2025 年 第 25 卷 第 17 期
2025, 25(17):07244-08

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2404338
引用格式:汤向荣, 边桂彬, 李桢, 等. 面向经腔道自主规划的残差强化学习方法[J]. 科学技术与工程, 2025, 25(17): 7244-7251.
Tang Xiangrong, Bian Guibin, Li Zhen, et al. Residual reinforcement learning for autonomous transluminal intervention[J]. Science Technology and Engineering, 2025, 25(17): 7244-7251.

面向经腔道自主规划的残差强化学习方法

汤向荣[1], 边桂彬[1,2], 李桢[2], 马睿宸[2*]
(1. 北京信息科技大学自动化学院, 北京 100192; 2. 中国科学院自动化研究所, 北京 100190)

摘 要 使用连续体机器人的经自然腔道介入面临介入路径曲折狭窄和腔道软组织挤压受力等挑战。针对介入递送过程中,现有规划方法难以兼顾多个控制目标,导致难以到达较深位置的问题,提出一种基于残差强化学习的自主规划方案。该方法能够实现柔性连续体机器人经自然腔道的自主递送。通过建立连续体机器人递送姿态与自然腔道空间状态间的反馈偏差模型来控制递送过程中的姿态目标。同时建立连续体机器人的整体运动过程的马尔可夫模型,用于强化学习算法的训练过程。利用姿态反馈控制与强化学习控制相结合产生的残差策略来输出连续体机器人递送过程的最优动作。在仿真支气管腔道中的实验表明,所提出的方法比现有方法的收敛速度快 60% 以上,能够以平滑、无碰撞的轨迹规划连续体机器人的经腔道介入过程,在多个指标方面优于现有方法。

关键词 连续体机器人; 自主规划; 残差策略; 经腔道介入; 强化学习
中图法分类号 TP242; **文献标志码** A

Residual Reinforcement Learning for Autonomous Transluminal Intervention

TANG Xiang-rong[1], BIAN Gui-bin[1,2], LI Zhen[2], MA Rui-chen[2*]
(1. School of Automation, Beijing Information Science and Technology University, Beijing 100192, China;
2. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

[**Abstract**] The natural orifice intervention using continuum robots faces challenges such as tortuous and narrow intervention paths, as well as compressive forces exerted by soft tissues in the orifice. To address the issue in the delivery process where existing planning methods struggle to balance multiple control objectives, resulting in difficulty in reaching deeper positions, an autonomous planning scheme based on residual reinforcement learning was proposed. The method enables the autonomous delivery of continuum robots through natural orifices. A feedback deviation model between the delivery posture of the continuum robot and the spatial state of the natural orifice was established to control the posture target during the delivery process. Simultaneously, a Markov model of the overall motion process of the continuum robot was constructed to train the reinforcement learning algorithm. A residual strategy, generated by combining posture feedback control with reinforcement learning control, was used to output the optimal actions for the continuum robot's delivery process. Experiments conducted in a simulated bronchial orifice show that the proposed method converges over 60% faster than existing methods and can plan smooth, collision-free trajectories for the continuum robot's intervention through the orifice, outperforming existing methods in several key metrics.

[**Keywords**] continuum robot; autonomous planning; residual strategy; transluminal intervention; reinforcement learning

具有多自由度且能够连续变形的柔性连续体机器人,因其优越的柔顺性能与交互安全性,已在医疗康复、海洋科考、资源勘探等领域引起关注[1]。在形态多变、曲折幽深的人体自然腔道(如呼吸道、消化道等)中,柔性连续体机器人能够实现顺应环境曲率变化的递送过程,已成为手术机器人领域的研究热点[2-3]。为了将连续体机器人递送到狭窄的自然腔道深处,需要保持连续体机器人的整体曲率与其所处的腔道相近,并将连续体机器人的远端维持在腔道截面的中心处。当控制连续体机器人进入曲折幽深的人体腔道时,需要避免机器人远端与腔道壁发生碰撞。鉴于新型连续体机器人所具有的高度灵活性,需要提出一种新颖的连续体机器人自主规划方法,以控制机器人在复杂曲折、直径大

收稿日期: 2024-06-11 **修订日期**: 2025-03-10
基金项目: 北京市杰出青年科学基金(JQ21016)
第一作者: 汤向荣(1997—),男,汉族,江苏泰州人,硕士研究生。研究方向:手术机器人自主规划。E-mail:2021020401@bistu.edu.cn。
* **通信作者**: 马睿宸(1993—),男,汉族,陕西西安人,博士,助理研究员。研究方向:手术机器人。E-mail:maruichen2016@ia.ac.cn。

小不一的自然腔道环境中运动到指定位置。

现有的机器人运动规划研究大多使用传统的搜索和采样规划方法[4-10]。肖瑶等[5]提出一种基于改进 A * 算法的四足机器人规划算法。Hawks 等[6]通过分析连续体机器人的构型空间结构确定约束条件,使用快速搜索随机树(rapid-exploration random tree, RRT)来生成连续体机器人规划过程中的运动学参数。Mbakop 等[7]提出一种具有给定长度参数的空间勾股速度图来建模形状运动学,并用人工势场法控制机器人形状在障碍物中自适应变化。现有的方法只能进入人体腔道中直径较大、曲率较小的位置,且难以兼顾多个控制目标。对于部分难以显式建模的控制目标,现有规划方法难以取得较好的控制效果。

