数据和规则混合驱动的全自动代客泊车轨迹规划

赵花蕊[1],曹仰杰[2]
(1. 河南省平台经济发展指导中心,河南 郑州 450008;
2. 郑州大学 网络空间安全学院,河南 郑州 450003)

摘　要:该文研究了全自动代客泊车系统中的轨迹规划,并提出了一种创新的基于深度强化学习的方法。当前的路径规划技术主要依赖几何算法,这些方法在复杂停车环境中面临诸多限制,尤其是在处理动态障碍物和环境变化不确定性时。此外,基于优化的策略虽然理论上有效,但其计算复杂度较高,难以实现实时响应,限制了其在实际应用中的可行性。该文提出了一种数据与规则混合驱动泊车轨迹规划方法。该方法通过利用历史数据和经验规则,显著提高了系统的可扩展性和泛化能力。值得注意的是,该方法不依赖于实时交互获取其他车辆的精确物理信息,使其更加适合当前实际应用场景中信息不完全或传感器受限的情况。此外,采用课程学习和混合 A * 算法来加速强化学习模型的收敛速度,通过逐步增加任务复杂度,提升模型对环境变化的适应能力。实验结果显示,该方法在复杂自动泊车任务中的表现优异,能够有效实现高效、安全的泊车操作,充分展现了其在全自动代客泊车系统中的应用潜力。

关键词:全自动代客泊车系统;深度强化学习;混合 A * ;课程学习;轨迹规划
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2025)07-0148-08
doi:10.20165/j.cnki.ISSN1673-629X.2025.0048

Data and Rule Hybrid-driven Motion Planning for Fully Automated Valet Parking

ZHAO Hua-rui[1],CAO Yang-jie[2]
(1. Platform Economy Development Guidance Center of Henan Province,Zhengzhou 450008,China;
2. School of Cyberspace Security,Zhengzhou University,Zhengzhou 450003,China)

Abstract:We investigate trajectory planning in fully automated valet parking systems and propose an innovative approach based on deep reinforcement learning. Current path planning techniques primarily rely on geometric algorithms,which face numerous limitations in complex parking environments,especially when dealing with dynamic obstacles and uncertainties in environmental changes. Although optimization-based strategies are theoretically effective,their high computational complexity hinders real-time responsiveness,limiting their feasibility in practical applications. We introduce a hybrid data and rule driven method for parking trajectory planning that significantly enhances the system's scalability and generalization capabilities by leveraging historical data and heuristic rules. Notably,this approach does not depend on real-time interactions to obtain precise physical information about other vehicles,making it more suitable for current practical scenarios,particularly in situations with incomplete information or limited sensors. Furthermore,we employ curriculum learning and a hybrid A * algorithm to accelerate the convergence speed of the reinforcement learning model by gradually increasing task complexity,enhancing the model's adaptability to environmental changes. Experimental results demonstrate that the proposed method performs exceptionally well in complex automated parking tasks,effectively achieving efficient and safe parking operations,thereby showcasing its potential application in fully automated valet parking systems.

Key words:fully automated valet parking system;deep reinforcement learning;hybrid A * ;curriculum learning;trajectory planning

0　引　言

虽然半自动泊车辅助系统已经在商用车市场上普及,但全自动代客泊车(Automated Valet Parking,AVP)[1-4]仍然具有挑战性。为了让人类司机在停车场外停车,自动停车系统必须既能在杂乱的环境中进行远程导航,又能在狭窄的停车位周围进行短程调整。

收稿日期:2024-10-09　　**修回日期**:2025-02-19
基金项目:国家自然科学基金(61972092)
作者简介:赵花蕊(1984-),女,硕士研究生,高级工程师,通讯作者,研究方向为人工智能、计算机视觉、电子政务;曹仰杰(1976-),男,博士,教授,研究方向为机器智能与高性能计算。

这需要多个智能组件的无缝集成,包括感知、定位、任务规划、轨迹规划和运动控制。后两个模块主要负责驾驶决策和车辆运动转向[5-6]。因此,它们通常被认为是反映 AVP 平台智能化水平的重要指标。本研究主要关注自动代客泊车系统的轨迹规划部分,该系统在给定已知的起点和目标姿态的情况下搜索无碰撞且满足目标姿态的轨迹。

目前,已经提出了许多完善的几何路径规划器[7-9]。这些方法的核心思想是利用样条或多项式技术来塑造和平滑车辆的机动轮廓。尽管基于几何的方法在计算上是友好的,但它们有许多限制。通常,几何路径规划器可以在特定结构的基础上产生解。当需要考虑复杂的交通环境时,这种规划可能无法满足特定场景下的车辆动态约束。此外,这可能导致实际机动轨迹与设计曲线之间存在较大差异。针对多车停车场景,文献[10]提出了一种在 AVP 系统框架下协调多车的控制方法。该系统依赖于基础设施服务器和车对基础设施(V2I)通信接口,难以实现单车部署。文献[11]采用双树结构,提出了一种增强型随机树状轨迹规划器来生成移动机器人系统的运动轨迹。同样,文献[12]提出了一种结合随机树和最优控制理论的轨迹规划新方法。这两种方法都属于样本搜索方法,首先使用有限网格集对搜索空间进行离散化。然后在初始姿态和目标姿态之间选择一个令人满意的连接。在文献[7]中,作者利用一种特定的动态优化算法来规划一类轮式移动机器人的时间最优停车轨迹。此外,利用初始猜测生成器,在文献[13]中建立了 AGV 的双环停车机动规划。这两项工作的主要目的是构建一个由 AGV 动力学、车辆相关约束和其他停车需求组成的最优控制公式。随后,利用完善的优化算法生成机动轨迹,然后将其用于自动驾驶控制器或建议人类驾驶员。与这种方法相关的一个优点是,不同的限制/需求可以很容易地包含在优化过程中。然而,这些基于优化的策略存在两个关键问题:由于巨大的计算需求,它们不太可能实时实现;算法的收敛性会受到干扰、不确定性、局部不可行区域等的显著影响。到目前为止,对 AVP 轨迹规划的研究主要集中在基于已知地图的采样或优化方法上,以获得连续轨迹[14-15]。然而,传统规划方法的计算复杂度取决于环境设置,并且在不同环境下的性能差异很大。基于学习的方法在自动驾驶领域逐渐显示出强大的优势。

