陕西科技大学

硕士研究生学位论文开题报告

论文题目	基于强化学习的足式机器人控制方法研究
学生姓名	陶亚凡
学科专业	控制科学与工程
学号	1706073
导师姓名	党宏社
所在学院	电气与信息工程学院
答辩日期	2018 年 10 月 29 日

2018 年 11 月 5 日

一、选题依据（论文选题的背景、目的、意义、国内外研究现状分析等）

1.1. 背景

1.1.1. 足式机器人的研究背景

“机器人”一词最早由 1920 年捷克剧作家卡雷尔·恰佩克提出[1]，原文称作“Robota”，后在西方统称为“Robot”。对于机器人的定义有很多，包括牛津字典、美国机器人协会、日本工业机器人协会和国际标准化组织等对机器人均有不同的定义，我国对机器人的定义为“一种拟人功能的机械电子装置”。对于不同的定义，有几项共同点，（1）像人或人的上肢，并能模仿人的动作；（2）具有智力或感觉与识别能力；（3）是人造的机器或机械电子装置[2]。机器人一般分为工业机器人和智能机器人（在后文中，若无特殊说明，机器人即表示智能机器人），工业机器人是一种能执行与人的上肢（手和臂）类似动作的多功能机器，智能机器人是一种具有感觉和识别能力，并能够控制自身行为的机器。工业机器人应用较为固定，多为工业的制造车间中，由人类编写程序完成相对固定的任务；而智能机器人应用较为广泛，包括医疗、航空、科研、军事还有娱乐等应用场景，是我们生活中最常见的机器人，也逐渐在各个领域变得不可缺少。

智能机器人一般具有自己的运动方式，比如轮式、足式、飞行式、履带式、蠕动式和震动冲击式等[3]。轮式具有结构简单、稳定性强、易于控制和移动速度快等优点，因此应用最为广泛，但是应用场景有一定要求，比如地面要平坦、连续，当一旦遇到一定高度障碍物，如地面凸起或楼梯等，便无法通过或陷入无法移动的困难境地。而足式机器人只需要离散的支撑点，可以在地面较为复杂的环境中移动，环境的适应性较好，具有较好的灵活性。但是足式机器人相对于轮式机器人更加难以实现，由于其运动的规划和控制以及稳定性都是需要考虑的问题，因此对足式机器人的研究目前仍然是热点之一。

足式机器人的研究起源于对自然界生物的模仿，比如人类使用双足进行运动，大部分哺乳动物如马、狗、猫、牛，羊等均采用 4 足的方式进行运动，一些爬虫类动物会采用 6 足、8 足等多足方式运动。而不同足数的机器人在稳定性、负载能力、机构复杂度、移动速度和控制难易程度等方面有着各自的优势[4]。一般对于双足和四足的研究较多，四足机器人稳定性较强，控制复杂度较为简单，因此相对于双足机器人研究成果更多，而双足机器人灵活性更高，类人的外形对大部分人来说显得更加友好。

1.1.2. 强化学习的研究背景

强化学习是一种学习方式，对于一个智能体（agent），比如一个机器人，强化学习就是赋予它学习如何去做的能力，也就是学习一个从自身状态到输出动作的一个映射，从环境中感知到信息推测自身所在状态，然后决定输出的动作来使自己获得最大的奖赏（reward）。教学者不告诉智能体如何做能获得更高的奖赏，取而代之的是智能体自己学习应该采取什么样的动作[5]。

强化学习是人工智能（AI）领域的重要组成部分之一，AI 领域的主要目标之一是产生出完全自主的智能体，它们与环境交互以学习最优行为，通过不断尝试和错误，随着时间的推移不断改进。AI 系统的设计是一个长期存在的挑战，从能够感知周围世界并做出反应的机器人，到能够与自然语言和多媒体互动的纯软件代理。经验驱动自主学习的一个基本框架就是强化学习。尽管 RL 在过去取得了一些成功，但以前的方法缺乏可伸缩性，并且天生局限于相当低维度的问题。近年来我们所看到的深度学习（Deep learning[6]）的兴起，依靠深度神经网络强大的函数逼近和表征学习特性，为我们提供了解决这些问题的新工具[7]。

在 2013 年，DeepMind 首次提出将强化学习与深度学习相结合，也就是后来广为人知的深度强化学习（在后文中若无特殊标注，强化学习则指代深度强化学习），用来教计算机玩 Atari 2600 的 7 种游戏，输入像素级别的游戏原始画面，输出游戏指定的动作[8]。近些年深度强化学习在很多领域非常成功，2015 年 DeepMind 在原有基础上稍加改进，正式提出 DQN 模型，在 49 种游戏中达到人类顶尖水平[9]；同年 AlphaGo 大放异彩，围棋程序先学习人类棋谱，后采用强化学习与自己对弈提升棋力，后来在围棋比赛中战胜世界冠军李世石[10]；AlphaGo 之后又提出 AlphaGo Master 和 AlphaGo Zero[11]，后者直接放弃学习人类棋谱，直接使用强化学习在自己对弈中学习；在自然语言处理领域强化学习同样有一席之地，2017 年 L Yu 等人提出的 SqeGAN 将 GAN[12]与强化学习相结合，在序列生成任务中达到了世界领先水平[13]。

