Программные системы и вычислительные методы
Правильная ссылка на статью:
Кашко В.В., Олейникова С.А. Общий алгоритм ликвидации критических состояний для решения задачи управления реальным шагающим роботом на основе методов глубокого обучения с подкреплением // Программные системы и вычислительные методы. 2025. № 3. DOI: 10.7256/2454-0714.2025.3.75996 EDN: OOVYNZ URL: https://nbpublish.com/library_read_article.php?id=75996

Общий алгоритм ликвидации критических состояний для решения задачи управления реальным шагающим роботом на основе методов глубокого обучения с подкреплением

Кашко Василий Васильевич

ORCID: 0009-0009-6146-9295

аспирант; кафедра автоматизированных и вычислительных систем; Воронежский государственный технический университет

394006, Россия, Воронежская обл., г. Воронеж, ул. 20-летия Октября, д. 84

vasya.kashko@mail.ru

Олейникова Светлана Александровна

ORCID: 0000-0002-0333-2313

доктор технических наук

профессор; кафедра автоматизированных и вычислительных систем; Воронежский государственный технический университет

394006, Россия, Воронежская обл., г. Воронеж, ул. 20-летия Октября, д. 84

s.a.oleynikova@gmail.com

Статья из рубрики "Базы знаний, интеллектуальные системы, экспертные системы, системы поддержки принятия решений"

DOI:

10.7256/2454-0714.2025.3.75996

EDN:

OOVYNZ

Дата направления статьи в редакцию:

24-09-2025

Аннотация: Объектом исследования является мобильный шагающий робот с двумя или

более подвижными конечностями шарнирного типа. Вводится понятие «критического состояния», при возникновении которого механизм балансирует на грани падения (но не падает) или возникает вероятность повреждения механических узлов, по причине генерации недопустимых углов сочленений. Предметом исследования является общий алгоритм ликвидации критических состояний, обеспечивающий возможность обучения агента, основанного на глубоком алгоритме обучения с подкреплением, напрямую на реальном роботе, без риска повреждения его механизмов и прерывания процесса взаимодействия с окружающей средой для восстановления устойчивого состояния. Целью данной работы является разработка общего алгоритма ликвидации критических состояний в контексте адаптивного управления шагающим роботом на основе алгоритмов глубокого обучения с подкреплением. Было произведено сравнение предлагаемого и стандартного способов применения глубокого ОП на реальном роботе. Эксперименты проводились на 6000 эпизодах, размерностью в 300 шагов каждый. Для оценки были выбраны следующие метрики качества: процент эпизодов без фактического падения, процент полностью завершённых эпизодов, максимальная длина эпизода. Формирование алгоритма основывается на понятии «критическое состояние» и использует следующие принципы и методы: метод «проб и ошибок», принцип обратной связи, удержание проекции точки центра тяжести в области многоугольника, образованного точками соприкосновения конечностей с рабочей поверхностью, что обеспечивает балансировку конструкции и позволяет определить пограничные области, в которых робот ещё устойчив. Научная новизна работы заключается в предлагаемом подходе, позволяющем интеллектуальному агенту управлять физическим роботом «напрямую», без предварительной настройки в имитационной среде с последующей реализацией переноса. Предлагаемый алгоритм не направлен на повышение производительности агента, а предназначается для обеспечения большей автономности в процессе обучения робота, непосредственно в «железе». Основная идея заключается в моментальном реагировании на возникшее критическое состояние в виде наискорейшего последовательного возврата на некоторое число шагов назад по траектории принятия решений, обеспечив агенту постоянное пребывание в стабильном безопасном состоянии. В качестве метода глубокого обучения с подкреплением был использован метод проксимальной оптимизации политики (PPO). В результате сравнительного анализа предлагаемый алгоритм продемонстрировал сто кратный прирост устойчивости механизма.

