空间机器人抓捕非合作目标的自主强化学习控制

刘帅¹, 邬树楠^{1*}, 刘宇飞², 吴志刚¹, 毛子铭³

1. 大连理工大学航空航天学院, 大连 116024;

2. 中国空间技术研究院钱学森空间技术实验室, 北京 100094;

3. 大连理工大学工程力学系, 大连 116024

*联系人, E-mail: wushunan@dlut.edu.cn

收稿日期: 2018-08-18; 接受日期: 2018-11-05; 网络出版日期: 2019-01-08

国家自然科学基金(编号: 91748203)资助项目

摘要 近年来, 空间机器人在轨服务已成为许多国家的研究热点. 本文针对空间机器人抓捕非合作目标任务, 提出了一种强化学习控制与PD控制组成的双回路控制方法, 对空间机器人基座平台姿态与机械臂运动进行控制. 首先, 对空间机器人的空间任务进行分析, 建立包含基座平台姿态与机械臂运动的空间机器人耦合动力学模型; 然后, 设计双回路控制系统分别对机械臂运动与基座平台姿态进行控制, 内回路中将强化学习与模糊理论结合在一起设计控制器对机械臂末端运动进行控制, 外回路中采用PD控制对基座平台姿态进行稳定控制; 最后, 使用所提控制方法进行数值仿真, 并与传统PD控制方法作对比, 验证所提控制方法的有效性. 结果表明, 强化学习控制下的机械臂运动过程平稳、控制精度高, 与传统PD控制方法相比, 具有一定的自主学习能力, 更加适应抓捕目标的非合作特性.

关键词 空间机器人, 非合作目标, 强化学习, 机械臂, 模糊理论

PACS: 47.27.-i, 47.27.Eq, 47.27.Nz, 47.40.Ki, 47.85.Gj

1 引言

近年来, 越来越多的卫星被发射升空, 尽管卫星任务失败大部分是因为运载火箭发射失败, 但在轨故障也是导致卫星任务失败的重要原因, 还有许多卫星因为燃料耗尽或电力供应不足而无法继续完成工作, 严重影响卫星的使用寿命^[1]. 由于恶劣的太空环境对于宇航员舱外作业有着较大风险, 而空间机器人完全可以替代人类完成这些舱外空间作业, 可广泛应用于在

轨维修、燃料加注与在轨组装等多方面的在轨服务任务中, 已成为多个国家的研究热点^[2-6].

空间机器人系统一般由两部分组成, 包括基座平台与附着在基座平台上的机械臂. 空间机器人在进行在轨工作时, 由于目标航天器运动的不确定性, 在接触目标航天器时会出现不确定的干扰. 此外, 机械臂与基座平台组成高度耦合的动力学系统, 机械臂的运动会影响空间机器人基座平台的姿态稳定, 这些问题都对空间机器人的在轨稳定控制提出了较大的挑战,

引用格式: 刘帅, 邬树楠, 刘宇飞, 等. 空间机器人抓捕非合作目标的自主强化学习控制. 中国科学: 物理学 力学 天文学, 2019, 49: 024512
Liu S, Wu S N, Liu Y F, et al. Autonomous reinforcement learning control for space robot to capture non-cooperative targets (in Chinese). Sci Sin-Phys Mech Astron, 2019, 49: 024512, doi: 10.1360/SSPMA2018-00287

关系到在轨服务任务的成功与否^[7-9].

针对这些问题,许多学者提出了针对空间机器人系统稳定控制的方法.Xu等人^[10]针对空间机器人基座平台可控的情况设计了自适应控制方法,并在关节空间中对机械臂控制进行了仿真计算.Chu等人^[11]提出了带扰动观测器的鲁棒控制策略,将模型的不确定性与外界的干扰集中在一起,在机械臂的每个关节处设计了扰动观测器,结合PD控制完成对机械臂的控制.Busoniu课题组^[12]提出了自适应神经网络控制方法,对不确定部分用神经网络近似,并证明了系统的稳定性.de Taveira等人^[13]将H ∞ 控制与神经网络结合在一起,对自由漂浮空间机器人的操作手进行了控制,并考虑了模型的不确定性与外界的扰动.魏承等人^[14]在进行空间机器人抓捕漂浮目标的研究中,引入末端装置抓取目标时的碰撞模型,提出了“动态抓取域”用于机械臂抓取目标时的控制,同时应用关节主动阻尼控制,以减小抓取碰撞激振对空间机器人冲击的影响.于潇雁和陈力^[15]针对载体位置、姿态均不受控并具有有界外部扰动的漂浮柔性两杆空间机械臂振动进行了不主动抑制的全局鲁棒Terminal滑模控制.王明等人^[16]在自由漂浮空间机器人系统中,考虑基座平台的反作用干扰,对机械臂运动进行了轨迹优化.Oda等人^[17]通过估计机械臂产生的角动量,然后对基座平台姿态进行补偿控制,对机械臂运动与基座平台姿态进行了协调控制.徐文福等人^[18]提出在自由漂浮基下,由空间机械臂自身的控制实现机械臂与基座平台协调运动的方法.Jayakody等人^[19]考虑反应轮的方面来重新描述自由飞行空间机器人的动力学方程,应用一种新的自适应变结构控制方法实现了不受系统不确定性影响的空间机器人鲁棒协调控制器.Hu等人^[20]针对空间机械手轨迹跟踪问题,使用分散递推控制策略,设计了分散鲁棒控制器.

