计算机测量与控制
Computer Measurement & Control
ISSN 1671-4598,CN 11-4762/TP

《计算机测量与控制》网络首发论文

题目：基于 D-DQN 强化学习算法的双足机器人智能控制研究
作者：李丽霞，陈艳
网络首发日期：2023-10-13
引用格式：李丽霞，陈艳．基于 D-DQN 强化学习算法的双足机器人智能控制研究[J/OL]．计算机测量与控制.
https://link.cnki.net/urlid/11.4762.TP.20231012.0946.014

网络首发时间：2023-10-13 15:50:11
网络首发地址：https://link.cnki.net/urlid/11.4762.TP.20231012.0946.014

基于 D-DQN 强化学习算法的双足机器人智能控制研究

李丽霞，陈艳
（广州华商学院，广东 广州 511300）

摘 要：针对现有双足机器人智能控制算法存在的轨迹偏差大、效率低等问题，提出了一种基于 D-DQN 强化学习的控制算法。先分析双足机器人运动中的坐标变换关系和关节连杆补偿过程，然后基于 Q 值网络实现对复杂运动非线性过程降维处理，采用了 Q 值网络权值和辅助权值的双网络权值设计方式，进一步强化 DQN 网络性能，并以 Tanh 函数作为神经网络的激活函数，提升 DQN 网络的数值训练能力。在数据训练和交互中经验回放池发挥出关键的辅助作用，通过将奖励值输入到目标函数中，进一步提升对双足机器人的控制精度，最后通过虚拟约束控制的方式提高双足机器人运动中的稳定性。实验结果显示：在 D-DQN 强化学习的控制算法，机器人完成第一阶段测试的时间仅为 115s，综合轨迹偏差 0.02m，而且步态切换极限环测试的稳定性良好。

关键词： D-DQN；强化学习；双足机器人；智能控制；经验回放池；虚拟约束控制

文献标识码： A

Research on Intelligent Control of Biped Robot Based on D-DQN Reinforcement Learning Algorithm

LI Lixia，CHEN Yan
（Guangzhou Huashang College, Guangzhou 511300, China）

Abstract: Aiming at the problems of large trajectory deviation and low efficiency of existing intelligent control algorithms for biped robots, a control algorithm based on D-DQN reinforcement learning is proposed. Firstly, the coordinate transformation relationship in the motion of biped robot and the compensation process of joint and link are analyzed, and then the dimensionality reduction of complex nonlinear motion process is realized based on Q-value network. The double weight design method of Q-value network weight and auxiliary weight is adopted to strengthen the performance of DQN network, and Tanh function is used as the activation function of neural network to improve the numerical training ability of DQN network. The experience playback pool plays a key auxiliary role in data training and interaction. By inputting the reward value into the objective function, the control accuracy of the biped robot is further improved. Finally, the stability of the biped robot is improved by virtual constraint control. The experimental results show that under the D-DQN reinforcement learning control algorithm, the time of the robot to complete the first stage test is only 115s, the comprehensive trajectory deviation is 0.02m, and the stability of the gait switching limit cycle test is good.

Key words: D-DQN; Reinforcement learning; Bipedal robot; Intelligent control; Experience playback pool; Virtual constraint control

1 引 言：

在计算机科学技术、自动化控制技术和人工智能技术的共同推动下，机器人智能化水平不断提高[1]。机器人的设计与控制极其复杂，融合了机械、电子、传感、软件控制、无线通信等多项技术[2-3]，代表了一个国家的科研实力和精密制造水平。在诸多机器人结构设计中，双足机器人的设计难度最大，对机械结构合理性及运动中自由度的分配提出了更高要求。鉴于双足机器人结构设计优势[4]，其具有更高的灵活性能够跨越和躲避障碍；由于双足机器人的运动自由度较高，因此可以做出类似于提拉、抓取等相对复杂的动作；与传统轮式机器人、爬行式机器人相比，双足机器人运动效率较高，而且具有更低的功耗[5]。双足机器人在控制中要先确保运动的平稳性，在此基础上提升机器人的自由度和运动效率，并纠正机器人的前进动作和就这和运动轨迹偏差。

在双足机器人运动控制方面，国内外领域内的专家和学者进行深入且广泛的研究。学者 Wang 提出一种基于实时混合控制的方案，机器人的底层采用了模块化的结构设计，而机器人顶层则采用了齐次坐标变换的控制及实时通信的方式，既解决了控制过程中数据冗余的问题，也能够确保机器人沿着设定轨迹前进[6]。但该种方法下机器人的控制效率过低且控制精度和灵活度均不足，导致该种控制方案无法得到广泛的推广和使用；国内学者张品等人提出一种基于模糊 PID 控制的方案，结合模糊控制宽容度高和 PID 控制简洁、精准性高等优点，来提升对机器人的控制效果[7]。但由于双足机器人对于多自由度变化的控制精度要求较高，模糊 PID 控制在上下台阶或跨越障碍物等复杂动作的控制精度较差。

由于双足机器人的复杂程度较高，为提升对机器人的控制精度和灵活度，本文引入了机器学习算法中的一种增强学习算法：D-DQN（DOUBLE-Deep Q Networks），