在机器人领域，实现机器人运动控制的一般方法为建立机器人的运动学模型和动力学模型，并设计运动时的步态和轨迹规划，然后按照步态规划、轨迹规划、运动学逆解和动力学逆解的步骤计算出每个执行器输出的力或力矩，这些任务往往会遇到这些问题：（1）建立精确的模型较为困难。（2）需要提前考虑机器人所遇到的任何环境和规划机器人的所有行为，一旦遇到非预期的困难时机器人难以独立解决。而强化学习可以提供给机器人学一个设计复杂和人为难以设定的工程的行为的工具集和框架[14]。

1.2. 目的

将强化学习引入足式机器人的控制过程中，使机器人具有一定的学习能力，经

过不断的自我尝试，学会站立或行走。

1.3. 意义

（1）降低机器人的设计难度，简化设计过程。设计机器人运动的过程很大程度上依赖于工程师们的手工设计，他们需要从已有的知识中发现适合于给定形态的步态[15]。尽管经验丰富的工程师可以为各种各样的机器人设计出惊人有效的步态，但是对于复杂或硬件精度低的机器来说，建立精确的模型是很困难的，特别是当试图考虑到身体可能发生的拓扑变化时，比如由损伤引起的变化[16]。使用强化学习可以使机器人具有学习能力，不需要建立模型或要求模型精度很高，可以由机器人自己学习自身的模型，并自己规划动作和决定输出动作输出的大小和幅度。

（2）增加机器人智能程度，提升环境适应能力。在遇到没有预期到的困难时，传统的机器人原有程序设定不含有解决问题的方法，但是，具有自学习功能的机器人可以通过观察预期结果与实际结果的差异，不断改善输出动作，学习解决方法，克服困难，因此适应性更强，智能程度更高。

1.4. 国内外研究现状

1.4.1. 足式机器人研究现状

在国外足式机器人的研究较早，在 1968 年，同时诞生了双足机器人[17]和四足机器人[18]。同在 1968 年，日本开始做双足机器人的研究工作，并先后在 1969 年、1967 年分别成功研制除了 WAP-1 平面自由度步行机和 WAP-3 双足机器人[19]。1977 年，McGhee 课题组创造出第一个由计算机控制的 4 足机器人，外形仿昆虫，单腿的 3 个自由度由电机控制[20]。他们又在 1986 年研制了 ASV hexapod 机器人，该机器人的自适应悬挂机制性能优越，使其成为目前体型如此大的机器人中自适应能力最强者。

进入 80 年代后随着计算机技术和传感器技术等领域的快速发展，足式机器人得到了飞跃式发展。福田机器人实验室研制了 PV-II 四足机器人，使用了姿态传感器和足端触觉传感器，研究一种运动控制算法，是世界上第一个能自主爬楼梯的四足机器人[22]。1990 年，美国的 Ohio 大学 ZHENG 等人提出用神经网络实现双足机器人的动态步行，并实现了 SD-1 双足机器人[23]。同年日本的 Kajita 研制出了能在不平整的面上稳定行走的五连杆双足机器人[24]。

从 1994 年起，木村•宏开展了基于中枢模式发生器（Central Pattern Generation, CPG）的步态控制策略，开发了 Patrush 系列机器人和 Tekken 系列小型放狗机器人，并实现了平坦地形下多种步态行走和自然地形下 trot 步态的行走[25]。

1995 年，日本早稻田大学开发了“HADALY-1”机器人，之后集合之前的研究

结果，又于 1997 年开发了“HADALY-2”和“WABIAN”机器人[26]。1996 年日本本田又推出了世界第一台拟人的自主双足不行机器人 P2，之后在 2000 年，其更是推出了令人震惊的 AMSIMO 机器人[27]。2002 年，日本索尼公司展示了 SDR-4X 小型机器人，之后在 2003 年，又推出了 QRIO 仿人机器人[28]。2004 年，法国 Aldebaran Robotics 公司开启了仿人机器人项目“NAO”，具有众多传感器和方便的开发环境，方便了很多相关的学术研究。

2006 年波士顿动力公司（DBI）成功研制出高动态液压四足机器人——第一代 BigDog[29]，在 2008 年发布了第二代 BigDog[30]，是迄今为止第一台真正意义上的能量自给、能够奔跑的跳跃的四足机器人。随后 DBI 又于 2016 年发布了双足机器人 Atlas[31]，并持续更新，在 2017 年实现在雪地上稳定行走和后空翻，在 2018 年实现单腿跳跃翻过障碍，代表了当前机器人领域的最先进水平。

