Т. Ю. Ким

ПРИМЕНЕНИЕ МЕТОДА ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ ДЛЯ АНТРОПОМОРФНОГО РОБОТА «АСТРОНАВТ» В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ

В данной работе рассматривается антропоморфный робот «астронавт» как реальный робот, так и его цифровой двойник. Цифровой двойник и алгоритм обучения будут реализовываться в среде MATLAB / Simulink с использованием среды визуализации Gazebo и программного обеспечения ROS. Предлагается применение метода обучения с подкреплением для робота в режиме реального времени с использованием высокоскоростного интернета. В качестве Агентов обучения были выбраны TD3 (Twin Delayed Deep Deterministic Policy Gradients) и DDPG (Deep Deterministic Policy Gradient). В качестве функции вознаграждения принимается сохранение вертикального положения в каждый момент времени.

Введение

Робототехника играет особенную и немаловажную роль в освоении ближнего и дальнего космоса с использованием специализированной роботизированной техники. Особое внимание следует уделить направлению антропоморфных роботов, будучи предназначенные непосредственно для замены человека в опасных ситуациях: невесомость, перепады температур, радиация и т. д. Антропоморфный робот имеет физические характеристики, свойственные человеку – наличие аналогов головы, пары рук и ног. Подобный класс роботов будет востребован при восстановлении космических аппаратов, а также сможет обеспечить подготовку условий для заселения человека, в том числе и освоения космического пространства [1].

I. Цель исследования

Данная работа планирует разработку реального антропоморфного робота-астронавта, и его цифрового аналога. Целью робота является обучение / повторение / выполнение действий, соответствующие человеческим, таких как преодоление неровных поверхностей (утесов, каньонов или скользкого льда) с минимальным усилением управления, подъем по лестнице, открытие люков, манипуляции с дверными ручками.

II. Постановка задачи

Для того, чтобы реализовать вышеуказанные задачи предполагается использовать метод глубокого обучения с подкреплением. Данный метод позволяет обучать агента в режиме реального времени, основываясь на собственном предыдущем опыте и на основе высокоточных (фотографических) наблюдений, больших данных об изменении окружающей среды и высокоскоростного моделирования [2].

Робот будет оснащен датчиками, камерами и множествами двигателями для передвижения и для выполнения различных действий с помощью верхней частью корпуса. Предполагается, что для ориентации и навигации в окружающей среде, и распознавания объектов робот будет ориентироваться на компьютерное зрение. Поэтому необходимо оценивать местность через распознавание объектов и ориентирование в пространстве в режиме реального времени, что в свою очередь, подразумевает эффект телеприсутствия (технология телеприсутствия), при условии, что только человек может повторить поведение робота [3].

Далее следует этап формирования функции вознаграждения. Вознаграждение состоит из основного компонента, пропорционально скорости по горизонтальной плоскости, побуждающего агента продвигаться вперед по заданной траектории, а также небольшого количества штрафных крутящих моментов, также робот-астронавт получает дополнительную награду за каждое отклонение во временных шагах, чтобы предотвратить падение. «Тем самым побуждая робота двигаться вперед, предоставляя положительное вознаграждение за положительную скорость движения вперед [1]».

Далее следует выбрать Агента для обучения. Выбор будет стоять между DDPG и TD3, где оба представителя имеют достаточно хорошие показатели. Агент DDPG аппроксимирует долгосрочное вознаграждение с учетом наблюдений и действий, используя представление функции критического значения. Агент DDPG решает, какое действие следует предпринять для данных наблюдений, используя представление актора [4]. Структура сетей акторов и критиков, используемых для агента TD3, такая же, как и для агента DDPG. Агент DDPG может переоценить значение Q, поскольку агент использует значение Q для обновления своей политики (актора). Результирующая политика может быть неоптимальной, а накопление ошибок обучения может привести к сходящемуся поведению. Алгоритм TD3 является расширением DDPG с улучшениями, которые делают его более надежным,

[1] https://www.mathworks.com/help/reinforcement-learning/ug/train-biped-robot-to-walk-using-reinforcement-learning-agents.html

предотвращая переоценку значений Q. Оба агента были применены ранее в работах [5,6] и показали хорошие результаты, где нашли оптимальные рабочие коэффициенты для поставленной для мобильного робота задачи. Агент обучается на основе полученных данных с датчиков на основе которого формирует действие для антропоморфного робота.

III. Выводы

Для реализации данного робота-астронавта планируется использовать следующие инструменты: Gazebo (симулятор – средство верификации); ROS (среда – разработка программного обеспечения); MATLAB/Simulink (среда разработки системы управления / разработка мехатроники).

Для реализации вышеуказанных задач необходим высокоскоростной интернет, который будет совершать сбор данных от датчиков реального робота и отправлять цифровому двойнику на базу.

Список литературы

1. Ярмолик, В. Н. Физически неклонируемые функции / В. Н. Яролик, Ю. Г. Вашинко // Информатика. – 2011. – №2. – С. 20-30.
2. Богданов, А. А. Космический эксперимент с антропоморфным роботом / А. А. Богданов, И. М. Кутлубаев, А. Ф. Пермяков. // Решетневские чтения. – 2017. – №21-1. URL: https://cyberleninka.ru/article/n/kosmicheskiy-eksperiment-s-antropomorfnym-robotom (дата обращения: 01.04.2022).
3. Клюшников, В. Ю. Робот-аватар – средство телеприсутствия человека в космосе / В. Ю. Клюшников, С. А. Родькина // ВКС. – 2020. – №1 (102). URL: https://cyberleninka.ru/article/n/robot-avatar-sredstvo-teleprisutstviya-cheloveka-v-kosmose (дата обращения: 30.03.2022).
4. Lillicrap, T. P. Continuous control with deep reinforcement learning / T. P. Lillicrap, J. J. Hunt, A. Pritzel [et al.] // Google Deepmind. – 2016. – №6. – P. 1-14. URL: https://arxiv.org/abs/1509.02971. (date of access: 30.03.2022). DOI: 10.48550/arXiv.1509.02971
5. Fujimoto, S. Addressing Function Approximation Error in Actor-Critic Methods / S. Fujimoto, Herke van Hoof, D. Meger // Artificial Intelligence (cs.AI). – 2018. – №3. – P. 1-15. DOI: 10.48550/arXiv.1802.09477
6. Tatyana, K., Prakapovich, R. Automatic Tuning of the Motion Control System of a Mobile Robot Along a Trajectory Based on the Reinforcement Learning Method / T. Kim, R. Prakapovich // In: Tuzikov, A. V., Belotserkovsky, A. M., Lukashevich, M. M. (eds) Pattern Recognition and Information Processing. PRIP 2021. Communications in Computer and Information Science, vol 1562. Springer, Cham. DOI: 10.1007/978-3-030-98883-8_17
7. Ким, Т. Ю. Применение алгоритма DDPG обучения с подкреплением для мобильного робота // II Международной научно-практической конференции Минск / Компьютерные технологии и анализ данных (CTDA'2022) // приняли в печать.

Ким Татьяна Юрьевна, аспирантка, лаборатория робототехнических систем, Объединенный институт проблем информатики Национальной академии наук Беларуси, tatyana_kim92@mail.ru.

Научный руководитель: Прокопович Григорий Александрович, заведующий лабораторией робототехнических систем, Объединенный институт проблем информатики Национальной академии наук Беларуси, кандидат технических наук, доцент, prakapovich@newman.bas-net.by.