近年来,深度强化学习(deep reinforcement learning, DRL)因其强大的特征提取和决策能力,在机器人和无人系统的控制任务中得到了广泛的应用[11-17]。现代异轨策略,如深度 Q 网络(deep Q-network, DQN)和深度确定性策略梯度(deep deterministic policy gradient, DDPG)等[11-12],可以在与环境实时交互的同时,保持高效的样本利用率。在自主无人系统领域,周治国等[13]提出一种用于无人艇避障规划的改进 DQN 算法。在机器人规划领域,Segato 等[14]提出一种在 GPU 上运行的异步优势演员批评家(asynchronous advantage actor-critic, A3C)算法,用于微创神经外科导管插入术。该场景下机器人的状态空间范围非常小,使得将该方法直接转移到具有不同角度的腔道中具有挑战性。Chi 等[15]提出一种近端策略优化(proximal policy optimization, PPO)结合生成对抗性模仿学习(generative adversarial imitation learning, GAIL)的血管内导丝介入框架。该方法在几种不同的模型中进行了测试。在人体腔道内窥镜检查领域,Li 等[16]使用注意力机制来解释超声图像中的内窥镜姿势,以进行经食管机器人导航。使用内窥镜或超声图像的规划方法需要高质量的医学成像并使用实际的机器人进行训练。该方法专注于在仿真环境中构建和训练模型的框架,以减少对真实机器人的依赖。Wohlke 等[17]提出一种将强化学习方法与演示相结合的分层学习方法。该方法利用以对象为中心的演示分割,将教学轨迹自动分割为片段,同时采用并行训练机制的两级分层模拟学习方法,以同时训练两级策略。该方法在稀疏奖励场景中有较好表现。然而,当在弯曲、狭窄和直径逐渐减小的腔道中运动时,不仅需要考虑角度偏差,还需要考虑连续体机器人远端在腔道中的位置。需要提出一种表征人体腔道空间结构的特征提取方法,用以表征连续体机器人的运动状态变化。

残差强化学习最近已成为解决复杂机器人控制问题的一种有前景的技术[18]。传统的反馈控制方法可以通过显式建模来解决各种机器人控制问题。这些控制方法可以有效缓解强化学习方法中由于状态空间过大带来的灾难性遗忘问题。DRL 方法已被证明能够从与环境的交互中提取状态特征并学习机器人控制策略,这可以解决用反馈控制方法难以建模的控制问题。将困难的连续体机器人控制问题分解为通过传统反馈控制方法有效解决的部分和通过 DRL 解决的残差问题。最终产生的控制策略是这两个控制器的叠加。

综上所述,针对现有规划方法难以兼顾多个控制目标,且控制目标难以显式建模的问题,现提出一种结合残差强化学习和人体腔道空间状态引导的残差 DDPG 算法(residual DDPG, ResDDPG),用于在直径逐渐减小的腔道环境中控制连续体机器人运动到指定位置。提出一种姿态反馈控制器,以建模连续体机器人运动规划问题中的姿态控制目标,并作为传统控制器。建立连续体机器人在人体腔道中运动过程的马尔可夫模型,以满足残差目标的控制要求。最后通过仿真试验证明所提出的算法在收敛速度、到达能力和过程指标方面的优越性。

1　连续体机器人运动环境建模

1.1　人体腔道空间结构特征表示

基于随机采样的运动规划方法需要在栅格地图上采样,并使用碰撞盒完成碰撞检测才能找到可行动作。在具有复杂空间结构的三维环境中,使用这种点云模型的成本过于高昂。使用人体腔道引导线的特征信息来表示人体腔道空间结构。腔道引导线被定义为从起点向腔道内部延伸,并在分叉处分叉开的连续曲线,用于引导递送装置在腔道内前进。

在医学 3D 软件,如 Mimics 中,可使用公开医学影像数据集提取 3D 空间中的多个坐标点数据,重建形成自然腔道中心线段。所提取的中心线段可以近似代替引导线的作用。使用的腔道引导线是由上述中心线段连接而成的空间线段集合。图 1 所示为在 Mimics 中提取的支气管中心线数据。根据支气管腔道的形状特征,可以建立该处腔道的空间特征信息模型,用于 DRL 算法训练。

如图 1 所示,将引导线段绕 x 轴旋转的 α 角定义为旋转角,绕 z 轴旋转的 β 角定义为扭转角。α 的范围为 $[-90°, 90°]$,β 的范围为 $[-180°, 180°]$。

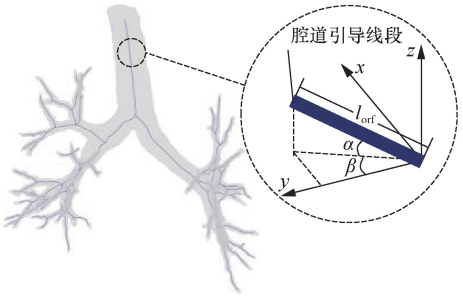


图 1 人体腔道空间结构表示

Fig. 1 Illustration of the human natural orifice spatial structure

腔道引导线段的空间特征信息可以用包含姿态,位置和腔道截面特征的空间信息四元组 $I_{SF}=\{\alpha,\beta,l_{orf},R_{orf}\}$ 来表示,其中,l_{orf} 为一段引导线段的长度,R_{orf} 为引导线段所在自然腔道的直径,即截面特征。

