Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Stochastic pedestrian avoidance for autonomous vehicles using hybrid reinforcement learning*

Huiqian LI†1, Jin HUANG†‡1, Zhong CAO†1, Diange YANG†1, Zhihua ZHONG2

[1]*School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China*

[2]*Chinese Academy of Engineering, Beijing 100088, China*

†E-mail: lihq20@mails.tsinghua.edu.cn; huangjin@tsinghua.edu.cn; caoc15@mails.tsinghua.edu.cn; ydg@tsinghua.edu.cn

Received Apr. 3, 2022; Revision accepted Aug. 10, 2022; Crosschecked Dec. 5, 2022

Abstract: Ensuring the safety of pedestrians is essential and challenging when autonomous vehicles are involved. Classical pedestrian avoidance strategies cannot handle uncertainty, and learning-based methods lack performance guarantees. In this paper we propose a hybrid reinforcement learning (HRL) approach for autonomous vehicles to safely interact with pedestrians behaving uncertainly. The method integrates the rule-based strategy and reinforcement learning strategy. The confidence of both strategies is evaluated using the data recorded in the training process. Then we design an activation function to select the final policy with higher confidence. In this way, we can guarantee that the final policy performance is not worse than that of the rule-based policy. To demonstrate the effectiveness of the proposed method, we validate it in simulation using an accelerated testing technique to generate stochastic pedestrians. The results indicate that it increases the success rate for pedestrian avoidance to 98.8%, compared with 94.4% of the baseline method.

Key words: Pedestrian; Hybrid reinforcement learning; Autonomous vehicles; Decision-making
https://doi.org/10.1631/FITEE.2200128 **CLC number:** TP18; U495

1 Introduction

Autonomous vehicles (AVs) are expected to avoid traffic accidents and improve road traffic safety. Interaction with other traffic participants is one of the major considerations in evaluating the AV safety, which can be intractable even for experienced human drivers (Li et al., 2021). According to the National Highway Traffic Safety Administration of the U.S. Department of Transportation, from 2017 to 2018, pedestrian fatalities increased from 6274 to 6482 in the United States, accounting for 20% of all motor vehicle fatalities (National Highway Traffic Safety Administration, 2019). The great uncertainty of pedestrians' behavior makes handling interaction with them a challenging task for AVs. A review illustrates that pedestrians' attention, speed, and trajectory can be affected by many factors, such as their age, the time of day, and road structure (Rasouli and Tsotsos, 2020). To promote the social acceptance of AVs, researchers must design a robust decision-making system to cope with stochastic and uncertain pedestrian behaviors.

In recent years, a rapidly growing amount of literature specifically studies the interaction between pedestrians and autonomous vehicles. Typical methods can be categorized as classical methods and learning-based methods. Classical methods include rule-based methods, optimization methods, and probabilistic methods (Yang et al., 2020; Bhattacharyya et al., 2021; Koç et al., 2021). Learning-based methods imply using machine learning

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 61872217, U20A20285, 52122217, and U1801263) and the Key R&D Projects of the Ministry of Science and Technology of China (No. 2020YFB1710901)

ORCID: Huiqian LI, https://orcid.org/0000-0003-4949-8834; Jin HUANG, https://orcid.org/0000-0001-8774-2936

techniques, such as deep learning (Li et al., 2022a), transfer learning (Li et al., 2022b), and reinforcement learning (Liu et al., 2021).

Classical methods gained popularity because of their advantages on interpretability, adjustability, and feasibility of implementation. Kapania et al. (2019) demonstrated that a hybrid controller with just four distinct modes allowed an autonomous vehicle to handle interaction with pedestrians successfully. Simulation and experimental results validated that the proposed controller outperformed an alternate partially observable Markov decision process (POMDP) based solution in Schratter et al. (2019). In addition, several researchers depicted pedestrian avoidance as a model predictive control (MPC) problem, by translating each pedestrian's predicted motion into inequality constraints (Batkovic et al., 2019). Jayaraman et al. (2020a) attempted to model pedestrians' crossing behavior using their gap acceptance behavior and constant velocity dynamics for long-term (>5 s) pedestrian trajectory prediction at crosswalks. Based on this work, they developed a behavior-aware MPC controller to efficiently plan for longer horizons to handle a wider range of pedestrian interaction scenarios (Jayaraman et al., 2020b). Rule-based strategies can be implemented easily in practice and work well in most situations. However, such strategies may fail in some situations where the assumptions in the design process are not consistent with the facts, and they cannot adjust themselves to avoid repeating these mistakes in the future. For example, in Kapania et al. (2019), the proposed hybrid controller fails if pedestrians stand still on the roadway.

Some researchers have used learning-based methods for decision-making on interaction with pedestrians. Bai et al. (2015) implemented a POMDP-based online planner on an autonomous car to drive near pedestrians safely, which indicated its effectiveness of planning under uncertainty. Bouton et al. (2018) leveraged a similar strategy and scaled it to avoid multiple road users. Everett et al. (2021) developed an algorithm, using deep reinforcement learning (DRL), to learn collision avoidance in a complex environment, without assuming any particular behavior rule for nearby agents. Pusse and Klusch (2019) proposed a hybrid solution named HyLEAP, combining the advantages of DRL and approximate POMDP planning, for collision-free navigation. The method was evaluated by simulation in multiple pedestrian–car accident scenarios in a German in-depth road accident study. The results revealed that the hybrid solution is superior to its individual methods. The previous work shows the strength of the learning-based algorithms in decision-making under significant uncertainties. However, the performance of learning-based methods cannot be guaranteed without sufficient training (Cao et al., 2022).