空间机械臂的运动与基座平台的运动相互耦合,为提高机械臂抓捕的精度与稳定度,需要同时控制机械臂和基座平台的运动.此外,机械臂的运动导致系统的动力学参数发生变化,因此抓捕过程中的空间机器人是一个时变的非线性耦合系统.若目标是失控或非合作航天器,在抓捕过程中它会自由漂移或主动做出规避动作.上述文献中的控制器虽然可获得良好的控制效果,但设计中需要已知空间机器人系统精确的动力学模型,且并没有考虑目标的非合作特性.针对

这些问题,本文提出一种基于强化学习的抓捕控制方法.所设计的控制系统包含两个回路,在外回路中设计PD控制器以保证抓捕过程中空间机器人基座平台的姿态稳定;在内回路中,提出一种基于强化学习的机械臂运动控制器以实现非合作目标的抓捕机动.所提出的强化学习控制方法不需要获知机械臂的动力学模型,是在与外界环境的交互中学习得到控制律,能提高空间机器人系统在轨工作的自主性.

2 问题描述与动力学模型

2.1 问题描述

空间机器人完成在轨抓捕任务一般分为三个阶段:观测与任务规划阶段;自主接近目标航天器阶段;抓捕目标及抓捕后阶段.本文研究的是抓捕过程中机械臂的运动控制及基座平台的稳定控制问题.

在机械臂展开接近目标航天器的过程中,空间机器人系统的质量参数会发生改变,当机械臂末端的操作手接触目标航天器时,由于目标航天器运动状态的不确定性,对于机械臂末端会造成一个不确定的瞬间扰动,另外机械臂的运动与空间机器人的基座平台运动组成一个高度耦合的动力学系统,会对基座平台的稳定平衡造成一定的影响,这些问题都可能导致空间任务的失败.因此有必要进行空间机器人机械臂运动及基座平台姿态的控制.

在进行空间机器人抓捕非合作目标问题研究前,做出以下假设:(1)空间机器人系统为刚体系统;(2)驱动关节采用力矩电机,忽略间隙及柔性等非线性因素;(3)空间机器人基座平台姿态需采用稳定控制,不考虑轨道控制.

2.2 自由飞行空间机器人动力学建模

根据空间机器人基座平台的不同,空间机器人主要分为自由飞行空间机器人与自由漂浮空间机器人两种类型.自由飞行空间机器人的基座平台能够进行主动控制,整个系统能够在空间中任意转移与定向;自由漂浮机器人的基座平台是不受控制的,机械臂的运动会导致基座平台完全自由地翻转和移动^[6].本文的研究对象是自由飞行的空间机器人.

如图1所示,空间机器人的可控基座平台与机械臂可以看作一个多体链,由 $n+1$ 个刚体和 n 个关节组成. n

图 1 空间机器人系统图

Figure 1 Space robot system.

个关节从1到 n 分别进行编号, $q = (q_1, q_2, \dots, q_n)^T$ 表示关节转动的角度; $\theta_B = (\theta_x, \theta_y, \theta_z)^T$ 表示基座平台的姿态角; $x_B = (x_{Bx}, x_{By}, x_{Bz})^T$ 表示基座平台的轨道位置.

