控制与决策
Control and Decision

基于MCPDDPG的智能车辆路径规划方法及应用

余伶俐， 魏亚东， 霍淑欣

引用本文:
余伶俐, 魏亚东, 霍淑欣. 基于MCPDDPG的智能车辆路径规划方法及应用[J]. 控制与决策, 2021, 36(4): 835−846.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.0460

基于近端强化学习的股价预测方法

Method of stock prices forecast based on proximal reinforcement learning

控制与决策. 2021, 36(4): 967−973 https://doi.org/10.13195/j.kzyjc.2019.1245

基于MCPDDPG的智能车辆路径规划方法及应用

余伶俐†, 魏亚东, 霍淑欣

(中南大学 自动化学院，长沙 410083)

摘 要: 针对智能车路径规划过程中常存在动态环境感知预估不足的问题,使用基于蒙特卡罗深度策略梯度学习(Monte Carlo prediction deep deterministic policy gradient, MCPDDPG)的智能车辆路径规划方法,设计一种基于环境感知预测、行为决策和控制序列生成的框架,实现实时的决策和规划,并输出连续的车辆控制序列. 首先,利用序贯蒙特卡罗预估他车行为状态量;然后,设计基于强化Q学习的行为决策方法,使智能车辆实时预知碰撞风险,采取合理的规避策略;最后,构建深度策略梯度学习网络框架,获取智能车辆规划路径的最优轨迹序列. 实验结果表明,所提方法能够缓解环境感知的预估不足问题,提升智能车辆行为决策的快速性,保障路径规划的主动安全,并输出连续的轨迹序列,为智能车辆导航控制提供前提.

关键词: 路径规划；蒙特卡罗预测；智能车辆；深度策略梯度；强化学习；决策

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.0460 **开放科学(资源服务)标识码(OSID):**

引用格式: 余伶俐,魏亚东,霍淑欣. 基于MCPDDPG的智能车辆路径规划方法及应用[J]. 控制与决策, 2021, 36(4): 835-846.

The method and application of intelligent vehicle path planning based on MCPDDPG

YU Ling-li†, WEI Ya-dong, HUO Shu-xin

(College of Automation，Central South University，Changsha 410083，China)

Abstract: Aiming at the problem of insufficient dynamic environment perception and estimation in the process of intelligent vehicle path planning, we design a frame based on environment perception prediction、behavior decision and control sequence generation with an intelligent vehicle path planning method based on MCPDDPG (Monte Carlo prediction deep deterministic policy cradient). The framework can realize a real-time decision-making and planning for intelligent vehicle, and output continuous vehicle control sequences. Firstly, we use sequential Monte Carlo to estimate the behavioral state of other cars; Then, we design a behavioral decision method based on reinforcement Q learning to enable intelligent vehicles to predict collision risks in real time and adopt reasonable avoidance strategies; Finally, we build a deep deterministic policy gradient learning network to obtain the optimal trajectory sequence of the intelligent vehicle planning path. Experimental results show that the proposed method can alleviate the problem of insufficient prediction of environmental perception, improve the speed of intelligent vehicle behavior decision-making, ensure the active safety of path planning, and output a continuous trajectory sequence, which provides a prerequisite for intelligent vehicle navigation control.

Keywords: path planning；Monte Carlo prediction；intelligent land vehicle；deep deterministic policy gradient；reinforcement learning；decision-making

0 引 言

随着智能驾驶技术研究持续推进,人们对智能驾驶系统主动安全和可靠性能提出了更高要求,这是智能驾驶研究者共同面对的问题[1]. 为此,在提升感知系统的高效建模基础上,若能提前对目标行为进行预测,将有助于及时作出高效决策. 为了规划智能车辆安全合理的行驶路径,根据感知数据实施对环境目标行为状态的预估. 预测前方车辆运动轨迹和状态信息,评估碰撞风险[2],并为决策规划提供前提. 其中最为常见的面向运动学模型的预测方法,即基于车辆位

收稿日期: 2019-04-13；**修回日期:** 2019-11-12.

基金项目: 国家重点研发计划项目(2018YFB1201602)；国家自然科学基金项目(61976224)；湖南省科技重大专项(2017GK1010).

责任编委: 魏秀琨.

†通讯作者. E-mail: llyu@csu.edu.cn.

置、速度、加速度等描述车辆状态[3]. 对目标行为进行预测的并不多见,如目标的转弯与否,路况中目标状态等[4]. 文献[5]和文献[6]分别利用高斯模型、卡尔曼滤波器对多目标进行有效的跟踪预测. 一般而言,在非线性和非高斯噪声的驾驶环境中,蒙特卡罗方法更为合适[7]. 文献[8]和文献[9]对交通流复杂的动态随机行为进行仿真,验证了蒙特卡罗方法对动态随机行为预估准确度的优势. 因此,针对交通路况动态不确定性特点,引入蒙特卡罗方法,将能够协助驾驶行为决策与路径规划提前预判.

强化学习按照策略更新方式主要分为基于值函数的强化学习方法和基于直接策略搜索的强化学习方法. 值函数更新方法主要有蒙特卡罗方法、时序差分学习方法(如同策略Sarsa算法和异策略Q学习算法)和利用神经网络拟合Q值的DQN算法;直接策略搜索方法主要包括基于策略梯度的强化学习方法(如policy gradient、Actor-Critic、蒙特卡罗策略梯度)和基于确定性策略搜索的强化学习方法(如DDPG). 文献[10]提出将深度神经网络与强化学习结合起来用以处理离散动作连续状态的深度Q网络(deep Q network, DQN). 文献[11]将强化学习运用在Atari游戏中,其最终游戏能力可以与专业人类游戏测试员相媲美. 在此基础上,文献[12]开发了具有优先级的经验框架,在DQN中使用优先体验重放(priority experience replay),以便更频繁地重复关键的转换. 文献[13]提出了一种新的无模型强化学习神经网络结构dueling,其包括两个独立的估计器,一个用于状态值函数,另一个用于状态相关的动作优势函数. 随后,文献[14]和文献[15]提出基于深度Q网络的离线深度强化学习算法,并将其扩展到连续高维的状态空间. DQN提高了高维状态空间的处理能力,但对于高维连续动作空间仍存在不足[16]. 为此,深度策略梯度(deep deterministic policy gradient, DDPG)以DQN为基础,采用了actor-critic框架,综合了文献[12]和文献[13]的思想,将基于值函数和基于策略梯度的算法结合起来[17],解决了连续状态空间和连续动作空间的问题. 与强化学习相比,通过端到端学习,DDPG能直接基于原始输入输出数据进行控制,扩展了强化学习[18]的使用范围. DDPG采用确定性策略,保证了网络的收敛性. 尽管DDPG的行为策略是确定的,却通常使用随机策略来获得经验,从而能够更好地探索环境. 这种探索性策略往往是通过在行为策略中添加噪声来实现的. 文献[19]在神经网络的参数空间中添加噪声,提高系统抗干扰能力,使其以更强健的方式学习. DDPG在连续控制问题上具有高效性能,文献[20-22]皆采用DDPG控制对象保持平衡或完成既定目标,文献[23]还综合了DDPG和异步优势动作评价(asynchronous advantage actor critic, A3C)算法,提高了汽车转向的平稳性能.

本文针对智能车路径规划过程中决策实时性及控制序列突变问题,设计一种基于环境感知预测、行为决策和控制序列生成的框架,并将多个单一的强化学习算法应用到智能车路径规划系统的不同子任务中. 通过提取环境信息预测前方目标车辆状态变化,并设计强化Q学习行为决策方法,预估风险并及时决策规避,预留足够安全距离完成避障或减速跟车,以提高安全性,实现决策快速性. 最后,构建深度强化学习最优智能驾驶策略,获取最优控制轨迹序列,对完成连续轨迹序列输出做到稳定可靠控制.

1 问题描述

基于环境感知预测、行为决策和控制序列生成的智能车辆路径规划框架如图1所示,目标是实现他车状态预测及无碰撞路径规划.