足式机器人在中国发展较晚，在上世纪 80 年代才开始了足式机器人的研究计划，很多高校投入了对双足的研究包括，包括哈尔滨工业大学[32]、浙江大学[33]、华南理工大学[34]、国防科技大学、北京理工大学、上海交通大学、清华大学等[35]。对四足机器人的研究包括了上海交通大学[36]、中国科学院合肥智能机械研究所[37]、华中科技大学[38]、西北工业大学[39]、中科院自动化研究所[40]、北京理工大学[41]、山东大学[42]、国防科技大学等[43]。

1.4.2. 强化学习在机器人领域的研究现状

在机器人领域，很多问题可以自然的表达为强化学习的问题，强化学习使得机器人在与环境的交互中不断的试错（trial-and-error）来自动的发现最优的动作[44]。在上世纪 90 年代，开始有人使用强化学习来做控制任务的研究，如 1994 美国的 Gullapalli 等人让机械臂学会了一项插杆的任务[45]，在 1997 年日本的 Schaal 让一个人形的机器人学会如何控制自己的手臂使得手上放的一个直立的杆子保持平衡[46]，在 2001 年美国斯坦福大学的 Bagnell 和 Schneider 使用一种基于模型的策略寻找方法，使得无人直升机习得鲁棒的飞行控制[47]。

随着深度学习的发展，近些年来在图像、视频、语音和音频处理方面取得了突破性进展，在文本和连续语音等连续数据上也发挥了重要作用[48]。在 2013 年，DeepMind 首次将深度学习与强化学习相结合，标志的深度强化学习的诞生[8]。深度强化学习的到来为连续控制任务带来了一段蓬勃的发展期，各种算法和应用层出不穷。

2015 年 DeepMind 在《Nature》上正式提出 DQN 算法，用神经网络来处理高维的状态[9]，随后，Hasselt 等人和 Wang 等人分别在 2016 年和 2015 年提出了 DQN 的改进版，Double DQN[49]和 Dueling DQN[50]来解决 DQN 估值过高的问题。为了使

机器输出连续的动作，Lillicrap 等人在 2015 年提出了 DDPG 算法[51]，同年，Schulman 等人提出了 TRPO 算法[52]。在 2016 年，Mnih 等人提出了 A3C 算法，同年 Gu 等人提出了 NAF 算法[54]。在 2017 年，Schulman 等人提出了 PPO 算法[55]，同年 Wu 等人提出了 ACKTR 算法[56]。

上述算法一般实现在各种仿真环境中，但也有一些研究针对真实的环境。2016 年 Levine 等人利用基于模型的 GPS 算法使一个带有 7 个自由度的机械臂实现了挂衣服、分拣各种形状的立方体和在瓶盖上拧螺丝等[57]。2017 年 Gu 等人提出一种异步的 NAF 算法，使机器人可以打开真实世界的门[58]。2018 年 Riedmiller 等人提出一种只从二元或稀疏的奖励中加速学习的方法，使得一个 9 自由度的机械臂可以获得一个放在有盖盒子里的物体，或把两块积木堆到一个塔上[59]。同年，Ha 等人在真实环境中使用了 TRPO 和 DDPG 两个算法，提出一种学习环境并实现了一种可扩展的多足机器人实现直线行走[60]。

在国内相关的研究较少，研究机构主要包括浙江大学[61]、华南理工大学[62]等。

二、研究内容及拟解决的关键问题

2.1. 研究内容

2.1.1. 机器人学习环境的设计

机器人的学习需要一个学习的环境，来保证机器人在学习过程中的“安全”，和对机器人的状态进行获取，并根据机器人的状态或动作来给予适当的奖励或批评，使机器人“知道”自己做的是否正确。

在仿真环境中，学习环境较易搭建，主要任务是选择合适的仿真环境，主要依据仿真环境的易扩展性、用户规模、说明文档的丰富程度和企业支持。

在真实环境中，学习环境负责的具体内容包括

（1）对机器人进行定位；

（2）判断机器人的状态（摔倒、直立和移动速度等）；

（3）根据机器人的状态判断机器人是否正常，若不正常，如快要离开监视范围，将对机器人发出停止指令，由人手动或自动将其拉至监视范围中心,在允许其继续学习；

（4）根据机器人的状态并结合人为设定的目标，反馈一个得分，告诉其当前动作或状态的好坏，如设定的目标是让机器人站立，结果机器人执行动作后倒地，则给一个低分告诉其当前的动作是错的，反之，若机器人正在站立，则给其一个高分表示当前的动作是对的。

2.1.2. 机器人硬件设计

为了实现真实环境下的机器人，因此机器人的机械结构和硬件电路也是需要考虑的。机械结构需要综合考虑学习环境确定机器人的大小，确定机器人的自由度个数，再确定机器人各个结构组件。