Ключевые слова:

система управления, обучение с подкреплением, глубокие нейронные сети, алгоритм, интеллектуальный агент, критическое состояние, шагающий робот, локомоторная программа, стабилизация, окружающая среда

Введение

Основная идея обучения с подкреплением (ОП) заключается в обучении агента (система управления, контроллер) достижению некоторой поставленной цели методом проб и ошибок на базе опыта, получаемого в процессе непосредственного взаимодействия с окружающей средой [1]. Взаимодействие осуществляется путём выполнения действий агентом, в ответ на которые среда продуцирует реакцию в виде сигналов вознаграждения [1 - 3]. Множество действий может быть дискретным или непрерывным, представимым в виде распределения вероятностей [1 - 3]. Среди решаемых задач

выделяются эпизодические и континуальные (непрерывные) [1]. В первом случае, агент взаимодействует со средой строго отведённое количество временных шагов, по завершении которого начинается новая итерация. Непрерывный вариант, зачастую, сводится к эпизодической форме, посредством введения некоторого искусственного значения длины эпизода - горизонта [1]. Основным принципом функционирования целенаправленного агента является выполнение поставленной задачи путём выбора действий, максимизирующих суммарное вознаграждение, именуемое доходом [1 - 3].

В настоящее время, глубокое обучение с подкреплением, являясь одной из разновидностей машинного обучения, приобрело наибольшую популярность в области робототехники, поскольку позволяет реализовать автономные самообучающиеся системы без построения сложных математических и динамических моделей управляемых объектов [4 - 6]. Одной из актуальных и наиболее сложных задач является управление локомоцией шагающего мобильного робота [7]. Несмотря на гибкость и очевидные преимущества, применение алгоритмов обучения с подкреплением в робототехнике сопряжено с множеством трудностей [8 - 16]. Одной из наиболее частых проблем является формирование действий, в процессе настройки политики агента, способных привести к поломке дорогостоящего механизма робота за счёт его падения или генерации недопустимых значений углов сочленений. По этой причине, наибольшей популярностью пользуется подход, основанный на предварительном обучении агента в имитационной среде, с последующим переносом обученной стратегии на реальный механизм [8, 9]. Поскольку имитационные среды и используемые в них модели не учитывают множество факторов реального мира, перенос сопряжён с нестабильной работой [8 - 10]. Были осуществлены различные попытки решения сложившихся проблем. К ним относятся: реализация моделирования приводов [11], применение модульного подхода в процессе обучения [12], использование рандомизации предметной области [9, 13], применение усиления обобщения политики, путём добавления шума в среду обучения [14]. Все предложенные идеи сопряжены со значительными временными затратами и наличием множества промежуточных операций. На сегодняшний день существуют несколько успешных демонстраций реализации обучения с подкреплением на реальном роботе без предварительной настройки. Но все они относятся к ограниченным областям [9, 15]. Анализ литературных источников позволил определить, что поиск подхода, обеспечивающего безопасное обучение, на основе взаимодействия робота со средой, без предварительной настройки политики агента в имитационной среде, является актуальной задачей.

1. Постановка задачи и ее особенности

Главной целью основного исследования является разработка адаптивной универсальной системы управления мобильным шагающим роботом с применением алгоритмов глубокого обучения с подкреплением для выполнения определённой локомоторной программы (например, движение прямо) в независимости от формы рельефа местности [17 - 20]. Постановка задачи выглядит следующим образом. Пусть задана некоторая окружающая среда, обладающая разнородным рельефом. В неё помещён шагающий робот, имеющий две конечности или более, состоящие из вращательных звеньев. Механизм взаимодействует со средой, получая награду за каждое действие, и накапливает полученное вознаграждение на протяжении всего процесса функционирования [17 - 20]. Необходимо разработать универсальную адаптивную систему управления шагающим роботом, которая способна автономно формировать стратегию, реализующую

поставленную локомоторную программу, путём использования полученного интерактивного опыта в независимости от формы и типа рельефа [17 - 20].

Объектом настоящего исследования является мобильный шагающий робот с двумя или более подвижными конечностями, состоящими из вращающихся шарниров, приводимых в движение путём использования сервоприводов или поступательных приводов пневматического или гидравлического типов [7]. В качестве предмета исследования выступает общий алгоритм ликвидации критических состояний, возникающих в процессе управления шагающим механизмом при выполнении некоторой локомоторной программы. Под *критическим состоянием* понимается такое, при котором возрастает вероятность повреждения механических узлов робота или потери его устойчивости. Целью данной работы является разработка общего алгоритма ликвидации критических состояний в контексте адаптивного управления шагающим роботом на основе алгоритмов обучения с подкреплением (ОП). Научная новизна заключается в предлагаемом подходе, позволяющем глубокому ОП агенту управлять физическим роботом «напрямую», без предварительной настройки в имитационной среде с последующей реализацией переноса.

