2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

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

Kyungdo Kim*[1], Yoon Kyung Lee*[2], Hyemin Ahn[1], Sowon Hahn[2] and Songhwai Oh[1]

***Abstract*— This paper proposes a multiple stakeholder perspective model (MSPM) which predicts the future pedestrian trajectory observed from vehicle's point of view. 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 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 pedestrian's perspective have been collected from the virtual reality experiment, and a network that can reflect 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 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.**

I. INTRODUCTION

Global autonomous vehicles market accounted is expected to reach $615.02 billion by 2026 growing at a compound annual growth rate of 41.5% during the period [1]. Now it is inevitable to bring autonomous vehicle in our traffic system. However, bringing autonomous vehicle to our society would cause several issues, such as object recognition error in driving situation since it is controlled by an intelligent system which handles interactions among vehicles, drivers and pedestrians. Among these interaction problems, we point out that existing works do not focus on vehicle-pedestrian interaction even if it is essential when autonomous cars move away from the motorway-centered system and go to the city center where encounter with pedestrians occur frequently.

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]–[5] extracted trajectory information from the scene images using deep neural networks. 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.

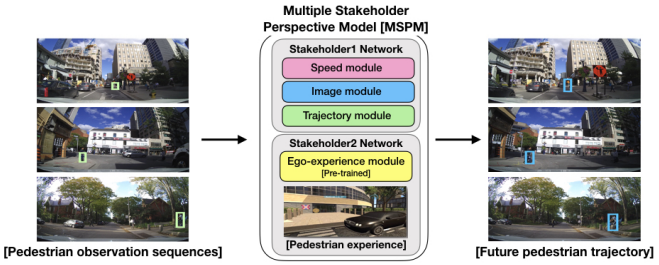

Fig. 1: An overview of the proposed Multiple Stakeholder Perspective Model (MSPM) for the pedestrian trajectory prediction. This model considers both information from the driver's and the pedestrian's point of view.

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 [6] and neuroscience [7], [8], 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.

Our main contributions 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 more reliable prediction is possible. Second, the proposed model estimates the pedestrian behavior using data collected in virtual reality system. Finally, our MSPM provides a cutting edge result in a recent pedestrian intention estimation dataset.

II. RELATED WORK

A. Pedestrian Intention Prediction

Recent works on action prediction have focused on the use of past and current scene information [9]. In these

*Equal Contribution
[1]Robot Learning Laboratory, Department of Electrical and Computer Engineering and ASRI, Seoul National University (e-mail: {kyungdo.kim, hyemin.ahn}@rllab.snu.ac.kr, songhwai@snu.ac.kr)
[2]Human Factors Psychology Laboratory, Department of Psychology, Seoul National University (e-mail: {yoonlee78, swhahn}@snu.ac.kr)

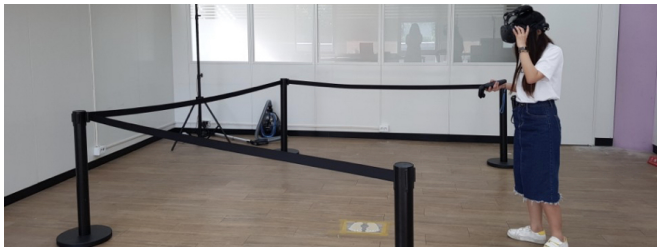

Fig. 2: The environment setting for the Virtual Reality (VR) experiment.

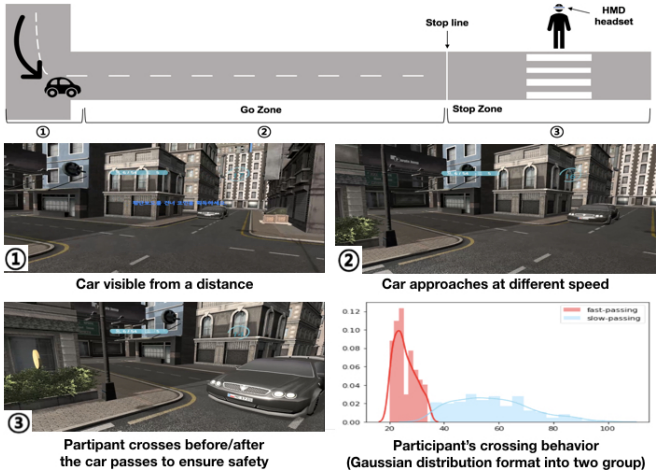

Fig. 3: Virtual Reality experiment scenario and participant's crossing behavior (In the graph of participant's crossing behavior, X axis denotes the timestamp and Y axis denotes the ratio of participants crossing during each timestamp).