In recent years, many explorations have been conducted to ensure the performance of learning-based approaches in safety-critical systems (García and Fernández, 2015; Cao et al., 2021). Combining learning- and rule-based policies is one of the typical approaches for guaranteeing safety and can integrate their strengths (Zhou et al., 2020). Yurtsever et al. (2020) proposed a hybrid approach for integrating a path planner into a vision-based DRL framework to mitigate the drawbacks of both approaches. By introducing a straying-away penalty in the reward function, the DRL agent is taught to oversee the planner and follow it when the planned path is safe. Simulation results showed that the proposed method can plan its path and navigate between randomly chosen origin–destination points. Cao et al. (2022) designed a framework named confidence-aware reinforcement learning (RL). In this framework, the RL agent works with a baseline rule-based policy and intervenes only when it has higher confidence. The framework has been applied in a two-lane roundabout scenario and shows better performance than both the pure RL policy and the baseline policy.

In this paper, to interact with pedestrians with safety-guaranteed performance for AVs, we propose a hybrid RL (HRL) strategy. The strategy starts with a baseline rule-based pedestrian-avoidance policy, and the RL policy is activated when the baseline policy has a lower score. We design the rule-based strategy based on the literature and focus on enhancing its performance through RL. To this end, we design a method to evaluate both policies and a function to decide the activation time. Finally, a simulation platform based on CARLA is built to train and test the proposed hybrid policy.

The main contributions of this paper are summarized as follows: (1) a hybrid RL approach for the autonomous vehicle to avoid pedestrians, combining the rule-based policy and RL policy, (2) a reliable

activation function design to determine whether to activate the rule-based policy or RL policy, and (3) a stochastic pedestrian generation method for accelerated evaluation in simulation.

2 Problem statement

In this paper, we focus on the traffic scenario illustrated in Fig. 1: An autonomous vehicle, called the ego vehicle, is driving on an urban road without traffic signs, where no control device explicitly guides the interaction between the ego vehicle and pedestrians. Pedestrians enter the road from the sidewalk at some reasonable speed and orientation. The ego vehicle is required to avoid collision when the pedestrians are on the roadway.

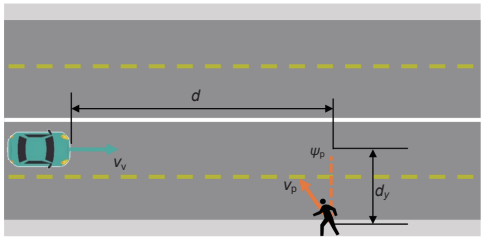

Fig. 1 Traffic scenario of the ego vehicle and pedestrian interaction, where the distances between the ego vehicle and pedestrian, along and vertical to the forward direction, are represented by d and d_y, respectively, the ego vehicle drives at the speed v_v along the road, and the pedestrian goes across the road at speed v_p and heading angle ψ_p

In the process of designing the pedestrian avoidance method, we follow several reasonable principles and assumptions: (1) Once a pedestrian enters the roadway, the ego vehicle can detect his/her position, speed, and heading direction. This assumption holds easily because AVs are usually equipped with various sensors and algorithms to realize pedestrian detection. (2) The ego vehicle is not explicitly required to stop when interacting with pedestrians. This principle relates to local traffic rules, but we consider only avoiding collision in this work. (3) In each vehicle–pedestrian interaction, only one pedestrian enters the roadway. This is the typical scenario that occurs most of the time. The proposed approach focuses on avoiding collision in this scenario. (4) The pedestrian does not change walking speed or orientation while crossing in the crosswalk. In this work, the reaction of the pedestrian is not modeled explicitly.

3 Hybrid reinforcement learning

In this section, we establish a hybrid reinforcement learning (HRL) framework in which the ego vehicle can safely interact with pedestrians. The HRL framework starts with a rule-based policy, and the RL policy is integrated to enhance performance. Section 3.1 introduces the rule-based policy adopted in the framework, and Section 3.2 explains how the HRL framework is implemented.

3.1 Rule-based policy

We select the hybrid control architecture in Kapania et al. (2019) as the fundamental rule-based policy. However, other rule-based policies can also be used in our framework. The basic policy designs a finite state machine (FSM) with four modes: keep speed, slow down, hard brake, and speed up, as shown in Fig. 2; the variables of the FSM are explained in Table 1.

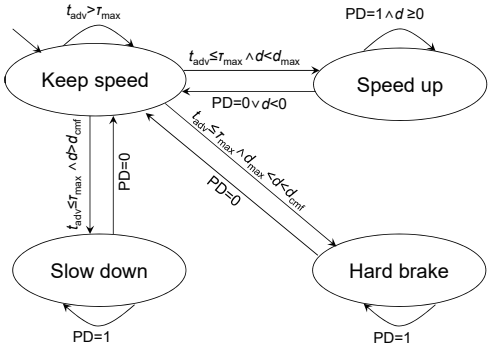

Fig. 2 Baseline rule-based policy

Table 1 Variables of the finite state machine

Symbol	Definition
t_{adv}	Time advantage, $\frac{d_y}{v_p}-\frac{d}{v_v}$
τ_{max}	Time advantage threshold
a_{max}	Value of maximum deceleration
a_{cmf}	Value of comfortable deceleration
d_{max}	Maximum braking distance, $\frac{v_v^2}{2a_{max}}$
d_{cmf}	Comfortable braking distance, $\frac{v_v^2}{2a_{cmf}}$
PD	Variable indicating whether the pedestrian is detected—1, detected; 0, not

In each mode, the baseline algorithm uses different methodologies to compute the desired acceleration. The transitions between modes are determined by the current state of the ego vehicle and whether a

pedestrian is detected. Details of the four modes are as follows:

Keep speed: In this mode, the ego vehicle detects no pedestrian on the road and attempts to drive at the desired speed v_{des}. The proportional speed control law is applied to compute the target acceleration:

$$a = k_{\mathrm{p}}(v_{\mathrm{v}} - v_{\mathrm{des}}), \tag{1}$$

where k_{p} represents the proportional coefficient, v_{v} is the current speed of the ego vehicle, and v_{des} is the desired speed of the vehicle, which is the same as the limit speed of the current lane v_{lim}. The Boolean variable PD is 1 when pedestrians are detected on the roadway. If PD is equal to 0, or time advantage t_{adv} exceeds a specified threshold $\tau_{\max}$, the ego vehicle keeps in the keep speed mode. Otherwise, the FSM decides which mode to enter next.

