Towards Personalized Autonomous Driving: Deep Reinforcement Learning from Human Feedback

Jiali Ling[1,a)] Jialong Li[1] Kenji Tei[†1,1] Shinichi Honiden[†1,1]

Abstract: In modern society, personalization is one of the important indicators to attract customers. And this is the same in the field of autonomous driving. Personalized autonomous driving can not only meet the different passengers' riding preferences but also relieve the pressure and distrust caused by autonomous driving to a certain extent. In this research, We regard human as another agent, and vehicles and humans are in a cooperative relationship. And we propose a composite reward model based on reinforcement learning, which combines the passenger's feedback on autonomous driving behavior. The system proposed in this study can learn personalized driving behavior based on passenger feedback .

Keywords: autonomous driving, DDPG, composite reward, human feedback, personalization

1. Introduction

For fully automated self-driving vehicles[1], The driver is no longer required to engage in the driving task. Thus the driver becomes a passive passenger or occupant. It is like we have a dedicated driver service. However, humans always hope that the ride experience can meet their preferences or expectations, whether it is for autonomous driving or dedicated driver service. An excellent driver can be familiar with and observe employers and grasp their ride preferences to improve ride comfort. However, today's autonomous driving cannot achieve this. Some studies have shown that the discomfort and distrust of passengers in autonomous driving are significantly higher than that of humans Driving[2].

In contrast, nowadays, self-driving research rarely considers the preferences of individual passengers[3]. Violating these preferences will lead to undesirable emotional or physical problems like passenger discomfort or anxiety. It will also affect passengers' trust in autonomous driving to a certain extent[4]. Passengers will have individual differences in ride comfort or satisfaction about the same driving behavior, so adjusting the driving behavior automatically for this difference will be necessary and valuable to solve these problems.

This paper considers individual passenger differences, which means that vehicle agent should respond to human agent preference. Thus we propose a novel system that can analyze passengers' favorability for driving behavior by human feedback. It lets the vehicle personalized learn the driving behavior, which aligns with the current passenger's preference.

2. Related work

In previous studies, autonomous driving usually only has one static driving mode, and it could not personally adjust driving behavior according to customers' actual preferences and feedback. In the existing studies, there are reinforcement learning methods that combine with human preferences. Christiano P et al.[5] proposed a very new idea, that is, instead of using the classic reward function to train the agent, it is a new method of reinforcement learning based on human feedback. Human preferences are used as weak supervision to accelerate the speed of agent learning. But it is mainly aimed at solving Atari games and requires a lot of labor costs.

In W. Lu et al.'s study, the researcher analyzes real-time human EEG data feedback and embodies it into actual operating instructions to help tractor drivers reduce manual operations and reduce their driving fatigue[6]. However, it did not consider the issue of driving preference.

This paper identified the issues and attempted to propose a system that can satisfy these needs.

3. Method

3.1 System design

Considering the characteristics of automatic driving control problems, Deep Deterministic Policy Gradient(DDPG), a type of Deep reinforcement Learning algorithm used to solve continuous action spaces problems, began to be used in autonomous driving areas [7].

On this basis, Multi-Agent Deep Deterministic Policy

[1] Faculty of Science and Engineering, Waseda University, 1 Chome-104 Totsukamachi, Shinjuku City, Tokyo 169-8050 Japan

[†1] Presently with National Institute of Informatics, 2 Chome-1-2 Hitotsubashi, Chiyoda City, Tokyo 101-8430 Japan

[a)] lingjiali0103@ruri.waseda.jp

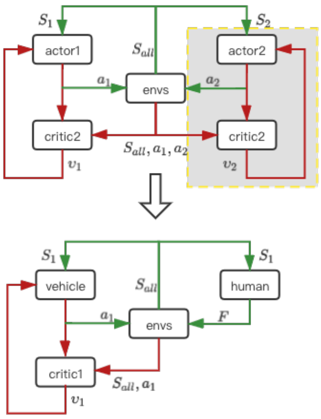

Fig. 1 MADDPG algorithm framework

Gradient(MADDPG)[8] is applied. The MADDPG algorithm framework is centralized training and decentralized execution.

During training, first, the Actor selects an action based on the current state, and then the Critic can calculate a Q value based on the state-action as feedback to the Actor's actions. Critic trains based on the estimated Q value and the actual Q value, and the Actor updates the strategy based on Critic's feedback.