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] reported that using the head orientation information can enhance the accuracy of this prediction. Since only depending on extracted image features can increase the noise, [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. 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.

B. Virtual Reality

Virtual reality (VR) has been used in psychotherapy, rehabilitation, and social skill training [12], since researchers can observe more realistic behavior in VR controlled environments where traditional IT methodologies could not offer [13]. Researchers also used VR for pedestrian safety education [14]. A recent approach using virtual reality has been applied to many fields such as psychology, communication, professional training and human-robot interaction [12].

III. Pedestrian Perspective Data Collection

We have built a pedestrian crossing scenario, in which participants interacted with an automated driving agent. The virtual reality scenario was created with Unity 2017. Participants wore HTC Vive Pro headphones as shown in Figure 2.

1) Preparation: Participants were informed of a brief description of the study. They were asked to check if they had ever suffered from nausea, illness and anxiety after a virtual reality experience. Participants have been advised that they can withdraw the experience at any time when they experience discomfort.

2) Scenario of Virtual Reality: As shown in Figure 3, 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 VR system for 3 times and we excluded this exercise from the analysis. 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. 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.

3) Survey: In the post-experience interview, participants rated the overall experience and the quality of the VR scenario. In particular, we included questions asking "Realism: how much they felt in the VR scenario", "Similarity: how much they walked in the VR scenario compared to their usual walking behavior", "Vehicle Speed Effect: how much the speed of a car affected their crossing behavior"

4) Data Extraction and Analysis: After the VR experiment, we extracted the information such as vehicle speed, distance between the pedestrian and vehicle, angle of head movement, and the position of the pedestrian for each timestamp. After analyzing this VR data, we found that the data can be divided into two groups-group with some people really care about the movement of the vehicle and move with caution, Otherwise, there are people 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 3). 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.

IV. Network Architecture

A. 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 [15]. In

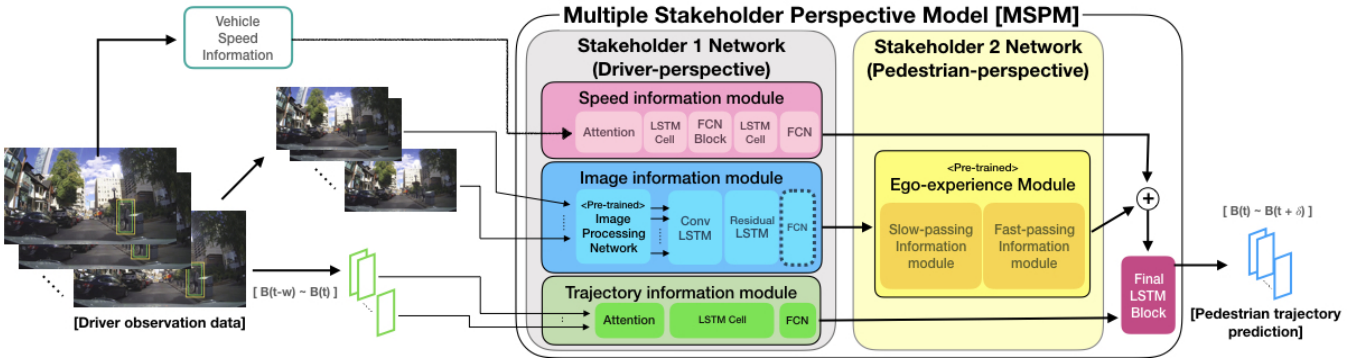

Fig. 4: Illustration of the stakeholder 2 network when it is in pretraining procedure. First, we pretrain an ego-experience module inside the stakeholder 2 (pedestrian-perspective) network. After training this network, the network is appended next to the stakeholder 1 (driver-perspective) network in the MSPM model. Note that the final FCN unit (dashed line) in image information module is deactivated when the stakeholder 2 network is implemented.

particular, human can build the *Theory of Mind* (ToM) model, which is a model of the physical and psychological states of others [16]. 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 [17], and recent works [18] 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 [19]. Existing works [20], [21] 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.