为研究方便, 将轨道坐标系看作惯性坐标系 $\sum_I$, 基座平台坐标系 $\sum_B$ 固联在空间机器人基座平台上, 其原点位于基座平台的质心, 如图1所示. R_i 与 r_i 是第 i 个连杆质心的位置矢量分别在惯性坐标系 $\sum_I$ 与基座平台坐标系 $\sum_B$ 下的表示, 因此

$$R_i = r_i + R_B, \quad (1)$$

其中, R_B 表示基座平台的质心位置矢量在惯性坐标系 $\sum_I$ 下的表示. V_i 和 Ω_i 分别是第 i 个刚体的线速度与角速度在惯性坐标系 $\sum_I$ 下的表示, v_i 和 ω_i 则分别是其在基座平台坐标系 $\sum_B$ 下的表示, 可以得到

$$V_i = v_i + V_B + \Omega_B \times r_p, \quad (2)$$

$$\Omega_i = \omega_i + \Omega_B, \quad (3)$$

其中的 V_B 与 Ω_B 分别代表基座平台的线速度与角速度在惯性坐标系 $\sum_I$ 下的表示. 同时, 在基座平台坐标系 $\sum_B$ 下的线速度 v_i 与角速度 ω_i 又可表示为

$$v_i = J_{Li}(q)\dot{q}, \quad (4)$$

$$\omega_i = J_{Ai}(q)\dot{q}, \quad (5)$$

其中, J_{Li} 与 J_{Ai} 分别是第 i 个连杆的雅克比矩阵 J_i 的线性部分与角度部分.

在惯性坐标系 $\sum_I$ 下, 整个空间机器人系统的动能主要由基座平台动能与机械臂动能组成, 其中每部分动能又由平动动能与转动动能两部分组成, 可写为

$$E = E_B + \sum_{i=1}^n E_i = \Omega_B^T I_B \Omega_B + V_B^T m_B V_B + \sum_{i=1}^n (\Omega_i^T I_i \Omega_i + V_i^T m_i V_i) = \dot{x}^T M \dot{x}, \quad (6)$$

其中, $M = \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{bmatrix}$ 为一个对称矩阵,

$$x = [x_B, \theta_B, q]^T.$$

$$M_{11} = \sum_{i=1}^n m_i + m_B,$$

$$M_{12} = M_{21} = \sum_{i=1}^n m_i r_i^\times,$$

$$M_{13} = M_{31} = \sum_{i=1}^n m_i J_{Li},$$

$$M_{22} = I_B + \sum_{i=1}^n (I_i + m_i [r_i^\times][r_i^\times]),$$

$$M_{23} = M_{32} = \sum_{i=1}^n (I_i J_{Ai} + m_i [r_i^\times] J_{Li}),$$

$$M_{33} = \sum_{i=1}^n (J_{Ai}^T I_i J_{Ai} + J_{Li}^T m_i J_{Li}).$$

本文主要针对空间机器人机械臂的运动控制与基座平台的姿态稳定控制进行研究, 不考虑基座平台的轨道运动, 另外由于空间机器人在微重力环境下, 忽略空间机器人的重力势能, 采用拉格朗日方法得到空间机器人系统动力学方程:

$$H\ddot{\phi} + C(\phi, \dot{\phi}) = T + \tau, \quad (7)$$

其中, $H = \begin{bmatrix} M_{22} & M_{23} \\ M_{32} & M_{33} \end{bmatrix}$; $\phi = [\theta_B, q]^T$; $C(\phi, \dot{\phi}) =$

$-\frac{\partial}{\partial \phi} \left\{ \frac{1}{2} \dot{\phi}^T H \dot{\phi} \right\}$; $T = [T_B, T_q]$, T_B 为空间机器人的姿态控制力矩, T_q 为机械臂的控制力矩; τ 为干扰力矩.

机械臂末端的位置可以表示为

$$X = \text{Tran}(\theta_B) R(q), \quad (8)$$

其中, $\text{Tran}(\theta_B)$ 是从基座平台坐标系 $\sum_B$ 到惯性坐标系 $\sum_I$ 的坐标转换矩阵, 与基座平台姿态角 θ_B 有关; $R(q)$ 为机械臂末端位置在基座坐标系 $\sum_I$ 下的表示, 与机械臂各个关节处的转角、连杆尺寸大小有关.

3 强化学习算法

强化学习是在个体与环境的相互作用中进行学习, 不断进行尝试, 找到最优的策略, 以获得长期最大

收益的学习算法. 不需要获知准确的动力学模型, 应用简单, 因此得到广泛应用^[21]. 强化学习一般将决策过程视为马尔可夫过程, 即当前时刻采取的动作对过去的状态没有影响, 只对未来的状态有影响. 根据马尔可夫决策过程, 可以得到策略迭代算法: 策略迭代、价值迭代和泛化迭代, 本文使用常用的 Q 学习算法. 一般来说, 强化学习的具体流程如图2所示, 个体Agent在当前状态下, 选择动作 a 执行, 在与环境的相互作用下, 状态 s 改变为新状态 s' , 并且获得在状态 s 采取动作 a 的奖赏值 r , 此时更新动作策略并且选择下一步的动作 a , 不断重复这个训练过程, 动作策略不断更新优化, 最终获得最优的动作策略.