最近,一些研究表明,可以利用基于深度神经网络(Deep Neural Network,DNN)的直接召回来规划最优机动轨迹[16-17]。该方法的核心思想是在预生成的轨迹集合上训练深度神经网络,使其能够学习和表示最优控制动作与机动轨迹之间的关系。通过迭代调用训练好的映射关系,实现了机动轨迹的在线规划。文献[18]采用基于模型的预测控制和强化学习方法停放车辆,实现平稳连续运动。文献[19]在基于深度学习的轨迹规划与控制的基础上,提出了一种针对单个车辆泊车机动的一体化实时轨迹规划与跟踪控制。另外,深度强化学习(Deep Reinforcement Learning,DRL)在机器人[20]和自动驾驶[21]等各种规划和决策任务上表现出了良好的性能。与传统的基于规则的轨迹规划方法相比[22],由于神经网络具有较高的表征能力,DRL 在复杂场景下具有更好的可扩展性和泛化潜力。当前已经有一些方法将 DRL 方法用在了自动代客泊车任务中。Chen 等人[23]提出了一种基于多智能体 DRL 的自动代客泊车轨迹规划方法。该方法利用传统的轨迹规划来加速学习过程,并引入碰撞冲突约束进行策略优化,以缓解路径冲突问题。但是该方法中需要车辆之间进行实时交互从而获取周围车辆的位置、速度等物理信息。然后基于精确的环境车辆以及障碍物的信息从而进行轨迹规划。但在当前背景下,依然有很多车辆并不是智能网联车,所以,每辆车仅能获得自己的精确物理信息,而无法实时获取到其他车辆的精确物理信息,那么上述方法将无法在实际环境中应用。另外,自动泊车任务是一个复杂度很高的任务,其成功条件非常苛刻,这就导致了强化学习算法在学习的过程中收敛速度慢,需要大量的试错,而且在某些极端困难的车位下会出现无法收敛的情况。

为了克服这些难题,该文提出了一种数据和规则混合驱动的自动代客泊车轨迹规划方法。该方法仅仅基于红外测距传感器来感知周围环境,然后以端到端的方式完成整个自动代客泊车任务。该文提出的基于深度强化学习的方法通过在仿真环境中采集泊车数据以及奖励信号来优化模型参数。另外为了解决模型收敛困难的问题,利用基于规则的混合 A * 算法生成参考轨迹,从而引导模型更好地学习泊车策略。通过在停车场场景中进行实验,验证了该方法的有效性,并通过消融实验验证了新增模块的作用。实验结果表明,该方法不仅在准确度上取得了很好的结果,而且在安全性上也很出色。

1　预备知识

1.1　强化学习

强化学习(RL)是一种解决决策问题的学习范式,它为行为建模提供了一种形式化方法,其中软件或物理代理通过试错学习如何在环境(即真实或模拟世界)中采取最佳行动,仅由积极或消极的标量奖励信号(有时称为强化)指导。形式上,环境是一个由元组 (S,A,P,R,γ) 表示的马尔可夫决策过程(MDP)。这

里，S 代表状态空间，A 表示动作空间，$P(s' \mid s,a)$ 表示转移模型，也称为环境动力学，可以预测环境状态的演变。奖励函数表示为 $R(s,a,s')$，它量化与状态-动作对相关联的即时奖励。最后，折扣因子 $\gamma \in (0,1]$ 用于权衡即时奖励和未来奖励之间的关系。

MDP 遵循马尔可夫性质，这意味着未来状态仅依赖于前面的状态和动作。然而，某些问题涉及部分可观察马尔可夫决策过程（POMDP），其中无法访问完全可观察的马尔可夫状态。POMDP 引入了观测集合 Ω 和观测函数 O。具体而言，$O(a,s',o)=p(o' \mid a,s')$ 表示当智能体执行动作 a_t 并达到状态 s' 时观测到 o_{t+1} 的概率。在每个时间步 t，MDP 中的智能体基于当前状态 s_t 选择动作 $a_t \in A$，随后在过渡到新状态 s_{t+1} 时获得数值奖励 r_{t+1}。由此产生的序列 $s_0,a_0,r_1,s_1,a_1,r_2,\cdots$ 通常被称为轨迹。在时间步骤 t 之后，未来的预期累积奖励，即预期回报 G_t，可以定义为：

$$G_t \doteq r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{T} \gamma^k r_{t+k+1}$$

