ПОДКРЕПЛЯЮЩЕЕ ОБУЧЕНИЕ В ЗАДАЧЕ УПРАВЛЕНИЯ МНОГОКОЛЕСНЫМ РОБОТОМ

Дёмин В. В., Кабыш А. С., Головко В. А.
Лаборатория СНИЛ «Робототехника», кафедра Интеллектуальных информационных технологий,
Брестский государственный технический университет
Брест, Республика Беларусь
E-mail: {spas.work, anton.kabysh}@gmail.com, gva@bstu.by

В статье рассматривается проблема эффективного управления многоколесными роботизированными платформами. Особенность разрабатываемого подхода к решению задачи энерго-эффективного управления состоит в том, что многоколесный робот рассматривается как многоагентная система, где каждый модуль представлен автономным агентом. Используя методологию многоагентных систем, разработаны адаптивные методы координации поведения агентов относительно друг друга. Для адаптивной настройки поведения модулей используется метод многоагентного обучения с подкреплением.

Введение

Эффективное управление мобильным роботом на производстве позволяет экономить множество ресурсов: время автономной работы, дальность перевозки грузов, грузоподъемность, маневренность при перевозке габаритных грузов в ограниченном пространстве. Важными задачами являются оптимизация энергопотребления и оптимальное планирование траектории.

Оптимальное планирование траектории современных систем управления, как правило, реализуется на уровне подсистем планирования [1], [2]. Подсистема такого типа строит траекторию до цели и разбивает ее на части, которые могут быть представлены в виде кривых определенного радиуса и прямолинейных промежутков. Система управления роботом позволяет передвигаться (по возможности без остановок) по этой траектории, затрачивая как можно меньше энергии батарей.

В данной работе рассматривается разработка интеллектуального метода эффективного управления производственным роботом, разработанным в лаборатории университета Равенсбург-Вайнгартен. Робот изображен на рисунке 1. Данная платформа построена на основе инновационных модулей с низким энергопотреблением [3]. Интеллектуальная система управления основана на методах мульти-агентных систем и обучения с подкреплением.

I. Интеллектуальная система управления модулей

Интеллектуальная система управления построена на основе обучения с подкреплением и решает две задачи: (1) позиционирует модули относительно точки вращения и (2) координирует согласованное движение модулей.

Обучение с подкреплением является методом обучения оптимальному управлению автономных агентов в неизвестной среде [4]. Используя Q-learning правило, ошибка временной разности между двумя последующими состояниями агентов вычисляется по следующей формуле:

$$\delta^t = r^t - \gamma \max_{a \in A(s^{t+1})} Q(s^{t+1}, a) - Q(s^t, a^t). \quad (1)$$

где r^t – значение награды полученное за выбор действия a^t в состоянии s^t, γ – коэффициент обесценивания отдаленных ценностей, $Q(s^t, a^t)$ – ценность выбора действия a^t в состоянии s^{t+1}. После каждого временного шага, ценность прошлого состояния корректируется согласно ошибке временной разницы:

$$Q(s^t, a^t) = Q(s^t, a^t) + \alpha\delta^t. \quad (2)$$

Рис. 1 – Роботизированная мобильная платформа на производственных испытаниях

Обучение агента позиционированию означает положительное подкрепление тех действий, которые минимизируют угол φ_{err}. Благодаря обучению и обобщению, агент способен поддерживать значение угла при больших отклонениях, позиционироваться относительно любых углов, даже если они динамически изменяются с течением времени движения. Для координации используется многоагентное расширение обучения с подкреплением. Основные аспекты подхода изложены в статьях [5–6]. Идея разработанного подхода заключается в использовании значения влияния для координации между модулями и виртуальным лидером платформы. Целью является определение последовательности правиль-

ных действий. Правильное влияние должно награждать, отрицательное – наказывать. Проблемой проектирования является определение таких влияний в рамках полученных индивидуальных наград. Архитектура подкрепляющего обучения, решающая задачу кооперативного движения, изображена на рисунке 2.