1.2 连续体机器人结构与描述

研究对象是一种由两段主动段和一段被动段组成的连续体机器人。机器人整体呈蛇形连续结构,可通过被动段末端的推送装置送入人体自然腔道深处。机器人主动段能够向所有方向弯曲 90°,两段共可弯曲 180°,因此机器人每段主动段的自由度数目为 3 个。主动段在连续体机器人结构中起到远端导向的作用,能够引导被动段在弯曲的腔道环境中改变方向,以在狭窄、幽深的环境内实现递送功能。被动段将主动段连接至推动装置,具有与主动段相同的弯曲性能。

如图 2 所示,连续体机器人的姿态可以通过曲率角 θ 和偏转角 φ 来描述,且 $\theta\in[0°,90°]$,$\varphi\in[0°,360°]$。因此连续体机器人的正向运动学可以通过分段常曲率(piecewise constant curvature, PCC)模型[19]来描述。

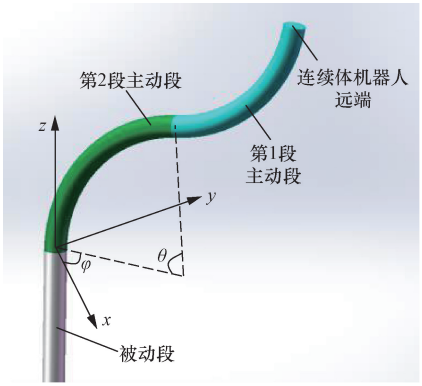
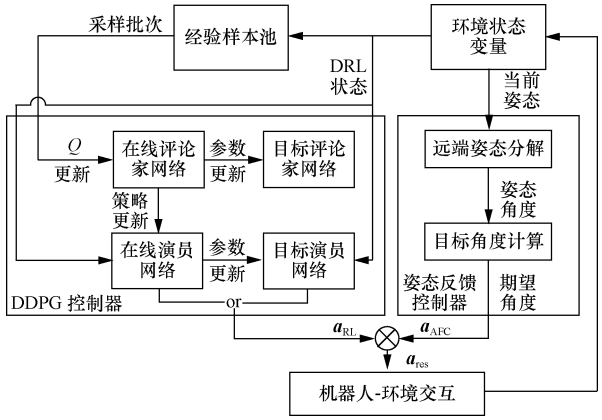


图 2 连续体机器人结构示意图

Fig. 2 Schematic of continuum robot

2 自主规划策略框架

在第 1 节建立的环境和机器人模型的基础上,提出一种基于残差强化学习的连续体机器人自主规划框架。图 3 显示了所提出策略的总体架构。所提出的方法由两个控制器组成。首先是基于传统的反馈控制理论,控制连续体机器人沿着输出与期望姿态之间的误差最小化的方向移动。在传统控制器的基础上叠加 DRL 控制器,用以为运动过程中其他难以显式建模的控制目标建立马尔可夫转移模型,并求解出整体的最优策略。



a_{res} 为算法的整体策略的输出;a_{RL} 为强化学习控制器的输出;a_{AFC} 为姿态反馈控制器的输出

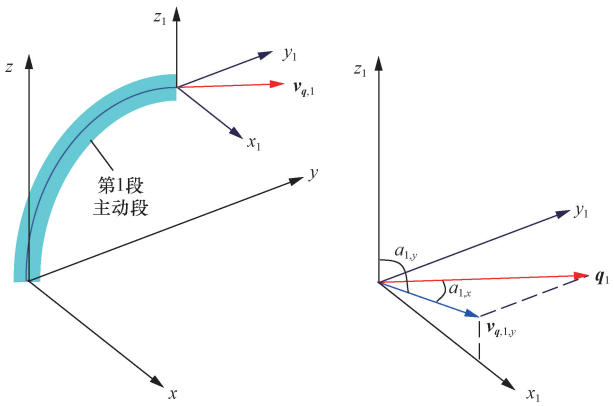
图 3 残差强化学习框架

Fig. 3 Residual reinforcement learning framework

2.1 姿态反馈控制器

提出一种姿态反馈控制器(attitude feedback controller, AFC)作为 ResDDPG 中的传统控制部分。该控制器建模经腔道介入中的期望姿态,以连续体机器人两段主动段远端的姿态与对应引导线角度之间的偏差作为最小化目标。具体来说,对每段主动段,当前所处引导线位置定义为到远端中点具有最短垂线段的引导线段。

连续体机器人主动段远端姿态表示如图 4 所示。以第 1 段主动段为例,单段主动段远端的姿态可以用姿态四元数 $\boldsymbol{q}_1$ 来表示。在标准 PCC 模型的



(a) 连续体机器人远端姿态 (b) 远端姿态分解示意图

图 4 连续体机器人远端姿态示意图

Fig. 4 Illustration of continuum robot distal end attitude

框架下,四元数旋转轴 $\boldsymbol{v}_{q,1}$ 为

$$\boldsymbol{v}_{q,1} = [x_{q,1}, y_{q,1}, z_{q,1}]^{\mathrm{T}} = \boldsymbol{T}\boldsymbol{v}_z \tag{1}$$

式(1)中:$\boldsymbol{v}_z = [0,0,1]^{\mathrm{T}}$ 为一个垂直于 z 轴的初始方向向量;$\boldsymbol{T}$ 为 PCC 模型中的姿态变换矩阵。