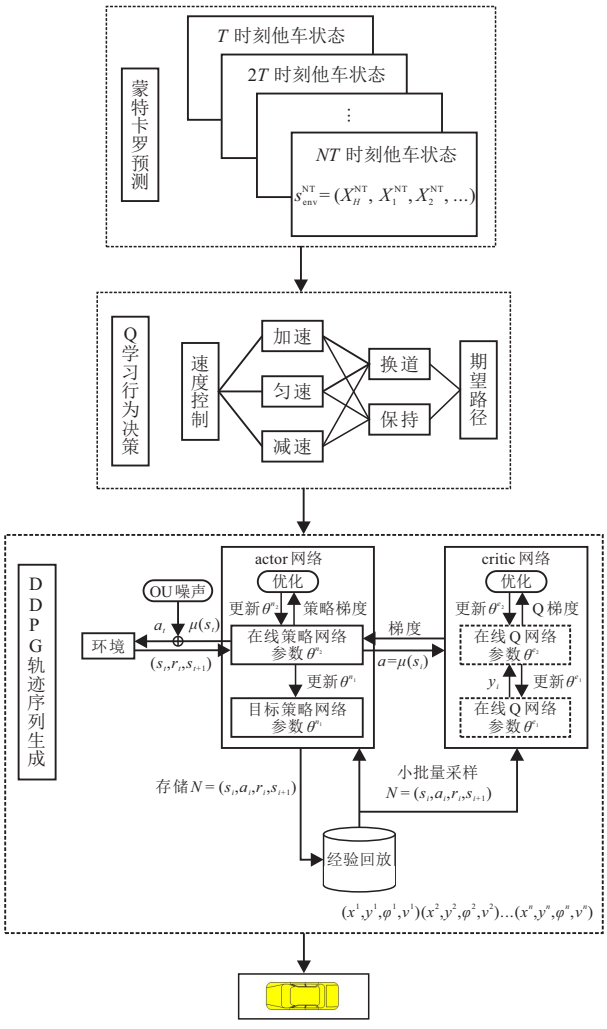


图1 智能车辆路径规划框架

智能车辆路径规划分为决策和规划两个部分,其中决策使用Q学习,规划使用深度策略梯度学习网络. 决策输入为蒙特卡罗输出的未来他车状态,决策输出为智能车辆是否换道及速度规划(加速、减速或匀速),并将决策结果作为规划的输入,规划输出为主车位姿和转角的控制序列,最后实现车辆行驶的路径规划. 本文第2节阐述了蒙特卡罗的前方目标车辆状态预测,着重于根据感知信息完成对环境车辆未来t时刻状态的推导过程. 其结果直接影响第3节强化Q学习的智能车辆行为决策估计. 在此基础上,第4节利用深度强化学习框架生成轨迹序列. 由于决策与轨迹序列生成均使用的是强化学习的方法,这里着重介绍各模型的状态定义及各自回报函数的选择.

考虑3个场景智能驾驶行为预测和路径规划: 直道超车、弯道行驶、匝道行驶. 主要涉及交通参与者包括主车X_H和两个环境车辆X_K,其中主车位姿$[x_H, y_H, \varphi_H]$表示主车横纵坐标位置及航向角,v_H表示主车的速度. 同理,另两个环境车辆的位姿及速度分别表示为$[x_1, y_1, \varphi_1]$、v_1及$[x_2, y_2, \varphi_2]$、v_2. $L_i(i=1,2)$表示主车和环境车辆相对位置,状态预测限制在$t \in [0, t_{\text{finish}}]$内.

2　基于蒙特卡罗的前方目标车辆状态预估方法

2.1　前方目标车辆状态的预测

目标车辆t时刻的状态$s_K^t = [x_K^t, y_K^t, v_K^t]^{\mathrm{T}}$包括横纵坐标$x_K^t$、$y_K^t$的位置和速度$v_K^t$,目标车辆在极短的预测周期$T$内,可看作近似匀速直线运动,所以忽略其航向角. 目标车辆t时刻的状态s_K^t取决于上一时刻的状态s_K^{t-1}和过程噪声$w(t)$,有状态方程

$$s_K^{t+1} = F s_K^t + Q w(t). \tag{1}$$

F是相邻时刻状态的关系系数,Q是过程噪声的调整系数. 假设本车t时刻的状态为s_H^t. 观测目标车辆需要求出本车与目标之间的角度,而这过程又受到观测噪声$v(t)$的污染,故观测方程Z_t表示为

$$Z_t = \arctan\frac{y_K^t - y_H^t}{x_K^t - x_H^t} + v(t). \tag{2}$$

根据观测数据$Z_{1:t}$(后验知识)递推计算当前t时刻目标车辆状态的可信度,用概率公式$p(s_K^t|Z_{1:t})$表示. 该过程分为预测和修正两个步骤: 首先,利用车辆模型的运动学方程预测状态的先验概率密度$p(s_K^t|Z_{1:t-1})$,假设车辆状态的转移满足一阶马尔科夫模型,即当前时刻的状态s_K^t只与上一时刻的状态s_{K-1}^t有关. 结合贝叶斯公式,则有

$$p(s_K^t|Z_{1:t-1}) = \int p(s_K^t|s_K^{t-1}) p(s_K^{t-1}|Z_{1:t-1}) \mathrm{d}s_K^{t-1}, \tag{3}$$

其中$p(s_K^t|s_K^{t-1})$由目标车辆状态方程决定. 而后,利用最新的测量值(或观测值)对先验概率密度进行修正,得到后验概率密度$p(s_K^t|Z_{1:t})$,进而对之前的预测进行修正. 该后验概率将代入下一次的预测,形成递推,有

$$p(s_K^t|Z_{1:t}) = \frac{p(Z_t|s_K^t) p(s_K^t|Z_{1:t-1})}{p(Z_t|Z_{1:t-1})}. \tag{4}$$

归一化

$$p(Z_t|Z_{1:t-1}) = \int p(Z_t|s_K^t) p(s_K^t|Z_{1:t-1}) \mathrm{d}s_K^t, \tag{5}$$

$p(Z_t|s_K^t)$也称为似然函数. 以上推导过程中采用了积分操作,对于非线性非高斯系统,难以得到后验概率的解析解,为此引入蒙特卡罗采样. 根据目标车辆的后验概率分布采样到一系列的样本$x_1, \ldots, x_N$,直接对采集到的N个样本求平均值,以代替样本的均值. 由于后验概率未知,为每个样本引入权重$W_t(s_K^t)$. 根据重要性密度函数$q(s_K^t|Z_{1:t})$进行采样,并计算采样点的权重

$$W_t(s_K^t) \propto \frac{p(s_K^t|Z_{1:t})}{q(s_K^t|Z_{1:t})}. \tag{6}$$

权重经过了归一化处理后,求出该目标车辆状态s_K^t的均值. 重复采样过程,更新采样点权重,完成目标车辆状态递推.

2.2　最小安全距离的设定

当预测到目标车辆运动状态后,分析智能驾驶车辆与目标车辆的碰撞关系,为智能车提供避障决策的可靠依据. 为了避免智能车与目标车辆相撞,设计最小安全距离,只要两车距离大于最小安全距离,则认为两车不会相撞. 最小安全距离与目标车辆的位置、相对速度、加速度相关,因此假设目标车辆沿直线行驶,将两车长度的影响融入阈值范围内考虑.

2.2.1　智能车跟车行驶的最小安全距离

在这种场景下,目标车辆位于智能车前方,且当智能车速度v_H或加速度a_H大于目标车辆速度v_K或加速度a_K时,方有可能发生碰撞. 设碰撞时间为t_{col1},则最小安全距离可表示为

$$S_{\text{MIN1}} = \max_t \int_0^t \int_0^\mu a_H(\tau)\mathrm{d}\tau\mathrm{d}\mu + v_H(0)t - x_K^t, \quad t \in [0, t_{\text{col1}}], \tag{7}$$

其中x_K^t为目标车辆在t时刻状态. 当智能车与目标车辆的距离大于最小安全距离时,两车不会相撞.

2.2.2　智能车变更车道行驶的最小安全距离

智能车执行超车或汇入车流等动作时,需要变更车道. 此时,注意是否与侧后方来车发生碰撞. 在这

种场景下,假设此时智能车具有固定航向角φ_H,速度v_H是恒定的,碰撞时间为$t_{\rm col2}$. 最小安全距离分为横向距离和纵向距离,纵向距离$S_{x\rm MIN2}$与跟车行驶情况下的相同,而横向距离$S_{y\rm MIN2}$表示为

$$S_{y\rm MIN2}=\max_t v_H\times t\times\tan\varphi_H,\ t\in[0,t_{\rm col2}]. \tag{8}$$

由于预测过程中会存在一定误差,在决策中设立碰撞条件时应考虑误差因素. 实际生活中,预测安全距离δ应考虑误差的最大值为阈值,同时考虑车身长度控制滞后性,本文决策规划中设定主车和环境车辆的相对距离$S_{\rm MIN}|_{t=N'T}+\delta>L_1$为发生碰撞条件.

3 基于强化Q学习的智能车辆行为决策估计

3.1 强化Q学习决策方法

强化学习的目的是使智能车辆在每一个不同的状态s,通过策略π选择相应的动作$a_{\rm action}$,并与环境发生交互,获取环境的回报值R及智能车辆的下一个状态s',迭代循环探索,直至寻找到最优的策略π^*完成相应的任务. 策略选择分为两种,即随机策略和确定性策略:随机策略是在状态s时为每个动作$a_{\rm action}$设置发生概率$p:\pi(a_{\rm action}|s)$,以概率选择作为策略选择;确定性策略则直接根据s选择动作$a_{\rm action}=\pi(s)$.