其中，T 表示一个有限值，对于无限时间问题，T 可以取 ∞。策略 $\pi(a \mid s)$ 将状态映射到概率，表示选择每个可能动作的可能性。值函数 $v_\pi(s)$ 表示从状态 s 按照策略 π 获得的预期回报。

$$v_t(s) \doteq \mathbb{E}_\pi[G_t \mid s_t = s]$$

类似地，动作值函数 $q_\pi(s,a)$ 定义为：

$$q_\pi(s_t,a_t) \doteq \mathbb{E}_\pi[G_t \mid s_t = s, a_t = a]$$

它满足递归贝尔曼方程：

$$q_\pi(s_t,a_t) \doteq \mathbb{E}_{s_{t+1}}[r_{t+1} + \gamma q_\pi(s_{t+1},\pi(s_{t+1}))]$$

强化学习的主要目标是确定最优策略 π^*，使得期望回报最大化。数学上可以表示为：

$$\pi^* = \operatorname{argmax}_\pi E_\pi[G_t \mid s_t = s]$$

1.2 车辆模型

车辆运动模型如图 1 所示。发动机提供加速度，而刹车则导致加速度为 0。车辆的转向由前轮控制。车辆的位置由前轴和后轴中点 $p=(x,y)$ 表示。对应的速度用 v 表示。前轮和后轮轴之间的纵向距离用参数 L 表示。变量 θ 和 β 分别表示车辆的角度和前轮转向角。

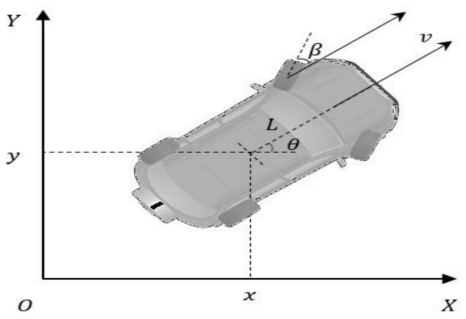

图 1 车辆运动模型

考虑到车辆可以调整前轮的转向角而不显著改变其位置，状态空间可以简化为 $x^{a1}=[x;y;\theta;v]^{\mathrm{T}}$，运动学方程可表示如下：

$$\frac{\mathrm{d}x_{a1}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x \\ y \\ \theta \\ v \end{bmatrix} = \begin{bmatrix} v\cos\theta \\ v\sin\theta \\ v\tan\beta/L \\ a \end{bmatrix} \tag{1}$$

2 方 法

在这一部分中，将车辆的泊车任务建模为多智能体强化学习问题。学习环境被表示为 $E=(S,O,A,P,R,\gamma)$，构成一个部分可观察马尔可夫决策过程（POMDP），定义了智能体的任务。这个环境是在 Unity 中实现的，利用了 mlagents[24] 框架和 OpenAI 的 gym 库[25]。本研究使用近端策略优化（PPO），这是一种同策略 RL 技术[26]；智能体的目标是优化它们的行为，以最大化指定奖励信号的预期累积总和。如图 2 所示，整个系统使用 PPO 算法作为主要框架来优化强化学习策略。考虑到泊车场景中空间有限导致任务实现难度大，从而出现奖励稀疏，模型收敛速度慢。提出使用基于规则的混合 A * 算法生成局部目标，为模型的策略学习提供准确方向，缓解稀疏奖励问题，从而加快模型训练速度。此外，为了更加接近真实场景，本研究不假设可以获取到周围障碍物的精确坐标、尺寸信息，而是通过车辆自身传感器进行感知。仅利用距离传感器作为感知设备，实现了良好的无人泊车效果。接下来的章节中，将详细介绍整个系统。

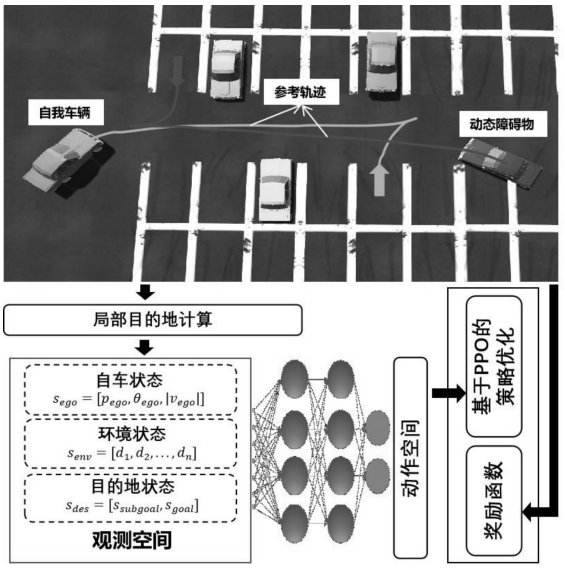

图 2 方法框架

2.1 观测空间

模型中的状态空间 S 包含了 Unity 中整个停车场的信息。然而，由于限制，智能体只能观察到部分状态。在每个时间步 t，真实状态 s_t 产生相应的观测 o_t，

表示智能体的感知。在车辆泊车任务中,考虑到自我车辆的状态 s_{ego} 、周围环境状态 s_{env} 以及目标状态 s_{des} 是至关重要的。因此,每辆车的观测空间表示为 $[s_{ego}, s_{env}, s_{des}]$ 。在接下来的章节中,将详细介绍观测空间的这三个主要组成部分。