在 ResDDPG 中,连续体机器人远端姿态表示是通过分解旋转轴 $\boldsymbol{v}_{q,1}$ 的姿态得到的。在标准 PCC 模型下,连续体机器人在 x 轴方向的姿态角度 $a_{1,x}$ 可以通过轴 $\boldsymbol{v}_{q,1}$ 及其投影向量的夹角来定义为

$$a_{1,x} = \arcsin \frac{y_{q,1}}{\| \boldsymbol{v}_{q,1} \|_2} \tag{2}$$

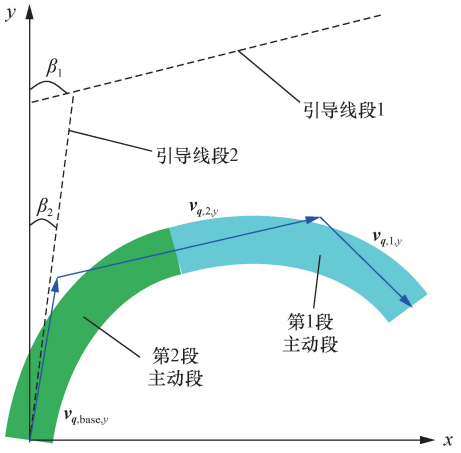
同样,在 y 轴方向的姿态角度 $a_{1,y}$ 可以通过轴 $\boldsymbol{v}_{q,1}$ 投影向量 $\boldsymbol{v}_{q,1,y}$ 与 z 轴的夹角来定义为

$$a_{1,y} = \arccos \frac{z_{q,1}}{\| \boldsymbol{v}_{q,1,y} \|_2} \tag{3}$$

由上述姿态角的求解过程可得两个角度的范围,即 $a_{1,x}, a_{1,y} \in [-90°, 90°]$。按照类似方法,可以定义第二段主动段与被动段的姿态角度 $a_{2,(x,y)}$ 和 $a_{\mathrm{base},(x,y)}$。

姿态反馈控制器的原理如图 5 所示。假设被动段尾部的固定坐标系与引导线方向相同。由于关节执行器的角度与中心线的角度相似,如图 5 所示,对于第二段主动段,下一时刻的期望角度是与引导线对应的角度。对于第一段主动段,下一时刻的期望角度需要考虑第二个主动段在下一时刻产生的角度。因此,连续体机器人的期望角度可以用 $a_{\mathrm{base},(x,y)}$ 与引导线角度之间的偏差来表示为

$$\begin{cases} a^{\mathrm{afc}}_{\mathrm{goal},2,x} = \alpha_2 - a_{\mathrm{base},x} \\ a^{\mathrm{afc}}_{\mathrm{goal},1,x} = \alpha_2 - a_{2,x} - a_{\mathrm{goal},2,x} + a_{\mathrm{base},x} \\ a^{\mathrm{afc}}_{\mathrm{goal},2,y} = \beta_2 - a_{\mathrm{base},y} \\ a^{\mathrm{afc}}_{\mathrm{goal},1,y} = \beta_2 - a_{2,y} - a_{\mathrm{goal},2,y} + a_{\mathrm{base},y} \end{cases} \tag{4}$$



β_1、β_2 分别为与第一、二段主动段对应的引导线扭转角;$v_{q,\mathrm{base},y}$ 为被动段姿态轴的投影向量

图 5　姿态反馈控制器原理

Fig. 5　Principle of attitude feedback controller

由此 AFC 的输出可以表示为

$$\boldsymbol{a}_{\mathrm{AFC}} = [a^{\mathrm{afc}}_{\mathrm{goal},1,x}, a^{\mathrm{afc}}_{\mathrm{goal},1,y}, a^{\mathrm{afc}}_{\mathrm{goal},2,x}, a^{\mathrm{afc}}_{\mathrm{goal},2,y}]^{\mathrm{T}} \tag{5}$$

式中:$a_{\mathrm{base},(x,y)}$ 为被动段在 x 轴和 y 轴方向的姿态角度;$a^{\mathrm{afc}}_{\mathrm{goal},1,x}$、$a^{\mathrm{afc}}_{\mathrm{goal},1,y}$、$a^{\mathrm{afc}}_{\mathrm{goal},2,x}$、$a^{\mathrm{afc}}_{\mathrm{goal},2,y}$ 分别为第一、二段主动段在 x 轴和 y 轴方向设定的期望角度。

2.2　强化学习控制器

强化学习控制器使用 DDPG[12] 实现,旨在建模传统控制器中难以建模的控制目标。DRL 的状态空间被设计为包括比角度偏差更精细的特征,包括当前角度、距离和当前腔道的半径 R_{orif},DRL 的状态空间 s 表达式为

$$\boldsymbol{s} = [\alpha_{1,2}, \beta_{1,2}, a_{(1,2,\mathrm{base}),(x,y)}, d, R_{\mathrm{orif}}, a^{\mathrm{afc}}_{\mathrm{goal},(1,2),(x,y)}]^{\mathrm{T}} \tag{6}$$

式(6)中:d 为连续体机器人远端与引导线段间的距离;$\alpha_{1,2}$ 和 $\beta_{1,2}$ 分别为第一、二段主动段对应引导线的旋转角和扭转角。DDPG 的状态空间同时还包括了 AFC 的输出,以使强化学习算法更好理解环境。