D-DQN 算法的优势在于可以更好地利用实时回馈的数据，对机器人连续发送动作指令，并实施纠正运动偏差，并利用 Q 值函数在较短的时间内获得最优的控制回报。深度 Q 值网络中包含了若干个卷积层和全连接层，通过实时映射和非线性变换获取到双足机器人运动中的特征参数，并将结果反馈到控制中心。而 D-DQN 算法中包含了两个控制权值，用其中的一个参数控制机器人的姿态动作，而使用另一个参数去估计函数模型中的 Q 值。D-DQN 优化算法能够更好地解决强化学习过程中的激活函数参数的估计问题，进而提升了改善对双足机器人的实时控制。

2.DQN 强化学习算法及优化

2.1 双足机器人结构设计

控制双足机器人的关键点是要保证机器人稳定行走，因此在设计机器人时通常采用连杆结构，在机器人运动中通过步态控制和自由度控制，实现机器人的协调行进。本文设计的四连杆双足机器人运动模型，如图 1 所示：

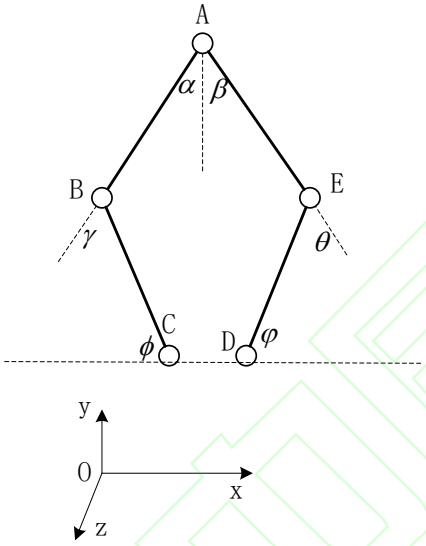

图 1 四连杆双足机器人的结构设计

四连杆双足机器人的核心运动关节为 A、B、C、D、E 五个关节，其中关节 A 模拟人体的髋关节，B、E 模拟人体的膝关节，C、D 模拟人体的踝关节；连杆 AB 和 AE 是机器人的大腿，连杆 BC 和 ED 是机器人的小腿，双足机器人动力系统包括电池、电机、舵机等。

2.2 机器人运动过程分析

双足机器人稳定行走关键要控制好重心[8-9]，可以将关节 A 大致视为机器人的重心，在机器人双腿交替前行时，重心落在关节 A 以确保机器人的稳定性和行进效率。根据齐次坐标转换[10]，先得出关键 C 的坐标公式和关节 D 的坐标公式：

$$\begin{cases} x_C = 0 \\ z_C = d_1 \end{cases}$$
$$\begin{cases} x_D = x_C + d_2 \sin\gamma \\ z_D = z_C + d_1 \cos\gamma \end{cases} \quad (1)$$

其中 d_1 是连杆 BC 的长度，d_2 是连杆 BA 的长度，同理得出 A、B、E 各点的坐标：

$$\begin{cases} x_A = x_C + d_2 \sin\gamma + d_1 \sin\alpha \\ z_A = z_C + d_1 \cos\gamma \cos\alpha \end{cases}$$
$$\begin{cases} x_B = x_A + d_2 \sin\alpha + d_3 \sin\alpha - d_3 \cos\beta \\ z_B = z_A + d_1 \cos\alpha + d_2 \sin\alpha + d_3 \sin\beta \end{cases} \quad (2)$$
$$\begin{cases} x_E = x_B + d_1 \sin\gamma + d_1 \sin\alpha - d_3 \cos\beta + d_4 \cos\theta \\ z_E = z_B + d_1 \cos\gamma + d_1 \sin\alpha - d_3 \sin\beta + d_4 \cos\theta \end{cases}$$

其中，d_3 和 d_4 分别为连杆 AE 和 DE 的长度，通过各关节坐标定位能够确定移动过程中机器人的质心及运动速度[11]，进而实现对双足机器人移动行走过程中的稳定性控制。在对机器人的控制中，规定绕 y 轴顺时针旋转为正向，逆时针旋转为逆向。以双足机器人右侧关节 C、B 为例进行描述和说明，当关节 C 沿 x 轴方向旋转 γ 角度时，其向 y 轴旋转的角度为 $-\alpha$，通过齐次坐标变换，得到了关节 C 相对于坐标系 O 的齐次坐标变换矩阵：

$$\begin{bmatrix} \cos\gamma & 0 & \sin\gamma & 0 \\ -\sin\gamma\cos\alpha & \cos\alpha & \sin\alpha\cos\gamma & 0 \\ \sin\gamma\cos\gamma & -\sin\gamma & -\sin\alpha\cos\gamma & d_1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (3)$$

当右侧膝关节 B 的旋转角度为 γ 时，相对于坐标系 O 的变化过程如下：

$$\begin{bmatrix} \cos\alpha & 0 & \sin\alpha & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\alpha & 0 & \cos\gamma & d_2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (4)$$

