν όμως τα αυτόνομα οχήματα αντικαταστήσουν πλήρως τα συμβατικά, χρειάζεται να γίνουν αποδεκτά από τους χρήστες του δρόμου και να αναπτυχθούν οι κατάλληλες υποδομές και τα αντίστοιχα νομικά πλαίσια για την ασφαλή ενσωμάτωσή τους στο οδικό και μεταφορικό δίκτυο.

1.1.1 Αυτόνομα Οχήματα

Αυτόνομο (autonomous) ή αυτοοδηγούμενο (self-driving) όχημα είναι αυτό που οδηγεί τους επιβάτες στον προορισμό τους χωρίς ανθρώπινη μεσολάβηση και εξασφαλίζοντας άνεση, με την χρήση συστημάτων και αισθητήρων για την αναγνώριση του περιβάλλοντος. Τα αυτόνομα οχήματα έχουν πολλά πλεονεκτήματα όπως μεγαλύτερη ασφάλεια, οικονομία καυσίμου, μείωση αναγκών στάθμευσης και ευκολότερη μετακίνηση για τις ευπαθείς ομάδες χρηστών (ηλικιωμένοι, ΑΜΕΑ).

Υπάρχουν 6 διαφορετικά επίπεδα αυτονομίας ελέγχου σύμφωνα με τη SAE (Society of Automotive Engineers), από το Επίπεδο 0 (χωρίς αυτοματισμό οδήγησης) έως το Επίπεδο 5 (πλήρης αυτοματισμός οδήγησης). Τα επίπεδα αυτοματισμού έχουν υιοθετηθεί και από την Εθνική Διοίκηση Κυκλοφοριακής Ασφάλισης Αυτοκινητοδρόμων των Η.Π.Α. (NHTSA) και είναι τα εξής:

- **Επίπεδο 0:** Ο οδηγός έχει την πλήρη ευθύνη του οχήματος, κανένας αυτοματισμός.

- **Επίπεδο 1:** Τα συστήματα λειτουργούν βοηθητικά (cruise control, lane assist) κατά την πορεία του οχήματος, αλλά ο οδηγός έχει τον απόλυτο έλεγχο της οδήγησης.
- **Επίπεδο 2:** Η οδήγηση αυτοματοποιείται με επιπλέον συστήματα πλοήγησης (πιο ενεργό cruise control) ιδιαίτερα σε μεγάλες οδικές αρτηρίες και κατά το παρκάρισμα, όμως ο οδηγός οφείλει να ελέγχει την κυκλοφορία και να παρακολουθεί το περιβάλλον.
- **Επίπεδο 3:** Το όχημα εκτελεί την οδήγηση και την παρακολούθηση του περιβάλλοντος εξ ’ολοκλήρου (ενεργό cruise control, αυτόματο φρενάρισμα για αποφυγή σύγκρουσης, αυτόματο παρκάρισμα), όμως ο οδηγός πρέπει να επέμβει σε περίπτωση έκτακτης ανάγκης.
- **Επίπεδο 4:** Το όχημα έχει υψηλή αυτοματοποίηση και μπορεί να κινείται μέσα σε μία χαρτογραφημένη περιοχή, υπό συγκεκριμένες συνθήκες, χωρίς ανθρώπινη παρέμβαση.
- **Επίπεδο 5:** Το όχημα είναι πλήρως αυτοματοποιημένο και μπορεί να κινηθεί σε οποιαδήποτε περιοχή, χωρίς την απαίτηση οδηγού.

Μια άλλη ονομασία που δίνεται σε αυτά τα οχήματα είναι το CAVs, δηλαδή συνδεμένα ή διασυνδεδεμένα αυτόνομα οχήματα. Τα οχήματα αυτά έχουν τη δυνατότητα να επικοινωνούν με τα κοντινά οχήματα (V2V) και με το περιβάλλον (οδικές υποδομές, βάσεις δεδομένων) (V2I) γύρω τους, π.χ. με φανάρια, καθιστώντας πιο εύκολη την μεταβίβαση πληροφοριών και βελτιστοποιώντας τον σχεδιασμό πορείας. Επιπλέον, αυξάνεται η ασφάλεια και μειώνεται η εκπομπή καυσαερίων. Οι τεχνολογίες αυτοματοποίησης και συνδεσιμότητας είναι αλληλένδετες και μελλοντικά όλα τα αυτόνομα οχήματα θα είναι και διασυνδεδεμένα οχήματα (Αυτόνομα αυτοκίνητα στην ΕΕ: από επιστημονική φαντασία...σε απτή πραγματικότητα, 2019).

Σήμερα υπάρχουν στην αγορά αρκετά οχήματα με επίπεδο αυτονομίας 2, όπως τα οχήματα της Tesla με το Tesla Autopilot και τα οχήματα της Ford με το Blue Cruise. Η Mercedes είναι η πρώτη εταιρεία με πιστοποίηση αυτοματισμού επιπέδου 3 από την SAE για το Drive Pilot ADAS στην Αμερική. Το Waymo της Google, πρώην Google Self-Driving Car Project, είναι επιπέδου αυτοματισμού 4 και ξεκίνησε ως εγχείρημα το 2016 στις ΗΠΑ και έχει επεκταθεί σε υπηρεσίες ταξί και μεταφοράς με φορτηγά. (https://waymo.com/waymo-via/). Επιπλέον, η Uber έχει επενδύσει σε στόλο αυτόνομων ταξί για τις μεταφορές ανθρώπων και αγαθών στις ΗΠΑ και παλαιότερα στην έρευνα για τα αυτόνομα οχήματα (https://www.uber.com/us/en/autonomous/).

Με Υπουργική απόφαση στην εφημερίδα της Κυβερνήσεως, τον Δεκέμβρη του 2022, επιτρέπεται η κυκλοφορία με επανδρωμένων οχημάτων στην Ελλάδα για ερευνητικούς σκοπούς, εφόσον είναι δυνατός ο απομακρυσμένος έλεγχος τους από χειριστή (Δικαιολογητικά, όροι, προϋποθέσεις και διαδικασία θέσης σε κυκλοφορία επιβατηγού οχήματος χωρίς την παρουσία οδηγού επ’ αυτού, 2022).

Η 1η ελληνική προσπάθεια έγινε το 2013 στα Τρίκαλα με ένα μικρό αυτόνομο λεωφορείο (minibus) με το πρόγραμμα CityMobil2, σταμάτησε όμως το 2015 λόγω ενός ατυχήματος. Τον Μάιο του 2023 ξεκίνησε ξανά η κυκλοφορία δύο minibus στα Τρίκαλα ως υπηρεσία μετακίνησης κατ' εντολή (MaaS).

1.1.2 Πεζοί

Σύμφωνα με έκθεση του Ευρωπαϊκό Συμβούλιο Ασφάλειας Μεταφορών ETSC, το 70% των νεκρών ή σοβαρά τραυματιών από τροχαία συμβάντα στις αστικές οδούς των Ευρωπαϊκών πόλεων είναι πεζοί, ποδηλάτες και μοτοσικλετιστές, κατά την περίοδο 2010-2017. (ΙΟΑΣ, 2019)

Οι πεζοί είναι οι πιο ευάλωτοι χρήστες του δρόμου, συνεπώς είναι απαραίτητη η ανάπτυξη στρατηγικών οδήγησης με επίκεντρο την ασφάλειά τους και η χρήση κατάλληλης τεχνολογίας για την πρόβλεψη της συμπεριφοράς των πεζών από τα αυτόνομα οχήματα.

Επομένως, η εισαγωγή των αυτόνομων οχημάτων μπορεί να συμβάλλει σημαντικά στην ασφαλέστερη αλληλεπίδραση AVs και πεζών και στην βέλτιστη διαχείριση της κυκλοφορίας, με αποτέλεσμα την κατακόρυφη μείωση των ατυχημάτων.

1.2 ΣΚΟΠΟΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Η παρούσα Διπλωματική εργασία έχει ως στόχο την ανάπτυξη ενός μοντέλου αντίστροφης ενισχυτικής μάθησης (inverse reinforcement learning) για τη διερεύνηση των παραγόντων που επηρεάζουν την αλληλεπίδραση των αυτόνομων οχημάτων με τους πιο ευάλωτους χρήστες τους δρόμου. Πιο συγκεκριμένα, εξετάζεται η αλληλεπίδραση ενός αυτόνομου οχήματος με ένα πεζό που βρίσκεται στο πεζοδρόμιο και θέλει να διασχίσει το δρόμο. Ο αλγόριθμος αντίστροφης ενισχυτικής μάθησης που επιλέχθηκε είναι η Μέγιστη Εντροπία (Maximum Entropy IRL) που αναπτύχθηκε από τον Ziebart et al. (2008). Το μοντέλο μαθαίνει την λειτουργία ανταμοιβής έχοντας ως δεδομένα τις καταστάσεις, τις ενέργειες και την βέλτιστη πολιτική που καθορίζεται από τις παρεχόμενες τροχιές ενός ειδικού.

Το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση και την αξιολόγηση του μοντέλου συλλέχθηκε μέσω του πιλοτικού έργου RO2 που πραγματοποιήθηκε στην Καρλσρούη της Γερμανίας, με επικεφαλής το FZI. Τα δεδομένα περιλαμβάνουν διάφορους τύπους δρόμων και κυκλοφορίας, όπως οχήματα, ποδήλατα και πεζούς σε αστικές και προαστιακές περιοχές, χώρους στάθμευσης αυτοκινήτων και αυτοκινητόδρομους. Δύο αυτοκίνητα Επιπέδου 3 με διαφορετικούς τρόπους ελέγχου και στρατηγικές αλληλεπίδρασης χρησιμοποιήθηκαν στις δοκιμές Drive2theFuture για αλληλεπίδραση με τους πεζούς. Εκτός από την πραγματικό κομμάτι δοκιμών, δεδομένα συλλέχθηκαν επίσης μέσω προσομοιώσεων επαυξημένης και εικονικής πραγματικότητας. Το σύνολο των δεδομένων που χρησιμοποιείται για την

εκπαίδευση του μοντέλου Max-Ent IRL ανήκει στην Φάση Ι και προέρχεται από το περιβάλλον προσομοίωσης.

Με την χρήση των παραπάνω δεδομένων επιχειρείται η ανάπτυξη ενός αλγόριθμου αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας που μοντελοποιεί την αλληλεπίδραση μεταξύ ενός διασυνδεδεμένου αυτόνομου οχήματος (CAV) και ενός πεζού.

1.3 ΔΙΑΡΘΡΩΣΗ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Η διπλωματική εργασία αποτελείται από τα ακόλουθα κεφάλαια, εκτός του Κεφαλαίου 1:

Το Κεφάλαιο 2, περιλαμβάνει την βιβλιογραφική ανασκόπηση σχετικά με τις διαθέσιμες μεθοδολογίες και τα υπάρχοντα μοντέλα πρόβλεψης τροχιών και οδήγησης για τα αυτόνομα οχήματα και την αλληλεπίδρασή τους με πεζούς. Παρατίθενται σχετικές έρευνες και επιστημονικά άρθρα για τις υπάρχουσες μεθόδους μοντελοποίησης των αυτόνομων οχημάτων και προσομοίωσης της οδηγικής συμπεριφοράς, καθώς και πρόβλεψης της συμπεριφοράς των πεζών σε αλληλεπίδραση με τα AVs. Γίνεται αναφορά στις διαφορετικές τεχνικές, στα πλεονεκτήματα και στους περιορισμούς του κάθε μοντέλου. Τέλος, επιλέγεται ο καταλληλότερος αλγόριθμος για την ανάπτυξη μίας αποτελεσματικής στρατηγικής οδήγησης για αυτόνομα οχήματα σε αλληλεπίδραση με πεζούς για την παρούσα Διπλωματική εργασία.

Στο Κεφάλαιο 3, περιγράφεται η μεθοδολογική προσέγγιση της Διπλωματικής εργασίας, η οποία περιλαμβάνει την χρήση αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας (Max-Ent IRL) για την μοντελοποίηση της αλληλεπίδρασης των αυτόνομων οχημάτων με τους πεζούς. Παρουσιάζεται η θεωρία για την μηχανική μάθηση, και έπειτα για την ενισχυτική και αντίστροφη ενισχυτική μάθηση, με τα στοιχεία και τις μαθηματικές σχέσεις που τις περιγράφουν. Στη συνέχεια, αναλύεται εκτενέστερα η αντίστροφη ενισχυτική μάθηση μέγιστης εντροπίας, καθώς και οι μέθοδοι χωρίς μοντέλο, οι παράμετροι της συνάρτησης και η τυχαιότητα.

Στο Κεφάλαιο 4, γίνεται η παρουσίαση των προσομοιώσεων και των δεδομένων εισαγωγής που χρησιμοποιήθηκαν στην παρούσα έρευνα. Περιλαμβάνεται η ανάλυση και η επεξεργασία της βάσης δεδομένων στο Excel, καθώς και ο καθορισμός των ομάδων των χαρακτηριστικών, που καθορίζουν τις καταστάσεις και τις ενέργειες που προέκυψαν από τις προσομοιώσεις. Έπειτα, περιγράφεται λεπτομερώς η εφαρμογή της μεθοδολογίας με την χρήση του αλγορίθμου αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας στη python και αναλύονται τα αποτελέσματα του αλγορίθμου. Ο αλγόριθμος Max-Ent IRL χρησιμοποιήθηκε για να συμπεράνει την υποκείμενη συνάρτηση ανταμοιβής στο προσομοιωμένο περιβάλλον από τις διατιθέμενες τροχιές αλληλεπίδρασης οχημάτων και πεζών και για να εξάγει τα βάρη

της συνάρτησης ανταμοιβής για τα χαρακτηριστικά του αλγορίθμου, καθώς και τις ανταμοιβές των καταστάσεων που ορίστηκαν.

Το Κεφάλαιο 5, περιλαμβάνει την σύνοψη των αποτελεσμάτων και τα συμπεράσματα που προέκυψαν από την παρούσα Διπλωματική εργασία. Γίνεται αναφορά στους περιορισμούς που αντιμετωπίστηκαν και πώς επηρέασαν τα αποτελέσματα της έρευνας. Τέλος, γίνονται προτάσεις για περαιτέρω μελλοντική έρευνα πάνω σε πιο περίπλοκες αλληλεπιδράσεις αυτόνομων οχημάτων και πεζών, με περισσότερα δεδομένα και με χρήση βαθιάς ενισχυτικής μάθησης.

ΚΕΦΑΛΑΙΟ 2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

2.1 ΕΙΣΑΓΩΓΗ

Στη βιβλιογραφική ανασκόπηση γίνεται αναφορά σε υπάρχουσες στρατηγικές οδήγησης, στην μοντελοποίηση της συμπεριφοράς των πεζών και στην αλληλεπίδρασή τους με τα αυτόνομα οχήματα με χρήση ενισχυτικής και αντίστροφης ενισχυτικής μάθησης. Παρουσιάζονται τα πλεονεκτήματα κάθε μεθοδολογίας, καθώς και τα συμπεράσματα που προέκυψαν από την ανάλυση της βιβλιογραφίας.

2.2 ΣΥΝΑΦΕΙΣ ΕΡΕΥΝΕΣ & ΜΕΘΟΔΟΛΟΓΙΕΣ

Σύμφωνα με έρευνα που πραγματοποιήθηκε στις ΗΠΑ, σε 723 συγκρούσεις το 99% οφειλόταν σε σφάλμα οδηγικής συμπεριφοράς (Hendricks, 2001). Αυτό θα μπορούσε να αποφευχθεί με την πλήρη αυτοματοποίηση των οχημάτων και την απελευθέρωση του οδηγού από το βάρος της οδήγησης σύμφωνα με τους You et al. (2019) .

Η έρευνα για τα αυτόνομα οχήματα βασίστηκε αρχικά στην ενισχυτική μάθηση, σύμφωνα με την οποία το όχημα μαθαίνει ποιες ενέργειες να επιλέξει με βάση την αλληλεπίδρασή του με το περιβάλλον και τις ανταμοιβές που λαμβάνει έπειτα. Με τον τρόπο αυτό, το όχημα βελτιστοποιεί την συμπεριφορά του σε δυναμικά και αβέβαια περιβάλλοντα, μέσω της διαδικασίας δοκιμής και λάθους.

Στη συνέχεια, η έρευνα στα αυτόνομα οχήματα επεκτάθηκε με την εισαγωγή της αντίστροφης ενισχυτικής μάθησης, η οποία βασίζεται στη μάθηση από επιδείξεις ειδικών ή στην παρατήρηση της συμπεριφοράς τους. Αυτή η τεχνική επιτρέπει στο όχημα να κατανοήσει τις προθέσεις, τις αξίες και τα κίνητρα πίσω από τις ανθρώπινες ενέργειες και να διαμορφώσει ανάλογα την στρατηγική οδήγησης για να αλληλοεπιδράσει με ασφάλεια με τους χρήστες του δρόμου. Αυτό δεν θα ήταν πάντα δυνατό με την ενισχυτική μάθηση και τα διαθέσιμα δεδομένα από το περιβάλλον.

Οι Sharifzadeh et al. (2016) προτείνουν την δημιουργία ενός μοντέλου αυτόνομης οδήγησης με τη χρήση αντίστροφης ενισχυτικής μάθησης, το οποίο θα συμπεριφέρεται σαν άνθρωπος και θα παρέχει άνεση και ασφάλεια στο χρήστη. Για να καταστεί δυνατό να εφαρμοστεί σε περιβάλλον πόλεων με περισσότερα χαρακτηριστικά, όπως διασταυρώσεις και πεζούς, προτείνεται η χρήση βαθιού (νευρωνικού) Q-δικτύου στο βήμα της ενισχυτικής μάθησης.

Το πρόβλημα προσδιορισμού των κινηματικών χαρακτηριστικών του οχήματος σχεδιάζεται ως διαδικασία λήψεως αποφάσεων Markov (MDP) και η εξαγωγή της συνάρτησης ανταμοιβής γίνεται με χρήση DQN λόγω της ύπαρξης πολλών, 2^{52}, καταστάσεων. Το όχημα έχει 13 αισθητήρες και 3 βαθμούς ελευθερίας. Αποτελείται από το στρώμα εισόδου χαρακτηριστικών, 2 ενδιάμεσα πλήρως συνδεδεμένα κρυφά στρώματα με 160 μέρη το καθένα , και ένα πλήρως συνδεδεμένο κρυφό στρώμα με τις ενέργειες που είναι 3 (δεξιά, αριστερά και διατήρηση πορείας). Η εκπαίδευση γίνεται από 90 επιδείξεις ειδικών από το περιβάλλον προσομοίωσης στην Python με χρήση του projection-based IRL και παρουσιάζει ικανοποιητικά αποτελέσματα μετά από 6 επαναλήψεις.

Η αποδοχή των αυτόνομων οχημάτων εξαρτάται από το επίπεδο άνεσης και αξιοπιστίας που παρέχουν μέσω της οδήγησή τους στον ίδιο το χρήστη, αλλά και στα περιβάλλοντα οχήματα και στους περαστικούς. Οι Kuderer et al. (2015) μελετούν την δυνατότητα εκμάθησης διαφορετικών στυλ οδήγησης από επιδείξεις οδηγών. Η εκμάθηση γίνεται με τη χρήση αντίστροφης ενισχυτικής μάθησης που βασίζεται στα χαρακτηριστικά της οδήγησης, τα οποία σχηματίζουν μια συνάρτηση κόστους. Η μέθοδος αντιστοιχίζει τα στοιχεία οδήγησης όπως ταχύτητα, επιτάχυνση, αλλαγή λωρίδας, επιθυμητή λωρίδα και την δεύτερη παράγωγο της ταχύτητας (jerk) σε κάθε στυλ οδήγησης και σκοπός είναι η διαμόρφωση των παραμέτρων του προβλήματος.

Τα δεδομένα συλλέχθηκαν με τη χρήση ενός οχήματος με αισθητήρες σε αυτοκινητόδρομο με κανονική κυκλοφορία στις ΗΠΑ. Ο αλγόριθμος μαθαίνει επιτυχώς τα διαφορετικά στυλ οδήγησης. Η επιλογή της τροχιάς με το χαμηλότερο κόστος γίνεται με τη χρήση του αλγορίθμου βελτιστοποίησης RPROP. Η πολιτική που μαθαίνει, παράγει τροχιές με παρόμοια χαρακτηριστικά με αυτά των επιδείξεων και μετά από 30 επαναλήψεις ο αλγόριθμος συγκλίνει. Περαιτέρω έρευνα μπορεί να γίνει σχετικά με τη πρόβλεψη συμπεριφοράς σε πιο περίπλοκα σενάρια οδήγησης.

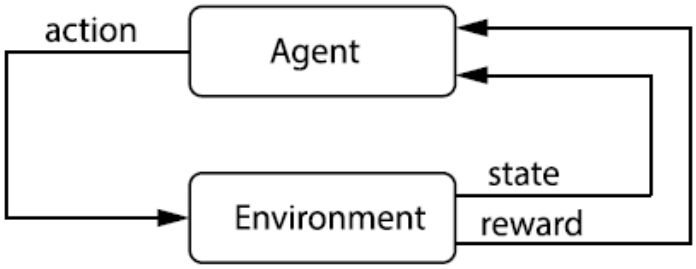
Οι Mantouka et al. (2019) διερευνούν την εφαρμογή τεχνικών μηχανικής εκμάθησης για την αξιολόγηση των προφίλ ασφάλειας οδήγησης με βάση τα δεδομένα από έξυπνα κινητά τηλέφωνα (smartphone). Οι συγγραφείς αναγνωρίζουν τις δυνατότητες των smartphone ως πηγών δεδομένων για τη μελέτη της συμπεριφοράς και της ασφάλειας των οδηγών. Στην έρευνα χρησιμοποιήθηκαν δεδομένα αισθητήρων smartphone από περισσότερα από 10.000 ταξίδια 129 οδηγών σε αστικές και αγροτικές περιοχές για να εντοπιστούν μη ασφαλείς συμπεριφορές οδήγησης. Συνέλεξαν δεδομένα σε πραγματικό χρόνο από smartphone, συμπεριλαμβανομένων δεδομένων GPS, επιταχυνσιόμετρου και γυροσκοπίου, για να δημιουργήσουν ένα ολοκληρωμένο σύνολο δεδομένων. Έγινε ομαδοποίηση δύο επιπέδων k-means, εντοπίζοντας πρώτα επιθετικά χαρακτηριστικά, όπως η απότομη επιτάχυνση και το απότομο φρενάρισμα, και στη συνέχεια εμβαθύνοντας σε ακατάλληλη ταχύτητα και περισπασμούς. Παραδόξως, μόνο το 8% των

ταξιδιών περιλάμβανε οδήγηση με απόσπαση της προσοχής, αλλά η πλειονότητα εμφάνισε επικίνδυνες συμπεριφορές με ασυνέπεια σε διάφορα ταξίδια.

Αυτή η προσέγγιση καθιστά δυνατή την παρακολούθηση και βελτίωση της οδικής ασφάλειας, χωρίς την ανάγκη δαπανηρού αποκλειστικού υλικού ή υποδομής. Αξιοποιώντας τις δυνατότητες των smartphone, αυτή η έρευνα προσφέρει μια πολλά υποσχόμενη οδό για την ανάπτυξη εξατομικευμένων παρεμβάσεων ασφάλειας και στοχευμένων στρατηγικών οδικής ασφάλειας, που μπορούν να συμβάλουν στη μείωση των ατυχημάτων και στη βελτίωση της συνολικής οδικής ασφάλειας. Οι μελλοντικές μελέτες στοχεύουν στη διερεύνηση πρόσθετων επικίνδυνων συμπεριφορών και στην κατανόηση των παραγόντων που επηρεάζουν τα διαφορετικά στυλ οδήγησης.

Οι You et al. (2019) συνδυάζουν την χρήση ενισχυτικής και αντίστροφης ενισχυτικής μάθησης για τον καλύτερο σχεδιασμό της διαδρομής ενός αυτόνομου οχήματος. Η αλληλεπίδραση του οχήματος με το περιβάλλον αναπαρίσταται με τη χρήση ενός στοχαστικού MDP και λαμβάνεται υπόψη η γεωμετρία του δρόμου. Στόχος είναι η αναπαράσταση της συμπεριφοράς ενός έμπειρου οδηγού, λαμβάνοντας υπόψη τις συμπεριφορές των γειτονικών οχημάτων στο δρόμο. Το μοντέλο μπορεί να προσαρμοστεί για διαφορετικό αριθμό λωρίδων και οχημάτων.

Στην έρευνα αναπαρίσταται η κίνηση σε αυτοκινητόδρομο και αναπτύσσονται τεχνικές οδήγησης για προσπέραση (overtaking) και ακολούθηση σε ουρά (tailgating) με τη χρήση ενισχυτικής μάθησης. (Εικ.1) Ο σχεδιασμός της ανταμοιβής γίνεται με τη χρήση βαθιού νευρωνικού δικτύου (DNN). Η βέλτιστη πολιτική επιτυγχάνεται με τη χρήση του αλγορίθμου Q-learning.

