이학석사 학위논문

Pedestrian Intention Prediction for Autonomous Driving and Mobile Robots Using a Multiple Stakeholder Perspective Model

다중 이해관계자 관점모델을 이용한 자율주행 및 이동로봇에서 관측한 보행자 행동 예측

2021 년 2 월

서울대학교 대학원

협동과정 인지과학전공

김 경 도

Pedestrian Intention Prediction for Autonomous Driving and Mobile Robots Using a Multiple Stakeholder Perspective Model

다중 이해관계자 관점모델을 이용한 자율주행 및 이동로봇에서 관측한 보행자 행동 예측

지도교수 오 성 회

이 논문을 이학석사 학위논문으로 제출함

2020 년 10 월

서울대학교 대학원

협동과정 인지과학전공

김 경 도

김경도의 이학석사 학위논문을 인준함

2021 년 1 월

위 원 장 최 진 영 (인)

부위원장 오 성 회 (인)

위 원 한 소 원 (인)

This dissertation is based on research published in the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2020* and deals with research that has been extended to mobile robots.

*Kyungdo Kim, Yoon Kyung Lee, Hyemin Ahn, Sowon Hahn, and Songhwai Oh, "Pedestrian Intention Prediction for Autonomous Driving Using a Multiple Stakeholder Perspective Model," in Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct. 2020.

Abstract

Pedestrian Intention Prediction for Autonomous Driving and Mobile Robots Using a Multiple Stakeholder Perspective Model

Kyungdo Kim
Interdisciplinary Program in Cognitive Science
The Graduate School
Seoul National University

This thesis proposes a multiple stakeholder perspective model (MSPM) that predicts the future pedestrian trajectory observed from the vehicle's point of view. The motivation of the MSPM is that a human driver exploits the experience of being a pedestrian when he or she encounters a pedestrian crossing over the street. With many studies focusing on vehicle-vehicle systems, the autonomous vehicle system on freeways is nearing completion to some extent. In order for existing technology in an autonomous vehicle to be applied in urban areas from highways, vehicle-pedestrian interaction must also develop at a rapid pace. For the vehicle-pedestrian interaction, the estimation of the pedestrian's intention is a key factor. However, even if this interaction is commonly initiated by both the human (pedestrian) and the agent (driver), current research focuses on developing a neural network trained by the data from the driver's perspective only. In this paper, we suggest a multiple stakeholder perspective model (MSPM) and apply this model for pedestrian intention prediction. The model combines the driver (stakeholder 1) and pedestrian (stakeholder 2) by separating the information based on the perspective. The dataset from the pedestrian's

perspective has been collected from the virtual reality experiment, and a network that can reflect the perspectives of both pedestrian and driver is proposed. Our model achieves the best performance in the existing pedestrian intention dataset while reducing the trajectory prediction error by an average of 4.48% in the short-term (0.5s) and middle-term (1.0s) prediction, and 11.14% in the long-term prediction (1.5s) compared to the previous state-of-the-art. Also, we collect an indoor pedestrian dataset, which includes various human behavior in indoor environments. With these data and models, we suggest a method when training and testing robots using data collected by different robot platforms.

Keywords: Intention prediction, Autonomous Driving, Mobile robots, Human-robot interaction, Cognitive modeling

Student Number: 2019-29337

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The global autonomous vehicle market is projected to reach $615.02 billion by 2026 and grow at an average annual growth rate of 41.5% over the period [1]. Now it is inevitable to bring an autonomous vehicle into our transportation system. However, bringing an autonomous vehicle into our society would cause various problems such as object recognition errors in the driving situation, as it is controlled by an intelligent system that handles the interactions between vehicles, drivers, and pedestrians. Among these interaction issues, we point out that existing work does not focus on vehicle-pedestrian interaction, although it is important as autonomous cars move away from the highway-centric system and head into the city center, where pedestrian encounters are frequent.

1.1 Conventional approach of pedestrian intention estimation

The ultimate goal of the pedestrian intention estimation study is to identify future trajectories of pedestrians through past patterns of their behavior. To fulfill this, [2, 3, 4, 5] extracted trajectory information from the scene images using deep neural networks, and [2, 6] segmented an image data into different features and fed it through different network architectures. However, the drawback of current existing research is that making a robust and precise behavioral

prediction is difficult since the performace of existing models only depends on the information observed from the driver's point of view. For example, the dataset from [2] also provides the movement information of the pedestrian only observed from the driver's perspective. In addition, the proposed network from [2] only focuses on how the driver can employ pedestrian's information based on his or her perspective.

There exists numerous researches on predicting future action of observed targets from sequential images (or video). Recent works on action prediction have focused on the use of past and current scene information [7, 8, 9]. In these cases, their ultimate goal is to predict the future action of targets through various neural network architectures. To predict the intention of pedestrians, [2, 10] suggested a dataset related to the trajectory of objects and pedestrians which are observed by the driver, [11, 12] reported that using the head orientation information can enhance the accuracy of this prediction. In addition, [13] targets the POMDP approach to estimate the future trajectory efficiently and robustly.

Since only depending on extracted image features can increase the noise, [7, 4] proposed to modularize the information in terms of the human body segmentation and activate each network module separately in order to reduce the noise effect. Also, [4, 5] proposed 2D skeleton pose estimation, [2, 7, 6] used both an image input and object motion data extracted from image segmentation via a separated network. However, current research remains based on a driver-centered approach, even if the estimation of pedestrian intention is based on interactions between autonomous vehicles and pedestrians.

1.2 Limitation of current PIE research

In a real traffic system, it is rational for a driver to estimate the observed pedestrian's behavior based on one's memory when he or she was a pedestrian. We claim that this concept can be applied when training a neural network model for pedestrian trajectory estimation. To sum up, our model takes advantage of perspective combination, which indicates the utilizing information from both driver's and pedestrian's perspectives. This inspiration is based on the existing studies related to the robotics[14] and neuroscience[15, 16], which have shown that combining information from different perspectives can improve the performance of the path prediction. In this paper, we empirically show that the performance of existing studies can be improved by employing *pedestrian experiences* when training neural network models. To the best of our knowledge, our model is the first to use both perspectives.

1.3 Main Contribution

The main contributions of this thesis are as follows: First, we propose a multiple stakeholder perspective model (MSPM) for the vehicle-pedestrian interaction problem. By adding a novel network and data from a pedestrian's perspective, it is empirically shown that a more reliable prediction is possible. Second, the proposed model uses pedestrian behavior using data collected in a virtual reality system. Although there exist several works using virtual reality, applying the virtual reality(VR) system to detect the intention of pedestrians is the first one as far as we know. Third, we collect an indoor navigation dataset that includes human-vehicle interaction information. While current research does not cover much of the indoor mobile robot dataset, the model cannot learn about these environments firmly. After we collect and train the data using mobile

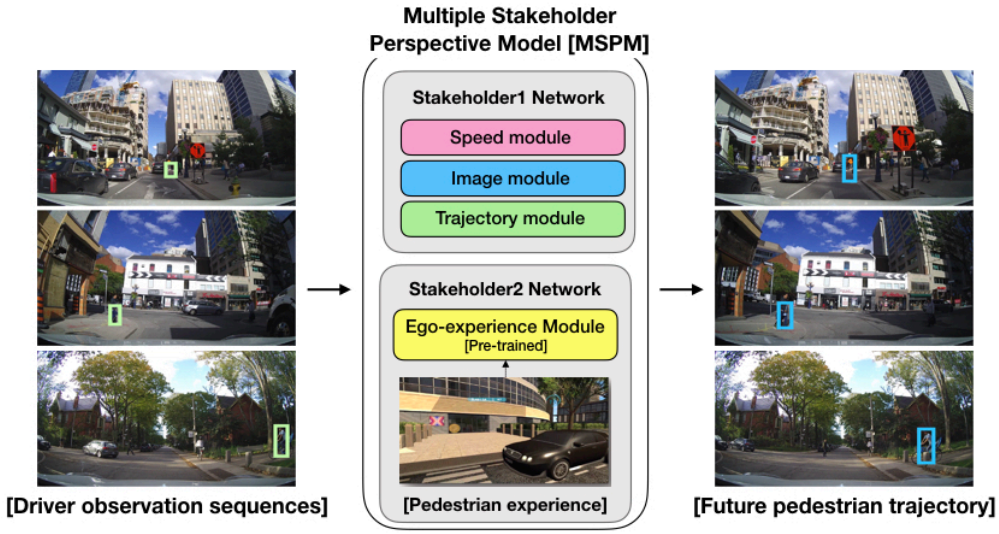


Figure 1.1: An overview of the proposed Multiple Stakeholder Perspective Model (MSPM) for the pedestrian trajectory prediction. This model considers both information from the driver's point of view and from the pedestrian's point of view. The stakeholder 2 network (pedestrian-perspective network) is pretrained by the pedestrian experience and implemented in the MSPM. Then, the overall MSPM model is trained by the driver observation data and predicts the future pedestrian trajectory.

robots, the prediction accuracy has enhanced even in harsh inputs. Finally, my MSPM provides a cutting edge result in a recent pedestrian intention estimation dataset. Using my model, the loss of trajectory prediction is reduced by 8.61% for middle-term prediction (1.0s) and 11.14% reduction for long-term prediction (1.5s) compared to the best-known result[2].

Chapter 2

Human behavior dataset analysis

2.1 Pedestrian perspective data

For this research, we can get data from the Human Factors Laboratory. They have built a pedestrian crossing scenario, in which participants interacted with an automated driving agent. Participants had to cross a crosswalk when a ready sign disappeared. The virtual reality(VR) scenario was created with Unity 2017. Participants wore HTC Vive Pro headphones. The virtual world had buildings, an engine noise of the vehicle and the occasional appearance of an avatar crossing together.