Q学习作为强化学习里的一种算法,使用贪婪策略选择动作,即选择奖励最大的状态s对应的动作$a_{\rm action}$,基于时序差分方法的异策略Q学习更新方式为

$$\begin{aligned}Q(s,a_{\rm action})=&Q(s,a_{\rm action})+\alpha(R+\\&\gamma\max_a Q(s',a_{\rm action})-Q(s,a_{\rm action})).\end{aligned} \tag{9}$$

其中:s为当前的状态,$a_{\rm action}$为当前状态采取的动作,s'为执行这次行动后到达的下一个状态,R为本次行动的奖励,γ为折扣因数,α为学习效率.

3.2 强化Q学习决策方法

决策的状态参量s包括:主车和环境车辆的速度、主车和环境车辆的车道、主车和环境车辆的相对距离,即状态s定义为三维状态. 训练过程为:状态输入后,首先根据ε-greedy策略选择对应动作,决策状态中主车共有6个动作,换道(0 or1)×车速变化(加速、减速、匀速). 接着通过环境模型的回报函数计算状态-动作序列对的回报值,并判定是否发生碰撞. 若判定会发生碰撞,则更新Q表后开始下一次训练过程;若判定不会发生碰撞,则更新Q表后进入下一状态,本次训练过程直至到达终点或判定会发生碰撞后结束. 重复训练过程,最终Q表达到稳定.

环境模型的回报函数为

$$R=R_l+R_v. \tag{10}$$

$$R_v=R_{v\_\rm change}+R_{v\_\rm level}+R_{v\_\rm scene}. \tag{11}$$

$$R_{v\_\rm change}=0.25a_{\rm accelerate}T.$$

$$R_{v\_\rm level}=0.1v_H-1.$$

$$R_{v\_\rm scene}=\begin{cases}0.5,\ v_H=10\ \&\ L_1<v_HNT\ \&\ L_2\geqslant v_HNT;\\-1,\ L_K\geqslant v_HNT\ \&\ v_H\neq10;\\-35,\ \text{collision}.\end{cases} \tag{12}$$

$$R_l=\begin{cases}-5,L_1<L_2\ \&\ v_1<v_2;\\-150,\ \text{collision}.\end{cases} \tag{13}$$

$$\text{collision}=\begin{cases}1,\ S_{\rm MIN}|_{t=N'T}+\theta>L_1;\\0,\ \text{else}.\end{cases} \tag{14}$$

其中:R_l表示车距奖励,R_v表示车速奖励,$R_{v_\rm change}$表示车速变化奖励,$R_{v_\rm level}$表示车速等级奖励,$R_{v_\rm scene}$表示车速对应场景奖励,T表示决策周期,a表示主车加速度$a_{\rm accelerate}=4\,\rm m/s^2$,主车同车道的环境车辆状态用1表示,邻车道车辆状态用2表示,N表示蒙特卡罗预测最佳周期数目,N'在换道和弯道行驶场景中为3,匝道会车场景中$N'=\text{ceil}(L_{\rm collision-point}/v_H)$且$3\leqslant N'\leqslant N$,ceil表示向上取整函数,$L_{\rm collision_point}$表示主车距离会车点的距离.

训练完成后,将动作选择策略更改为贪婪策略,输入状态,即可输出对应决策结果. 决策结果为:期望路径即根据换道与否确定期望路径为本车道中线或邻车道中线,且不考虑在换道过程中执行下一次换道动作;期望速度即根据速度选择确定期望速度值. 针对测试场景直道超车、弯道行驶及匝道减速让行的决策目的,虚拟环境构建以单向双车道为背景(见图2). 地图长1 000 m,车道宽3.5 m,主车尺寸为2 m×4 m,速度变化范围为6 m/s、8 m/s、10 m/s,环境车辆为匀速运动,速度为8 m/s、10 m/s中随机选取,主车加速度限制为4 m/s^2,决策规划以预测周期$T=0.5\,\rm s$为决策更新周期,以预测误差阈值θ为预测误差输入.

图2 行为决策训练场景

4 基于MCPDDPG轨迹序列生成

智能车辆轨迹序列生成不同于决策的离散机制,其连续的状态空间及动作空间定义,促使轨迹序列的

结果是连续的位姿状态. 基于值函数的强化学习方法, 如采用时序差分法的异策略Q学习, 及同样使用神经网络拟合Q值的DQN算法, 假设考虑文中双车道(车道宽3.6 m)换道场景, 选择纵向道路50 m范围, 以0.1 m×0.1 m×1°(横向位置×纵向位置×航向角)为间隔划分道路, 则对应状态空间有12 960 000种子状态, 值函数方法无法对状态及动作空间很大进行有效求解. 基于直接策略搜索的MCPDDPG既可以利用神经网络拟合策略完成智能车辆在连续动作空间策略拟合, 同时并非使用“hard”更新策略更新网络参数(直接复制所有网络权值属于“hard”更新策略), 而是使用“soft”更新策略(缓慢地追随学习的网络: $\theta' \leftarrow \tau\theta + (1-\tau)\theta', \tau \ll 1$)来限制目标Q值的变化, 而目标Q值的变化缓慢能极大提高算法的收敛性, 故采用深度强化学习的方法来生成轨迹序列. 仿真环境中沿用单向双车道作为直道部分, 在直道最后插入弯道部分如图3所示, 在弯道不进行超车动作且减速动作应在入弯前, 故决策并未加入弯道场景. 转角最大输出 $\delta_{\max} = 0.3$ rad, 期望路径 path_d 为中间车道中心线(如图3所示的点划线), 随机初始位置 $x_H \in [50, 90], y_H \in [0, 7], \varphi_H \in [-\pi/2, \pi/2]$; 最大行驶步数 $\text{Num}_{\max} = 100$, 步长0.1 s.

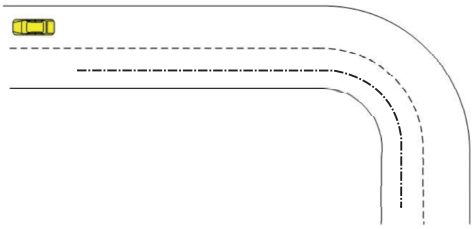


图 3　轨迹序列训练场景

轨迹序列环境模型的回报函数为

$$\text{Reward} = \begin{cases} -1, \text{ when terminates;} \\ R_{\text{action}} + R_{\text{money}}, \text{ otherwise.} \end{cases} \tag{15}$$

$$\begin{cases} R_{\text{action}} = -\lambda_1 \times \|\delta_{\text{old}} - \delta\|^2. \\ R_{\text{money}} = \begin{cases} 0, \text{ get\_money} = \text{False;} \\ 0.1, \text{ get\_money} = \text{True.} \end{cases} \end{cases} \tag{16}$$

$$\text{get\_money} = \begin{cases} \text{True, } |\Delta x| \leqslant \varepsilon_1 \,\&\, |\Delta y| \leqslant \varepsilon_2 \,\&\, |\Delta\varphi| \leqslant \varepsilon_3; \\ \text{False, otherwise.} \end{cases} \tag{17}$$

其中: λ_1 表示正惩罚系数; R_{action} 表示智能车辆前后连续前轮转角 δ_{old} 与 δ 之间的差额惩罚, 连续动作之间的变化越小惩罚越小; R_{money} 表示智能车在期望路径上行驶所获得的奖励; $(\Delta x, \Delta y, \Delta\varphi)$ 表示智能车当前位姿与期望路径 $\text{path}_d = (X_d, Y_d, \phi_d)$ 的差值; $\varepsilon_1, \varepsilon_2, \varepsilon_3$ 表示容错误差. 基于MCPDDPG轨迹序列生成方法由3部分组成, 分别为预测、 决策和深度强化学习轨迹序列生成. 预测部分提供他车的状态信息 s_{env} 结合速度 v 及上一时刻的动作 δ_{old} 合并为 $s_a = (s_{\text{env}}, v, \delta_{\text{old}})$, 并将其作为深度强化学习中Actor策略网络的输入, 其中深度强化学习的策略网络隐藏层使用3个全连接网络, 每层网络包含512个神经元, 且全连接后接BN, 然后使用Relu作为激活函数. 同时, 最后一层网络改用tanh作为激活函数, 将网络的输出映射到区间 $[-1, 1]$ 中. 网络输出为动作 $\delta \in \sum(\delta)$, 状态 s_a 与动作 δ 合并后为 $s_c = (s_a, \delta)$, 将其作为Critic评价网络的输入. 评价网络隐藏层与策略网络结构相同, 最后一层网络用 $y = kx + b$ 的线性函数激活, 网络的输出为 s_c 对应的Q值 $Q(s_a, \delta)$. 其中: x 为最后一层的输入, y 为预测的Q值, k、b 为网络训练后的权重与偏置.

