第51卷第11期
2023年11月

华南理工大学学报（自然科学版）
Journal of South China University of Technology
（Natural Science Edition）

Vol. 51 No. 11
November 2023

doi：10. 12141/j. issn. 1000-565X. 220747

基于改进DDPG的无人驾驶避障跟踪控制

李新凯　虎晓诚　马萍　张宏立

（新疆大学 电气工程学院，新疆 乌鲁木齐 830017）

摘　要：无人驾驶汽车在跟踪避障控制过程中，被控对象具有非线性特征且被控参数多变，线性模型及固定的无人驾驶车辆数学模型难以保证车辆在复杂环境下的安全性和稳定性，并且无人驾驶离散化控制过程增加了控制难度。针对此类问题，为提高无人驾驶汽车实时控制跟踪轨迹的精度，同时降低整个控制过程的难度，文中提出了一种基于蒙特卡洛–深度确定性策略梯度（MC-DDPG）的无人驾驶汽车避障跟踪控制算法。该算法基于深度强化学习网络搭建控制系统模型，在策略学习采样过程中采用优秀的训练样本，使用蒙特卡洛方法优化网络训练梯度，对算法的训练样本采取优劣区分，使用优异的样本通过梯度算法寻找最优的网络参数，从而增强网络算法的学习能力，实现无人驾驶汽车的更优连续控制。在计算机模拟环境TORCS中对该算法进行仿真实验，结果表明，应用MC-DDPG算法可以有效地实现无人驾驶汽车的避障跟踪控制，其控制的无人驾驶汽车的跟踪精度及避障效果均优于深度Q网络算法和DDPG算法。

关键词：无人驾驶；动态避障；深度确定性策略梯度；轨迹跟踪；梯度优化

中图分类号：TP273^{+}.5；U461.1　　**文章编号：**1000-565X（2023）11-0044-12

随着人工智能的发展，交通方式也在不断地发展进步，智慧交通逐渐进入到生活当中[1-2]。无人驾驶作为智慧交通的重要特征受到了越来越多的关注[3]，避障跟踪控制是无人驾驶技术的重要性能之一，无人驾驶汽车需要根据捕捉到的环境信息、障碍物信息，并结合自身当前状态信息灵活控制车速和转向，以实现稳定的避障跟踪[4-5]。

针对无人驾驶车辆在复杂环境工况下的避障跟踪控制，众多学者进行了深入研究[6]。在经典控制方法中，张家旭等[7]提出了汽车紧急换道避障的路径规划与跟踪控制方法，实现在紧急情况下的避障跟踪；Wang等[8]采用PID控制器设计了闪避障碍物驾驶控制模型，并通过Carsim车辆仿真软件验证了该模型的鲁棒性，但PID控制的最优参数难以确定，在变化复杂的环境工况下难以精确控制；樊晓平等[9]提出了基于人工势场的机器人避障方法，根据障碍物在人工势场中的势来判断障碍物，使得机器人可以自主避障，但该方法应用的广泛性以及误差判断的精准度未得到验证；Katsuki等[10]提出了基于图搜索方法的障碍物采样及局部路径规划算法，该算法在障碍物周围密集采样，提高了对障碍物位置的鲁棒性，但图搜索方法的连续性较差，并且在高维空间中图搜索方法的性能并不可靠；Wang等[11]提出了一种基于双层非线性模型预测控制的自动驾驶卡车避障跟踪控制器，实现了平稳避障并重新跟踪上原路径，但模型预测控制需要车辆的高精度数学模型，而精确的数学模型很难建立；Zong等[12]设计了一种移动机器人的差分运动学

收稿日期：2022-11-14

基金项目：国家自然科学基金资助项目（62263030）；新疆维吾尔自治区自然科学基金青年科学基金资助项目（2022D01C86）

Foundation item：Supported by the National Natural Science Foundation of China（62263030）

作者简介：李新凯（1991-），男，博士，讲师，主要从事智能控制、复杂非线性控制研究。E-mail：lxk@xju. edu. cn

模型，提出了一种基于扩展卡尔曼滤波的多传感器信息融合，并利用模糊神经网络控制的避障算法，通过仿真验证了算法的鲁棒性；Yang等[13]根据车辆动力学模型设计了一种基于神经网络补偿的前馈反馈控制器，实现了无人驾驶的姿态和障碍物控制；姚强强等[14]提出了一种基于最优前轮侧向力和附加横摆力矩协同的力驱动模型预测控制路径跟踪控制策略，实现了较好的控制效果。以上控制算法都是基于规则化的，这种根据人工经验编程的控制算法很难应对突发情况，故需要更加智能的算法[15]。

传统的控制算法都是基于一个先要条件，即需要一个精确的数学模型来代替研究对象，通过算法调整模型参数，即算法的作用是为模型提供对应准确的参数。而在实际状态下模型具有可变性，即使当前状态下的参数为模型最优参数，但到下一状态模型的变化导致了所使用的参数不一定为最优参数。基于深度强化学习的控制算法则不需要模型，基于当前的环境信息，根据学习到的相关经验做出动作并进入到下一状态，整个过程中模型的变化并不会影响到控制策略。因此，基于深度强化学习的控制算法更加适用于现实环境中。

随着深度强化学习的发展，越来越多的学者将该方法运用到无人驾驶控制中[16-18]。卢笑等[16]提出了一种联合图像与单目深度特征的强化学习端到端自动驾驶决策，Wang等[19]设计了一种基于异步监督学习方法的强化学习端到端自动驾驶模型，解决了强化学习在初始训练中性能差的问题，并经过仿真验证了算法的鲁棒性。但以上智能算法都是离散型的学习算法，而无人驾驶控制是连续的控制过程，因此，Google DeepMind团队提出了一种确定性动作的DDPG算法，有效地解决了连续性控制问题。

深度强化学习是在强化学习的基础上，采用深度学习的神经网络代替连续状态空间的值函数，弥补了强化学习对状态空间和离散动作行为存在的缺陷[20]。深度强化学习应用较广泛的DDPG算法可以实现无人驾驶汽车的跟踪控制[21-22]，该算法是将确定性策略梯度(DPG)算法和行动者-评论家(Actor-Critic，AC)框架相结合[23]，并在算法中加入了DQN算法的经验回放机制，使其在连续状态空间和动作的控制问题中取得了很好的效果。

针对无人驾驶汽车在智能决策方法研究的现状和目前存在的问题，为提高无人驾驶汽车实时控制跟踪轨迹的精度，同时降低整个控制过程的难度，文中提出了一种基于蒙特卡洛-深度确定性策略梯度(MC-DDPG)的无人驾驶汽车避障跟踪控制算法。该算法首先对用于训练的经验设置临界值，保留优秀经验用于网络训练；然后采用Q函数的蒙特卡洛策略评价训练网络参数，以达到更好的控制效果。文中最后在计算机模拟环境TORCS中对该控制算法进行仿真实验，并对比分析DQN算法、DDPG算法和MC-DDPG算法的跟踪效果。

1　环境及无人车设置

1.1　传感器

无人驾驶汽车自主驾驶过程中，首先要实现对环境的感知以及确定初始状态。无人驾驶汽车通过各类传感器的相互融合实现环境及初始状态信息的采集。传感器的相关类型有全球定位系统(GPS)、摄像头、超声波雷达和激光雷达。GPS用以判断无人驾驶汽车的位置信息；摄像头用以收集车道线信息，判断车辆与车道线的相对位置；超声波雷达则用以检测道路边缘信息；激光雷达用以检测车辆周围的障碍物大小以及运动状态。无人驾驶汽车的初始状态是由传感器采集的信息经过处理得到的，无人驾驶汽车的基本状态信息如表1所示。