由上述过程可知, 强化学习具有较强的在线自适应特性和对复杂系统的自主学习能力. 空间机器人机械臂的动力学参数具有非线性时变特点, 另外抓捕目标具有非合作特性, 外界环境在不断地发生变化, 共同构成了一个较为复杂的系统. 将强化学习应用到机械臂控制器设计中, 使控制器具有自主学习能力, 主动调整控制策略适应自身动力学结构的变化与外界环境的变化, 在变化中进行自我强化学习, 不断学习最优的控制策略, 使得机械臂运动更具有智能性, 以获得最优的控制效果.

4 强化学习控制器设计

为提高空间机器人抓捕的精度和稳定度, 需要同时控制机械臂的运动以及基座平台姿态的稳定. 因此, 本文提出一种双回路控制系统, 如图3所示. 内回路是基于强化学习的机械臂运动控制系统, 外回路为基座平台的姿态稳定控制系统, 采用PD控制器.

在内回路强化学习控制系统中, 忽略操作手的体积与尺寸, 即认为机械臂的末端位置 $\mathbf{X}$ 为操作手位置. 将机械臂末端位置误差 $\mathbf{E}_c$ 、速度误差 $\dot{\mathbf{E}}_c$ 作为强化学习系统的输入, 空间机械臂控制力矩 $\mathbf{T}_q$ 作为输出, 内部控制系统主要由模糊推理系统与强化学习 Q 值更新两部分组成, 从而对机械臂末端进行运动控制.

$$\mathbf{E}_c = \mathbf{X}_d - \mathbf{X}, \quad (9)$$

$$\dot{\mathbf{E}}_c = \dot{\mathbf{X}}_d - \dot{\mathbf{X}}, \quad (10)$$

其中, $\mathbf{X}_d$ 为机械臂末端期望位置.

图2 强化学习

Figure 2 Reinforcement learning.

图3 控制方案示意图

Figure 3 Control scheme.

4.1 模糊推理系统设计

模糊推理系统设计是强化学习算法与模糊控制结合的关键, 主要是将强化学习中的 Q 学习技术与模糊控制中的模糊规则设计融合在一起. 针对机械臂末端接近目标航天器运动过程控制, 若将机械臂末端的位置误差 $\mathbf{E}_c$ 与速度误差 $\dot{\mathbf{E}}_c$ 作为强化学习算法中的状态量, $\mathbf{E}_c$ 与 $\dot{\mathbf{E}}_c$ 均为实时变化的连续变量, 则产生的状态量将会有无数种, 构成的 Q 表变得异常庞大, 使用强化学习算法将变得不再现实. 为了解决这个问题, 先对连续变化的状态量进行模糊化处理, 将其变为有限的状态量, 再进行 Q 表设计.

采用高斯隶属度函数对连续变量进行模糊化处理,

$$Z_l(y(i)) = e^{\frac{-(y(i)-y^l(i))^2}{\sigma^2(i)}}, \quad (11)$$

其中, l 为输入变量的模糊层数, 这里 $l=1, 2, 3$, 即将输入变量模糊处理为3层; y 为输入状态变量, 即 $y = [\mathbf{E}_c, \dot{\mathbf{E}}_c]$; y^l 为每层模糊层的中心值, $y^l(i) = m(i) + n(i) * (l-1)$; σ 控制着高斯隶属度函数的宽度, 如图4所示.

Q 学习算法中的 Q 表由机械臂离散状态量与机械臂末端的执行动作构成. Q 表的每一行代表机械臂模糊化处理后的不同状态, 由于分别将 $\mathbf{E}_c$ 与 $\dot{\mathbf{E}}_c$ 模糊处理为3层, 则机械臂末端总共构成9种不同状态, 即 Q 表有

图 4 高斯隶属度函数示意图

Figure 4 Gaussian function diagram.

9行; Q 表的每一列代表机械臂末端的执行动作, 将执行动作划分为模糊集 $U\{NB, N, Z, P, PB\}$, 这里的NB, N, Z, P, PB分别代表负大、负、零、正、正大5个模糊级别, 即 Q 表有5列; 这样就构成了一个大小为 9×5 的 Q 表, 表格中的值代表了表格所在行(机械臂状态)采取这个表格所在列(机械臂末端执行动作)的好坏程度。

与传统模糊控制不同, 具体模糊规则的制定并不是人为制定, 而在机械臂末端运动过程中, 不断地学习训练得到。在机械臂一种状态下, 有许多执行动作可以选择, 至于选择哪一个动作则由 Q 表格中的值决定, 即 Q 表中的值代替了模糊规则的制定。因此, 设计的模糊规则的主要形式如下^[22]。

如果 $y(1), y(2)$ 满足第 p 条模糊规则 L_p , 那么 $a=a_1, Q(p, 1)$

或者 $a=a_2, Q(p, 2)$,

.....