Slow down: The FSM enters the slow down mode if the time advantage is too small for the ego vehicle to pass directly. Considering the smoothness of driving, the ego vehicle will yield to the pedestrian at a comfortable deceleration a_{cmf} when the distance is sufficient, namely $d > d_{\mathrm{cmf}} = \frac{v_{\mathrm{v}}^2}{2a_{\mathrm{cmf}}}$. In this mode, the desired deceleration is a_{cmf} and the desired speed is given by

$$v_{\mathrm{des}}(d) = \sqrt{2a_{\mathrm{cmf}}(d - d_{\mathrm{o}}) + v_{\mathrm{o}}^2}, \tag{2}$$

where d_{o} and v_{o} are the values of d and v at the beginning time when FSM turns to the slow down mode, respectively. The vehicle yields at a deceleration of a_{cmf}, with an additional feedback term as shown in Eq. (3):

$$a = -a_{\mathrm{cmf}} + k_{\mathrm{p}}(v - v_{\mathrm{des}}). \tag{3}$$

Hard brake: If the distance between the ego vehicle and the pedestrian satisfies $d_{\max} < d < d_{\mathrm{cmf}}$, the ego vehicle needs to decelerate more quickly than a_{cmf}. In this mode, the desired speed is given by

$$v_{\mathrm{des}} = \frac{v_{\mathrm{o}}'}{\sqrt{d_{\mathrm{o}}'}}\sqrt{d}, \tag{4}$$

where d_{o}' and v_{o}' are the values of d and v when the FSM first enters the hard brake mode, respectively. The target deceleration in this mode can be computed by

$$a = -\frac{v^2}{2d} + k_{\mathrm{p}}(v - v_{\mathrm{des}}). \tag{5}$$

Speed up: The condition under which the FSM enters this mode is $d < d_{\max}$, which means that there is no sufficient place for the ego vehicle to slow down to avoid the pedestrian. In this situation, speeding up and passing quickly make more sense. The acceleration in this mode is set as a_{cmf}.

The FSM policy used in this study shows its effectiveness for most pedestrian avoidance situations in previous work (Kapania et al., 2019; Jayaraman et al., 2020b). However, it may fail in some cases where the given assumptions do not hold. Starting with this policy, our HRL policy can achieve a continuously improving performance.

3.2 Hybrid reinforcement learning policy

The framework of the designed HRL strategy is shown in Fig. 3. According to the current traffic environmental state, we can obtain a rule- or learning-based policy. We evaluate both policies with a uniform criterion. Then we design an activation function to select the final policy. In this way, we can guarantee that the performance of the final policy is not worse than that of the rule-based policy. Moreover, by combination with the self-learning algorithm, the autonomous vehicle can learn from previous failures and improve safety. We use deep Q learning (DQN) as the learning-based method in our HRL framework (Mnih et al., 2015). To apply the learning-based method, we should model the pedestrian avoidance problem as a Markov decision process (MDP). First, we define the state, action, and reward model for the problem.

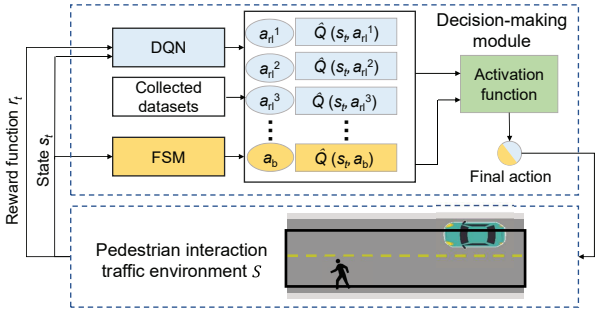

Fig. 3 Framework of the proposed hybrid reinforcement learning (HRL) pedestrian avoidance strategy

State: The state space should contain all the considered elements of the scenario. The state space

$\mathcal{S}$ can be defined as follows:

$$s=(d,d_y,\psi_{\mathrm{p}},v_{\mathrm{v}},v_{\mathrm{p}})\in\mathcal{S},$$

and the explanations of each state variable can be referred to in Section 2.

Action: In this work, we assume that the ego vehicle handles interaction with pedestrians by simply adjusting its longitudinal acceleration. For RL policy generation, the ego vehicle can use the same modes as the baseline FSM policy, namely keep speed, slow down, hard brake, and speed up. Therefore, action space $\mathcal{A}$ can be defined as

$$\mathcal{A}=\{a_{\mathrm{ks}},a_{\mathrm{sd}},a_{\mathrm{hb}},a_{\mathrm{su}}\},$$

where each action corresponds to a mode in the FSM. At each time t, according to the current state s_t, the ego vehicle can obtain two candidate actions, a_{rule} by the FSM and a_{rl} by RL exploration. The calculation of the target acceleration taking each action is the same as that in the FSM.