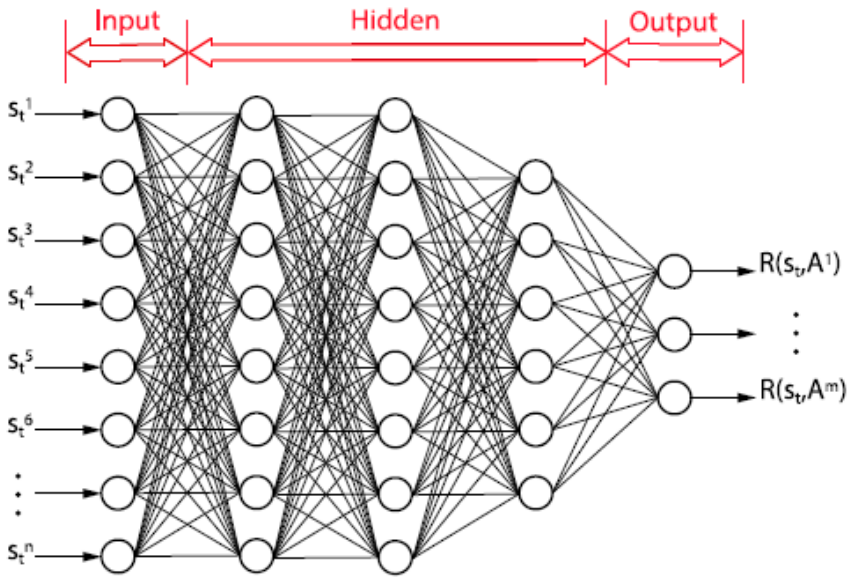


Εικόνα 1: Αλληλεπίδραση πράκτορα με το περιβάλλον (Πηγή: You et al., 2019)

Στη συνέχεια, γίνεται χρήση αντίστροφης ενισχυτικής μάθησης για την εξαγωγή της συνάρτησης ανταμοιβής από επιδείξεις οδήγησης από ειδικό. Χρησιμοποιείται η αρχή της μέγιστης εντροπίας και παρουσιάζονται 3 νέοι αλγόριθμοι μέγιστης εντροπίας βαθιάς ενισχυτικής μάθησης για την αναπαράσταση ενός model-free σεναρίου. Για πρώτη φορά, η συνάρτηση ανταμοιβής παίρνει μη γραμμική μορφή και

γίνεται χρήση DNN για την αναπαράσταση του μητρώου κατάσταση-ενέργεια-ανταμοιβή αντί για κατάσταση-ανταμοιβή για να προβλέπει περισσότερες οδηγικές συμπεριφορές (Εικ.2). Η επιλογή από τους 3 αλγόριθμους γίνεται με βάση τα διαθέσιμα δεδομένα και το μοντέλο MDP του προβλήματος (στοχαστικό ή ντετερμινιστικό).

Ο αλγόριθμος της ενισχυτικής μάθησης φτάνει σε σύγκλιση μετά από 5000 με 6000 επεισόδια και σε λιγότερο από 5 λεπτά και για τις δύο οδηγικές συμπεριφορές- προσπέραση και ακολούθηση σε ουρά. Για τον αλγόριθμο αντίστροφης ενισχυτικής μάθησης επιλέγεται ένα νευρωνικό δίκτυο με διαστάσεις [10,20,20,20,5], δηλαδή 10 κανάλια εισόδου για τις καταστάσεις s_t, 5 κανάλια εξόδου για τις ενέργειες A και 3 κρυμμένα στρώματα με 20 νευρώνες το καθένα. Ο 1ος αλγόριθμος, που προτείνεται, δεν συγκλίνει λόγω πολλών δεδομένων και στοχαστικού MDP, ο 2ος και ο 3ος συγκλίνουν, όμως ο 2ος απαιτεί λιγότερη ώρα, εφόσον δεν μαθαίνει το μοντέλο και δεν υπολογίζει τις αναμενόμενες επισκέψεις κατάστασης-ενέργειας. Η αξιολόγηση των αλγορίθμων γίνεται σε προσομοίωση.



Εικόνα 2: Δομή Βαθιού Νευρωνικού Δικτύου για Συνάρτηση Ανταμοιβής (Πηγή: You et al., 2019)

Η έρευνα συμπεραίνει ότι οι επιθυμητές οδηγικές συμπεριφορές επιτεύχθηκαν με τη χρήση και των δύο μεθόδων. Το μοντέλο είναι εύκολα κλιμακούμενο, όμως τα οχήματα αντιμετωπίζονται ως σημεία μάζας, χωρίς να λαμβάνεται υπόψη η ταχύτητα, κάτι που θα μπορούσε να γίνει σε μελλοντική έρευνα για πιο ρεαλιστικά σενάρια. Στο πείραμα χρησιμοποιήθηκε δρόμος με 5 λωρίδες και κάθε περιβαλλοντικό όχημα είχε μία τυχαία πολιτική. Για την αντίστροφη ενισχυτική μάθηση δεν είναι προτεινόμενες οι μακροχρόνιες επιδείξεις, διότι μπορεί να μην είναι επαρκείς για την αναπαράσταση της στοχαστικής συμπεριφοράς του

συστήματος και είναι πιο πιθανό το σφάλμα πρόβλεψης να γίνει μεγαλύτερο για προβλήματα χωρίς μοντέλο.

Αυτό λύνεται με την μεγιστοποίηση της εντροπίας σε μικρότερα κομμάτια δεδομένων. Μελλοντική έρευνα θα μπορούσε να γίνει σε πραγματικά σενάρια οδήγησης, με χρήση μερικώς παρατηρήσιμου MDP για την αβεβαιότητα του περιβάλλοντος, καθώς και με πολλαπλούς πράκτορες για καλύτερο έλεγχο της κυκλοφορίας.

Οι Ziebart et al. (2008) διερευνούν την χρήση της αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας (Max-Ent IRL) για την ανάπτυξη μια στρατηγικής οδήγησης για αυτόνομα οχήματα για την πλοήγηση σε πραγματικό χρόνο και την λήψη αποφάσεων που βασίζονται στις ανθρώπινες προτιμήσεις.

Οι πολιτικές που μαθαίνει ο αλγόριθμος αναπαρίστανται ως διαδικασία λήψεως αποφάσεων Markov (MDP) και στόχος είναι η έρευση των βαρών της υποκείμενης συνάρτησης ανταμοιβής (reward weights), που κάνουν την συμπεριφορά επίδειξης βέλτιστη. Η μέγιστη εντροπία καλείται να λύσει τα προβλήματα αβεβαιότητας, λόγω θορύβου στις παρατηρούμενες επιδείξεις και ατελών συμπεριφορών οδήγησης, καθώς και της ασάφειας στην επιλογή κατανομής αποφάσεων.

Με την χρήση της μέγιστης εντροπίας στην αντίστροφη ενισχυτική μάθηση το μοντέλο αναπτύσσει μία ενιαία στοχαστική πολιτική και επιλέγει την κατανομή απόφασης, που αντιστοιχεί στα χαρακτηριστικά των συμπεριφορών, χωρίς να επιδεικνύει επιπλέον προτίμηση σε κάποια διαδρομή, πέραν από ότι επιβάλλουν οι περιορισμοί του προβλήματος. Η κατανομή παραμετροποιείται με τα βάρη επιβράβευσης. Στο πείραμα χρησιμοποιείται ο ανανεούμενος σε πραγματικό χρόνο διαδικτυακός αλγόριθμος καθόδου με εκθετική κλίση (online exponentiated gradient descent algorithm).

Σκοπός της έρευνας είναι η ανάκτηση μίας συνάρτησης χρησιμότητας για την πρόβλεψη της οδηγικής συμπεριφοράς και την πρόταση διαδρομών με την χρήση της μάθησης μίμησης (imitation learning) από τις επιλογές διαδρομής των οδηγών. Η συγκεκριμένη έρευνα αποτελεί το μεγαλύτερης κλίμακας πρόβλημα IRL από άποψη μεγέθους των δεδομένων επίδειξης. Γίνεται μοντελοποίηση του οδικού δικτύου στην περιοχή του Pittsburgh της Pennsylvania με ένα ντετερμινιστικό MDP που αποτελείται από 300.000 καταστάσεις (οδικά τμήματα κ.α.) και 900.000 ενέργειες (επιλογές σε διασταυρώσεις κ.α.). Θεωρείται ότι ο οδηγός θέλει να επιτύχει ένα στόχο, με βελτιστοποίηση του κόστους της διαδρομής, που περιλαμβάνει τον χρόνο, την ασφάλεια, το κόστος καυσίμου και άλλα. Το βάρος ανταμοιβής είναι ανεξάρτητο από τον προορισμό και επομένως πολλά MDPs μπορούν να μάθουν το ίδιο βάρος για διαφορετικούς προορισμούς (goal state).

Η συλλογή των δεδομένων έγινε από 25 οδηγούς ταξί με χρήση GPS και συλλέχθηκαν 100.000 μίλια ταξιδιού. Στη συνέχεια χωρίστηκαν σε διαδρομές και αφαιρέθηκαν οι κυκλικές και κάποιες μικρής διάρκειας με πολύ θόρυβο. Το 20% χρησιμοποιήθηκε για εκπαίδευση (training set) του αλγόριθμου και το 80% για δοκιμές (testing set) από τα 7403 παραδείγματα. Οι διαδρομές αποτελούνταν από 4 χαρακτηριστικά: είδος δρόμου (διαπολιτειακός -τοπικός), ταχύτητα (υψηλή -χαμηλή), λωρίδες και μεταβάσεις (ευθεία, αριστερά, δεξιά, hard left, hard right) συμβάλλοντας σε 22 διαφορετικά χαρακτηριστικά.

Η χρήση του μοντέλου Max-Ent IRL για την μάθηση της συλλογικής συνάρτησης χρησιμότητας των οδηγών ταξί γίνεται πιο αποτελεσματική και γρήγορη με την αύξηση της πιθανότητας των διαδρομών επίδειξης σε μια μικρότερη τάξη σχετικά καλών μονοπατιών. Το μοντέλο συγκρίνεται με το Maximum Margin Planning (Ratliff et al., 2006), το οποίο μπορεί να προβλέψει νέες διαδρομές, όχι όμως να υπολογίσει την πιθανότητα των δοσμένων διαδρομών. Έπειτα, γίνεται σύγκριση με το action-based distribution model, που χρησιμοποιήθηκε για Bayesian IRL και hybrid IRL (Ramachandran et al., 2017 και Neu et al., 2012), όμως είναι μεροληπτικό απέναντι σε ορισμένες διαδρομές (label bias). Το Max-Ent IRL μοντέλο επιλέγει το μονοπάτι με το βέλτιστο κόστος και ορίζει ίσες πιθανότητες σε συμπεριφορές με ίση αναμενόμενη επιβράβευση. Το μοντέλο μέγιστης εντροπίας IRL παρουσιάζει τα καλύτερα αποτελέσματα στην πιο πιθανή εκτίμηση διαδρομής και εκτίμηση πυκνότητας διαδρομής σε σχέση με τα υπόλοιπα μοντέλα.

Το μοντέλο Max-Ent IRL προσφέρει χαμηλότερα κόστη στις επιθυμητές οδούς και υψηλότερα στις μη επιθυμητές οδούς. Επιπρόσθετα, η παραμετρική προσέγγιση επιτρέπει την εξαγωγή, από τις προτιμήσεις του οδηγού, προσθέτων πληροφοριών και, στη συνέχεια, την γενίκευση σε καινούρια οδικά τμήματα. Εκτός από την εφαρμογή του μοντέλου για την πρόταση διαδρομών, ανοίγονται επιπλέον δυνατότητες για εξατομίκευση των διαδρομών, μέσω της παρατήρησης ενός χρήστη, όπως προτάσεις για αποφυγή της κίνησης- κάτι, που δεν έχει διερευνηθεί εδώ.

Η εφαρμογή του Max-Ent IRL μοντέλου έγινε σε περιορισμένο σύνολο χαρακτηριστικών και αποτέλεσε μία νέα πρόταση για την αντιμετώπιση της ασάφειας προηγούμενων μεθόδων, προσφέροντας σημαντική εξασφάλιση απόδοσης με χρήση μιμητικής μάθησης. Σε μελλοντική έρευνα, θα μπορούσε να γίνει προσθήκη επιπλέον χαρακτηριστικών (π.χ. ώρα της ημέρας, καιρός) για την βελτιστοποίηση του μοντέλου.

Σε επόμενο άρθρο τους, οι Ziebart et al. (2009), διερευνούν την εξέλιξη του μοντέλου αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας και τα αποτελέσματα της εφαρμογής του στην μοντελοποίηση οδηγικών προτιμήσεων. Επιπλέον, η έρευνα επεκτείνεται στην εφαρμογή του αλγορίθμου σε δυναμικά

περιβάλλοντα. Παρουσιάζονται τα αποτελέσματα της μοντελοποίησης της βάσης δεδομένων χρονικής χρήσης και οι προκλήσεις για εφαρμογή σε περαιτέρω τομείς της ανθρώπινης συμπεριφοράς.

Η ανθρώπινη συμπεριφορά αντιμετωπίζεται ως μια δομημένη σειρά αποφάσεων, που επηρεάζονται από τα δεδομένα του περιβάλλοντος. Για περίπλοκες συμπεριφορές, η μοντελοποίηση γίνεται από παρατηρούμενες συμπεριφορές με χρήση του αλγόριθμου Max-Ent IRL. Σκοπός της έρευνας είναι η αναπαράσταση της διαδικασίας λήψης ανθρώπινων αποφάσεων.

Η ανθρώπινη συμπεριφορά είναι σκόπιμη και οι άνθρωποι ενεργούν για να πραγματοποιήσουν αποτελεσματικά ένα στόχο. Η παρατήρηση της ομοιότητας των στόχων σε διαφορετικούς τομείς επιτρέπει την μοντελοποίηση με κοινές έννοιες χρησιμότητας και αποτελεσματικότητας. Η διαδικασία λήψης αποφάσεων Markov (MDP) επιτρέπει την αναπαράσταση αυτών των στόχων και της αντίστοιχης έννοιας αποτελεσματικότητας.

Τα πιθανολογικά γραφικά μοντέλα χρησιμοποιούνται για την αντιμετώπιση της αβεβαιότητας των δεδομένων στην μηχανική μάθηση (Bayesian networks, Markov random fields, conditional random fields/ CRF). Το μοντέλο αντίστροφου βέλτιστου ελέγχου μέγιστης εντροπίας (Maximum Entropy Inverse Optimal Control) είναι συνώνυμο της αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας (Max-Ent IRL). Το Max-Ent IOC μοντέλο είναι MDP με ντετερμινιστικά αποτελέσματα ενεργειών και μοιάζει με chain CRF, που εστιάζει σε δεδομένα δομημένα σε αλυσίδα, όπως ακολουθίες ή χρονοσειρές, με τη διαφορά ότι η σειρά αποφάσεων εξαρτάται από τα συνολικά χαρακτηριστικά των καταστάσεων και των ενεργειών, αντίθετα από τοπικές παρατηρήσεις μέρους της σειράς αποφάσεων.

Ο αλγόριθμος μέγιστης εντροπίας χρησιμοποιεί δυναμικό προγραμματισμό (forward-backward για CRFs ή value iteration για MDPs) για να υπολογίσει αποτελεσματικά τις αναμενόμενες συχνότητες κατάληψης κατάστασης, αξιοποιώντας αναδρομικά την χρήση της συνάρτησης διχοτόμησης (partition function).

Η ανθρώπινη συμπεριφορά χαρακτηρίζεται από τυχαιότητα. Στην προηγούμενη έρευνα (Ziebart et al., 2008) δόθηκε μια κατά προσέγγιση μέθοδος για την μοντελοποίηση της συμπεριφοράς. Στη συνέχεια, στην παρούσα έρευνα (Ziebart et al., 2009) παρουσιάζεται η ακριβής προσέγγιση με χρήση της αρχής της μέγιστης εντροπίας. Ορίζεται μια μη ελεγχόμενη κατανομή- τυχαία διαδικασία, για τις τροχιές και μεγιστοποιείται η εντροπία της κατανομής των τροχιών. Από την επίλυση των παραπάνω, προκύπτει ένας αναδρομικός τύπος για τον υπολογισμό της συνάρτησης διάτμησης (partition function), επιτρέποντας την εξαγωγή των πιθανοτήτων δράσης και έπειτα τον υπολογισμό των συχνοτήτων κατάστασης.

Στο παρόν άρθρο, προστέθηκαν συναφείς πληροφορίες στο μοντέλο σχετικά με το περιβάλλον, όπως η ώρα της ημέρας, η κίνηση κ.α., και έγινε σύγκριση με κατευθυνόμενα γραφικά μοντέλα που μοντελοποιούν την συνήθη μετακίνηση (Liao et al., 2007) και πιο συγκεκριμένα με άλλα μοντέλα απόφασης Markov σε επόμενη διασταύρωση, εξαρτώμενα από την τοποθεσία-στόχο και από τα προηγούμενα οδικά τμήματα (Simmons et al.,2006, Krumm et al., 2008). Φαίνεται ότι το μοντέλο Max-Ent IRL έχει σημαντικά καλύτερη απόδοση σε σχέση με τα υπόλοιπα μοντέλα.

Στη συνέχεια, γίνεται αξιολόγηση του μοντέλου στην πρόβλεψη προορισμού από μερική διαδρομή, με τη χρήση του κανόνα Bayes και ενσωματώνοντας μία προηγούμενη κατανομή στους προορισμούς. Συγκρίνονται τα μοντέλα Predestination (Krumm & Horvitz 2006) και ένα destination-based μοντέλο Markov (Simmons et al. 2006). Το Predestination υποθέτει μια σταθερή μέτρηση (χρόνος ταξιδιού) και μοντελοποιεί την αποτελεσματικότητα (δηλαδή την προτίμηση) δεδομένης αυτής της μέτρησης, ενώ το μοντέλο Max-Ent IRL υποθέτει ένα μοντέλο σταθερής προτίμησης και μαθαίνει την μέτρηση με βάση τον οδηγό. Το Max-Ent IRL και το Predestination είναι εμφανώς καλύτερα από το μοντέλο Markov. Επιπλέον, το μοντέλο μέγιστης εντροπίας έχει λιγότερα σφάλματα για μικρότερο ποσοστό ολοκληρωμένης διαδρομής (0-40%) και για μεγαλύτερο ποσοστό (80-100%) σε σχέση με το Predestination.

Έπειτα, παρουσιάζονται τα αποτελέσματα από τα προκαταρκτικά πειράματα μοντελοποίησης δεδομένων χρονικής χρήσης με Max-Ent IRL που συλλέχθηκαν από 12248 άτομα στην αμερικανική time-use έρευνα (ATUS), τα οποία είναι πολύτιμα για την μοντελοποίηση της ανθρώπινης συμπεριφοράς, λόγω της ποικιλομορφίας και της έκτασής τους. Γίνεται μοντελοποίηση μίας ημέρας χρονικής χρήσης ενός ατόμου βάση των δημογραφικών του χαρακτηριστικών (φύλο, ηλικία) με 3 διαφορετικά μοντέλα (Naive, Stationary Markov και Max-Ent IRL). Το 80% των δεδομένων χρησιμοποιείται για training και το 20% για testing. Η αξιολόγηση γίνεται με εμπειρική μέση πιθανότητα καταγραφής (empirical average log probability) με βάση το 2 και υπολογίζεται πόσα bits απαιτούνται ανά μοντέλο για την αναπαράσταση μίας ημέρας δραστηριοτήτων. Το naive μοντέλο θεωρείται ως baseline. Το Max-Ent IRL μοντέλο έχει την καλύτερη απόδοση και βασίζεται στις δημογραφικές πληροφορίες, ενσωματώνοντας την ώρα της ημέρας.

Στον τομέα της οδήγησης, η διάρκεια που δαπανάται σε τμήματα του δρόμου επηρεάζεται από τυχαίους εξωτερικούς παράγοντες. Στο μοντέλο μέγιστης εντροπίας IRL, η διάρκεια αφαιρέθηκε ως μέρος της συνάρτησης κόστους και προστέθηκε η σειρά αποφάσεων στο χώρο καταστάσεων. Τα Μαρκοβιανά μοντέλα τείνουν να υποθέτουν εκθετικά φθίνουσες διάρκειες, ενώ οι ψευδο-Γκαουσιανές κατανομές είναι πιο κατάλληλες για τη μοντελοποίηση της διάρκειας των ανθρώπινων συμπεριφορών. Η επέκταση της λογικής των υφιστάμενων προσεγγίσεων, όπως οι διαδικασίες απόφασης ημι-Markov και τα τυχαία πεδία ημι-Markov- για ημι-παρατηρούμενα περιβάλλοντα, για την feature-based συνάρτηση χρησιμότητας του

μοντέλου Max-Ent IRL με ικανή απόδοση, αποτελεί σημαντική πρόκληση στη μοντελοποίηση της ανθρώπινης συμπεριφοράς.

Σε ορισμένους τομείς, η απόκτηση ακριβών προβλέψεων απαιτεί εξατομικευμένα μοντέλα. Η κατασκευή μοντέλου για ένα μεμονωμένο άτομο γίνεται με χρήση των δεδομένων εκπαίδευσης του ατόμου, αντί της ομάδας, στο μοντέλο μέγιστης εντροπίας IRL. Ωστόσο, σε σενάρια όπου τα διαθέσιμα δεδομένα δεν είναι επαρκή για το άτομο, η αξιοποίηση δεδομένων από άτομα με παρόμοια δημογραφικά χαρακτηριστικά για τη δημιουργία ενός πιο ακριβούς μοντέλου συμπεριφοράς, φαίνεται ελκυστική. Αυτή η προσέγγιση επιτρέπει την ανάπτυξη αξιόπιστων μοντέλων συμπεριφοράς για νέα άτομα, χωρίς την προηγούμενη παρατήρηση της συμπεριφορά τους. Η διερεύνηση εναλλακτικών μεθόδων για την αξιοποίηση της ομοιότητας των ατόμων και η αξιολόγηση τους παραμένουν σημαντικές μελλοντικές κατευθύνσεις έρευνας.

Συμπερασματικά, η νέα προσέγγιση μέγιστης εντροπίας συνδυάζει αποτελεσματικά τη βέλτιστη μοντελοποίηση αποφάσεων και τα πιθανοτικά γραφικά μοντέλα για να δημιουργήσει ένα συμπαγές και αποτελεσματικό πιθανολογικό μοντέλο ανθρώπινης συμπεριφοράς. Αυτό το μοντέλο μπορεί να ενσωματωθεί εύκολα σε ένα Bayesian πλαίσιο και προσφέρει ισχυρές εγγυήσεις απόδοσης. Εφαρμόζοντας τη μέθοδο μέγιστης εντροπίας στο πρόβλημα της μοντελοποίησης των προτιμήσεων διαδρομής και της χρήσης χρόνου, με βάση τα δημογραφικά δεδομένα, παρατηρήθηκαν αποτελέσματα με υψηλή ακρίβεια πρόβλεψης. Ωστόσο, οι προκλήσεις της μοντελοποίησης της διάρκειας συμπεριφοράς και της αξιοποίησης ομάδων με παρόμοια χαρακτηριστικά για την επέκταση εφαρμογής της μεθόδου σε άλλους τομείς μοντελοποίησης συμπεριφορών συνεχίζουν να υπάρχουν.

Οι Martinez-Gil et al. (2020) παρουσιάζουν μια νέα προσέγγιση για τη βελτίωση της ακρίβειας και της αξιοπιστίας των προσομοιώσεων πεζών με τη χρήση αντίστροφης ενισχυτικής μάθησης (IRL) και με δεδομένα από πραγματικές τροχιές πεζών. Η μέθοδος της αντίστροφης ενισχυτικής μάθησης προτείνεται για την αναπαράσταση πιο περίπλοκων και ποικιλόμορφων συμπεριφορών των πεζών, με την μάθηση της επιθυμητής συμπεριφοράς από επιδείξεις που συλλέχθηκαν στο εργαστήριο.

Βασικός στόχος της έρευνας είναι η προσαρμογή της μεθόδου IRL με χρήση μέγιστης περιστασιακής εντροπίας (maximum casual entropy) για την αναπαράσταση της ανθρώπινης πλοήγησης. Στα θετικά της αντίστροφης ενισχυτικής μάθησης ανήκουν ο αυτόματος σχεδιασμός της συνάρτησης ανταμοιβής και η ρεαλιστή απεικόνιση των παραγόμενων συμπεριφορών, καθώς βασίζονται σε πραγματικά παραδείγματα. Οι προκλήσεις που παρουσιάζονται είναι η μη γραμμικότητα της συνάρτησης ανταμοιβής και ο υψηλός

φόρτος υπολογισμού για εκδοχές χωρίς μοντέλο (model-free), όπου χρησιμοποιείται RL στην διαδικασία. Επιπλέον, η ύπαρξη συνεχόμενου χώρου καταστάσεων χρήζει έρευνας.

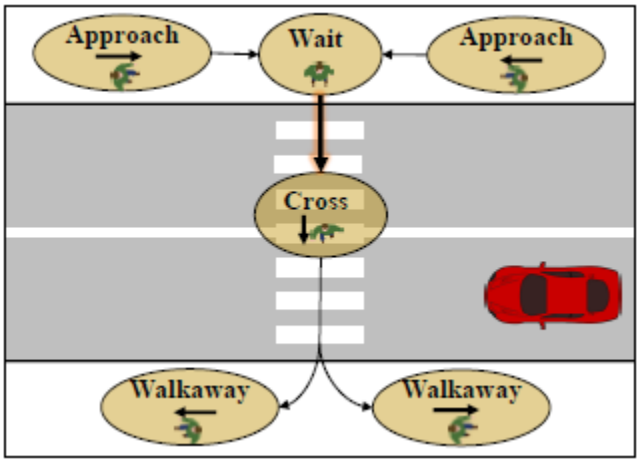
Πιο συγκεκριμένα, η έρευνα απεικονίζει την πορεία ενός πεζού σε ένα τρισδιάστατο περιβάλλον, το οποίο είναι ένας λαβύρινθος. Στην περιγραφόμενη ρύθμιση (setup) ένας ενσωματωμένος 3D πράκτορας πλοηγείται σε ένα εικονικό περιβάλλον με στόχο να φτάσει σε ένα σημείο-στόχο μέσα σε περιορισμένο αριθμό βημάτων. Η κατάσταση του πράκτορα αντιπροσωπεύεται από ένα διάνυσμα χαρακτηριστικών με πραγματική αξία και επιλέγει ενέργειες για να ελέγξει την ταχύτητα και την κατεύθυνσή του. Ο χώρος κατάστασης διακριτοποιείται χρησιμοποιώντας μια μέθοδο γενίκευσης, που ονομάζεται κωδικοποίηση πλακιδίων (tile coding), η οποία επιτρέπει την αντιστοίχιση συνεχών χαρακτηριστικών σε μια διακριτή συνάρτηση αξίας (value function).