When testing, we only need the Actor to complete it, and the Critic's feedback is not needed at this time. Therefore, during training, we can add some additional information in the Critic phase to get a more accurate Q value, such as the state and actions of other agents, which is the meaning of centralized training. That is, the vehicle agent is not only based on itself. In this case, the value of the current action is evaluated based on the state feedback of the human agent. Moreover, human feedback would directly affect the reward that the vehicle can obtain.

Because the agent human is in the vehicle, the actions are consistent with the agent vehicle, so the decentralized execution targets the agent vehicle, just like shown in **Fig. 1**.

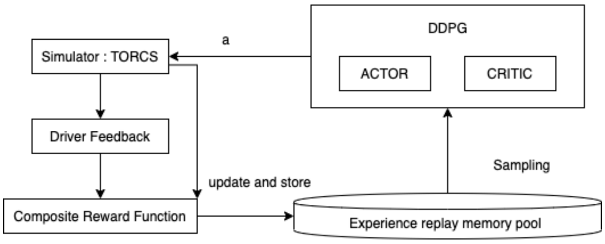

Fig. 2 Framework via composite reward function.

In this study, the proposal is to integrate a composite reward function into this algorithm framework. The overall structure shows in **Fig. 2**, the composite reward function will replace the actor network to update and store the transition data in the experience pool.

3.2 Composite reward Function

The composite reward function consists of following two parts:

3.2.1 Traditional reward:

The traditional reward is The rewards for vehicle driving status, traffic rules, and common sense constraints. The vehicle driving status constraints formula shows like 1 and the parameters are show in **Fig. 3**. $V_x \cos\theta$ represents the velocity along the track. $\mathrm{V_x} \sin\theta$ represents the velocity of the vertical track. TrackPos Measure the distance between the vehicle and the track. When the agent deviates from the track, the last item shows the punishment. The data will be provided by simulator sensors[9]. At the same time, there will be other restrictions such as collision penalty, etc.

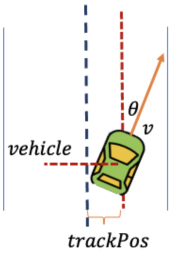

Fig. 3 Tradition reward parameters

$$\mathrm{R_t} = \alpha V_x \cos\theta - \beta \mathrm{V_x} \sin\theta - \gamma \mathrm{V_x} \mid trackPos \mid \qquad (1)$$

3.2.2 Human Feedback reward:

This part can also be understood as a personal favorability reward model. In this research, we will first analyze the collected related human Feedback signals, then train and classify them[10], finally get a human feedback reward after fuzzy evaluation.The evaluation result would be like formula 2.

$$\mathrm{R_f} = \{1_{\mathrm{positive}}, 0_{\mathrm{natural}}, -1_{\mathrm{negative}}\} \qquad (2)$$

Therefore, the final reward is the sum of traditional rewards and feedback rewards as formula 3. In this study, the proportion of traditional rewards should be greater than the proportion of feedback rewards(δ>0.5), and the essential requirement for autonomous driving is always safety.

$$\mathrm{R} = \delta \mathrm{R_t} + (1 - \delta)\mathrm{R_f} \qquad (3)$$

4. Evaluation design

The purpose of our research is to automatically adjust the driving behavior in line with the passenger's preference by their feedback, so it is necessary to verify whether our method can meet the individual preference of the passenger.

The system will train twice, the first turn without passengers involved. The purpose of the first training is to obtain the initial parameters and sample data. The second turn adds the Emotion feedback parts. We can get the passenger's emotion through some physiological signals, such as Electroencephalogram(EEG) signals[10]. However, due to the difficulty of testing and obtaining EEG signals in front of a virtual screen, at this stage, we chose DEAP[11], an emotional EEG research data set, as emotional feedback dataset for training. Then we trained a KNN classifier to predict human emotions from EEG data through valence-arousal mode. In this research, we roughly divide emotions into three categories, Positive, Negative, and Neutral. The following is the emotional feedback data in **Fig. 4** obtained by the participants numbered 11 in the test data set for movie fragments 1 to 30. The time required for a single feedback analysis during the test is about 0.0168 seconds.

In the preliminary stage of this research, We will first experiment with emotion feedback obtained from the dataset to demonstrate the feasibility of this system. Moreover, we will evaluate our methods in two ways with virtual and purposeful feedback.