对于硬件电路，需要设计动作执行机构和控制电路，动作执行机构负责机器人每个关节的转动，而控制电路控制执行机构；机器人需要传感器来感知环境和自身的信息，因此需要选择传感器和设计相应的传感器获取电路；而强化学习的算法部分需要一个算力较高的平台，因此要设计信息处理单元，专门对算法部分进行实现。

2.1.3. 机器人学习过程抽象模型的设计

这部分内容属于学习算法的范畴。在使用强化学习之前，首先要对机器人的学习过程进行抽象，得到符合强化学习的模型，及马尔科夫决策过程，需要将机器人的学习过程抽象为：状态——动作——奖赏——状态——...——动作——结束。下面是这个过程中三个部分的具体内容：

（1）状态是机器人通过自身传感器的数据得到的信息和学习环境反馈的信息进行综合的得到的，对于马尔科夫决策过程，要求当前的状态与过去的任意一个时刻的状态独立，及当前的状态包含了过去的所有信息。因此需要选择合适的信息作为状态，且状态信息不能太多，太多则难以处理。

（2）动作便是机器人为了达到目标使得各个执行器做出相应的运动，要求理论上存在一系列的动作使得机器人达到人为设定的目标，高维的动作的学习目前一直是强化学习中研究的重点，也是本课题的重要研究内容之一。

（3）奖赏是告诉机器人每个动作执行的好坏，一般正值表示好的反馈，负值表示坏的反馈，值越大表示程度越大。但是根据学习环境看到的机器人状态来得到奖赏是研究的重要内容之一。

2.1.4. 策略函数建立与学习的研究

策略函数是根据机器人的状态，得到执行动作的一个函数，由于状态和动作均为高维，传统的强化学习算法（蒙特卡洛算法（MC）、时序差分算法（TD）等）无法处理，因此需要与深度学习相结合，为了处理高维的状态，需要使用卷积神经网络（CNN），而该 CNN 的结构和超参数的选取需要反复的设计验证，但强化学习不像监督学习，没有确定的期望输出结果，即直接根据误差函数进行梯度下降来学习是不行的，因此需要用到策略梯度法，对多次的奖赏进行采样，将奖赏最大设为目标函数对网络模型的参数求梯度，进行网络参数的学习。

2.2. 拟解决的关键问题

2.2.1. 真实环境中的机器人状态反馈

机器人在真实环境中，可以学习的要求就是有外部环境对其行为进行反馈，反馈装置需要利用摄像头等传感器获取机器人的信息，并进行处理得到机器人的状态，再根据奖赏机制确定得分。而如何获得机器人的信息、获得哪些信息以及如何对获得的信息进行处理都是本课题的关键问题。

2.2.2. 奖赏最优化或自适应奖赏

奖赏的确定一直是强化学习中的研究重点，一个好的奖赏的设定带来的效果非常明显，一般采用稀疏的奖赏，如围棋程序中，每次落子后获胜奖赏为 1，失败奖赏为-1，否则奖赏为 0。虽然大部分落子后奖赏为 0，但是每个棋局总会结束，因此每局棋均可以判断每个落子的好坏。但是在机器人学习站立的时候，若设定为稀疏的奖励，如执行一个动作后站立，奖赏为 1，倒地后奖赏为-1，其他为 0，则机器人可能一直都无法获得正的反馈，因此无法学习到站立。所以奖赏的设立可能需

要设定连续的值。或者针对奖赏设计一个网络，根据机器人当前状态自己计算奖赏的值，比如人类做一件事正确后，更多是内部分泌的多巴胺使得自己高兴，即内部给自己一定的奖赏。自适应奖赏这也是研究的关键问题。

三、研究方案及可行性分析（研究思路与方法、技术路线、实验或调查方案及可行性分析，从事自然科学研究所需主要仪器设备和试剂，从事人文社科类研究所需要的工作条件）

3.1. 研究思路与方法

通过查阅文献与书籍，学习强化学习相关算法，主要专注于机器人控制领域。先搭建仿真平台，对现有算法进行试验比较，选出最优算法并进行改进，在仿真环境中实现机器人的移动。仿真实现后在实际环境中搭建测试环境，使用优化后的算法，在实际场景中实现机器人的运动控制。