5　实验与分析

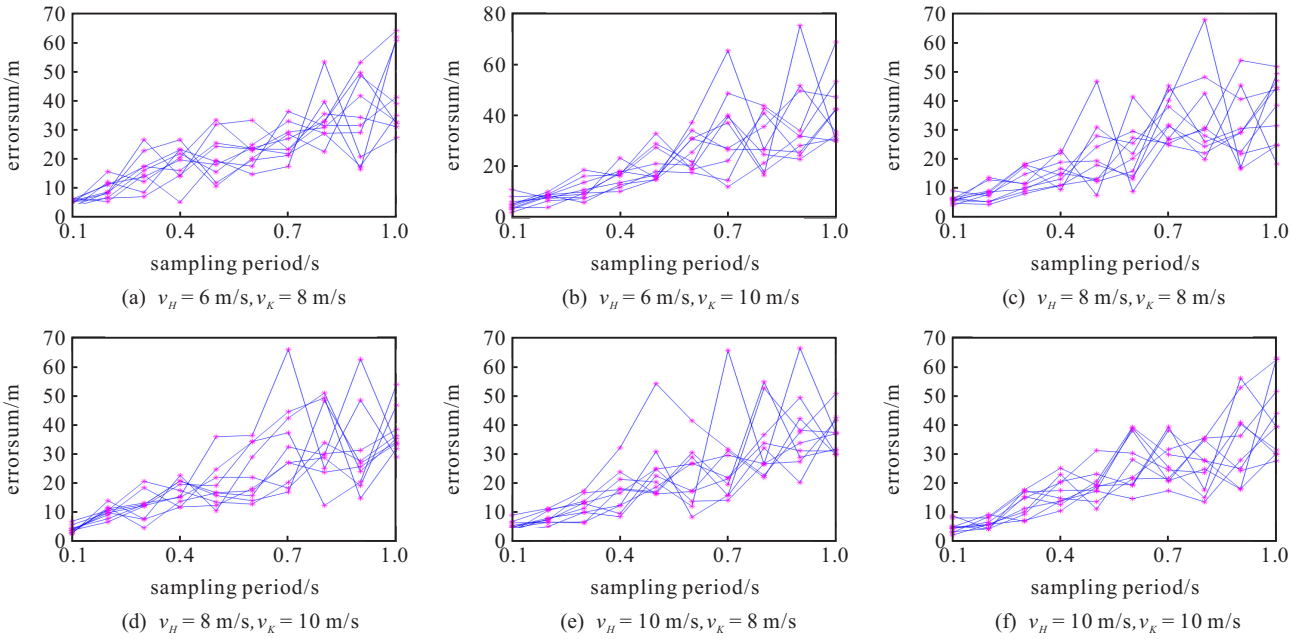
5.1　基于蒙特卡罗车辆状态预测实验分析

5.1.1　最佳预测周期估计

定义目标车辆的预测位置与实际位置之间距离为 R, 则 R 表示为

$$R = \sqrt{\left(x_K^{t'} - x_K^t\right)^2 + \left(y_K^{t'} - y_K^t\right)^2}, \tag{18}$$

其中 $x_K^{t'}$ 和 $y_K^{t'}$ 分别是目标车辆的预测纵向位移和横向位移. 距离误差 R 评判预测结果优劣的指标, 其应在一定的范围 $((0, \theta])$ 内, 故而误差阈值越小, 预测结果越准确. 预测误差与采样周期、预测距离、主车和环境速度有关. 通过确定最佳的采样周期和预测时间窗, 便可得出不同驾驶速度下的误差阈值. 假定10个观测点, 各观测点预测30个周期, 采样周期 T 在0.1 s～1 s范围内. 由于在决策规划训练中, 智能车的速度在6 m/s～10 m/s之间变化. 假定智能车速度是6 m/s、 8 m/s、 10 m/s, 目标车辆速度分别是8 m/s、10 m/s, 共6组实验, 探究 R 与采样周期 T 之间的联系. 图4表示在不同驾驶速度设定下, 当采样周期取0.1 s～1 s时, 10个观测点分别预测30个周期的预测误差总和. 其中: 横坐标表示采样周期 T, 以0.1 s为间距在0.1 s～1 s的范围内取值, 纵坐标表示在某一观测点预测多个周期的误差总和. 以图4(a)为例, 该图表示智能车以6 m/s速度行驶, 目标车辆以8 m/s速度行驶, 当采样周期在0.1 s～1 s的范围内取值时, 10个观测点分别预测30个周期的预测误差总和. 图4(b)～4(f)则是在图4(a)实验基础上改变智能



(a) $v_H = 6$ m/s, $v_K = 8$ m/s (b) $v_H = 6$ m/s, $v_K = 10$ m/s (c) $v_H = 8$ m/s, $v_K = 8$ m/s

(d) $v_H = 8$ m/s, $v_K = 10$ m/s (e) $v_H = 10$ m/s, $v_K = 8$ m/s (f) $v_H = 10$ m/s, $v_K = 10$ m/s

图 4 不同采样周期下的预测误差和

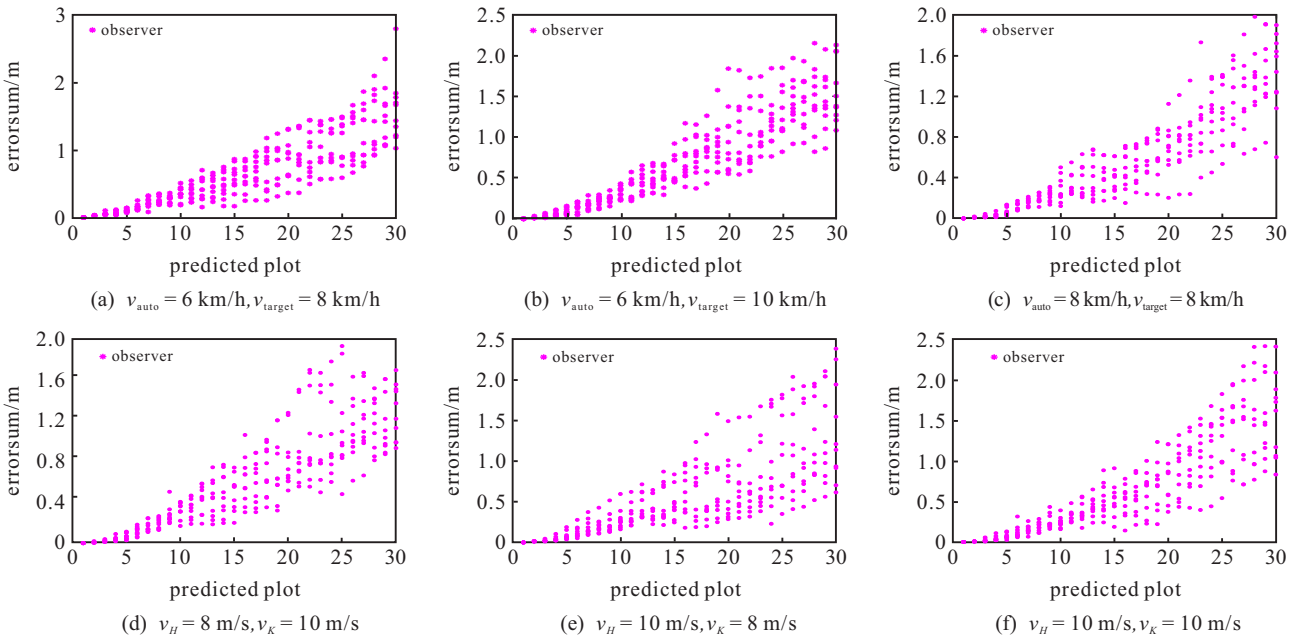
车和目标车辆的初始速度后的实验结果. 某一观测点的预测误差总和随着采样周期增大而增大. 为了满足规划周期的需求,选取采样周期$T = 0.5$ s.

5.1.2 预测时间窗估计

图5是在不同观测点下各预测点的误差均值,横坐标表示30个预测点,纵坐标表示某一观测点对某一预测点多次预测后的误差取平均值. 以图5(a)为例,该图表示智能车以6 m/s速度行驶,目标车辆以8 m/s速度行驶,10个观测点分别对30个预测点做多次预测实验后,各预测点的误差平均值. 同样地,图5(b)~5(f)是在图5(a)实验基础上改变智能车和目标车辆的初始速度后的实验结果. 由图5可知,随着预测距离的增大,误差逐渐积累,预测的准确度会下降,所以在采用预测结果前需要剔除若干点只保留前$N_{\rm pre}$个预测点. $N_{\rm pre}$的约束条件有

$$0 < N_{\rm pre} \leqslant \frac{20}{T \times v_K} N_{\rm pre} \in N^*. \tag{19}$$

其中: v_K表示环境车辆的速度,20表示预测距离(m),最大有效预测点数$N_{\rm pre}$在4或5中选取,预测时间窗$T_{\rm wd} = N_{\rm pre} \times T$.