Πραγματικές τροχιές, που έχουν συλλεχθεί από πεζούς στο εργαστήριο, χρησιμοποιούνται ως παραδείγματα για την διαδικασία αντίστροφης ενισχυτικής μάθησης (IRL) με σκοπό την εκμάθηση παρόμοιων μονοπατιών. Χρησιμοποιείται επίσης μια βασική προσέγγιση ενισχυτικής μάθησης (RL) με τον αλγόριθμο Sarsa(λ) και την κωδικοποίηση πλακιδίων, όμως παράγει ανεπιθύμητες τροχιές που αποφεύγουν τους διαδρόμους του λαβύρινθου. Η διαδικασία IRL, αντιθέτως, μαθαίνει αποτελεσματικά συμπεριφορές που ευθυγραμμίζονται με τις πραγματικές τροχιές των πεζών.

Η ανάλυση Procrustes, που είναι στατιστική τεχνική για τη σύγκριση και την ευθυγράμμιση δύο συνόλων πολυμεταβλητών δεδομένων, συχνά στο πλαίσιο της ανάλυσης σχήματος, επιβεβαιώνει την ομοιότητα μεταξύ των μαθησιακών διανυσμάτων χαρακτηριστικών και των διανυσμάτων πραγματικών χαρακτηριστικών, επικυρώνοντας την αποτελεσματικότητα του πλαισίου IRL. Ο υπολογιστικός χρόνος για την προσαρμογή της επιβράβευσης του IRL είναι 100 επαναλήψεις, ενώ θα μπορούσε να χρησιμοποιηθεί μια πιο στοχευμένη επιβράβευση με την προσέγγιση RL. Το υπολογιστικό κόστος της μεθόδου IRL είναι υψηλό και η χρήση παραδειγμάτων ειδικών επισημαίνεται ως εναλλακτική σε περιπτώσεις, όπου ο σχεδιασμός της συνάρτησης ανταμοιβής δεν είναι απλός για το RL.

Η έρευνα των Jayaraman et al. (2020) προτείνει ένα μοντέλο υβριδικών συστημάτων για την πρόβλεψη των μακροπρόθεσμων τροχιών πεζών (5-10 sec), όταν αλληλοεπιδρούν με αυτοματοποιημένα οχήματα (AVs) στις διαβάσεις πεζών. Οι τρέχουσες προσεγγίσεις στο σχεδιασμό κίνησης AV προβλέπουν μόνο βραχείς χρονικούς ορίζοντες (1-2 sec) με βάση δεδομένα από αλληλεπιδράσεις πεζών με οχήματα, που οδηγούνται από ανθρώπους (HDV) για την αποφυγή συγκρούσεων. Το υβριδικό μοντέλο συνδυάζει τη συμπεριφορά αποδοχής του χάσματος των πεζών (gap acceptance) και τη δυναμική σταθερής ταχύτητας για να προβλέψει με ακρίβεια τις τροχιές των πεζών για μεγάλες διάρκειες (> 5 secs) στις διαβάσεις πεζών.

Ορίζονται τέσσερις διακριτές καταστάσεις: προσέγγιση της διάβασης πεζών, αναμονή, διάσχιση και απομάκρυνση από τη διάβαση. Η συμπεριφορά των πεζών στη διάβαση ορίζεται ως η ακολουθία ενεργειών και μεταβάσεων μεταξύ των καταστάσεων, μαζί με την προκύπτουσα τροχιά θέσης. Το μοντέλο συνδυάζει μεταβάσεις ενεργειών υψηλού επιπέδου και εξέλιξη συνεχούς κίνησης χαμηλού επιπέδου, επιτρέποντας τη μακροπρόθεσμη πρόβλεψη της συμπεριφοράς διάβασης των πεζών. (Εικόνα 3)



Εικόνα 3: Ενέργειες των πεζών όταν διασχίζουν τον δρόμο. (Πηγή: Jayaraman et al., 2020)

Το υβριδικό μοντέλο χρησιμοποιεί τυπικά μοντέλα, όπως το Support Vector Machine (SVM) για την πρόβλεψη αποδοχής χάσματος και τα μοντέλα κίνησης σταθερής ταχύτητας, για να καθορίσει την εξέλιξη της θέσης και της ταχύτητας των πεζών. Αξιοποιώντας αυτά τα μοντέλα, το υβριδικό σύστημα προβλέπει με ακρίβεια τις τροχιές διέλευσης των πεζών. Η έρευνα συγκρίνει επίσης τα μέτρα των συμπεριφορών διέλευσης πεζών σε ένα περιβάλλον εικονικής πραγματικότητας (IVE) σε αλληλεπίδραση με AV, με εκείνα του πραγματικού κόσμου (από δημοσιευμένες μελέτες πεζών που αλληλοεπιδρούν με HDV). Οι ομοιότητες μεταξύ των δύο περιβαλλόντων δείχνουν τη δυνατότητα εφαρμογής του υβριδικού μοντέλου για σενάρια πραγματικού κόσμου που αφορούν τόσο AV όσο και HDV.

Συμπερασματικά, η συνεισφορά της έρευνας είναι διπλή: αναπτύσσεται ένα μοντέλο υβριδικών συστημάτων για μακροπρόθεσμη πρόβλεψη τροχιών των πεζών και επιδεικνύει τη δυνατότητα εφαρμογής του μοντέλου σε σενάρια του πραγματικού κόσμου. Το μοντέλο βελτιώνει την ακρίβεια πρόβλεψης για τις τροχιές διέλευσης πεζών και παρέχει πληροφορίες για τη συμπεριφορά των πεζών, όταν αλληλοεπιδρούν με αυτόνομα οχήματα.

Μελλοντική έρευνα θα μπορούσε να συμπεριλάβει περισσότερες σχετικές πληροφορίες στο μοντέλο μετάβασης διακριτών καταστάσεων, όπως περιβαλλοντικούς παράγοντες, καθώς και να μελετήσει την αλληλεπίδραση των πεζών σε εικονικό περιβάλλον με οχήματα οδηγούμενα από ανθρώπους.

Οι Deshpande et al. (2021) διερευνούν την πλοήγηση σε αστικές περιοχές, όπου τα αυτόνομα οχήματα αλληλοεπιδρούν με τους πεζούς ως ευάλωτους χρήστες του δρόμου. Το πρόβλημα πλοήγησης διατυπώνεται με την χρήση ενισχυτικής μάθησης πολλαπλών στόχων (Multi-objective RL ή MORL), λαμβάνοντας υπόψη στόχους, όπως η ασφάλεια, η ταχύτητα, η άνεση και η τήρηση των κανόνων κυκλοφορίας. Προτείνεται μια παραλλαγή βαθιάς μάθησης, Deep threshold lexicographic Q-learning, ως μέθοδος για την αυτόνομη πλοήγηση με την παρουσία πεζών.

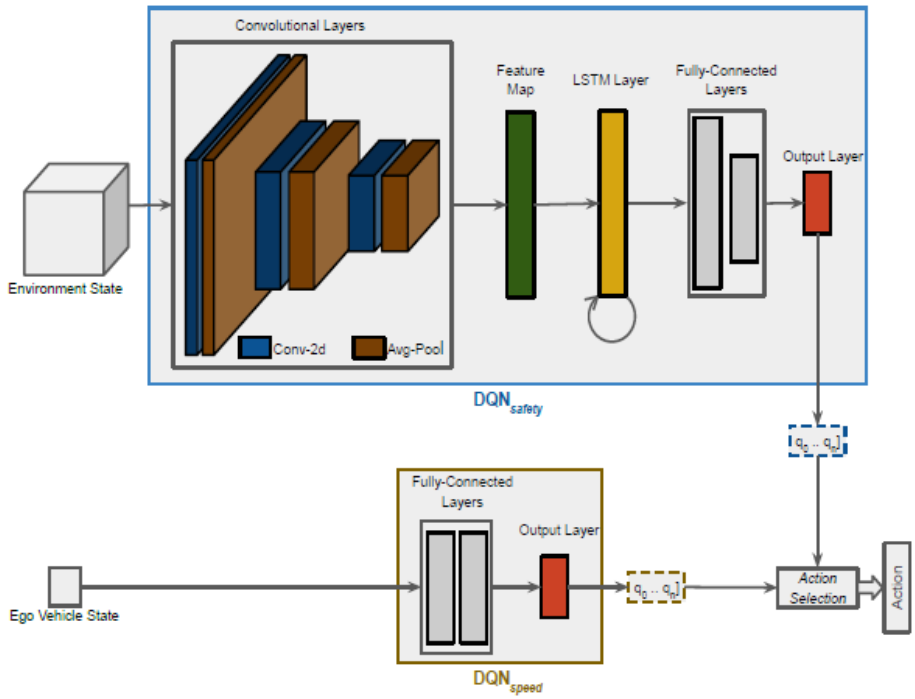
Η προτεινόμενη προσέγγιση αξιολογείται χρησιμοποιώντας ένα προσαρμοσμένο αστικό περιβάλλον που αναπτύχθηκε στον προσομοιωτή CARLA. Η απόδοση του παράγοντα DQN πολλαπλών στόχων συγκρίνεται με μια παραλλαγή ενός παράγοντα DQN μεμονωμένου στόχου (Single-Objective RL ή SORL) με νέα συνάρτηση επιβράβευσης, τόσο σε γνωστά όσο και σε άγνωστα περιβάλλοντα. Τα αποτελέσματα της αξιολόγησης δείχνουν ότι η προσέγγιση πολλαπλών στόχων υπερτερεί της προσέγγισης ενός στόχου σε όλες τις πτυχές.

Η εργασία υπογραμμίζει τους περιορισμούς των συμβατικών συστημάτων πλοήγησης στην αντιμετώπιση της πολυπλοκότητα των αστικών περιβαλλόντων με τους πεζούς. Η ύπαρξη διαφορετικών αντιφατικών στόχων, όπως η ασφάλεια και η ταχύτητα, αντιμετωπίζεται με την χρήση ενισχυτικής μάθησης πολλαπλών στόχων, η οποία επιτρέπει στο AV να εξετάζει και να εξισορροπεί πολλαπλούς στόχους, διασφαλίζοντας την ασφαλή και αποτελεσματική πλοήγηση μεταξύ των πεζών.

Στόχος του αλγορίθμου MORL είναι η αποφυγή των συγκρούσεων με τους πεζούς, και επιπλέον η απόκτηση της επιθυμητής ταχύτητας από το αυτόνομο όχημα. Στην ενισχυτική μάθηση πολλαπλών στόχων κάθε στόχος μαθαίνεται από ένα ξεχωριστό πράκτορα. Έπειτα, αναπτύσσεται μία κοινή πολιτική από τους πράκτορες για όλους τους στόχους.

Δύο πράκτορες MORL με διαφορετική αρχιτεκτονική νευρωνικών δικτύων (NN) στο DQNsafety και όμοια δομή στο DQNspeed συγκρίνονται στη μελέτη με 2 αντίστοιχους πράκτορες SORL. Ο πράκτορας 1 έχει συνελεκτικά στρώματα (CNN) που συνδέονται με 2 πλήρως συνδεδεμένα στρώματα, ενώ ο πράκτορας 2 (Εικ.4) έχει στρώμα LSTM, είδος επαναλαμβανόμενου νευρωνικού δικτύου (RNN), που συνδέεται με πλήρως συνδεδεμένα στρώματα. Τα δεδομένα εισόδου αποτελούνται από 3 συνελεκτικά στρώματα (32,64,64), που περιγράφουν το περιβάλλον και το στρώμα εξόδου αποτελείται από 4 νευρώνες, που αναπαριστούν το σύνολο ενεργειών. Το DQNspeed δέχεται ως δεδομένα εισόδου την κατάσταση του

εξεταζόμενου οχήματος, δηλαδή την ταχύτητά του. Στα δεδομένα εξόδου εφαρμόζεται ο αλγόριθμος threshold lexicographic Q-learning για να παραχθεί η τελική ενέργεια.



Εικόνα 4: Διαφορετική δομή DQN για Στόχο Ταχύτητας και Ασφάλειας (Πηγή: Deshpande et al., 2021)

Οι πράκτορες εκπαιδεύονται στον προσομοιωτή CARLA, που είναι ανοιχτό λογισμικό προσομοίωσης, για 500.000 βήματα, όμως το επεισόδιο τελειώνει, όταν ο πράκτορας επιτύχει τον στόχο του ή αντιμετωπίσει κατάσταση σύγκρουσης με χρονικό βήμα t=0,1 sec. Το όχημα ξεκινά από τη θέση εκκίνησης σε κάθε επεισόδιο και το επιθυμητό όριο ταχύτητας είναι 8 m/s. Σε κάθε χρονικό βήμα 35 πεζοί πρέπει να βρίσκονται σε ακτίνα 35 μέτρων με τυχαίους στόχους και ταχύτητες 0,4- 1,2 m/s. Το 80% των πεζών μπορεί να διασχίσει τον δρόμο και διατηρείται συνεχώς αυξημένη κίνηση πεζών γύρω από το εξεταζόμενο όχημα, αντικαθιστώντας τους πεζούς που απομακρύνονται με νέους.

Όλοι οι πράκτορες εξετάζονται για 100 επεισόδια στο γνωστό περιβάλλον εκπαίδευσης και σε νέα άγνωστα περιβάλλοντα. Η προτεινόμενη μέθοδος πολλαπλών στόχων RL έχει καλύτερα αποτελέσματα από αυτά της RL μεμονωμένου στόχου. Βασικός στόχος είναι η αποφυγή συγκρούσεων σε αστικά περιβάλλοντα με πεζούς, το οποίο επιτυγχάνεται και από τους 2 πράκτορες σε ποσοστό 100%. Οι πράκτορες SORL φρενάρουν περισσότερο από τους MORL, διότι η επιβράβευση τους είναι βαθμωτό μέγεθος που προσπαθούν συνεχώς να εξισορροπήσουν, ενώ οι MORL επιτυγχάνουν πιο φυσική οδήγηση με λιγότερες

στάσεις. Όλοι οι πράκτορες διατηρούν το όριο ταχύτητας. Οι πράκτορες με LSTM στρώματα διατηρούν υψηλότερες ταχύτητες, καθώς η μνήμη τους επιτρέπει να προβλέπουν τις προθέσεις των πεζών με αποτέλεσμα τη μείωση του χρόνου αναμονής στις διασταυρώσεις και την αύξηση της απόστασης από τους πεζούς.

Οι Predhumeau et al. (2021) προτείνουν τη χρήση ενός υβριδικού πρακτορικού μοντέλου για την πρόβλεψη τροχιών των πεζών σε κοινόχρηστους χώρους κατά τη διάρκεια αλληλεπιδράσεων με αυτόνομα οχήματα (AVs). Το μοντέλο συνδυάζει το μοντέλο κοινωνικής δύναμης (SFM) με έναν αλγόριθμο λήψης αποφάσεων για να καταγράψει την πολυπλοκότητα της συμπεριφοράς των πεζών. Το SFM, που χρησιμοποιείται ευρέως στην προσομοίωση πεζών, ενσωματώνει απωθητικές και ελκυστικές δυνάμεις με βάση τις θέσεις, τις ταχύτητες και τις κοινωνικές συμπεριφορές. Το μοντέλο έχει βαθμονομηθεί και επικυρωθεί χρησιμοποιώντας διάφορα σύνολα δεδομένων, συμπεριλαμβανομένων ελεγχόμενων πειραμάτων και πραγματικών σεναρίων.

Το προτεινόμενο υβριδικό μοντέλο στοχεύει να ξεπεράσει τους περιορισμούς του SFM και να βελτιώσει την ακρίβεια πρόβλεψης. Με την ενσωμάτωση ενός αλγόριθμου λήψης αποφάσεων, λαμβάνει υπόψη ατομικές και ομαδικές συμπεριφορές κατά τη διάρκεια αλληλεπιδράσεων με AV. Οι πεζοί σε μια ομάδα παίρνουν κοινές αποφάσεις, ενώ οι μεμονωμένοι πεζοί λαμβάνουν υπόψη τη δική τους θέση, αλλά ακολουθούν την απόφαση της ομάδας, όταν διστάζουν. Σε καταστάσεις επικείμενης σύγκρουσης, οι πεζοί λαμβάνουν ατομικές αποφάσεις ανεξάρτητα από την επιλογή της ομάδας. Το μοντέλο διασφαλίζει επίσης τη συνοχή της ομάδας, με όλα τα μέλη να στρέφονται ή να τρέχουν προς την ίδια κατεύθυνση.

Για την αξιολόγηση του προτεινόμενου μοντέλου, οι προσομοιωμένες τροχιές συγκρίνονται με δεδομένα του πραγματικού κόσμου. Πιο συγκεκριμένα, γίνεται χρήση των ανοικτών συνόλων δεδομένων DUT για ποσοτική αξιολόγηση και CITR για ποιοτική αξιολόγηση. Το 1ο σύνολο δεδομένων περιλαμβάνει δεδομένα πλευρικής αλληλεπίδρασης μεταξύ οχήματος και ροής πεζών και αξιολογεί την δυνατότητα αναπαραγωγής διαφορετικών συμπεριφορών πεζών που παρατηρήθηκαν, όπως η επιτάχυνση, το τρέξιμο, η επιβράδυνση, ο δισταγμός και η κίνηση σε ομάδες. Το 2ο σύνολο δεδομένων περιλαμβάνει 4 σενάρια αλληλεπίδρασης μεταξύ 8 πεζών και ενός AV. Το όχημα φθάνει από μπροστά, πίσω από τους πεζούς, από τη δεξιά τους πλευρά και από την μεριά των πεζών με 4 πεζούς να κοιτούν τους άλλους 4. Αξιολογείται η ακρίβεια του μοντέλου και συγκρίνεται με αυτή του SFM με μετρήσεις του τελικού σφάλματος: μετατόπισης (FDE), γραμμικής ταχύτητας (FLVE), προσανατολισμού (FOE) και απόστασης πλησιέστερης προσέγγισης (CADE) για την ποσοτικοποίηση της ακρίβειας της πρόβλεψης.

Οι προβλέψεις του υβριδικού μοντέλου συγκρίνονται με αυτές του SFM. Τα αποτελέσματα δείχνουν ότι το προτεινόμενο μοντέλο υπερέχει του SFM στην πρόβλεψη μετατόπισης πεζών, γραμμικής ταχύτητας, προσανατολισμού και απόστασης προσέγγισης από τα AV. Το μοντέλο αναπαράγει με ακρίβεια διάφορες συμπεριφορές πεζών που παρατηρούνται σε πραγματικά σενάρια, όπως το τρέξιμο για να διασχίσει, το σταμάτημα για να περιμένει χωρίς να παρεκκλίνει και η παραμονή σε ομάδες χωρίς συγκρούσεις. Μειώνει, επίσης, τις αποκλίσεις από τις επιθυμητές τροχιές και αποφεύγει τις συγκρούσεις. Το μοντέλο SFM παρουσίασε συγκρούσεις σε ποσοστό 0-5% σε διαφορετικά σενάρια, καθώς οι δυνάμεις αντισταθμίζονται μεταξύ τους δημιουργώντας μη ρεαλιστικές συμπεριφορές. Αντίθετα, στο υβριδικό μοντέλο δεν παρατηρήθηκαν συγκρούσεις με τη χρήση των ίδιων δεδομένων. Αυτό συνέβη, διότι οι πεζοί στην προσομοίωση αντιλαμβάνονται το AV πριν φτάσει στο σημείο διέλευσης και έχουν αρκετό χρόνο για την αποφυγή της σύγκρουσης.

Ο προσομοιωτής υλοποιώντας το προτεινόμενο μοντέλο επιτρέπει τη δοκιμή αλγορίθμων πλοήγησης AV μέσα σε προσομοιωμένα πλήθη και τη δυναμική προσαρμογή στη συμπεριφορά των πεζών. Η υπολογιστική απόδοση του υβριδικού μοντέλου επιτρέπει προσομοιώσεις σε πραγματικό χρόνο με μεγάλο αριθμό πεζών, πιο συγκεκριμένα: αλληλεπίδραση AV με πλήθος 100 πεζών και πυκνότητα 0,5 πεζοί/m^2. Επιπλέον, ο προσομοιωτής μπορεί να χρησιμοποιηθεί για διαδικτυακές προβλέψεις και να τρέξει γρηγορότερα από ότι σε πραγματικό χρόνο (real-time). Συμπερασματικά, το υβριδικό μοντέλο ενισχύει τον ρεαλισμό και τις προγνωστικές δυνατότητες των προσομοιώσεων πεζών, στο πλαίσιο των αλληλεπιδράσεων με AV, παρέχοντας ένα πολύτιμο εργαλείο για το σχεδιασμό ασφαλών και αποτελεσματικών κοινόχρηστων χώρων. Περαιτέρω έρευνα θα μπορούσε να περιλαμβάνει περισσότερα σενάρια μελέτης και κοινωνικές ομάδες.

2.3 ΣΥΜΠΕΡΑΣΜΑΤΑ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Τα συμπεράσματα που προέκυψαν από την ανάλυση της βιβλιογραφίας σχετικά με την ανάπτυξη μοντέλων οδήγησης με χρήση ενισχυτικής και αντίστροφης ενισχυτικής μάθησης, καθώς και μοντέλων πρόβλεψης τροχιών των πεζών μέσα από πραγματικά δεδομένα, σε αλληλεπίδραση με αυτόνομα οχήματα, είναι τα εξής:

- Αρχικά είναι εμφανής η ανάγκη αύξησης της ασφάλειας στους δρόμους, η οποία μπορεί να επιτευχθεί με την αύξηση των αυτόνομων οχημάτων και την ανάπτυξη αποτελεσματικών στρατηγικών οδήγησης, οι οποίες λαμβάνουν υπόψη τις αλληλεπιδράσεις με τους πιο ευάλωτους χρήστες του δρόμου και δίνουν προτεραιότητα στην αποφυγή των συγκρούσεων.

- Η ομαλή ενσωμάτωση των αυτόνομων οχημάτων στην κυκλοφορία μπορεί να πραγματοποιηθεί με την ανάπτυξη μοντέλων που αναγνωρίζουν τις ανθρώπινες αξίες και μιμούνται τις ανθρώπινες συμπεριφορές. Αυτό μπορεί να επιτευχθεί με την χρήση αντίστροφης ενισχυτικής μάθησης, η οποία επιτρέπει την μοντελοποίηση των ανθρώπινων προτιμήσεων και επομένως την ευκολότερη αποδοχή από τους χρήστες του δρόμου.
- Σε σχέση με την μοντελοποίηση, το μοντέλο μέγιστης εντροπίας αντίστροφης ενισχυτικής μάθησης επιτρέπει την αντιμετώπιση της αβεβαιότητας και της ασάφειας στην λήψη αποφάσεων, ενώ δεν εμφανίζει προτιμήσεις στην επιλογή ορισμένων διαδρομών (label-bias). Επιπλέον, το μοντέλο μέγιστης εντροπίας μπορεί να χρησιμοποιηθεί για την μοντελοποίηση της ανθρώπινης συμπεριφοράς και την προσθήκη πληροφορίων σχετικά με το περιβάλλον και με το άτομο.
- Η χρήση πραγματικών δεδομένων για την εκπαίδευση και την αξιολόγηση μοντέλων οδήγησης, σε συνδυασμό με τη χρήση αντίστροφης ενισχυτικής μάθησης, εξασφαλίζει μεγαλύτερη ακρίβεια και πιο ρεαλιστική απεικόνιση των συμπεριφορών των οδηγών και των πεζών σε μια πληθώρα πιθανών καταστάσεων.
- Η διερεύνηση διαφορετικών σεναρίων αλληλεπιδράσεων πεζών με οχημάτων είναι αναγκαία για την δυνατότητα έγκαιρης πρόβλεψης της συμπεριφοράς και των τροχιών των πεζών και, επομένως, την αποφυγή των συγκρούσεων και την εξασφάλιση της ασφάλειας σε δυναμικά κυκλοφοριακά περιβάλλοντα.

Η μελέτη της αλληλεπίδρασης των αυτόνομων οχημάτων με τους πεζούς είναι ακόμα σε πρώιμο στάδιο, επομένως είναι απαραίτητη η ενδελεχής κατανόηση των μικροσκοπικών και δυναμικών χαρακτηριστικών ,που διέπουν την αλληλεπίδραση αυτή και η μετέπειτα ανάπτυξη μοντέλων, που θα περιγράφουν τη βέλτιστη συμπεριφορά του οχήματος με βάση τη συμπεριφορά του πεζού.