或者 $a=a_5, Q(p, 5)$ 。

每一条模糊规则权重 v_p 可由输入变量机械臂末端位置误差 $\mathbf{E}_c$ 、速度误差 $\dot{\mathbf{E}}_c$ 的模糊隶属度相乘得到

$$v_p = Z(y(1)) * Z(y(2)). \quad (12)$$

4.2 强化学习 Q 值更新

通过4.1节分析可知, Q 表中的值决定了模糊规则的具体制定, 为获得最优的控制效果, 就要制定最优的模糊规则, 即获取最优 Q 表。 Q 值的更新优化是通过强化学习不断迭代更新计算得来, 具体训练过程如图5所示。

在每一条的模糊规则中, 选取 Q 表中与模糊规则所对应的一行中, 使得 Q 值最小的执行动作作为最优

图 5 强化学习训练流程

Figure 5 Reinforcement learning training process.

执行动作 a_p^* , 如式(13)所示。通过对每条模糊规则中的最优执行动作去模糊化计算, 得到连续性机械臂最优控制力矩 $\mathbf{T}_q^*$, 如式(14)所示。

$$a_p^* = \arg \min_{a_j \in U} Q(p, a_j), \quad (13)$$

$$\mathbf{T}_q^* = \mathbf{J}^T \frac{\sum_{p=1}^9 v_p a_p^*}{\sum_{p=1}^9 v_p}. \quad (14)$$

在强化学习控制系统训练过程中, 为兼顾探索与利用两方面因素, 每一次的迭代训练计算中并不是都使用最优控制力矩, 而是采用贪心算法选择控制力矩作为实际的输出力矩 $\mathbf{T}_q^r$, 具体步骤如式(15)和(16)所示:

$$a_p^r = \begin{cases} a_p^\varepsilon & \varepsilon \text{ 概率,} \\ a_p^* & 1-\varepsilon \text{ 概率,} \end{cases} \quad (15)$$

$$\mathbf{T}_q^r = \mathbf{J}^T \frac{\sum_{p=1}^9 v_p a_p^r}{\sum_{p=1}^9 v_p}. \quad (16)$$

同理, 通过去模糊计算可得此时的 Q 值:

$$Q(s, \mathbf{T}_q^r) = \frac{\sum_{p=1}^9 v_p Q(p, a_p^r)}{\sum_{p=1}^9 v_p}. \quad (17)$$

在施加机械臂控制力矩 $\mathbf{T}_q^r$ 后, 机械臂末端由状态 s 变为状态 s' , 在状态 s' 时, 使用最优控制力矩 $\mathbf{T}_q^*$ 代替实际控制力矩 $\mathbf{T}_q^r$, 去模糊计算获得此时目标 Q 值为

$$Q(s, \mathbf{T}_q^*) = \frac{\sum_{p=1}^9 v_p Q(p, a_p^*)}{\sum_{p=1}^9 v_p}. \quad (18)$$

机械臂末端位置距离目标位置点越远, 奖赏函数越大, 并考虑机械臂末端的速度与目标航天器之间的速度差, 强化学习目标是使得奖赏函数最小. 因此, 将奖赏函数 $\mathbf{r}$ 定义为

$$\mathbf{r} = 0.5 * (\mathbf{E}_c^2 + \dot{\mathbf{E}}_c^2). \quad (19)$$

因此, 得到 Q 值的更新公式:

$$Q^t(p, a_q^r) = Q^{t-1}(p, a_q^r) + \lambda(\mathbf{r}^{t-1} + \gamma Q^{t-1}(s', \mathbf{T}_q^*) - Q^{t-1}(s', \mathbf{T}_q^r)), \quad (20)$$

其中, t 表示强化学习训练中的时刻; λ 为强化学习算法学习率, 表示强化学习算法学习的快慢; γ 为强化学习算法的折扣系数, 使得长期奖励有意义.

经过强化学习的训练过程, Q 表中的值不断更新收敛, 最后稳定到最优 Q 值, 同时获得最优模糊规则, 从而得到基于强化学习的机械臂运动控制器:

$$\mathbf{T}_q = \mathbf{J}^T \frac{\sum_{p=1}^P v_p a_p^r}{\sum_{p=1}^P v_p}. \quad (21)$$

5 数值仿真

本节进行数值仿真验证本文所提空间机器人控制方案的有效性. 为方便计算, 仿真算例中, 空间机器人的机械臂具有两根连杆, 而且每个关节都只绕 z 轴旋转. 空间机器人系统参数见表1.

