一种基于强化学习的仿真光电平台的驱动方法

丁 琰[1] 张文琼[1] 陈 颖[1] 张天琪[1] 王宣林[1] 熊华星[1]
DING Yan ZHANG Wenqiong CHEN Ying ZHANG Tianqi WANG Xuanlin XIONG Huaxing

摘 要 提出一种基于强化学习的仿真光电平台的驱动方法，用于替代在光电平台控制中经常用到的 PID 控制算法，以实现驱动光电平台快速跟踪海面舰船目标的目的。所提出的仿真光电平台的驱动方法可以提高光电平台的响应速度，快速驱动光电平台跟踪目标，同时具备无需手动调节参数、对速度快速变化的舰船目标适应性强的优点。

关键词 光电平台；仿真；强化学习；机器学习；海面目标

doi：10.3969/j.issn.1672-9528.2024.09.040

0 引言

目标跟踪是光电平台系统重要的功能之一。目标识别系统识别到目标后向光电平台输出目标所在的位置，或者输出目标相对于光电平台视野中央的偏移量。光电平台根据接收到的目标位置或偏移量向伺服电机系统发出信号，从而驱动光电平台在 x 轴和 y 轴上向目标的方向运动，并将目标重新定位到视野中央。在实践中，光电平台经常使用具有三个环路的 PID 算法进行控制，并通过调节 PID 算法的参数使光电平台能够快速、平稳地跟随目标。

具有三个环路的 PID 算法包括：最外层的位置反馈环，用以确定光电平台的目标位置；中间层的速度反馈环，用以调节光电平台的运行速度；最内层的转矩反馈环，用以控制光电平台运行过程中的转矩。其中，每层 PID 环都有 P、I 和 D 三个参数，三个 PID 环共有 9 个可调节参数。

具有三个环路的 PID 算法因为其结构简单在实践中得到了广泛的应用。但是，PID 控制算法存在以下一些问题：第一，参数调节复杂，光电平台具有 x 和 y 两个自由度，每个自由度需要使用一个三环 PID 调节，每个三环 PID 具有 9 个可调节参数，所以共需要对 18 个参数进行调整，而这些参数之间还存在一定的耦合性，可能会互相影响；第二，性能不高，PID 算法性能有限，即使在较为合理的参数下，特别是对于速度快速变化的目标，也容易出现跟踪延迟较大或者跟踪不平稳的情况；第三，鲁棒性不高，当光电平台的硬件参数发生改变时，往往需要对 PID 算法重新调参。

随着近年来机器学习的快速发展，机器学习方法在许多领域得到了成功应用并取得了良好的效果。因此，本文考虑使用强化学习方法作为光电平台的控制算法，以克服上述三环 PID 控制中存在的诸多问题，提高光电平台的跟踪性能。

1 仿真环境

用于仿真的环境有很多，其中常见的有 MuJoCo 和 Issac 以及 PyBullet 等，用于强化学习时，还需要 OpenAI Gym 等强化学习框架的支持。这些仿真环境的共同点是它们都由物理仿真引擎驱动，即对模型的物理属性，例如质量、体积、摩擦力、加速度等物理量进行模拟，从而驱动模型运行。由于本文提出的方法需要驱动电机运动，为使得仿真中电机的运动可以容易迁移到实际环境中，本文使用可以精确仿真电机运动的物理引擎进行仿真。

1.1 MuJoCo 环境

MuJoCo 是多关节接触动力学（multi-Joint dynamics with contact）的缩写，它是一个通用的物理引擎，旨在促进机器人学、机器学习及其他需要物理仿真的学科的研究。该引擎后来被 deepmind 收购并开源。MuJoCo 是机器人操作系统（ROS）中最常用的物理仿真引擎，与机器人操作系统（ROS）很好地兼容。MuJoCo 可以读取 URDF 格式的机器人模型文件，并进行仿真，MuJoCo 也有自己的数据格式 MJCF。MJCF 格式是 URDF 的扩展，语法结构与 URDF 类似，同时扩充了更加丰富的对模型和环境的描述，例如环境光照。

