ПРИМЕНЕНИЕ АЛГОРИТМА DDPG ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ ДЛЯ МОБИЛЬНОГО РОБОТА

Т. Ю. Ким

*Лаборатория робототехнических систем, Объединенный институт
проблем информатики Национальной академии наук Беларуси, Минск,
Республика Беларусь
E-mail: tatyana_kim92@mail.ru*

В данной работе предлагается использование алгоритма, основанный на глубоком обучении с подкреплением на уровне управления разработанного цифрового двойника мобильного робота, который следует по цветоконтрастной линии. В частности, алгоритмы глубокого обучения с подкреплением, такие как TD3 (Twin Delayed Deep Deterministic Policy Gradients) и DDPG (Deep Deterministic Policy Gradient), реализованы для сравнения результатов. Результатом данной работы является обученная модель с рабочими значениями коэффициентов для ПИД-регулятора, способная подавать сигнал мобильному роботу для движения с минимальным отклонением от цветоконтрастной линии при заданной скорости, построенная с применением алгоритма DDPG для более быстрого обучения.

Ключевые слова: *система управления, ПИД-регулятор, имитационная модель, моделирование, цифровой двойник, MATLAB/Simulink, обучение с подкреплением, TD3, DDPG.*

Введение. В настоящее время на производстве помимо классических роботов-манипуляторов весьма востребованными являются мобильные роботы в виде роботизированных тележек [1, 2]. Для системы управления подобных мобильных роботов в большей степени предпочтение отдается ПИД-регуляторам [3]. Порой подбор коэффициентов регулятора является достаточно долгим процессом, который не всегда приводит к успеху, так как есть вероятность наступления перерегулирования [4].

Алгоритм DDPG, предполагаемый как вариант решения данной проблемы, совмещает детерминированный градиент политики и глубокую Q-сеть.

В настоящей работе рассматривается применение алгоритма TD3, с помощью которого были определены рабочие коэффициенты для ПИД-регулятора [1, 2], и алгоритма DDPG для разработанного цифрового двойника мобильного робота, целью которого является держать минимальное расстояние между мобильным роботом и цветоконтрастной линией со скоростью 1 м/с.

Описание мобильного робота. В качестве модели рассматривается разработанный цифровой двойник мобильного робота RoboCoke [5]. Полученные рабочие значения коэффициентов для ПИД-регулятора помогут мобильному роботу двигаться без человеческого вмешательства по

заданной линии. Данный способ, применяемый для автоматизации систем управления, в последнее десятилетие обрел обширную популярность [6].

Для данного эксперимента были применены 3 датчика света, которые предоставляют данные в виде матрицы 9×9 .

В качестве вознаграждения была разработана целевая функция [7]. Вознаграждение формируется в два этапа: минимизация расстояния между роботом и цветоконтрастной линией, вычисление угла между касательной к эллипсу и направляющей мобильного робота [1, 2].

Рассматриваемые алгоритмы. Линейная скорость мобильного робота – 1 м/с – является постоянной, а сигнал с ПИД-регулятора выдает значения, обработанные после получения результатов с датчиков. При подобном автономном управлении действия должны быть непрерывными, следовательно, приводящими к низкой производительности для алгоритмов, основанных на полученных значениях. В данной работе предлагается применить алгоритмы с детерминированными акторами, используемыми для непрерывных действий. Преимуществом данных алгоритмов заключается в обеспечении прямого отображения наблюдений на значения действий детерминированных субъектов. В данных субъектах заложен алгоритм политики детерминированного градиента [8]. Далее рассмотрим 2 алгоритма:

1. Алгоритм TD3 характеризуется тем, что помимо Агента, предлагающего конкретные действия актора в ответ на определённые показания сенсорной системы, используются две новые сущности – два критика, формирующие долгосрочное вознаграждение. В работе [2] алгоритмом TD3 были найдены рабочие коэффициенты.

2. Алгоритм DDPG сочетает в себе идеи DPG (детерминированный градиент политики), акторов-критиков и DQN (глубокая Q-сеть). Преимущество данного алгоритма (DDPG) основывается на том, что Агент, используя нейронные сети, позволяет уменьшать время поиска значений коэффициентов для мобильного робота при заданной скорости 1 м/с. Алгоритм DDPG в основном следует алгоритму DPG, за исключением того, что аппроксимация функций как для актора, так и для критика, представлена глубокими нейронными сетями [8]. В качестве входных данных мы решили взять входные данные датчиков света.