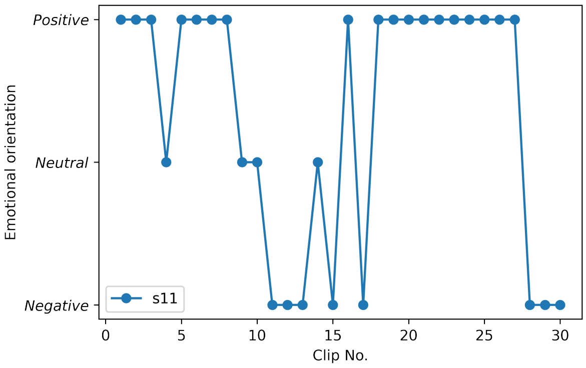

Fig. 4 Emotion feedback

Quantitatively: Because the elements which can affect the passengers' emotion are complex and changeable. It may not only includes driving status, road condition but also includes weather factors, etc. So in the experiment, we will select specific emotional feedback to make the results more intuitive and readable. For example, suppose that passengers are sensitive to speed and always give positive Emotion feedback to high speeds, and the opposite for low speeds. Then the change in speed can be used as an intuitive comparison.

Qualitatively: Since the actual feedback is multi-factorial, it is not easy to judge from the qualitative evaluation whether the driving behavior after training meets the passenger's preference without setting prerequisites. Sometimes we do not have a quantitative evaluation of behavior—reward function. We can only qualitatively evaluate the degree to which driving behavior meets human preferences. Let the passenger subjectively judge the satisfaction of the two driving.

5. Conclusion

In this paper, we propose a novel system for adjusting the driving behaviors by human feedback. We try to treat humans and vehicles as two different agents. They have different reward functions, but we only train and execute their agent vehicle and only use the feedback of agent Human as one of its state inputs. Traditional autonomous driving rewards usually only consider the impact or rewards brought by the vehicle itself. We consider the interaction between the vehicle and the environment and consider the feelings of the passenger in the vehicle. The passenger determines this part of the reward in response to driving behavior.

Compared with existing research, the method we designed does not require passengers to actively operate or select something but only uses the Emotion feedback of passengers when they are riding in the vehicle so that autonomous driving can return to the essence of autonomy.Because it is challenging to collect real emotional feedback on a ride, we only use the data set to obtain emotional feedback at this stage. Although the emotional feedback does not represent the actual thoughts of the passengers, the principles are the same, and the feasibility of the system can be verified. We believe that our research shows the possibility of personalized autonomous driving in the future. Also, we will continue to experiment and evaluate this system in future research.

Acknowledgments The research was partially supported by JSPS KAKENHI.

References

[1] T. S. of Automotive Engineers(SAE), "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," https://www.synopsys.com/automotive/autonomous-driving-levels.html, 2016.

[2] M. Seet, J. Harvy, R. Bose, A. Dragomir, A. Bezerianos, and N. Thakor, "Differential impact of autonomous vehicle malfunctions on human trust," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2020.

[3] M. Elbanhawi, M. Simic, and R. Jazar, "In the passenger seat: Investigating ride comfort measures in autonomous cars," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 3, pp. 4–17, 2015.

[4] N. Dillen, M. Ilievski, E. Law, L. E. Nacke, K. Czarnecki, and O. Schneider, *Keep Calm and Ride Along: Passenger Comfort and Anxiety as Physiological Responses to Autonomous Driving Styles*. Association for Computing Machinery, 2020, p. 1–13.

[5] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *arXiv preprint arXiv:1706.03741*, 2017.

[6] W. Lu, Y. Wei, J. Yuan, Y. Deng, and A. Song, "Tractor assistant driving control method based on eeg combined with rnn-tl deep learning algorithm," *IEEE Access*, vol. 8, pp. 163 269–163 279, 2020.

[7] W. Huang, F. Braghin, and S. Arrigoni, "Autonomous vehicle driving via deep deterministic policy gradient," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 59216. American Society of Mechanical Engineers, 2019, p. V003T01A017.

[8] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *CoRR*, vol. abs/1706.02275, 2017. [Online]. Available: http://arxiv.org/abs/1706.02275

[9] D. Loiacono, L. Cardamone, and P. L. Lanzi, "Simulated car racing championship: Competition software manual," *CoRR*, vol. abs/1304.1672, 2013.

[10] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu, "Continuous emotion detection using eeg signals and facial expressions," in *2014 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2014, pp. 1–6.

[11] M. G. F. Fortin, F. Rainville, M. Parizeau, and C. Gagné, "DEAP: Evolutionary algorithms made easy," *Journal of Machine Learning Research*, vol. 13, pp. 2171–2175, jul 2012.