B. 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. 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]. 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, 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 $([(x_1, y_1), (x_2, y_2)])$. 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 $I \in \mathcal{I} \subset R^{n_i \times n_j \times 3}$ 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.

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] 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 III, we have empirically found that the vehicle speed is the key factor that affects pedestrian's crossing behavior as shown in Table I.

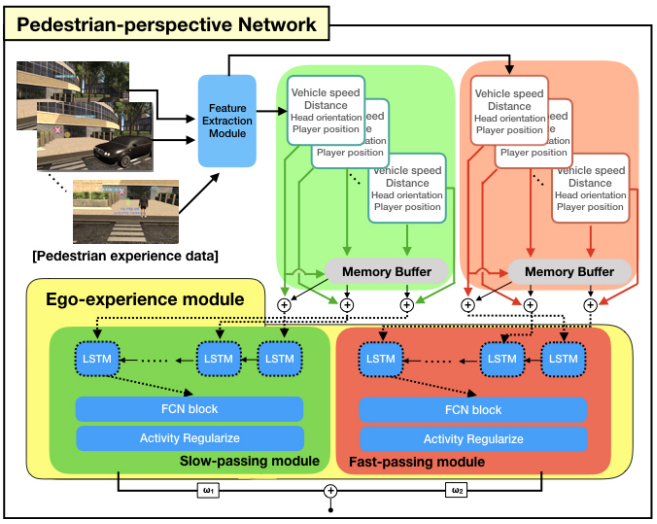

Fig. 5: The overall structure of the stakeholder 2 (pedestrian-perspective) network and pretraining procedure of the ego-experience module. Note that dashed lines in ego-experience module are activated only when pretraining process, and are deactivated after the network is implemented in the MSPM.

For network details, the speed information module (the pink block in Figure 4) receives the vehicle speed information (S_{obs}) as inputs, the image information module (the blue block in Figure 4) gets the image sequences (I_{obs}) as inputs, and the trajectory information module (the green block in Figure 4) 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). 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) [22] 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.

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 [6] has shown this empirically. In addition, this can be also supported by the "mirror neuron" theory in neuroscience, as [8] reported that humans have been shown to possess viewpoint-invariant representations of objects and other agents [7], [8].

Keywords	Survey Questions	Criteria	Mean Rating (Standard Deviation)
Realism	I found the virtual environment setting is very similar to that in real life	1: Not very unrealistic. 5: Very realistic.	**3.98 (0.80)**
Behavior Similarity	My crossing behavior in the experiment was similar to how I usually cross in real life	1: Not very similar. 5: Very similar.	**4.03 (0.81)**
Vehicle Speed Effect	Vehicle speed affected my crossing behavior during the experiment	1: Not at all. 5: Very much.	**4.05 (0.81)**

TABLE I: Survey result after VR experiment for pedestrian crossing behavior.

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 III. The network receives the vehicle speed information $s_{past} = \{s_i^{t-w}, s_i^{t-w+1}, \ldots, s_i^t\}$, the distance information between the vehicle and pedestrian $d_{past} = \{d_i^{t-w}, d_i^{t-w+1}, \ldots, d_i^t\}$, and the head orientation information of pedestrian $o_{past} = \{o_i^{t-w}, o_i^{t-w+1}, \ldots, o_i^t\}$ as an input. This network is trained to generate the future position of the pedestrian P_{pred} by learning distribution of $p(P_{pred}|s_{past}, d_{past}, o_{past})$. After training, we argue that the feature vector extracted from the last layer represents the feature space of the circumstances of vehicle-pedestrian interaction.

While pretraining, since our dataset gathered from the VR experiment can be divide into two groups as mentioned in Section III, 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 5. Our network combines the past information extracted from the memory buffer with current information and exploits them to generate solid and contextual information. 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 3. 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}.