(a) $v_{\rm auto} = 6$ km/h, $v_{\rm target} = 8$ km/h (b) $v_{\rm auto} = 6$ km/h, $v_{\rm target} = 10$ km/h (c) $v_{\rm auto} = 8$ km/h, $v_{\rm target} = 8$ km/h

(d) $v_H = 8$ m/s, $v_K = 10$ m/s (e) $v_H = 10$ m/s, $v_K = 8$ m/s (f) $v_H = 10$ m/s, $v_K = 10$ m/s

图 5 在不同观测点下各预测点的误差均值

5.2　基于深度强化学习的路径规划实验

5.2.1　基于强化Q学习的行为决策实验分析

主车初始车道为右侧车道,初始位置为(1, 1.75),初始速度为 6 m/s. 环境车辆为匀速运动,速度为 8 m/s或10 m/s,初始位置以[20, 30]之间随机整数作为纵向间隔,随机搭配右侧车道或左侧车道,分布在[0, 1 000]范围内,以主车与环境车辆发生碰撞或主车运行至1 000 m处结束每一次训练,训练次数为3 000,训练结果如图6所示.

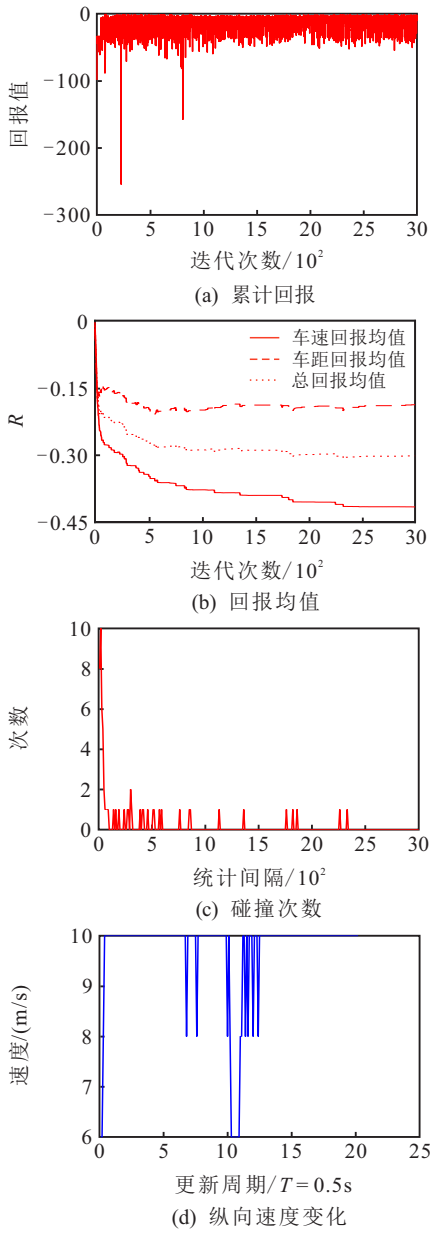


图 6　基于强化Q学习的行为决策实验

图6表示在经历3 000次训练后,主车已经基本识别前方障碍车的信息,并基本作出超车决策. 图6(a)表示累计回报值,强化学习的最终目的是获得最大累计奖励,而设计的奖励基本都是负值,所以整体系统的奖励为负,并会趋于稳定值 −80 左右. 图6(b)中的虚线表示车距的回报均值,实线表示车速的回报. 图7表示将车速和车距奖励依次放大0、0.1、0.5、1、2、4、10倍下训练过程中碰撞的次数统计. 可见,如果将车速等级奖励和车速改变奖励减小,则主车会以最低速跑完全程;如果加大奖励值,则主车会趋向于以高速跑完全程. 随着训练次数的增加,主车逐渐掌握避让策略,Q表也逐渐趋于稳定,系统稳定程度随迭代次数逐次提高.

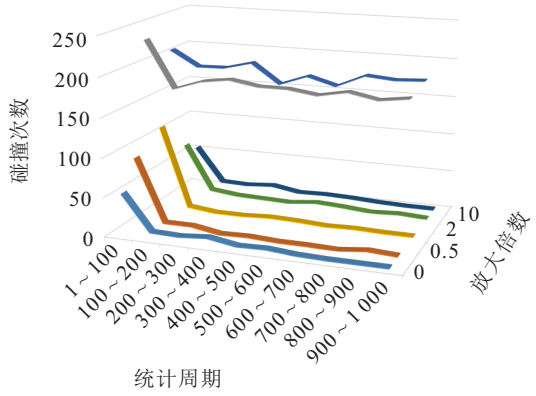


图 7　不同速度回报值比例下训练碰撞次数统计

5.2.2　基于深度强化学习的轨迹序列生成参数分析

深度强化学习的超参数对实验影响如下:学习速率越大,保留之前训练的效果越少. 折扣因子越大,越重视以往经验;折扣因子越小,越重视当前回报. 隐层层数、隐层神经元数量过少,不能很好地对数据进行拟合;隐层层数、隐层神经元数量过多,容易导致过拟合. 经多次试验尝试,选择一较优的网络结构和网络参数,深度强化学习模型中的超参数设置为:折扣因子 $\gamma = 0.9$, Actor与Critic网络的学习率均为 10^{-4},优化方法使用Adam,软更新率为 $\tau = 0.001$,隐藏层均使用3个全连接网络,每层隐藏层神经元个数均为512,经验回放池大小为 10^4,批量为64,误差生成使用高斯过程,初始方差 $\mathrm{var}_{\max} = 2$,最小方差 $\mathrm{var}_{\min} = 0.01$,衰减率为 10^{-4}. 实验环境为Linux操作系统,内存16 G,显卡为GTX1080Ti,深度学习框架为Tensorflow.

如图8所示,在训练过程中智能车辆所获累积奖励和平均Q值随着迭代次数增加逐渐上升,最后趋于稳定. 损失随着迭代次数的增加逐渐降至0附近,说明评价网络愈加有效. 噪音随着迭代次数增加不断降低,在早期提供智能车辆足够的勘探探测能力,而在后期提供智能车辆足够的挖掘开采能力. 由于随机选择因数设置波动,导致策略选择出现了小范围的改变,致使回报均值和损失在70 000步和90 000步时出现了抖动. 以上说明模型在训练时从经验中学习,向最优策略逼近.

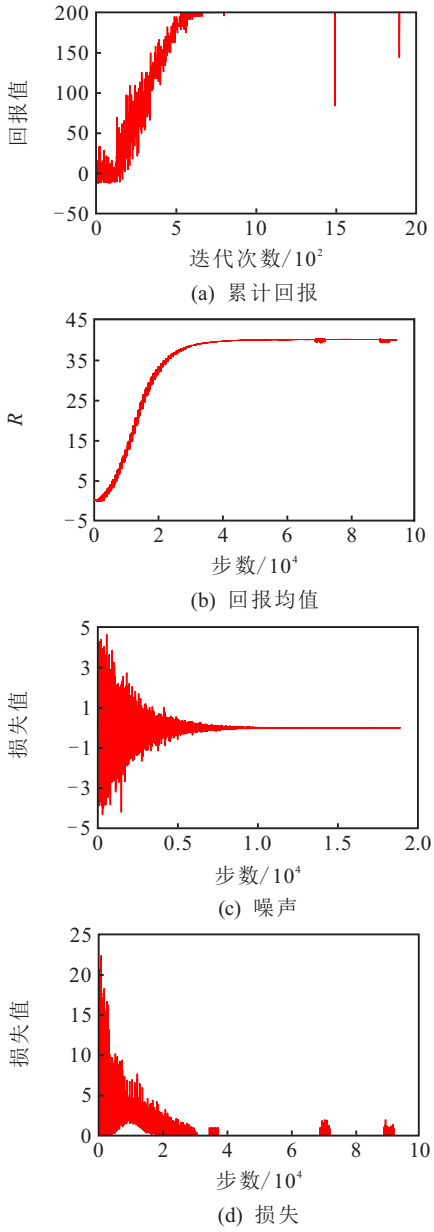


图 8 基于深度强化学习轨迹序列实验

5.3 面向不同场景的轨迹序列生成对比实验分析

强化学习的路径规划算法是通过智能车不断探索环境,以环境回报值大小的反馈,最终学习“驾驶”技能;端到端路径规划方法是模仿人类驾驶员驾驶行为,枚举出车辆可以行驶的所有轨迹,并且对于任意的起始/目标状态对都能学习并建立其与对应方向盘控制动作之间的直接映射关系,完成控制动作.

与MCPDDPG算法所拥有相同框架的另一种可以应用于连续领域的强化学习A3C算法,在打破训练数据具有相关性的本质上有所不同,MCPDDPG利用了经验回放的技巧,而A3C采用的是异步学习方法,故本文将MCPDDPG算法与A3C算法及端到端轨迹规划算法在直道超车、弯道跟随、匝道减速避让3个场景进行对比实验,且使用Q学习行为决策产生相同的决策结果. 其中A3C实验所有环境场景及回报值设定均与文中所设一致,且同时探索环境智能体为2个.

5.3.1 直道超车

直道超车场景如图9所示,智能车检测到本车道前方和侧向车道内均有环境车辆. 其中:主车速为$V_H = 10\,\text{m/s}$,环境车辆速度为$V_1 = V_2 = 8\,\text{m/s}$,$L_1 = 6\,\text{m}$,$L_2 < 0$. 根据预测及决策结果,3个决策周期$T$后有碰撞风险,执行减速超车的决策动作,减速到$V_H' = 8\,\text{m/s}$,并执行换道动作. 原期望路径为$Y = 2.5\,\text{m}$,换道期望路径为$Y = 6\,\text{m}$.