2. Формализация задачи управления локомоцией шагающего робота с учётом наличия критических состояний

Поскольку рассматриваемый механизм состоит из вращательных звеньев, управление его движением представляет собой поочерёдное чередование векторов значений их углов. В результате, состояние робота в момент времени t можно представить в виде кортежа следующего вида [17 - 19]:

$$d_t = (\theta_1, \theta_2, \theta_3, \ldots, \theta_N), \quad (1)$$

где N - количество сочленений робота, θ_i - угол i - ого сочленения, d_t - состояние робота в момент времени t.

Совокупность всех возможных кортежей d образует множество состояний робота D.

С другой стороны, система управления шагающей платформой, в контексте планирования и реализации движения, дополнительно должна решать задачу удержания равновесия на каждом шаге. Исходя из этого, управление роботом, вне зависимости от количества имеющихся у него конечностей, можно свести к классической проблеме стабилизации перевёрнутого маятника [17 - 19]. В результате, среди состояний робота возникают два подмножества. Первое G - *безопасные*, и второе B - *неустойчивые*, приводящие к падению механизма, такие, что:

$$D = G \bigcup B \quad (2)$$

Как ранее упоминалось, каждое действие характеризуется вознаграждением r со стороны окружающей среды, представляющим её реакцию на оказанное воздействие [1]. Так же, результатом выбора агента (системы управления) является факт падения или удержания равновесия, представимый в виде флага f, принимающего значения из множества $F: \{0,1\}$. В результате состояние окружающей среды в момент времени t имеет следующий вид [17 - 19]:

$$s_t = (d_t, f_t, r_t), \quad (3)$$

где d_t - состояние робота в момент времени t, f_t - флаг удержания равновесия в момент времени t, r_t - вознаграждение, полученное в момент времени t.

Совокупность всех возможных кортежей s образует множество всех состояний окружающей среды S.

Исходя из того факта, что позиционирование конечностей робота в момент времени t можно представить в виде картежа из формулы (1), становится возможным определение множества всех допустимых действий механизма. Пусть Δ - некоторая константная добавочная величина приращения угла. Представив состояние робота в виде вектора, длины N, в момент времени t, действием будет являться прибавление или вычитание Δ к углу i - ого сочленения. Так же возможен случай «холостого хода», при котором текущее позиционирование не изменяется в следующий момент времени $t+1$. В результате множество всех возможных действий робота можно представить в виде совокупности векторов, длины N:

$$A: \left\{ \begin{bmatrix} \Delta \\ 0 \\ \dots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \Delta \\ \dots \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ \dots \\ \Delta \end{bmatrix}, \begin{bmatrix} -\Delta \\ 0 \\ \dots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -\Delta \\ \dots \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ \dots \\ -\Delta \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix} \right\}$$
$$|A| = 2N+1 \quad , (4)$$

где N - количество сочленений робота.

Знак, расположенный рядом с Δ в (4), показывает тип арифметической операции. В результате, состояние робота в момент времени $t+1$ имеет следующий вид:

$$d_{t+1} = d_t + a_t, \quad (5)$$

где a_t - выбранное действие в момент времени t, причём $a_t \in A$.