在仿真过程中考虑目标航天器的非合作特性, 即在抓捕过程中, 目标航天器会做出逃逸动作规避空间机器人的抓捕. 初始时刻, 目标航天器位于 A 点, 距离空间机器人基座平台质心距离为 $\mathbf{X}_{d_A} = (10, 0, 0) \text{ m}$, 500 s时刻目标航天器为逃避抓捕运动至 B 点, 距离初始时刻空间机器人基座平台质心距离 $\mathbf{X}_{d_B} = (15, 5, 0) \text{ m}$. 机械臂各关节初始角度为 $\mathbf{q} = (90^\circ, 180^\circ)$. 在机械臂末端接近目标航天器运动过程中, 由于摩擦等干扰因素, 各关节的干扰力矩随时间变化为 $\boldsymbol{\tau} = 0.01 \sin(t) \text{ N m}$. 学习率 $\lambda=0.1$, 折扣系数 $\gamma=0.9$, 贪心率 $\varepsilon=0.4-0.01$, 在强化学习训练过程中逐渐减小. 基座平台PD控制参数分别取为 $K_p = -10$, $K_d = -100$. 另外, 采用PD控制方法控制机械臂运动, 进行仿真计算并将其仿真结果与强化学

习(Reinforcement Learning, RL)控制作对比.

图6是机械臂末端的三维运动轨迹. 从图中可以看出, RL控制下的机械臂末端能够准确到达目标位置点, 动态性能较好; 而采用PD控制时, 机械臂也逐渐接近目标, 但在目标位置点附近机械臂末端出现来回摆动的现象. 图7是机械臂末端位置 x , y , z 三个方向随时间变化图. 图7(a)可以看到RL控制下的机械臂末端 x 方向约在180 s时到达目标航天器 A 点, PD控制需要200 s左右, 但最终机械臂末端 x 方向稳定在10.01 m处, 而RL控制下的机械臂末端 x 方向最终稳定在10.001 m附近, 控制精度比PD控制高一个数量级, 同样在500 s时刻开始运动至 B 点的过程中, RL控制比PD控制下的机械臂末端更快接近目标点, 且控制精度较高; 图7(b)可以看到RL控制下的机械臂末端 y 方向分别约在180和600 s左右到达 A 与 B 点, PD控制下的机械臂末端 y 方向分别约在300和800 s左右到达 A 与 B 点, 与 x 方向一样, RL控制比PD控制较快到达目标位置点; 图7(c)可看到在 z 方向上, RL控制与PD控制下的机械臂末端到达目标位置 A ,

表1 空间机器人系统参数

Table 1 Space robot system parameters

	基座平台	连杆1	连杆2
$m \text{ (kg)}$	1000	20	20
$I_{xx} \text{ (kg m}^2\text{)}$	1500	0.133	0.133
$I_{yy} \text{ (kg m}^2\text{)}$	1500	0.133	0.133
$I_{zz} \text{ (kg m}^2\text{)}$	1500	666.733	666.733
长 (m)	3	10	10
宽 (m)	3	0.2	0.2
高 (m)	3	0.2	0.2

图6 机械臂末端三维运动轨迹

Figure 6 Three-dimensional trajectory of robot arm end.

图7 机械臂末端 x , y , z 三方向的运动轨迹. (a) 机械臂末端 x 方向轨迹; (b) 机械臂末端 y 方向轨迹; (c) 机械臂末端 z 方向轨迹

Figure 7 Robot arm end x , y , z three-direction motion trajectory. (a) x -direction trajectory at the end of the arm; (b) y -direction trajectory at the end of the arm; (c) z -direction trajectory at the end of the arm.

B 点的时间差不多, 但可明显发现在到达 B 点时, PD控制下的机械臂末端 z 方向在目标位置上下抖动, 这主要是因为基座平台的姿态抖动造成的.

图8是空间机器人基座平台在 x , y , z 三个方向上的姿态扰动随时间变化图, 可以看到三个方向上都具有一定的姿态扰动, RL控制下的姿态扰动较PD控制姿态扰动幅度小, 由于机械臂的两个连杆分别绕 z 轴旋转, 所以基座平台 z 方向上的姿态扰动更大, 最大扰动角达到 1.3° , 最终也稳定控制在 0.1° 以内, 而 x , y 两个方

图8 基座平台姿态角. (a) 基座平台 x 方向姿态角; (b) 基座平台 y 方向姿态角; (c) 基座平台 z 方向姿态角

Figure 8 Base platform attitude angle. (a) x -direction attitude angle of base platform; (b) y -direction attitude angle of base platform; (c) z -direction attitude angle of base platform.