表1　无人驾驶汽车的状态信息

Table 1　Status information for unmanned vehicles

名称	取值范围	定义
SpeedX	$(-\infty,+\infty)$	车辆纵向(车行驶方向)车速
Angle	$-\pi\sim\pi$ rad	汽车行驶方向和道路轴方向的夹角
Track	0～200 m	200 m范围内车辆与道路边缘的距离
TrackPos	$(-\infty,+\infty)$	车辆与道路中心线之间的距离，利用道路宽度将其归一化处理，0表示在道路中心线上，1和-1都表示车辆越过道路边缘线
Opponents	0~100 m	100 m范围内障碍物的距离

表1所示的多类传感器信息多样且复杂，因此需要将这些数据进行融合，以字典的方式作为状态输入到无人驾驶系统中。其基本融合过程如下：①传感器观测采集数据；②对采集数据进行特征提取，得到观测数据的特征值；③对特征值数据进行关联，形成对相同目标的描述；④对不同目标的特征值进行组合，以字典的方式作为状态信息输入到强化学习模型。

图1所示为无人驾驶车辆的传感器测量输入值。b为车辆中心与道路中线的距离，由传感器的状态信息TrackPos表征，将其进行归一化处理，大于1或小于-1，表示越过道路边缘；d为本车与其他车辆的相对距离，由传感器的状态信息Opponents

表征；v_x为车辆纵向速度，由传感器的状态信息SpeedX表征，v_y为车辆横向速度，v_z为车辆垂向速度；α为车辆行驶方向与道路中心线的夹角，由传感器的状态信息Angle表征。

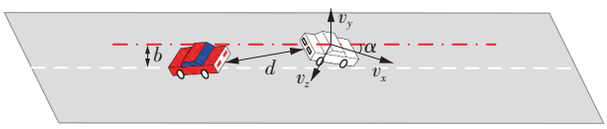

图1 无人驾驶汽车测量输入值

Fig. 1 Measurement inputs for driverless vehicle

1.2 模型框架

为验证算法的鲁棒性，文中设计了一种基于深度强化学习方法的系统模型。利用深度学习处理收集到的信息，包括环境、障碍物、无人驾驶汽车的运动状态等信息，将其以特征形式作为状态输入。强化学习根据当前状态信息做出相应的行为决策并得到一定的回报值，通过回报值判断行为决策的优越性，并根据结果误差对动作决策行为做出相应的奖惩，以提高决策能力。无人驾驶汽车的基本深度强化学习模型如图2所示。

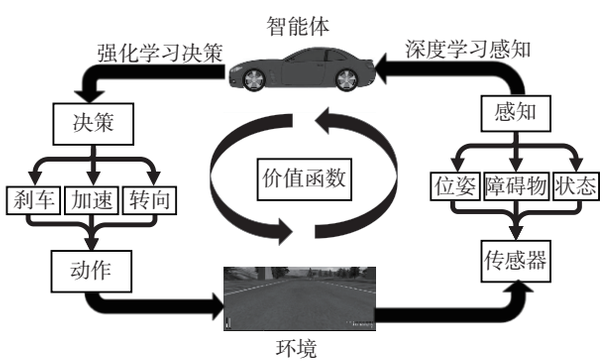

图2 无人驾驶汽车的基本深度强化学习模型

Fig. 2 Basic deep reinforcement learning model for driverless vehicle

文中采用DDPG算法解决无人驾驶避障跟踪的连续控制问题，基于DDPG算法的无人驾驶避障跟踪控制算法流程图如图3所示。

1.3 控制量

在车辆行驶过程中，车辆和环境提供状态信息。无人驾驶汽车根据状态信息提供控制决策，文中根据无人驾驶跟踪控制选择了3个基本控制量，分别为加速、转向和刹车，通过这3个基本控制量实现相对的无人驾驶汽车的避障跟踪控制，其基本定义如表2所示。

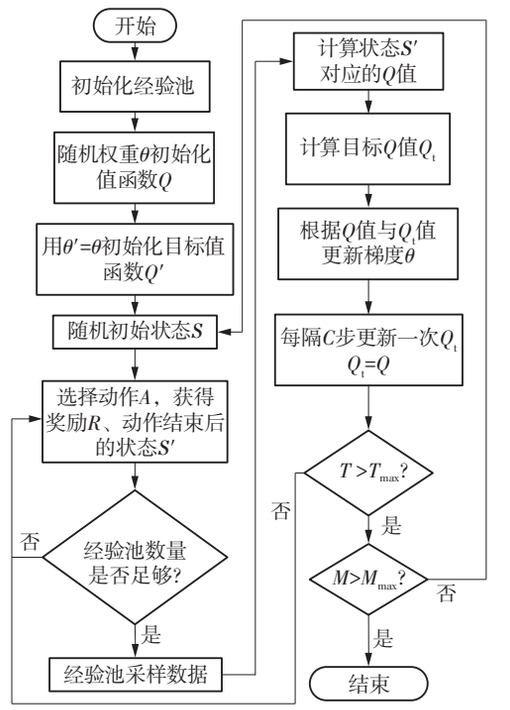

图3 基于DDPG算法的无人驾驶避障跟踪控制算法流程图

Fig. 3 Flow chart of driverless obstacle avoidance tracking control algorithm based on DDPG algorithm

表2 输出控制量定义

Table 2 Definition of output control volumes

名称	取值范围	定义
加速	[0,1]	0表示不加速,1表示全加速
转向	[-1,1]	0表示不转向,-1和1分别表示向左、向右最大转向
刹车	[0,1]	0表示不刹车,1表示最大刹车

2 无人驾驶避障策略设计

2.1 DDPG算法

DDPG算法是根据DPG算法提出来的，属于无模型的行动者-评论家方法中的离线策略学习算法，结合DQN算法的基础，突破了DQN算法只能解决离散型和低维度动作空间的难题。以无人车作为智能体与DDPG算法进行交互，根据传感器收集到的状态信息作为输入，行动者网络根据状态输入做出相应动作决策，并将当前状态信息储存用以采样更新网络参数。评论家网络通过计算行动者网络动作决策的Q值判断动作决策的优劣。为使训练稳定使用目标-行动者网络和目标Q网络定时从经验池中采样数据，目标Q网络通过梯度算法缓慢更新网络参数。DDPG算法的基本框架如图4所示。

为了使网络训练更加稳定，文中加入奥恩斯坦-乌伦贝克(Ornstein-Uhlenbeck，OU)随机过程作

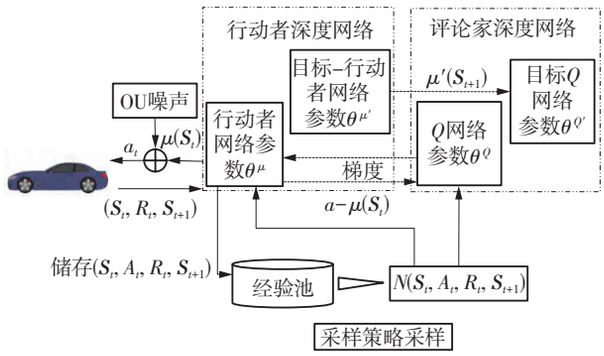

图4　DDPG算法的基本框架

Fig. 4　Basic framework of DDPG algorithm

为噪声输入，OU随机过程定义如下：

$$\mathrm{d}x_t = \theta\left(\mu - x_t\right)\mathrm{d}t + \sigma\mathrm{d}W_t \tag{1}$$