Модуль i, находясь в состоянии ${s_i}^t$, выбирает действие ${a_i}^t$, используя текущую стратегию выбора действий, и переходит в следующее состояние ${s_i}^{t+1}$. Платформа получает данные об изменениях после выполнения действия, вычисляет и присваивает награду ${r_i}^{t+1}$ модулю как обратную связь успешности данного действия. Схожий Q-learning алгоритм (3) может быть использован для обновления политики модуля. Главное их отличие заключается в том, что во втором случае награда назначается виртуальным лидером, вместо окружающей среды:

$$\Delta Q_i({s_i}^t, {a_i}^t) = \alpha[{r_{p\to i}}^{t+1} + \gamma \max_{a\in A({s_i}^{t+1})} Q_i({s_i}^{t+1}, a) - Q_i({s_i}^t, {a_i}^t)]. \quad (3)$$

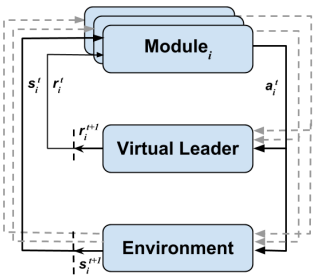

Рис. 2 – Архитектура подкрепляющего обучения для мульти-агентной системы

II. Результаты моделирования

Первый этап моделирования заключается в позиционировании модулей относительно маяка. Таким образом, они занимают правильное положение для езды по кругу. Обучение происходит один раз для одного модуля перед кооперативным этапом моделирования. Изученные правила сохраняются и копируются для других агентов. Топология Q-функции, которая обучалась в течении 720 эпох, показана на рисунке 3.

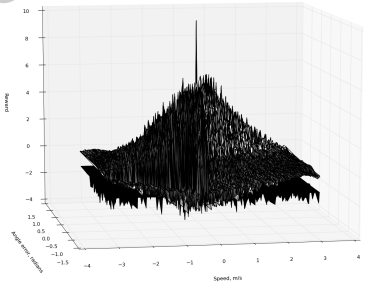

Рис. 3 – Топология Q-функции после обучения

На рисунке 4 в позиции 1 показано начальное положение модулей платформы, в позиции 2 – результат автоматического позиционирования агентов, используя обученную политику. В позиции 3 показан результат эксперимента совместного движения платформы после обучения. В среднем подготовка агентов занимает 11000 эпох.

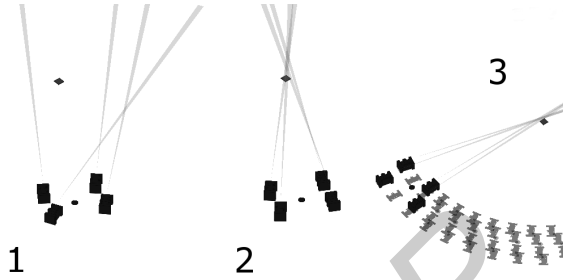

Рис. 4 – Совместные действия модулей платформы после обучениея

Внешние параметры моделирования: шаг обучения $\alpha = 0.4$, коэффициент обесценивания $\gamma = 0.7$, оптимальная скорость $\omega_{opt} = 0.8$ рад/с, угол торможения $\varphi_{stop} = 0.16$ рад.

III. Заключение

Экспериментальная часть демонстрирует успешное применение мульти-агентного подхода на основе подкрепляющего обучения для задачи эффективного управления многоколесной роботизированной платформы. Предлагаемый подход включает множество Q-learning агентов, которые определяют оптимальное управление модулями относительно виртуального лидера. Достоинства разработанного подхода заключаются в адаптивности к изменению целей и масштабируемости по количеству агентов.

IV. Список литературы

1. Balkcom, D. J. Extremal trajectories for bounded velocity differential drive robots. / D. J. Balkcom, T. M. Matthew // Robotics and Automation. Proceedings of IEEE International Conference on ICRA'00. IEEE, 2000. Vol. 3. P. 2479–2484.
2. Kaliukhovich, D. Control algorithms for the mobile robot «Max» on a task of line following provided by intelligent image processing / D. Kaliukhovich, V. Golovko, A. Paczynski // Solid state phenomena. 2009. Vol 147. P. 35–42.
3. Stetter, R. Realization and Control of a Mobile Robot / R. Stetter, P. Ziemniak, A. Pachinski // Research and Education in Robotics-EUROBOT 2010, Communication in Computer and Information Science. Springer, 2011. Vol. 156. P. 130–140.
4. Sutton, R. S. Reinforcement Learning: An Introduction / R. S. Sutton, A. G. Barto // MIT Press, 1998. 322 pages.
5. Dziomin, U. A Multi-Agent Reinforcement Learning Approach for the Efficient Control of Mobile Robot / U. Dziomin, A. Kabysh, V. Golovko, R. Stetter // Proceeding of th 7th IEEE International conference IDAACS–2013, Berlin 12–14 September 2013. – Berlin, 2013. – P. 867–774.
6. Kabysh, A. Influence Learning for Multi-Agent Systems Based on Reinforcement Learning / A. Kabysh, V. Golovko // International Journal of Computing. 2012. Vol. 11. Issue 1. P. 39–44.