ΚΕΦΑΛΑΙΟ 3. ΜΕΘΟΔΟΛΟΓΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

3.1 ΕΙΣΑΓΩΓΗ

Στο παρόν κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο, στο οποίο βασίστηκε η ανάλυση των δεδομένων και η εφαρμογή της μεθόδου για την παρούσα Διπλωματική Εργασία. Αρχικά, παρουσιάζεται η μεθοδολογική προσέγγιση που ακολουθήθηκε. Στη συνέχεια, ορίζεται η ενισχυτική μάθηση και γίνεται η περιγραφή του προβλήματος της αντίστροφης ενισχυτικής μάθησης και της μεθόδου μέγιστης εντροπίας. Έπειτα, αναλύονται οι έννοιες της συνάρτησης ανταμοιβής, της συνάρτησης αξίας, της πολιτικής και της διαδικασίας απόφασης Markov. Τέλος, γίνεται αναφορά στις παραμέτρους, στις μεθόδους χωρίς μοντέλα και στην τυχαιότητα του μοντέλου.

3.2 ΠΡΟΤΕΙΝΟΜΕΝΗ ΠΡΟΣΕΓΓΙΣΗ

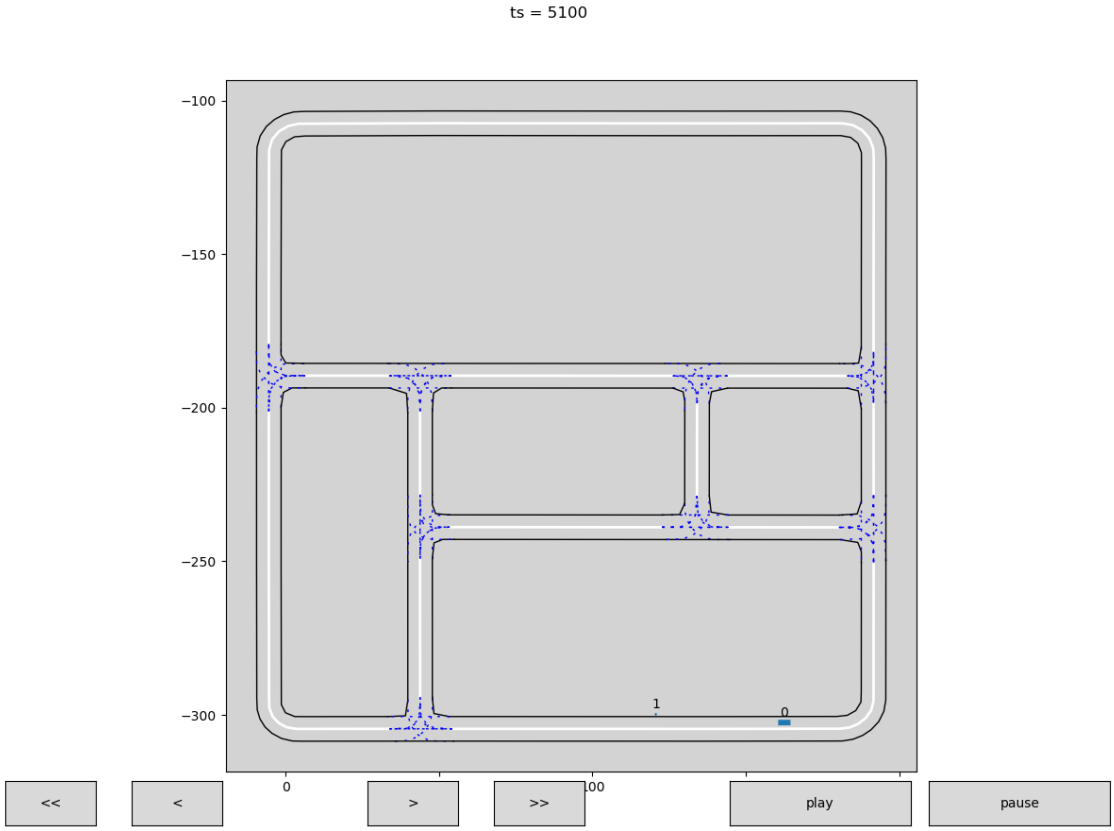
Η μεθοδολογική προσέγγιση που ακολουθήθηκε για την ανάπτυξη του αλγορίθμου αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας από δεδομένα συλλεγμένα από ένα πείραμα εικονικής πραγματικότητας, περιλαμβάνει 3 βασικά βήματα:

Βήμα 1: Προετοιμασία δεδομένων

Τα δεδομένα έχουν ληφθεί μέσω ενός πειράματος εικονικής πραγματικότητας που πραγματοποιήθηκε στην Καρλσρούη της Γερμανίας από το Ερευνητικό Κέντρο Πληροφορικής FZI. (Εικόνα 5). Στα δεδομένα περιλαμβάνονται η θέση x και y του αυτόνομου οχήματος και του πεζού, οι τιμές ταχύτητας στις διαστάσεις x και y του οχήματος, η γωνία εκτροπής του πράκτορα σε rad, και το μήκος και το πλάτος του πράκτορα. Για την εισαγωγή των δεδομένων στον αλγόριθμο Max-Ent IRL γίνεται ο υπολογισμός των τιμών της επιτάχυνσης στους άξονες x και y, της χρονικής και χωρικής απόστασης από το προηγούμενο όχημα, της πλευρικής θέσης και τέλος της πλευρικής απόστασης του οχήματος, μέσω της χρήσης χάρτη υψηλής ευκρίνειας του οδικού δικτύου και των δεδομένων των τροχιών της προσομοίωσης. Επιπλέον, υπολογίζονται η ταχύτητα του πεζού, η ταχύτητα του οχήματος, καθώς και η επιτάχυνση του οχήματος μέσω των διαθέσιμων δεδομένων από τις τροχιές.

Βήμα 2: Ανάπτυξη αλγορίθμου Max-Ent IRL

Για την ανάπτυξη του αλγορίθμου IRL μέγιστης εντροπίας, είναι απαραίτητος ο καθορισμός του χώρου καταστάσεων και ενεργειών του πράκτορα. Με βάση τα δεδομένα που συλλέχθηκαν από το εικονικό πείραμα καθορίστηκαν τα χαρακτηριστικά που περιγράφουν την αλληλεπίδραση μεταξύ του οχήματος και του πεζού. Αυτά αποτελούνται από την ταχύτητα του οχήματος, τη διαφορά ταχύτητας οχήματος-πεζού και τη χωρική απόσταση μεταξύ οχήματος και πεζού. Στη συνέχεια, προσδιορίζονται οι ενέργειες που μπορεί να εκτελέσει το αυτόνομο όχημα- πράκτορας, οι οποίες είναι η ομαλή και απότομη επιτάχυνση και η ομαλή, μέση και απότομη επιβράδυνση. Έπειτα, γίνεται στατιστική ανάλυση των χαρακτηριστικών των καταστάσεων και των ενεργειών και καθορίζονται οι κλάσεις των χαρακτηριστικών για τον αλγόριθμο. Τέλος, τα δεδομένα εισάγονται στον αλγόριθμο αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας.



Εικόνα 5: Οπτικοποίηση των δεδομένων που συλλέγονται στον προσομοιωτή εικονικής πραγματικότητας (VR) (Πηγή: INTERACTION dataset)

Βήμα 3: Αξιολόγηση αλγορίθμου

Το αποτέλεσμα του αλγόριθμου αξιολογείται με βάση το βάρη επιβράβευσης των χαρακτηριστικών, τα οποία πρέπει να εμφανίζουν όμοιες τιμές στις τελευταίες επαναλήψεις του αλγορίθμου, δηλαδή να συγκλίνουν. Η σύγκλιση αποτελεί ένδειξη ότι η εκπαίδευση του αλγορίθμου έχει φτάσει σε ένα ικανοποιητικό επίπεδο. Επιπλέον, η φυσική σημασία των βαρών επιβράβευσης είναι σημαντική για την αξιολόγηση του αλγορίθμου, καθώς δείχνουν πώς ο πράκτορας αξιολογεί τα διαφορετικά χαρακτηριστικά. Επιπρόσθετα, η σημασία του τελικού διανύσματος ανταμοιβών καταστάσεων r που προκύπτει από τον αλγόριθμο αξιολογείται με βάση τις ορισμένες καταστάσεις και την σημασία των εκάστοτε χαρακτηριστικών για το σενάριο έρευνας. Η ερμηνευσιμότητα των αποτελεσμάτων του αλγορίθμου έχει πολλαπλή σημασία, γιατί υποδεικνύει την τεχνική απόδοση, την χρησιμότητα και την δυνατότητα εφαρμογής του αλγόριθμου σε πραγματικά σενάρια. Ένας άλλος τρόπος αξιολόγησης του αλγορίθμου είναι ο έλεγχος με την χρήση νέων δεδομένων, στα οποία δεν έχει εκπαιδευθεί ο αλγόριθμος. Επίσης, μπορεί να ελεγχθεί η ταχύτητα και η δυνατότητα σύγκλισης του αλγορίθμου, καθώς και η δυνατότητα χρήσης του σε σενάρια μεγαλύτερης κλίμακας, χωρίς μεγάλη αύξηση του χρόνου σύγκλισης για την ποιότητα και την απόδοσή του.

3.3 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

3.3.1 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση (reinforcement learning) ανήκει στην ευρύτερη κατηγορία της Μηχανικής Μάθησης (ML) και αφορά αλγόριθμους που εκπαιδεύονται μέσω της αλληλεπίδρασης με το περιβάλλον. Στην ενισχυτική μάθηση (RL) ο πράκτορας επιλέγει ενέργειες και λαμβάνει ανατροφοδότηση με τη μορφή ανταμοιβών ή κυρώσεων από το περιβάλλον. Στόχος του αλγορίθμου είναι να μάθει μια πολιτική που να μεγιστοποιεί την ανταμοιβή με την πάροδο του χρόνου. Μέσω μιας διαδικασίας δοκιμής και λάθους, ο πράκτορας διερευνά διαφορετικές ενέργειες, παρατηρεί τα αποτελέσματα και ενημερώνει την πολιτική του με βάση τις ανταμοιβές που έχει λάβει.

3.3.2 Αντίστροφη Ενισχυτική Μάθηση

Στόχος των αλγόριθμων RL είναι η μάθηση της βέλτιστης πολιτικής, που μεγιστοποιεί τη σωρευτική ανταμοιβή σε ένα δεδομένο περιβάλλον, όμως κάποιες φορές η συνάρτηση ανταμοιβής είναι δύσκολο ή αδύνατο να οριστεί. Σε τέτοιες περιπτώσεις, οι αλγόριθμοι αντίστροφης ενισχυτικής μάθησης (inverse reinforcement learning) χρησιμοποιούνται για την ανάκτηση της συνάρτησης ανταμοιβής από την παρατήρηση της συμπεριφοράς ενός πράκτορα ή από επιδείξεις ειδικών σε συνδυασμό με το μοντέλο του

περιβάλλοντος. Η αντίστροφη ενισχυτική μάθηση (IRL) είναι ένα πρόσφατα αναπτυγμένο υποπεδίο της ενισχυτικής μάθησης και μπορεί να λύσει το αντίστροφο πρόβλημα της ενισχυτικής μάθησης.

Η αντίστροφη ενισχυτική μάθηση αφορά τη μάθηση από τους ανθρώπους και πιο συγκεκριμένα την εκμάθηση των στόχων, των αξιών ή των ανταμοιβών ενός πράκτορα με την παρατήρηση της συμπεριφοράς του. Ο αλγόριθμος IRL υποθέτει ότι η συμπεριφορά του ειδικού είναι βέλτιστη σε σχέση με κάποια άγνωστη συνάρτηση ανταμοιβής. Στα προβλήματα αντίστροφης ενισχυτικής μάθησης περιλαμβάνονται επιδείξεις ειδικών, που αποτελούνται από ακολουθίες καταστάσεων και ενεργειών μεταξύ του πράκτορα και του περιβάλλοντος. Αναλύοντας τις ενέργειες του ειδικού, ένας αλγόριθμος IRL προσπαθεί να συμπεράνει την υποκείμενη συνάρτηση ανταμοιβής, που εξηγεί καλύτερα την παρατηρούμενη συμπεριφορά. Έπειτα, η συνάρτηση ανταμοιβής μπορεί να χρησιμοποιηθεί για να καθοδηγήσει την εκμάθηση πολιτικών ή της συμπεριφοράς ενός πράκτορα, ο οποίος εκτελεί εργασίες στο ίδιο ή σε ένα παρόμοιο περιβάλλον. Αυτό γίνεται είτε για σκοπούς μοντελοποίησης, είτε για να παραχθεί μία μέθοδος που επιτρέπει την μίμηση μιας συγκεκριμένης συμπεριφοράς επίδειξης από τον πράκτορα (Ramachandran & Amir, 2007).

Η ενισχυτική και αντίστροφη ενισχυτική μάθηση χαρακτηρίζονται από τα ακόλουθα στοιχεία:

Χώρος καταστάσεων (S): Το σύνολο των πιθανών καταστάσεων στο περιβάλλον έρευνας. Κάθε κατάσταση αντιπροσωπεύει μια συγκεκριμένη διαμόρφωση ή κατάσταση, στην οποία μπορεί να βρεθεί ο πράκτορας.

Χώρος ενεργειών (A): Το σύνολο των πιθανών ενεργειών που μπορεί να ακολουθήσει ο πράκτορας σε μια δεδομένη κατάσταση. Οι ενέργειες είναι οι επιλογές ή οι αποφάσεις που λαμβάνονται από τον πράκτορα για τη μετάβαση από τη μια κατάσταση στην άλλη.

Συνάρτηση ανταμοιβής (R): Αντιστοιχίζει ένα ζεύγος κατάστασης- δράσης (s, a) σε μια κλιμακωτή τιμή, υποδεικνύοντας την επιθυμία ή την ποιότητα της ανάληψης ενέργειας a στην κατάσταση s.

Συνάρτηση πιθανότητας μετάβασης (P ή T): καθορίζει τις πιθανότητες μετάβασης από τη μια κατάσταση στην άλλη, όταν πραγματοποιείται μια συγκεκριμένη ενέργεια. Ενσωματώνει τη στοχαστική φύση του περιβάλλοντος και καταγράφει την αντίδρασή του στις ενέργειες του πράκτορα.

Τροχιές (τ) ή επιδείξεις ειδικών (D): Η παρατηρούμενη συμπεριφορά ή οι επιδείξεις ειδικών που παρέχονται. Αυτές οι επιδείξεις αποτελούνται συνήθως από τροχιές ή χρονικές ακολουθίες ζευγών

κατάστασης-δράσης, που δείχνουν πώς ο ειδικός αλληλοεπιδρά με το περιβάλλον και επιτυγχάνει τα επιθυμητά αποτελέσματα.

Πολιτική (π): Η πολιτική καθορίζει τη συμπεριφορά ή τη στρατηγική λήψης αποφάσεων του πράκτορα. Αντιστοιχίζει καταστάσεις σε ενέργειες και αναπαριστά πώς ο πράκτορας επιλέγει ενέργειες σε διαφορετικές καταστάσεις.

Συνάρτηση αξίας (V ή Q): Η συνάρτηση αξίας εκτιμά τις αναμενόμενες αθροιστικές ανταμοιβές ή τιμές, που σχετίζονται με την ύπαρξη σε μια συγκεκριμένη κατάσταση ή ζεύγος κατάστασης-ενέργειας. Παρέχει ένα μέτρο της μακροπρόθεσμης επιθυμίας ή χρησιμότητας των καταστάσεων ή των ζευγών κατάστασης-ενέργειας.

3.3.3 Συνάρτηση Ανταμοιβής

Η συνάρτηση ανταμοιβής ποσοτικοποιεί την επιθυμία ή την αξία του πράκτορα, όταν βρίσκεται σε μια συγκεκριμένη κατάσταση ή όταν αναλαμβάνει μια συγκεκριμένη ενέργεια. Η συνάρτηση ανταμοιβής επηρεάζει τη συμπεριφορά ενός πράκτορα στην ενισχυτική μάθηση και χρησιμοποιείται για να συμπεράνει τις υποκείμενες προτιμήσεις του πράκτορα στην αντίστροφη ενισχυτική μάθηση.

Στην αντίστροφη ενισχυτική μάθηση, η συνάρτηση ανταμοιβής συνάγεται από τροχιές ή επιδείξεις ειδικών. Αποτυπώνει τις υποκείμενες προτιμήσεις του ειδικού και έχει καθοριστική σημασία στη διαδικασία λήψης αποφάσεων. Η συνάρτηση ανταμοιβής που προκύπτει χρησιμοποιείται για την κατανόηση της συμπεριφοράς του ειδικού ή για να επιτρέψει σε έναν πράκτορα να μιμηθεί τη συμπεριφορά του. (Εικ.6)

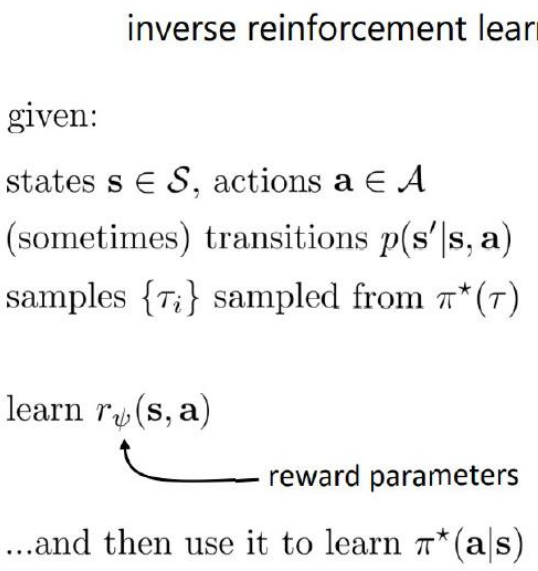
Ο στόχος της αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας είναι να βρεθεί μια συνάρτηση ανταμοιβής που να συμφωνεί με την παρατηρούμενη συμπεριφορά, μεγιστοποιώντας ταυτόχρονα την εντροπία ή την αβεβαιότητα στον χώρο της πολιτικής. Αυτό ενθαρρύνει την εξερεύνηση διαφορετικών πολιτικών, που ταιριάζουν με τη συμπεριφορά του ειδικού.

Οι τύποι για τη συνάρτηση ανταμοιβής στη Max-Ent IRL μπορούν να εξαχθούν χρησιμοποιώντας την αρχή της μέγιστης εντροπίας. Η συνάρτηση ανταμοιβής συχνά παραμετροποιείται ως ένας γραμμικός συνδυασμός συνάρτησης χαρακτηριστικών, το οποίο συμβολίζεται ως $\varphi(s, a)$ ή f_ζ. Το ζ συμβολίζει ένα μονοπάτι ή μία τροχιά από την κατάσταση s στην ενέργεια a. Οι συναρτήσεις χαρακτηριστικών καταγράφουν σχετικές πληροφορίες για την κατάσταση και τη δράση που επηρεάζουν την ανταμοιβή. Η συνάρτηση ανταμοιβής μπορεί να αναπαρασταθεί ως:

$$R(s, a) = \theta^T * \phi(s, a) \text{ ή } r(f_\zeta) = \theta^T * f_\zeta \qquad (1)$$

Το R(s, a) ή r(f_ξ) αντιπροσωπεύει την ανταμοιβή που σχετίζεται με την εκτέλεση της ενέργειας "a" στην κατάσταση "s". Το θ αντιπροσωπεύει το διάνυσμα βάρους που καθορίζει τη σημασία ή τη συμβολή του κάθε χαρακτηριστικού.

Ο αλγόριθμος Max-Ent IRL στοχεύει να βρει το διάνυσμα βάρους θ, που μεγιστοποιεί την πιθανότητα της παρατηρούμενης συμπεριφοράς των ειδικών, καθώς και την εντροπία της πολιτικής. Το πρόβλημα βελτιστοποίησης μπορεί να διατυπωθεί χρησιμοποιώντας εκτίμηση μέγιστης πιθανότητας (maximum likelihood estimation) και να λυθεί μέσω διαφόρων τεχνικών, όπως μεθόδων βελτιστοποίησης gradient ascent ή convex.



Εικόνα 6: Περιγραφή του αλγορίθμου IRL (Πηγή: Sergey Levine)

3.3.4 Συνάρτηση Αξίας

Η συνάρτηση αξίας χρησιμοποιείται για να αξιολογήσει την ποιότητα ή την αναμενόμενη απόδοση σε μια συγκεκριμένη κατάσταση ή σε ένα συγκεκριμένο ζεύγος κατάστασης-ενέργειας, βάσει μιας δεδομένης πολιτικής. Η συνάρτηση αξίας παρέχει μια εκτίμηση των σωρευτικών ανταμοιβών, που ένας πράκτορας αναμένεται να λάβει με την πάροδο του χρόνου, ακολουθώντας μια συγκεκριμένη πολιτική.

Στο Max-Ent IRL η συνάρτηση αξίας χρησιμοποιείται συνήθως ως βασικό στοιχείο στη διαδικασία εκμάθησης για την εκτίμηση των ανταμοιβών, που σχετίζονται με διαφορετικές καταστάσεις. Με την ενσωμάτωση της συνάρτησης αξίας το Max-Ent IRL μπορεί να συμπεράνει την υποκείμενη συνάρτηση ανταμοιβής, που εξηγεί καλύτερα την παρατηρούμενη συμπεριφορά ενός ειδικού.

Η συνάρτηση κατάστασης-αξίας, που συμβολίζεται ως V(s), αντιπροσωπεύει τις αναμενόμενες σωρευτικές ανταμοιβές που μπορεί να λάβει ένας πράκτορας, ξεκινώντας από μια συγκεκριμένη κατάσταση και ακολουθώντας μια δεδομένη πολιτική. Προσδιορίζει ποσοτικά τη μακροπρόθεσμη αξία του να βρίσκεσαι σε μια συγκεκριμένη κατάσταση.

Η συνάρτηση κατάστασης-αξίας μπορεί να οριστεί ως το αναμενόμενο άθροισμα των μελλοντικών ανταμοιβών με έκπτωση:

$$V_{\pi}(s) = E_{\pi}(\sum_{k=0} \gamma^{k*} R_{t+k=} \mid s_t = s) \qquad (2)$$

Εδώ, το γ είναι ο παράγοντας έκπτωσης (discount factor ή discount) που καθορίζει τη σημασία των άμεσων ανταμοιβών σε σύγκριση με τις μελλοντικές ανταμοιβές.

Το R_t αντιπροσωπεύει την ανταμοιβή που σχετίζεται με την ύπαρξη στην κατάσταση "s" στο χρονικό βήμα "t". Η προσδοκία αναλαμβάνεται σε πιθανές μελλοντικές τροχιές ξεκινώντας από την κατάσταση «s» και ακολουθώντας την πολιτική π.

Η συνάρτηση ενέργειας-αξίας ή συνάρτηση Q-value, που συμβολίζεται ως Q(s, a), αντιπροσωπεύει τις αναμενόμενες σωρευτικές ανταμοιβές που μπορεί να επιτύχει ένας πράκτορας, κάνοντας μια συγκεκριμένη ενέργεια σε μια συγκεκριμένη κατάσταση και ακολουθώντας μια δεδομένη πολιτική. Εκτιμά την αξία της λήψης μιας συγκεκριμένης ενέργειας σε μια συγκεκριμένη κατάσταση.

Η συνάρτηση ενέργειας-αξίας μπορεί να οριστεί ως το αναμενόμενο άθροισμα των μελλοντικών ανταμοιβών με έκπτωση, λαμβάνοντας υπόψη τη δράση που έγινε:

$$Q_{\pi}(s, a) = E_{\pi}(\sum_{k=0} \gamma^{k*} R_{t+k=} \mid s_t = s, a_t = a) \qquad (3)$$

Το γ είναι ο συντελεστής έκπτωσης, το R_t αντιπροσωπεύει την ανταμοιβή που λαμβάνεται στο χρονικό βήμα "t" και η προσδοκία λαμβάνεται για πιθανές μελλοντικές τροχιές ξεκινώντας από την κατάσταση "s", κάνοντας την ενέργεια "a" και ακολουθώντας την πολιτική π.

Η συνάρτηση αξίας παρέχει ένα μέτρο της αναμενόμενης απόδοσης, που μπορεί να επιτύχει ένας πράκτορας στο πλαίσιο μιας συγκεκριμένης πολιτικής. Χρησιμοποιείται για να καθοδηγήσει τη διαδικασία μάθησης και λήψης αποφάσεων σε αλγόριθμους ενισχυτικής μάθησης. Η συγκεκριμένη εκτίμηση ή ο υπολογισμός της συνάρτησης αξίας στο Max-Ent IRL μπορεί να εξαρτάται από τον αλγόριθμο που χρησιμοποιείται. Μια κοινή προσέγγιση είναι η επαναληπτική ενημέρωση της συνάρτησης αξίας μέσω μεθόδων, όπως η επανάληψη αξίας (value iteration) ή η αξιολόγηση πολιτικής (policy evaluation). Αυτές

οι τεχνικές εκτιμούν τη συνάρτηση αξίας με βάση τις παρατηρούμενες τροχιές και την εκτιμώμενη συνάρτηση ανταμοιβής.

3.3.5 Πολιτική

Στην ενισχυτική μάθηση, η πολιτική αναπαριστά μια χαρτογράφηση από την παρατήρηση του τρέχοντος περιβάλλοντος σε μια κατανομή πιθανοτήτων των ενεργειών που πρέπει να γίνουν. Συνοψίζει τη διαδικασία λήψης αποφάσεων ενός πράκτορα με βάση την παρατηρούμενη κατάσταση, καθορίζοντας πώς αντιδρά ο πράκτορας σε διάφορες συνθήκες περιβάλλοντος.