3.2. 技术路线

（1） 仿真平台搭建

（2） 建立机器人学习过程的抽象模型

（3） 选择合适的强化学习算法

（4） 算法优化

（5） 机器人硬件实现

（6） 真实测试环境的搭建与实现

（7） 机器人实际场景中运动的测试与实现

3.3. 研究方案

3.3.1. 机器人学习环境的设计

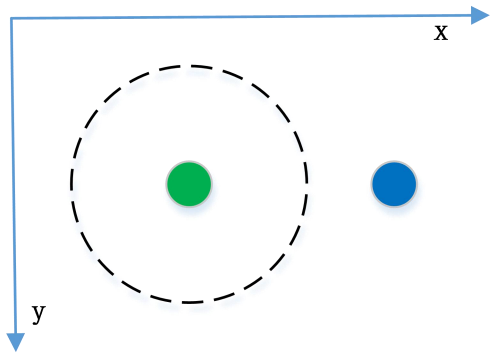

图 3-1 机器人学习环境示意图

在仿真时，采用现已成熟的仿真环境，如 OpenAI Gym，做机器人仿真的话需要 Mujoco 作为物理引擎，也有人增加了 DART 作为物理引擎，或使用 roboschool，该环境使用 bullet 作为物理引擎，开源免费。

真实环境中拟采用摄像头实现机器人的状态信息的获取，因此可监视的范围为

一个矩形，机器人的所在的平面示意图如图 3-1 所示，左侧实心区域是机器人起始位置，虚线圆圈是机器人在学习站立的时候不能超过的区域，右侧实心区域是机器人学习行走时的目的地，要求机器人初始方向正对蓝色区域。

为了检测机器人的位置和高度，一共需要 3 个摄像头，分别有一个在顶部，两个在侧面，示意图如图 3-2 所示，顶部摄像头检测机器人所在位置，判断有没有超出学习区域；侧面摄像头判断机器人的高度，可以判断机器人是否倒下。

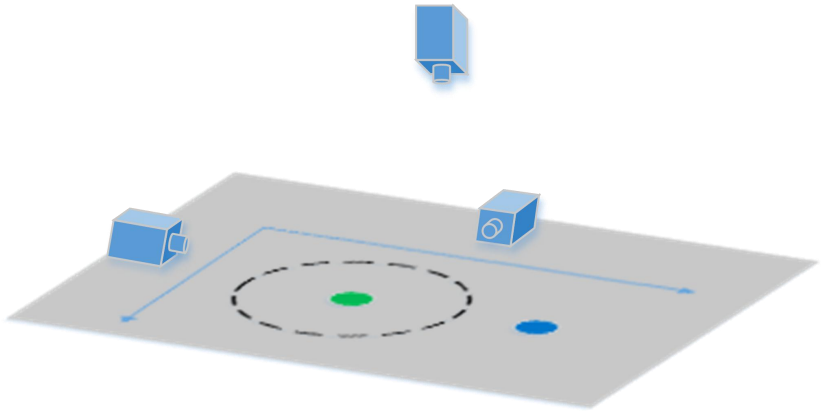

图 3-2 机器人学习环境三维示意图

3.3.2. 机器人硬件设计

机械结构拟采用较为成熟的方案，不自己单独设计，机器人头部采用特殊的形状或特别的颜色，方便摄像头识别。执行器硬件电路框图如图 3-3 所示，每个部分的介绍如下：

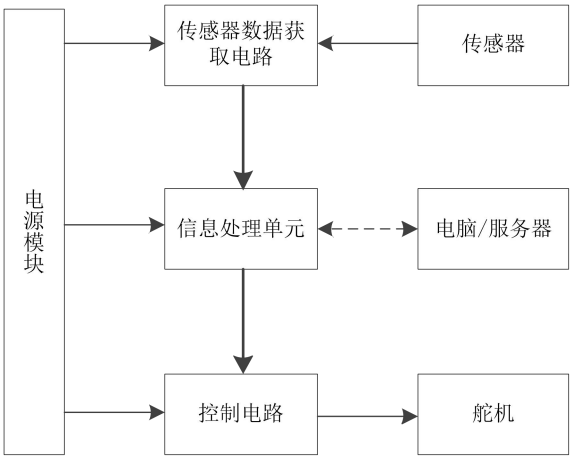

图 3-3 硬件组成框图

（1）电源模块的设计，电源作为机器人的能量的唯一来源，是首先需要考虑的。电源拟采用锂电池，但是电源的大小需要结合机器人的动力机构和整体机械结构来考虑，需要保证重量和能量适中，使得机器人可以独立移动，且尽量维持时间要久。电源模块也需要考虑其他硬件电路，保证可以为每一个硬件电路供电。电源的安全也是需要保证的，如过流保护等是最基本的要求。电源拟采用 DC-DC 的结构，电流大，效率高。

（2）动作执行机构和控制电路，动作执行机构及机器人运动的执行器，暂不考虑液压机构，主要考虑步进电机、伺服电机或舵机。控制电路的作用是控制执行器运动，需要考虑执行器的控制方式，拟使用舵机，控制方式为通过 PWM 占空比来控制角。控制电路采用单片机来实现控制信号的输出，通过一些常用的接口如串口、SPI、I2C 或 CAN 总线等实现与信息处理单元的通讯。拟采用 STM32 作为核心，使用串口与信息处理单元通讯。