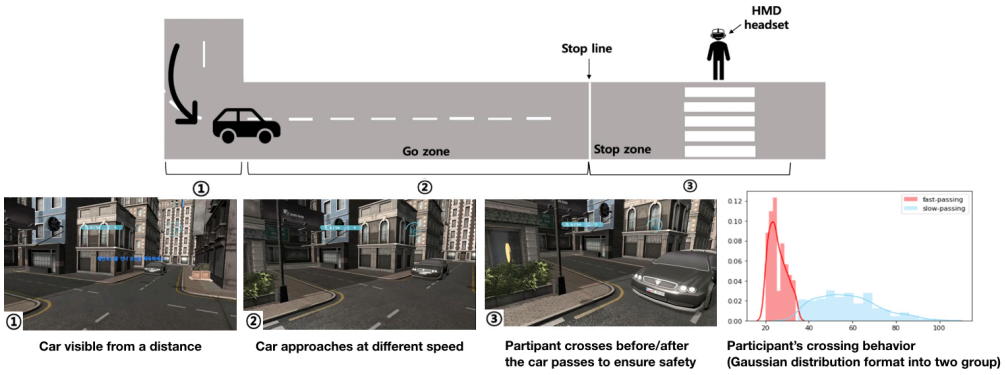


Figure 2.1: VR experiment scenario and participant's crossing behavior (In the graph of participant's crossing behavior, X axis denotes the timestamp and Y axis denotes the ratio of participants crossing during each timestamp).

As shown in Figure 2.1, a test consists of (1) a car appearing at the end of the turn, (2) a car approaching the pedestrian crossing, and (3) a car stopping or overtaking. Each test is initiated by a participant standing in the designated area indicated by a green arrow. Each test includes a "Ready" and a "Start" phase. The "Ready" phase ended after 3 seconds and the "Start" phase started with a delay of 20 seconds. Before starting the experiment, participants had a chance to practice his/her crossing behavior in a VR system for 3 times and we excluded this exercise from the analysis. Participants were asked to take a time and cross at any time; however, they aim to cross the road without being hit by a car. The cars were different in size, speed, and initial departure direction (left and right) at random in each test. We measured the head orientation of the participants every 200 ms. Participants were asked to follow the rules accordingly: (1) cross the crosswalk safely, (2) avoid being struck by the approaching car, (3) cross within a time limit. We put a time limit of 20 seconds on a single trial to continue to the next trials without delaying too much time. After the participant ended the entire experiment, We asked basic and post-demographic questions. The participants were then debriefed and left. A total of 39 participants (17 females) conducted 54 trials each. Including installation and post-maintenance time, the entire experience took approximately one hour.

2.2 Data extraction and Analysis

After the VR experiment, we extracted the information such as vehicle speed, the distance between the pedestrian and vehicle into our feature vector for each timestamp. In addition, we extracted the pedestrian behavior data such as the angle of head movement, and the position of the pedestrian. After analyzing these VR data, we found that subjects can be divided into two groups. Ac-

cording to the participants, some people (slow-passing group, Group 1) really care about the movement of the vehicle and move with caution. Otherwise, there are people (fast-passing group, Group 2) who do not care much about the movement of the vehicle and who move forward. Therefore, we have divided the collected data into these two groups. Group 1 with people who crossed the road with 35 timestamps (7 seconds) and Group 2 with people who crossed the road with less than 35 timestamps (see Figure 2.1). Using Bayes' statistical processing, we concluded that this is a reasonable approach to divide our data into these two groups. We defined each group as "slow-passing group (Group 1)" and "fast-passing group (Group 2)" and prepared data for each group.

Chapter 3

Mobile Robot Approach

3.1 Motivation

At the same time as the process of making a prediction model and measuring the result for the existing autonomous vehicle dataset, the work of constructing a dataset for navigation in an indoor environment was conducted. Previously, most of the navigation model and pedestrian intention estimation have been carried out using the classical statistical method. However, with the recent development of hardware and various researches in the field of deep learning, the overall model is being changed to a deep neural network. What makes this deep neural network different from existing statistical methods or machine learning is that a vast amount of data must be constructed. The weight of the layer can be freely adjusted, and the more data there are, the more sophisticated the posterier distribution can be built based on this, so big data is needed along with the scalability of the network and hardware.

In particular, more diverse cases of data collection is required for pedestrian behavior, but the existing open-source navigation dataset has not yet reflected this. Representative pedestrian behavior datasets are the PIE_dataset from ICCV 2019[2] and the Honda dataset from CVPR 2020[17]. Both of these datasets contain pedestrian information (image, trajectory, etc.) and vehicle information (speed, steering, etc.) for pedestrians passing by from a vehicle

perspective.

Of course, we can also use these datasets to train and perform pedestrian intention estimation. However, the model trained in this way focuses only on the intention prediction of the case of observing people on the road due to the nature of the training data. For delivery robots and indoor mobile robots to be finally applied in this study, more focused data on the indoor environment was required.

3.2 Data Collection Pipeline

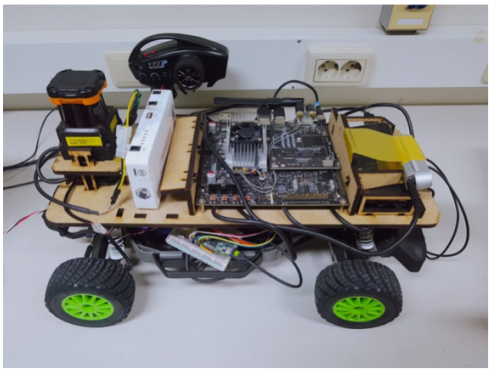
So, in this study, the indoor navigation dataset was also built. After installing and remodeling the sensor on the mobile robot platform in the laboratory, data collection was carried out by calculating various scenarios. The overall plan is shown in the figure below.

Figure 3.1: Schematic diagram of the mobile robot data pipeline. Since it is a data structure that did not exist in the past, it was made in an optimal form through trial and error from hardware setting.

As shown in this Figure 3.1, the entire collection and prediction process consists of 1. Hardware setting, 2. Test driving, 3. Hardware modification, 4. Collection scenario setting, 5. Data collection, 6. Draw bounding box, 7. Make model & prediction.

3.2.1 Hardware setting

The most important feature of the robot platform for use as an indoor mobile robot is mobility. The degree of freedom of human movement is very high, and in the real field, the robot should move in the middle of a lot of these people. Therefore, the complexity of the environment in which the robot moves is very high. Among these, the movement of the robot must be determined appropriately in accordance with the movement of the person in order to go toward the designated goal without conflicting with the person. What differentiates it from simple vehicle driving is that it is unlikely to collide with people on the road, but it is very high in indoor environments. Besides, due to the nature of robots, if it collides with a person, the damage to the person is large, and the mobile robot must move more agilely. Considering that the environment to which the actual research will be applied is also considered, the mobile robot platform was carefully selected from the time of collection.



Mobile robot 1. RC Car

Mobile robot 2. Husky

Figure 3.2: Mobile robots used in this study. An RC (Radio control) car with good mobility and a Husky robot with good stability were used.

In this study, data collection was carried out using RC Car and Husky

robot as shown in the Figure 3.2. First of all, RC Car can be equipped with various sensors and smooth control is possible. In this experiment, in particular, measurements were carried out using Realsense's camera and depth sensor. In addition, vehicle information was collected by separately placing a speed sensor and a steering sensor. A husky robot is characterized by being able to drive stably because it has a heavier body and a lower center of gravity than an RC car. Realsense cameras and depth sensors are also placed in Husky, so that even if the robot platform changes, data processing is unified. Since this experiment required a robot that can move freely between people, an appropriate robot was selected in consideration of both mobility and stability when selecting an experimental robot.

3.2.2 Test driving

Figure 3.3: Results of trial run in various environments using RC Car and Husky robot.

The trial operation was conducted based on the robot designed in this way, and the results are as follows.

RC Car As shown in Figure 3.3, The biggest problem that occurred during the trial run was that it was not possible to properly take the shape of a person. The purpose of the existing RC car is to focus on fast driving. Therefore, the body of the RC Car is designed very low. When images are collected through a camera in such a place, it is difficult to properly capture a person's appearance, and this would lead to a result that makes prediction difficult. Besides, the training of this model is carried out through a dataset taken from an actual car, and it will be applied to the indoor driving dataset. But if the height is too different, the learning and testing process would not proceed properly. This is similar to a person's cognitive structure, if a person is asked to predict other people's movements in a dataset with a low body based on what they have experienced at the car's eye level, they will not do well. This is because it is a dataset that has not been experienced before, and if it is collected by raising the camera a little, similar to the eye level of an existing car, high prediction accuracy could be expected. Also, the final goal is not to estimate the pedestrian intention through the data seen from the low vehicle body, but because the vehicle body of most mobile indoor robots is near the waist of a person, the position of the camera and sensor is adjusted according to the actual driving environment.

Husky In the case of Husky, it was possible to install the camera in a high place because the vehicle body was larger and more stable. Therefore, in Husky, the camera was installed near 1.25m, which is the height installed on the existing support, and the experiment could be conducted.