向上的姿态扰动角均在 0.05° 以内.

图9是空间机器人基座平台的控制力矩随时间变化曲线. RL控制下基座平台姿态与PD控制基座平台姿态所需要的力矩大小基本一致, RL控制力矩略小于PD控制力矩. 在 x , y 两个方向上的姿态控制力矩均在 $0.05 \text{ N}\cdot\text{m}$ 以内, 而 z 方向上的姿态控制力矩较大一些, 这与 z 方向上的姿态扰动较大有关. 图10和11分别是空间机器人两个连杆上的关节控制力矩变化曲线, 考虑到实际情况, 对机械臂关节处的控制力矩进行了限幅处理, 最大控制力矩不超过 $2 \text{ N}\cdot\text{m}$.

图 9 基座平台控制力矩. (a) 基座平台 x 方向控制力矩; (b) 基座平台 y 方向控制力矩; (c) 基座平台 z 方向控制力矩

Figure 9 Base platform control torque. (a) x -direction control torque of base platform; (b) y -direction control torque of base platform; (c) z -direction control torque of base platform.

6 结论

本文针对抓捕非合作目标任务, 首先建立了自由飞行空间机器人耦合动力学模型; 然后提出了一种基

图 10 关节1控制力矩

Figure 10 Link 1 control torque.

图 11 关节2控制力矩

Figure 11 Link 2 control torque.

于强化学习的双回路控制系统, 内外回路分别对机械臂运动与基座平台姿态进行控制; 最后, 给出数值仿真验证控制器的有效性, 得到以下结论: (1) 机械臂运动会会对空间机器人的姿态运动产生干扰, 因此需要在抓捕过程中对基座平台进行稳定控制; (2) 强化学习控制与PD控制仿真结果作对比, 两者的基座平台姿态扰动都在允许范围内, 但强化学习控制下的基座平台姿态扰动更小, 机械臂末端运动过程更平稳, 控制精度更高; (3) 针对抓捕目标的非合作特性, 强化学习控制下的机械臂运动灵活性好, 更具有自主智能性。

参考文献

- Ellery A, Kreisel J, Sommer B. The case for robotic on-orbit servicing of spacecraft: Spacecraft reliability is a myth. *Acta Astron*, 2008, 63: 632–648
- Cui N G, Wang P, Guo J F, et al. A review of on-orbit servicing (in Chinese). *J Astron*, 2007, 28: 805–811 [崔乃刚, 王平, 郭继峰, 等. 空间在轨服务技术发展综述. *宇航学报*, 2007, 28: 805–811]