采用相同的坐标转换方法能够识别双足机器人左侧关节 D 和关键 E 的坐标变化情况。双足机器人采用双腿交替前行的方式运动，通过对机器人运动模型的分析及对运动过程的实时控制，在确保机器人平稳运行的前提下，实现机器人合理避障、轨迹纠偏及运行效率的最大化。

2.3 深度 Q 值网络

双足机器人运动过程中涉及到多个连杆、关节的同步协同运动，机器人质心变化、速度变化等都会影响到机器人的运动平稳性。而双足机器人的运动控制是一个极其复杂非线性控制过程[11-12]，通过大量实时数据交互形成完整的数据通信网络，持续对双足机器人发送实时指令，并提升对机器人运动中坐标变换的控制精度，进而实现对机器人步态和运动轨迹控制的精度。对双足机器人的步态稳定性控制和运动效率控制问题，最终转换为一个复杂的神经网络数值训练问题[13]，与传统深度学习[14]、机器学习[15]及 Q 学习算法[16]相比，深度 Q 值网络在损失函数设计和经验回放机制等方面做了同步优

化。例如，Q 学习算法在实际的数据训练中利用 Q 表格对高维空间内的数据做降维处理[17]，数据训练的工作量巨大导致训练时间过长及训练效率低下，反应在对双足机器人控制方面，会导致控制精度变差和控制效率降低。而深度 Q 值网络算法基于值函数解决高维数据的降维问题，将实时采集到机器人运动轨迹数据，姿态数据，步频步幅等数据作为训练样本输入到深度 Q 值网络中，判定机器人的实际动作与理论动作之间的差距，并通过调整损失函数值的方式[18]，控制机器人的动作指令及调整和缩小实际值与理论值之间的差距，Q 值更新的过程描述如下：

$$Q_t(s,a)=Q_{t-1}(s_{t-1},a_{t-1})+\alpha\left[r+\max Q_{t-1}(s_{t-1},a_{t-1})\right] \quad (5)$$

其中，$Q_t(s,a)$ 表示当前 t 时刻的 Q 值，s,a 分别表示当前 t 时刻机器人的状态和动作，而 $Q_{t-1}(s_{t-1},a_{t-1})$ 分别表示 t-1 时刻的 Q 值状态和动作，α 为循环系数取值范围在 0-2 之间，r 为 t-1 时刻 Q 值神经网络训练的奖励值，r 会根据模型的不断迭代而累加。基于贝尔曼方程来表示当前 t 时刻双足机器人的 Q 值，y 为模型的输出项：

$$y=r+\alpha\max Q_t(s,a,\varsigma) \quad (6)$$

其中，ς 为 Q 学习过程中 Q 值对应的权值，模型输出项的目标函数，即为对应权值的损失函数 $L(\varsigma)$：

$$L(\varsigma)=E\left[\left(y-Q_t(s,a,\varsigma)\right)^2\right] \quad (7)$$

由公式（7）可知，损失函数本质上是一种数学期望，损失函数的值越低表明对双足机器人的实际控制偏差越小。

2.4 Q 值的预估

通过网络迭代，不断地优化 Q 值神经网络的参数，对应的权值增量 $\Delta\varsigma$ 表示如下：

$$\Delta\varsigma=E\left[y-Q_t(s_t,a_t,\varsigma_t)\right]\nabla Q_t(s_{t-1},a_{t-1},\varsigma_{t-1}) \quad (8)$$

在针对双足机器人运动姿态和动作数值反馈的过程中，及目标网络降维中，采用第 t 期的参数表示当前网络参数，而主干网络的预测中采用第 t-1 期参数表示上一期的网络参数。在参数调整和数值训练中目标 Q 值保持不变，从而提升了算法的稳定性。

深度学习和强化学习中样本之间具有较强的关联性，会对样本属性判断和特征提取构成不利影响[19]，甚至会到数值训练收敛速度降低和出现函数的损失值波动等问题。深度 Q 值网络算法的另一个优势是设置了数据存储的缓冲区，并形成了一种经验回收机制，有助于在训练中提取并应用成熟准确的数据样本。从双足机器人数值传感器反馈回的数据具有高维属性和海量性特征，设置数据缓冲区一方面，能够鉴别出反馈数据的完整性和真实性，另一方面利用缓冲区的先验知识也能够提升学习的效率。双足机器人在运动中其步频、步幅和步态具有较强的重复性，通过信息存储和筛选大量关于机器人状态描述的样本被存储于经验池中；在数据的重复训练和特征提取过程中，相同的有价值的数据会被提取出并重复利用。经验回放模式相当于保留了机器人重复动作指令，也打破了传统模式下动作，状态及指令之间一一对应的情况。一方面解决了重复指令数据量过大的问题，另一方面在下一个动作选择时，也能够优先地筛选成熟和准确的指令方案。深度 Q 值网络从损失函数优化和数据缓冲存储区设置等两个视角，改善了传统深度学习算法存在的问题，解决了传统深度学习算法和 Q 学习算法网络收敛慢，参数集选择复杂和数值训练中所面临的维数灾难问题[20]。但深度 Q 值网络在处理海量数据时容易产生高估 Q 值的问题，一方面影响数据训练的迭代效率和收敛效率，另一方面 Q 值预估过高还将影响到对双足机器人运动控制精度。