3.2.3 Hardware modify

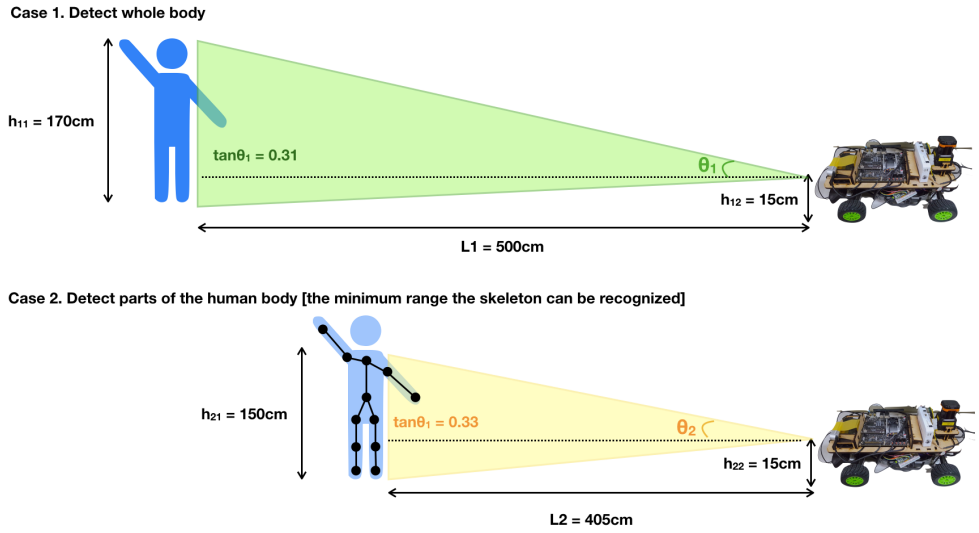


Figure 3.4: The appearance of a person that can be observed according to distance when measured in a previous RC car. When experimenting with this RC car, a proper image can be obtained only when a person is placed at 4 5m as shown in the picture above.

RC Car As shown in Figure 3.4, based on the height of the current RC car, it was possible to identify the overall appearance of a person from a distance of 4-5m. In fact, we experimented with various possibilities as to whether a person will recognize it in advance when a person is located, and it was concluded that if a person can be identified from about 2m, it can respond in advance, and even a person can easily recognize it as a robot. In addition, even if the entire human body does not come out, the skeleton recognizes that it is a human and the model design that responds accordingly is possible with the current research, so the height was adjusted using this as a Maginot Line.

In order to install the camera at a high place in the existing vehicle body,

the height was adjusted using a little calculation formula. If the camera was raised too high, similar to the height of the camera measured in an actual car or human eye level, there was too much vibration and turbulence in the vehicle body, and the image was rather blurred. Therefore, a design was needed to collect data from an appropriately high position, but to prevent the RC car from shaking stably when driving.

Finally, by installing the camera at a height of 60cm from the RC car, it was possible to establish an environment suitable for both data and pedestrian positions. After that, as a result of investigating the indoor mobile robot by Delivery Heroes which used in the actual field, it was confirmed that the camera was positioned and set in a similar place to that used in this study.

In addition, we tried to offset the vibration by using a gimbal that is widely used in video recording so that the RC car can run stably. As a result of pulling out the support designed in this way by 3D printing and applying it to an actual RC car, data with some degree of shaking was obtained.

Husky Husky had the disadvantage of having a large body and slow movement, but since it was able to stably collect data, it proceeded without modifying the hardware.

3.2.4 Collection scenario

Until now, most of the navigation datasets simply observed humans from the vehicle's point of view (case 1), so this time we tried to collect a lot of shapes that reflect actual interactions (cases 2 and 3) as shown in Figure 3.5.

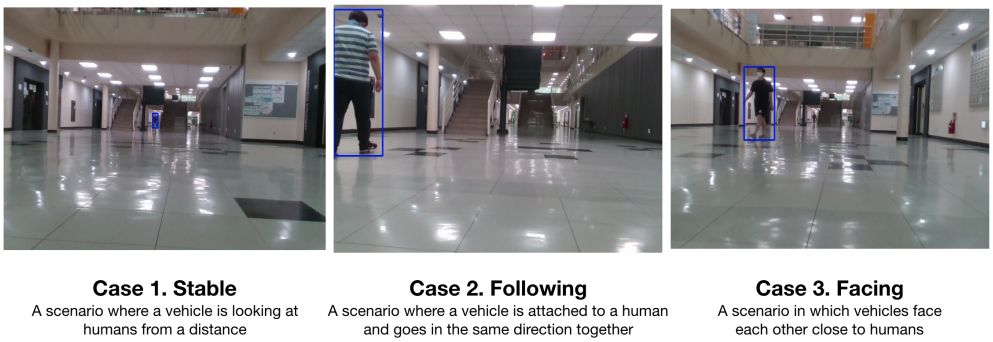


Figure 3.5: Possible scenarios when a mobile robot faces a person.

In applying to mobile robots like delivery robots, there are many situations in which the robots will actually encounter cases 2 and 3, not case 1, but these cases are rarely included in the current open dataset. Therefore, in this study, a scenario that can collect many of these cases 2 and 3 were set and proceeded.

3.2.5 Data collection

RC Car The main purpose of gathering this dataset was to contain human-robot interaction. Therefore, for this purpose, the following scenarios in which pedestrians and mobile robots can interact were calculated and experimented as shown in Figure 3.6.

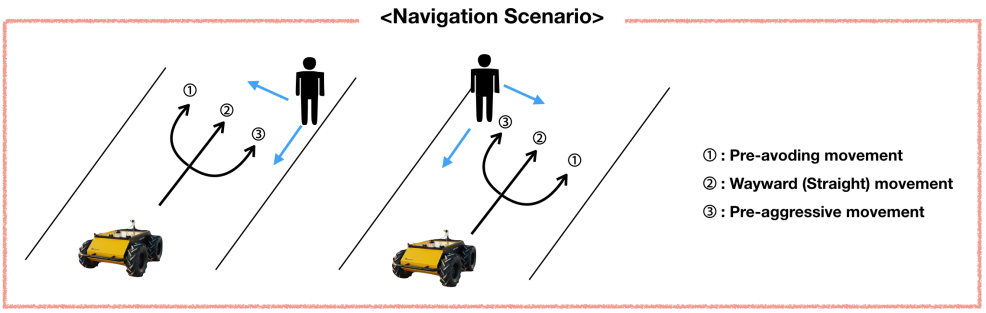


Figure 3.6: Scenarios in which mobile robots can interact with pedestrians.

The experiment was conducted after sufficient control exercises were performed so as not to make movements that are too threatening to humans. The experiment was conducted from 2020.07 to 2020.11. The test site was first conducted in buildings such as 38, 39, 133, and 301 buildings inside Seoul National University. These buildings were selected based on criteria such as various environments, places where it is easy to carry and experiment with robots, and places where there are some floating populations. However, due to the global COVID-19 pandemic, the amount of data compared to the experiment time was very small because there were not many people on campus. Therefore, the experiment was conducted according to various cases and scenarios after waiting for people to come at a time when as many people gather as possible.

And as a result, a total of 32388 frames of data could be collected. In fact, considering that the amount of data used as a test in the existing open-source dataset is 36,000 frames (10 minutes), it can be said that enough data has been collected for the experiment. This frame contains pedestrian information and vehicle information (speed, steering, etc.).

Husky Although Husky collected thousands of frames of data, it was discovered that there was an error in the code in the process of sequentially stacking data in the middle, and it was not possible to proceed after that. In addition, the weight of the Husky robot itself was too heavy, and it was rather bulky to operate with when people actually move and interact in the indoor environment, so it was delayed for the next study.

3.2.6 Draw bounding box

Based on the measured data, all areas where the real person is located were extracted in the form of bounding boxes. In the previous end-to-end learning,

image, and sensor data observed by automobiles were collectively inserted to perform pattern recognition and processing, but in recent research, research is being conducted to separate and process them. This study also focused on modularizing information, so it was necessary to properly annotate the collected dataset. It is very difficult to annotate data of tens of thousands of frames. Therefore, in the existing research, large amounts of money were spent by employing Amazon Mechanical Terk or part-time employees. However, in this study, since there was no room for such a large-scale experiment, the results after running the existing pedestrian detection algorithm were fine-tuned.

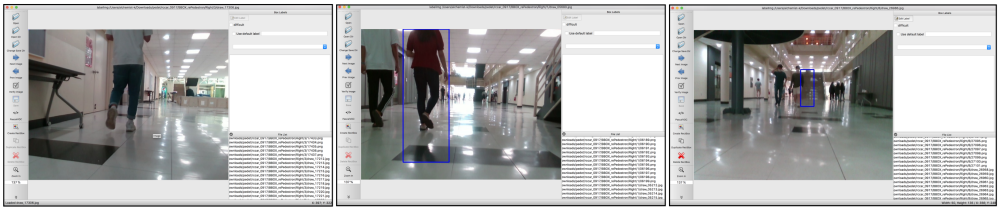


Figure 3.7: Fine-tuning (Image labeling) based on collected data

The pedestrian detection algorithm used in this study is Pedestron[18]. It showed the best performance among the algorithms available so far and is easy to modify in various ways, so this algorithm was selected and proceeded. Initially, this algorithm did not properly detect pedestrians. This is because the existing algorithm is mainly trained on a clean dataset observed in automobiles. Since the data collected this time contains a lot of vibrations and rapid movements, there are many blurring values. In addition, in an indoor environment different from the outdoor, a lot of images were included in which shadows were generated or reflected depending on the lighting. In this harsh dataset,

the existing algorithm could not properly detect pedestrians. Therefore, in this study, while engineering based on this algorithm, the detection algorithm was revised every time until the desired pedestrian for each scenario appeared, and the baseline pedestrian trajectory was extracted for each case.

After that, we observed the frames one by one as shown in the picture and corrected any errors as shown in Figure 3.7 by using [74]. As a result, a total of 11619 annotated frames were obtained. In fact, even in the case of an open dataset, assuming there are 36,000 frames, a person actually appears and the annotated frame is only about 20-30% of them. Therefore, the data collected this time can be said to have collected much better data in terms of purity than the existing dataset. In sum, fine-tuning was completed for the pedestrian trajectory and vehicle information for the measured data, and the additional information extracted in this way helps the model predict the actual human movement.