式中，$\theta > 0$，μ为均值，方差$\sigma > 0$，W_t为维纳过程(布朗运动)。

OU随机过程是时序相关的，因此在无人驾驶汽车前一步和后一步的动作过程中利用OU随机过程产生时序相关的探索。训练时目标–行动者网络和目标–评论家网络的参数按照式(2)进行更新，即

$$\begin{cases}\theta^{Q'} = \tau\theta^{Q} + (1-\tau)\theta^{Q'} \\ \theta^{\pi'} = \tau\theta^{\pi} + (1-\tau)\theta^{\pi'}\end{cases} \tag{2}$$

式中，τ为奖励考虑长远的回报值(文中取为0.01)，θ^{Q}为动作价值Q函数的参数。

由图4和DDPG算法的基本原理可以得到DDPG算法的以下特征：①行动者和评论家网络分别由训练网络和目标网络构成，总共有4个网络；②DDPG引入了DQN的经验回放机制，用于储存智能体和环境交互数据信息(S_t, A_t, R_t, S_{t+1})；③DDPG采用软更新方法缓慢更新目标网络参数，使网络学习更加稳定；④在神经网络中对每一组数据进行归一化处理；⑤DDPG算法采用向动作网络的输出中添加随机噪声的方式实现信号输入。

行动者网络用以实现当前动作的确定性策略，使无人驾驶汽车执行当前动作。行动者网络是由传感器收集到的状态信息作为输入，有两个节点数分别为300和600的隐含层，用以逼近策略模型$\pi(a|\boldsymbol{s})$并输出相应动作。其网络结构如图5所示。

评论家网络通过一组参数θ^{Q}估计当前状态动作的Q值，准确的Q值对网络的收敛至关重要。评论家网络结构如图6所示。

评论家网络由传感器传入的状态信息$\boldsymbol{s}$和动作值a作为输入，以两个节点数为600的隐含层逼近价值Q函数，输出为当前网络价值Q。

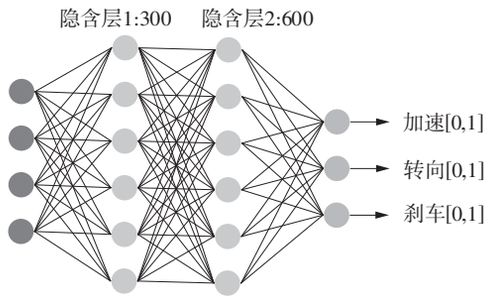

图5　行动者网络结构

Fig. 5　Actor network structure

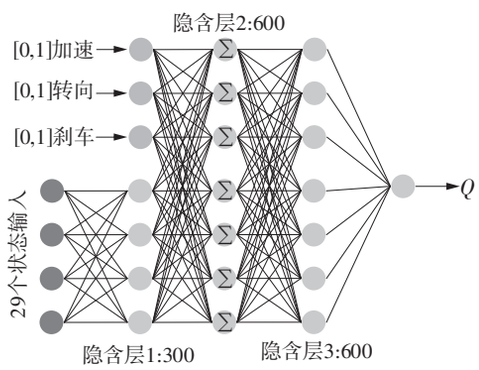

图6　评论家网络结构

Fig. 6　Critic network structure

2.2　奖励设计

文中基于DDPG算法的无人驾驶汽车避障跟踪控制的奖励设计是根据马尔可夫决策的奖励过程，是由一个四元组$\langle S, \boldsymbol{P}_{ss'}, R_s, \gamma\rangle$构成的奖励。其中$S$为状态空间集，$\boldsymbol{P}_{ss'}$为从状态$\boldsymbol{s}$到状态$\boldsymbol{s}'$的状态转移矩阵，$\boldsymbol{P}_{ss'} = \boldsymbol{P}\left[\boldsymbol{S}_{t+1} = \boldsymbol{s}' \middle| \boldsymbol{S}_t = \boldsymbol{s}\right]$，$R_s$为状态$\boldsymbol{s}$的奖励，$R_s = E\left[R_{t+1} \middle| \boldsymbol{S}_t = \boldsymbol{s}\right]$，$\gamma$为折扣因子。

累积奖励是从当前时间到最终状态的所有奖励，即

$$R = R_{t+1} + R_{t+2} + \cdots + R_n \tag{3}$$

由于环境是随机的或者环境变化是不确定的，因此下一个状态的发生也是随机的，当前执行的动作在下一状态下不一定同样执行，且无法确定得到相等的奖励。未来的状态具有不确定性，在累积奖励的过程中需要考虑未来奖励的变化对奖励值的影响，需要对未来奖励状态进行调整，使用折扣未来奖励G_t代替累积奖励，即

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^n R_n \tag{4}$$

奖励回报的期望用状态$\boldsymbol{s}$的价值来表示，其价值函数由两部分组成，包括即时奖励R_{t+1}和下一状态的折扣状态价值$\gamma V\left(\boldsymbol{S}_{t+1}\right)$。因此，状态奖励过程通过贝尔曼方程推导，得

$$V(\boldsymbol{s})=E\left[G_t|\boldsymbol{S}_t=\boldsymbol{s}\right]=$$
$$E\left[R_{t+1}+\gamma R_{t+2}+\gamma^2R_{t+3}+\cdots|\boldsymbol{S}_t=\boldsymbol{s}\right]=$$
$$E\left[R_{t+1}+\gamma G_{t+1}|\boldsymbol{S}_t=\boldsymbol{s}\right]=$$
$$E\left[R_{t+1}+\gamma V(\boldsymbol{S}_{t+1})|\boldsymbol{S}_t=\boldsymbol{s}\right] \tag{5}$$

式中，$V(\boldsymbol{S}_{t+1})$为下一状态预估价值。

奖励函数的贝尔曼方程推导是对所有可能性进行加权平均，其函数表达式为

$$V(\boldsymbol{s})=R_s+\gamma\sum_{s'\in S}\boldsymbol{P}_{ss'}V(\boldsymbol{s}') \tag{6}$$

因此，奖励函数R的定义为

$$R=\begin{cases}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{v}, & 0.1<|d|<D/2\\ -200, & |d|\leqslant 0.1\\ -100, & |d|\geqslant D/2\end{cases} \tag{7}$$

式中：D为道路宽度；$\boldsymbol{C}$为权重系数向量，$\boldsymbol{C}=(c_1,c_2,c_3,c_4,c_5)^{\mathrm{T}}$；$\boldsymbol{v}$为状态参数向量，$\boldsymbol{v}=\left(v_x\cos\alpha,\ -|v_y|,\ -|v_z|,\ -|b|,\ -v_xd\right)^{\mathrm{T}}$。当无人驾驶车辆与其他车辆发生碰撞(即用传感器测量与其他车辆的间距$d<0.1$ m)时，给予最大惩罚，结束该回合训练；当无人驾驶车辆越过道路线时，给予相应的惩罚，结束该回合训练。

2.3 蒙特卡洛优化的DDPG算法策略

强化学习本身就是一个决策过程，在奖励过程中引入一个动作，可以形成一个无人驾驶的决策过程。因此，无人驾驶的决策过程是由一个五元组$\langle S,A,\boldsymbol{P}^a_{ss'},R^a_s,\gamma\rangle$构成，其中$S$为状态集，$A$为动作集，$P^a_{ss'}$为在当前动作下从状态$\boldsymbol{s}$到状态$\boldsymbol{s}'$的状态转移矩阵，$\boldsymbol{P}^a_{ss'}=\boldsymbol{P}\left[\boldsymbol{S}_{t+1}=\boldsymbol{s}'|\boldsymbol{S}_t=\boldsymbol{s},A_t=a\right]$，$R^a_s$为当前动作和状态下的奖励，$R^a_s=E\left[R_{t+1}|\boldsymbol{S}_t=\boldsymbol{s},A_t=a\right]$，$\gamma$为折扣因子。