2.5 D-DQN 网络模型的构建

为进一步提升深度 Q 值网络数据训练性能，同时解决 Q 值预估过高的问题，本文设计了一种 D-DQN 强化学习算法，并采用双 Q 值权值的网络模型结构设计（ς 和 ς'），其中权值 ς 的功能不变，仍旧用于控制机器人的指令动作，而辅助权值 ς' 则用于控制对 Q 值的预估，用模型的输出项作为样本的目标 Q 值，具体表示如下：

$$y=r+\alpha\max Q_t(s,a,\varsigma,\varsigma') \quad (9)$$

基于 D-DQN 强化学习算法训练数据集时，先通卷积神经网络构建最优的状态值函数，经过模型的强化学习，大多数情况下估计出的 Q 值均大于零。在激活函数的选择方面，本文选择了 Tanh 函数，该函数是一种双切线函数，取值范围在-1,1 之间。D-DQN 强化学习算法由两个部分组成，即训练部分和交互学习部分，具体如图 2 所示：

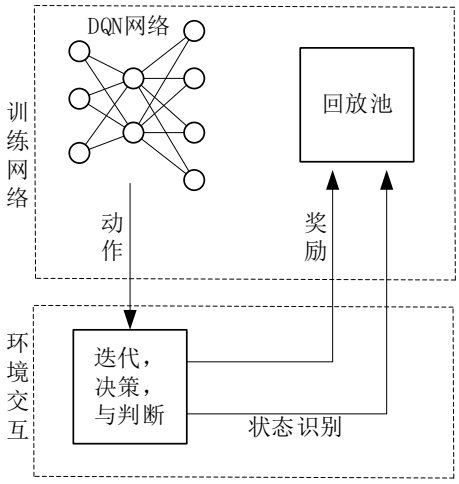

图 2 D-DQN 强化学习算法的构成

算法包含了两个部分，即训练网络和环境交互部分，在训练中由于采用了双 Q 值权值的结构设计，提升对 Q 值预估的准确率而且优化了模型的目标函数，通过反向传播的方式来纠偏的指令动作，并将其传递给双足机器人的控制中心；在环境交互部分根据输入的机器人状态迭代分析结果，给出动作指令值和奖励值，协作优化模型参数并提升对机器人的控制精度及控制效率。双足机器当前位置信息和动作信息，作为 D-DQN 算法模型输入项，经过多轮的迭代和训练，对机器人步态动作指令的奖励值增加，机器人会把训练时被证明正确的元数据组回放到经验池，经验池中的有效指令数据逐渐增加，对机器的控制精度也能够得到同步改善。而双足机器人下一步的动作指令，会根据奖励值和经验池的真实指令做出，因此还要奖励值输入到目标函数中做进一步优化，通过反复训练得到现有参数体系下最优 Q 值。由于辅助权值ς'的存在，在 Q 值寻优中避免了最高峰值的出现。

2.6 奖惩函数的优化和改进

机器人动作的奖励值与动作的控制精度相关，通常情况下动作的精度越高，回馈后得到的奖励值就越高而且回放池中的数值精度也在同步提升，从理论上来看回放池呈现出一种自组织优化的发展趋势。但根据对双足机器人动作状态而做出奖励的方式，容易出现奖励值稀疏的问题，这是因为动作的奖励值不是孤立存在的而是不断累加的，而奖励值设定的目标也不相同。为解决奖励值的稀疏问题，最有效的方法是对 D-DQN 算法模型的奖惩函数进行优化。根据机器人动作精度分配奖励值，动作指令反馈到神经网络中如果判定动作正确则分配正奖励值，而且奖励值的多少与机器人动作的精度正相关；如果动作错误则赋予负的奖励值，奖惩值的计算过程如下：

$$r = v \cdot \cos(\omega) \cdot n \quad (10)$$

其中v代表对应的机器人运动位移线速度，ω代表关节的旋转角速度，n代表重复运动的次数。从公式（10）中可知，双足机器人的线速度和角速度都会影响到奖励值，经过多次迭代和训练后，根据奖励值的多少系统会给机器人设定一个新的路径，但无法避免地会出现奖惩稀疏的情况。考虑到经验回放池在机器人动作指令寻优中的作用，在给机器人发送指令时参考经验回放池的已有数据。机器人运动位移线速度和关节的旋转角速度时决定奖惩回报值的关键参量，为加速机器运动中的自主寻优能力，通过提升线速度和角速度倍数的方式优化奖惩函数，优化后的奖惩值计算过程如下：

$$r = \rho v^2 \cdot \cos(2\omega) \cdot n \quad (11)$$

其中，ρ为奖惩参数，所采取的优化方式是线速度乘方及角速度翻倍，以提升双足机器人行进中躲避障碍和自主选择最优路径的能力

3. 双足机器人运动控制模型构建与智能控制

3.1 强化学习及信任推理

在双足机器人的 Q 值计算中，$Q_t(s,a)$为t时刻的 Q 值，s和a分别为t时刻的机器人的动作和状态，假设双足机器人发出了m个动作，且具有n种状态，则动作集合S和状态集合A表示如下：

$$\begin{cases} S = \{s_1, s_2, ..., s_i, ..., s_m\} \\ A = \{a_1, a_2, ..., a_j, ..., a_n\} \end{cases} \quad (12)$$