1.2 PyBullet 环境

PyBullet 是另一个常用的物理仿真环境，它基于开源的物理仿真引擎 Bullet Physics SDK 开发，这是一个成熟的、广泛使用的开源物理引擎开发。得益于它提供的 Python 接口，PyBullet 具有高度的易用性，它同样支持加载和仿真 URDF 文件，同时提供了逆向动力学、运动规划等完善的功能，可以精确仿真多体系统的动态行为。同时，作为开源软

1. 中国船舶集团有限公司系统工程研究院 北京 100094

件，PyBullet 具有良好的社区支持。但是相对于 MuJoCo，PyBullet 内置的渲染器能力较弱，在某些复杂的场景可能需要额外的渲染工具支持。

1.3 Nvidia Issac 环境

Issac 是英伟达推出的高性能物理仿真环境。Issac 由 PhysX 物理引擎驱动，使用张量作为内部的数据交换格式，从而减少数据在 GPU 和 CPU 之间的传输时间。Issac 对 GPU 提供了良好的支持，它可以在一个仿真环境下创建数万个智能体，同时在多块 GPU 上进行训练，这对于强化学习驱动的机器人设计是非常重要的。得益于对 GPU 的优化，Issac 具有很高的仿真性能，在本文的实验中，使用 Issac 环境和一块 GTX 3070，仅仅需要 6 个小时就可以让人形机器人学会行走。

2 方案设计

强化学习是机器学习的一个重要分支 [1]，它解决的是强化学习代理在一个外部环境中应该如何给出动作，从而可以使得累计的来自环境的奖励（回报）最大。强化学习并不需要给出标注的数据，它在一次次的动作尝试中找到最优的动作序列。

强化学习本质上是一个马尔可夫决策过程（MDP）[2]，它建立在这样一个假设之上，即当前的决策只与上一步的状态有关，这极大地简化了强化学习的求解过程 [3]。强化学习在过去几年里得到了很大的发展，其成功的应用范例包括 Atari 和 AlphaGO。

2.1 强化学习的环境设置

机器学习的训练需要大量数据，为了能够获得足够的样本进行训练，本文提出的方法首先通过使用三维仿真引擎进行仿真，并从仿真环境中获取实景数据作为训练样本，对强化学习代理进行快速原型验证和训练，再将训练好的算法模型移植到实际的光电平台上。

本文提出的方法在三维仿真引擎中建模了一个三维实景环境，实景环境包括海平面、天空以及在海平面上行驶的舰船。该实景环境主要关注点在于其中的舰船模型需要尽量接近真实舰船的外观，以方便目标识别系统的识别，通过在三维仿真引擎中对舰船的模型进行贴图，此关注点可以较为容易地被解决。整个实景环境结构简单，属于成熟技术，在本文中不再赘述。通过三维仿真引擎提供的摄像机模组，可以模拟光电平台以及光电平台的运动，以采集实景环境中的二维图像数据。光电平台采集到的图像数据包含海平面和在海平面上运动的舰船目标。图像的分辨率为 1920×1080，可选的，以图像的中心点作为坐标原点建立横向的 x 轴和纵向的 y 轴并取向右为 x 轴正方向、向上为 y 轴正方向。

在强化学习与环境的每一次交互中，上述环境返回一张实景环境的二维图像。通过使用目标识别算法例如 YOLO-v5 对图像进行处理 [4]，识别图像中的舰船目标，在舰船目标周围绘制包围框，并返回包围框的中心像素坐标作为环境的最终输出。