策略π是状态到动作的映射，是在某个状态采取相应的动作，其表达式为

$$\pi(a|\boldsymbol{s})=\boldsymbol{P}\left[A_t=a|\boldsymbol{S}_t=\boldsymbol{s}\right] \tag{8}$$

式中，$\boldsymbol{P}$为状态转移矩阵。

基于强化学习的无人驾驶决策包括$M=\langle S,A,\boldsymbol{P},R,\gamma\rangle$和策略$\pi$两部分，因此状态和奖励序列可以表示为$\langle S,\boldsymbol{P}^\pi,R^\pi,\gamma\rangle$，其状态转移矩阵和奖励函数表示为

$$\boldsymbol{P}^\pi_{ss'}=\sum_{a\in A}\pi(a|\boldsymbol{s})\boldsymbol{P}^a_{ss'} \tag{9}$$

$$R^\pi_s=\sum_{a\in A}\pi(a|\boldsymbol{s})R^a_s \tag{10}$$

无人驾驶汽车根据状态输入采取相应策略后，其累计回报服从一个分布，因此得到状态值函数

$$V_\pi(\boldsymbol{s})=E_\pi\left[\sum_{k=0}^{\infty}\gamma^kR_{t+k+1}|\boldsymbol{S}_t=\boldsymbol{s}\right] \tag{11}$$

及状态-行为值函数

$$q_\pi(\boldsymbol{s},a)=E_\pi\left[\sum_{k=0}^{\infty}\gamma^kR_{t+k+1}|\boldsymbol{S}_t=\boldsymbol{s},\ A_t=a\right] \tag{12}$$

将式(11)和式(12)转换成贝尔曼方程，得

$$V_\pi(\boldsymbol{s})=E_\pi\left[G_t|\boldsymbol{S}_t=\boldsymbol{s}\right]=$$
$$E_\pi\left[R_{t+1}+\gamma R_{t+2}+\cdots|\boldsymbol{S}_t=\boldsymbol{s}\right]=$$
$$E_\pi\left[R_{t+1}+\gamma V_\pi(\boldsymbol{S}_{t+1})|\boldsymbol{S}_t=\boldsymbol{s}\right]=$$
$$\sum_{a\in A}\pi(a|\boldsymbol{s})q_\pi(\boldsymbol{s},a) \tag{13}$$

$$q_\pi(\boldsymbol{s},a)=E_\pi\left[R_{t+1}+\gamma q(\boldsymbol{S}_{t+1},A_{t+1})|\boldsymbol{S}_t=\boldsymbol{s},A_t=a\right] \tag{14}$$

所有策略中的最大值函数为最优状态值函数$V^*(\boldsymbol{s})$，即$V^*(\boldsymbol{s})=\max_\pi V_\pi(\boldsymbol{s})$，最大的状态-行为值函数为最优状态-行为值函数$q^*(\boldsymbol{s},a)$，即$q^*(\boldsymbol{s},a)=\max_\pi q_\pi(\boldsymbol{s},a)$。因此，可得到最优状态值函数和最优状态-行为值函数的贝尔曼最优方程：

$$V^*(\boldsymbol{s})=\max_a R^a_s+\gamma\sum_{s'\in S}\boldsymbol{P}^a_{ss'}V^*(\boldsymbol{s}') \tag{15}$$

$$q^*(\boldsymbol{s},a)=R^a_s+\gamma\sum\boldsymbol{P}^a_{ss'}\max_{a'}q^*(\boldsymbol{s}',a) \tag{16}$$

DDPG算法是一种无模型连续控制的深度强化学习方法[23]，采用AC网络结构，由行动者网络$\mu(\boldsymbol{s}|\theta^\mu)$、目标-行动者网络$\mu(\boldsymbol{s}|\theta^{\mu'})$及评论家网络$Q(\boldsymbol{s},a|\theta^Q)$和目标-评论家网络$Q(\boldsymbol{s},a|\theta^{Q'})$组成，还增加了随机噪声以加强无人驾驶汽车对环境的感知能力，以及经验回放池使系统可以离线训练学习。

累计回报与动作有关，可以通过改进梯度策略的方法对确定性动作策略的参数ω进行更新，传统的参数ω的更新由累计回报作为更新权重，文中设计当前的策略参数更新只与采取动作之后的结果有关，与曾经和将来所有奖励的总和无关，以使训练过程学习效果更加准确。其基本过程如下：

(1)在当前环境下进行N个回合实验，记录状态数据；

(2)计算总的奖励回报值；

(3)丢弃总奖励回报值后70%回合的经验；

(4)剩余30%的优秀经验以观测值为输入，作为梯度策略数据训练网络的输出；

(5)返回步骤(2)，直到精确稳定地跟踪路径。

在避障学习过程中以跟踪精度作为目标，通过

不断地训练学习，在选择从左侧超车或者从右侧超车时选择合适的超车路径，并在超车过程中避免了因避障而出现的较大误差。MC-DDPG避障跟踪算法的网络结构示意图如图7所示。

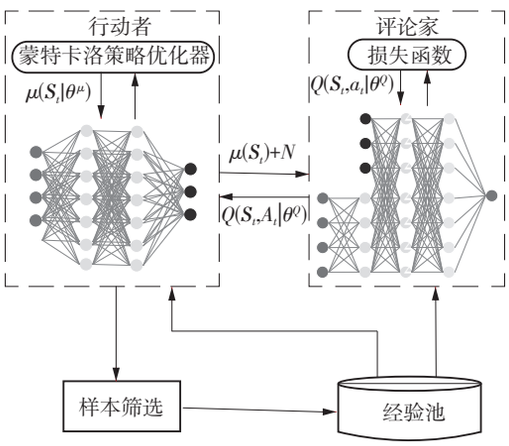

图7　MC-DDPG算法的网络结构示意图

Fig. 7　Schematic diagram of the network structure of MC-DDPG algorithm

用最小化的损失函数对参数θ^Q进行更新，使得Q值更加准确。其最小化损失函数为

$$L=\frac{1}{M}\sum_t\left[r_t+\lambda Q'\left(S_{t+1},\mu'\left(S_{t+1}|\theta^{\mu'}\right)|\theta^{Q'}\right)-Q\left(S_t,A_t|\theta^Q\right)\right]^2 \tag{17}$$

式中，M为从小批量经验中选取的经验值。

当前行动者网络中，由于是确定性策略定义的损失梯度，利用蒙特卡洛近似将式(17)梯度策略中的未知项进行近似处理。文中使用观测到的r_t近似$Q_\pi(S_t,A_t)$，其过程是观测到一个轨迹后，计算出回报，就可以得到回报R_t。更新价值网络就是用该价值奖励去拟合回报R_t，因为之前将状态价值函数近似为R_t，就可以得到预测的误差，使用误差的平方作为损失函数，使得误差变小，然后对其求导，就可以得到梯度。本质上，$Q(S_t,A_t,\omega)$表示在相应策略下给定S_t、A_t后回报G_t的期望，即

$$Q(S_t,A_t,\omega)=E\left[G_t|S_t,A_t,\omega\right] \tag{18}$$

最后根据梯度算法更新网络参数。因此，可以得到优化后的策略梯度公式为

$$\nabla_\omega J(\omega)=E_\omega\left[\sum_{t=0}^{T}\nabla_\omega\lg\pi_\omega\left(A_t|S_t\right)\nabla_\omega Q\left(S_t,A_t,\omega\right)\right]=E_\omega\left[\sum_{t=0}^{T}\nabla_\omega\lg\pi_\omega\left(A_t|S_t\right)E\left[G_t|S_t,A_t,\omega\right]\right]=E_\omega\left[\sum_{t=0}^{T}\nabla_\omega\lg\pi_\omega\left(A_t|S_t\right)G_t\right] \tag{19}$$