动作空间 $a \in A$:$a \in \mathbf{R}^4$ 包括两个主动段的期望姿态角度 $a^{\mathrm{rl}}_{\mathrm{goal},(1,2),(x,y)}$,共 4 个参数。强化学习控制器的输出为

$$\boldsymbol{a}_{\mathrm{RL}} = [a^{\mathrm{rl}}_{\mathrm{goal},1,x}, a^{\mathrm{rl}}_{\mathrm{goal},1,y}, a^{\mathrm{rl}}_{\mathrm{goal},2,x}, a^{\mathrm{rl}}_{\mathrm{goal},2,y}]^{\mathrm{T}} \tag{7}$$

2.3　残差策略框架

在自主介入任务中,连续体机器人的总体奖励 r_{res} 可以表示为

$$r_{\mathrm{res}} = r_{\mathrm{RL}} + r_{\mathrm{AFC}} \tag{8}$$

式(8)中:r_{AFC} 为使用 AFC 能够实现优化的目标;r_{RL} 为使用 DRL 能够更好学习的部分目标。AFC 能引导连续体机器人以最小化姿态误差的方式完成递送过程,仅能作为单一目标的控制器。而 DDPG 能够建模其他控制目标的误差。以 AFC 为基础,引导 DDPG 算法的训练过程,就形成了 ResDDPG 算法。如图 4 所示,算法的整体策略输出为

$$\boldsymbol{a}_{\mathrm{res}} = \boldsymbol{a}_{\mathrm{RL}} + \boldsymbol{a}_{\mathrm{AFC}} \tag{9}$$

在残差策略中,经验样本池存储了交互产生的状态-动作-奖励对。其中 DDPG 产生的 $\boldsymbol{a}_{\mathrm{RL}}$ 被作为动作存储,而 $\boldsymbol{a}_{\mathrm{AFC}}$ 被作为状态空间的一部分存储,以用于强化学习算法的训练。

经验样本池中的样本经采样回放后,评论家网络的更新遵循确定性贝尔曼方程,即

$$Q^{\pi}(\boldsymbol{s}_t, \boldsymbol{a}^t_{\mathrm{RL}}) = E_{\boldsymbol{s}_{t+1} \sim E}[r(\boldsymbol{s}_t, \boldsymbol{a}^t_{\mathrm{RL}}) + \gamma Q^{\pi}(\boldsymbol{s}_{t+1}, \boldsymbol{\pi}_{\boldsymbol{s}_{t+1}})] \tag{10}$$

式(10)中:Q 为状态价值函数;π 为给定策略;r 为奖励函数;γ 为折扣系数。因此更新评论家网络的损失函数 L 为

$$L(\omega) = E_{\pi'}\{[Q(\boldsymbol{s}_t, \boldsymbol{a}^t_{\mathrm{RL}}) - y_t]^2\} \tag{11}$$

式(11)中:ω 为评论家网络参数,优化目标 y_t 为

$$y_t = r_{\mathrm{RL}} + \gamma Q[s_{t+1}, \pi_{\Phi_{t+1}}(s_{t+1})] \tag{12}$$

经过训练的评论家网络被用于指导策略网络的训练。策略网络的更新遵循确定性策略梯度定理,即

$$\nabla_{\Phi} J(\Phi) = E_{s\sim p_{\pi}}[\nabla_{a_{\mathrm{RL}}} Q^{\pi}(s, a_{\mathrm{RL}})\, \nabla_{\Phi} \pi_{\Phi}(s)] \tag{13}$$

式(13)中:J 为策略梯度定理的目标函数;Φ 为网络参数。在更新了在线网络之后,采用滑动平均方式更新目标评论家和策略网络,得

$$\begin{cases} \omega' = \tau\omega + (1-\tau)\omega' \\ \varphi' = \tau\varphi + (1-\tau)\varphi' \end{cases} \tag{14}$$

式(14)中:τ 为滑动平均系数;φ 为策略网络参数。

3 实验与结果

3.1 实验设置

在一段典型的人体支气管腔道仿真环境中进行训练和实验,以验证所提出腔道内自主规划算法的性能。所建立的支气管腔道的路径从气管入口开始,到左上叶前节段支气管终止,包括气管、左主支气管、左上叶支气管和左上叶前段支气管。直径分别为 18、16、8 和 5 mm。该路径的总体弯曲角度接近 180°,由于其较大的曲率和较小的直径,这段路径被外科医生认为是最难递送的位置之一。在仿真环境中进行的实验验证了算法到达叶支气管和段支气管等位置的能力。

将 ResDDPG 与当前最先进的连续体机器人运动规划算法进行了比较,同时测试了 ResDDPG 的两个组成部件的性能。进行对比的算法如下。

(1)ResDDPG 算法。

(2)原始 DDPG 算法。

(3)文献[9]中的轨迹跟踪算法。

(4)文献[10]中的交叉熵快速搜索随机树(cross-entropy RRT, CE-RRT)算法。

此外,还比较了两种初始引导策略,即 AFC 控制器和随机初始策略的性能差异。

3.2 训练结果

图 6 显示了算法训练过程中奖励曲线的变化情况。在样本采集阶段,ResDDPG 使用 AFC 和随机动作的混合输出与环境交互,而 DDPG 使用单纯的随机策略与环境交互。此时两智能体的得分都较低,而由于 ResDDPG 的动作建模了连续体机器人姿态误差,因此其得分高于 DDPG。在 250 回合后,两种算法的奖励值开始上升,且 ResDDPG 分数高于原始 DDPG。