（3）传感器数据获取电路，这个电路的功能是获取各个传感器数据，进行解析，通过常用的接口与信息处理单元通讯，将当前机器人的状态反馈给它。传感器拟使用陀螺仪、压力传感器和超声波传感器。陀螺仪用来检测机器人的三轴加速度和三轴角速度，为了使机器人知道当前的运动情况；压力传感器可以放在足端等位置、为了使机器人感知到足部或者身体个别部位的压力。拟采用 STM32 作为核心，使用串口与纤细处理单元通讯。

（4）信息处理单元，机器人也向人一样，需要一个“大脑”，这个大脑就是信息处理单元，接受传感器获取电路的数据，得到机器人当前的状态，并决策应该执行什么样的动作，并把要执行的动作告诉控制电路。如何决策便是这个大脑要学习的最重要的事情，使用强化学习算法来实现这个学习的过程。信息处理单元需要有较高的运算速度，单片机一般不能满足要求，选用更高级的嵌入式系统，并具备常见的无线连接设备，在训练过程中可以连接电脑或服务器来加速学习过程。信息处理单元拟使用树莓派，在机器人学习时，树莓派处理能力不足，可通过无线将数据发送给电脑或服务器，由电脑或服务器进行网络结构的学习，在机器人学习达标后可以单独使用树莓派进行信息的处理。

3.3.3. 机器人学习过程抽象模型的设计

如研究内容部分所介绍，主要是如何抽象状态、动作和奖赏。对于状态，要尽量多的包含信息，拟结合上一个时刻和当前时刻的自身传感器数据和环境反馈的头部高度和前进方向与速度以及上一个时刻的动作作为一个次的状态，既包含了当前的信息又包含了环境信息。对于动作，设定为输入每个舵机的占空比，即转动的角度。对于奖赏，不同的任务可以设定不同的奖赏，如站立的任务，可以根据头部的

高度来判断站立的效果，高度越高说明站立的效果越高，奖赏越高，还需要设定两个阈值，超过阈值一表示站立成功，此时给予正值最高奖赏，低于阈值二表示站立失败，给予负值最低奖赏。若学习其他任务，奖赏的设定方式也如此。

3.3.4. 策略函数建立与学习的研究

策略函数拟采用深度残差网络作为模型，输入机器人当前的状态，输出每个舵机转动的角度（控制器输出的 PWM 占空比）。对于策略函数的学习，OpenAI 在 2017 年提出了 PPO 算法，这种算法用在强化学习中时表现能达到甚至超过现有算法的顶尖水平，同时还更易于实现和调试，且 OpenAI 已经把 PPO 作为自己强化学习研究中首选的算法。因此本课题首先将尝试 PPO 算法，在未来再做算法改进。

3.4. 完成课题所具备的条件

（1）在校期间完成了线性系统、模式识别以及最优控制等相关课程的学习，具备完成本课题需要的理论基础。

（2）实践方面，本人做过智能车、无人机等控制任务，参加过多次电子设计竞赛，对机器人的控制任务有一定基础；在算法方面，做过深度学习、强化学习相关课题有一定基础。

（3）阅读了国内外与足式机器人的运动控制相关的文献和与强化学习相关的文献，做过基本的编程任务，对该领域有一定的了解。

（4）在校可得到相关领域的老师和同学的帮助，有问题和困难时，可以进行沟通和交流。

四、预期成果、创新之处、成果预期社会效益

4.1. 预期成果

（1）在仿真环境中实现足式机器人简单的控制任务，如站立行走等。

（2）完成机器人状态监测的环境的搭建。

（3）完成机器人实物，关节可控，并尝试学习站立行走。

（4）发表论文至少 1 篇。

4.2. 创新之处

将强化学习应用在足式机器人的控制过程中。

4.3. 成果预期社会效益

组成机器人的硬件可以由目前的高精度高成本降低为低精度低成本，将机器人的使用门槛降低，可以应用在更广泛的地方。且降低了新型机器人的研发成本，对新型机器人的提出及研发起到促进作用。

五、工作进度安排及经费预算

5.1. 工作进度安排

本课题计划分五个阶段实施：

（1）2018.11-2019.2

搭建仿真环境，学习相关算法，并在仿真环境中进行试验验证，尝试优化算法。

（2）2019.2-2019.4

机器人机械结构的选择和硬件电路的设计，实现每个关节简单的控制任务。

（3）2019.4-2019.5

搭建真实场景下的测试环境搭建，实现机器人状态检测等相关功能。

（4）2019.5-2019.8

实现机器人在真实环境下的学习，反复结合仿真研究算法。

（5）2019.8-2019.12

总结前期所做的工作，撰写论文。

5.2. 经费预算

本课题的经费包括机器人机械结构、硬件电路、测试环境中的摄像头和一些复位装置所用机械结构。

具体的经费预算如下表所示

项目	预算	备注
机器人机械结构套件	300 元	
舵机	1700 元	单价 100 元*17 个
硬件电路	500 元	
摄像头	600 元	单价 200 元*3 个
通用机械结构组件	200 元	
合计	3300 元	

六、参考文献

[1] Bryson B, Roberts W. A short history of nearly everything[M]. New York: Broadway Books, 2003.