使用蒙特卡洛策略梯度Reinforce算法，将价值函数$V(s)$近似代替策略梯度公式中的$Q_\pi(s,a)$。用蒙特卡洛法计算序列每个时间位置t的状态价值V_t，并使用梯度上升法更新策略函数的参数θ：

$$\theta=\theta+\alpha\nabla_\theta\lg\pi_\theta\left(S_t,A_t\right)V_t \tag{20}$$

同样地，状态价值函数的更新公式为

$$V_\pi(s)=V_\pi(s)+\alpha\left[G_t-V_\pi(s)\right] \tag{21}$$

式中，α为学习率。

3　仿真实验

3.1　训练环境

TORCS平台是根据模拟车辆物理模型建立的模拟器，因此可以实现车辆与环境的交互。文中通过TORCS平台验证文中所提算法的可行性，基于TORCS的算法训练路线环境如图8所示，其中B_1为起始点(终点)、B_2为连续转弯起点、B_3为急转弯起点，B_4为直线行驶起点。

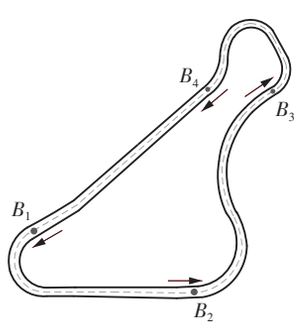

图8　无人驾驶避障训练路线图

Fig. 8　Driverless obstacle avoidance training roadmap

动态障碍物作为影响无人驾驶汽车跟踪效果的因素，可出现在跟踪路径中的任意位置。为验证算法的可行性，设置5种工况(3、5、8、10、15个动态障碍物)，以研究无人驾驶汽车的稳定状态。

本实验设置优化DDPG算法中的超参数如下：样本储存空间大小BUFFER_SIZE=100 000，小批量选取样本数量BATCH_SIZE=32，折扣系数γ=0.99，目标网络超参数τ=0.001，行动者网络学习率和评论家网络学习率分别为0.000 1和0.001。

3.2　仿真结果分析

网络训练累计价值回报表示无人驾驶汽车从一个状态选择一个动作后走到最终状态，最后获得的累计奖励总和的平均值。不同障碍物数量下，3种算法的累计价值回报比较如图9所示。从图中可知：DQN算法的回报价值较低，算法的稳定

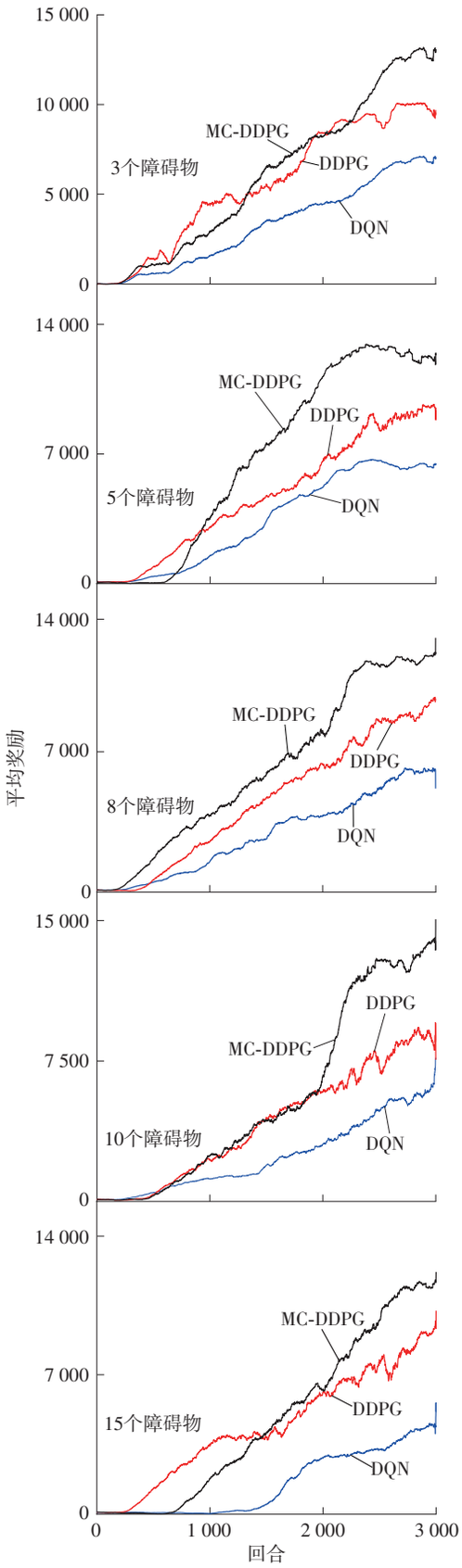

图9 训练场景下3种算法的累计奖励曲线

Fig. 9 Cumulative reward curves of three algorithms in the training scenario

性差；DDPG算法比较稳定，但随着障碍物的增加，算法的训练效果达到最佳时需要的回合数也在增加；随着障碍物的增加，MC-DDPG算法收敛到最佳效果后的跟踪效果基本稳定，且优于DQN算法和DDPG算法。由图9可以计算出蒙特卡洛策略梯度改进前后3种算法的目标Q值回报性能，如表3所示。

表3 3种算法的累计回报性能对比

Table 3 Comparison of cumulative return performance among three algorithms

障碍物数量	Q值		
	MC-DDPG	DDPG	DQN
3	12787	9895	6778
5	12506	9307	6289
8	12035	9535	6141
10	12751	8944	5572
15	11385	8107	4226

为验证算法的跟踪避障控制效果，通过误差直观判断跟踪效果的实际情况，文中设计了两种无人驾驶跟踪误差，分别是车辆与实际道路角度的误差、车辆位置与实际道路中心线的误差，通过这两种误差可基本判断无人驾驶车辆的实际跟踪效果。不同障碍物数量下，3种算法的角度误差和位置误差曲线分别如图10、图11所示。

障碍物的存在使无人驾驶汽车在避开障碍物时与道路线的夹角(即角度误差)有着较大的变化。由图10可以知道：在跟踪过程中，使用DQN算法的无人驾驶汽车的角度误差变化最大，其跟踪稳定性差；使用DDPG算法的无人驾驶汽车的角度误差变化较大，跟踪轨迹的准确性大大降低；使用MC-DDPG算法的无人驾驶汽车在遇到障碍物时，利用小角度的转向避开障碍物，大大提高了无人驾驶汽车轨迹跟踪的控制精度。由图10可以计算出3种算法的角度误差，如表4所示。

障碍物的存在使无人驾驶汽车在避障过程中远离中心线。由图11可以看出，MC-DDPG算法控制的角度变化比DQN算法和DDPG算法小，使得无人驾驶汽车在安全避障的同时行驶稳定性高，位置误差变化小，控制效果更佳。由图11可以计算出3种算法的位置误差，如表5所示。

3.3 结果验证与分析

为验证训练数据的可靠性，对训练好的模型进行验证，验证路线如图12所示。3种算法的跟踪速度如图13所示。验证训练效果是否使车辆稳定的驾驶，可以通过速度分析无人驾驶汽车的跟踪稳