Chapter 4

Network Architecture

4.1 Cognitive Motivation

The architecture of the proposed model has been inspired by the human cognitive structure. Human cognition can be treated as an information processing system[19]. In particular, human can build the *Theory of Mind* (ToM) model, which is a model of the physical and psychological states of others[20, 21]. This model assumes that a human has an ability to build a representation of mental states and assess the unknown intention of others (human or artificial agent). This concept has been applied in multi-agent systems[22, 23, 24, 25, 26], and recent works [23, 24, 27, 28] based on the concept of ToM have shown better performance in human-agent and human-robot interaction field.

In a traffic system, we assume that pedestrians and the autonomous driving agent can have a *theory of mind* for each other. In this case, the autonomous driver would possess both representations of driver itself, and of pedestrian interacting with itself[29, 30]. Existing works[31, 32, 33, 34] have also shown that a driver agent with the ability to build a mental model of pedestrians can lead to the better performance when estimating pedestrian's crossing intention. Inspired by this, we propose a multiple stakeholder perspective model (MSPM), which utilizes the information from both driver's and pedestrian's perspective.

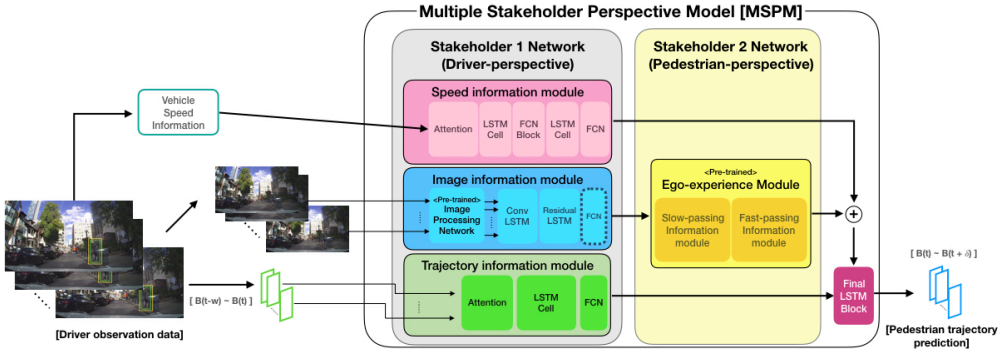


Figure 4.1: Illustration of the stakeholder 2 network when it is in pretraining procedure. First, we pretrain an ego-experience module inside the stakeholder 2 (pedestrian-perspective) network. Based on the pedestrian-perspective data collected from the virtual reality experience, the stakeholder 2 network is trained to generate a feature vector which represents the behavior and position of pedestrians. After training this network, the network is appended next to the stakeholder 1 (driver-perspective) network in the MSPM model. Through the entire MSPM model, driver observation data is used for training and test to predict the pedestrian's future trajectory. Note that the final FCN unit (dashed line) in image information module is deactivated when the stakeholder 2 network is implemented.

4.2 Multiple stakeholder perspective model (MSPM)

A multiple stakeholder perspective model (MSPM) is designed to reflect all perspectives of stakeholders involved in a given interaction situation. In particular, in vehicle-pedestrian situation, we have set up a network of a vehicle (stakeholder 1) and a pedestrian (stakeholder 2). As shown in Figure 4.1, the stakeholder 1 (driver-perspective) network consists of a speed, image, and trajectory information modules for processing data observed from a driver. The

stakeholder 2 (pedestrian-perspective) network is a network pretrained with pedestrian experience data detailed in Section 2.1. These two networks are combined to predict the future pedestrian trajectory from driver-perspective information. Previous works have been conducted to combine data from different angles to improve the accuracy and robustness [6, 7, 14]. In this paper, by combining first-person (driver) and third-person (pedestrian) narrative scene data, we have achieved robust and competitive results compared to the previous works [2, 3] which only focus on single-perspective scene data.

The overall structure of MSPM follows the encoder-decoder scheme. As shown in Figure 4.1, the stakeholder 1 network and the stakeholder 2 network work as an encoder to build a feature representation space, and final LSTM block works as a decoder to predict a pedestrian trajectory.

Future trajectory prediction can be defined as an optimization process that finds the best future prediction given past information[2]. In this case, the model receives the trajectory information $B_{obs} = \left\{B_i^{t-w}, B_i^{t-w+1}, ..., B_i^t\right\}$, where B_i^t is a 2D bounding box around the pedestrian in ith scene at time t, defined by top-left and bottom-right points ($[(x_1, y_1), (x_2, u_2)]$). Also, the model receive the vehicle speed information $S_{obs} = \left\{S_i^{t-w}, S_i^{t-w+1}, ..., S_i^t\right\}$, and the image information $I_{obs} = \left\{I_i^{t-w}, I_i^{t-w+1}, ..., I_i^t\right\}$ where $I \in \mathcal{I} \subset \mathbb{R}^{n_i \times n_j \times 3}$ as inputs. Here, S_i^t, I_i^t denote the vehicle speed and image in ith scene at time t, $\mathcal{I}$ is a set of images observed by the driver point of view, and n_i, n_j is the size of the image. And the model generates the future trajectory B_{pred} by learning distribution $p\left(B_{pred}|B_{obs}, S_{obs}, \mathcal{I}_{obs}\right)$, while B_{pred} is defined by top-left and bottom-right corner points in the form of a 2D bounding box.

Before training the entire MSPM model based on the driver observation data, the stakeholder 2 network is trained in a supervised way with our VR dataset. In this pretraining procedure, the ego-experience module composing

the stakeholder 2 network maps the raw VR-based input data into the high dimensional space $\mathcal{Z}$, which would represent the vehicle-pedestrian interaction information. When the stakeholder 2 network is used in a test phase, the output feature from the image information module in the stakeholder 1 network is projected into the $\mathcal{Z}$ and used as an input to the stakeholder 2 network. To enhance the performance of this feature projection procedure, we feed the pretrained stakeholder 2 network into the entire MSPM, and re-train the entire network so that it can perform the pedestrian trajection prediction robustly.

4.2.1 Stakeholder 1 (Driver-perspective) network

When designing a driver-perspective network, we have focused on dividing information and network module so that the entire network can manage a complex set of driver-perspective data. Existing works[3, 7, 6] have empirically shown that the categorization of information and its divided processing is effective when processing a complex dataset. Thus, several recent studies[2, 3] related to the pedestrian trajectory estimation also follow this concept, and their model's performance has been increased after this *modularization*. In our model, the stakeholder 1 network also follows this information and network modularization.

We divided the information and module into three parts: the *speed* of the vehicle, the *image* of the scene and the annotated *trajectory* of the observed targets in the scene. Regarding the reason of dividing the vehicle *speed* module, through the survey after the VR experiment as described in Section 2.1, we have empirically found that the vehicle speed is the key factor that affects pedestrian's crossing behavior.

For network details, the speed information module (the pink block in Figure 4.1) receives the vehicle speed information (S_{obs}) as inputs, the image infor-

mation module (the blue block in Figure 4.1) gets the image sequences (I_{obs}) as inputs, and the trajectory information module (the green block in Figure 4.1) employs the bounding box of pedestrian (B_{obs}) as inputs. A stakeholder 1 network is trained based on a supervised way, where inputs are speed, image, and trajectory information, and the output is the future pedestrian trajectory (B_{pred}). After supervised learning, it is expected that the output of speed information module will be the feature vector of the future vehicle speed, the output of the image information module will be the feature of pedestrian dynamics, the output of trajectory information module will be the feature of future pedestrian position, and finally, these outputs are concatenated and fed into the decoder unit, a Final LSTM block (the purple block in Figure 4.1). For each module, the LSTM Cell is applied to process the sequential information, and the FCN block is employed to generate each feature vector. Through this process, the model is able to generate a pedestrian trajectory prediction in a format of bounding box.

Furthermore, we have focused on implementing a residual function (Residual LSTM in Figure 4.1)[35] and an attention-based model when processing a sequential information. Based on these, we expect the model to efficiently learn the way of giving an attention to the past experiences, so that it can determine how much past information is related to the current timestamp. We have empirically shown that this process gives better results compared to the conventional RNN model. For the entire network, we extracted the image features from a pretrained VGG network and fed it to the convLSTM layers[36] with 64 filters, 64 hidden layers and 3 residual blocks for the residual network, and trained with the RMSprop optimizer.

4.2.2 Stakeholder 2 (Pedestrian-perspective) network

Compared to previous studies, the most important network in our model is the pedestrian-perspective network. This network is trained based on the information experienced by pedestrian, and embedded to enhance the robustness and accuracy in processing the information observed by a driver. It is rational that employing the data from different perspectives can increase the robustness and the accuracy of the entire network, and existing study related to the reinforcement learning [14] has shown this empirically. In addition, this can be also supported by the "mirror neuron" theory in neuroscience, as [16] reported that humans have been shown to possess viewpoint-invariant representations of objects and other agents [16, 15].

The ego-experience module consisting the stakeholder 2 network is pretrained in a supervised way to estimate the future position of pedestrian. For pretraining this module, we employ the pedestrian-perspective data which have been obtained from the VR experiment mentioned in Secion 2.1. The network receives the vehicle speed information $s_{past} = \left\{s_i^{t-w}, s_i^{t-w+1}, \ldots, s_i^t\right\}$, the distance between the vehicle and pedestrian $d_{past} = \left\{d_i^{t-w}, d_i^{t-w+1}, \ldots, d_i^t\right\}$, and the head orientation information of pedestrian $o_{past} = \left\{o_i^{t-w}, o_i^{t-w+1}, \ldots, o_i^t\right\}$ as an input. This network is trained to generate the future position of the pedestrian P_{pred} by learning distribution of $p\left(P_{pred}|s_{past}, d_{past}, o_{past}\right)$. In the network, LSTM and FCN block are used to process sequential information, and activity regularize unit is employed to generate an internal representation of raw observations in the neural network. After training, we argue that the feature vector extracted from the last layer represents the feature space of the circumstances of vehicle-pedestrian interaction.