在 600 回合后,ResDDPG 算法收敛到最大值,并在奖励最大值附近作小幅震荡。而 DDPG 算法在约 1 400 回合训练后收敛。可以看出,ResDDPG 的收敛速度比原有 DDPG 快 60% 左右。这体现了在强化学习算法训练过程中采用传统控制器引导策略搜索产生的显著加速效果。

从图 7 可以看出,在训练过程中,与 DDPG 相比,ResDDPG 能够更频繁且更早地到达直径为5 mm 的段支气管。这凸显了残差策略引导带来的训练效果的提高。

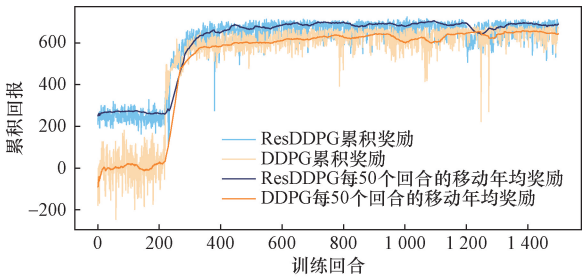


图 6 训练累积奖励

Fig. 6 Training cumulative rewards

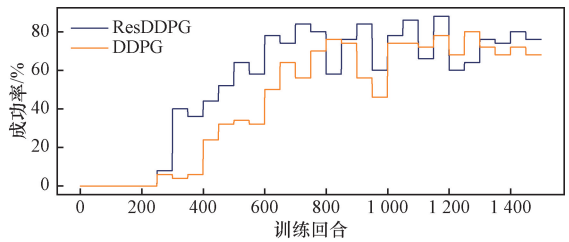


图 7 训练回合与成功率

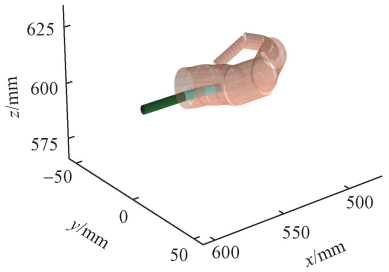
Fig. 7 Training epochs and success rate

3.3 自主规划方法验证

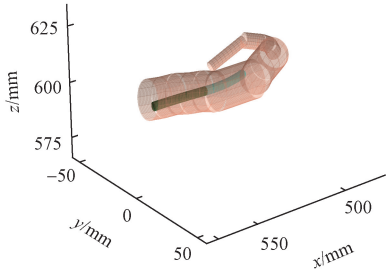
在具有随机 I_{SF} 参数的环境中测试了上述方法的性能。图 8 和图 9 展示了连续体在支气管内的运动过程和远端轨迹。可以看出,所提出的方法能够控制连续体机器人跟随路径曲率变化,运动到目标位置处。

表 1 显示了测试过程中智能体达到目标 1(左上叶支气管,直径为 7 mm)和目标 2(左上叶前段支气管,直径为 5 mm)的成功率。4 种方法均能以较大概率到达直径较大、曲率较小的目标 1 处的概率较高,其中 ResDDPG 最高。当进入直径更加细小的目标 2 时,只有 ResDDPG 的成功率超过 90%,而对比方法的成功率均有较大下降。

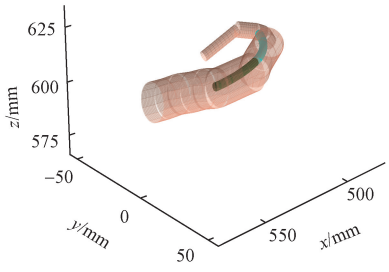
图 10 显示了连续体机器人远端与支气管之间碰撞的累积次数。碰撞的测量是通过将远端与引导线之间的距离与远端所在支气管的直径进行比较来完成的。可以看出,ResDDPG 产生的碰撞比其他方法少得多,并且仅在进入叶支气管后发生接触。由于较浅位置处的支气管直径较大,因此经过训练的方法很少在此产生碰撞。



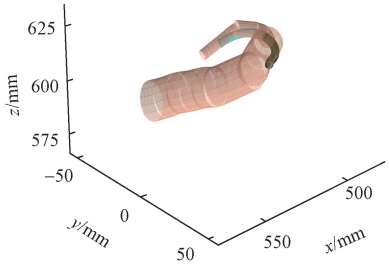
(a) 回合参数为20



(b) 回合参数为60



(c) 回合参数为100



(d) 回合参数为140

图 8　支气管自主介入过程

Fig. 8　Bronchial autonomous interventional procedure

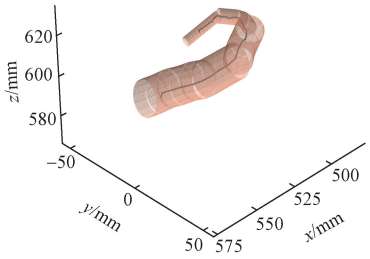
表 1　智能体到达结果

Table 1　The reach results of agents

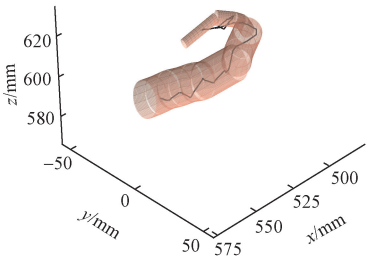
智能体	测试次数	成功次数		成功率/%	
		目标 1	目标 2	目标 1	目标 2
ResDDPG	2 000	1 956	1 809	97.80	90.45
DDPG	2 000	1 931	1 646	96.55	82.30
轨迹跟踪[9]	2 000	1 722	1 473	86.10	73.65
CE-RRT[10]	2 000	1 279	1 137	63.95	56.85