首先,自我车辆的状态表示为 $s_{ego}=[p_{ego},\theta_{ego},|v_{ego}|]$,$p_{ego}=(x_{ego},y_{ego})$,如图 3 所示。这些变量表示车辆的位置坐标、方向角和速度大小。因此,为车辆配备了距离传感器,在感知环境方面起到至关重要的作用。

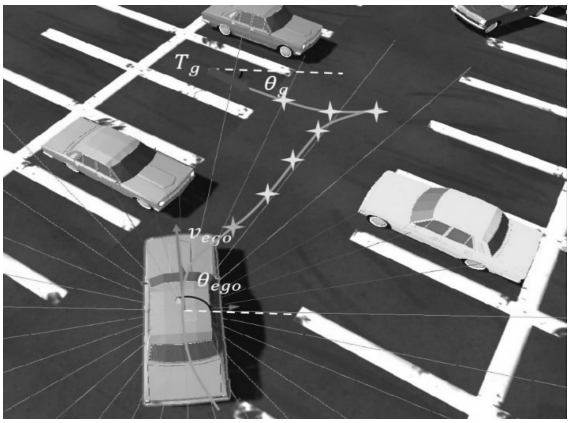

图 3　观测空间

如图 3 所示,它能够检测到其他车辆、墙壁等障碍物,为防止碰撞提供了重要的安全措施。距离传感器提供的感知结果表示为 $s_{env}=[d_1,d_2,\cdots,d_n]$,$n=30$,均匀分布在 360 度的视野范围内。最大检测距离为 30 米。最后,目的地的状态表示为 $s_{des}=[s_{subgoal},s_{goal}]$ 。由于车辆泊车的最终目标是到达具有特定方向角的目的地,与传统的目的地引导任务相比,该任务的奖励更加稀疏。这种稀疏性可能使得深度强化学习模型的学习变得极具挑战性,甚至可能无法学习泊车任务。因此,采用混合 A * [27] 来规划参考轨迹,然后根据车辆的运动模型和目的地的状态,得到车辆的局部目标状态。这种方法将整个任务分解为多个步骤,从而减轻了稀疏奖励的影响。局部目标在每个时间步计算,计算过程在算法 1 中说明。集合 $\boldsymbol{T}_{ego}=\{T_1,T_2,\cdots,T_N\}$ 表示一个 episode 之前通过轨迹规划算法计算得到的轨迹。每个 $T_N=(x_N,y_N,\theta_N,|v_N|)$ 包含了位置坐标 x_N 和 y_N 、速度大小 $|v_N|$ 和方向角 θ_N 。最后,得到了局部目标状态 T_{loc} 。

算法 1:局部目标计算

输入:$p_{ego}=(x_{ego},y_{ego})$ 自我车辆的位置;$\boldsymbol{T}_{ego}=\{T_1,T_2,\cdots,T_N\}$;i_{pre} 表示上一时间步的局部目的地

初始化:$d_{loc}=\infty$;$i_{loc}=0$

输出:T_{loc}

for ($i=1$ to $N-1$) {

$d_i,h_i=\mathrm{Dist}(p_{ego},T_iT_{i+1})$

函数 Dist 用于计算当前位置 p 到线段 T_iT_{i+1} 的距离 d ,以及垂直投影点 $h=(hx,hy)$

if { $x_i<=hx_i<=x_{i+1}$ & $y_i<=hy_i<=y_{i+1}$ }

{

if{ $d_i<d_{loc}$ }

{

$i_{loc}=i$;

}

}

}

if{ $i_{loc}<i_{pre}$ }

{

$i_{loc}=i_{pre}$

}

$T_{loc}=\boldsymbol{T}[i_{loc}]$

$s_{goal}=(x_g,y_g,\theta_g,|v_g|)$ 表示目的地的位置坐标 (x_g,y_g) ,最终方向角 θ_g 和最终速度大小 $|v_g|$ 。

2.2　动作空间

动作空间由两个连续动作组成。第一个动作表示驱动力,记为 F 。它的取值范围为 $[-1,+1]$ 。小于 0 的值表示后退,大于 0 的值表示前进。根据牛顿第二定律,$a=\frac{F}{m}$,其中 m 表示车辆的质量。如公式 1 所示,这个值直接影响车辆的加速度和速度。第二个动作是控制车辆的转向角度。这个值的变化会影响角度 β 的大小。当这个值为 0 时,车轮朝向前方。当值为 1 或-1 时,表示最大右转或左转角度。从公式 1 可以看出,这个动作可以影响车辆的运动角度和方向角。

2.3　奖励函数

强化学习算法旨在通过最大化随时间累积奖励的期望来优化代理的决策。奖励函数在定义强化学习算法优化的任务方面起着关键作用。因此,仔细选择和平衡不同的奖励信号对于实现理想结果至关重要。为了使车辆尽快到达目标点,同时确保避免碰撞和安全导航,并满足方向角的约束条件,以下章节详细解释了奖励函数的每个组成部分。

终止状态奖励 R_{des} 是一种稀疏奖励,包括四个方程,由三个部分组成,如公式 2 所示。在第一部分中,如果车辆位置 p_{ego} 与目标位置 p_g 之间的距离 $D(p_{ego},p_g)$ 小于一个超参数 d_δ ,则认为车辆已经到达目标地点。在这种情况下,分配奖励 R^d_{des} 。