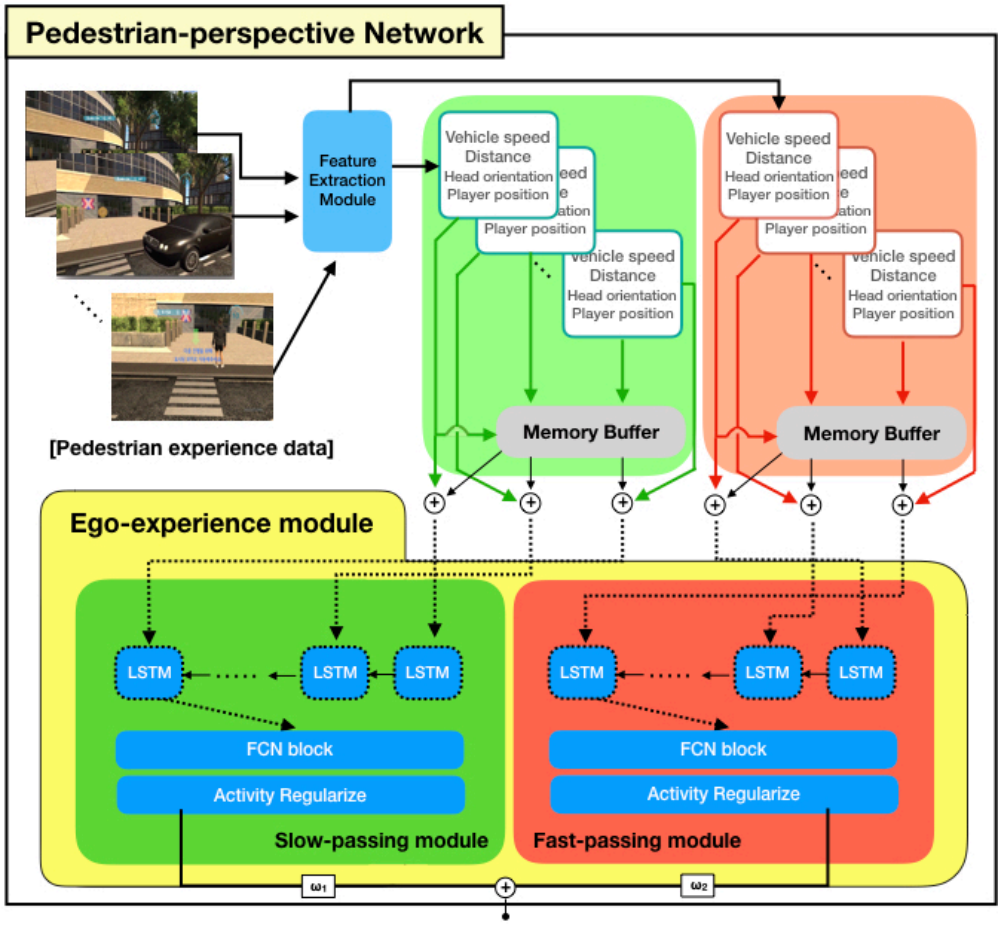


Figure 4.2: The overall structure of the stakeholder 2 (pedestrian-perspective) network and pretraining procedure of the ego-experience module. Note that dashed lines in ego-experience module are activated only when pretraining process, and are deactivated after the network is implemented in the MSPM.

While pretraining, since our dataset gathered from the VR experiment can be divide into two groups as mentioned in Section 2.1, which are slow-passing and fast-passing group, we have built two modules (slow-passing module and fast-passing module) with memory buffer as seen in Figure 4.2. Our network

combines the past information extracted from the memory buffer with current information and exploits them to generate solid and contextual information. Regarding the memory buffer, existing researches have shown that memorizing the past information and combining it with current information can increase the performance of the entire network [37, 38, 39]. The outputs from each module are finally converged with the weight (w_1, w_2) according to the ratio from the analysis of Gaussian distribution shown in Figure 2.1. Through pretraining the stakeholder 2 network in a supervised way, the output from this module is a high dimensional feature vector which is fed to the last fully connected layer to generate the P_{pred}.

In practice, we use an LSTM unit with 64 hidden units, and an activity regularization unit in Tensorflow, whose parameter of l_1=0.0001, l_2=0.0002. In addition, we divided the data from VR experiment into 200ms timestamp and used the MSE loss function. All hyper-parameter value has been empirically found. We also used the MSE loss function for pretraining.

4.2.3 For mobile robot experiment

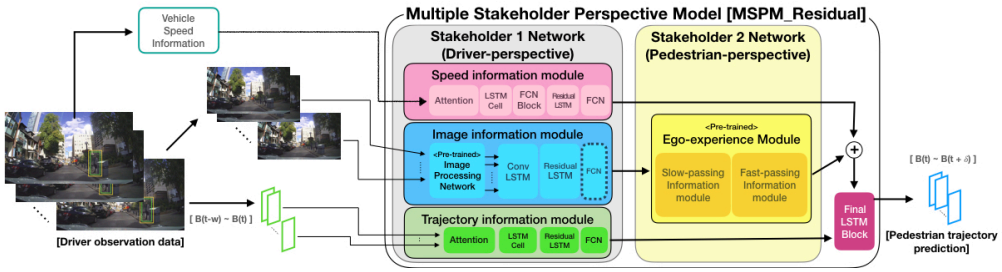


Figure 4.3: Illustration of MSPM_R with reinforced residual structure than a previous MSPM structure

In end-to-end learning, which has been mainly studied so far, we proposed an MSPM model structure that improved efficiency by modularizing information. To this end, rather than simply creating a model for learning, pre-training was performed that embodies the cognitive structure of real people. In this mobile robot experiment, some modifications were made in the previous experiment. The motivation of this model was focused on the fact that the more memorable data influenced the learning/test results more than the person's driving experience (data) sequentially. In the study to be described later, an experiment was conducted by designing a model reinforcing the residual structure (ex. Resnet structure) as shown in Figure 4.3.

Chapter 5

Evaluation & Result

5.1 Pedestrian intention estimation on automobile dataset

5.1.1 Trajectory Prediction

The prediction performance of the proposed MSPM is assessed using the pedestrian intention estimation (PIE) dataset[2]. This dataset provides information about the scene image, the past trajectory and the speed of the vehicle. We use this dataset because this dataset is the most recent open dataset on estimation of pedestrian intention. Some of previous studies to estimate pedestrian intentions[40, 4] used the JAAD dataset[41, 42]. But as the authors of the PIE dataset who also deployed the JAAD dataset indicated that the PIE dataset includes more features than the JAAD dataset and reflects more diverse environments with annotations[2]. Therefore in this paper, we evaluate our model through the PIE dataset.

Of course, the VR environment for pretraining the stakeholder 2 network is not exactly the same as the environment of [2] which we use for comparison experiment. However, we use both the information from the VR experiment and pedestrian intention estimation dataset[2] since the elements composing both dataset is identical. Also, we designed the traffic scene in our VR experiment

to be similar with the one in [2]. This utilization of two dataset improves the model's performance in the pedestrian intention estimation.

For the evaluation, we use the PIE dataset[2] with the image and the selection frame. The data has been resized to 224×224 and padded with zeros to keep the aspect ratio of VGGnet[43].

Method	**MSE-0.5s**	**MSE-1s**	**MSE-1.5s**	**C_MSE-1.5s**	**CF_MSE-1.5s**
Linear [2]	123	477	1365	950	3983
LSTM [2]	172	330	911	837	3352
B-LSTM [3]	101	296	855	811	3259
PIE_traj [2]	58	200	636	596	2477
MSPM [Ours] (%)	**57.80 (0.344%)**	**182.77 (8.615%)**	**565.15 (11.14%)**	**526.83 (11.60%)**	**2191.78 (11.51%)**

Table 5.1: Pedestrian trajectory prediction errors over varying future time steps. CMSE and CFMSE are the MSEs calculated over the center of the bounding boxes for the entire predicted sequence and only the last time step respectively. (%) means improvement percentage compared to previous state-of-the-art model.

As shown in Table 5.1, we evaluated our model with previous models, which are a linear Kalman filter[2] denoted as Linear, a vanilla LSTM model denoted as LSTM, a Bayesian LSTM[2, 3] denoted as B-LSTM, the previous state-of-the-art model[2] denoted as PIE_traj. Every model is trained and tested on 15 frames (0.5s) observation, and predicts the future trajectory of pedestrian on 30 frames (1.0s), 45 frames (1.5s).

As a result, our network is showing better results compared to the PIE_traj network[2], which is the state-of-the-art network. Our network predicts precise trajectory results at different times, and achieves greater accuracy from the cen-

ter point of the trajectory bounding box as shown in Table 5.1. Above all, as the estimated time goes from 0.5s (short term) to 1.5s (long term), our network has predicted accurately compared to other networks. In case of predicting pedestrian trajectory, 1.0s is considered as long term prediction. Even if a prudent and conservative driver is driving at a speed of 40km/h in a residential area, the distance covered in 1 second roughly corresponds to the braking distance. The anticipation of traffic scenes in a time horizon of at least 1 second would therefore enable safe driving at such speeds[3]. While short term prediction can be done by relatively small networks by learning the information presented, for long term prediction it is important to consider more variables and situations. From this point of view, our network displays a more robust and precise prediction in a test environment. This is also reflected in the central value of the loss results. The central value is the point where the person actually exists in the bounding box. In an autonomous vehicle, the system can see not only the front view of the pedestrian, but also the side and rear view. It is therefore important to understand in particular which point to look inside the information in the bounding box. From this point of view, the central point is the crucial point. As the result showed, our network is good at predicting the average central position of the timestamp and the last timestamp.