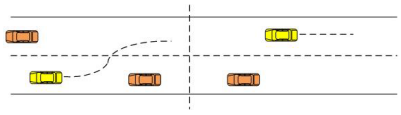


图 9 直道超车场景

图10表示直道超车使用MCPDDPG轨迹规划方法所得结果,图10(a)表示主车行驶轨迹,图10(b)表示主车航向角随周期更新的变化,图10(c)表示主车方向盘转角随周期更新的变化. 图11表示使用A3C轨迹规划方法所得结果,图12表示使用端到端轨迹规划方法所得结果. 根据行驶轨迹显示,MCPDDPG轨迹规划方法在直道超车时,能在较小纵向行驶距离下完成换道,且完成换道后相较于A3C及端到端轨迹规划方法横向偏差波动范围较小,轨迹更加平稳;根据航向角和转角控制序列显示,MCPDDPG轨迹规划方法在换道超车过程中,方向盘控制更加平滑,且换道成功后相较于其他两种算法的方向盘调整幅度更小,不会出现方向盘震荡,舒适度更高.

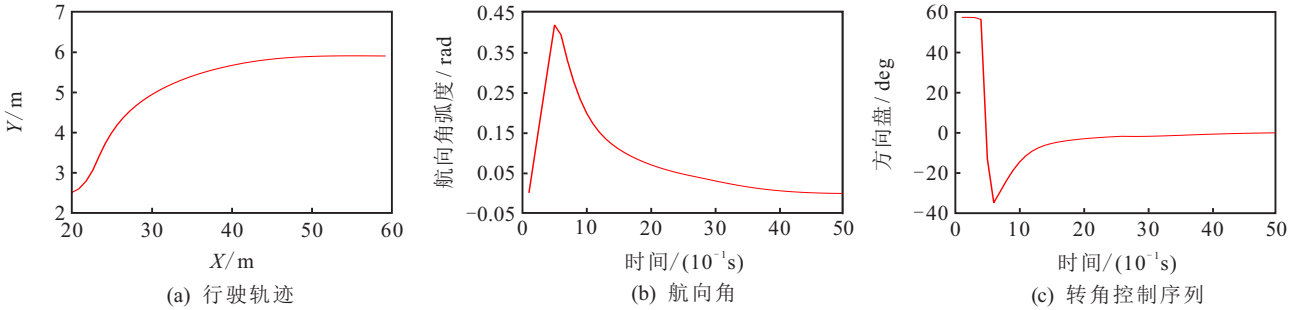
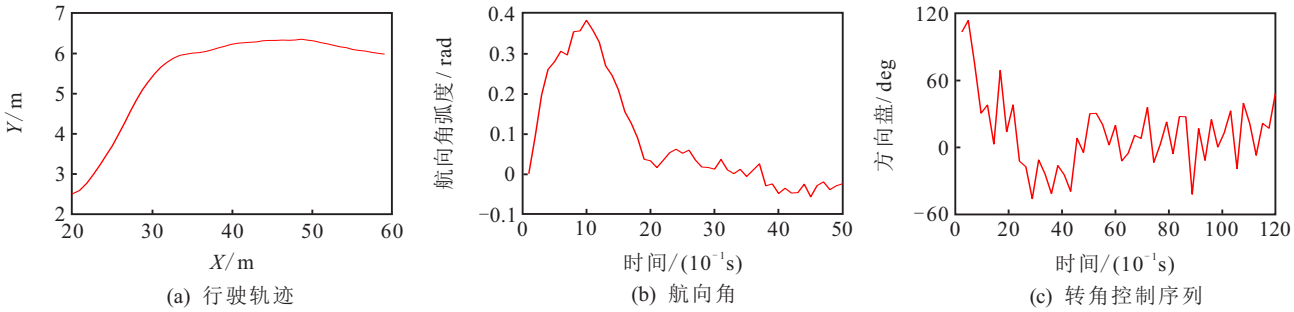
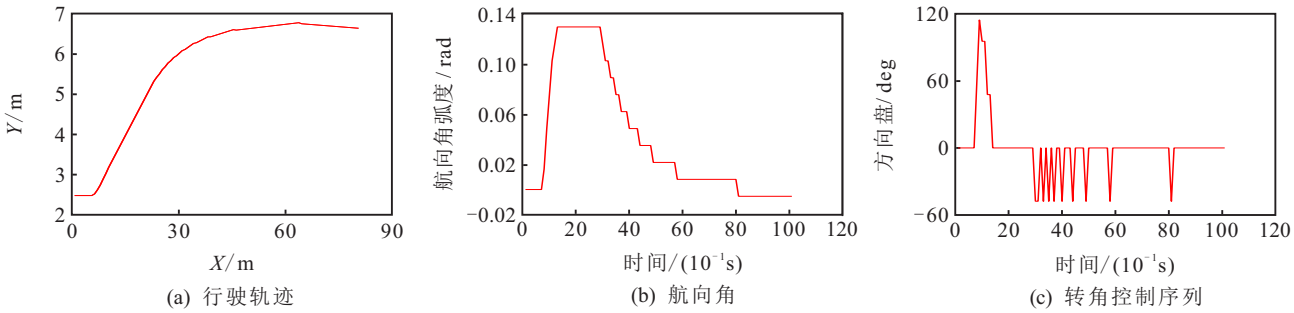


图 10 MCPDDPG-直道超车结果



(a) 行驶轨迹　(b) 航向角　(c) 转角控制序列

图 11　A3C-直道超车结果



(a) 行驶轨迹　(b) 航向角　(c) 转角控制序列

图 12　端到端-直道超车结果

5.3.2　弯道跟随

弯道跟随前车场景如图13所示,智能车检测到本车道前方道路环境为弯道不允许超车,同时本车道前方有低速行驶的车辆. 其中: 主车速为$V_H = 10\,\mathrm{m/s}$,环境车辆速度为$V_1 = 8\,\mathrm{m/s}$,$L_1 = 6\,\mathrm{m}$. 根据预测及决策结果,3个决策周期T后有碰撞风险,执行减速跟随动作,速度减到$V_H' = 8\,\mathrm{m/s}$.

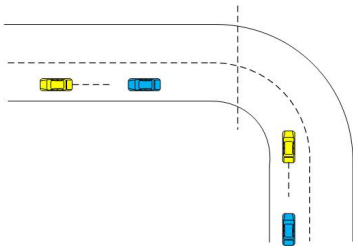
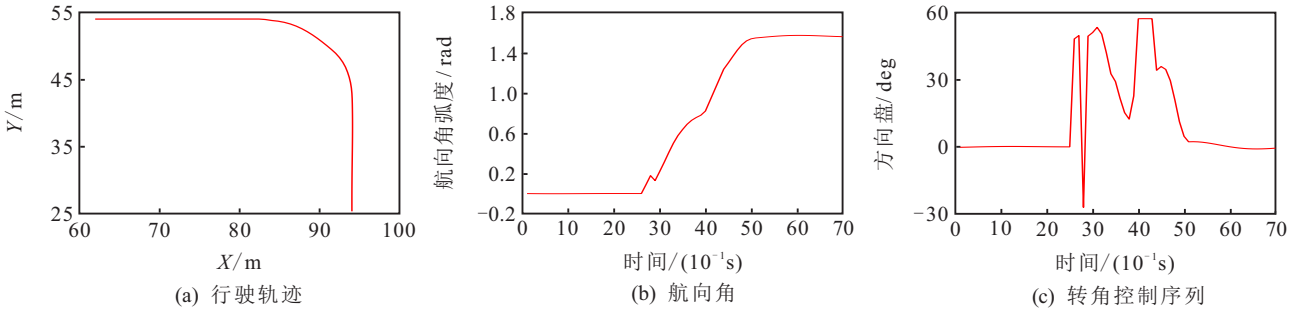
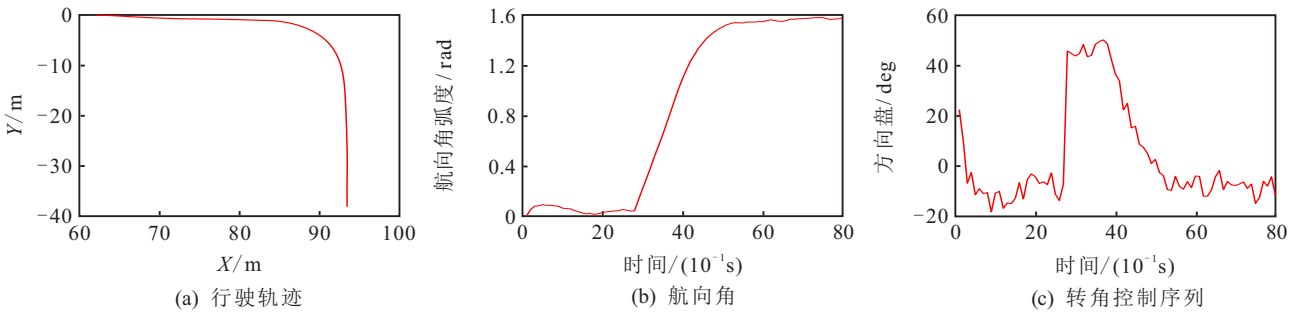