在第二部分中,已到达的车辆速度的大小 $|v_{ego}|$ 还需要接近最终速度 $|v_g|$ 。如果 $\|v_g|-|v_{ego}\|<|v_{delta}|$ 且 $D(p_{ego},p_d)<d_\delta$,则分配奖励 $R^{|v|}_{des}$ 。

最后一部分是为了满足车辆泊车任务中的方向角约束,需要满足条件 $D(p_{ego},p_d)<d_\delta$,并且自身方向角 θ_{ego} 与目标方向角 θ_g 之间的差值 $|\theta_{ego}-\theta_g|$ 小于一个超参数 θ_δ 。如果满足这些条件,额外分配奖

励 R^{θ}_{des}。

$$R_{\text{des}}=\begin{cases}R^{d}_{\text{des}}, & D(p_{\text{ego}},p_g)<d_\delta\\ R^{d}_{\text{des}}+R^{|v|}_{\text{des}}, & D(p_{\text{ego}},p_g)<d_\delta\\ & \&\ \|v_{\text{ego}}|-|v_g\|<|v_\delta\\ R^{d}_{\text{des}}+R^{\theta}_{\text{des}}, & D(p_{\text{ego}},p_g)<d_\delta\\ & \&\ |\theta_{\text{ego}}-\theta_g|<\theta_\delta\\ R^{d}_{\text{des}}+R^{|v|}_{\text{des}}+R^{\theta}_{\text{des}}, & D(p_{\text{ego}},p_g)<d_\delta\\ & \&\ \|v_{\text{ego}}|-|v_g\|<|v_\delta|\\ & \&\ |\theta_{\text{ego}}-\theta_g|<\theta_\delta\end{cases}\tag{2}$$

为了解决终止状态奖励稀疏性的问题,引入了局部目标奖励 R_{loc},该奖励由三个部分组成,如公式 3 所示。与公式 2 不同的是,将目标替换为局部目标,并相应地降低了奖励值。这种降低是必要的,因为车辆需要学会与环境进行交互、避免碰撞,并在泊车过程中执行其他任务。

$$R_{\text{loc}}=\begin{cases}R^{d}_{\text{loc}}, & D(p_{\text{ego}},p_{\text{loc}})<d_\delta\\ R^{d}_{\text{loc}}+R^{|v|}_{\text{loc}}, & D(p_{\text{ego}},p_{\text{loc}})<d_\delta\\ & \&\ \|v_{\text{ego}}|-|v_{\text{loc}}\|<|v_\delta|\\ R^{d}_{\text{loc}}+R^{\theta}_{\text{loc}}, & D(p_{\text{ego}},p_{\text{loc}})<d_\delta\\ & \&\ |\theta_{\text{ego}}-\theta_{\text{loc}}|<\theta_\delta\\ R^{d}_{\text{loc}}+R^{|v|}_{\text{loc}}+R^{\theta}_{\text{loc}}, & D(p_{\text{ego}},p_{\text{loc}})<d_\delta\\ & \&\ \|v_{\text{ego}}|-|v_g\|<|v_\delta|\\ & \&\ |\theta_{\text{ego}}-\theta_{\text{loc}}|<\theta_\delta\end{cases}\tag{3}$$

为了增强局部目标的引导效果,引入了一个奖励项表示为 R_{cloc}。在每个时间步骤中,如果到达局部目标的距离小于上一个时间步骤的距离,则获得奖励;否则,产生惩罚 R_{floc}。此外,在泊车过程中,如果车辆与其他车辆或静态障碍物发生碰撞,则会施加一个惩罚,记为 R_o。此外,为了使车辆尽快完成泊车任务,引入了一个生存惩罚 R_l,鼓励车辆迅速完成泊车任务。

在模拟步骤 t 中,总奖励 R_t 由以下公式给出:

$$R_t=R_{\text{des}}+R_{\text{loc}}+R_o+R_l+R_u+R_{\text{cloc}}+R_{\text{floc}}$$

其中,奖励信号的权重和具体奖励值在表 1 中提供。

表 1 奖励设置

奖励符号	值	奖励符号	值
R^{d}_{des}	1	R_o	-1
$R^{|v|}_{\text{des}}$	0.5	R_l	-0.000 3
R^{θ}_{des}	0.5	R_{floc}	-0.000 25
R^{d}_{loc}	0.01	R_u	-0.002
$R^{|v|}_{\text{loc}}$	0.005	R_{cloc}	-0.000 75
R^{θ}_{loc}	0.005		

2.4 训练策略

在开始训练过程之前,需要配置车辆的运动学参数和环境约束参数等,这些参数如表 2 的上半部分所示。

表 2 训练参数

参数	值	描述
$F_{\max}$/N	3 000	最大驱动力
m/kg	1 500	车辆质量
$\beta_{\max}$/°	40	最大转向角
d_δ/m	0.5	是否到达目标点
θ_δ/°	5	是否满足方向角约束
T/s	0.04	仿真时间步
学习率	$3e-4$	梯度下降更新
γ	0.99	衰减因子
H	15 000	每个 episode 得最大步数
Epochs	3	训练轮数
Batch size	1 024	批尺寸
Buffer Size	10 240	缓冲池尺寸
β	$5e-3$	熵正则化强度
ϵ	0.2	发散阈值