Στην ενισχυτική μάθηση, οι πολιτικές χρησιμοποιούνται από έναν πράκτορα που αλληλοεπιδρά με το περιβάλλον προκειμένου να μάθει τη βέλτιστη πολιτική μέσω δοκιμής και λάθους. Ο στόχος του πράκτορα είναι να εντοπίσει την βέλτιστη πολιτική, η οποία συμβολίζεται με π*, και μεγιστοποιεί ένα σωρευτικό σήμα ανταμοιβής με την πάροδο του χρόνου.

Πιο συγκεκριμένα, μια πολιτική π: s → a μπορεί να αναπαρασταθεί ως συνάρτηση που παίρνει μια κατάσταση ως είσοδο και εξάγει την ενέργεια, που πρέπει να γίνει σε αυτήν την κατάσταση. Οι πολιτικές μπορεί να είναι ντετερμινιστικές, δηλαδή αντιστοιχίζουν πάντα μια κατάσταση σε μια συγκεκριμένη ενέργεια ή στοχαστικές, δηλαδή αποδίδουν πιθανότητες σε διαφορετικές ενέργειες με βάση την κατάσταση.

- π (s) = a : Ντετερμινιστική πολιτική
- π(a|s) = P(a|s) : Στοχαστική πολιτική

Η softmax είναι ένας τύπος στοχαστικής πολιτικής που χρησιμοποιεί μια συνάρτηση για να εκχωρήσει πιθανότητες σε ενέργειες. Συχνά παραμετροποιείται από ένα σύνολο προτιμήσεων ενέργειας-αξίας το $Q_\pi(s, a)$. Η συνάρτηση softmax κανονικοποιεί αυτές τις προτιμήσεις για την απόκτηση μιας κατανομής πιθανότητας πάνω στις ενέργειες.

$$\pi(a|s) = \exp(Q_\pi(s, a)) / \sum \exp(Q_\pi(s, a')) \qquad (4)$$

3.3.6 Διαδικασία Απόφασης Markov

Η διαδικασία απόφασης Markov είναι μια διαδικασία στοχαστικού ελέγχου διακριτού χρόνου. Παρέχει ένα μαθηματικό πλαίσιο για τη μοντελοποίηση διαδοχικών προβλημάτων λήψης αποφάσεων στον τομέα της ενισχυτικής μάθησης. Ο όρος "Markov" αναφέρεται στην ιδιότητα της έλλειψης μνήμης, όπου οι μελλοντικές καταστάσεις μίας στοχαστικής διαδικασίας εξαρτώνται μόνο από την τρέχουσα κατάσταση

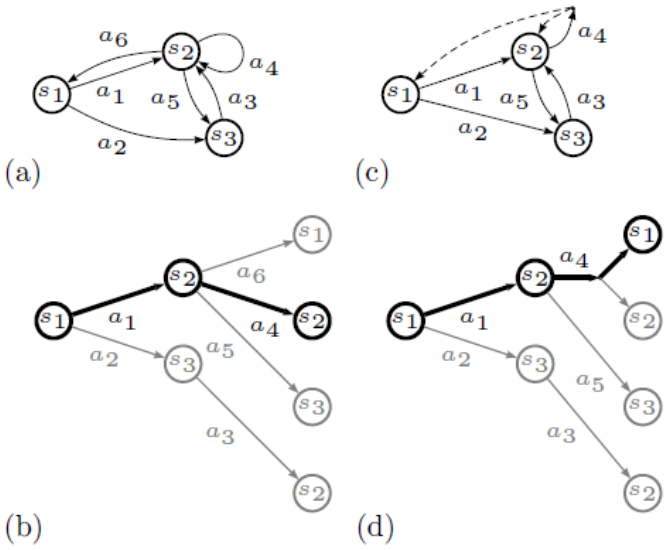
και την τρέχουσα δράση, που αναλαμβάνεται και όχι από τις παρελθοντικές. Αυτό σημαίνει ότι το παρελθόν μπορεί να αγνοηθεί μόλις γίνει γνωστό το παρόν.

Σε μία MDP ο πράκτορας αλληλοεπιδρά με το περιβάλλον σε μια σειρά από διακριτά χρονικά βήματα. Σε κάθε χρονικό βήμα, ο πράκτορας παρατηρεί την τρέχουσα κατάσταση του περιβάλλοντος και επιλέγει μια ενέργεια προς εκτέλεση. Στη συνέχεια, το περιβάλλον μεταβαίνει σε μια νέα κατάσταση με βάση την επιλεγμένη ενέργεια και ο πράκτορας λαμβάνει ένα σήμα ανταμοιβής με βάση την μετάβαση.

Η διαδικασία απόφασης Markov αποτελείται από ένα σύνολο (S,A,P,R).

$$P(s,s')=P(s_{t+1}=s'|s_t=s, a_t=a) \qquad (5)$$

Είναι η πιθανότητα μετάβασης από μία κατάσταση s με ενέργεια α σε χρόνο t, σε μία κατάσταση s' κατά τον χρόνο t+1. Όταν μεταβεί στην κατάσταση s' από την κατάσταση s η R(s,s') είναι η άμεση ανταμοιβή που λαμβάνεται από τη δράση α. Συνεπώς, η επόμενη κατάσταση s' εξαρτάται από την τρέχουσα κατάσταση s και την ενέργεια του πράκτορα α, οι οποίες είναι ανεξάρτητες από τις προηγούμενες καταστάσεις και ενέργειες, άρα οι μεταβάσεις ικανοποιούν την ιδιότητα Markov. (Εικ.7)



Εικόνα 7: Απεικόνιση Διαδικασίας Markov σε ένα ντετερμινιστικό (a) και στοχαστικό σύστημα (c) & μία αντίστοιχη διαδρομή (b, d) (Πηγή: Ziebart et al., 2008)

Στον αλγόριθμο Max-Ent IRL, μια διαδικασία απόφασης Markov χρησιμοποιείται ως υποκείμενο πλαίσιο για την εκμάθηση μιας συνάρτησης ανταμοιβής σύμφωνα με τις παρατηρούμενες τροχιές. Το διάνυσμα ανταμοιβής εκχωρεί μια τιμή ανταμοιβής σε κάθε κατάσταση, η οποία πρέπει να είναι συνεπής με τις

παρατηρούμενες τροχιές και τη δυναμική του MDP. Για να βρει τη συνάρτηση ανταμοιβής, ο αλγόριθμος χρησιμοποιεί βελτιστοποίηση gradient descent, προσαρμόζοντας τις τιμές ανταμοιβής επαναληπτικά με βάση τις παρατηρούμενες τροχιές. Το διάνυσμα ανταμοιβής που προκύπτει θα συλλάβει την επιθυμία ή τη χρησιμότητα κάθε κατάστασης, παρέχοντας ένα μέτρο των αναμενόμενων αθροιστικών ανταμοιβών, που μπορεί να αποκτήσει ένας πράκτορας, όταν βρίσκεται σε μια συγκεκριμένη κατάσταση στο δεδομένο πλαίσιο MDP.

3.3.7 Αντίστροφη Ενισχυτική Μάθηση Μέγιστης Εντροπίας

Η εντροπία είναι μια θερμοδυναμική ποσότητα, που αντιπροσωπεύει τη μη διαθεσιμότητα της θερμικής ενέργειας ενός συστήματος για μετατροπή σε μηχανικό έργο, που συχνά ερμηνεύεται ως ο βαθμός αταξίας ή τυχαιότητας στο σύστημα.

Η αρχή της μέγιστης εντροπίας δηλώνει ότι η κατανομή πιθανοτήτων που αντιπροσωπεύει την τρέχουσα κατάσταση γνώσης για ένα σύστημα, είναι αυτή με τη μεγαλύτερη εντροπία στο πλαίσιο των προηγούμενων δεδομένων. Η αρχή της μέγιστης εντροπίας χρησιμοποιείται στην αντίστροφη ενισχυτική μάθηση για την επίλυση των ασαφειών στην επιλογή κατανομών (Jaynes 1957). Με τον τρόπο αυτό, οι κατανομές των συμπεριφορών ταιριάζουν με τις προσδοκίες των χαρακτηριστικών, ενώ δεν είναι περισσότερο μεροληπτικές από όσο απαιτείται από τον περιορισμό.

Στην αντίστροφη ενισχυτική μάθηση μέγιστης εντροπίας, ο αλγόριθμος ακολουθεί μια διαδικασία δύο βημάτων: αρχικά εκτιμά μια βέλτιστη πολιτική, και στη συνέχεια συμπεραίνει την συνάρτηση ανταμοιβής. Ακολουθεί η περιγραφή των βημάτων του αλγορίθμου.

Ο αλγόριθμος ξεκινά υποθέτοντας μια τυχαία ή αρχική πολιτική. Χρησιμοποιεί έναν αλγόριθμο ενισχυτικής μάθησης, συχνά επανάληψη αξίας ή επανάληψη πολιτικής, για να βελτιώσει επαναληπτικά την πολιτική. Αυτοί οι αλγόριθμοι αξιολογούν και ενημερώνουν την πολιτική με βάση τις εκτιμώμενες τιμές ή τις Q- τιμές των ζευγών κατάστασης-ενέργειας. Ο στόχος είναι να βρεθεί μια πολιτική, που να μεγιστοποιεί τις αναμενόμενες αθροιστικές ανταμοιβές ή αξίες.

Όταν επιτευχθεί η βέλτιστη πολιτική, ο αλγόριθμος προχωρά στην εξαγωγή της υποκείμενης συνάρτησης ανταμοιβής. Υποθέτοντας ότι η συμπεριφορά του ειδικού μπορεί να εξηγηθεί από μια συνάρτηση ανταμοιβής, που είναι συνεπής με την παρατηρούμενη συμπεριφορά, ο αλγόριθμος μαθαίνει αυτήν την συνάρτηση ανταμοιβής συγκρίνοντας τη συμπεριφορά του ειδικού (επιδείξεις ή τροχιές) με αυτή που παράγεται από την εκτιμώμενη πολιτική. Ο αλγόριθμος προσαρμόζει τη συνάρτηση ανταμοιβής, ώστε ελαχιστοποιήσει την διαφορά μεταξύ των δύο συμπεριφορών.

Η βασική ιδέα του αλγορίθμου Max-Ent IRL είναι η εύρεση της συνάρτησης ανταμοιβής, η οποία εξηγεί τη συμπεριφορά του ειδικού, ενώ παράλληλα μεγιστοποιεί την εντροπία του χώρου πολιτικής. Αυτή η μεγιστοποίηση της εντροπίας ενθαρρύνει τον αλγόριθμο να συλλάβει ένα ευρύ φάσμα πιθανών πολιτικών, που συμφωνούν με τη συμπεριφορά του ειδικού.

Συνεπώς, το Max-Ent IRL συνδυάζει την εκτίμηση της πολιτικής και την διαδικασία εξαγωγής της συνάρτησης ανταμοιβής για να μάθει τόσο τη βέλτιστη πολιτική, όσο και τη συνάρτηση ανταμοιβής που εξηγεί τη συμπεριφορά του ειδικού. Ο αλγόριθμος βελτιώνει επαναληπτικά την πολιτική και προσαρμόζει τη συνάρτηση ανταμοιβής προκειμένου να επιτύχει μία ακριβή αντιστοίχιση μεταξύ της συμπεριφοράς του ειδικού και αυτής από την εκτιμώμενη πολιτική.

3.3.8 Μέθοδοι Χωρίς Μοντέλο

Στον τομέα της Μηχανικής Μάθησης, οι μέθοδοι χωρίς μοντέλα (model-free) αναφέρονται σε προσεγγίσεις, που μαθαίνουν άμεσα από δεδομένα, χωρίς να μοντελοποιούν ρητά τη δυναμική ή το περιβάλλον του συστήματος. Αυτές οι μέθοδοι επικεντρώνονται στην εκμάθηση βέλτιστων πολιτικών ή συναρτήσεων αξίας, χωρίς να απαιτούν την προηγούμενη γνώση των πιθανοτήτων μετάβασης ή της δομής της ανταμοιβής.

Οι μέθοδοι χωρίς μοντέλα στοχεύουν στην εκμάθηση της βέλτιστης συμπεριφοράς αλληλοεπιδρώντας με το περιβάλλον, παρατηρώντας τις καταστάσεις και τις ανταμοιβές, και βελτιώνοντας επαναληπτικά την πολιτική ή τις εκτιμήσεις αξίας με βάση τα δεδομένα που συλλέγονται. Είναι ιδιαίτερα χρήσιμες, όταν η δυναμική του συστήματος είναι άγνωστη, πολύπλοκη ή δύσκολο να μοντελοποιηθεί με ακρίβεια.

Στις μεθόδους χωρίς μοντέλα, η έμφαση δίνεται στη διαδικασία μάθησης που βασίζεται αποκλειστικά στα δεδομένα. Αυτές οι μέθοδοι χρησιμοποιούν διάφορες τεχνικές εκμάθησης, όπως ενισχυτική μάθηση, Q-learning, διαβαθμίσεις πολιτικής (policy gradients) ή μεθόδους Monte Carlo. Είναι ευέλικτες και εφαρμόζονται ευρέως σε τομείς, όπου η ρητή μοντελοποίηση της δυναμικής του συστήματος είναι δύσκολη ή ανέφικτη.

Το Max-Ent IRL είναι μία μέθοδος χωρίς μοντέλο, που χρησιμοποιείται για να συμπεράνει την υποκείμενη συνάρτηση ανταμοιβής από την παρατηρούμενη συμπεριφορά των ειδικών, χωρίς να απαιτεί γνώση της δυναμικής του συστήματος ή των πιθανοτήτων μετάβασης. Αντίθετα, αναζητά μία συνάρτηση ανταμοιβής, που είναι συνεπής με την παρατηρούμενη συμπεριφορά των ειδικών, ενώ παράλληλα μεγιστοποιεί την εντροπία ή την αβεβαιότητα στον χώρο της πολιτικής. Αυτό καθιστά το Max-Ent IRL κατάλληλο για σενάρια, όπου η ρητή μοντελοποίηση της δυναμικής του περιβάλλοντος είναι δύσκολη ή δεν είναι δυνατή,

επιτρέποντάς του να μαθαίνει από επιδείξεις ειδικών σε ένα ευρύ φάσμα εφαρμογών του πραγματικού κόσμου.

3.3.9 Παράμετροι & Υπερπαράμετροι Συνάρτησης

Οι παράμετροι αντιπροσωπεύουν τις εσωτερικές ρυθμίσεις ενός μοντέλου, καθορίζοντας τη συμπεριφορά και την απόδοσή του. Αυτές οι ρυθμίσεις προσαρμόζονται κατά τη διάρκεια της εκπαίδευσης για να βελτιστοποιηθεί η απόδοση του μοντέλου σε μια συγκεκριμένη εργασία. Τροποποιώντας τις παραμέτρους, ελέγχεται ο τρόπος με τον οποίο το μοντέλο επεξεργάζεται δεδομένα και βελτιώνεται η ακρίβεια και η ποιότητα των προβλέψεων του.

Οι υπερπαράμετροι είναι οι ρητά καθορισμένες παράμετροι που επηρεάζουν την διαδικασία εκπαίδευσης του μοντέλου και είναι κρίσιμες για τη βελτιστοποίησή του. Οι υπερπαράμετροι, συχνά αναφέρονται και ως απλά παράμετροι, απαιτούν χειροκίνητη ρύθμιση και συντονισμό για την αποτελεσματική εκπαίδευση του μοντέλου.

Ο συντελεστής έκπτωσης (discount factor) ή έκπτωση (discount) γ, χρησιμοποιείται για τον προσδιορισμό της σημασίας των άμεσων ανταμοιβών σε σύγκριση με τις μελλοντικές, στη διαδικασία λήψης αποφάσεων ενός πράκτορα και είναι μία υπερπαράμετρος. Η τιμή της είναι μεταξύ του 0 και του 1, όπου μια υψηλότερη τιμή δίνει μεγαλύτερη έμφαση στις μακροπρόθεσμες ανταμοιβές, ενώ μια χαμηλότερη τιμή δίνει προτεραιότητα στις άμεσες ανταμοιβές. Βοηθά τον πράκτορα να εξισορροπήσει τα βραχυπρόθεσμα κέρδη με τα πιθανά μακροπρόθεσμα οφέλη.

Ο ρυθμός εκμάθησης (learning rate) είναι μια υπερπαράμετρος, που καθορίζει το μέγεθος του βήματος σε κάθε επανάληψη, κατά την οποία ενημερώνονται οι παράμετροι του μοντέλου. Ελέγχει την ταχύτητα και το μέγεθος των ενημερώσεων των παραμέτρων, με τον υψηλότερο ρυθμό εκμάθησης να επιτρέπει την ταχύτερη σύγκλιση, αλλά μπορεί να οδηγήσει σε απόκλιση, ενώ ο χαμηλότερος ρυθμός εκμάθησης οδηγεί σε πιο αργή σύγκλιση, αλλά δυνητικά και σε πιο ακριβείς προσαρμογές.

Οι εποχές (epochs) αντιπροσωπεύουν τον αριθμό επαναλήψεων, που διατρέχουν ολόκληρο το σύνολο των δεδομένων εκπαίδευσης. Ο αριθμός των εποχών είναι μια σημαντική υπερπαράμετρος για τον αλγόριθμο. Κατά τη διάρκεια κάθε εποχής, οι παράμετροι του μοντέλου ενημερώνονται με βάση τα δεδομένα εισόδου και τις αντίστοιχες τιμές-στόχους, επιτρέποντας στο μοντέλο να μάθει από τα παραδείγματα εκπαίδευσης και να βελτιώσει την απόδοσή του. Είναι απαραίτητο να επιτευχθεί μια ισορροπία μεταξύ υποπροσαρμογής (underfitting) και υπερπροσαρμογής (overfitting): πολύ λίγες εποχές μπορεί να οδηγήσουν σε ένα υποεκπαιδευμένο μοντέλο, που αποτυγχάνει να καταγράψει πολύπλοκα μοτίβα, ενώ πάρα πολλές εποχές

μπορεί να οδηγήσουν σε υπερβολική προσαρμογή, όπου το μοντέλο έχει κακή απόδοση σε δεδομένα διαφορετικά από αυτά της εκπαίδευσής του. Ο προσδιορισμός του κατάλληλου αριθμού εποχών συνήθως περιλαμβάνει την παρακολούθηση της απόδοσης του μοντέλου σε ένα ξεχωριστό σύνολο αξιολόγησης για να βρεθεί το σημείο βέλτιστης γενίκευσης.

3.3.10 Τυχαιότητα

Ο αλγόριθμος μπορεί να αρχικοποιηθεί σε μια τυχαία κατάσταση και να δίνει διαφορετικά αποτελέσματα με τα ίδια δεδομένα εισόδου σε διαφορετικές εκτελέσεις του ίδιου αλγορίθμου. Με ρύθμιση του τυχαίου σπόρου numpy (numpy random seed) μπορεί να επιτευχθεί αναπαραγωγιμότητα των αποτελεσμάτων του αλγορίθμου. Το numpy random seed είναι μια αριθμητική τιμή που δημιουργεί ένα νέο σύνολο ή επαναλαμβάνει ψευδοτυχαίους αριθμούς. Η τιμή στον τυχαίο σπόρο numpy αποθηκεύει την κατάσταση τυχαιότητας. Εάν καλέσουμε τη συνάρτηση για παράδειγμα με random seed(1), δηλαδή χρησιμοποιώντας την τιμή 1 πολλές φορές, ο υπολογιστής θα εμφανίσει τους ίδιους τυχαίους αριθμούς σε κάθε εκτέλεση.

ΚΕΦΑΛΑΙΟ 4. ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΟΛΟΓΙΑΣ & ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 ΠΕΡΙΓΡΑΦΗ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ

Τα δεδομένα για την εκπαίδευση και την αξιολόγηση του μοντέλου συλλέχθηκαν μέσω του πιλότου RO2, με επικεφαλής τον FZI, στην Καρλσρούη της Γερμανίας στο πλαίσιο του Ευρωπαϊκού έργου Drive2theFuture (Orfanou et al., 2021). Η πραγματική πίστα δοκιμών περιλαμβάνει όλους τους σχετικούς τύπους δρόμων, π.χ. μεικτής κυκλοφορίας σε αστικές περιοχές, ζώνες ενδοαστικές 30 km/h και 50 km/h, δημοτικούς χώρους στάθμευσης, κατοικημένες περιοχές, κρατικούς δρόμους και αυτοκινητόδρομους. Στις δοκιμές Drive2theFuture, δύο Αυτοκίνητα Επιπέδου 3 χρησιμοποιήθηκαν για την αλληλεπίδραση με πεζούς, με διαφορετικά είδη συμπεριφορών - συμπεριφορά οχήματος ελεγχόμενου από τον άνθρωπο (HDV) και αυτοματοποιημένου οχήματος (AV).

Τα δεδομένα συλλέγονται αρχικά μέσω προσομοίωσης επαυξημένης και εικονικής πραγματικότητας, όπου το πραγματικό δίκτυο δοκιμών ενσωματώνεται σε μια πλατφόρμα προσομοίωσης, και στη συνέχεια από την πραγματική πίστα δοκιμής. Πρακτορικά μοντέλα οχημάτων και πεζών χρησιμοποιήθηκαν για τη δημιουργία δυναμικής συμπεριφοράς, καθώς και μοντέλα αισθητήρων και επικοινωνίας για την αναπαράσταση της ψηφιακής υποδομής (digital twin, Εικ.8), η οποία παρέχει το ίδιο αποτέλεσμα με την περιοχής δοκιμής του πραγματικού κόσμου.

Το σύνολο των δεδομένων για την εκπαίδευση του μοντέλου της παρούσας έρευνας προέρχεται από το προσομοιωμένο περιβάλλον (Φάση Ι). Στη Φάση Ι, δημιουργήθηκε ένα περιβάλλον VR βασισμένο στο Unreal Engine και τον προσομοιωτή CARLA, ο οποίος είναι ένας προσομοιωτής ανοιχτού κώδικα για έρευνα της αυτόνομης οδήγησης, που παρέχει ένα ρεαλιστικό και ευέλικτο περιβάλλον με ποικιλία ψηφιακών στοιχείων και αισθητήρων για την ανάπτυξη αλγορίθμων για AVs. Το περιβάλλον VR υποστηρίζει την εικονική πραγματικότητα και τη βύθιση ως «οδηγού» ενός οχήματος (είτε χειροκίνητη οδήγηση με πραγματικό τιμόνι, είτε σε αυτοματοποιημένη λειτουργία), όσο και ως «πεζού».

Εικόνα 8: Ψηφιακό δίδυμο της περιοχής δοκιμών (Πηγή: Drive2theFuture)

Οι τροχιές των οχημάτων και των πεζών συλλέχθηκαν από την προσομοίωση εικονικής πραγματικότητας. Τα δεδομένα περιλαμβάνουν θέσεις με χρονική αναφορά (frame_id, timestamp_ms), δισδιάστατες θέσεις (x,y) και ταχύτητες του οχήματος (vx, vy), διαφορετικούς παράγοντες παρακολούθησης (track_id, agent_type), τον προσανατολισμό του πράκτορα (psi_rad), το μήκος (length) και το πλάτος των πλαισίων οριοθέτησης (width) του οχήματος. Πιο συγκεκριμένα:

- **χρονικό πλαίσιο:** τα δεδομένα συλλέγονται κάθε 100 ms (=0,1 second).
- **είδος πράκτορα:** το πείραμα διερευνά την αλληλεπίδραση μεταξύ επιβατικών αυτοκινήτων και πεζών, επομένως η παράμετρος "agent-type" παίρνει τις τιμές: "car"/ όχημα ή "pedestrian" / πεζός.
- **track_id:** Το αναγνωριστικό **0** είναι πάντα το **όχημα** που μελετάται (είτε mode='automated' για την Φάση Ι, είτε mode='manually'), ενώ το αναγνωριστικό ίχνους **1** είναι πάντα ο **πεζός VR**.
- **x, y:** η θέση x και y του πράκτορα (m) .
- **vx, vy:** οι τιμές ταχύτητας του οχήματος στις διαστάσεις x και y (m/s).
- **psi_rad:** η γωνία εκτροπής του πράκτορα (rad). Υπολογίζεται στο παγκόσμιο σύστημα συντεταγμένων (SI). Η τιμή μηδέν σημαίνει ότι ο πράκτορας κατευθύνεται προς τα ανατολικά και οι τιμές αυξάνονται αριστερόστροφα με την περιστροφή του πράκτορα.
- **μήκος:** μήκος του πράκτορα (m) .
- **πλάτος:** πλάτος του πράκτορα (m).
- **ax, ay:** τιμές επιτάχυνσης/ επιβράδυνσης του AV στις διαστάσεις x και y (m/s^2).
- **time-headway:** η χρονική απόσταση του πράκτορα από το προηγούμενο όχημα (s).
- **κενό:** η χωρική απόσταση του πράκτορα από το προηγούμενο όχημα (m).

- **πλευρική θέση:** η απόσταση του κεντρικού άξονα του οχήματος από τον κεντρικό άξονα της λωρίδας (m).
- **πλευρική απόσταση:** η απόσταση του κεντρικό άξονα του πράκτορα από το πλευρικό αντικείμενο (m).
- **mode:** κατάσταση λειτουργίας, περιγράφει εάν η λειτουργία αυτοματισμού είναι ενεργοποιημένη ή εάν εφαρμόζεται χειροκίνητος έλεγχος του οχήματος. Δεδομένου ότι η αυτοματοποίηση μπορεί να ενεργοποιηθεί μόνο στην περίπτωση, που ο τύπος πράκτορα είναι το όχημα, αυτή η παράμετρος λαμβάνει τις τιμές "automated"/ αυτόνομο και "simulation"/ προσομοίωση, εάν η αυτόματη λειτουργία είναι ενεργοποιημένη και απενεργοποιημένη αντίστοιχα. Στα δεδομένα της Φάσης I, η κατάσταση λειτουργίας/ mode για το αυτοκίνητο είναι πάντα ίση με το «αυτόνομο». Για τους πεζούς, η μόνη τιμή για αυτήν την παράμετρο είναι η προσομοίωση/"simulation".