图 11 显示了以连续体机器人远端距引导线距离表示的位置误差随时间变化的情况。可以看出,ResDDPG 在整个过程中的误差值均较低,表现出较好的跟踪引导线运动的能力。与 ResDDPG 相比,原有 DDPG 和轨迹跟踪方法表现出稍高的位置误差,在到达终点时与支气管引导线的偏差明显更大。CE-RRT 方法产生的位置误差较大。

图 12 显示了运动过程中角度误差绝对值的总体情况。姿态误差使用曲率角 θ 表示。可以看出,仅依靠角度偏差作为控制标准的 AFC 表现出最小的整体角度误差。然而,其整体控制性能不如其他方法。除 AFC 之外,ResDDPG 的整体角度误差最小,且 AFC 的姿态误差变化情况与 ResDDPG 相似,表明 ResDDPG 学习到了与 AFC 相近的姿态调整策略。



(a) ResDDPG 轨迹



(b) AFC 轨迹

图 9　连续体机器人远端轨迹

Fig. 9　Continuum robot distal end trajectory

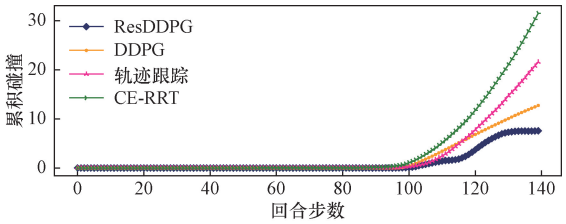


图 10　自主规划策略累积碰撞

Fig. 10　Autonomous planning policy cumulative collision

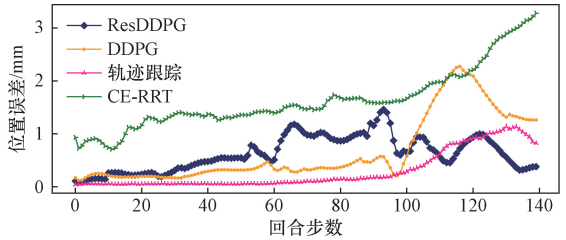


图 11　自主规划策略位置误差

Fig. 11　Autonomous planning policy position error

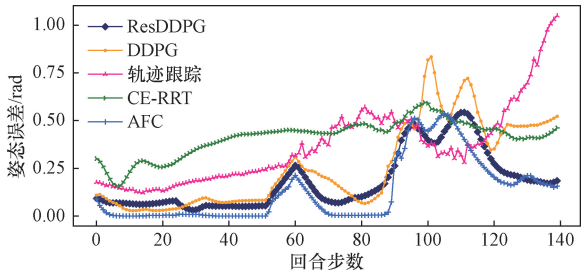


图 12 自主规划策略姿态误差

Fig. 12 Autonomous planning policy attitude error

与 ResDDPG 相比,轨迹跟踪方法通过跟踪路径上不断变化的轨迹点来控制连续体机器人向目标位置的移动[9]。这种跟踪确保了机器人远端的高位置精度。然而,这种方法在运动过程中不考虑整体姿势,导致在整个运动过程中产生了较大的姿态误差。

CE-RRT 使用交叉熵方法来优化随机参数间隔,从而提高路径质量[10]。然而,用随机参数生成的动作往往是不稳定且频繁振荡的,导致连续体机器人在狭窄的支气管中发生许多碰撞和接触。作为全局规划算法,RRT 每次只能搜索一条路径,当路径参数改变时,就必须重新规划,这也限制了其应用。

与原有 DDPG 相比,ResDDPG 将传统控制器与确定性策略梯度相结合。利用初始引导策略为 DRL 提供高质量初始样本,加速最优策略的搜索。实验结果表明,采用残差引导策略可以提高 DRL 的动作质量,从而提高了过程指标和到达能力。

3.4 姿态反馈控制器性能分析

图 13 和图 14 比较了两种初始引导策略(即姿态反馈控制器和随机初始策略)的性能。可以看出,AFC 可以产生更好的动作,促进算法快速收敛到最优策略。在生成动作的质量方面,AFC 输出动作引起的碰撞和位置误差远小于随机策略产生的碰撞和位置误差。因此在图 6 中,当两种算法处于样本采集阶段时,由姿态反馈控制器产生的策略所获得的分数远高于随机策略的分数。由于随机策略产生的优秀动作较少,因此智能体找到最优动作的速度也降低了,收敛速度也相应变慢。

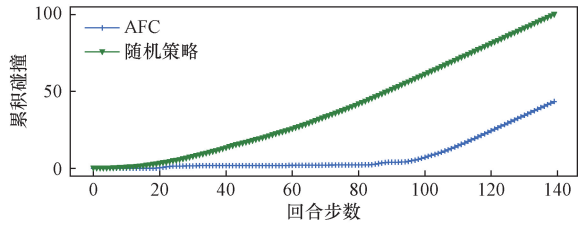


图 13 引导策略累积碰撞

Fig. 13 Guide policy cumulative collisions

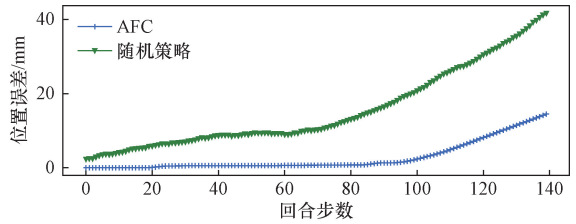


图 14 引导策略位置误差

Fig. 14 Guide policy position error

4 结论

提出一种基于残差策略引导的深度强化学习算法,以规划连续体机器人经自然腔道自主介入的运动过程。实验结果表明,自主规划方法可以连续无碰撞地规划在具有不同曲率的腔道中的运动过程。与现有策略相比,所提出的策略在收敛速度、到达能力和过程指标方面都有显著提高。在未来的工作中,将研究动态环境中连续体机器人的规划方法,以满足现实世界中的机器人控制需求。