采用简单的全连接神经网络(Fully Connected Neural Network,FCNN)来表示策略,该网络由两个隐藏层组成,每个隐藏层包含 128 个节点。该 FCNN 以智能体的观测(如 3.1 节所述)作为输入,并生成智能体的相应动作。为了训练该网络,采用了 Proximal Policy Optimization(PPO)算法,使用表 2 的下半部分中给出的参数值。在训练过程中,根据不同的泊车任务,为车辆分配不同的起始状态和终止状态。通过考虑起始和终止状态以及环境信息,生成参考轨迹,为车辆提供局部目的地,从而减轻终止状态奖励的稀疏性。

由于在有限空间和机动性条件下执行泊车任务的复杂性,本研究从课程学习(Curriculum Learning)[28]中汲取灵感,并采用了分阶段学习策略来训练 DRL 模型。在训练开始时,由于环境空间较大,车辆的行为类似于在环境中进行随机探索。因此,如果获得奖励的条件过于严格,可能会阻碍模型有效学习观测状态和奖励值之间的对应关系。因此,模型可能难以为车辆做出最佳决策以达到目标状态,从而影响模型的收敛性。因此,将整个深度强化学习过程分为三个阶段,难度逐渐增加。难度主要体现在获得奖励值的条件上,主要是参数 d_δ、θ_δ 和 $|v_\delta|$ 的大小。在第一阶段,设置 $d_\delta=2.5$ m,$\theta_\delta=15^\circ$,$|v_\delta|=0.5$ m/s。在第二阶段,设置 $d_\delta=1.5$ m,$\theta_\delta=10^\circ$,$|v_\delta|=0.25$ m/s。在第三阶段,设置 $d_\delta=0.5$ m,$\theta_\delta=5^\circ$,$|v_\delta|=0$ m。这样的分阶段学习策略使模型能够从简单到困难的场景中学习,

并增进对观测状态和奖励之间关系的理解,从而做出更加令人满意的决策。

3　实　验

本节中,对模型进行了训练,并进行了实证实验以验证其有效性,并与其他现有方法进行了比较。实验在一台高性能图形工作站上进行,该工作站配置如下:2.9 GHz Xeon Gold 6226R CPU、250 GB RAM 和 Quadro RTX 6000 显卡。采用了基于 Unity 开发的 ml-agents 框架提供的 Proximal Policy Optimization (PPO)实现。实验环境配置如表 3 所示。

表 3　实验环境配置

实验环境名称	具体配置
CPU	AMD Ryzen 7 5800H with RadeonGraphics (3.20 GHz)
RAM	16 GB
GPU	Nvidia GeForceGTX 3070
OS	Windows 10
深度学习框架	Pytorch1.5.0
Python	Python3.7.3

3.1　指　标

泊车任务的主要目标是在确保安全和满足最终方向角的前提下,尽快将车辆停放在目的地。因此,成功率、安全性和效率是评估泊车有效性的关键指标。泊车成功率是指在泊车任务中成功停放的车辆在所有参与泊车任务的车辆中所占的比例。假设车辆的停放位置与目的地之间的距离小于 δ_d ,并且停放方向与最终方向角的偏差在 δ_{angle} 范围内,就被认为是成功完成泊车任务。本研究将 δ_d 设为 0.5 m, δ_{angle} 设为 5 °。安全性是指车辆在泊车过程中是否与静态障碍物或其他车辆发生碰撞,通过碰撞次数来衡量泊车过程的安全性。终止方向角是自主泊车任务中的一个关键约束。为了更精确地分析算法对方向角的控制能力,使用另一个成功率指标:无碰撞成功率。与成功率不同,该指标仅考虑在所有到达目的地的车辆中,达到目的地并满足终止方向角的车辆所占的比例,而不考虑由于碰撞而未能到达目的地的车辆。泊车效率是指车辆执行泊车任务的速度。在给定起始和终止状态的情况下,在最大速度的约束下,较高的速度和较短的距离会导致较高的泊车效率。因此,在保持泊车任务一致的同时,根据泊车的持续时间来衡量效率。

3.2　泊车效果

图 4 展示了停车场场景,其中停车位由白色线条组成。没有车辆的车位都可以进行停车。将停车场视为一个二维平面,其中向右为正 x 轴方向,向上为正 y 轴方向,正 y 轴方向的角度定义为 0 °。

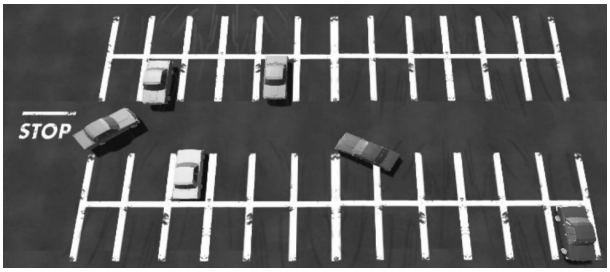

图 4　泊车场景

进行了四组实验,分别是单车、两辆车、四辆车、八辆车同时进行泊车时的泊车效果。实验开始时,在整个停车场中的没有车辆的位置为泊车车辆生成初始状态,然后从停车位中随机为其选择一个作为其泊车任务的终止状态。表 4 为四组实验中泊车车辆的起止状态。