Με την χρήση ενός χάρτη υψηλής ευκρίνειας (HD) του οδικού δικτύου σε μορφή Lanelet, πρόσθετες τιμές υπολογίστηκαν για τις τροχιές, όπως οι δισδιάστατες επιταχύνσεις (ax, ay) του οχήματος, η χρονική απόσταση (time headway) των πρακτόρων, το κενό από το προπορευόμενο όχημα (gap) των πρακτόρων-AV και πεζού, η πλευρική θέση/μετατόπιση προς το κέντρο της λωρίδας (lateral position) και η απόσταση από τα πλευρικά αντικείμενα εκτός της λωρίδας (side distance). Τα στοιχεία αυτά παρουσιάζονται παρακάτω στους Πίνακες 1 και 2.

- Το χωρικό κενό s είναι μονοδιάστατο και υπολογίζεται χρησιμοποιώντας τον υποκείμενο χάρτη. Αντιστοιχεί στη συντομότερη διαδρομή από την τρέχουσα θέση έως τη θέση του πλησιέστερου οχήματος μπροστά και ορίζεται στο 100 ως μέγιστη τιμή, εάν δεν υπάρχει όχημα μπροστά.
- Η ταχύτητα v υπολογίζεται από τον τύπο: $v = sqrt (vx^2 + vy^2)$.
- Η επιτάχυνση α υπολογίζεται από τον τύπο: $\alpha = sqrt (\alpha x^2 + \alpha y^2)$.

Κατά την διάρκεια του πειράματος διερευνήθηκαν τρία σενάρια:

1. Οδηγώντας ευθεία, ενώ οι πεζοί μπορεί απροσδόκητα να διασχίσουν το δρόμο μπροστά από το όχημα.
2. Στροφή στα δεξιά, ενώ οι πεζοί μπορεί να διασχίσουν το δρόμο μετά τη στροφή.
3. Ελεύθερη οδήγηση, όπου ολόκληρος ο χάρτης χρησιμοποιείται για οδήγηση, ενώ το όχημα αλληλοεπιδρά με άλλους συμμετέχοντες στην κυκλοφορία.

Αυτόματη οδήγηση πραγματοποιήθηκε μόνο στα σενάρια 1 και 3, ενώ στην παρούσα διπλωματική εργασία αναλύθηκαν μόνο τα δεδομένα, που συλλέχθηκαν για το σενάριο της ευθείας οδήγησης, καθώς η ύπαρξη

ενός αυτόματου οχήματος και ενός πεζού διευκόλυνε την ενδελεχή κατανόηση της συμπεριφοράς των δύο χρηστών και των χαρακτηριστικών, που διέπουν τη συμπεριφορά τους κατά τη μεταξύ τους αλληλεπίδραση.

Πίνακας 1: Στοιχεία καταγεγραμμένα από τις τροχιές της προσομοίωσης

	A	B	C	D	E	F	G	H	I	J	K
1	track_id	frame_id	timestamp_ms	agent_type	x	y	vx	vy	psi_rad	length	width
2	0	1	0	car	170	-302.6	-0.5652	0.04028	3.071	4	1.8
3	0	3	100	car	169.9	-302.6	-0.5932	0.04037	3.074	4	1.8
4	0	5	200	car	169.9	-302.6	-0.6038	0.03969	3.075	4	1.8
5	0	7	300	car	169.8	-302.6	-0.6381	0.04354	3.077	4	1.8
6	0	9	400	car	169.8	-302.6	-0.6652	0.04226	3.079	4	1.8
7	0	11	500	car	169.7	-302.6	-0.6917	0.04265	3.081	4	1.8
8	0	13	600	car	169.6	-302.6	-0.7208	0.0387	3.084	4	1.8
9	0	15	700	car	169.6	-302.6	-0.7435	0.04246	3.085	4	1.8
10	0	17	800	car	169.5	-302.6	-0.7657	0.04518	3.086	4	1.8
11	0	19	900	car	169.4	-302.6	-0.8046	0.04454	3.09	4	1.8
12	0	21	1000	car	169.3	-302.6	-0.8242	0.04206	3.091	4	1.8
13	0	23	1100	car	169.2	-302.6	-0.8643	0.03938	3.095	4	1.8
14	0	25	1200	car	169.1	-302.6	-0.9066	0.04075	3.098	4	1.8
15	0	27	1300	car	169.1	-302.6	-0.9258	0.04076	3.099	4	1.8
16	0	29	1400	car	168.9	-302.6	-0.9721	0.03849	3.102	4	1.8
17	0	31	1500	car	168.9	-302.6	-0.9979	0.03588	3.104	4	1.8
18	0	33	1600	car	168.8	-302.6	-1.028	0.03605	3.106	4	1.8
19	0	35	1700	car	168.6	-302.6	-1.073	0.03678	3.109	4	1.8
20	0	37	1800	car	168.6	-302.5	-1.095	0.03397	3.11	4	1.8

L	M	N	O	P	Q	R
ax	ay	time_headway	gap	lateral_position	side_distance	mode
-0.2146	-0.008949	-1	100	0.08173	-1	automated
-0.1919	0.004119	-1	100	0.08162	-1	automated
-0.2395	-0.07401	-1	100	0.08163	-1	automated
-0.0382	0.04715	-1	100	0.08149	-1	automated
-0.2979	-0.004526	-1	100	0.0815	-1	automated
-0.2968	0.01074	-1	100	0.08139	-1	automated
-0.2603	-0.06106	-1	100	0.08125	-1	automated
-0.9294	0.08372	-1	100	0.08125	-1	automated
-0.4197	-0.02956	-1	100	0.08114	-1	automated
0.119	-0.09473	-1	100	0.081	-1	automated
0.2892	-0.1698	-1	100	0.08088	-1	automated
0.3125	-0.1169	-1	100	0.08075	-1	automated
-0.9408	0.1	-1	100	0.08064	-1	automated
-0.9135	0.03668	-1	100	0.08064	-1	automated
-0.4696	0.015	-1	100	0.08035	-1	automated
-0.9465	0.1118	-1	100	0.08035	-1	automated
0.3209	-0.2845	-1	100	0.08024	-1	automated
-0.3082	-0.03782	-1	100	0.07997	-1	automated
-0.2008	-0.02416	-1	100	0.01998	-1	automated
-1.016	-0.1567	-1	100	0.02025	-1	automated

Πίνακας 2: Στοιχεία υπολογίσιμα από τον χάρτη & τα δεδομένα προσομοίωσης

4.2 ΕΠΕΞΕΡΓΑΣΙΑ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ

4.2.1 Δεδομένα τροχιών

Τα δεδομένα τροχιών που εξετάστηκαν από το προσομοιωμένο σενάριο ήταν σε μορφή csv. Αρχικά, χρησιμοποιήθηκε ο αλγόριθμος **main_visualize_data** στην **Python** για την αναπαράσταση των σεναρίων σε χάρτες της μορφής Open Street Map (OSM) και επιλέχθηκαν τα χρονικά πλαίσια για την έρευνα αλληλεπίδρασης του αυτόνομου οχήματος με τον πεζό, ενώ το όχημα κινείται σε ευθεία. Η επιλογή των χρονικών πλαισίων έγινε με βάση την χωρική απόσταση, πιο συγκεκριμένα όταν η τιμή της ξεκινά να είναι μικρότερη του 100 και ο πεζός ξεκινάει να κινείται, ενώ παράλληλα έγινε επαλήθευση από την αναπαράσταση των τροχιών στην Python. Η διάρκεια της αλληλεπίδρασης οχήματος και πεζού θεωρήθηκε ίση με 7,4 sec. Η ταχύτητα των πεζών δεν καταγράφηκε κατά τη διάρκεια του πειράματος, και συνεπώς υπολογίστηκε στους άξονες x και y- vx_{ped}, vy_{ped} , ως το πηλίκο της απόστασης που έχει διανύσει ο πεζός

σε ένα χρονικό πλαίσιο διά το χρόνο (τα δεδομένα συλλέχθηκαν ανά 100 ms= 0.1 s), άρα: $vx_{ped}= \Delta x/0,1$ sec και $vy_{ped}= \Delta y/0,1$ sec.

4.2.2 Καταστάσεις

Τα δεδομένα που συλλέχθηκαν από τα εικονικά πειράματα χρησιμοποιήθηκαν για την εξαγωγή και την εκτίμηση των μεταβλητών για τον καθορισμό του χώρου καταστάσεων S, σε συνδυασμό με την διαθέσιμη βιβλιογραφία (Jayaraman et al., 2020). Τρία χαρακτηριστικά επιλέχθηκαν για τον καθορισμό των καταστάσεων αλληλεπίδρασης μεταξύ του οχήματος και του πεζού: η ταχύτητα του οχήματος, η διαφορά ταχύτητας μεταξύ οχήματος και πεζού, και το χωρικό χάσμα μεταξύ των δύο χρηστών του δρόμου. Κάθε χαρακτηριστικό χωρίστηκε σε διάφορα επίπεδα με βάση τα αποτελέσματα ομαδοποίησης k-means. Πιο συγκεκριμένα, χρησιμοποιήθηκαν η elbow method και η silhouette coefficient προκειμένου να καθοριστεί ο καταλληλότερος αριθμός συστάδων των χαρακτηριστικών και τα όρια κάθε συστάδας. Οι μέθοδοι έδειξαν ότι τα χαρακτηριστικά "ταχύτητα οχήματος" και "διαφορά ταχύτητας" θα πρέπει να διακριθούν σε δύο επίπεδα το καθένα, ενώ το χαρακτηριστικό "χωρική απόσταση" σε τρία επίπεδα. Το silhouette score δείχνει πόσο καλά τα δεδομένα συγκεντρώνονται σε διακριτές ομάδες και κυμαίνεται από -1 μέχρι 1, όπου μια υψηλή βαθμολογία σιλουέτας δείχνει ότι το αντικείμενο είναι καλά προσαρμοσμένο στη δική του συστάδα. Η βαθμολογία σιλουέτας για την ταχύτητα οχήματος υπολογίστηκε στο 0.78 για τις 2 κλάσεις, για την διαφορά ταχύτητας στο 0.715 για τις 2 κλάσεις και για την χωρική απόσταση στο 0.62 για τις 3 κλάσεις. Έπειτα, έγινε ο χωρισμός των χαρακτηριστικών στις κατάλληλες κλάσεις. Τα αποτελέσματα, που φαίνονται στον παρακάτω Πίνακα 3, οδηγούν σε 12 καταστάσεις.

Πίνακας 3: Καθορισμός Καταστάσεων

Επίπεδα	Ταχύτητα Οχήματος	Διαφορά Ταχύτητας	Χωρική Απόσταση
1	(0.003, 10.50)	(-11.38, 7.31)	(4.27, 11.84)
2	[10.50, 25.38)	[7.31, 25.19)	[11.84, 19.50)
3	-	-	[19.50, 33.09)

4.2.3 Ενέργειες

Για τον χώρο ενεργειών Α, θεωρούμε την επιτάχυνση ως την κρίσιμη τιμή για τον καθορισμό του τρόπου, με τον οποίο ένας οδηγός- αυτόνομο όχημα θα αντιδράσει σε εξωτερικά ερεθίσματα, όπως είναι η αλληλεπίδραση με τον πεζό. Βάση της ανάλυσης K-means που έγινε στις ενέργειες, για την επιτάχυνση ο βέλτιστος αριθμός κλάσεων είναι ίσος με 2 και η τιμή 1,79 m/s^2 προέκυψε ως το ανώτατο όριο, ώστε να ληφθεί υπόψη ότι το όχημα επιταχύνει ομαλά. Για την επιβράδυνση, οι τιμές χωρίστηκαν σε τρεις κατηγορίες, με βάση τη μέθοδο K-means, με το 1,627 m/s^2 ως ανώτατο όριο για την ομαλή επιβράδυνση, και το 4,32 m/s^2 ως ανώτατο όριο για την μέση επιβράδυνση. Για τιμές μεγαλύτερες από 4,32 m/s^2 το όχημα θεωρείται ότι πραγματοποιεί απότομη επιβράδυνση. Τα ανώτατα όρια για την ομαλή επιτάχυνση και επιβράδυνση συμφωνούν με την βιβλιογραφία, καθώς βρέθηκαν να είναι περίπου 0,16g – 0,36g ($g=9,81 m/s^2$), όπως περιγράφεται στους Vlahogianni & Barbounakis (2017). Οι τιμές του silhouette score υπολογίστηκαν ίσες με 0.68 για την επιτάχυνση με τις 2 κλάσεις και 0.61 για την επιβράδυνση με τις 3 κλάσεις. Επιπλέον, υπάρχει η πιθανότητα ο οδηγός να μην προβεί σε καμία ενέργεια παραμένοντας στην τρέχουσα κατάσταση, το οποίο δεν το λαμβάνουμε υπόψη. Με βάση τα παραπάνω, διακρίνονται 5 ενέργειες για το όχημα, οι οποίες φαίνονται στον Πίνακα 4.

Πίνακας 4: Καθορισμός Ενεργειών

Δράσεις	Τιμές (m/s^2)
Περιπλάνηση	-
Ομαλή επιτάχυνση	(0, 1.79]
Απότομη επιτάχυνση	(1.79, 6.10]
Ομαλή επιβράδυνση	(-1.627, 0]
Μέση επιβράδυνση	(-4.32, -1.627]
Απότομη επιβράδυνση	[-9,-4.32]

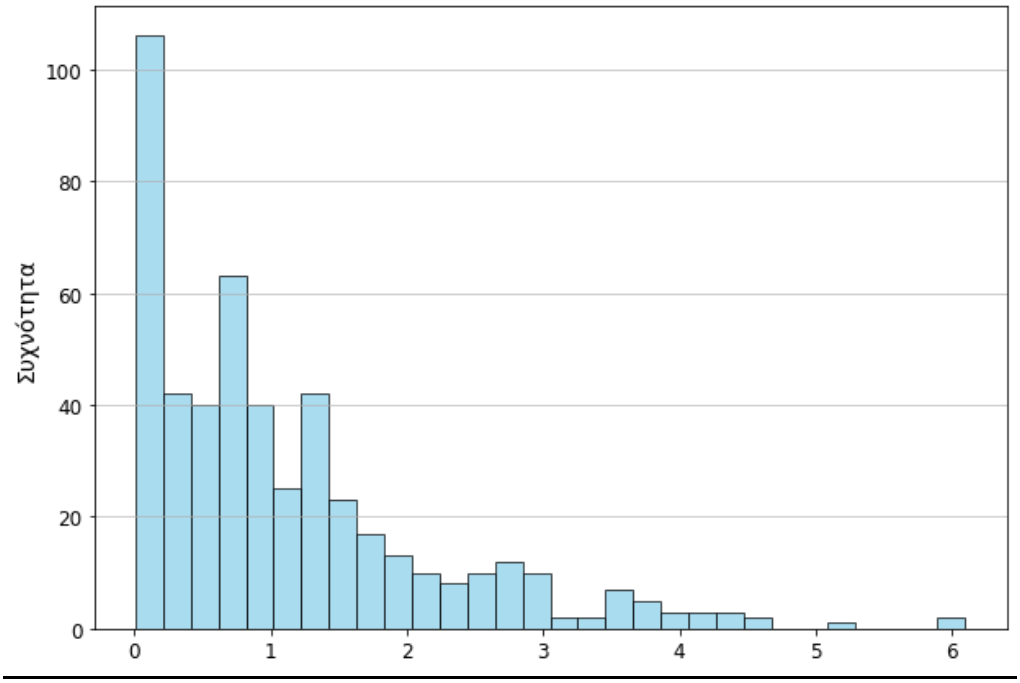
4.2.4 Στατιστική Ανάλυση Χαρακτηριστικών Καταστάσεων & Ενεργειών

Η ταχύτητα του οχήματος, η διαφορά ταχύτητας και το χωρικό κενό αποτελούν τα τρία χαρακτηριστικά που χρησιμοποιήθηκαν για την περιγραφή της κατάστασης του οχήματος, ενώ η δράση του οδηγού ορίστηκε από την επιτάχυνση ή την επιβράδυνση του οχήματος. Στον Πίνακα 5 φαίνονται τα περιγραφικά στατιστικά στοιχεία των πέντε παραμέτρων που αναφέρθηκαν παραπάνω. Από την στατιστική ανάλυση εξαιρέθηκαν οι επιβραδύνσεις με τιμές μεγαλύτερες των 9 m/s^2 (πέδηση έκτακτης ανάγκης). Στη συνέχεια, παρουσιάζονται τα ιστογράμματα των χαρακτηριστικών των καταστάσεων και των ενεργειών, ξεχωριστά για κάθε στοιχείο, για την καλύτερη απεικόνιση της κατανομής των τιμών τους. Η στατιστική ανάλυση και απεικόνιση των κινηματικών χαρακτηριστικών κυκλοφορίας έγινε στο περιβάλλον της **Python** με χρήση των λειτουργιών της βιβλιοθήκης Pandas, Matplotlib και Sklearn.

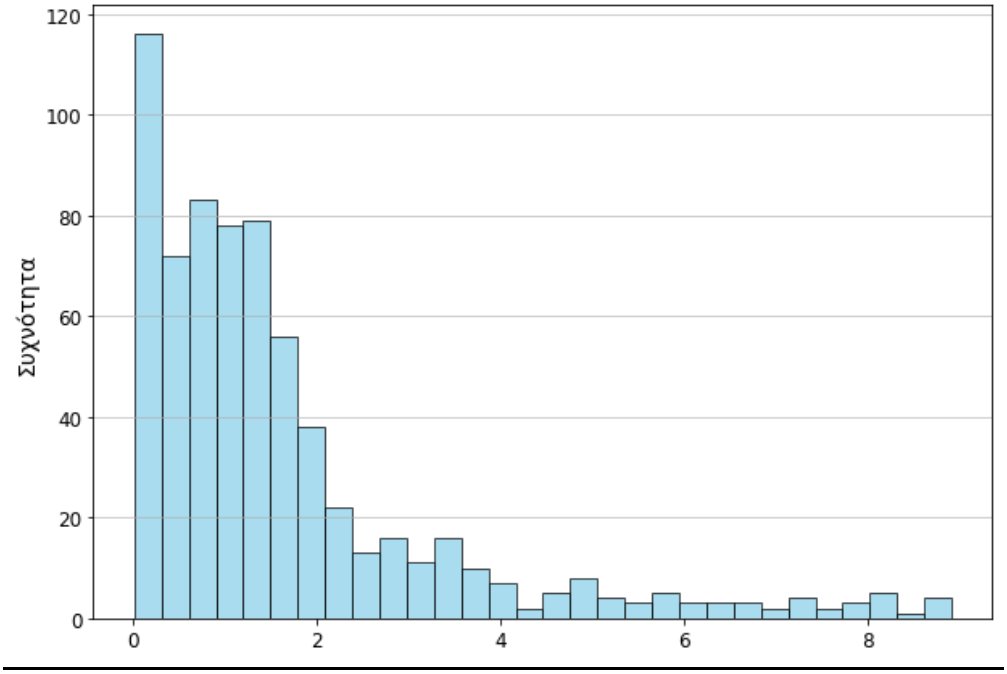
Πίνακας 5: Περιγραφική Στατιστική Κινηματικών Χαρακτηριστικών Κυκλοφορίας

	Επιτάχυνση	Επιβράδυνση	Ταχύτητα Οχήματος	Διαφορά Ταχύτητας	Χωρικό Χάσμα
	(m/s2)	(m/s2)	(km/h)	(km/h)	(m)
Μέσος όρος	1.112	1.643	7.285	4.009	13.114
Τυπική απόκλιση	1.070	1.741	8.671	9.174	5.764
Διάμεσος	0.792	1.152	2.040	0.348	11.736
Ελάχιστη τιμή	0.005	0.019	0.003	-11.377	4.272
25%	0.299	0.535	0.130	-3.472	8.889
50%	0.792	1.152	2.040	0.348	11.736
75%	1.497	1.926	14.805	11.197	16.026
Μέγιστη τιμή	6.096	8.916	25.373	25.189	33.086

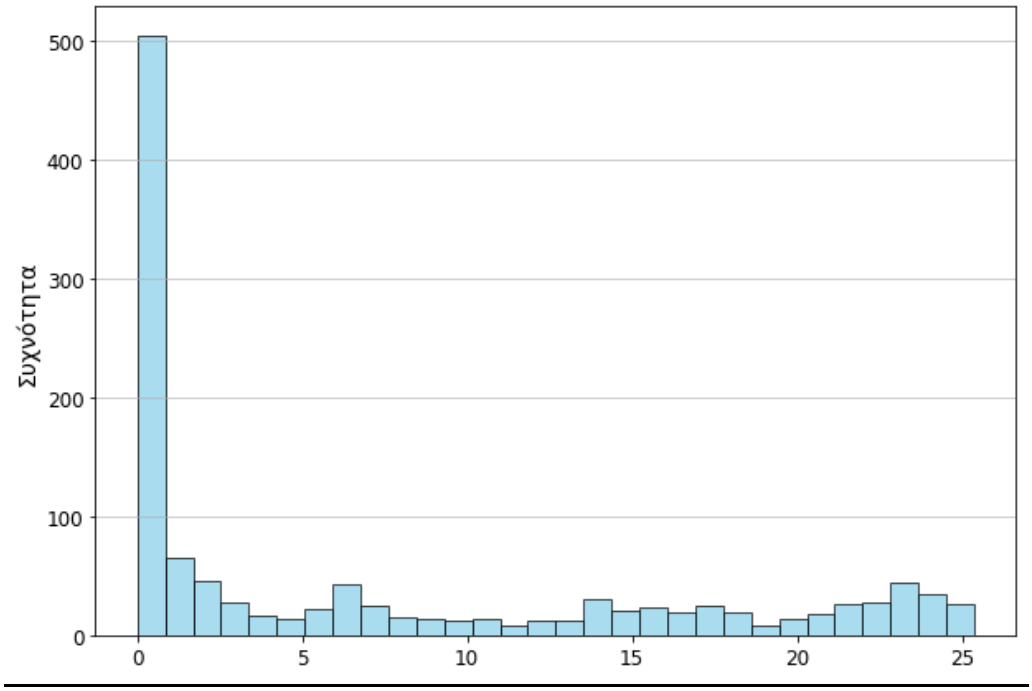
Στα παρακάτω Διαγράμματα (Διαγράμματα 1-5) παρουσιάζονται οι στατιστικές κατανομές των τιμών των παραπάνω κινηματικών χαρακτηριστικών, που χρησιμοποιήθηκαν για τον προσδιορισμό των καταστάσεων και των ενεργειών του αυτόνομου οχήματος.