图 13　弯道场景



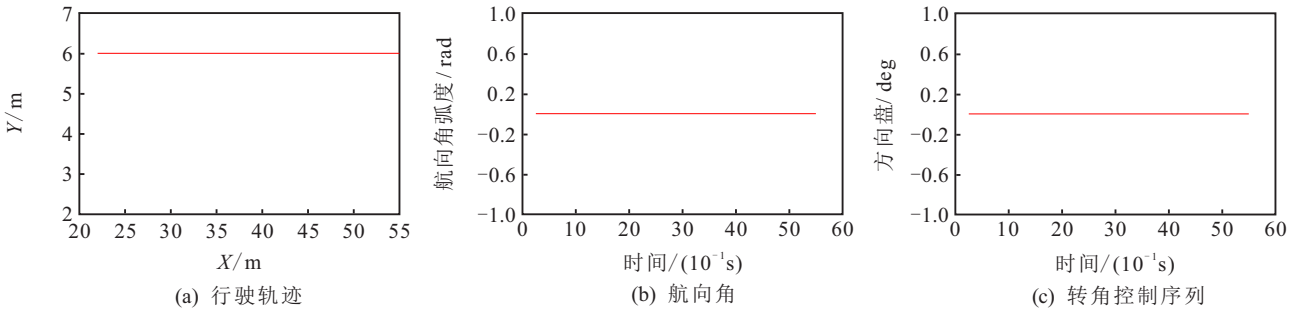
(a) 行驶轨迹　(b) 航向角　(c) 转角控制序列

图 14　MCPDDPG-弯道减速跟随结果



(a) 行驶轨迹　(b) 航向角　(c) 转角控制序列

图 15　A3C-弯道减速跟随结果



(a) 行驶轨迹 (b) 航向角 (c) 转角控制序列

图 16 端到端-弯道减速跟随结果

图14表示弯道行驶使用MCPDDPG轨迹规划方法所得结果,图14(a)表示主车行驶轨迹,图14(b)表示主车航向角随周期更新的变化,图14(c)表示主车方向盘转角随周期更新的变化. 图15表示使用A3C轨迹规划方法所得结果,图16表示使用端到端轨迹规划方法所得结果. 根据行驶轨迹显示,MCPDDPG轨迹规划方法在弯道行驶时,相较于A3C及端到端轨迹规划方法行驶轨迹更加贴合道路几何描述,轨迹更加平稳,横向误差更小;根据航向角和转角控制序列显示,MCPDDPG轨迹规划方法在弯道行驶时,相较于其他两种算法的方向盘控制更加平滑,同时航向角变化更加平滑,舒适度更高.

5.3.3 匝道会车

匝道减速避让场景如图17所示,智能车检测侧向车道和辅道内均有环境车辆. 其中:主车速为$V_H = 10\,\mathrm{m/s}$,距离匝道会车点$L_0 = 20\,\mathrm{m}$,侧向车辆速度为$V_1 = 8\,\mathrm{m/s}$,$L_1 = 4\,\mathrm{m}$,辅道内车辆$V_2 = 6\,\mathrm{m/s}$,$L_2 = 5\,\mathrm{m}$. 根据预测及决策结果,判定有碰撞风险,执行减速避让动作,速度减到$V_H' = 6\,\mathrm{m/s}$,期望路径为$Y = 6\,\mathrm{m}$.

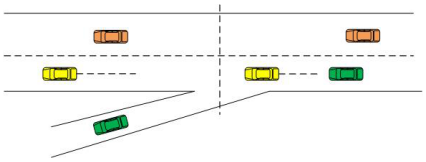
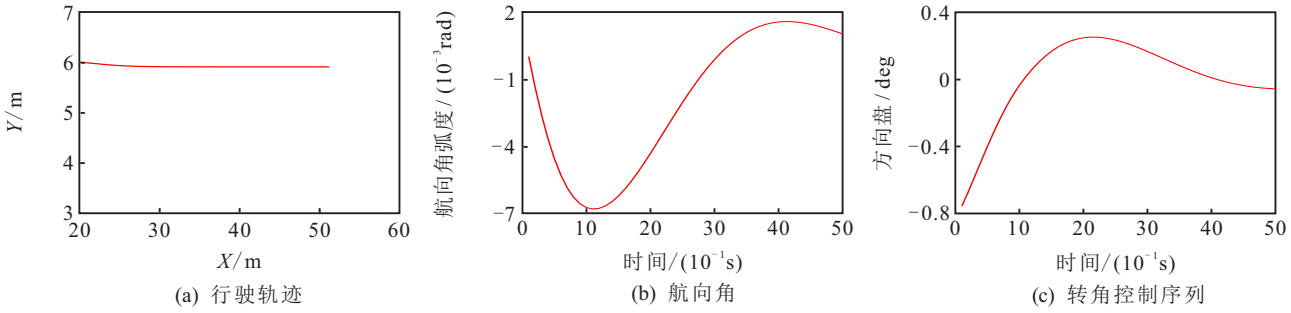


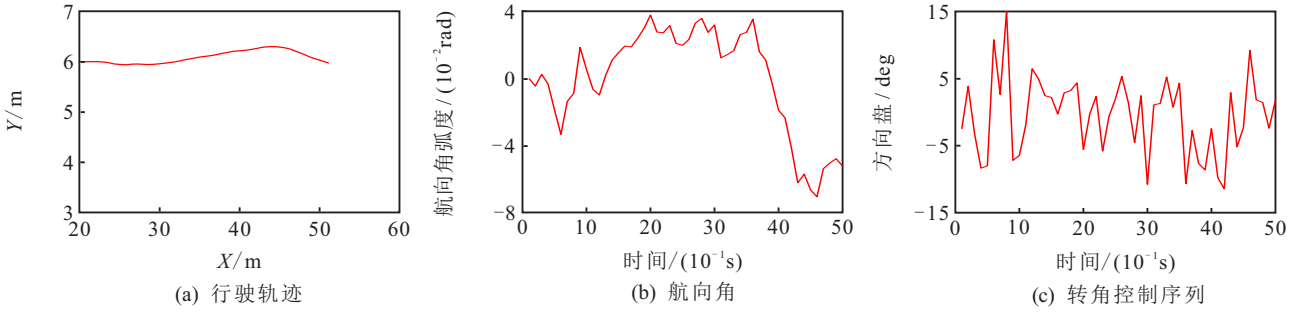
图 17 匝道场景

图18表示匝道行驶使用MCPDDPG轨迹规划方法所得结果,图18(a)表示主车行驶轨迹,图18(b)表示主车航向角随周期更新的变化,图18(c)表示主车方向盘转角随周期更新的变化. 图19表示使用A3C轨迹规划方法所得结果,图20表示使用端到端轨迹规划方法所得结果. 根据行驶轨迹显示,MCPDDPG轨迹规划方法在匝道减速行驶时,相较于A3C及端到端轨迹规划方法行驶轨迹更加平稳,横向误差更小;根据航向角和转角控制序列显示,MCPDDPG轨迹规划方法在匝道减速行驶时,方向盘控制更加平滑,震荡范围较小,且航向角变化更加平滑,舒适度更高.