前后两个动作之间的连贯性取决于后一个动作对前一个动作的信任值，在第j个状态下机器人对第i个动作（下一个预期动作）的信任值表示为τ_{ij}，信任值与动作的预测精度高度相关，也与经验回放池中的经验数据相关。双足机器人做出下一个指令动作，要通过神经网络的输入层、中间隐含层到网络体系中，并依赖于激活函数映射到输出层，最后得到 Q 值。深度 Q 值神经网络的结构相同，层数可以更加指令集合的复杂程度来确定，激活函数选择 ReLU 函数，各层的网络参数随着迭代次数的增加而不断调整。双足机器人只有具备了最优的状态，才能根据信任度推理而选择偏差最小的下一个指令动作，推理的过程如下：

STEP1：根据机器人前一个动作的奖励值和信任值确定下一个动作的指定及机器人的整个行进路径，同时考虑到经验回放池中已经数据，以提升机器人运动中的效率。

STEP2：模拟机器人前后动作之间的交互关系，以提升对前一个动作指令的信任度，并通过推理的模式动态更新 Q 值网络的参数。

STEP3：由于信任度会出现衰减，因此要考虑到经验池中已存储的路径，并通过不断地迭代学习识别出新的信任路径。

STEP4：最后还要考虑到采样时间的复杂度和测试

样本的规模，实时评估预估值与信任值之间的差距，和算法的有效性。

3.2 基于优化 D-DQN 的双足机器人步态稳定控制

基于 D-DQN 强化学习算法能够确保机器人向前行进的每一步都得到有效控制，同时在行进的过程中根据 Q 值和信任度，来实时调整行进动作出现的偏差和路径轨迹的偏差。由于双足机器人的运动过程是一个极其复杂的非线性过程，本文在 D-DQN 强化学习模型的基础上，采用了虚拟约束控制的方式，将双足机器人多自由度控制融合成为一种完整的约束关系，以机器人的侧视步态为例进行分析，如图 3 所示：

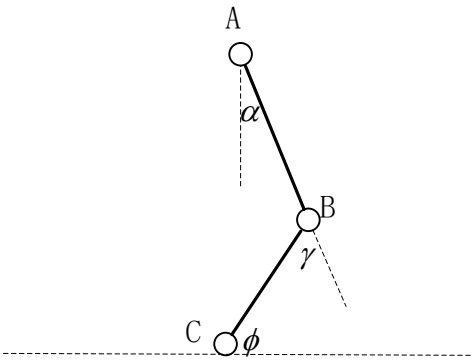

图 3 机器人虚拟控制约束示意图

机器人连杆 AB 的长度为l_1和 BC 的长度为l_2，机器人运动过程中的约束表达式，具体表示如下：

$$\begin{cases} l_1\cos\phi + l_2\cos(\phi+\gamma) = 0 \\ 2\phi + \gamma = \pi\sqrt{2\alpha} \end{cases} \quad (13)$$

将各关节都加上驱动，基于 D-DQN 强化学习算法得到的模型输出项表示如下：

$$\begin{cases} \gamma = \pi - \phi + \arccos\left(\dfrac{l_1}{l_2}\cos\phi\right) \\ \alpha = \arcsin\left(\dfrac{l_1}{l_2}\cos\phi\right) \\ y = \sin\gamma - \arcsin\left(\dfrac{l_1}{l_2}\cos\phi\right) \end{cases} \quad (14)$$

从双足机器人侧视的角度观察，虚拟控制的过程将对机器人三个关节控制过程联合起来，以实现对输出量的等价控制，同步控制过程是一个虚拟的控制过程，各关节具有同步的控制特征，实施向后台反馈控制结果并添加到经验池。图 3 中的融合控制量包括了 5 个独立约束量（3 个关节和 5 个连杆，可以视为 5 维），独立约束量越多对双足机器人的非线性控制过程就越复杂，为实现虚拟控制过程精度的提升，要基于 D-DQN 强化学习算法地独立约束向量进行降维处理。包括 5 个独立约束量的线性映射过程表示H_5，从 5 维到 4 维的降维过程如下：

$$H_5(\phi,\gamma,\varphi,l_1,l_2) = \phi H_4(\gamma,\varphi,l_1,l_2)\cdot c(q_n) \quad (15)$$

其中，$c(\ )$为一个从高维降到低维的线性函数，q_n表示双足机器人的自由度，具体独立约束量降维的过程表示如下：

$$H_4 = [q_5 \quad q_4 \quad q_3 \quad q_2] \quad (16)$$

在一个独立的机器人行走周期内，$c(q_n)$始终为一个单调递增函数，再通过设定和调整两个无量纲参数κ_1、κ_2和最大提髋距离τ，来提升双足机器人运行的稳定性：

$$\begin{cases} \kappa_1 = \dfrac{l_1+l_2}{D} \\ \kappa_2 = \dfrac{l_1}{l_1+l_2} \end{cases} \quad (17)$$