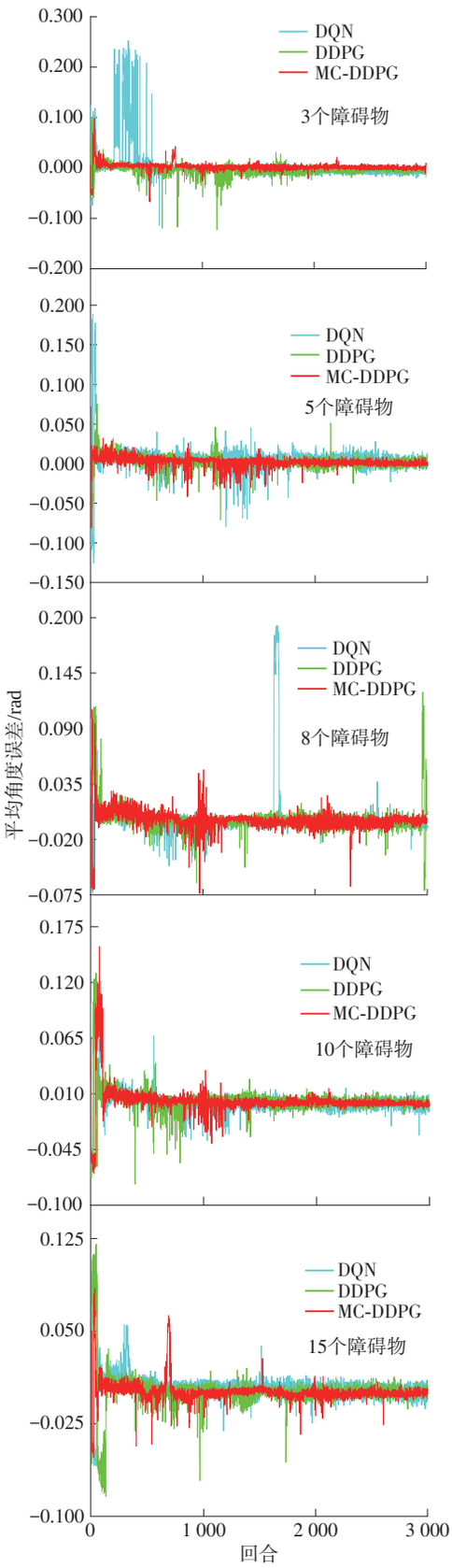

图 10　不同障碍物数量下 3 种算法的角度误差曲线

Fig. 10　Angular error curves of three algorithms for different numbers of obstacles

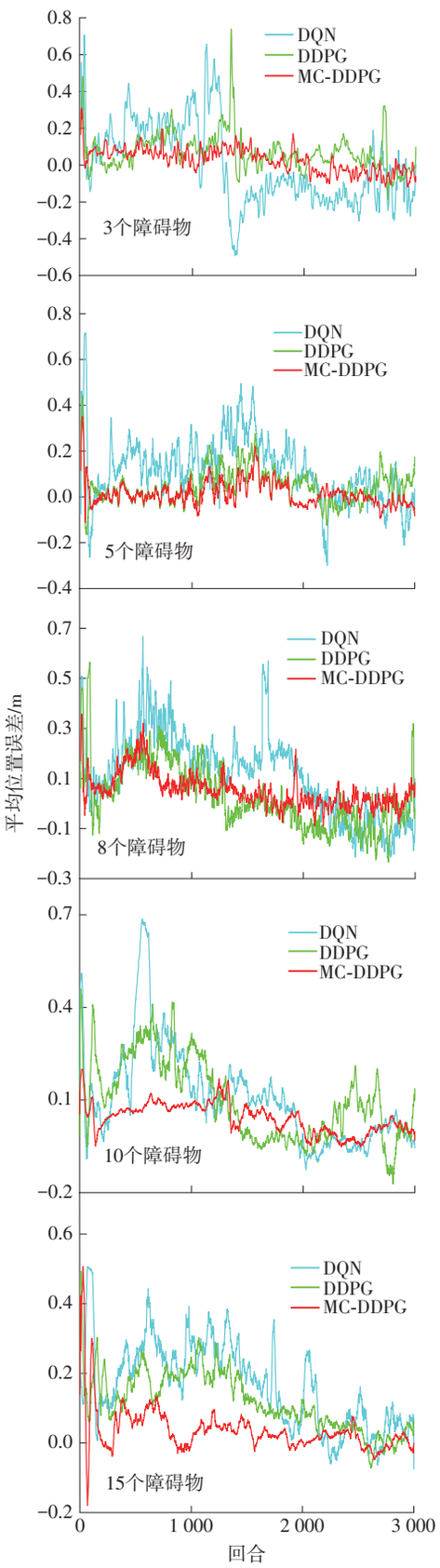

图 11　不同障碍物数量下 3 种算法的位置误差曲线

Fig. 11　Position error curves of the three algorithms for different numbers of obstacles

表4 3种算法的角度误差对比

Table 4 Comparison of angular error among three algorithms

障碍物数量	角度误差/rad		
	MC-DDPG	DDPG	DQN
3	0.00195	0.00217	0.00262
5	0.00141	0.00237	0.00316
8	0.00143	0.00216	0.00221
10	0.00154	0.00216	0.00361
15	0.00164	0.00223	0.00415

表5 3种算法的位置误差对比

Table 5 Comparison of position error among three algorithms

障碍物数量	位置误差/m		
	MC-DDPG	DDPG	DQN
3	0.1457	0.2011	0.3164
5	0.1435	0.1815	0.2457
8	0.1650	0.2418	0.3085
10	0.1872	0.1754	0.3518
15	0.1437	0.2130	0.2916

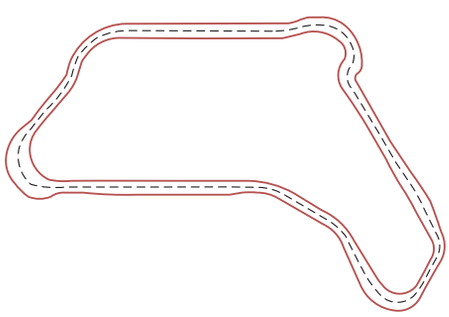

图12 无人驾驶避障验证路线图

Fig. 12 Driverless obstacle avoidance validation roadmap

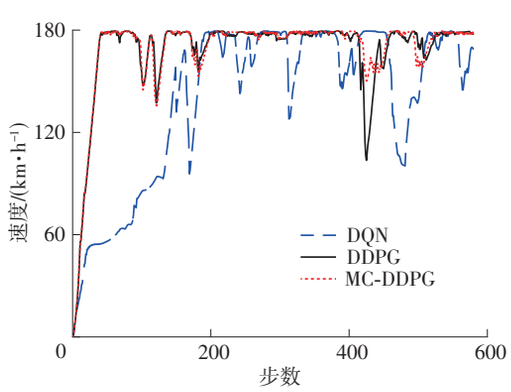

图13 3种算法的无人驾驶跟踪速度曲线

Fig. 13 Driverless tracking speed curves of three algorithms

定性，速度的稳定性意味着无人驾驶汽车正常驾驶。无人驾驶汽车在每一步跟踪过程都存在相应的奖励，该奖励值可以反映无人驾驶汽车的避障跟踪效果，3种算法的奖励回报曲线如图14所示。

由图13、图14可以看出：在DQN算法控制下，无人驾驶汽车可以学习到基本的控制效果，但其跟踪稳定性差；在DDPG算法和MC-DDPG算法控制下，无人驾驶汽车都可以稳定的驾驶，MC-DDPG算法的控制效果最佳。