V. Evaluation & Result

A. Pedestrian perspective experiment through Virtual Reality

After the participant ended the entire experiment session mentioned in Section III, we conducted a survey since the observed data on pedestrian behavior could give us information with limited intention. Through this survey, we were able to obtain a qualitative assessment of the data as shown in Table I. We wanted to know if the VR scenario was well designed to immerse them in the environment. Most participants reported that the VR settings felt realistic, rating an average score of 3.98 out of 5 (80%). When asked how similar their crossing behaviors was to what they had shown

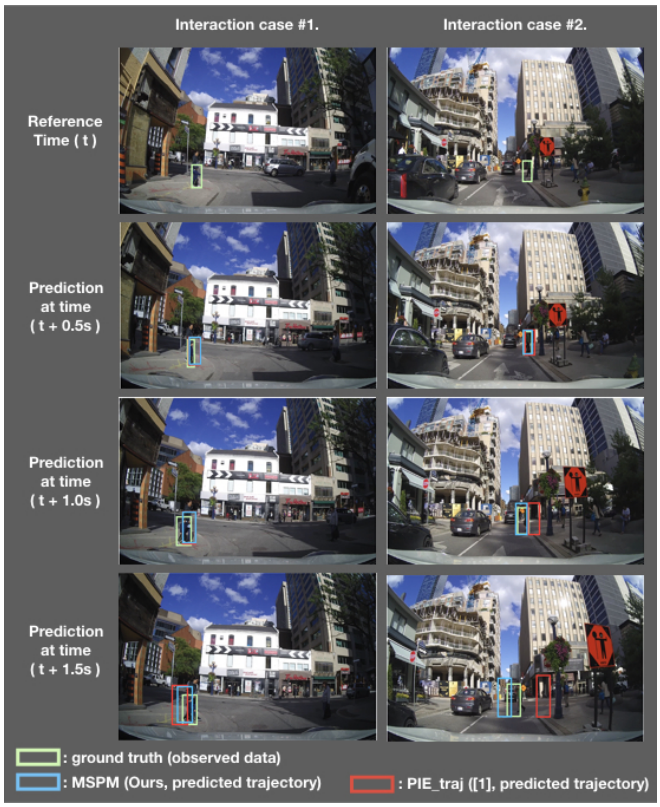

Fig. 6: Example of predicted trajectory using the proposed model (MSPM) and previous state-of-the-art model (PIE_traj) [2]. Implement reference time data and predict the path of the pedestrian through different times. Each color of bounding box means: ground truth (green), MSPM (blue), and PIEtraj (red). Each interaction situation is: #1. A woman and child are crossing, #2. A man passing in front of the car.)

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 II: 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.

in the VR experiment, participants rated 4.03 out of 5 (81%). Participants reported that VR scenario seemed a reproduction of a typical road in a real world that induced them to cross as if they were in a real crosswalk.

Participants also rated average of 4.05 out of 5 (81%) for the speed of a car to affect their crossing behaviors. Most participants in our study also felt the need to adjust their behaviors and decisions to cross or stop according to their judgment of vehicle speed. To avoid any cultural biases, we also included a replication of foreign roads and buildings.

B. Pedestrian Intention Estimation

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 [4] used the JAAD dataset [23], [24]. 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.

As shown in Table II, 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 on 15 frames (0.5s) observation, and predicts the future trajectory of a pedestrian on 15 frames (0.5s), 30 frames (1.0s) and 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 trajectories at different times, and achieves greater accuracy from the center point of the trajectory bounding box as shown in Table II. Above all, as the estimated time goes from 0.5s (short term) to 1.5s (long term), our network predicts accurately compared to other networks. In case of predicting pedestrian trajectory, 1 second 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, it is important to understand in particular which point to look inside the information in the bounding box. As the result showed, our network is good at predicting the average central position of the timestamp and the last timestamp.

VI. Ablation Study

We also conducted an experiment by removing some part of the suggested model. For each condition, we remove the stakeholder 2 network (condition 1), head orientation information while pretraining the stakeholder 2 network

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 III: 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).

(condition 2). Also, we concatenate slow/fast-passing module in the stakeholder 2 network (condition 3). As shown in Table III, 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.

1) 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 II and III, 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.

2) 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.

3) 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.

VII. CONCLUSIONS

In this paper, we have proposed model of 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. In addition, we found that the head orientation data is crucial for improving the performance of the pedestrian trajectory estimation. Since pedestrian dataset can be further improved by accounting culture factors, we will have future work of this.

ACKNOWLEDGEMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant funded by the Korea government (MSIT) (2019-0-01190, [SW Star Lab] Robot Learning: Efficient, Safe, and Socially-Acceptable Machine Learning)

REFERENCES

[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] 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.

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

[8] 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.

[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] 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.

[13] 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.

[14] 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.

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

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

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

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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