Reward: The proposed strategy aims to avoid collisions and improve efficiency, so the reward model of RL, $r(s,a):\mathcal{S}\times\mathcal{A}\to\mathcal{R}$, should be relevant to these two factors. The reward function consists of two portions, safety reward r_1 and efficiency reward r_2. The final reward is $r=r_1+r_2$.

The agent obtains a safety penalty when collisions occur.

$$r_1=\begin{cases}0, \text{ no collision},\\ -1, \text{ there is a collision}.\end{cases}$$

The safety reward encourages the ego vehicle to interact with pedestrians safely. However, if we set only a safety reward, the vehicle may drive overly conservatively to avoid collisions. To keep a normal road capacity, we introduce an efficiency reward

$$r_2=\frac{v}{v_{\mathrm{des}}}-1,$$

where the desired speed v_{des} is equal to the limit speed of the current lane. The agent receives a penalty when the speed is lower than the desired speed.

Data collection: To generate an HRL policy, we need to collect datasets defined as follows:

$$\begin{cases}\tau_\pi(s):=\{s_1=s,a_1^\tau,s_2^\tau,a_2^\tau,...,s_H^\tau\},\\ D_\pi:=\{\tau_\pi(s_i)\},s_i\in\mathcal{S},\\ G(\tau_\pi(s)):=\sum\limits_i\gamma^n(r(s_i^\tau,a_i^\tau)),\end{cases}\tag{6}$$

where $\tau_\pi(s)$ stands for a trajectory of length H with policy π, which is a sequence of states and actions starting with state s. D_π denotes the dataset that saves all these trajectories. For each trajectory, the return value means the sum of the discounted rewards, represented as $G(\tau_\pi(s))$.

In the process of data collection, the sliding window with a fixed horizon is used to collect the trajectory and its value return. To evaluate two different policies, we define two sub-datasets as

$$\begin{cases}D(s,a)=D_{\mathrm{rule}}(s)\cup D_{\mathrm{rl}},\\ D_{\mathrm{rule}}:=\{\tau(s_1=s,a_1=\pi_{\mathrm{rule}}(s_1))\},\\ D_{\mathrm{rl}}:=\{\tau(s_1=s,a_1=\pi_{\mathrm{rl}}(s_1))\},\end{cases}\tag{7}$$

where the total dataset $D(s,a)$ consists of two sub-datasets D_{rule} and D_{rl}. The former contains the trajectories that use the rule-based policy as the first action, while the latter uses the RL policy.

RL policy generation: The objective of RL is training the ego vehicle to execute optimal actions to handle the interaction with pedestrians so that it will obtain a high cumulative reward in a finite expectation horizon. We employ DQN as the learning method in our HRL approach to solve the MDP problem described above.

The DQN updates the Q value by sampling environmental data, rather than setting up a state transition probability template. Therefore, the DQN can solve the MDP faster. The principle of updating the Q value is the Bellman equation:

$$\begin{aligned}Q(s_t,a_t)\leftarrow&\,Q(s_t,a_t)\\&+\alpha\left[r(s_{t+1})+\gamma\max_a Q(s_{t+1},a)-Q(s_t,a_t)\right],\end{aligned}\tag{8}$$

where $Q(s_k,a_k):\mathcal{S}\times\mathcal{A}\to\mathcal{R}$ represents the Q function at state s_k with action a_k. α is the learning rate and γ denotes the discount factor, which is a scalar in $[0,1]$ that demonstrates the relative importance of the next reward concerning the current.

In Eq. (8), we notice that the update of the Q function depends on the state and the action. However, in our problem, the state space is continuous, where we cannot visit a specified state repeatedly. Therefore, we use a neural network with parameters θ to approximate the Q function. The updating

principle is

$$\begin{cases} \theta_{j+1} \leftarrow \theta_j - \alpha \nabla_\theta \mathbb{E}\left[(Q(s_t, a_t, \theta_j) - Q^+(s_t, a_t))^2\right], \\ Q^+(s_t, a_t) = r(s_{t+1}) + \gamma \max\limits_a Q(s_{t+1}, a, \theta^-), \end{cases} \tag{9}$$

where θ and θ^- represent current adjusting parameters and parameters from some previous iteration, respectively. θ^- is updated after a specified number of iterations, which is helpful to keep the network stable. The term $(Q(s_t, a_t, \theta_j) - Q^+(s_t, a_t))^2$ denotes a training error, where Q^+ means the value calculated by the Bellman equation. The action generated by the deep Q learning algorithm is

$$a_{\mathrm{rl}} = \operatorname*{argmax}_{a \in \mathcal{A}} Q(s_t, a). \tag{10}$$

Training process: To collect the required data efficiently, the proposed HRL method designs a training process with two stages, baseline evaluation stage and RL agent exploration stage. In the first stage, the ego vehicle uses only the baseline rule-based policy to collect datasets and train the value net. In the second stage, the RL policy may explore different actions for better performance. The ϵ-greedy algorithm is used for exploration. Two conditions prevent switching to the second stage: (1) The baseline policy has not been evaluated sufficiently; namely, the number of times the state is visited is smaller than a threshold n_{thre}. (2) The baseline policy performs well. For the second condition, the exploration probability is set equal to $-Q(s, a_{\mathrm{rule}})$. These conditions help the agent reduce unnecessary exploration.