2.2 强化学习的奖励（reward）设置

本文提出的方法在接收到环境返回的舰船目标的像素坐标 (x,y) 后，分别求取目标在 x 轴和 y 轴上与坐标原点的像素距离，得到像素偏移量，得到像素偏移量后对像素偏移量取负，之后对负的偏移量进行归一化处理，归一化后，将 x 轴和 y 轴的归一化后的偏移量加权取和作为最终的奖励值。完整的奖励的计算公式为：

$$R = -0.8 \times \frac{|x|}{960} - 0.2 \times \frac{|y|}{540} \quad (1)$$

上述公式中的归一化值和奖励系数的值为经验值，在本文中取归一化的值为图像像素尺寸的一半即在 x 轴上取 1920/2、在 y 轴上取 1080/2；取 x 轴的奖励系数为 0.8，y 轴的奖励系数为 0.2。对于火目标跟踪系统而言，光电平台采集到的图像中沿舰艇沿水面横向（x 轴）移动的情况较多，沿水面纵向（y 轴）移动的情况较少，因而能够在 x 轴上快速地跟踪目标移动的价值相对较大。

2.3 强化学习代理的选择

强化学习中，有两种主要的学习形式，一种是基于值的学习，一种是基于策略的学习 [5]。基于值的学习方法通过强化学习代理输出期望的动作值，随后根据期望的动作获得的回报对动作的价值进行评估，从而确定在某个状态下价值最大的动作，并在该状态下选择回报最大的动作执行；基于策略的强化学习方法则是通过回报对强化学习代理进行训练，使得代理可以直接输出获得最大回报的动作值。这两种学习形式各有优劣，但都可以用于驱动本文中的光电平台 [6]，其中基于值的方法的典型代表算法是 Q-Learning，基于策略的学习方法的典型代表是近端策略优化（PPO），本文对这两种算法分别进行了实验，并分别给出了性能评估。

2.4 目标跟踪器的选择

目标跟踪算法不在本文的探讨范围之内，但是为了方案的完整性，本文在实验中选取了 YOLOv5 作为目标跟踪器。目标跟踪器在光电平台返回的实景图像中找到舰船目标，使用矩形包围框标注舰船目标的外部轮廓，并返回包围框的像素中心点作为目标的位置信息。

2.5 强化学习的整体设计

环境返回的值是舰船所在位置的中心像素坐标，得到像素坐标后，通过对像素坐标取绝对值并作归一化和加权取和等操作获得最终的奖励[7]。在本文提出的方法中，强化学习代理的输入并不是由环境返回的舰船目标的像素坐标值，而是像素坐标值与光电平台的当前位置和当前速度进行的拼接，以提供更完整的环境信息[8]。这样设计的目的是使代理在发出下一个动作时可以充分考虑到当前光电平台的状态，因为实际的光电平台由电机驱动，而取决于环境采样的速度，电机的速度在接收到下一个动作指令时并不一定为 0。

在本文提出的方法中，强化学习代理使用三层全连接网络，接收拼接后的环境输出作为输入，并最终输出两个动作值。这两个动作值分别作为施加到光电平台 x 轴上的转矩和 y 轴上的转矩，并驱动三维仿真环境中的摄像机以模拟光电平台的运动。

3 实验设置和结果

本文提出的方法对虚拟实景环境采样产生训练数据，并采用 Q-Learing 和 PPO 算法分别在 Nvidia GTX 3090 上训练了 100 000 次得出结果。为减小优化器和学习率对两种方法的结果的影响，在 Q-Learing 和 PPO 中，均使用 Adam 优化器，学习率随训练过程线性减小。

为进一步提高算法对速度快速变化的目标的适应性，仿真实景环境模拟舰船在水平面上做随机的变加速运动，速度在 0 pixels/s ～ 100 pixels/s 的范围内随机变化，加速度在 -25 pixel/s^2 ～ 25 pixels/s^2 的范围内随机变化。这种变加速运动在实际场景中几乎不会发生，但是会较大提高算法的跟踪性能和鲁棒性[9]，如果强化学习算法在这种随机的变加速运动中能够表现出良好的跟踪性能，那么在实际场景中也可以很好地跟踪舰船的运动。

训练结果如图 1 和图 2 所示，图 1 和图 2 分别为 x 轴和 y 轴上的像素跟踪误差与训练步数之间的关系。其中，图像的纵轴表示舰船目标与坐标原点的像素偏差，图像的横轴表示训练次数。