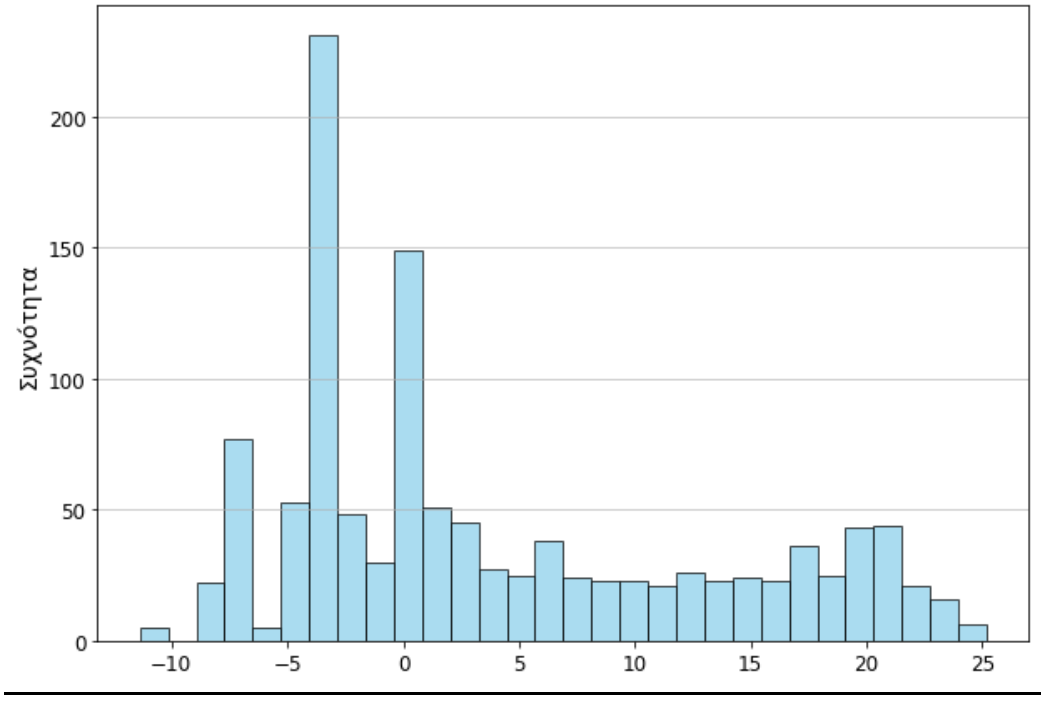
Διάγραμμα 1: Κατανομή των τιμών της επιτάχυνσης του οχήματος



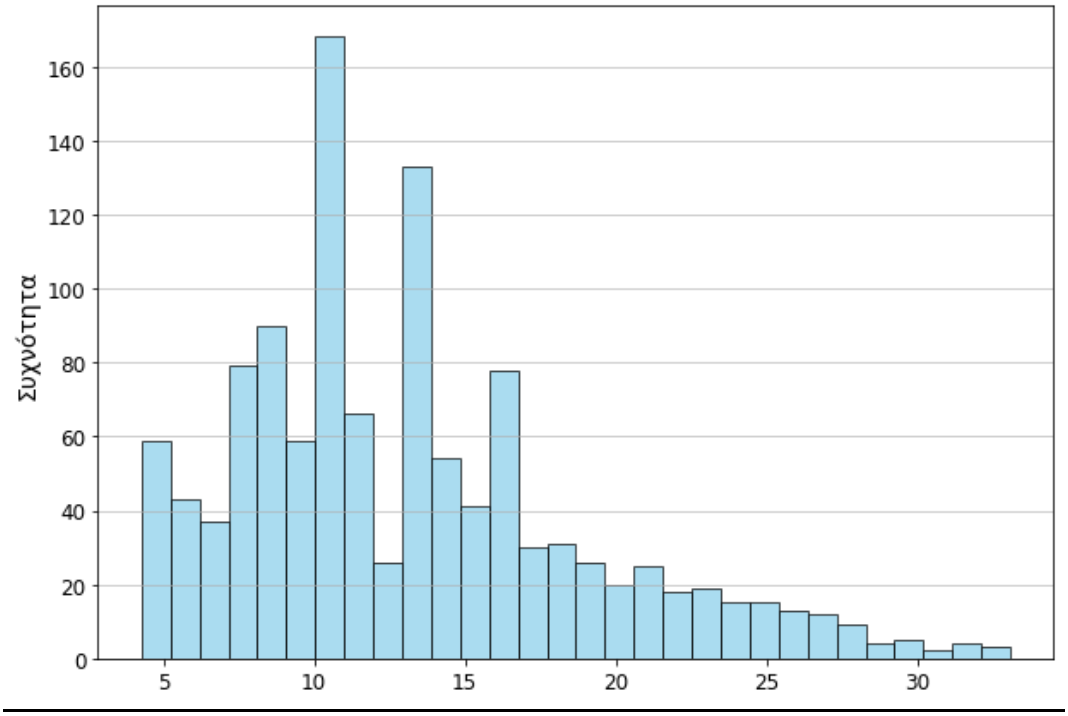
Διάγραμμα 2: Κατανομή των τιμών της επιβράδυνσης του οχήματος



Διάγραμμα 3: Κατανομή των τιμών της ταχύτητας του οχήματος



Διάγραμμα 4: Κατανομή των τιμών της διαφοράς ταχυτήτων οχήματος-πεζού



Διάγραμμα 5: Κατανομή των τιμών της χωρικής απόστασης οχήματος-πεζού

4.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΛΓΟΡΙΘΜΟΥ MAX-ENT IRL

Η επεξεργασία των δεδομένων από τις τροχιές των οχημάτων έγινε στο περιβάλλον προγραμματισμού της **Python**. Χρησιμοποιήθηκαν οι λειτουργίες των βιβλιοθηκών NumPy, Pandas και Matplotlib, καθώς επίσης και η συνάρτηση αποθήκευσης από το NumPy.

Με βάση των ορισμό των καταστάσεων και των ενεργειών του οχήματος, όπως περιγράφηκε και παρουσιάστηκε στις προηγούμενες ενότητες, οι τροχιές του αυτόνομου οχήματος μετατράπηκαν σε τροχιές αλληλουχίας ζεύγους (κατάστασης, ενέργειας).

Στη συνέχεια, εκτελούνται οι επαναληπτικοί βρόγχοι για τον υπολογισμό των πιθανοτήτων μετάβασης από το σύνολο των δεδομένων τροχιάς. Αυτοί οι βρόγχοι μετρούν πόσες φορές συμβαίνουν συγκεκριμένες μεταβάσεις από μια κατάσταση (ορίζεται από τα trajectories[i, j, 0] και τα trajectories[i, j, 1]) σε μια άλλη κατάσταση (trajectories[i, j+1, 0]). Έπειτα, υπολογίζεται η πιθανότητα μετάβασης (transition_probability) διαιρώντας τον αριθμό των μεταβάσεων με το συνολικό αριθμό μεταβάσεων από την ίδια κατάσταση. Ο πίνακας μετάβασης πιθανότητας αποτυπώνει την πιθανότητα μετάβασης από την κατάσταση s_i στην s_k υπό την ενέργεια α με διαστάσεις (12, 5, 12), όπου το 12 αντιπροσωπεύει τον αριθμό των καταστάσεων και το 5 το πλήθος των ενεργειών. Τέλος, δημιουργείται ένας πίνακας χαρακτηριστικών (feature_matrix) που

καθορίζει τα χαρακτηριστικά, που προσδιορίζουν την κατάσταση του οχήματος (στην παρούσα έρευνα τα χαρακτηριστικά είναι: η ταχύτητα του οχήματος, η διαφορά ταχύτητας των δύο χρηστών και η χωρική τους απόσταση).

Στη συνέχεια εφαρμόστηκε ο αλγόριθμος Αντίστροφης Ενισχυτικής Μάθησης Μέγιστης Εντροπίας (Max-Ent IRL) του Ziebart et al., 2008. Το Max-Ent IRL είναι μια τεχνική, που χρησιμοποιείται για την εκτίμηση της συνάρτησης ανταμοιβής με την μεγιστοποίηση της εντροπίας της κατανομής της πολιτικής, συλλαμβάνοντας την υποκείμενη αβεβαιότητα και ενθαρρύνοντας την εξερεύνηση. Στόχος είναι η εύρεση της συνάρτησης ανταμοιβής που ερμηνεύει καλύτερα την παρατηρούμενη συμπεριφορά, διατηρώντας παράλληλα μια ποικιλόμορφη κατανομή πολιτικής. Οι παράμετροι που εισήχθησαν στον αλγόριθμο περιλαμβάνουν: τον πίνακα χαρακτηριστικών (feature_matrix), τον αριθμό των ενεργειών, τον συντελεστή έκπτωσης, τις πιθανότητες μετάβασης, τις παρατηρούμενες τροχιές, τον αριθμό εποχών και τον ρυθμό εκμάθησης του αλγορίθμου. Ο αλγόριθμος ενημερώνει επαναλαμβανόμενα ένα διάνυσμα ανταμοιβής, που αντιπροσωπεύει τα βάρη επιβράβευσης (reward weights), χρησιμοποιώντας gradient descent, με στόχο να ελαχιστοποιήσει τη διαφορά μεταξύ της παρατηρούμενης συμπεριφοράς και της συμπεριφοράς που προβλέπεται από την τρέχουσα συνάρτηση ανταμοιβής. Το gradient descent είναι ένας επαναληπτικός αλγόριθμος βελτιστοποίησης, που χρησιμοποιείται για την ελαχιστοποίηση μιας συνάρτησης προσαρμόζοντας τις παραμέτρους της προς την κατεύθυνση της πιο απότομης καθόδου κλίσης, βρίσκοντας αποτελεσματικά τις βέλτιστες τιμές των παραμέτρων. Έπειτα, δίνει τα τελικά βάρη επιβράβευσης για κάθε χαρακτηριστικό που ορίστηκε, καθώς και τις τελικές τιμές ανταμοιβής για την κάθε κατάσταση και ένα διάγραμμα με τις ανακτώμενες ανταμοιβές.

Για την έρευση της συνάρτησης ανταμοιβής, ο αλγόριθμος υπολογίζει τις προσδοκίες των χαρακτηριστικών και τις αναμενόμενες συχνότητες επίσκεψης καταστάσεων, με βάση τις παρατηρούμενες τροχιές αλληλουχίας (κατάσταση, ενέργεια, κατάσταση). Οι προσδοκίες των χαρακτηριστικών αντιπροσωπεύουν τις μέσες τιμές των χαρακτηριστικών κατάστασης σε πολλαπλές τροχιές, παρέχοντας πληροφορίες για τα τυπικά χαρακτηριστικά κατάστασης που παρατηρούνται κατά τις αλληλεπιδράσεις του πράκτορα με το περιβάλλον. Η συχνότητα επίσκεψης κατάστασης εκφράζει το πόσο συχνά γίνεται επίσκεψη σε κάθε κατάσταση για την παρατηρούμενη συμπεριφορά.

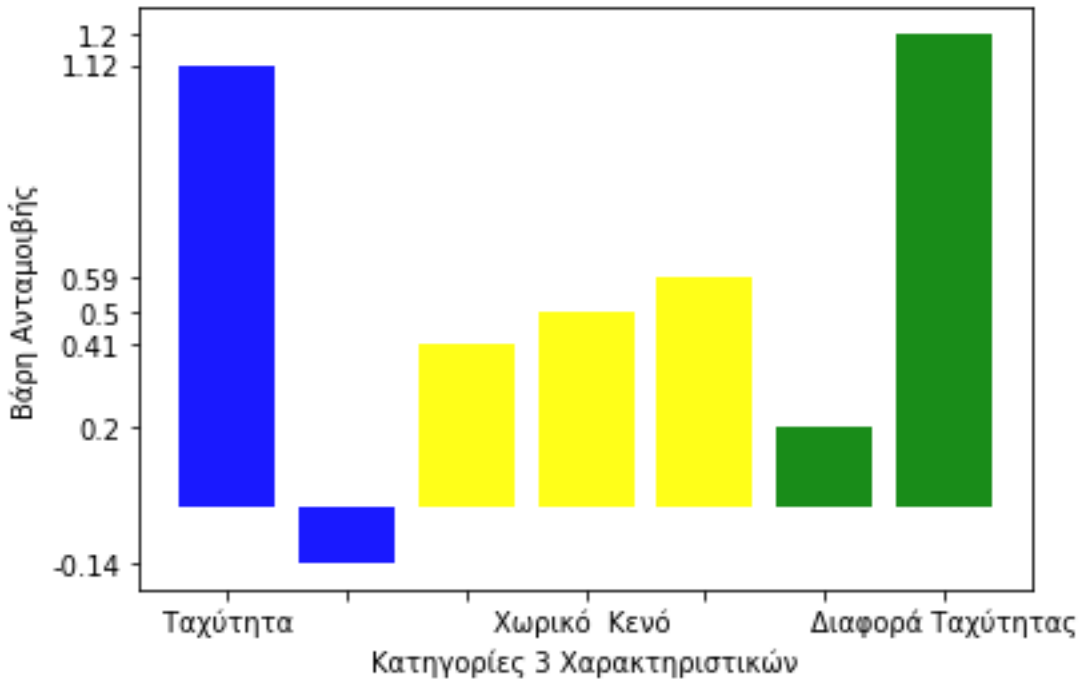
Τέλος, στόχος του αλγορίθμου είναι η έρευση της βέλτιστης πολιτικής, που περιγράφεται καλύτερα από τη συνάρτηση ανταμοιβής με βάση τα δεδομένα των τροχιών του οχήματος και της συνάρτηση βέλτιστης αξίας, η οποία αντιπροσωπεύει την αναμενόμενη αθροιστική ανταμοιβή, που μπορεί να επιτύχει ένας πράκτορας για την κάθε κατάσταση.

Ο προσδιορισμός της βέλτιστης πολιτικής γίνεται με βάση των αριθμό καταστάσεων, τον αριθμό ενεργειών, τις πιθανότητες μετάβασης, τις ανταμοιβές, τον συντελεστή έκπτωσης, ενός ορίου σύγκλισης και της συνάρτηση αξίας. Η συνάρτηση αξίας, ανάλογα με την καθορισμένη στοχαστικότητα, στην παρούσα εργασία χρησιμοποιείται μία στοχαστική πολιτική, επιστρέφει έναν πίνακα πιθανοτήτων ενέργειας για κάθε κατάσταση, ενώ για μία ντετερμινιστική πολιτική επιστρέφει ένα πίνακα του βέλτιστου δείκτη δράσης για κάθε κατάσταση. Η στοχαστική πολιτική προκύπτει από την αξιολόγηση των αναμενόμενων τιμών των ενεργειών χρησιμοποιώντας τις πιθανότητες μετάβασης, τις ανταμοιβές και την υπολογισμένη συνάρτηση αξίας.

Εφαρμόζοντας τον **αλγόριθμο Max-Ent IRL**, το σενάριο στοχεύει να μάθει μια συνάρτηση ανταμοιβής που καταγράφει τους υποκείμενους παράγοντες, που οδηγούν στην παρατηρούμενη συμπεριφορά. Αυτό μπορεί να είναι πολύτιμο σε διάφορες εφαρμογές, όπως στην αλληλεπίδραση των αυτόνομων οχημάτων με πεζούς, όπου η κατανόηση των ανταμοιβών που σχετίζονται με διαφορετικές συμπεριφορές μπορεί να βελτιώσει τη λήψη αποφάσεων και την πρόβλεψη συμπεριφοράς, και επομένως να αυξήσει την ασφάλεια.

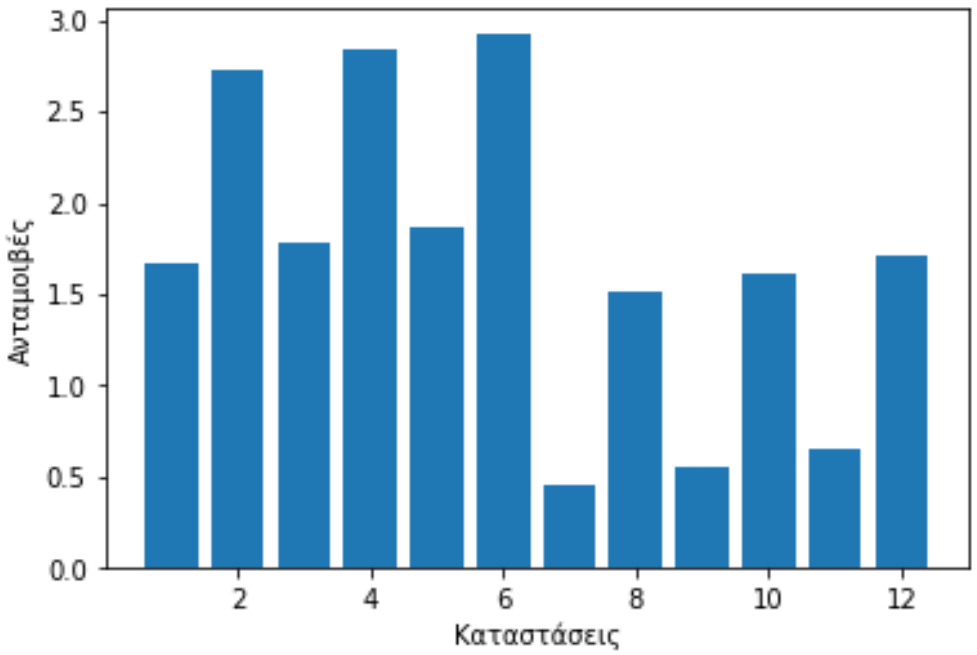
Για την εκπαίδευση και την προσπάθεια σύγκλισης του αλγορίθμου έγινε fine-tuning των υπερπαραμέτρων της συνάρτησης irl και πιο συγκεκριμένα της έκπτωσης, των εποχών και του ρυθμού εκμάθησης. Στόχος ήταν η σύγκλιση των βαρών επιβράβευσης, η οποία δείχνει ότι η εκπαίδευση του αλγορίθμου είναι ικανοποιητική. Λόγω του περιορισμένου αριθμού τροχιών χρησιμοποιήθηκαν και οι 16 τροχιές για την εκπαίδευση του μοντέλου. Τελικά, επιλέχθηκαν οι τιμές discount = 0,9 , epochs = 35 και learning_rate = 0,02 για τις υπερπαραμέτρους. Τα αποτελέσματα των βαρών επιβράβευσης παρουσιάζονται παρακάτω.(Διαγρ.6)

Στον συγκεκριμένο αλγόριθμο χρησιμοποιήθηκε επιπλέον ένας τυχαίος σπόρος (random.seed), ο οποίος διασφαλίζει την αναπαραγωγιμότητα των αποτελεσμάτων του αλγορίθμου ελέγχοντας τη δημιουργία τυχαίων αριθμών. Ορίζοντας ένα συγκεκριμένο τυχαίο σπόρο, ο αλγόριθμος παράγει την ίδια ακολουθία τυχαίων τιμών κάθε φορά που εκτελείται. Ο τυχαίος σπόρος προστέθηκε μετά την ρύθμιση των παραμέτρων της συνάρτησης. Έγιναν δοκιμές για την έρευση του κατάλληλου τυχαίου σπόρου, και τελικά επιλέχθηκε ο **np.random.seed(123)**.



Διάγραμμα 6: Βάρη ανταμοιβής για τα διαφορετικά επίπεδα χαρακτηριστικών

Σύμφωνα με τα αποτελέσματα που φαίνονται στο Διάγραμμα 6, ο αλγόριθμος αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας παρέχει τα βάρη της συνάρτησης ανταμοιβής για τα διαφορετικά χαρακτηριστικά που χρησιμοποιούνται στον ορισμό καταστάσεων, δηλαδή την ταχύτητα οχήματος, το χωρικό χάσμα και την διαφορά ταχύτητας οχήματος-πεζού. Από το γράφημα φαίνεται ότι οι καταστάσεις με χαμηλότερες ταχύτητες οχήματος (επίπεδο 1) λαμβάνουν σημαντικά υψηλότερες ανταμοιβές σε σύγκριση με τις καταστάσεις με υψηλότερες ταχύτητες που λαμβάνουν αρνητικές ανταμοιβές (επίπεδο 2). Όσον αφορά το χωρικό κενό, οι υψηλότερες τιμές ανταμοιβής παρατηρούνται για το επίπεδο 3, υποδεικνύοντας ότι προτιμώνται μεγαλύτερες αποστάσεις μεταξύ του πεζού και του οχήματος που πλησιάζει. Παράλληλα, οι καταστάσεις με πολύ χαμηλά χωρικά κενά (επίπεδο 1) έχουν το χαμηλότερο βάρος συνάρτησης ανταμοιβής, γεγονός που υποδηλώνει ότι οι οδηγοί αποφεύγουν να διατηρούν εξαιρετικά κοντινές αποστάσεις με τους πεζούς. Τέλος, όσον αφορά τη διαφορά ταχύτητας, οι υψηλότερες τιμές ανταμοιβής συνδέονται με τις μέτριες με υψηλές διαφορές ταχύτητας (επίπεδο 2) και όχι με τις χαμηλότερες (επίπεδο 1), υποδηλώνοντας την προτίμηση των οδηγών να έχουν μεγαλύτερη ταχύτητα από τους πεζούς.



Διάγραμμα 7: Ανακτώμενες ανταμοιβές καταστάσεων

Στο Διάγραμμα 7 φαίνονται οι ανακτώμενες ανταμοιβές για τις 12 καταστάσεις, οι οποίες ορίστηκαν για τον αλγόριθμο. Οι πρώτες 6 καταστάσεις αντιπροσωπεύουν τις τροχιές με χαμηλές ταχύτητες οχήματος (επίπεδο 1) και είναι φανερό ότι οι ανταμοιβές τους είναι σημαντικά μεγαλύτερες σε σχέση με τις υπόλοιπες 6 καταστάσεις.

Επιπλέον, οι ζυγές καταστάσεις 2, 4, 6, 8, 10 αντιστοιχούν στις καταστάσεις με μέτριες και μεγαλύτερες διαφορές ταχύτητας οχήματος και πεζού (επίπεδο 2) και είναι σημαντικά υψηλότερες από τις μονές καταστάσεις που αντιστοιχούν σε μικρές διαφορές ταχύτητας (επίπεδο 1).

Στη συνέχεια, οι καταστάσεις 5, 6 και 11, 12 αντιστοιχούν στο υψηλότερο χωρικό χάσμα (επίπεδο 3), ενώ οι 3, 4 και 9, 10 σε μέτριο χωρικό χάσμα (επίπεδο 2) και οι υπόλοιπες, (1, 2 και 7, 8) σε μικρό χωρικό κενό (επίπεδο 1) και φαίνεται η σταδιακή μείωση της ανταμοιβής με τη μείωση του αντίστοιχου χωρικού κενού.

Οι τρεις καταστάσεις με τις μεγαλύτερες ανταμοιβές είναι οι 2,4 και 6, ενώ οι τρεις καταστάσεις με τις μικρότερες ανταμοιβές είναι οι 7,9 και 11. Όλες οι καταστάσεις αντιστοιχούν σε μικρά (2, 7), μέτρια (4, 9) και μεγάλα χωρικά κενά (6, 11). Οι κύριες διαφορές μεταξύ των καταστάσεων με τις υψηλότερες και χαμηλότερες ανταμοιβές είναι ότι: οι πρώτες τρεις (2, 4, 6) χαρακτηρίζονται από χαμηλότερες ταχύτητες οχήματος σε σχέση με τις δεύτερες (7, 9, 11), και ότι η διαφορά ταχύτητας είναι μεγαλύτερη για τις πρώτες (κλάση 2), ενώ είναι μικρότερη ή αρνητική για τις δεύτερες (κλάση 1).

Ακολουθούν κάποια παραδείγματα καταστάσεων με ένα διαφορετικό χαρακτηριστικό για την καλύτερη σύγκριση και κατανόηση. Η ανταμοιβή της κατάστασης 1 είναι 1.67, ενώ της κατάστασης 2 είναι 2.73. Και οι δύο καταστάσεις χαρακτηρίζονται από χαμηλή ταχύτητα οχήματος και μικρό χωρικό κενό, ενώ η μόνη διαφορά τους είναι ότι η κατάσταση 1 περιγράφεται από διαφορά ταχύτητας κλάσης 1, ενώ η κατάσταση 2 από κλάσης 2. Ο αλγόριθμος δείχνει προτίμηση στην υψηλή διαφορά ταχύτητας και επιβραβεύει τον οδηγό, όταν την διατηρεί.

Η ανταμοιβή της κατάστασης 1 είναι 1.67 και της κατάστασης 7 είναι 0.46. Τα κοινά χαρακτηριστικά τους περιλαμβάνουν την διαφορά ταχύτητας κλάσης 1 και το μικρό χωρικό κενό, όμως η κατάσταση 1 χαρακτηρίζεται από χαμηλή ταχύτητα σε αντίθεση με την κατάσταση 7. Ο οδηγός στην κατάσταση 1 επιβραβεύεται για την χαμηλή ταχύτητα σε σχέση με την υψηλή ταχύτητα στην 7.

Τέλος, η κατάσταση 1 έχει ανταμοιβή 1.67, η κατάσταση 3 1.78 και η κατάσταση 5 1.86. Οι τρεις αυτές καταστάσεις χαρακτηρίζονται από χαμηλή ταχύτητα και διαφορά ταχύτητας κλάσης 1, ενώ η 1 περιγράφεται από χαμηλό χωρικό κενό, η 3 από μεσαίο και η 5 από υψηλό χωρικό κενό, το οποίο φαίνεται αντίστοιχα να επηρεάζει την επιβράβευση αυξητικά, όπως αυξάνεται το χωρικό κενό. Ο αλγόριθμος επιβραβεύει το μεσαίο, και περισσότερο το μεγαλύτερο χωρικό κενό σε σχέση με το μικρότερο.

Από τα παραπάνω, φαίνεται πόσο καθοριστική είναι η μεγαλύτερη διαφορά ταχύτητας του οχήματος σε σχέση με τον πεζό για την απόκτηση μεγαλύτερης επιβράβευσης από τον αλγόριθμο (επίπεδο 2). Στο επίπεδο 1 της διαφοράς ταχύτητας περιλαμβάνονται μικρότερες διαφορές ταχύτητας, καθώς και αρνητικές τιμές, που σημαίνει ότι η ταχύτητα του πεζού είναι μεγαλύτερη από αυτή του οχήματος, με βάση τον ορισμό της διαφοράς ταχύτητας. Από τις ανταμοιβές φαίνεται ότι οι οδηγοί προτιμούν να έχουν μεγαλύτερη ταχύτητα από τον πεζό, και ενώ επιβραδύνουν κοντά στον πεζό, δεν προτιμούν να εφαρμόζουν μέγιστη επιβράδυνση, αλλά να διατηρούν μια λογική ταχύτητα μεγαλύτερη από αυτή του πεζού. Αυτό μπορεί να ερμηνευτεί, διότι οι οδηγοί προσπαθούν να διατηρήσουν σταθερή πορεία και μικρότερη κατανάλωση καυσίμου, και εφόσον ο αλγόριθμος μπορεί να εντοπίσει τον πεζό έγκαιρα, το όχημα δεν χρειάζεται να επιβραδύνει σημαντικά την τελευταία στιγμή, αλλά προγραμματίζει καλύτερα την πορεία του.

Έπειτα, φαίνεται ότι ακολουθεί η σημαντικότητα της ταχύτητας του οχήματος, δείχνοντας την προτίμηση για χαμηλές με μέτριες ταχύτητες οχήματος με μεγαλύτερη ανταμοιβή (επίπεδο 1). Τέλος, φαίνεται ότι όσο αυξάνεται το χωρικό χάσμα, ανά 2 καταστάσεις, υπάρχει μια αναλογική αύξηση της ανταμοιβής με προτίμηση για το μεγαλύτερο χωρικό χάσμα πεζού και AV (επίπεδο 3).