HRL policy: The core idea of the HRL policy is to determine which policy is used according to a uniform criterion. To generate hybrid policies, we first use the datasets collected by rule-based policy $D(s_t, a_{\mathrm{rule}})$ to compute the value function distribution of the rule-based policy. Further, based on the datasets $D(s_t, a_{\mathrm{rl}})$, we can estimate the distribution of the learning-based policy's value function. Then the hybrid policy can be designed as

$$\pi_{\mathrm{hrl}} = \pi_{\mathrm{rule}} + \frac{\pi_{\mathrm{rl}} - \pi_{\mathrm{rule}}}{1 + \exp(-w\mathcal{C}(\pi_{\mathrm{rl}}, \pi_{\mathrm{rule}}, s))}, \tag{11}$$

where π_{rule}, π_{rl}, and π_{hrl} represent the rule-based policy, RL policy, and hybrid policy, respectively. w is a constant that tends to ∞. The activation function $\mathcal{C}(\pi_{\mathrm{rl}}, \pi_{\mathrm{rule}}, s)$ is

$$\mathcal{C}(\pi_{\mathrm{rl}}, \pi_{\mathrm{rule}}, s) = Q(s, \pi_{\mathrm{rl}}(s)) - Q(s, \pi_{\mathrm{rule}}(s)) - c_{\mathrm{thre}}, \tag{12}$$

where c_{thre} is the activation threshold that ranges from 0 to 1. A larger activation threshold makes the policy more conservative. Previous work has demonstrated by experiments that when the threshold is set as 0.5, the maximum expectation of the performance enhancement can be achieved (Cao et al., 2021). Nevertheless, the activation threshold is still selected by trial and error, and further research is needed for theoretical foundation. When $\mathcal{C} > 0$ the learning-based policy is triggered; otherwise, the rule-based policy is adopted.

4 Simulation

In this section, we describe a simulation test we conducted to validate the effectiveness of the proposed HRL policy. Section 4.1 describes the details of the simulation setup. The simulation results are displayed and discussed in Section 4.2.

4.1 Simulation setup

To test the performance of the HRL policy, we designed a scenario with stochastic pedestrians. The ability of AVs to handle interaction with the pedestrian relates to the distance at the moment the pedestrian is detected and its behavior. We verified our method in a four-lane urban road scenario constructed using the CARLA simulator. At the beginning the ego vehicle drove in the left lane, and the pedestrian was spawned near the road and started to cross the walk at some initial condition. If the ego vehicle arrived at the destination or hit the pedestrian, the next test epoch started.

The algorithm architecture of the ego vehicle was based on an open-source autonomous driving platform CLAP (Zhong et al., 2020). The desired acceleration was calculated by the proposed algorithm, and a proportional integral derivative (PID) controller was used to track the desired acceleration in the platform. The proposed method was implemented in Python based on an open-source RL framework named Stable Baselines. The parameters used in the baseline policy and training process are listed in Table 2.

To simulate the randomness of pedestrians, we set several types of pedestrians with different behavior patterns and risk levels, as listed in Table 3. Owing to the rareness of safety-critical events, we designed a pedestrian generation method for

Table 2 Parameter setup in simulation

Parameter	Symbol	Value
Limit speed	v_{lim}	8 m/s
Comfortable acceleration	a_{cmf}	2 m/s^2
Maximum acceleration	$a_{\max}$	6 m/s^2
Feedback factor	k_s	−2
Activation threshold	c_{thre}	0.5
Number of visited times	n_{thre}	30
Discount factor	γ	0.99

accelerated evaluation to demonstrate the performance of the algorithm (Feng et al., 2021). There are two categories of pedestrian behavior patterns, normal behavior and random behavior. Normal pedestrians walk at a relatively low speed, while random pedestrians walk faster and may often act abruptly, such as naughty children. The pedestrians are also divided into four risk levels according to the amount of deceleration required of the ego vehicle to avoid the pedestrian (Wang et al., 2019). In each test, we sampled a pair of time-to-collision (TTC $= \frac{d}{v_{\text{v}}}$) and speed value as the initial condition of the pedestrian. TTC is calculated as the initial distance divided by the initial speed of the vehicle. We also set normal or random behavior patterns for each pedestrian. In the training process, the initial conditions of pedestrians were generated randomly. The HRL policy was trained for 1500 epochs in total. For testing, we sampled 1000 cases with different risk levels, as shown in Fig. 4. Both the baseline FSM policy and the HRL policy were tested in these cases.

Table 3 Pedestrian behavior patterns and risk levels

Type	Speed v_{p} (m/s)	Orientation ψ_{p} (°)
Normal	$[1, 2]$	0
Random	$[1.5, 4]$	$[-30, 30]$
Risk level	**Required deceleration (m/s^2)**	
High	$[-a_{\max}, -4.1)$	
Medium	$[-4.1, -2.3)$	
Low	$[-2.3, 0)$	
Trivial	$[0, +\infty)$	

4.2 Results

In this subsection, the simulation results of the HRL algorithm are shown and compared with those of the baseline method. Fig. 5 shows the relationship between the success rate and the number of simulation runs for both HRL and the baseline FSM method. The success rate converged after about 400 simulation runs. Figs. 6 and 7 show the results of test cases. Every marker in the figure represents 1 of 1000 simulation cases.