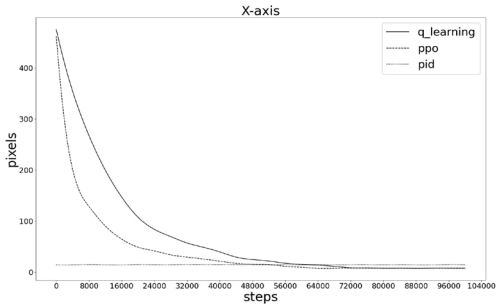

图 1 在 x 轴上的训练结果

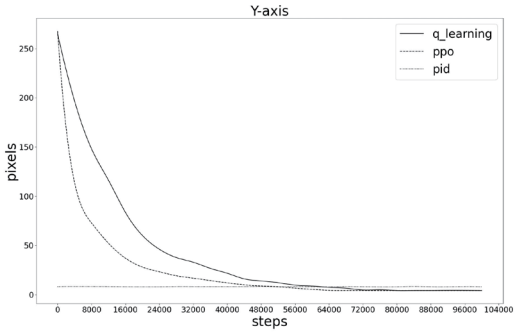

图 2 在 y 轴上的训练结果

4 性能评估

在有限次的迭代后，无论是基于值的 Q-Learing 还是基于策略的 PPO，性能均明显高于两组参数的 PID 算法的性能。相较于 Q-Learning，PPO 的收敛速度更快，这可能得益于策略梯度方法本身所具有的更高的稳定性[10]。

策略梯度方法直接根据累计回报求取策略，直接地生成最大回报对应的动作。相对于 Q-Learning 方法，策略梯度方法的输出可以表示为一个分布，并从分布中采样，从而实现连续的动作空间。策略梯度方法面临方差较大的问题，为了降低样本方差，出现了 actor-critic 方法，通过使用价值函数作为基线，提高了训练的稳定性，加快了收敛的速度。根据策略梯度定理，每一步输出的动作引起的环境分布的变化必须足够小，才能够在无穷步时达到马尔科夫链的平稳分布。基于这种思想，出现了限制动作幅度的 TRPO 和近端策略优化（PPO）等新的算法，大幅度地提高了策略梯度方法的稳定性，使强化学习趋向稳定和易用。

5 结语

在本文的实验中，并未考虑舰船目标与光电平台的距离，本文的方法使用像素坐标，实际上是假设舰船目标均投影在无穷远的平面上。这种假设对跟踪的精度影响不大，实际场景中舰船目标通常距离光电平台很远，因此投影后的距离误差也很小。后续可以通过使用单目的深度视觉算法估计舰船与光电平台的距离，对目标在像素坐标系下的位置做校正，从而进一步补偿精度。下一步工作还包括将在仿真环境下训练好的模型移植到物理的光电平台上，进行相应的性能评估。

参考文献：

[1] 李连民，孙立功，孙士保．一种改进的视觉词包模型的船舶识别方法 [J]. 河南科技大学学报 (自然科学版), 2024, 45(4): 10-16+115+4.

[2] 于乐凯，曹政，孙艳丽，等．海上舰船目标可见光 / 红外图像匹配方法 [J/OL]. 海军航空大学学报 ,1-11[2024-06-07]. http://kns.cnki.net/kcms/detail/37.1537.v.20240628.1533.008.html.