Figure 5.1: Example of predicted trajectory using the proposed model (MSPM) and previous state-of-the-art model (PIE_traj)[2]. Implement reference time data and predict the path of the pedestrian through different times. Each color of bounding box means: ground truth (green), MSPM (blue), and PIEtraj (red). Each interaction situation is: #1. A woman and child are crossing, #1-2. Same situation in #1-1, focused on a woman who walks on the right side, #2. A man passing in front of the car.)

In addition, as shown in Figure 5.1, our model can make a robust and precise trajectory prediction compared to the previous model[2]. Especially, we found that our model can even follow the child's movement and correct the prediction of false negatives (both in the case of Interaction #1-1 in Figure 5.1). This false negative prediction, which means that the pedestrian crosses while the model predicts not to cross, can represent a serious danger when the autonomous

vehicle system adapts to society. Thus, these characteristics, which predict well in short-term sequences and also in long-term sequences, while avoiding false negatives can give better information to the system to deal with a complex pedestrian-vehicle interaction.

5.1.2 Ablation Study

Method	MSE-0.5s	MSE-1s	MSE-1.5s	C_MSE-1.5s	CF_MSE-1.5s
MSPM (Ours)	**57.80**	**182.77**	**565.15**	**526.83**	**2191.78**
MSPM - condition 1	59.10	188.53	587.84	549.55	2282.07
MSPM - condition 2	59.50	187.09	580.35	540.27	2231.40
MSPM - condition 3	60.18	191.57	594.40	554.68	2278.58

Table 5.2: Pedestrian trajectory prediction errors over removing part of the MSPM (Condition 1: MSPM without the stakeholder 2 (pedestrian-perspective) network, Condition 2: MSPM without using head orientation information when training the stakeholder 2 network, Condition 3: MSPM with concatenating slow/fast-passing module in the stakeholder 2 network).

We also conducted an experiment by removing some part of the suggested model. For each condition, we remove the stakeholder 2 (pedestrian-perspective) network (condition 1), head orientation information while pretraining the stakeholder 2 network (condition 2). Also, we concatenate slow/fast-passing module in the stakeholder 2 network (condition 3). As shown in Table 5.2, the loss of the pedestrian trajectory prediction increased compared to the original MSPM, and in particular, the margin of the loss value is higher in the long-term prediction compared to the short-term prediction.

Condition 1 By default, our MSPM model can work only with the stakeholder 1 network without the pedestrian experience data and the stakeholder 2 network. As shown in Table 5.1 and 5.2, even with the stakeholder 1 network, our model can show better performance than the previous model because our model has advantage of reinforcing the residual structure through data pre-processing and network architecture. However, there is a limit to the performance improvement, and the best performance can be achieved after combining the stakeholder 2 network.

Condition 2 After removing head orientation information of pedestrian during pretraining of the stakeholder 2 network, the loss increases. This result can show that the information of head orientation can improve the performance of predicting the future position of the pedestrian.

Condition 3 In addition, after the convergence of two groups (fast-passing group and slow-passing group) in the stakeholder 2 network, the loss of future trajectory increased. This means that the separation of the pedestrian behavior model and its application across the network is important for the future estimation of the trajectory.

5.2 Pedestrian intention estimation in indoor mobile robots dataset

5.2.1 A study on how to effectively make predictions in novel indoor situations based on models learned using only previous datasets

First, a model was trained using a dataset that measured pedestrian behavior using a vehicle, and the model was applied to the newly collected indoor model

and tested. Previously, the dataset collected using automobiles was the recently released dataset and the PIE_dataset[2], which has the most related features. Of course, there are existing datasets with pedestrian behavior such as the JAAD dataset and Stanford dataset, but there are no datasets that have as diverse as the PIE_dataset and open-sourced feature information. Therefore, in this study, a model was trained through this PIE_dataset. PIE_dataset is largely composed of 6 sets, and in this study, training and validation were performed with 5 training sets and 1 validation set. There are two models trained, the PIE_traj model that previously showed state-of-the-art performance and the MSPM_residual model that was studied this time. The results of training these two models and testing the indoor dataset are as follows in Table 5.3.

	PIE_traj	**MSPM_R (ours)**
MSE-0.5s	181.99	**121.99**
MSE-1.0s	633.98	**355.56**
MSE-1.5s	1517.6	**949.5**
C_MSE-1.5s	1413.66	**834.09**

Table 5.3: Compare between previous state-of-the-art model and our model without training indoor dataset. Each second mean prediction time.

As shown from the MSE and the C_MSE (center value of MSE), our model showed better performance than the previous model that showed the highest performance. In particular, it can be seen that the error is greatly reduced not only in short-term predictions such as 0.5 seconds but also in long-term predictions such as 1.0 seconds and 1.5 seconds. In automobile research, predicting the future of 1.0 seconds or more is defined as a long-term prediction. Therefore, it can be confirmed that this model made a more robust and precise prediction.

	PIE_traj	**MSPM_R (ours)**
F1 score-0.5s	0.4449	**0.4576**
F1 score-1.0s	0.4141	**0.4264**
F1 score-1.5s	0.3829	**0.3931**

Table 5.4: F1 score for each model. Each second mean prediction time.

Based on the above result, the result of obtaining the F1 Score for all the tested cases is as follows. The F1 Score used in this evaluation is a form of IoU and is a metric that is widely used in vision research. After calculating the ground-truth bounding box and the predicted bounding box in pixel units, the area of the intersection area where the two bounding boxes overlap is investigated. After that, the value obtained by dividing the intersection area x 2 by the sum of the area of the ground-truth bounding box and the prediction bounding box is defined as an F1 score.

Unlike conventional IoU, in pedestrian prediction with variable bounding box size, the F1 score was investigated by determining that the F1 score metric could better reflect the model prediction. Also in the F1 score, as shown in the Table 5.4, it can be seen that the model we studied shows better performance. In particular, not only the overall F1 score increased, but also more cases were confirmed when the detection was evaluated based on a high threshold. This means that the detection performance is improved overall, rather than being well detected in some special cases and not well detected in the rest.

Overall, based on the results of this study, it was confirmed that it is important to use the same shape as my model and pretrain with appropriate data for more robust and stable prediction than the existing model.

5.2.2 A study on what prediction result comes out when the newly collected data is also trained

In general, in deep learning, if the size of the network is sufficient, the more data is collected and trained, the higher the accuracy of prediction. Of course, the overfitting issue caused by the network memorizing data features in this process inevitably arises. Therefore, it is important to properly select a dataset for pre-training and training and to proceed with learning based on this. In this evaluation, not only the previously collected dataset but also the newly collected dataset were trained and the results were analyzed. If this dataset, which has features different from the existing PIE_dataset, has a positive influence on the network, the network will be able to learn to handle more robust features. However, if this dataset gives more fluctuation to the existing network weight, the learning performance will decrease.

Metric	PIE_traj	**MSPM_R(ours)**
MSE-0.5s	106.79	**88.72**
MSE-1.0s	291.87	**270.02**
MSE-1.5s	731.86	**774.21**
C_MSE-1.5s	584.51	**640.64**
F1 score-0.5s	0.5811	**0.5905**
F1 score-1.0s	0.5817	**0.5936**
F1 score-1.5s	0.5601	**0.5673**

Table 5.5: MSE and F1 score for each model.

In this study, 5 training sets, 2 validation set, and 1 test set were used. In

the training set, the 4 sets were from existing PIE_dataset, and 1 set from this dataset were used, the validation was PIE_dataset, and the test was conducted with this dataset. Of course, the training and test set were separated in advance so that they were not mixed. The result of this measurement can be shown in the Table 5.5.

There are two things that can be known based on this result. First, we compared the MSE and F1 scores of PIE_traj and MSPM_R shown in this table. My model showed better MSE values in both short-term prediction and long-term prediction than existing PIE_traj. In particular, it can be seen that the long-term prediction showed excellent results. Also, when we collected the results for all cases through the F1 score, my model showed better performance. Hence, we can confirm that my model enables more robust prediction.

The second thing that can be confirmed through this result is a comparison with the previous evaluation 1. In Evaluation 1, the indoor dataset measured this time was not trained and the test was conducted directly. However, as in evaluation 2, if training is performed with the indoor dataset, it can be confirmed that the prediction performance is improved. From this result, it was confirmed that the indoor dataset has different features from the dataset measured in automobiles such as the existing PIE_dataset, and therefore, to apply this in an indoor environment, it was confirmed that a separate dataset must be collected and trained. Also, the indoor dataset is a very harsh dataset. This is because vibration and fluctuation were very severe in the measurement process because it was data collected through a mobile robot, and this was caused by blur images and the like. Even if it is such harsh data, it was confirmed that even better results were obtained by training by splitting some data.

Threshold	Prediction time	PIE_traj	MSPM_R(ours)
F1 score = 0.6	0.5 second after	59.77%	**61.24%**
	1.0 second after	56.38%	**59.80%**
	1.5 second after	51.86%	**54.56%**
F1 score = 0.65	0.5 second after	51.86%	**54.46%**
	1.0 second after	47.03%	**51.19%**
	1.5 second after	43.18%	**46.31%**
F1 score = 0.7	0.5 second after	43.50%	**45.71%**
	1.0 second after	38.16%	**41.38%**
	1.5 second after	34.80%	**37.48%**
F1 score = 0.75	0.5 second after	33.05%	**36.10%**
	1.0 second after	28.98%	**31.69%**
	1.5 second after	26.52%	**28.93%**
F1 score = 0.8	0.5 second after	22.88%	**26.27%**
	1.0 second after	19.92%	**22.80%**
	1.5 second after	17.78%	**20.73%**
F1 score = 0.85	0.5 second after	14.97%	**16.78%**
	1.0 second after	12.49%	**13.67%**
	1.5 second after	10.56%	**12.52%**
F1 score = 0.9	0.5 second after	8.02%	**9.10%**
	1.0 second after	5.85%	**6.53%**
	1.5 second after	4.39%	**5.18%**

Table 5.6: The success ratio measured after determining the threshold based on the F1 score and determining that it was detected when it exceeds this threshold.