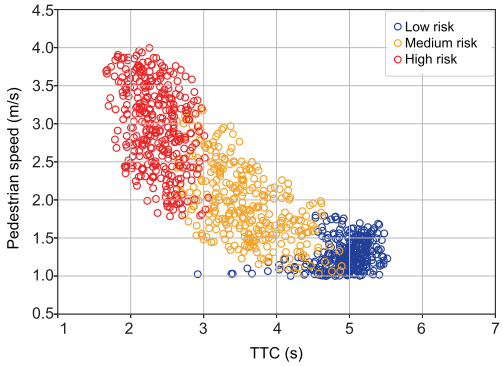

Fig. 4 Initial conditions generated according to different risk levels (References to color refer to the online version of this figure)

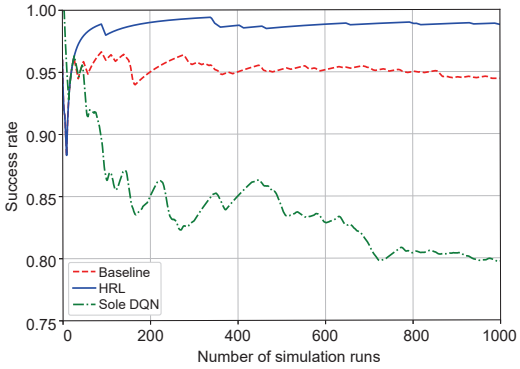

Fig. 5 Success rate of the ego vehicle passing the crosswalk (References to color refer to the online version of this figure)

4.2.1 Safety

Safety is the most significant factor to consider for autonomous vehicles. This work evaluates the degree of safety based on how successful the pedestrian is in crossing the street. From Fig. 8, the overall success rate of the baseline FSM method was 94.4%, while that of our HRL method was up to 98.8%. For both the normal pedestrian case and random pedestrian case, our HRL method (99.8%, 97.8%) outperformed the baseline FSM method (96.3%, 92.5%). The minimum distance between the ego vehicle and the pedestrian in each test case is shown in Fig. 6. For the baseline results, most collisions happened when the ego vehicle interacted with pedestrians

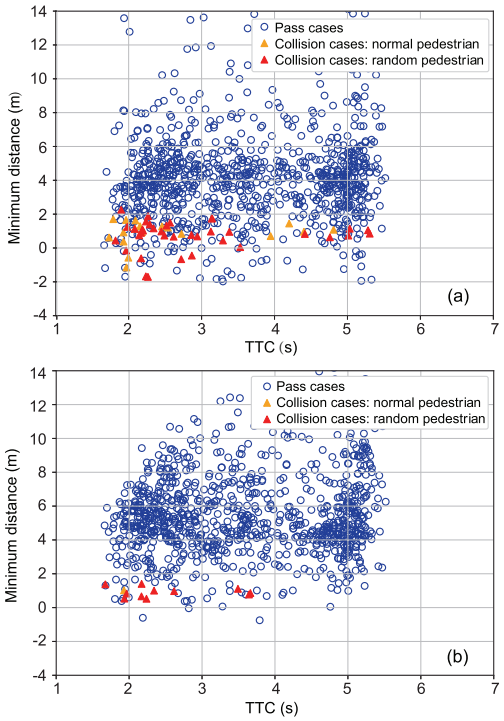

Fig. 6 Relationship between the minimum distance and TTC (References to color refer to the online version of this figure)

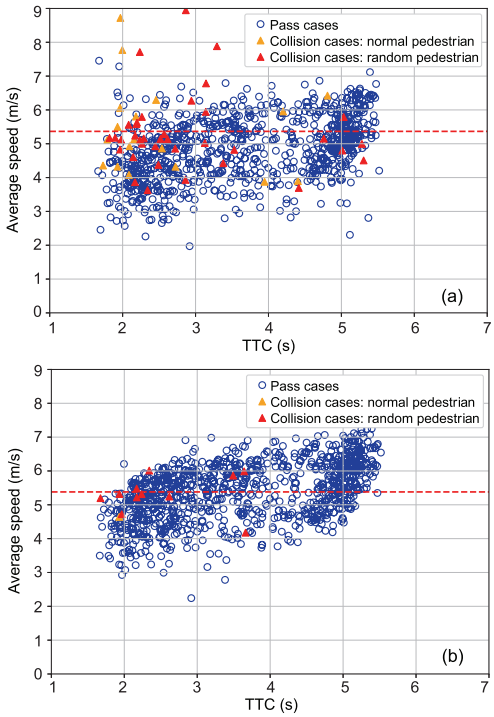

Fig. 7 Relationship between the average speed of the ego vehicle and TTC, where the dashed line stands for the mean value of the average speed in all test cases (References to color refer to the online version of this figure)

with random behavior. This is because their behavior does not match the baseline assumptions. The initial TTC of the collision cases was about 2–3 s, because in this case the ego vehicle did not have enough time to react to the suddenly detected pedestrian. To demonstrate the effectiveness of the proposed method in providing a lower bound of the performance, we conducted the simulation test using solely the learning-based policy. The action was selected by the policy used in the hybrid method to exclude the influence of the training process. As shown in Fig. 5, the success rate of the DQN alone was lower than those of the baseline and HRL, because the learning-based method is not sufficiently trained. Thus, the introduction of the learning-based method enables the ego vehicle to attempt to avoid the accidents that occurred before.

4.2.2 Efficiency

Fig. 7 shows the average speed of the ego vehicle for each simulation case. For large TTC, the distance between the ego vehicle and the pedestrian is sufficient to pass with less or even no deceleration. Conversely, small TTC indicates a high collision risk, so the ego vehicle needs to slow down to yield to pedestrians. The dashed line indicates the mean value of the average speed in all test cases, and shows that the two methods had similar pass performance efficiency (5.37 m/s for FSM and 5.38 m/s for HRL). This result means that the proposed HRL method does not improve safety using overly conservative driving. Moreover, we discuss the HRL computational efficiency to prove its usability in real-world AVs. The average runtime of the proposed HRL policy was 1.79 ms, which is greater than FSM's 0.51 ms. Obviously, the introduction of the learning policy increases the computational burden of the autonomous vehicle. The increase in computation is acceptable for real-world AVs, because most of them are equipped with huge computing power and can meet the real-time requirements.