[2] 蔡自兴. 机器人学[M]. 北京: 清华大学出版社, 2000.

[3] 张秀丽. 四足机器人节律运动及环境适应性的生物控制研究[D]. 北京: 清华大学, 2004.

[4] 王立鹏. 液压四足机器人驱动控制与步态规划研究[D]. 北京: 北京理工大学, 2014.

[5] Sutton R S, Barto A G. Reinforcement learning: An introduction(Second edition)[M]. Cambridge: MIT press, 2018.

[6] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436.

[7] Arulkumaran K, Deisenroth M P, Brundage M, et al. A brief survey of deep reinforcement learning[J]. arXiv preprint arXiv:1708.05866, 2017.

[8] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.

[9] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529.

[10] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484.

[11] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354.

[12] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.

[13] Yu L, Zhang W, Wang J, et al. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient[C]//AAAI. 2017: 2852-2858.

[14] 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.

[15] Ha S, Kim J, Yamane K. Automated Deep Reinforcement Learning Environment for Hardware of a Modular Legged Robot[C]//2018 15th International Conference on Ubiquitous Robots (UR). IEEE, 2018: 348-354.

[16] Bongard J, Zykov V, Lipson H. Resilient machines through continuous self-modeling[J]. Science, 2006, 314(5802): 1118-1121.

[17] 谢涛, 徐建峰, 张永学, 等. 仿人机器人的研究历史, 现状及展望[J]. 机器人,

2002, 24(4): 367-374.
[18] De Santos P G, Garcia E, Estremera J. Quadrupedal locomotion: an introduction to the control of four-legged robots[M]. Springer Science & Business Media, 2007.
[19] Hashimoto S, Narita S, Kasahara H, et al. Humanoid robots in Waseda university—Hadaly-2 and WABIAN[J]. Autonomous Robots, 2002, 12(1): 25-38.
[20] McGhee R B. Vehicular legged locomotion[J]. Advances in automation and robotics, 1985, 1: 248-259.
[21] De Santos P G, Garcia E, Estremera J. Quadrupedal locomotion: an introduction to the control of four-legged robots[M]. Springer Science & Business Media, 2007.
[22] Hirose S, Kato K. Study on quadruped walking robot in tokyo institute of technology[C]//Proceedings of the 2000 IEEE International Conference on Robotics and Automation. 2000: 414-419.
[23] Zheng Y F, Shen J. Gait synthesis for the SD-2 biped robot to climb sloping surface[J]. IEEE Transactions on Robotics and Automation, 1990, 6(1): 86-96.
[24] Kajita S, Tani K. Study of dynamic biped locomotion on rugged terrain-derivation and application of the linear inverted pendulum mode[C]//Robotics and Automation, 1991. Proceedings., 1991 IEEE International Conference on. IEEE, 1991: 1405-1411.
[25] Kimura H, Fukuoka Y, Cohen A H. Adaptive dynamic walking of a quadruped robot on natural ground based on biological concepts[J]. The International Journal of Robotics Research, 2007, 26(5): 475-490.
[26] Hashimoto S, Narita S, Kasahara H, et al. Humanoid robots in Waseda university—Hadaly-2 and WABIAN[J]. Autonomous Robots, 2002, 12(1): 25-38.
[27] Akachi K, Kaneko K, Kanehira N, et al. Development of humanoid robot HRP-3P[C]//Humanoid Robots, 2005 5th IEEE-RAS International Conference on. IEEE, 2005: 50-55.
[28] Nagasaka K, Kuroki Y, Suzuki S, et al. Integrated motion control for walking, jumping and running on a small bipedal entertainment robot[C]//Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on. IEEE, 2004, 4: 3189-3194.
[29] Raibert M, Blankespoor K, Nelson G, et al. Bigdog, the rough-terrain quadruped robot[J]. IFAC Proceedings Volumes, 2008, 41(2): 10822-10825.
[30] Kim T J, So B, Kwon O, et al. The energy minimization algorithm using foot

rotation for hydraulic actuated quadruped walking robot with redundancy[C]//Robotics (ISR), 2010 41st International Symposium on and 2010 6th German Conference on Robotics (ROBOTIK). VDE, 2010: 1-6.

[31] Kuindersma S, Deits R, Fallon M, et al. Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot[J]. Autonomous Robots, 2016, 40(3): 429-455.