其中，D为双足机器人的步长，最大提髋距离τ表示如下：

$$\tau = \sqrt{(D-l_1)^2 - l_2^2/4} \quad (18)$$

机器人稳定性的控制过程是一个极为复杂的非线性过程，对双足机器人的智能控制过程步骤如下：

STEP1：确定双足机器人运动过程中的空间坐标体系，明确其运动过程和空间坐标转换关系。

STEP2：实时传递机器人的空间坐标位置信息，并通过 D-DQN 强化学习算法训练传递回的数据，并将有效数据存储于经验池。

STEP3：基于奖惩函数确定双足机器人运动过程中的奖惩值和信任值，以实现对机器人运动动作和轨迹的纠偏。

STEP4：通过 D-DQN 强化学习算法的信任推理过程，一方面明确算法执行于理论动作和轨迹的差距，另一方面评估算法的有效性，

STEP5：采用 D-DQN 强化学习算法和虚拟约束控制的方式，实现对负责非线性运动过程的降维，并通过虚拟约束确保机器人的平稳高效运行。

4.实验结果与分析

4.1 实验环境设置

本文在实验室环境下来验证 D-DQN 强化学习算法对双足机器人智能控制效果，选用的实验用机器人为 Agility Robotics 公司生产的 Cassie 型号双足机器人，具体如下图 4 所示：

图 4 实验用 Cassie 型双足机器人

Cassie 型双足机器人的大腿、小腿长度和质量分别为 0.40m、0.35m、7.5kg 和 3.0kg，双足机器人的其他参数设置如表 1 所示：

表 1 双足机器人相关参数设置

序号	参数	取值
1	ϕ 角度区间	45-135°
2	α 角度区间	0-180°
3	γ 角度区间	45--180°
4	固定步长	0.5m
5	控制周期	0.1s
6	调节参数值范围	0.1-0.9
7	最大驱动力	200Nm

对双足机器人智能控制实验中的软硬件环境设置包括：CPU Intel corei9 9900h，CPUDE 最高主频 3.1GHz，RAM16GB，ROM1TB，操作系统选择兼容性更好的开源系统 LINUX。本文使用的机器人控制软件为 RobotArt，该软件支持多格式的机器人控制模型，包括三维 CAD 模型，能够根据现场的情况主动生成程序而且支持复合外部轴系统，以更好地实现机器人的轨迹控制和避碰。

实验过程中采用了两种实验场景：第一种是 30m 的平整路面，机器人的行进路线上设置了 10 个路障，通过该场景主要分析双足机器人，在不同算法控制下步态稳定性和躲避障碍物的智能化水平，及运动效率。

第二种是台阶环境（50 级台阶），主要验证机器人在不同算法控制下能否完成台阶的攀爬，及完成任务后的耗时情况。为了使实验结果更加直观，引入了文献 6 的混合控制方案和文献 7 的模糊 PID 控制方案参与对比。实验场景设置按照各算法的需求参数进行匹配，双足机器人的执行流程如图 4 所示：

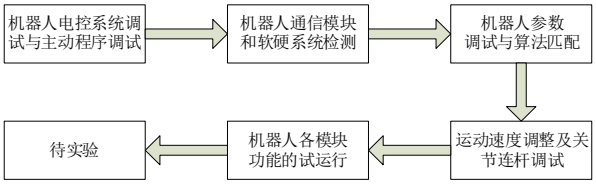

图 5 机器人的执行流程

4.2 实验数据分析

（1）平整路面实验

双足机器人平整路面测试，主要测试各算法对于机器人运动过程中的稳态控制，通过传感器采集机器人 6 个关节和 4 个连杆运动中的角度和位移偏差均值，同时分析完成 30m 测试后各算法躲避障碍物的情况，及最终的完成时间，关节和连杆的偏差均值分析情况，如表 2 所示（采集实际回馈的关节数据及连杆数据，与理论情况下机器人的控制数据进行对比）：

表 2 双足机器人平整路面行进测试关节与连杆轨迹偏差

关节/连杆偏差均值	控制算法控制偏差			
	D-DQN	传统 DQN	混合控制	模糊 PID
髋 1	0.05°	0.11°	0.13°	0.16°
髋 2	0.00°	0.20°	0.25°	0.17°
膝 1	0.02°	0.17°	0.18°	0.10°
膝 2	0.01°	0.16°	0.15°	0.14°
踝 1	0.04°	0.18°	0.20°	0.21°
踝 2	0.00°	0.16°	0.19°	0.19°
大腿 1	0.00m	0.20m	0.25m	0.30m
大腿 2	0.01m	0.12m	0.18m	0.31m
小腿 1	0.00m	0.17m	0.21m	0.25m
小腿 2	0.02m	0.18m	0.26m	0.32m

由于传感器采集的机器人运动数据可知，D-DQN 算法控制下双足机器人各关节的角度偏差和连杆的位移偏差值均较小，且趋近于理论值。完成平路测试后轨迹偏差、总耗时、及碰撞障碍物的相关统计数据，如表 3 所示