4.2.3 Effectiveness of the proposed method

To demonstrate the effectiveness of the HRL policy in enhancing safety when interacting with pedestrians, the cases where the rule-based policy fails are plotted in Fig. 9. The ego vehicle can

succeed in most cases where the baseline policy fails. This is because the rule-based policy cannot adjust its actions according to the environment, while the proposed strategy can avoid previous failures by self-learning. In addition, the HRL strategy never fails as long as the baseline policy succeeds, which demonstrates that the proposed strategy takes the performance of the baseline as a lower bound. Theoretically, if the RL policy is sufficiently trained, most failures can be avoided by HRL. Nevertheless, there are several cases in which HRL still fails. By observing the simulation process, we conclude two reasons: (1) The RL policy is not sufficiently trained and is not trained in some cases that FSM fails to handle; (2) The modeling of the pedestrians' behavior is simple. There are some cases where the pedestrian still goes forward even though the ego vehicle has stopped to yield. It is hoped that HRL will be safer in the future when a more reasonable behavior model is introduced.

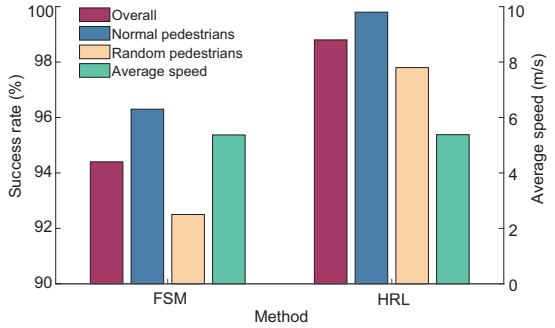

Fig. 8 Performance comparison of the finite state machine and hybrid reinforcement learning (References to color refer to the online version of this figure)

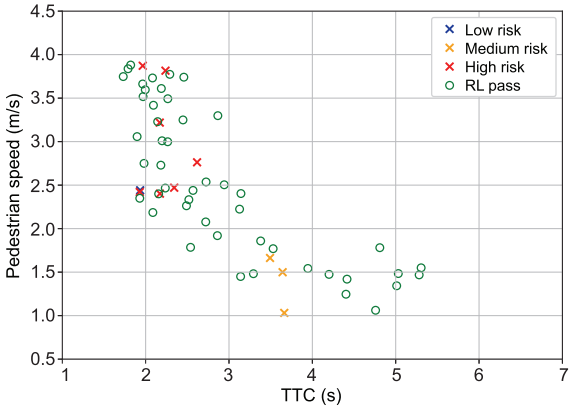

Fig. 9 Performance of the proposed method in the cases where the baseline method fails (References to color refer to the online version of this figure)

5 Conclusions

In this paper, we propose a hybrid reinforcement learning strategy for autonomous vehicles that must handle interaction with stochastic pedestrians. The proposed strategy can adjust the ego vehicle's longitudinal acceleration to improve the success rate when interacting with pedestrians in different risk levels and behavior models.

Our proposed strategy is a combination of the finite state machine and DQN. The baseline FSM method provides rule-based actions and state transition in most situations, and the learning-based action should be activated when the rule-based action may fail. The method is tested on an urban road on the CARLA simulator. The results show that the hybrid reinforcement learning method can increase the success rate by 4.4% compared with the basic rule-based method, and benefits from the introduction of reinforcement learning. Meanwhile, the ego vehicle can maintain a similar average speed. To conclude, the proposed hybrid reinforcement learning control strategy can generate stable actions against uncertain pedestrian behavior and outperform the baseline. Nevertheless, we only verify the effectiveness of the proposed method against the various speeds and orientations of pedestrians. In fact, pedestrian behavior is complex, and current research shows that precise behavior modeling and prediction are significant (Li et al., 2020). In the future, we will consider the pedestrian model, test our method in a real-world environment, and scale it to handle multiple pedestrians.

Contributors

Jin HUANG and Zhong CAO designed the research. Huiqian LI processed the data and drafted the paper. Diange YANG helped organize the paper. Huiqian LI and Zhihua ZHONG revised and finalized the paper.

Compliance with ethics guidelines

Huiqian LI, Jin HUANG, Zhong CAO, Diange YANG, and Zhihua ZHONG declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Bai HY, Cai SJ, Ye N, et al., 2015. Intention-aware online

POMDP planning for autonomous driving in a crowd. IEEE Int Conf on Robotics and Automation, p.454-460. https://doi.org/10.1109/ICRA.2015.7139219

Batkovic I, Zanon M, Ali M, et al., 2019. Real-time constrained trajectory planning and vehicle control for proactive autonomous driving with road users. 18th European Control Conf, p.256-262. https://doi.org/10.23919/ECC.2019.8796099

Bhattacharyya A, Reino DO, Fritz M, et al., 2021. Euro-PVI: pedestrian vehicle interactions in dense urban centers. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6404-6413. https://doi.org/10.1109/CVPR46437.2021.00634

Bouton M, Nakhaei A, Fujimura K, et al., 2018. Scalable decision making with sensor occlusions for autonomous driving. IEEE Int Conf on Robotics and Automation, p.2076-2081. https://doi.org/10.1109/ICRA.2018.8460914

Cao Z, Yang DG, Xu SB, et al., 2021. Highway exiting planner for automated vehicles using reinforcement learning. *IEEE Trans Intell Transp Syst*, 22(2):990-1000. https://doi.org/10.1109/tits.2019.2961739

Cao Z, Xu SB, Peng HE, et al., 2022. Confidence-aware reinforcement learning for self-driving cars. *IEEE Trans Intell Transp Syst*, 23(7):7419-7430. https://doi.org/10.1109/TITS.2021.3069497

Everett M, Chen YF, How JP, 2021. Collision avoidance in pedestrian-rich environments with deep reinforcement learning. *IEEE Access*, 9:10357-10377. https://doi.org/10.1109/ACCESS.2021.3050338

Feng S, Yan XT, Sun HW, et al., 2021. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nat Commun*, 12(1):748. https://doi.org/10.1038/s41467-021-21007-8

García J, Fernández F, 2015. A comprehensive survey on safe reinforcement learning. *J Mach Learn Res*, 16(1):1437-1480.