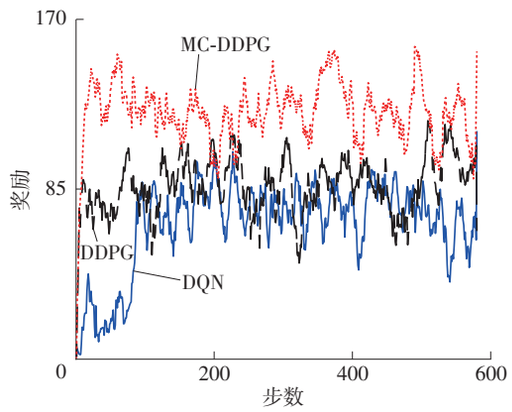

图14 3种算法的奖励回报曲线

Fig. 14 Reward return curves of three algorithms

为验证文中算法的跟踪效果和跟踪精度，文中对3种算法跟踪过程的角度误差和位置误差进行了对比，结果如图15所示。

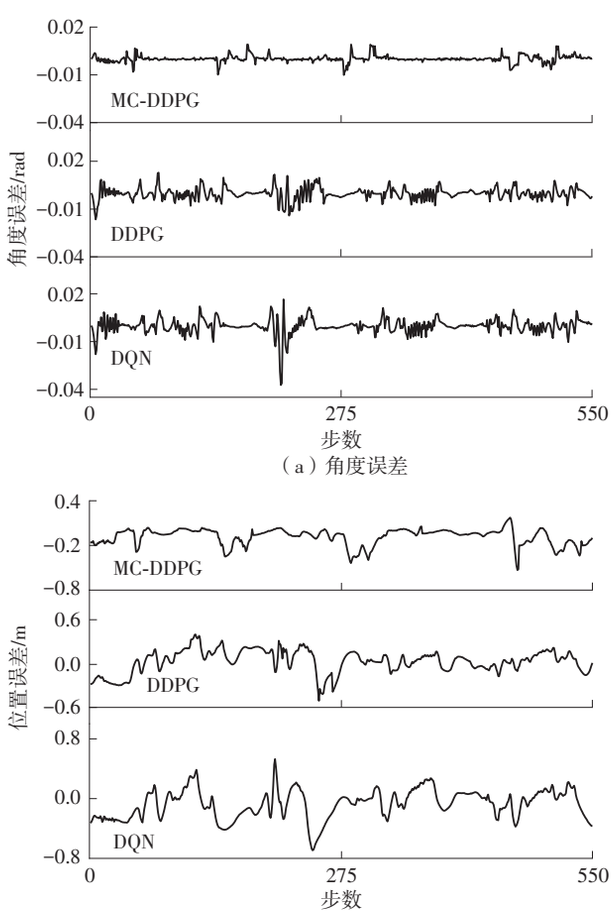

图15 在不同避障步数下3种算法的角度误差和位置误差曲线

Fig. 15 Angular and positional error curves of three algorithms at different obstacle avoidance steps

由图 15 可知：无人驾驶汽车与道路夹角的变化小，其中角度变化大的主要原因是障碍物的存在，在避障过程中不可避免地使无人驾驶汽车与道路的夹角发生变化；MC-DDPG 算法控制的无人驾驶汽车角度变化和位置误差变化均较小，其跟踪效果最佳。从整体误差看，DQN 算法和 DDPG 算法能达到相应的跟踪控制，但其控制稳定性差，误差变化大，而 MC-DDPG 算法的控制效果更加平稳，稳定性更强。

3 种算法的避障轨迹跟踪效果如图 16 所示。图中障碍物在轨迹上的位置不断发生变化，颜色越深表示障碍物与实际位置越接近，而浅色表示障碍物曾经的位置。在整个轨迹跟踪过程中，MC-DDPG 算法的跟踪精度高，可以很好地跟踪轨迹；DDPG 算法及 DQN 算法控制的无人驾驶汽车可以完成相关的跟踪任务，但跟踪误差较大，在轨迹跟踪中心线上来回浮动；MC-DDPG 算法控制的无人驾驶汽车避障效果最佳，避障路径最优。

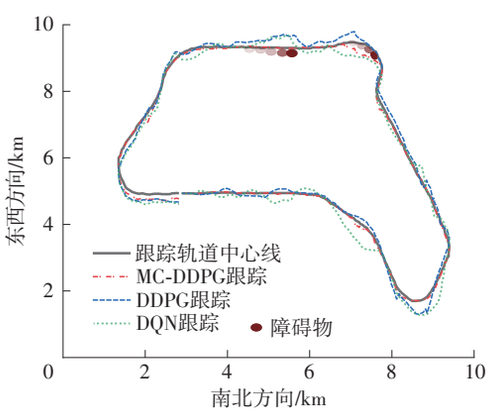

图 16　避障轨迹跟踪路线图

Fig. 16　Obstacle avoidance track tracking roadmap

避障跟踪过程中 3 种算法的性能比较如表 6 所示。由表中可以看出，在整个避障过程中，MC-DDPG 算法控制的无人驾驶汽车与由 DDPG 算法及 DQN 算法控制的无人驾驶汽车相比，平均位置误差最小，偏离跟踪轨迹的范围最小，并且避障过程采用的步数最短。因此，整个过程中 MC-DDPG 算法的控制效果更优，控制精度更高。

表 6　避障跟踪过程 3 种算法的性能对比

Table 6　Performance comparison of three algorithms for obstacle avoidance tracking process

算法	平均位置误差/m	最远避障距离（归一化值）	平均避障步数
MC-DDPG	0. 090	0. 35	23
DDPG	0. 154	0. 40	31
DQN	0. 184	0. 39	32

4　结论

文中提出了一种基于 MC-DDPG 的无人驾驶汽车避障跟踪控制算法，建立了单步状态策略学习的 AC 学习网络。针对车道跟随工况，设计相应的条件约束和价值回报函数，使无人驾驶汽车可稳定地沿着车道行驶方向收敛并合理地避开障碍物。在 TORCS 仿真环境下的结果表明，文中 MC-DDPG 算法在训练 2 500 回合后收敛，无人驾驶汽车可有效避开动态障碍物并跟踪车道中心线驾驶，且最高时速可达 180 km/h，实现了高速稳定避障跟踪效果。与原始 DDPG 算法对比，改进算法的控制精度及跟踪效果均较优，实现了更加精准的控制跟踪。

参考文献：

[1] JAN B, FARMAN H, KHAN M. Designing a smart transportation system: an internet of things and big data approach [J]. IEEE Wireless Communications, 2019, 26(4): 73-79.

[2] 徐向阳，胡文浩，董红磊. 自动驾驶汽车测试场景构建关键技术综述 [J]. 汽车工程，2021，43(4)：610-619.
XU Xiangyang, HU Wenhao, DONG Honglei. Overview of key technologies for autonomous vehicle test scenario construction [J]. Automotive Engineering, 2021, 43(4): 610-619.

[3] 熊璐，杨兴，卓桂荣，等. 无人驾驶车辆的运动控制发展现状综述 [J]. 机械工程学报，2020，56(10)：127-143.
XIONG Lu, YANG Xing, ZHUO Guirong, et al. Overview on motion control of autonomous vehicles [J]. Journal of Mechanical Engineering, 2020, 56(10): 127-143.

[4] ZHANG X L, ZHANG W X, ZHAO Y Q. Personalized motion planning and tracking control for autonomous vehicles obstacle avoidance [J]. IEEE Transactions on Vehicular Technology, 2022, 71(5): 4733-4747.

[5] 于向军，槐元辉，姚宗伟. 工程车辆无人驾驶关键技术 [J]. 吉林大学学报(工学版)，2021，51(4)：1153-1168.
YU Xiang-jun, KUI Yuan-hui, YAO Zong-wei. Key technologies in autonomous vehicle for engineering [J]. Journal of Jilin University (Engineering and Technology Edition), 2021, 51(4): 1153-1168.

[6] GRUYER D, MAGNIER V, HAMDI K, et al. Perception information processing and modeling: critical stages for autonomous driving applications [J]. Annual Reviews in Control, 2017, 41(10): 323-341.