- 3 Liang B, Xu W F, Li C, et al. The status and prospect of orbital servicing in the geostationary orbit (in Chinese). *J Astron*, 2010, 31: 1–13 [梁斌, 徐文福, 李成, 等. 地球静止轨道在轨服务技术研究现状与发展趋势. 宇航学报, 2010, 31: 1–13]
- 4 Diftler M A, Mehling J S, Abdallah M E, et al. Robonaut 2-the first humanoid robot in space. In: *IEEE International Conference on Robotics and Automation*. Shanghai, 2011. 2178–2183
- 5 Rekleitis I, Martin E, Rouleau G, et al. Autonomous capture of a tumbling satellite. *J Field Robot*, 2006, 24: 275–296
- 6 Flores-Abad A, Ma O, Pham K, et al. A review of space robotics technologies for on-orbit servicing. *Prog Aerospace Sci*, 2014, 68: 1–26
- 7 Longman R W. The kinetics and workspace of a satellite-mounted robot. *J Astron Sci*, 1993, 38: 423–440
- 8 Aslanov V, Kruglov G, Yudin V. Newton-Euler equations of multibody systems with changing structures for space applications. *Acta Astron*, 2011, 68: 2080–2087
- 9 Zhang L J, Chen L. Fuzzy robust H_∞ control and flexible vibration optimal control for free-floating flexible space manipulator (in Chinese). *China Mech Eng*, 2016, 27: 2447–2452 [张丽娇, 陈力. 漂浮基柔性空间机械臂的模糊 H_∞ 鲁棒控制及柔性振动最优控制. 中国机械工程, 2016, 27: 2447–2452]
- 10 Xu Y S, Shum H Y, Kanade T, et al. Parameterization and adaptive control of space robot systems. *IEEE Trans Aerosp Electron Syst*, 1994, 30: 435–451
- 11 Chu Z Y, Sun F C, Cui J. Disturbance observer-based robust control of free-floating space manipulators. *IEEE Syst J*, 2008, 2: 114–119
- 12 Busoniu L, Schutter B D, Babuska R. Decentralized reinforcement learning control of a robotic manipulator. In: *International Conference on Control, Automation, Robotics and Vision*. New York: IEEE, 2007. 1–6
- 13 de Taveira T, Siqueira A, Terra M, et al. Adaptive nonlinear H-infinity controllers applied to a free-floating space manipulator. In: *Proceedings of the 2006 IEEE International Conference on Control Applications*. New York, 2006. 929–934
- 14 Wei C, Zhao Y, Tian H. Grasping control of space robot for capturing floating target (in Chinese). *Acta Aeron Et Astron Sin*, 2010, 31: 622–637 [魏承, 赵阳, 田浩. 空间机器人捕获漂浮目标的抓取控制. 航空学报, 2010, 31: 622–637]
- 15 Yu X Y, Chen L. Globally robust terminal sliding mode control for a free-floating two-flexible-link space manipulator (in Chinese). *Manned Spaceflight*, 2016, 22: 512–519 [于潇雁, 陈力. 漂浮基柔性两杆空间机械臂的全局鲁棒Terminal滑模控制. 载人航天, 2016, 22: 512–519]
- 16 Wang M, Huang P F, Liu Z X, et al. Trajectory planning for minimizing base reaction of free-floating space robot (in Chinese). *J Astron*, 2011, 32: 2152–215 [王明, 黄攀峰, 刘正雄, 等. 自由漂浮空间机器人最小基座反作用轨迹规划. 宇航学报, 2011, 32: 2152–2157]
- 17 Oda M, Ohkami Y. Coordinated control of spacecraft attitude and space manipulators. *Control Eng Prac*, 1997, 5: 11–21
- 18 Xu W F, Meng D S, Xu C, et al. Coordinated control of a free-floating space robot for capturing a target (in Chinese). *Robot*, 2013, 35: 559–567 [徐文福, 孟得山, 徐超. 自由漂浮空间机器人捕获目标的协调控制. 机器人, 2013, 35: 559–567]
- 19 Jayakody H S, Shi L, Katupitiya J, et al. Robust adaptive coordination controller for a spacecraft equipped with a robotic manipulator. *J Guid Control Dyn*, 2016, 39: 2699–2711
- 20 Hu Q, Guo C, Zhang Y, et al. Recursive decentralized control for robotic manipulators. *Aerospace Sci Tech*, 2018, 76: 374–385
- 21 Montague P R. Reinforcement learning: An introduction, by Sutton R S and Barto A G. *Trends Cogn Sci*, 1999, 3: 360
- 22 Kumar A, Sharma R. Linguistic Lyapunov reinforcement learning control for robotic manipulators. *Neurocomputing*, 2018, 272: 84–95

Autonomous reinforcement learning control for space robot to capture non-cooperative targets

LIU Shuai¹, WU ShuNan^{1*}, LIU YuFei², WU ZhiGang¹ & MAO ZiMing³

¹ School of Aeronautics and Astronautics, Dalian University of Technology, Dalian 116024, China;

² Tsien Hsueshen Laboratory of Space Technology, China Academy of Space Technology, Beijing 100094, China;

³ Department of Engineering Mechanics, Dalian University of Technology, Dalian 116024, China

In recent years, space robot on-orbit service has become a research hotspot in many countries. Aiming at the task of capturing non-cooperative targets for space robot, a dual loop control method consisting of reinforcement learning control and PD control is proposed in this paper, which is used to control the attitude of space robot platform and the motion of manipulator arm. Firstly, a coupled dynamic model of the space robot including the motion of the base platform and the robot arm is established. Then, a dual loop control system is designed to control the movement of the robot arm and the attitude of the base platform. In the inner loop, the controller is designed by combining the reinforcement learning and fuzzy theory to control the motion of the end of the robot arm. In the outer loop, the attitude of the base platform is stabilized by PD control. Finally, the proposed control method is used for numerical simulation and compared with the traditional PD control method to verify the effectiveness of the proposed control method. The results show that the robot arm movement process under the control of reinforcement learning is stable and the control precision is high. Compared with the traditional PD control method, it has a certain self-learning ability and is more suitable for the non-cooperative characteristics of the catching targets.

space robot, non-cooperative target, reinforcement learning, robot arm, fuzzy theory

PACS: 47.27.-i, 47.27.Eq, 47.27.Nz, 47.40.Ki, 47.85.Gj

doi: [10.1360/SSPMA2018-00287](https://doi.org/10.1360/SSPMA2018-00287)