Jayaraman SK, Tilbury DM, Yang XJ, et al., 2020a. Analysis and prediction of pedestrian crosswalk behavior during automated vehicle interactions. IEEE Int Conf on Robotics and Automation, p.6426-6432. https://doi.org/10.1109/icra40945.2020.9197347

Jayaraman SK, Robert LP, Yang XJ, et al., 2020b. Efficient behavior-aware control of automated vehicles at crosswalks using minimal information pedestrian prediction model. American Control Conf, p.4362-4368. https://doi.org/10.23919/ACC45564.2020.9147248

Kapania NR, Govindarajan V, Borrelli F, et al., 2019. A hybrid control design for autonomous vehicles at uncontrolled crosswalks. IEEE Intelligent Vehicles Symp, p.1604-1611. https://doi.org/10.1109/IVS.2019.8814116

Koç M, Yurtsever E, Redmill K, et al., 2021. Pedestrian emergence estimation and occlusion-aware risk assessment for urban autonomous driving. IEEE Conf on Intelligent Transportation Systems, p.292-297. https://doi.org/10.1109/ITSC48978.2021.9565071

Li ZR, Gong JW, Lu C, et al., 2020. Importance weighted Gaussian process regression for transferable driver behaviour learning in the lane change scenario. *IEEE Trans Veh Technol*, 69(11):12497-12509. https://doi.org/10.1109/TVT.2020.3021752

Li ZR, Gong JW, Lu C, et al., 2021. Interactive behavior prediction for heterogeneous traffic participants in the urban road: a graph-neural-network-based multitask learning framework. *IEEE/ASME Trans Mechatron*, 26(3):1339-1349. https://doi.org/10.1109/TMECH.2021.3073736

Li ZR, Lu C, Yi YT, et al., 2022a. A hierarchical framework for interactive behaviour prediction of heterogeneous traffic participants based on graph neural network. *IEEE Trans Intell Transp Syst*, 23(7):9102-9114. https://doi.org/10.1109/TITS.2021.3090851

Li ZR, Gong J, Lu C, et al., 2022b. Personalized driver braking behavior modeling in the car-following scenario: an importance-weight-based transfer learning approach. *IEEE Trans Ind Electron*, 69(10):10704-10714. https://doi.org/10.1109/TIE.2022.3146549

Liu Q, Li XY, Yuan SH, et al., 2021. Decision-making technology for autonomous vehicles: learning-based methods, applications and future outlook. IEEE Conf on Intelligent Transportation Systems, p.30-37. https://doi.org/10.1109/ITSC48978.2021.9564580

Mnih V, Kavukcuoglu K, Silver D, et al., 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529-533. https://doi.org/10.1038/nature14236

National Highway Traffic Safety Administration, 2019. 2018 Fatal Motor Vehicle Crashes: Overview. Traffic Safety Facts Research Note, U.S. Department of Transportation, p.1-9.

Pusse F, Klusch M, 2019. Hybrid online POMDP planning and deep reinforcement learning for safer self-driving cars. IEEE Intelligent Vehicles Symp, p.1013-1020. https://doi.org/10.1109/IVS.2019.8814125

Rasouli A, Tsotsos JK, 2020. Autonomous vehicles that interact with pedestrians: a survey of theory and practice. *IEEE Trans Intell Transp Syst*, 21(3):900-918. https://doi.org/10.1109/TITS.2019.2901817

Schratter M, Bouton M, Kochenderfer MJ, et al., 2019. Pedestrian collision avoidance system for scenarios with occlusions. IEEE Intelligent Vehicles Symp, p.1054-1060. https://doi.org/10.1109/IVS.2019.8814076

Wang XP, Peng HE, Zhao D, 2019. Combining reachability analysis and importance sampling for accelerated evaluation of highly automated vehicles at pedestrian crossing. Proc ASME Dynamic Systems and Control Conf, Article V003T18A011. https://doi.org/10.1115/DSCC2019-9179

Yang DF, Redmill K, Özgüner Ü, 2020. A multi-state social force based framework for vehicle-pedestrian interaction in uncontrolled pedestrian crossing scenarios. IEEE Intelligent Vehicles Symp, p.1807-1812. https://doi.org/10.1109/IV47402.2020.9304561

Yurtsever E, Capito L, Redmill K, et al., 2020. Integrating deep reinforcement learning with model-based path planners for automated driving. IEEE Intelligent Vehicles Symp, p.1311-1316. https://doi.org/10.1109/IV47402.2020.9304735

Zhong YX, Cao Z, Zhu MH, et al., 2020. CLAP: cloud-and-learning-compatible autonomous driving platform. IEEE Intelligent Vehicles Symp, p.1450-1456. https://doi.org/10.1109/IV47402.2020.9304828

Zhou WT, Jiang K, Cao Z, et al., 2020. Integrating deep reinforcement learning with optimal trajectory planner for automated driving. IEEE 23rd Int Conf on Intelligent Transportation Systems, p.1-8. https://doi.org/10.1109/ITSC45102.2020.9294275