This time, the F1 score threshold was set, and if it exceeded this threshold, it was judged as detect. In that case, the detect ratio according to the threshold is shown in Table 5.6. It can be seen that not only the average detection success rate is high, but also more distributions in the section where the F1 score is higher than the existing model.

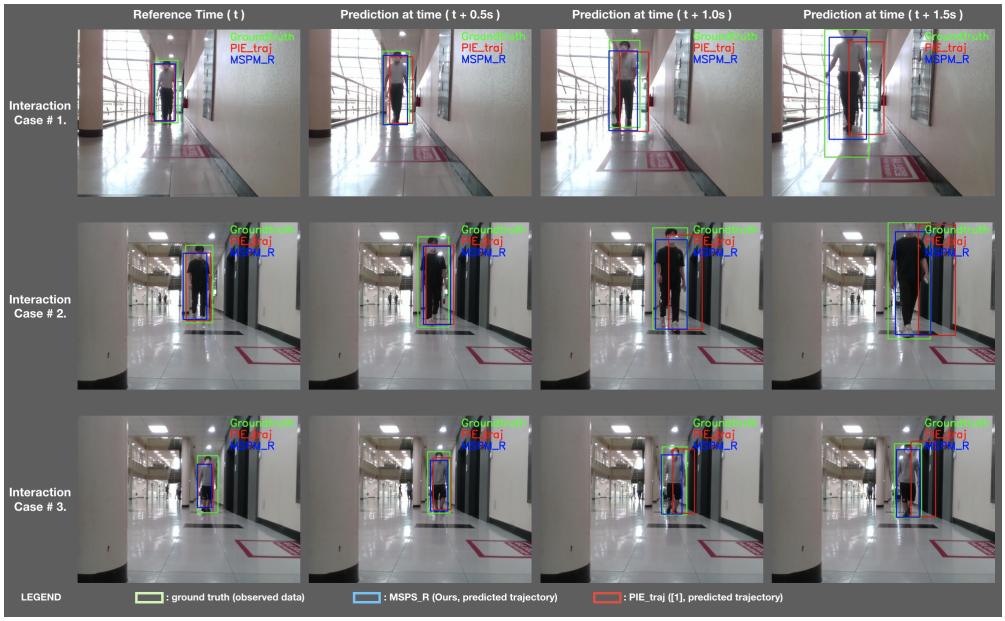


Figure 5.2: Example of predicted trajectory using the proposed model (MSPM_R) and previous state-of-the-art model (PIE_traj)[2]. Each color of bounding box means: ground truth (green), MSPM_R (blue), and PIE_traj (red). Each interaction situation is: #1. A man is walking along the narrow road, #2. A man is moving forward, #3. A man is walking with slow speed.

These results can also be confirmed through the trajectory of the Figure 5.2. The previous model is unable to follow complex human movements as they become longer-term, but we can confirm that our model can predict pedestrian trajectory with robustness.

Chapter 6

Conclusion

In this paper, we have proposed a model of a combined network that can reflect the perspective of both the pedestrian and the driver. Our model has shown cutting-edge results to predict the future trajectory of pedestrian behavior. Above all, our model has an advantage in detecting the long-term (1.5s) future trajectory. This indicates that our model can provide better information to the autonomous vehicle system in the event of unexpected pedestrian behavior or complicated pedestrian-vehicle interactions. From an ablation study removing the ego-experience module from the suggested network, we empirically found that the loss of trajectory increased an average of 3.14% after the stakeholder 2 network was removed, which can be interpreted that the data collected from the VR experiment have a similar context with information on classic pedestrian datasets like [2], and this affected the whole network in a positive way. In addition, we found that head orientation data is crucial for improving the performance of the pedestrian trajectory estimation. We collected data in an indoor environment using a mobile robot. Existing open-source data is mainly measured on roads. However, in order for a mobile robot or a delivery robot to work in an indoor environment, a dataset that can be trained in this environment is required. In this study, pipelines and features required for these datasets were designed, collected, and even trained. In the process of training

and testing in different robot platforms, it was confirmed that a model with a residual structure can make predictions more effectively. Consequently, since the pedestrian dataset can be further improved by accounting culture factors and more complex factors related to the current traffic system, we will have future work on this.

국문초록

본 학위논문에서는 차량의 관점에서 관찰되는 미래 보행자 궤적을 예측하는 다중 이해관계자 관점 모델(MSPM)을 제안한다. MSPM은 운전자가 보행자를 마주했을때, 자신이 보행자였을때의 경험을 되살린다는 점에서 착안하여 설계되었다. 지금까지 자율주행 관련 연구는 차량-차량 시스템에 초점을 맞춰 진행되었다. 때문에 고속도로에서의 자율주행자동차는 어느 정도 완성에 가까워지고 있지만, 도심으로 들어오려면 차량-보행자 상호작용의 연구가 필요하다. 이번 논문에서는 차량-보행자 상호작용에서 보행자 의도 파악을 핵심 요인으로 보고 연구를 진행하였다. 차량-보행자 상호작용은 서로 다른 2개의 개체(운전자, 보행자) 사이에서 이루어짐에도 불구하고, 지금까지의 연구는 운전자의 관점에서만 데이터에 의해 훈련된 신경망 개발에 초점을 맞추고 있다. 본 논문에서는 다중 이해관계자 관점 모델(MSPM)을 제안하고 보행자 의도 예측에 이 모델을 적용한다. 이를 위해 심리학 연구실에서 가상현실을 통해 측정한, 보행자 관점에서 자동차를 바라보았을때의 데이터를 받아서 분석한 뒤 네트워크에 반영하였다. 제안하는 모델은 기존 보행자 의도파악 데이터셋에서 최고 성능을 달성하는 동시에 오차율을 단기(0.5초)와 중기(1.0초) 예측에서 평균 4.48%, 장기 예측(1.5초)에서 기존 최고성능 모델 대비 11.14% 줄일 수 있었다. 또한 이번 논문에서는 실내 환경에서의 모바일 로봇에 필요한 데이터와 모델을 제안한다. 기존 연구에서는 이러한 형태의 데이터가 없었기에, 파이프라인을 구축하고 사람들의 다양한 행동 방식을 포함하여 실내 보행자 데이터 세트를 수집했다. 그리고 이러한 데이터와 모델을 토대로, 모델을 학습시키고 기존 모델보다 더 좋은 성능을 확인했다. 이번 학위논문에서 진행한 연구는 추후 서로 다른 로봇 플랫폼에서 수집된 데이터를 손쉽게 로봇에 적용할 수 있는 모델 연구로 이어질 수 있을 것이다.

주요어: 사람 행동예측, 자율주행자동차, 모바일로봇, 사람-로봇 상호작용, 인지모델링

학번: 2019-29337

Bibliography

[1] S. M. R. C. P. Ltd, "Autonomous vehicles - global market outlook (2017-2026)," 2019.

[2] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6262–6271.

[3] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4194–4202.

[4] Z. Fang and A. M. López, "Is the pedestrian going to cross? Answering by 2d pose estimation," in *IEEE Intell. Veh. Symp.(IV)*, 2018, pp. 1271–1276.

[5] Z. Fang, D. Vázquez, and A. M. López, "On-board detection of pedestrian intentions," *Sensors*, vol. 17, no. 10, p. 2193, 2017.

[6] C. G. Keller, C. Hermes, and D. M. Gavrila, "Will the pedestrian cross? probabilistic path prediction based on learned motion features," in *Joint. Patt. Recogn. Symp.* Springer, 2011, pp. 386–395.

[7] L. Chen, J. Lu, Z. Song, and J. Zhou, "Part-activated deep reinforcement learning for action prediction," in *Proc. Euro. Conf. Comp. Vis.*, 2018, pp. 421–436.

[8] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[9] Y. Kong, Z. Tao, and Y. Fu, "Deep sequential context networks for action prediction," in *Proc. IEEE Conf. Comp. Vis. Pat. Rec.*, 2017, pp. 1473–1481.

[10] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Joint attention in autonomous driving (jaad)," *arXiv preprint arXiv:1609.04741*, 2016.

[11] D. Lee, M.-H. Yang, and S. Oh, "Head and body orientation estimation using convolutional random projection forests," *IEEE. Trans. Pat. Analy. Machine. Intell.*, vol. 41, no. 1, pp. 107–120, 2017.

[12] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *ECCV*, 2014, pp. 618–633.

[13] H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee, "Intention-aware online pomdp planning for autonomous driving in a crowd," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2015, pp. 454–460.

[14] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *IEEE Int. Conf. Robot. Autom.*, 2018, pp. 1134–1141.

[15] G. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annu. Rev. Neurosci.*, vol. 27, pp. 169–192, 2004.

[16] V. Caggiano, L. Fogassi, G. Rizzolatti, J. K. Pomper, P. Thier, M. A. Giese, and A. Casile, "View-based encoding of actions in mirror neurons

of area f5 in macaque premotor cortex," *Current Biology*, vol. 21, no. 2, pp. 144–148, 2011.

[17] S. Malla, B. Dariush, and C. Choi, "Titan: Future forecast using action priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 186–11 196.

[18] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Pedestrian detection: The elephant in the room," *arXiv preprint arXiv:2003.08799*, 2020.

[19] V. Y. Tsvetkov, "Cognitive information models," *Life Science Journal*, vol. 11, no. 4, pp. 468–471, 2014.