表 3 平路测试环境数据对比

项目	控制算法控制偏差			
	D-DQN	传统 DQN	混合控制	模糊 PID
轨迹总偏差	0.02m	0.17m	0.31m	0.28m
完成时间	115s	135	145s	152s
碰撞障碍物	0 次	1	1 次	2 次

两种传统算法下双足机器人在每个行进中各关节和连杆都会产生偏差，且无法实现动作及轨迹上的纠偏，进而导致机器人总体轨迹偏差较大，完成 30m 运动轨迹的时间 D-DQN 算法耗时最短为 115s，且没有碰撞到设置的障碍物，而三种传统算法分别碰撞了 1 次障碍物，1 次障碍物和 2 次障碍物。不同算法下双足机器人运动轨迹偏差、控制时间和避碰情况表明，D-DQN 算法的智能控制效果优于三种传统算法。

（2）攀爬实验

具有良好的攀爬能力是双足机器人相对于其他类型机器人的优势所在，在一些特殊场景下，双足机器人垂直运动灵活且能够完成更多的指定任务。先分析

D-DQN 算法，随着机器人不断攀爬台阶，其奖励值的变化情况，如图 5 所示

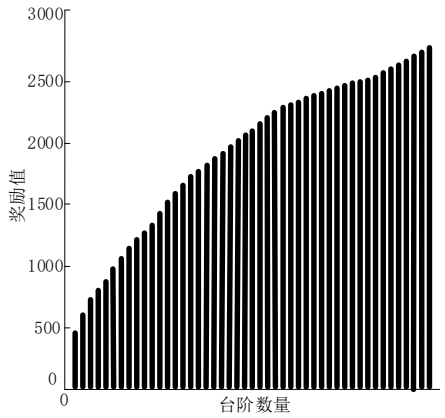

图 6D-DQN 算法下双足机器人运动中奖励值变化

D-DQN 算法下初始阶段奖励值快速增长，随着双足机器人平稳登台阶，算法的奖励值增速下降，但整个过程中算法的奖励值始终呈现出增长的趋势，这表明 D-DQN 算法在不断地纠正机器关节旋转角度偏差和位移偏差，以达到更平稳地控制机器人行走的目的。

双足机器人运动过程中的平稳性可以通过步态切换极限环来表示，机器人在攀爬 50 个台阶过程中重心三个轴向会不断地发生偏差，而智能机器人主动控制算法就是通过实时指令实现三个轴向的纠偏，用双足机器人步态切换极限环，来模拟对双足机器人的重心纠偏过程，D-DQN 算法下的步态切换极限环变化如下：

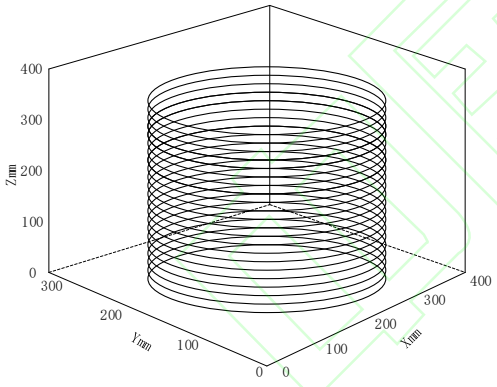

图 7 D-DQN 强化学习算法下步态切换极限环

图 7 中，相邻近的步态切换极限环并未出现较大的偏差，表明在 D-DQN 强化学习算法下双足机器人的重心控制良好；在相同的实验环境下，分析传统 DQN 算法、混合控制算法和模糊 PID 控制算法下极限环的变化情况，分别如图 8-图 10 所示：

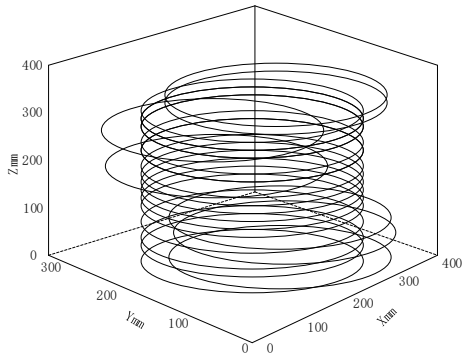

图 8 传统 DQN 算法下的步态切换极限环

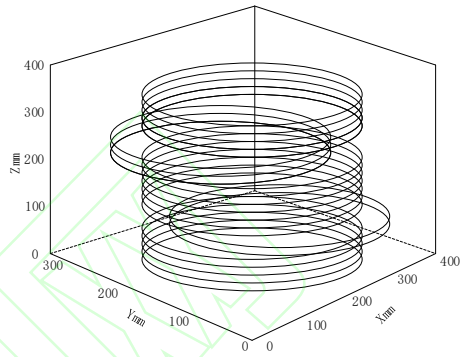

图 9 混合控制算法下的步态切换极限环

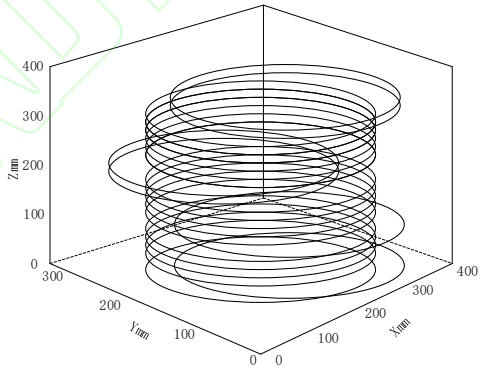

图 10 模糊 PID 算法下的步态切换极限环

如图 8-图 10 所示，在三种传统算法下双足机器人的步态切换极限环，均出现了较为明显的偏差，表明三种传统算法对双足机器人步态和重心控制存在一定问题，会影响到双足机器人运动轨迹和行进效率。