卫星通信系统网管主备切换研究

高 杨[1] 周志伟[2] 王 沛[1] 孙万海[1] 刘 萍[1] 张克龙[1]
GAO Yang ZHOU Zhiwei WANG Pei SUN Wanhai LIU Ping ZHANG Kelong

摘 要 针对高通量卫星通信网络管理系统主备切换应用的需求，分别对高通量卫星通信的网络管理系统和主备切换结构进行设计，并根据通信网络节点多、用户终端量大、高可靠性的特点，提出了一种分层、模块化的灵活架构，同时对网管系统进行容灾备份、主备切换，最后提出一种卫星通信系统网管主备切换场景应用。

关键词 卫星通信；高通量；网络管理；主备切换；模块化

doi：10.3969/j.issn.1672-9528.2024.09.041

0 引言

近年来，我国卫星通信应用范围不断扩大、产业规模化。相较于传统卫星通信方式，高通量卫星通信呈现容量大、覆盖广、速度高、组网便捷等特点，因此更具有价值优势。随着中星 16、中星 19、中星 26 等高通量卫星的成功发射，我国境内及周边区域已具备高通量卫星互联网覆盖能力。伴随着卫星通信网络的深度融合、多层次覆盖、端到端统一编排调度、智能化发展，传统的卫星网络的管理方式已无法满足当前复杂的业务场景。针对目前卫星通信多卫星、多信关站和多频段卫星资源的管理，以及网络设备种类多和用户终端量大、应用场景多样化、用户需求定制化等的现状，需要采取综合管控手段进行资源调配，这对网络管理体系的集中化、智能化、可视化有较高的要求。针对卫星通信系统目前发展过程中急需解决的问题，本文主要研究卫星通信网络管理系统中的主备切换问题。

1. 航天恒星科技有限公司 北京 100086
2. 亚太卫星宽带通信 (深圳) 有限公司 广东深圳 518126
[基金项目] 课题编号 D010203

1 高通量卫星通信网络管理系统结构设计

卫星通信网络管理系统负责全网的卫星资源统一管理和调度，对卫星通信网络中所有设备进行监视和管理，对终端设备进行控制及资源分配。根据具体的业务场景和用户需求，系统的结构可有所不同，这里以卫星通信系统通用场景为例，设计一种卫星通信网络管理系统结构，主要功能模块包括配置管理、监控管理、故障管理、性能管理、系统管理、用户管理等。同时，为了给业务运营系统提供信息服务，这里通过标准的北向接口，比如告警、流量等相关接口，为维护提供支撑。根据高通量卫星通信网络的特点，结合实际卫星通信网络规模的变化和终端规模不断扩大，以及多星、多信关

[3] 李丽，梁继元，张云峰，等 . 基于改进 YOLOv5 的复杂环境下柑橘目标的精准检测与定位方法 [J/OL]. 农业机械学报 ,1-10[2024-06-07].http://kns.cnki.net/kcms/detail/11.1964.s.20240624.1519.017.html.
[4] 陈真，朱嘉晟，陈潇潇，等 . 水下捕捞机器人视觉系统发展趋势分析 [J]. 造船技术 ,2024,52(3):49-54+77.
[5] 赵增辉，唐明 . 基于深度学习的移动机器人目标自动跟随控制系统设计 [J/OL]. 计算机测量与控制 ,1-12[2024-06-07].http://kns.cnki.net/kcms/detail/11.4762.TP.20240625.1748.002.html.
[6] 杨俊秀 . 基于机器视觉的多视角水面船舶识别方法研究 [D]. 厦门 : 集美大学 ,2024.
[7] 马中静，韩佳蓉，赵祖欣，等 . 新工科背景下的水下机器人实践教学平台 [J]. 电气电子教学学报 ,2024,46(1):196-200.
[8] 艾小锋，吴静，张静克，等 . 空天目标雷达智能识别仿真系统设计与实现 [J]. 现代防御技术 ,2024,52(2):151-162.
[9] 张明容，喻皓，吕辉，等 . 面向自动驾驶的多模态信息融合动态目标识别 [J]. 重庆大学学报 ,2024,47(4):139-156.
[10] 丁士强，梁玉峰，代昌盛，等 . 基于显著性检测的舰船目标识别方法研究 [J]. 科技视界 ,2024,14(10):56-59.

【作者简介】
丁琰（1989—），通信作者（email: dingyanht@hotmail.com），男，山东济南人，硕士，工程师，软件架构师，研究方向：机器学习。

（收稿日期：2024-07-08）