Τα αποτελέσματα που παρουσιάστηκαν παραπάνω δείχνουν ότι ο αλγόριθμος έχει εκπαιδευτεί αποτελεσματικά, διότι τα βάρη ανταμοιβής και οι ανακτώμενες ανταμοιβές αντικατοπτρίζουν τις

ανθρώπινες αξίες των οδηγών, αλλά και των πεζών. Οι οδηγοί προτιμούν να έχουν μεγαλύτερη ταχύτητα κατά την οδήγηση σε σχέση με τους πεζούς, αλλά επίσης αποφεύγουν να φρενάρουν απότομα κοντά σε πεζό, ενώ προτιμούν να επιβραδύνουν και να διατηρήσουν μια λογική ταχύτητα. Οι πεζοί, ακολούθως, προτιμούν να έχουν μεγαλύτερη απόσταση από το όχημα, όταν πρόκειται να διασχίσουν τον δρόμο, καθώς και το όχημα να έχει μικρότερη ταχύτητα, ώστε να νιώσουν μεγαλύτερη ασφάλεια στο δρόμο.

ΚΕΦΑΛΑΙΟ 5. ΣΥΜΠΕΡΑΣΜΑΤΑ & ΠΡΟΤΑΣΕΙΣ

5.1 ΕΙΣΑΓΩΓΗ

Η αύξηση της οδικής ασφάλειας και η μείωση των κινδύνων για τους πιο ευάλωτους χρήστες του δρόμου-τους πεζούς, αποτελούν το βασικό αντικείμενο της συγκεκριμένης Διπλωματικής εργασίας. Τα περισσότερα οδικά ατυχήματα οφείλονται σε ανθρώπινο σφάλμα, επομένως με την εισαγωγή των αυτόνομων οχημάτων στην κυκλοφορία θα αυξηθεί σημαντικά η ασφάλεια στον δρόμο. Για την ομαλή εισαγωγή τους στην κυκλοφορία και την αποδοχή τους από τους χρήστες του δρόμου τα αυτόνομα οχήματα πρέπει να συμπεριφέρονται με τρόπο, που να αντικατοπτρίζει τις ανθρώπινες συμπεριφορές και αξίες.

Η ανάπτυξη μίας αποτελεσματικής στρατηγικής οδήγησης για τα διασυνδεδεμένα αυτόνομα οχήματα, όταν αλληλοεπιδρούν με τους πεζούς είναι το επόμενο λογικό βήμα. Αυτό επιτυγχάνεται με την χρήση της αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας, με την οποία το μοντέλο εξάγει την συνάρτηση ανταμοιβής από τις παρατηρούμενες τροχιές των οχημάτων, οδηγώντας σε πιο ρεαλιστικές προβλέψεις. Τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση προέρχονται από ένα πείραμα εικονικής πραγματικότητας που πραγματοποιήθηκε στην Καρλσρούη της Γερμανίας, κατά το οποίο καταγράφηκαν οι τροχιές των οχημάτων και των πεζών, ενώ ένα AV κινείται σε ευθεία και ένας πεζός επιχειρεί να διασχίσει τον δρόμο.

5.2 ΣΥΜΠΕΡΑΣΜΑΤΑ ΕΡΕΥΝΑΣ

Η χρήση αντίστροφης ενισχυτικής μάθησης για τον σχεδιασμό στρατηγικών οδήγησης για τα αυτόνομα οχήματα αποτελεί μία προσέγγιση, που μπορεί να εξασφαλίσει πιο ρεαλιστικά αποτελέσματα μέσω της μάθησης από επιδείξεις ειδικών. Επιπλέον, δύναται να μειώσει τον χρόνο σχεδιασμού σε πολύπλοκα προβλήματα με την απευθείας εξαγωγή της συνάρτησης ανταμοιβής, καθώς και να μειώσει το κόστος σχεδιασμού.

Η στοχαστική φύση της αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας επιτρέπει την προσαρμοστικότητά του μοντέλου σε ποικίλες συμπεριφορές πεζών, ενισχύοντας την απόκριση και την ευρωστία του. Με την ενσωμάτωση πολλαπλών χαρακτηριστικών και πληροφοριών περιβάλλοντος, το Max-Ent IRL μοντέλο συμβάλει στη λήψη συναφών αποφάσεων οδήγησης. Αυτό οδηγεί σε βελτιωμένη ασφάλεια, μεγαλύτερη αποδοχή και αποτελεσματικότητα, ενώ παράλληλα αντικατοπτρίζει τις αναμενόμενες συμπεριφορές των οδηγών και των πεζών.

Η συγκεκριμένη Διπλωματική εργασία αποτελεί μία από τις πρώτες έρευνες, που επιχειρούν να αξιοποιήσουν την χρήση αντίστροφης ενισχυτικής μάθησης μέγιστη εντροπίας για την ανάπτυξη μιας πιο αποτελεσματικής στρατηγικής οδήγησης για την αλληλεπίδραση των αυτόνομων οχημάτων και των πεζών. Από τα αποτελέσματα που παρουσιάστηκαν στο Κεφ.4, προκύπτει ότι ο αλγόριθμος Max-Ent IRL επιδεικνύει ικανοποιητική απόδοση στην εκπαίδευση και αναπαράσταση των ανθρώπινων αξιών και προτιμήσεων. Επομένως, αποτελεί μια αποτελεσματική προσέγγιση για την αναπαράσταση των αλληλεπιδράσεων μεταξύ των AVs και των πεζών.

Ο αλγόριθμος Max-Ent IRL με την εκπαίδευση με τις διατιθέμενες τροχιές και την ρύθμιση των παραμέτρων του φτάνει σε σύγκλιση, όπως φαίνεται από τα βάρη ανταμοιβής. Έπειτα, τα βάρη ανταμοιβής που εξάγονται από την συνάρτηση, για τα χαρακτηριστικά που περιγράφουν την αλληλεπίδραση του αυτόνομου οχήματος με τον πεζό που διασχίζει τον δρόμο, αντικατοπτρίζουν τις βασικές προτιμήσεις και τη συμπεριφορά τόσο των οδηγών όσο και των πεζών. (Διαγρ.6) Επιπλέον, οι ανακτηθείσες ανταμοιβές τον καταστάσεων από τον αλγόριθμο αιτιολογούν τις ανθρώπινες προτιμήσεις των οδηγών και των πεζών, αποδεικνύοντας την ικανότητά του να αποτυπώσει επιτυχώς τις ανθρώπινες προτιμήσεις μέσω των παρεχόμενων τροχιών. (Διαγρ.7) Συνεπώς, ο αλγόριθμος είναι ικανός να διαμορφώσει μια αποτελεσματική στρατηγική οδήγησης για ασφαλή σενάρια αλληλεπίδρασης μεταξύ αυτόνομων οχημάτων και πεζών.

5.3 ΠΕΡΙΟΡΙΣΜΟΙ

Οι περιορισμοί της παρούσας έρευνας περιλαμβάνουν τον *περιορισμένο* αριθμό των διαθέσιμων δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση του αλγορίθμου. Τα δεδομένα αυτά αποτελούνται από 16 τροχιές οχημάτων και, λόγω της περιορισμένης φύσης τους, δεν ήταν δυνατή η εκπαίδευση του αλγορίθμου σε πολλά πιθανά σενάρια και καταστάσεις, ούτε ο χωρισμός των δεδομένων για την εκπαίδευση και την αξιολόγηση του αλγορίθμου (training, testing set). Πραγματοποιήθηκε όμως μια αρχική διερεύνηση πάνω στα βασικά χαρακτηριστικά που επηρεάζουν την οδήγηση του αυτόνομου οχήματος, όπως η ταχύτητα, το χωρικό χάσμα και η διαφορά ταχύτητας μεταξύ οχήματος και πεζού.

5.4 ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΠΕΡΑΙΤΕΡΩ ΕΡΕΥΝΑ

Η διαθέσιμη βιβλιογραφία σχετικά με τις αλληλεπιδράσεις πεζών και αυτόνομων οχημάτων εξελίσσεται και επεκτείνεται συνεχώς. Αυτό το πεδίο εξακολουθεί να είναι σχετικά νέο, επομένως υπάρχει αυξανόμενο ενδιαφέρον και ενεργή έρευνα για την κατανόηση και τη βελτίωση της δυναμικής, της ασφάλειας και της εμπειρίας χρήστη κατά τις αλληλεπιδράσεις αυτόνομων οχημάτων και πεζών. Καθώς η τεχνολογία προχωρά και τα AVs γίνονται πιο διαδεδομένα, η έρευνα σε αυτόν τον τομέα αναμένεται να συνεχίσει να

εξελίσσεται, αντιμετωπίζοντας διάφορες προκλήσεις και τελειοποιώντας στρατηγικές για αποτελεσματικές και ασφαλείς αλληλεπιδράσεις AVs και πεζών. Επομένως, είναι λογικό να διερευνηθεί περαιτέρω η χρήση της Max-Ent IRL για αυτές τις αλληλεπιδράσεις, καθώς αυτή η μέθοδος έχει πολλά πλεονεκτήματα και δεν έχει αξιοποιηθεί επαρκώς.

Σε μελλοντικές έρευνες, πάνω στη χρήση της αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας (Max-Ent IRL) για την ανάπτυξη στρατηγικών οδήγησης AV με την παρουσία πεζών, θα μπορούσαν να χρησιμοποιηθούν περισσότερα δεδομένα εισόδου σε μορφή τροχιών οχημάτων και πεζών. Εάν υπάρχουν περισσότερες διαθέσιμες τροχιές, ένα ποσοστό αυτών μπορεί να χρησιμοποιηθεί για την εκπαίδευση του μοντέλου και το υπόλοιπο για την μετέπειτα αξιολόγησή του (π.χ. 80% εκπαίδευση και 20% εξέταση). Αυτή η προσέγγιση θα επέτρεπε την διαμόρφωση ενός πιο αξιόπιστου μοντέλου, καθώς και πιθανόν πιο γρήγορη σύγκλιση για το μοντέλο. Επιπλέον, με την χρήση περισσότερων δεδομένων το μοντέλο θα κάνει καλύτερη γενίκευση και θα είναι πιο σταθερό και αξιόπιστο.

Για περαιτέρω βελτίωση της ακρίβειας και της αξιοπιστίας των μοντέλων αλληλεπίδρασης μεταξύ αυτόνομων οχημάτων και πεζών, μπορούν να εξεταστούν περισσότερες καταστάσεις και χαρακτηριστικά για το μοντέλο. Η προσθήκη επιπλέον πληροφοριών θα επιτρέψει στο μοντέλο να εκπαιδευτεί σε περισσότερες πιθανές καταστάσεις, συμβάλλοντας σε πιο γρήγορες και αξιόπιστες αποφάσεις. Για παράδειγμα, μπορεί να εξεταστεί η προσθήκη περισσότερων πεζών και διαφορετικών σεναρίων μετάβασης, όπως η διάσχιση σε διάβαση ή σε κοινόχρηστο χώρο (ελεύθερη ροή). Επιπλέον, μπορεί να προστεθούν σχετικές πληροφορίες περιβάλλοντος, όπως κυκλοφοριακές συνθήκες, καιρικές συνθήκες, προφίλ πεζών, και κανόνες κυκλοφορίας. Μια άλλη παράμετρος που θα μπορούσε να μελετηθεί είναι η προσθήκη περισσότερων ενεργειών στο μοντέλο, όπως περισσότερες κλάσεις επιταχύνσεων και επιβραδύνσεων, κρίσιμες για την αποφυγή ατυχημάτων, καθώς και η προσθήκη της διατήρησης σταθερής πορείας ως επιπλέον ενέργειας και πώς αυτές θα επηρεάσουν τις ανταμοιβές τους αλγορίθμου.

Καθώς τα μοντέλα αντίστροφης ενισχυτικής μάθησης εξελίσσονται συνεχώς και επεκτείνονται με νέες πληροφορίες, η προσθήκη σεναρίων, όπως περιπτώσεις με σημαντική επιτάχυνση, όπου οι πεζοί αποφεύγουν τη διάσχιση του δρόμου, θα οδηγήσει σε ανταμοιβές που αντικατοπτρίζουν την ανεπιθύμητη συμπεριφορά. Επιπλέον, μπορούν να προστεθούν περισσότερα σενάρια συμπεριφοράς πεζών, επιτρέποντας στα αυτόνομα οχήματα να αντιμετωπίζουν πεζούς με διαφορετικές συμπεριφορές.

Θα μπορούσε, επίσης, να ερευνηθεί η χρήση βαθιάς αντίστροφης ενισχυτικής μάθησης μέγιστης εντροπίας (Deep Max-Ent IRL). Η χρήση βαθιάς μάθησης επιτρέπει τον χειρισμό χώρων χαρακτηριστικών υψηλών διαστάσεων, και επομένως την αντιμετώπιση πιο σύνθετων καταστάσεων, επιτρέποντας πιο λεπτομερή

μοντελοποίηση. Με αυτό τον τρόπο, το μοντέλο μπορεί να ανακτήσει μη γραμμικές συναρτήσεις ανταμοιβής, επιτρέποντας πιο ευέλικτη και διαφοροποιημένη διαμόρφωση της συμπεριφοράς του. Επιπλέον, το Deep Max-Ent IRL αξιοποιεί τη δύναμη των βαθιάς μάθησης νευρωνικών δικτύων για την αυτόματη εξαγωγή σχετικών χαρακτηριστικών και την καταγραφή σύνθετων μοτίβων, ενισχύοντας την ικανότητά του να μαθαίνει από διαφορετικά και μη δομημένα δεδομένα.

Τα αποτελέσματα του αλγορίθμου αντιστροφής ενισχυτικής μάθησης μέγιστης εντροπίας, που αναπτύχθηκε στην παρούσα διπλωματική, μπορούν να αποτελέσουν κρίσιμο εργαλείο για τη βελτιστοποίηση της συμπεριφοράς ενός οχήματος. Ο προσδιορισμός της βέλτιστης πολιτικής απαιτεί την αξιοποίηση σωστών ανταμοιβών για να εκπαιδεύσει το όχημα να λαμβάνει ασφαλείς και αποτελεσματικές αποφάσεις. Ο αλγόριθμος που βασίζεται στη μέγιστη εντροπία, διαμορφώνει αυτές τις ανταμοιβές, παρέχοντας μια αξιόπιστη βάση για το ευθύ πρόβλημα ενισχυτικής μάθησης. Με αυτήν την προσέγγιση, το αυτόνομο όχημα μπορεί να μάθει να προσαρμόζεται σε διάφορες καταστάσεις και να λαμβάνει αποφάσεις που βελτιστοποιούν την ασφάλεια και την αποτελεσματικότητά του κατά την πλοήγηση για ποικίλες συνθήκες κυκλοφορίας και περιβάλλοντα, βασιζόμενο στα αποτελέσματα της παρούσας έρευνας. Αυτή η προσέγγιση παρέχει ένα θεμέλιο για περαιτέρω εξέλιξη των συστημάτων αυτόνομης οδήγησης, προσφέροντας προοπτικές για πιο ασφαλείς και πιο αποδοτικές μελλοντικές κινήσεις στον τομέα της αυτόνομης κυκλοφορίας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Abdelkader, G., Elgazzar, K., & Khamis, A. (2021). Connected Vehicles: Technology Review, State of the Art, Challenges and Opportunities. Sensors, 21(22), 7712. https://doi.org/10.3390/s21227712

Aghasadeghi, N., & Bretl, T. (2011, September). Maximum entropy inverse reinforcement learning in continuous state spaces with path integrals. In 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 1561-1566). IEEE.

Deep Reinforcement Learning, Sergey Levine, διάλεξη από: http://rail.eecs.berkeley.edu/deeprlcourse-fa17/f17docs/lecture_12_irl.pdf

Deshpande, N., Vaufreydaz, D., & Spalanzani, A. (2021, September). Navigation in urban environments amongst pedestrians using multi-objective deep reinforcement learning. In 2021 IEEE International Intelligent Transportation Systems Conference (ITSC) (pp. 923-928). IEEE.

Elallid, B. B., Benamar, N., Mrani, N., & Rachidi, T. (2022, November). DQN-based Reinforcement Learning for Vehicle Control of Autonomous Vehicles Interacting with Pedestrians. In 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) (pp. 489-493). IEEE.

Feng, C., Cunbao, Z., & Bin, Z. (2019, July). Method of pedestrian-vehicle conflict eliminating at unsignalized mid-block crosswalks for autonomous vehicles. In 2019 5th international conference on transportation information and safety (ICTIS) (pp. 511-519). IEEE.

Great Learning Team, 2023, What is Machine Learning? Defination, Types, Applications, and more, από: https://www.mygreatlearning.com/blog/what-is-machine-learning/

Hendricks, D. L., Freedman, M., & Fell, J. C. (2001). The relative frequency of unsafe driving acts in serious traffic crashes (No. DOT-HS-809-206). United States. National Highway Traffic Safety Administration.

Hsu, Y. C., Gopalswamy, S., Saripalli, S., & Shell, D. A. (2018, August). An MDP model of vehicle-pedestrian interaction at an unsignalized intersection. In 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall) (pp. 1-6). IEEE.

Jayaraman, S. K., Tilbury, D. M., Yang, X. J., Pradhan, A. K., & Robert, L. P. (2020, May). Analysis and prediction of pedestrian crosswalk behavior during automated vehicle interactions. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 6426-6432). IEEE.

Kato, S., Takeuchi, E., Ishiguro, Y., Ninomiya, Y., Takeda, K., & Hamada, T. (2015). An open approach to autonomous vehicles. IEEE Micro, 35(6), 60-68.

Krumm, J. (2008, April). A markov model for driver turn prediction. In Society of Automotive Engineers (SAE) 2008 World Congress, April 2008.

Krumm, J., & Horvitz, E. (2006). Predestination: Inferring destinations from partial trajectories. In UbiComp 2006: Ubiquitous Computing: 8th International Conference, UbiComp 2006 Orange County, CA, USA, September 17-21, 2006 Proceedings 8 (pp. 243-260). Springer Berlin Heidelberg.

Kuderer, M., Gulati, S., & Burgard, W. (2015, May). Learning driving styles for autonomous vehicles from demonstration. In 2015 IEEE International Conference on Robotics and Automation (ICRA) (pp. 2641-2646). IEEE.

Letchner, J., Krumm, J., & Horvitz, E. (2006, July). Trip router with individualized preferences (trip): Incorporating personalization into route planning. In AAAI (pp. 1795-1800).

Liao, L., Patterson, D. J., Fox, D., & Kautz, H. (2007). Learning and inferring transportation routines. Artificial intelligence, 171(5-6), 311-331.

Mantouka, E. G., Barmpounakis, E. N., & Vlahogianni, E. I. (2019). Identification of driving safety profiles from smartphone data using machine learning techniques. Safety Science, 119.

Martinez-Gil, F., Lozano, M., García-Fernández, I., Romero, P., Serra, D., & Sebastián, R. (2020). Using inverse reinforcement learning with real trajectories to get more trustworthy pedestrian simulations. Mathematics, 8(9), 1479.

Nageshrao, S., Tseng, H. E., & Filev, D. (2019, October). Autonomous highway driving using deep reinforcement learning. In 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC) (pp. 2326-2331). IEEE.

Neu, G., & Szepesvári, C. (2012). Apprenticeship learning using inverse reinforcement learning and gradient methods. arXiv preprint arXiv:1206.5264.

Ng, A. Y., & Russell, S. (2000, June). Algorithms for inverse reinforcement learning. In Icml (Vol. 1, p. 2).

Orfanou, F. P., Vlahogianni, E. I., Yannis, G., & Mitsakis, E. (2022). Humanizing autonomous vehicle driving; understanding, modeling and impact assessment. Transportation research part F: traffic psychology and behaviour, 87, 477-504.

Orfanou, F. P., Vlahogianni, E. I., Yannis, G, Unai Hernandez, Christelle Al Haddad, Constantinos Antoniou, Lars Töttel, Evangelos Mintsis, (2021, April). Behavioural modelling of autonomous vehicle "drivers", funded from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 815001.

Pieter Abbeel and Andrew Y. Ng. (2004). Apprenticeship learning via inverse reinforcement learning. In Proceedings of the twenty-first international conference on Machine learning (ICML '04). Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/1015330.1015430

Prédhumeau, M., Mancheva, L., Dugdale, J., & Spalanzani, A. (2021, May). An agent-based model to predict pedestrians' trajectories with an autonomous vehicle in shared spaces. In AAMAS 2021-20th International Conference on Autonomous Agents and Multiagent Systems (pp. 1-9).

Ramachandran, D., & Amir, E. (2007, March). Bayesian Inverse Reinforcement Learning. In IJCAI (Vol. 7, pp. 2586-2591).

Ratliff, N. D., Bagnell, J. A., & Zinkevich, M. A. (2006, June). Maximum margin planning. In Proceedings of the 23rd international conference on Machine learning (pp. 729-736).

Sharifzadeh, S., Chiotellis, I., Triebel, R., & Cremers, D. (2016). Learning to drive using inverse reinforcement learning and deep q-networks. arXiv preprint arXiv:1612.03653.

Simmons, R., Browning, B., Zhang, Y., & Sadekar, V. (2006, September). Learning to predict driver route and destination intent. In 2006 IEEE intelligent transportation systems conference (pp. 127-132). IEEE.

Simonelli, F., Bifulco, G. N., Martinis, V. D., & Punzo, V. (2009). Human-like adaptive cruise control systems through a learning machine approach. In Applications of Soft Computing (pp. 240-249). Springer, Berlin, Heidelberg.

Site Uber: (https://www.uber.com/us/en/autonomous/)

Site Waymo: (https://waymo.com/waymo-via/)

Sun, L., Zhan, W., & Tomizuka, M. (2018, November). Probabilistic prediction of interactive driving behavior via hierarchical inverse reinforcement learning. In 2018 21st International Conference on Intelligent Transportation Systems (ITSC) (pp. 2111-2117). IEEE.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Szarvas, M., Yoshizawa, A., Yamamoto, M., & Ogata, J. (2005, June). Pedestrian detection with convolutional neural networks. In IEEE Proceedings. Intelligent Vehicles Symposium, 2005 (pp. 224-229). IEEE.

Talpaert, V., Sobh, I., Kiran, B. R., Mannion, P., Yogamani, S., El-Sallab, A., & Perez, P. (2019). Exploring applications of deep reinforcement learning for real-world autonomous driving systems. arXiv preprint arXiv:1901.01536.

Vitelli, M., & Nayebi, A. (2016). Carma: A deep reinforcement learning approach to autonomous driving. Tech. rep. Stanford University, Tech. Rep.

Vlahogianni, E. & Barmpounakis, E. (2017). Driving analytics using smartphones: Algorithms, comparisons and challenges, Transportation Research Part C: Emerging Technologies, Volume 79, Pages 196-206, ISSN 0968-090X: https://doi.org/10.1016/j.trc.2017.03.014.

Wang, P., Chan, C. Y., & de La Fortelle, A. (2018, June). A reinforcement learning based approach for automated lane change maneuvers. In 2018 IEEE Intelligent Vehicles Symposium (IV) (pp. 1379-1384). IEEE.

Wu, Z., Sun, L., Zhan, W., Yang, C., & Tomizuka, M. (2020). Efficient sampling-based maximum entropy inverse reinforcement learning with application to autonomous driving. IEEE Robotics and Automation Letters, 5(4), 5355-5362.

Wulfmeier, M., Ondruska, P., & Posner, I. (2015). Maximum entropy deep inverse reinforcement learning. arXiv preprint arXiv:1507.04888.

Yannis, G., Papadimitriou, E., & Theofilatos, A. (2013). Pedestrian gap acceptance for mid-block street crossing. Transportation planning and technology, 36(5), 450-462.

Ye, Y., Zhang, X., & Sun, J. (2019). Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment. Transportation Research Part C: Emerging Technologies, 107, 155-170.

You, C., Lu, J., Filev, D., & Tsiotras, P. (2019). Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning. Robotics and Autonomous Systems, 114, 1-18.

Zheng, R., Liu, C., & Guo, Q. (2013, July). A decision-making method for autonomous vehicles based on simulation and reinforcement learning. In 2013 International Conference on Machine Learning and Cybernetics (Vol. 1, pp. 362-369). IEEE.

Zhou, M., Yu, Y., & Qu, X. (2019). Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: A reinforcement learning approach. IEEE Transactions on Intelligent Transportation Systems, 21(1), 433-443.

Zhou, Y., Fu, R., & Wang, C. (2020). Learning the car-following behavior of drivers using maximum entropy deep inverse reinforcement learning. Journal of advanced transportation, 2020.

Zhu, M., Wang, X., & Wang, Y. (2018). Human-like autonomous car-following model with deep reinforcement learning. Transportation research part C: emerging technologies, 97, 348-368.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2009, March). Human Behavior Modeling with Maximum Entropy Inverse Optimal Control. In AAAI spring symposium: human behavior modeling (Vol. 92).

Ziebart, B. D., Ratliff, N., Gallagher, G., Mertz, C., Peterson, K., Bagnell, J. A., & Srinivasa, S. (2009, October). Planning-based prediction for pedestrians. In 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 3931-3936). IEEE.

Ziebart, B.D., Maas, A.L., Bagnell, J.A., & Dey, A.K. (2008). Maximum Entropy Inverse Reinforcement Learning. AAAI.

Δικαιολογητικά, όροι, προϋποθέσεις και διαδικασία θέσης σε κυκλοφορία επιβατηγού οχήματος χωρίς την παρουσία οδηγού επ' αυτού, (2022). Ανακτήθηκε από: https://www.gcdata.gr/data/files/a6b641261fddfd880acc0031a13979d1.pdf

Οι πιο ευάλωτοι χρήστες του οδικού δικτύου στην Ευρώπη (2019), Ανάκτηση από ΙΟΑΣ: https://www.ioas.gr/deltia_typou/4231/Oi_pio_eualotoi_christes_tou_odikou_diktuou_stin_Europi.htm/

Σύνολο δεδομένων INTERACTION για οπτικοποίηση των τροχιών της μελέτης: https://github.com/interaction-dataset/interaction-dataset