表 4　实验设置

实验组	车辆编号	起始状态	终止状态
实验一	1	238.4,150.5,-128.2°	225.4,144.9,0.6°
实验二	1	238.4,150.5,-128.2°	225.4,144.9,0.6°
	2	218.1,152.8,118.4°	235.1,145.6,0°
实验三	1	238.4,150.5,-128.2°	225.4,144.9,0.6°
	2	218.1,152.8,118.4°	235.1,145.6,0°
	3	234.4,171.1,241.9°	228.4,160.8,0.1°
	4	204.9,171.2,447.5°	209.2,160.7,0°
实验四	1	238.4,150.5,-128.2°	225.4,144.9,0.6°
	2	218.1,152.8,118.4°	235.1,145.6,0°
	3	234.4,171.1,241.9°	228.4,160.8,0.1°
	4	204.9,171.2,447.5°	209.2,160.7,0°
	5	216.6,170.6,558.7°	206.1,155.3,181.1°
	6	250.8,137.5,265.3°	228.7,139.3,179.1°
	7	248.4,166.9,265.3°	212.4,155.2,181.3
	8	197.1,150.8,63.4	206.4,155.2,181.3

“无碰撞成功率”是通过将完成泊车的车辆数量除以在所有车辆中没有发生碰撞的车辆数量得到的。使用所有一次泊车任务中的碰撞次数来表示安全性。为方便起见,将使用 SR 表示成功率,SRWC 表示无碰撞成功率,TT 表示总时间,CN 表示碰撞次数。

如表 5 所示,列出了四组实验的结果。在实验一和实验二中,由于泊车车辆较少,泊车任务几乎不需要处理动态障碍物,仅需要考虑静态障碍物和终止姿态,任务相对简单。因此,这两组实验在成功率和时间方面都取得了很好的效果,并且没有发生碰撞等不安全情况。然而,随着同时进行泊车的车辆数量增加,整个场景的任务复杂度迅速上升。

表 5 调运实验结果

实验组	TT/s	CN	SRWC/%	SR/%
实验一	14.32	0	100	100
实验二	15.72	0	100	100
实验三	16.66	1	100	75
实验四	20.32	2	83.3	62.5

在实验三中,泊车车辆增加到 4 辆,任务的复杂度进一步提高,开始出现少量碰撞。尽管该方法在控制终止状态方面仍然表现良好,但在考虑终止状态的同时出现了一些碰撞情况。在实验四中,泊车车辆增至 8 辆,任务更加复杂,车辆之间会有更多的交互和潜在碰撞。首先,时间增长迅速,与前三组实验相比,实验四的时间增长较快,这是由于更多的碰撞避免过程所导致的。此外,由于需要躲避动态障碍物,实际终止状态出现偏差,因此成功率下降。

3.3 对比实验

为了验证提出的数据和规则混合驱动的全自动代客泊车轨迹规划方法的有效性,与现有的经典深度强化学习 PPO 和 DDPG 进行了实验比较。

图 5 训练曲线

分别用 PPO、DDPG 方法对泊车规划任务进行训练。另外,文中方法是在 PPO 的基础上,增加了基于子目标点计算的课程学习方法,记为 Ours。这三种方法的训练结果如图 5 所示,该图反映了训练步数和奖励函数之间的关系。从图中可以看出,文中方法由于增加了参考轨迹以及课程学习,训练效率大大提升,模型的收敛速度更快,同时,最终的奖励也略高于 PPO 和 DDPG 方法。

另外,为了验证实际的泊车效果,在实验四上进行了实验,进一步验证基于子目标点的课程学习对强化学习泊车规划方法的影响。表 6 为最终的实验结果。从实验结果可以看出,课程学习对模型最终的泊车效果产生了很大影响,尤其是在复杂环境中的泊车成功率和安全性方面都有很大提升。

表 6 消融实验结果

模型	TT/s	CN	SRWC/%	SR/%
Ours	20.32	2	83.3	62.5
DDPG	21.44	5	80	50
PPO	20.89	4	75	50

4 结束语

针对当前基于强化学习的代客泊车方法存在的问题—对周围车辆精确信息的需求以及算法收敛困难,进行了深入研究。为解决这些问题,提出了一种创新的数据和规则混合驱动的强化学习自动代客泊车方法。该方法仅依赖距离传感器,并结合基于规则的轨迹规划方法和课程学习思想,在不同复杂程度的泊车任务中取得了显著的效果。然而,该方法仍面临一些挑战。采用混合 A 算法生成全局轨迹规划,以支持后续子目标点的计算。然而,混合 A 算法偶尔可能无法找到解决方案。为缓解这一问题,通过放宽角度阈值的操作来减少无解情况的发生,但仍存在一定的概率。此外,未来的研究中,计划探索更复杂的泊车场景,例如规模更大的停车场和多层停车场。研究结果表明,采用数据和规则混合驱动的强化学习方法可以克服现有方法的局限性,同时在代客泊车任务中取得优秀的效果。该工作为进一步改进和推进自动驾驶技术提供了有益的启示。未来的研究方向包括进一步优化算法的稳定性和鲁棒性,以应对更复杂的泊车场景,并加强对车辆周围环境信息的处理能力。

参考文献:

[1] CONNER D C,KRESS-GAZIT H,CHOSET H,et al. Valet parking without a valet[C]//Proceedings of the 2007 IEEE/RSJ international conference on intelligent robots and systems. San Diego:IEEE,2007.