[32] 王强, 纪军红, 强文义, 等. 基于自适应模糊逻辑和神经网络的双足机器人控制研究[J]. 高技术通讯, 2001, 11(7): 76-78.

[33] 查望华. 双足机器人运动控制系统的研究[D]. 杭州: 浙江大学, 2016.

[34] 陈奇石. 强化学习在仿人机器人行走稳定控制上的研究及实现[D]. 广州: 华南理工大学, 2016.

[35] 陈强. 基于改进粒子群算法的仿人机器人步态多目标优化[D]. 广州: 华南理工大学, 2011.

[36] Zhang J, Gao F, Han X, et al. Trot gait design and CPG method for a quadruped robot[J]. Journal of Bionic Engineering, 2014, 11(1): 18-25.

[37] Sun L, Zhou Y, Chen W, et al. Modeling and robust control of quadruped robot[C]//Information Acquisition, 2007. ICIA'07. International Conference on. IEEE, 2007: 356-360.

[38] 余联庆. 仿马四足机器人机构分析与步态研究[D]. 武汉: 华中科技大学, 2007.

[39] 姬昌睿, 王润孝, 罗振元,等. 复杂环境下作业四足式机器人外形结构设计研究[J]. 机械设计与制造工程, 2008, 37(1):43-46.

[40] Li X, Wang W, Li B, et al. Central pattern generators based adaptive control for a quadruped robot[C]//Robotics and Biomimetics (ROBIO), 2009 IEEE International Conference on. IEEE, 2009: 2068-2072.

[41] 王立鹏, 王军政, 汪首坤,等. 基于足端轨迹规划算法的液压四足机器人步态控制策略[J]. 机械工程学报, 2013, 49(1):39-44.

[42] 荣学文. SCalf 液压驱动四足机器人的机构设计与运动分析[D]. 山东: 山东大学, 2013.

[43] Zhang T, Wei Q, Ma H. Position/force control for a single leg of a quadruped robot in an operation space[J]. International Journal of Advanced Robotic Systems, 2013, 10(2): 137.

[44] 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.

[45] Gullapalli V, Franklin J A, Benbrahim H. Acquiring robot skills via reinforcement learning[J]. IEEE Control Systems, 1994, 14(1): 13-24.

[46] Schaal S. Learning from demonstration[C]//Advances in neural information processing systems. 1997: 1040-1046.

[47] Bagnell J A, Schneider J G. Autonomous helicopter control using reinforcement learning policy search methods[C]//Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164). IEEE, 2001, 2: 1615-1620.

[48] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436.

[49] Van Hasselt H, Guez A, Silver D. Deep Reinforcement Learning with Double Q-Learning[C]//AAAI. 2016, 2: 5.

[50] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning[J]. arXiv preprint arXiv:1511.06581, 2015.

[51] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint arXiv:1509.02971, 2015.

[52] Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization[C]//International Conference on Machine Learning. 2015: 1889-1897.

[53] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning[C]//International conference on machine learning. 2016: 1928-1937.

[54] Gu S, Lillicrap T, Sutskever I, et al. Continuous deep q-learning with model-based acceleration[C]//International Conference on Machine Learning. 2016: 2829-2838.

[55] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.

[56] Wu Y, Mansimov E, Grosse R B, et al. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation[C]//Advances in neural information processing systems. 2017: 5279-5288.

[57] Levine S, Finn C, Darrell T, et al. End-to-end training of deep visuomotor policies[J]. The Journal of Machine Learning Research, 2016, 17(1): 1334-1373.

[58] Gu S, Holly E, Lillicrap T, et al. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates[C]//Robotics and Automation (ICRA), 2017 IEEE International Conference on. IEEE, 2017: 3389-3396.

[59] Riedmiller M, Hafner R, Lampe T, et al. Learning by Playing-Solving Sparse Reward Tasks from Scratch[J]. arXiv preprint arXiv:1802.10567, 2018.

[60] Ha S, Kim J, Yamane K. Automated Deep Reinforcement Learning Environment for Hardware of a Modular Legged Robot[C]//2018 15th International Conference on Ubiquitous Robots (UR). IEEE, 2018: 348-354.

[61] 彭自强. 基于 Q 学习和神经网络的双足机器人控制[D]. 杭州：浙江大学, 2012.

[62] 陈奇石. 强化学习在仿人机器人行走稳定控制上的研究及实现[D]. 广州：华南理工大学, 2016.

七、导师审查意见

导师签字：

年　　月　　日

八、答辩专家组对开题报告的评议：

1．对选题依据、研究思路的科学性、可行性、及创新性的评价：

2．存在的主要问题和改进措施：

3．□通过　　□不通过

专家组组长签名：

年　月　日

九、学位评定分委员会审查意见

学位评定分委员会主席签字：

年　月　日

备注：此表一式三份，研究生本人、研究生指导教师及研究生所在学院各保存一份。