参 考 文 献

[1] 梅栋, 赵鑫, 唐刚强, 等. 软体机器人建模与控制技术研究进展[J]. 机器人, 2024, 46(2): 234-256.
Mei Dong, Zhao Xin, Tang Gangqiang, et al. A review of soft robot modeling and control[J]. Robot, 2024, 46(2): 234-256.

[2] Runciman M, Darzi A, Mylonas G. Soft robotics in minimally invasive surgery[J]. Soft Robotics, 2019, 6(4): 423-443.

[3] Bian G B, Wang S, Li Z, et al. Design and Nonlinear error compensation of a multi-segment soft continuum robot for pulmonary intervention[J]. IEEE Transactions on Medical Robotics and Bionics, 2023, 5(4): 832-842.

[4] 张振国, 毛建旭, 谭浩然, 等. 重大装备制造多机器人任务分配与运动规划技术研究综述[J]. 自动化学报, 2024, 50(1): 21-41.
Zhang Zhenguo, Mao Jianxu, Tan Haoran, et al. A review of task allocation and motion plan-ning for multi-robot in major equipment manuf-acturing[J]. Acta Automatica Sinica, 2024, 50(1): 21-41.

[5] 肖瑶, 王强, 金仲平, 等. 基于改进 A * 算法的燃气微泄漏四足巡检机器人路径规划[J]. 科学技术与工程, 2024, 24(13): 5421-5426.
Xiao Yao, Wang Qiang, Jin Zhongping, et al. Path planning of gas micro-leakage quadruped inspection robot based on improved A * algorithm[J]. Science Technology and Engineering, 2024, 24(13): 5421-5426.

[6] Hawks Z, Frazelle C, Green K E, et al. Motion planning for a continuum robotic mobile lamp: defining and navigating the configuration space[C]//. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Macau: IEEE, 2019: 2559-2566.

[7] Mbakop S, Tagne G, Drakunov S V, et al. Parametric pH curves model based kinematic control of the shape of mobile soft manipulators in unstructured environment[J]. IEEE Transactions on Indus-

trial Electronics, 2022, 69(10): 10292-10300.

[8] 龙琴, 袁森, 李魏魏. 基于快速平稳幂次趋近律 AGV 滑模轨迹跟踪控制研究[J]. 科学技术与工程, 2024, 24(8): 3276-3283.
Long Qin, Yuan Sen, Li Weiwei. AGV sliding mode trajectory tracking control based on fast stationary power reaching law[J]. Science Technology and Engineering, 2024, 24(8): 3276-3283.

[9] Ni Y, Sun Y, Zhang H, et al. Data-driven navigation of ferromagnetic soft continuum robots based on machine learning[J]. Advanced Intelligent Systems, 2023, 5(2): 202200167.

[10] Chen J, Yan J, Qiu Y, et al. A cross-entropy motion planning framework for hybrid continuum robots[J]. IEEE Robotics and Automation Letters, 2023, 8(12): 8200-8207.

[11] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J]. ArXiv Preprint ArXiv, 2013:1312.5602.

[12] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[C]//4th International Conference on Learning Representations. San Juan: Ithaca, 2016: 1-14.

[13] 周治国, 余思雨, 于家宝, 等. 面向无人艇的 T-DQN 智能避障算法研究[J]. 自动化学报, 2023, 49(8): 1645-1655.
Zhou Zhiguo, Yu Siyu, Yu Jiabao, et al. Research on T-DQN intelligent obstacle avoidance algorithm of unmanned surface vehicle [J]. Acta Automatica Sinica, 2023, 49(8): 1645-1655.

[14] Segato A, Sestini L, Castellano A, et al. GA3C reinforcement learning for surgical steerable catheter path planning[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). Paris: IEEE, 2020: 2429-2435.

[15] Chi W, Dagnino G, Kwok T, et al. Collaborative robot-assisted endovascular catheterization with generative adversarial imitation learning[C]//2020 IEEE International Conference on Robotics and Automation. Paris: IEEE, 2020: 2414-2420.

[16] Li K, Li A, Xu Y, et al. RL-TEE: autonomous probe guidance for transesophageal echocardiography based on attention-augmented deep reinforcement learning[J]. IEEE Transactions on Automation Science and Engineering, 2024, 21(2): 1526-1538.

[17] Wohlke J, Schmitt F, Hoof H. Hierarchies of planning and reinforcement learning for robot navigation[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an: IEEE, 2021: 10682-10688.

[18] Johannink T, Bahl S, Nair A, et al. Residual reinforcement learning for robot control[C]//2019 International Conference on Robotics and Automation (ICRA). Montreal: IEEE, 2019: 6023-6029.

[19] Webster R, Jones B. Design and kinematic modeling of constant curvature continuum robots: a review[J]. The International Journal of Robotics Research, 2010, 29(13): 1661-1683.