Агент состоит из двух сетей: сеть актора и сеть критика. Сеть актора являющаяся политикой, принимает состояние от окружающей среды и полагаясь на ее основу формирует действие. Сеть критика также принимает наблюдение от окружающей среды и действие, сформированное актором. На основе действия критик оценивает ценность следования текущей политике в данной ситуации. Следующим действием агента, влия-

ющим на окружающую среду, а среда, в свою очередь, предоставляет агенту вознаграждение, является наблюдение за новым состоянием среды. Две оценочные стадии – исходная и обновленная – используются алгоритмом обучения для обновления как актора, так и критика. Критик обновляется, чтобы дать оценку и сравнить ее с фактическим вознаграждением. Чтобы достичь состояния с более высокими значениями, актору необходимо обновление для выполнения действий.

Проведенное обучение и заключительный эксперимент. Обучение проходило на базе цифрового двойника мобильного робота в среде MATLAB/Simulink. Разработанный цифровой двойник применялся в качестве среды для агента. Обучение происходило на компьютере с ОС Windows 10, с применением Rapid Accelerator на 6-ядерном микропроцессоре Intel(R) Core (TM) i5-9600KF и с 32 ГБ оперативной памяти.

Цифровой двойник был обучен в MATLAB/Simulink. Характерными отличия гиперпараметров для DDPG являются скорость обучения (learning rate) – 0,001 и дисконтирующий фактор (Discount Factor) – 1.

Результаты обучения. Полученные в [2] рабочие коэффициенты алгоритмов TD3 и DDPG рассмотренные в данной работе, показывают, что оба алгоритма достигают поставленной цели – скорости движения 1м/с. Результаты сравнения работы алгоритмов TD3 и DDPG, представленные в таблице ниже, показывают, что алгоритм DDPG имеет более высокую скорость сходимости при минимальном количестве особей.

Таблица

Результаты обучения с подкреплением алгоритмами TD3 и DDPG

П (пропорциональный)	I (интегральный)	D (дифференциальный)	АЛГО-РИТМ обучения	Количество особей	Пройденное время, мс
1.359	1.406	2.039	TD3	500	41
3.676	-0.159	-4.660		350	48
0.090	0.008	0.769		500	43
0.079	0.511	0.163	DDPG	390	42
0.050	0.086	0.456		420	40.5
0.212	0.040	0.358		340	41.3

Заключение. В данной работе рассматривались два основных алгоритма обучения с подкреплением: алгоритм TD3 и DDPG. Результаты исследования показали, что с помощью алгоритма DDPG Агент достигает полученных целей за более короткое время, чем с применением алгоритма TD3. В результате, настроенный ПИД-регулятор позволяет мобильному роботу безошибочно двигаться вдоль криволинейной цветовой контрастной линии со скоростью 1 м/с.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Kim T.Yu. Prakupovich R. A. Employing of RL technology to develop an adaptive motion controller for a line follower robot // 15th International conference on pattern recognition and information processing. Artificial intelligence: facing the challenges, 2021. С. 159-163.
2. Tatyana K., Prakupovich R. Automatic Tuning of the Motion Control System of a Mobile Robot Along a Trajectory Based on the Reinforcement Learning Method // 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
3. Филиппов А. В., Косолапов М. А., Маслов И. А., Тарасова Г. И. Автоматизированная настройка ПИД-регулятора для объекта управления следящей системы с использованием программного пакета Matlab Simulink // Наука, техника и образование, № 12 (18), 2015. С. 53-59.
4. Martins F. G. Tuning PID controllers using the ITAE criterion // International Journal of Engineering Education 21(5). Int. J. Engng Ed. Vol. 21, No. 5, 2005. P. 867-873.
5. Ким Т. Ю., Прокопович Г. А. Разработка цифрового двойника мобильного робота для исследовательских и учебных целей на базе MATLAB/SIMULINK // XVIII Международная научная конференция молодых ученых «Молодежь в науке – 2.0'21», 2021. С. 580-584.
6. Chan C-Y. Advancements, prospects, and impacts of automated driving systems // California partners for advanced transportation technology program, university of California at Berkeley, 2017. С. 208-216. DOI: 10.1016/j.ijst.2017.07.008
7. Ким Т. Ю., Прокопович Г. А. Оптимизация коэффициентов ПИД-регулятора системы управления движением по цветоконтрастной линии мобильного робота на основе генетических алгоритмов // Журнал “Информатика”, 2021. С. 53-68.
8. Wang S., Jia D., Weng X. Deep Reinforcement Learning for autonomous driving // CMU 10703 Deep Reinforcement Learning and control course project, 2017. - arXiv preprint, 2016