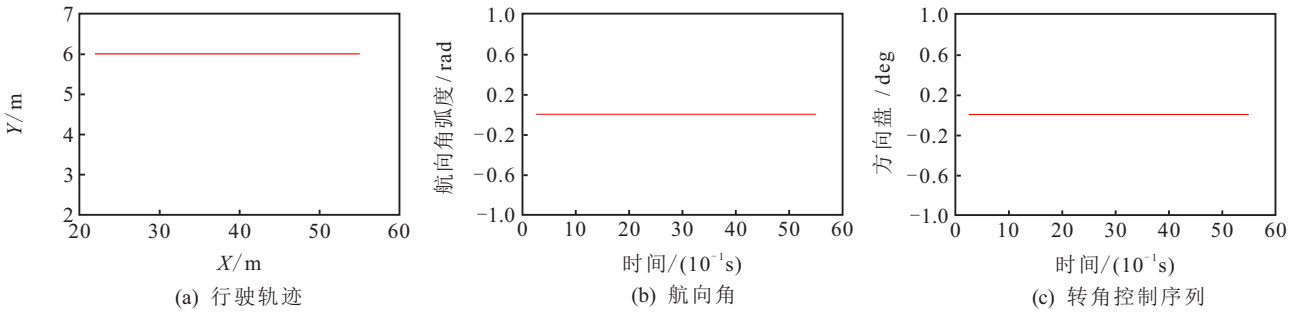
(a) 行驶轨迹 (b) 航向角 (c) 转角控制序列

图 18 MCPDDPG-匝道减速避让结果



(a) 行驶轨迹 (b) 航向角 (c) 转角控制序列

图 19 A3C-匝道减速避让结果

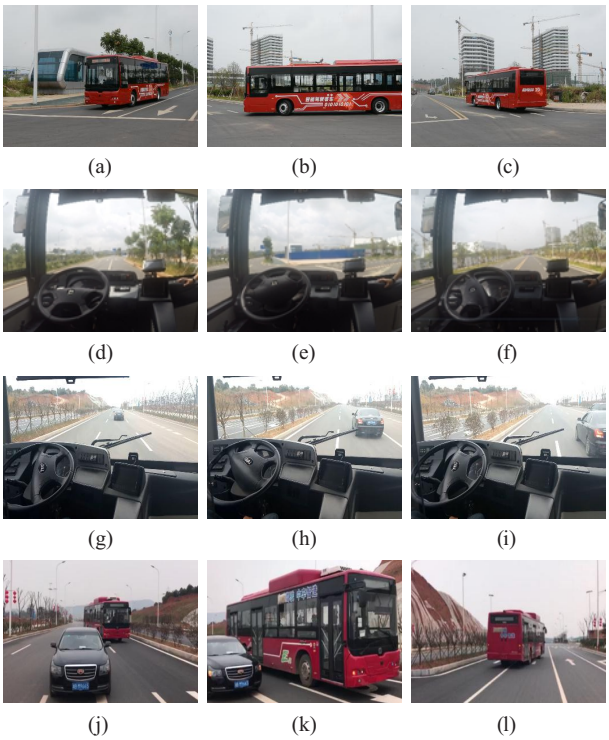


(a) 行驶轨迹　(b) 航向角　(c) 转角控制序列

图 20　端到端-匝道减速避让结果

5.4　实车实验分析

智能车辆通过车载GPS/IMU组合惯导,实时获取车辆位姿,并在车上装载摄像头、激光雷达、毫米波雷达、超声波雷达等,以获取周围环境数据. 图21为实车验证过程,其中图21(a)～图21(f)为转弯过程,图21(a)～图21(c)为车外视角展示,图21(d)～图21(f)为车内视角展示. 智能车在转弯过程中利用本文所述方法生成的连续轨迹序列及平滑的控制量,驾驶平稳、舒适性好. 图21(g)～图21(l)为超车换道过程组图,图21(g)～图21(i)为车内视角展示,图21(j)～图21(l)为车外视角展示. 当智能车检测到前方存在环境车辆,由感知数据输入经过蒙特卡罗预测及行为决策估计后,存在碰撞风险,执行换道避让策略,实现安全智能驾驶.



(a)　(b)　(c)

(d)　(e)　(f)

(g)　(h)　(i)

(j)　(k)　(l)

图 21　实车验证

6　结　论

针对动态环境感知预估不足引发的行为决策快速性不足、控制序列易突变问题,使用基于蒙特卡罗预测及深度策略梯度学习的智能车辆路径规划方法,设计一种基于环境感知预测、行为决策和控制序列生成的框架,并将多个单一的强化学习算法应用到智能车路径规划系统的不同子任务中. 首先,提取环境信息预测他车的状态变化;其次,设计Q学习行为决策方法,使智能车学会预估风险并及时作出决策规避,预留足够安全距离完成避障或减速跟车提高安全性;最后,构建深度强化学习最优智能驾驶策略,获取最优轨迹序列对完成连续轨迹序列输出,做到平稳控制. 通过多个驾驶场景功能验证实验MCPDDPG的优化性能,车辆拥有预估风险能力,争取尽量大的安全距离调整车辆状态,行驶时有更加连续的转角控制序列,实现既快速、安全又平稳的路径规划.

参考文献(References)

[1] 邵淑漫. 无人驾驶核心技术及其未来影响[J]. 科技经济导刊, 2019, 27(2): 76-77.
(Shao S M. Core technology and future impact foruUnmanned vehicles[J]. Technology and Economic Guide, 2019, 27(2): 76-77.)

[2] Kim J, Kum D. Collision risk assessment algorithm via lane-based probabilistic motion prediction of surrounding vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(9): 2965-2976.

[3] Rajamani R. Vehicle dynamics and control[M]. Boston: Springer, 2006: 1-13.

[4] Funfgeld S, Holzapfel M, Frey M, et al. Stochastic forecasting of vehicle dynamics using sequential monte carlo simulation[J]. IEEE Transactions on Intelligent Vehicles, 2017, 2(2): 111-122.

[5] Chae H S, Lee M S, Yi K S. Probabilistic prediction based automated driving motion planning algorithm for lane change[C]. The 17th International Conference on Control, Automation and Systems. Piscataway: IEEE, 2017: 1640-1645.

[6] Du Z L, Li X M. Strong tracking tobit kalman filter with model uncertainties[J]. International Journal of Control, Automation and Systems, 2019, 17(2): 345-355.

[7] Sanchez-Rico M T, Garcia-Rodenas R, Espinosa-Aranda J L. A Monte Carlo approach to simulate the stochastic demand in a continuous dynamic traffic network loading problem[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(3): 1362-1373.

[8] Wong X C, Ahmed S K, Zulkift F, et al. An approach for analyzing queuing systems using markov chain monte carlo methods: A traffic flow case study[C]. 2009 Student Conference on Research and Development. New York: IEEE, 2009: 41-44.

[9] Wong X C, James J A, Ramasamy A K, et al. Effect of highway ramp separation on traffic flow: An investigation using monte carlo simulation[C]. 2010 IEEE Student Conference on Research and Development. Piscataway: IEEE, 2010: 213-217.

[10] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518 (7540): 529-533.

[11] McKenzie M, Loxley P, Billingsley W, et al. Competitive reinforcement learning in atari games[C]. The 30th Australasian Joint Conference on Artificial Intelligence. Cham: Springer International Publishing, 2017: 14-26.

[12] Hou Y N, Liu L F, Wei Q, et al. A novel DDPG method with prioritized experience replay[C]. 2017 IEEE International Conference on Systems, Man and Cybernetics. Piscataway: IEEE, 2017: 316-321.

[13] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning[C]. The 33rd International Conference on Machine Learning. New Jersey: International Machine Learning Society, 2016: 2939-2947.

[14] Yang Z Y, Merrick K, Jin L W, et al. Hierarchical deep reinforcement learning for continuous action control[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(11): 5174-5184.

[15] Gu S X, Lillicrap T, Sutskever U, et al. Continuous deep q-learning with model-based acceleration[C]. The 33rd International Conference on Machine Learning. New Jersey: International Machine Learning Society, 2016: 4135-4148.

[16] Yu L L, Shao X Y, Wei Y D. Intelligent land-vehicle model transfer trajectory planning method based on deep reinforcement learning[J]. Sensors, 2018, 18(9): 2905-2926.

[17] Iriondo A, Lazkano E, Susperregi L, et al. Pick and place operations in logistics using a mobile manipulator controlled with deep reinforcement learning[J]. Applied Sciences-Basel, 2019, 9(2): 54-62.

[18] Liu Q, Zhai J W, Zhang Z Z, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 14(1): 1-27.

[19] Behzadan V, Munir A. Mitigation of policy manipulation attacks on deep q-networks with parameter-space noise[C]. Computer Safety, Reliability, and Security. Cham: Springer International Publishing, 2018: 406-417.

[20] Wu X G, Liu S W, Zhang T C, et al. Motion control for biped robot via DDPG-based deep reinforcement learning[C]. 2018 WRC Symposium on Advanced Robotics and Automation. Piscataway: IEEE, 2018: 40-45.

[21] Tuyen L P, Chung T. Controlling bicycle using deep deterministic policy gradient algorithm[C]. The 14th International Conference on Ubiquitous Robots and Ambient Intelligence. Piscataway: IEEE, 2017: 413-417.

[22] Kim S J, Kim H S, Kang D J. Vibration control of a vehicle active suspension system using a DDPG algorithm[C]. The 18th International Conference on Control, Automation and Systems. Piscataway: IEEE, 2018: 1654-1656.

[23] Yang F, Wang P, Wang X H. Continuous control in car simulator with deep reinforcement learning[C]. The 2nd International Conference on Computer Science and Artificial Intelligence. New York: Association for Computing Machinery, 2018: 566-570.

作者简介

余伶俐 (1983−), 女, 副教授, 博士生导师, 从事智能驾驶、智能机器人导航规划与决策控制、移动机器人感知融合等研究, E-mail: llyu@csu.edu.cn;

魏亚东 (1996−), 男, 硕士生, 从事智能驾驶、智能机器人导航控制应用的研究, E-mail: 13477011934@163.com;

霍淑欣 (1997−), 女, 硕士生, 从事移动机器人决策规划的研究, E-mail: huoshuxin@csu.edu.cn.

(责任编辑：齐 霁)