[2] MIN K W,CHOI J D. Design and implementation of autonomous vehicle valet parking system[C]//Proceedings of the 16th international conference on intelligent transportation sys-

tems (ITSC 2013). The Hague:IEEE,2013.
[3] 刘宏麾. 基于安全走廊的自主泊车轨迹规划方法研究[D]. 成都:电子科技大学,2024.
[4] 李志晋. 自动泊车的路径规划和跟踪控制研究[D]. 芜湖:安徽工程大学,2023.
[5] DRAGANJAC I,MIKLIĆ D,KOVAČIĆ Z,et al. Decentralized control of multi-AGV systems in autonomous warehousing applications[J]. IEEE Transactions on Automation Science and Engineering,2016,13(4):1433-1447.
[6] JIANG C,HU Z,MOURELATOS Z P,et al. R2-RRT*:reliability-based robust mission planning of off-road autonomous ground vehicle under uncertain terrain environment [J]. IEEE Transactions on Automation Science and Engineering,2021,19(2):1030-1046.
[7] LI B,WANG K,SHAO Z. Time-optimal maneuver planning in automatic parallel parking using a simultaneous dynamic optimization approach[J]. IEEE Transactions on Intelligent Transportation Systems,2016,17(11):3263-3274.
[8] FRAICHARD T,SCHEUER A. From Reeds and Shepp´s to continuous-curvature paths[J]. IEEE Transactions on Robotics,2004,20(6):1025-1035.
[9] 田　杰,叶　青. 自动泊车发展现状及运动规划研究进展[J]. 科学技术与工程,2024,24(21):8825-8836.
[10] KNEISSL M,MADHUSUDHANAN A K,MOLIN A,et al. A multi-vehicle control framework with application to automated valet parking[J]. IEEE Transactions on Intelligent Transportation Systems,2020,22(9):5697-5707.
[11] CHEN L,SHAN Y,TIAN W,et al. A fast and efficient double-tree RRT $^* $-like sampling-based planner applying on mobile robotic systems[J]. IEEE/ASME Transactions on Mechatronics,2018,23(6):2568-2578.
[12] LI Y,CUI R,LI Z,et al. Neural network approximation based near-optimal motion planning with kinodynamic constraints using RRT[J]. IEEE Transactions on Industrial Electronics,2018,65(11):8718-8729.
[13] CHAI R,TSOURDOS A,SAVVARIS A,et al. Two-stage trajectory optimization for autonomous ground vehicles parking maneuver[J]. IEEE Transactions on Industrial Informatics,2018,15(7):3899-3909.
[14] LIU W,LI Z,LI L,et al. Parking like a human:a direct trajectory planning solution[J]. IEEE Transactions on Intelligent Transportation Systems,2017,18(12):3388-3397.
[15] LI B,ACARMAN T,ZHANG Y,et al. Optimization-based trajectory planning for autonomous parking with irregularly placed obstacles:a lightweight iterative framework[J]. IEEE Transactions on Intelligent Transportation Systems,2021,23(8):11970-11981.
[16] CHEN L,AI H,ZHUANG Z,et al. Real-time multiple people tracking with deeply learned candidate selection and person re-identification[C]//Proceedings of the 2018 IEEE international conference on multimedia and expo (ICME). San Diego:IEEE,2018.
[17] CHAI R,TSOURDOS A,SAVVARIS A,et al. Six-DOF spacecraft optimal trajectory planning and real-time attitude control:a deep neural network-based approach[J]. IEEE Transactions on Neural Networks and Learning Systems,2019,31(11):5005-5013.
[18] SHI J,LI K,PIAO C,et al. Model-based predictive control and reinforcement learning for planning vehicle-parking trajectories for vertical parking spaces[J]. Sensors,2023,23(16):7124.
[19] CHAI R,LIU D,LIU T,et al. Deep learning-based trajectory planning and control for autonomous ground vehicle parking maneuver[J]. IEEE Transactions on Automation Science and Engineering,2023,20(3):1633-1647.
[20] KOBER J,BAGNELL J A,PETERS J. Reinforcement learning in robotics:a survey[J]. The International Journal of Robotics Research,2013,32(11):1238-1274.
[21] SAXENA D M,BAE S,NAKHAEI A,et al. Driving in dense traffic with model-free reinforcement learning[C]//Proceedings of the 2020 IEEE international conference on robotics and automation (ICRA). Paris:IEEE,2020.
[22] KESTING A,TREIBER M,HELBING D. Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity[J]. Philosophical Transactions of the Royal Society A:Mathematical,Physical and Engineering Sciences,2010,368(1928):4585-4605.
[23] CHEN S,WANG M,YANG Y,et al. Conflict-constrained multi-agent reinforcement learning method for parking trajectory planning[C]//Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, F, 2023. IEEE.
[24] JULIANI A,BERGES V P,TENG E,et al. Unity:a general platform for intelligent agents[J]. arXiv:1809.02627,2018.
[25] BROCKMAN G,CHEUNG V,PETTERSSON L,et al. Openai gym[J]. arXiv:1606.01540,2016.
[26] SCHULMAN J,WOLSKI F,DHARIWAL P,et al. Proximal policy optimization algorithms[J]. arXiv:1707.06347,2017.
[27] DOLGOV D,THRUN S,MONTEMERLO M,et al. Practical search techniques in path planning for autonomous driving [J]. Ann Arbor,2008,1001(48105):18-80.
[28] BENGIO Y,LOURADOUR J,COLLOBERT R,et al. Curriculum learning[C]//Proceedings of the 26th annual international conference on machine learning. New York:[s. n.],2009.