[7] 张家旭, 杨雄, 施正堂, 等. 汽车紧急换道避障的路径规划与跟踪控制 [J]. 华南理工大学学报(自然科学版), 2020, 48(9): 86-93, 106.
ZHANG Jiaxu, YANG Xiong, SHI Zhengtang, et al. Path planning and tracking control for emergency lane change and obstacle avoidance of vehicles [J]. Journal of South China University of Technology (Natural Science Edition), 2020, 48(9): 86-93, 106.

[8] WANG T, JIANG J F, LIN Y T, et al. Driver model for obstacle avoidance based on CarSim [J]. Transactions of the Chinese Society of Agricultural Engineering, 2010, 26(5): 159-163.

[9] 樊晓平, 李双艳, 陈特放. 基于新人工势场函数的机器人动态避障规划 [J]. 控制理论与应用, 2005, 22(5): 703-707.
FAN Xiao-ping, LI Shuang-yan, CHEN Te-fang. Dynamic obstacle-avoiding path plan for robots based on a new artificial potential field function [J]. Control Theory & Applications, 2005, 22(5): 703-707.

[10] KATSUKI R, TASAKI T, WATANABE T. Graph search based local path planning with adaptive node sampling [C]// Proceedings of 2018 IEEE Intelligent Vehicles Symposium. Changshu: IEEE, 2018: 2084-2089.

[11] WANG Hong-chao, ZHANG Wei-wei, WU Xun-cheng, et al. A double-layer nonlinear model predictive control based control algorithm for local trajectory planning for automated trucks under uncertain road adhesion coefficient conditions [J]. Frontiers of Information Technology & Electronic Engineering, 2020, 21(7): 1059-1074.

[12] ZONG C G, JI Z J, YU Y, et al. Research on obstacle avoidance method for mobile robot based on multi-sensor information fusion [J]. Sensors and Materials, 2020, 32(4): 1159-1170.

[13] YANG Z C, FENG Y T, ZHANG L X, et al. Obstacle avoidance control of underactuated robot based on neural network feedforward compensation [J]. Measurement & Control Technology, 2017, 36(11): 89-97.

[14] 姚强强, 田颖, 王圣渊, 等. 基于力驱动的智能汽车路径跟踪控制策略 [J]. 华南理工大学学报(自然科学版), 2022, 50(2): 33-41, 57.
YAO Qiangqiang, TIAN Ying, WANG Shengyuan, et al. Research on path tracking control strategy of intelligent vehicles based on force drive [J]. Journal of South China University of Technology (Natural Science Edition), 2022, 50(2): 33-41, 57.

[15] SALLAB A E, ABDOU M, PEROT E, et al. Deep reinforcement learning framework for autonomous driving [J]. Electronic Imaging, 2017, 29(19): 70-76.

[16] 卢笑, 竺一薇, 阳牡花, 等. 联合图像与单目深度特征的强化学习端到端自动驾驶决策方法 [J]. 武汉大学学报(信息科学版), 2021, 46(12): 1862-1871.
LU Xiao, ZHU Yiwei, YANG Muhua, et al. Reinforcement learning based end-to-end autonomous driving decision-making method by combining image and monocular depth features [J]. Geomatics and Information Science of Wuhan University, 2021, 46(12): 1862-1871.

[17] 张守武, 王恒, 陈鹏, 等. 神经网络在无人驾驶车辆运动控制中的应用综述 [J]. 工程科学学报, 2022, 44(2): 235-243.
ZHANG Shou-wu, WANG Heng, CHEN Peng, et al. Overview of the application of neural networks in the motion control of unmanned vehicles [J]. Chinese Journal of Engineering, 2022, 44(2): 235-243.

[18] 董豪, 杨静, 李少波, 等. 基于深度强化学习的机器人运动控制研究进展 [J]. 控制与决策, 2022, 37(2): 278-292.
DONG Hao, YANG Jing, LI Shao-bo, et al. Research progress of robot motion control based on deep reinforcement learning [J]. Control and Decision, 2022, 37(2): 278-292.

[19] WANG Y P, ZHENG K X, TIAN D X, et al. Asynchronous supervised learning pre-training methods for reinforcement learning autonomous driving models [J]. Frontiers of Information Technology & Electronic Engineering, 2021, 22(5): 673-687.

[20] 吕帅, 龚晓宇, 张正昊, 等. 结合进化算法的深度强化学习方法研究综述 [J]. 计算机学报, 2022, 45(7): 1478-1499.
LÜ Shuai, GONG Xiao-yu, ZHANG Zheng-hao, et al. Survey of deep reinforcement learning methods with evolutionary algorithms [J]. Chinese Journal of Computers, 2022, 45(7): 1478-1499.

[21] 张新钰, 高洪波, 赵建辉, 等. 基于深度学习的自动驾驶技术综述 [J]. 清华大学学报(自然科学版), 2018, 58(4): 438-444.
ZHANG Xinyu, GAO Hongbo, ZHAO Jianhui, et al. Overview of deep learning intelligent driving methods [J]. Journal of Tsinghua University (Science and Technology), 2018, 58(4): 438-444.

［22］ 陈红名，刘全，闫岩，等. 基于经验指导的深度确定性多行动者-评论家算法［J］. 计算机研究与发展，2019，56(8)：1708-1720.
CHEN Hongming，LIU Quan，YAN Yan，et al. An experience-guided deep deterministic actor-critic algorithm with multi-actor［J］. Journal of Computer Research and Development，2019，56(8)：1708-1720.

［23］ 陈亮，梁宸，张景异，等. Actor-Critic框架下一种基于改进DDPG的多智能体强化学习算法［J］. 控制与决策，2021，36(1)：75-82.
CHEN Liang，LIANG Chen，ZHANG Jing-yi，et al. A multi-intelligence reinforcement learning algorithm based on improved DDPG in the Actor-Critic framework［J］. Control and Decision，2021，36(1)：75-82.

Driverless Obstacle Avoidance and Tracking Control Based on Improved DDPG

LI Xinkai　HU Xiaocheng　MA Ping　ZHANG Hongli
（School of Electrical Engineering，Xinjiang University，Urumqi 830017，Xinjiang，China）

Abstract：In the process of tracking and obstacle avoidance control of driverless vehicles, the controlled object has nonlinear characteristics and variable control parameters. The linear model and the fixed mathematical model of driverless vehicles are difficult to ensure the safety and stability of the vehicle in complex environments, and the driverless discrete control process increases the difficulty of control. To address such problems, in order to improve the accuracy of real-time control tracking trajectory of driverless vehicles, and at the same time reduce the difficulty of the whole control process, the paper proposed a Monte Carlo-depth deterministic policy gradient-based obstacle avoidance tracking control algorithm for driverless vehicles. The algorithm builds a control system model based on a deep reinforcement learning network, and adopts excellent training samples in the strategy learning sampling process. It optimizes the network training gradient with the Monte Carlo method, and makes a distinction between good and bad training samples for the algorithm. The excellent samples are used to find the optimal network parameters through a gradient algorithm, so as to enhance the learning ability of the network algorithm and realize a better and continuous control of the driverless vehicle. Simulation experiments of the control method were carried out in the computer simulation environment TORCS. The results show that the proposed improved DDPG algorithm can be applied to effectively achieve the obstacle avoidance tracking control of the driverless vehicle, and the tracking accuracy and obstacle avoidance effect of the unmanned car under its control is better than that of the deep *Q* network algorithm and the DDPG algorithm.

Key words：self-driving；dynamic obstacle avoidance；depth deterministic policy gradient；trajectory tracking；gradient optimization