[20] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behav. Brain Sciences*, vol. 1, no. 4, pp. 515–526, 1978.

[21] S. Baron-Cohen, "Mindblindness: An essay on autism and theory," *Mind*, 1995.

[22] N. C. Rabinowitz, F. Perbet, H. F. Song, C. Zhang, S. Eslami, and M. Botvinick, "Machine theory of mind," *arXiv preprint arXiv:1802.07740*, 2018.

[23] S. Sturgeon, A. Palmer, J. Blankenburg, and D. Feil-Seifer, "Perception of social intelligence in robots performing false-belief tasks," in *IEEE Int. Conf. RO-MAN*, Oct 2019, pp. 1–7.

[24] A. Tabrez, S. Agrawal, and B. Hayes, "Explanation-based reward coaching to improve human performance via reinforcement learning," in *ACM/IEEE Int. Conf. HRI*, 2019, pp. 249–257.

[25] Y. Demiris, "Prediction of intent in robotics and multi-agent systems," *Cognitive processing*, vol. 8, no. 3, pp. 151–158, 2007.

[26] R. Raileanu, E. Denton, A. Szlam, and R. Fergus, "Modeling others using oneself in multi-agent reinforcement learning," *arXiv preprint arXiv:1802.09640*, 2018.

[27] J. Jara-Ettinger, "Theory of mind as inverse reinforcement learning," *Current Opinion in Behavioral Sciences*, vol. 29, pp. 105–110, 2019.

[28] S. Shamsuddin, H. Yussof, L. Ismail, F. A. Hanapiah, S. Mohamed, H. A. Piah, and N. I. Zahari, "Initial response of autistic children in human-robot interaction therapy with humanoid robot nao," in *IEEE. Int. Col. Signal Processing. Appl.*, 2012, pp. 188–193.

[29] A. Rasouli and J. K. Tsotsos, "Joint attention in driver-pedestrian interaction: from theory to practice," *arXiv preprint arXiv:1802.02522*, 2018.

[30] D. B. Miller and W. Ju, "Joint cognition in automated driving: Combining human and machine intelligence to address novel problems," in *AAAI Spring Symposium Series*, 2015.

[31] N. Guéguen, S. Meineri, and C. Eyssartier, "A pedestrian's stare and drivers' stopping behavior: A field experiment at the pedestrian crossing," *Safety science*, vol. 75, pp. 87–89, 2015.

[32] Y.-C. Lee, J. D. Lee, and L. Ng Boyle, "The interaction of cognitive load and attention-directing cues in driving," *Human factors*, vol. 51, no. 3, pp. 271–280, 2009.

[33] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in *IEEE Trans. Intell. Veh.(IV)*, 2017, pp. 264–269.

[34] S. Schmidt and B. Faerber, "Pedestrians at the kerb–recognising the action intentions of humans," *Transportation research part F: traffic psychology and behaviour*, vol. 12, no. 4, pp. 300–310, 2009.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comp. Vis. Pat. Rec.*, 2016, pp. 770–778.

[36] C. Lu, M. Hirsch, and B. Scholkopf, "Flexible spatio-temporal networks for video prediction," in *Proc. IEEE Conf. Comp. Vis. Pat. Rec.*, 2017, pp. 6523–6531.

[37] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.

[38] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, 2015, pp. 802–810.

[39] R. Liu and J. Zou, "The effects of memory replay in reinforcement learning," in *IEEE. Ann. Allerton. Conf. Commu. Cont. Comp.(Allerton)*, 2018, pp. 478–485.

[40] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *arXiv preprint arXiv:2002.08945*, 2020.

[41] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 206–213.

[42] ——, "It's not all about size: On the role of data properties in pedestrian detection," in *Euro. Conf. Comp. Vis. Workshop*, 2018.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[44] N. Rauh, T. Franke, and J. F. Krems, "Understanding the impact of electric vehicle driving experience on range anxiety," *Human factors*, vol. 57, no. 1, pp. 177–187, 2015.

[45] H. S. Kim, Y. Hwang, D. Yoon, W. Choi, and C. H. Park, "Driver workload characteristics analysis using eeg data from an urban road," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 4, pp. 1844–1849, 2014.

[46] K. A. Brookhuis and D. De Waard, "Monitoring drivers' mental workload in driving simulators using physiological measures," *Accident Analysis & Prevention*, vol. 42, no. 3, pp. 898–903, 2010.

[47] J. D. Hill and L. N. Boyle, "Driver stress as influenced by driving maneuvers and roadway conditions," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 10, no. 3, pp. 177–186, 2007.

[48] M. Ringhand and M. Vollrath, "Effect of complex traffic situations on route choice behaviour and driver stress in residential areas," *Transportation research part F: traffic psychology and behaviour*, vol. 60, pp. 274–287, 2019.

[49] W. Chen, X. Zhuang, Z. Cui, and G. Ma, "Drivers' recognition of pedestrian road-crossing intentions: Performance and process," *Transportation research part F: traffic psychology and behaviour*, vol. 64, pp. 552–564, 2019.

[50] J. McComas, M. MacKay, and J. Pivik, "Effectiveness of virtual reality for teaching pedestrian safety," *CyberPsychology & Behavior*, vol. 5, no. 3, pp. 185–190, 2002.

[51] H. Luo, T. Yang, S. Kwon, M. Zuo, W. Li, and I. Choi, "Using virtual reality to identify and modify risky pedestrian behaviors amongst chinese children," *Tra. Inj. Pre.*, vol. 21, no. 1, pp. 108–113, 2020.

[52] D. Whitney, E. Rosen, D. Ullman, E. Phillips, and S. Tellex, "ROS reality: A virtual reality framework using consumer-grade hardware for ros-enabled robots," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 07 2018.

[53] E. Matsas, G.-C. Vosniakos, and D. Batras, "Prototyping proactive and adaptive techniques for human-robot collaboration in manufacturing using virtual reality," *Robot. Comp. Manu.*, vol. 50, pp. 168–180, 2018.

[54] D. Freeman, S. Reeve, A. Robinson, A. Ehlers, D. Clark, B. Spanlang, and M. Slater, "Virtual reality in the assessment, understanding, and treatment of mental health disorders," *Psychological medicine*, vol. 47, no. 14, pp. 2393–2400, 2017.

[55] S. Nijman, W. Veling, C. Geraets, A. Aleman, and G. M. Pijnenborg, "Su104. reading the mind of the avatar: Dynamic interactive social cognition virtual reality training (discovr) for people with a psychotic disorder," *Schizophrenia bulletin*, vol. 43, no. Suppl 1, p. S198, 2017.

[56] C. A. Zanbaka, A. C. Ulinski, P. Goolkasian, and L. F. Hodges, "Social responses to virtual humans: implications for future interface design," in *Proc. SIGCHI. Conf. Hum. Fac. Comp. Syst.*, 2007, pp. 1561–1570.

[57] M. Sucha, D. Dostal, and R. Risser, "Pedestrian-driver communication and decision strategies at marked crossings," *Accident Analysis & Prevention*, vol. 102, pp. 41–50, 2017.

[58] K. Mahadevan, S. Somanath, and E. Sharlin, "Communicating awareness and intent in autonomous vehicle-pedestrian interaction," in *Proc. CHI. Conf. Human Factors. Comp. Syst.*, 2018, pp. 1–12.

[59] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Understanding pedestrian behavior in complex traffic scenes," *IEEE Trans. Intell. Veh.*, vol. 3, no. 1, pp. 61–70, 2017.

[60] D. Rothenbücher, J. Li, D. Sirkin, B. Mok, and W. Ju, "Ghost driver: A field study investigating the interaction between pedestrians and driverless vehicles," in *IEEE. Int. Symp. RO-MAN.* IEEE, 2016, pp. 795–802.

[61] A. Nematzadeh, K. Burns, E. Grant, A. Gopnik, and T. L. Griffiths, "Evaluating theory of mind in question answering," *arXiv preprint arXiv:1808.09352*, 2018.

[62] S. Narang, A. Best, and D. Manocha, "Inferring user intent using bayesian theory of mind in shared avatar-agent virtual environments," *IEEE. Trans. Vis. Comp. Graphics.*, vol. 25, no. 5, pp. 2113–2122, 2019.

[63] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 961–971.

[64] T. Bandyopadhyay, K. S. Won, E. Frazzoli, D. Hsu, W. S. Lee, and D. Rus, "Intention-aware motion planning," in *Algorithmic foundations of robotics X*. Springer, 2013, pp. 475–491.

[65] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3457–3464.

[66] W. Byeon, Q. Wang, R. Kumar Srivastava, and P. Koumoutsakos, "Contextvp: Fully context-aware video prediction," in *Proc. Euro. Conf. Comp. Vis. (ECCV)*, 2018, pp. 753–769.

[67] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 1468–1476.

[68] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE. Trans. Pat. Analy. Machine. Intell.*, vol. 34, no. 4, pp. 743–761, 2011.

[69] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 336–345.

[70] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Computer graphics forum*, vol. 26, no. 3. Wiley Online Library, 2007, pp. 655–664.

[71] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion gan for future-flow embedded video prediction," in *Proc. IEEE Conf. Comp. Vis. Pat. Rec.*, 2017, pp. 1744–1752.

[72] D. Lee, M.-H. Yang, and S. Oh, "Fast and accurate head pose estimation via random projection forests," in *Proc. IEEE. Int. Conf. Comp. Vis.*, 2015, pp. 1958–1966.

[73] D. Lee, G. Cha, M.-H. Yang, and S. Oh, "Individualness and determinantal point processes for pedestrian detection," in *Euro. Conf. Comp. Vis.* Springer, 2016, pp. 330–346.

[74] Tzutalin, "Labelimg," https://github.com/tzutalin/labelImg, 2015.