5.结束语

与传统轮式机器人、多足机器人等相比，双足机器人在平衡性、稳定性、仿真效果、垂直运动能力等方面均具有较大的优势。但双足机器人由于特殊用途在结构设计上更加复杂，会包括多个关节和连杆，整个运动的过程也是一个复杂度较高的非线性过程，因此对双足机器人控制算法的控制精度要求较高。本文设计了一种基于 D-DQN 强化学习的算法，融合 Q 学习、Q 值网络模型、神经网络和强化学习等多种算法的优势，采用双权值的设计方式，并通过实时的补偿控制纠正各关节和连杆复合运动中出现的偏差，在确保机器人运动效率的同时，准确控制机器人的运动轨迹和复杂动作。实验结果显示，在 D-DQN 强化学习的算法控制下，双足机器人

的轨迹偏差较小，且在完成复杂攀爬动作时机器人的重心较为稳定。

参考文献：

［1］霍淑珍，何志超.协作机器人在智能制造中的应用[J].机床与液压, 2021, 49(9):62-66.

［2］白克,王龙.基于 Simulink 的采摘机器人智能控制系统云平台仿真[J].农机化研究, 2021, 43(8):225-229.

［3］李肖,李世其,韩可,等.面向实时自避碰的双臂机器人力矩控制策略[J].信息与控制，2023, 52(2):211-219,234.

［4］韩连强,陈学超,余张国,等.面向离散地形的欠驱动双足机器人平衡控制方法[J].自动化学报，2022, 48(9):1976-1987.

［5］冯春,张祎伟,黄成,等.双足机器人步态控制的深度强化学习方法[J].计算机集成制造系统，2021, 27(8):2341-2349.

［6］Wang C，Shen J，Ran H .Imagining robots of the future: Examining sixth-graders perceptions of robots through their literary products[J].Journal of Research on Technology in Education, 2023, 55(4):684-709.DOI:10.1080/15391523.2022.2030264.

［7］张品,李长勇.基于改进模糊 PID 的全向搬运机器人路径跟踪控制研究[J].食品与机械，2021, 37(6):114-119,190.

［8］廖发康,周亚丽,张奇志.变长度柔性双足机器人行走控制及稳定性分析[J].计算机应用, 2023, 43(1):312-320.

［9］黎晴亮,张志安,马豪男,等.四足机器人抗重心偏移步态优化[J].计算机工程与应用, 2022, 58(7):303-310.

［10］韩连强,陈学超,余张国,等.面向离散地形的欠驱动双足机器人平衡控制方法[J].自动化学报，2022, 48(9):2164-2174.

［11］孟芸,周福娜,卢志强,等.双足机器人五质心模型的预测控制实现方法[J].机械设计与制造, 2022(3):254-257.

［12］高家昌，高建设，陶征,等.被动步行平足机器人动力学参数研究[J].机械传动, 2022, 46(12):22-30.

［13］徐征,张弓,汪火明,等.基于深度循环神经网络的协作机器人动力学误差补偿[J].工程科学学报，2021, (17):995-1002.

［14］高扬,张传玺,王晨,等.基于深度学习的激光同步定位与地图构建移动机器人可定位性研究[J].科学技术与工程, 2021, 21(32):13774-13780.

［15］刘鑫，王忠，秦明星,等.多机器人协同 SLAM 技术研究进展[J].计算机工程, 2022, 48(5):1-10.

［16］冯春,张祎伟,黄成,等.双足机器人步态控制的深度强化学习方法[J].计算机集成制造系统，2021, 27(8):2341-2349.

［17］王瑗珲,胡宁宁,喻俊,等.基于步态数据的机器人鲁棒自适应 PD 控制[J].控制工程, 2021, 28(9):1928-1939

［18］赵春华,李谦,胡恒星,等.一种新联合损失函数优化的迁移学习神经网络磨粒识别研究[J].润滑与密封, 2021, 46(4):26-31.

［19］张振宇,林沐阳.人工神经网络中的一种 Krylov 子空间优化算法[J].工程数学学报, 2022, 39(5):681-694.

［20］赵营鸽,李颖,王灵月,等.基于均值点展开的单变元降维法在 EIT 不确定性量化研究中的应用[J].电工技术学报, 2021, 36(18):3776-3786.

基金项目： 2022 年度广州华商学院高等教育教学改革项目（HS2022ZLGC71）

作者简介：

李丽霞（1983—），女，汉，山西长治人，硕士研究生，讲师。

陈艳（1979—），女，汉，湖北石首人，硕士研究生，讲师。