Задача управления шагающим роботом представляет собой непрерывный процесс принятия решений. В классической теории обучения с подкреплением в таких случаях используют некоторую величину T, называемую горизонтом, предназначенную для искусственного разделения процесса взаимодействия механизма со средой на эпизоды [1]. В контексте решаемой задачи, эпизод считается завершённым, если робот стабильно отработал T временных интервалов, потерял равновесие до достижения соответствующего горизонта или агентом была сгенерирована такая конфигурация сочленений, которая повлекла столкновение близлежащих конечностей робота. Исходя из этого, критические состояния можно подразделить на: *несбалансированные* и *травмирующие*. Среди возможных шагающих конфигураций, наибольшим количеством несбалансированных состояний обладает конструкция с двумя конечностями. С ростом числа педипуляторов (ног) возрастает устойчивость платформы, но тем сложнее становится определение факта её падения. При этом вероятность возникновения травмирующих ситуаций растёт пропорционально количеству конечностей. Для определения критических состояний предполагается использование специального аналитического блока. В процессе его реализации необходимо решить две основные задачи: определение факта столкновения конечностей или частей конечности между собой и установление критериев удержания сбалансированного положения во время движения. В биологических системах любая полученная травма сопровождается болевым ощущением, величина которого пропорционально зависит от масштаба

повреждений и нанесённого урона организму. С другой стороны, она выступает в качестве обратной связи, ограничивающей движения конечностей в рамках допустимых положений. Для определения травмирующих состояний необходимо обеспечить робота аналогом тактильной обратной связи, возникающей от механического воздействия. В настоящее время ведутся исследования, связанные с созданием различных тактильных датчиков и искусственных аналогов кожных покровов, обеспечивающих роботов способностью к осязанию [21 - 23]. В контексте основной задачи необходимо фиксировать непосредственный факт касания частей конечности или соседних конечностей. Поэтому сенсорная система осязания упрощена до уровня использования матриц кнопок-концевиков, закреплённых на бедрах и голенях, обеспечивающих фиксацию касания и последующую генерацию сенсорного сигнала, символизирующего болевое ощущение.

Основным критерием удержания равновесия шагающей платформы, вне зависимости от количества конечностей, выступает расположение проекции точки центра тяжести на поверхность земли внутри многоугольника, образованного путём соединения точек соприкосновения педипуляторов с рабочей поверхностью [24]. При этом учитываются форма и размер ступней. Необходимо, чтобы в процессе движения, проекция центра тяжести не выходила за пределы границ соответствующей безопасной области. В противном случае происходит опрокидывание корпуса. Исходя из этого, к несбалансированным критическим состояниям относятся такие, которые обеспечивают нахождение проекции центра тяжести в некоторой подобласти, входящей в безопасную область, близко расположенной к линии границы многоугольника. В результате, робот, находясь в соответствующем состоянии, находится на грани падения, но по-прежнему удерживает равновесие корпуса. Это позволяет исключить необходимость в применении третьих сторон в процессе обучения для восстановления механизма в рабочее состояние.

В связи с тем, что помимо вероятности падения необходимо учитывать так же возможность травматизации механических узлов, кортеж состояния окружающей среды примет следующий вид:

$$s_t = (d_t, c_t, r_t), \quad (6)$$

где d_t - состояние робота в момент времени t, c_t - флаг возникновения критического состояния в момент времени t, r_t - вознаграждение, полученное в момент времени t.

3. Общий алгоритм ликвидации критических состояний

Основная идея предлагаемого в данной работе алгоритма заключается в моментальном реагировании на возникшее критическое состояние в виде наискорейшего последовательного возврата на k шагов назад по траектории принятия решений, обеспечив агенту пребывание в стабильном безопасном состоянии. Состояния, входящие в подмножество критических, определяются как фиксирующие факт касания соседних конечностей (либо частей конечности) или находящиеся на грани падения, но не допускающие его. Их достижение резко прерывает операцию выбора следующего действия. Поскольку существует необходимость в предотвращении падения или повреждения робота, вводится некоторое фиксированное число шагов k, на которое необходимо обернуть вспять текущее состояние механизма. Использование глубокого обучения с подкреплением подразумевает наличие буфера, накапливающего опыт, полученный в процессе взаимодействия со средой [2, 3]. Благодаря этому можно с лёгкостью получить предыдущее состояние механизма на шаге $t-k$. После возврата в

d_{t-k}, оно рассматривается в качестве стартового и счётчик интервалов времени для определения достижения горизонта T сбрасывается до начального значения. Общий алгоритм ликвидации критических состояний в процессе решения задачи управления локомоцией шагающего робота представлен на рисунке 1.

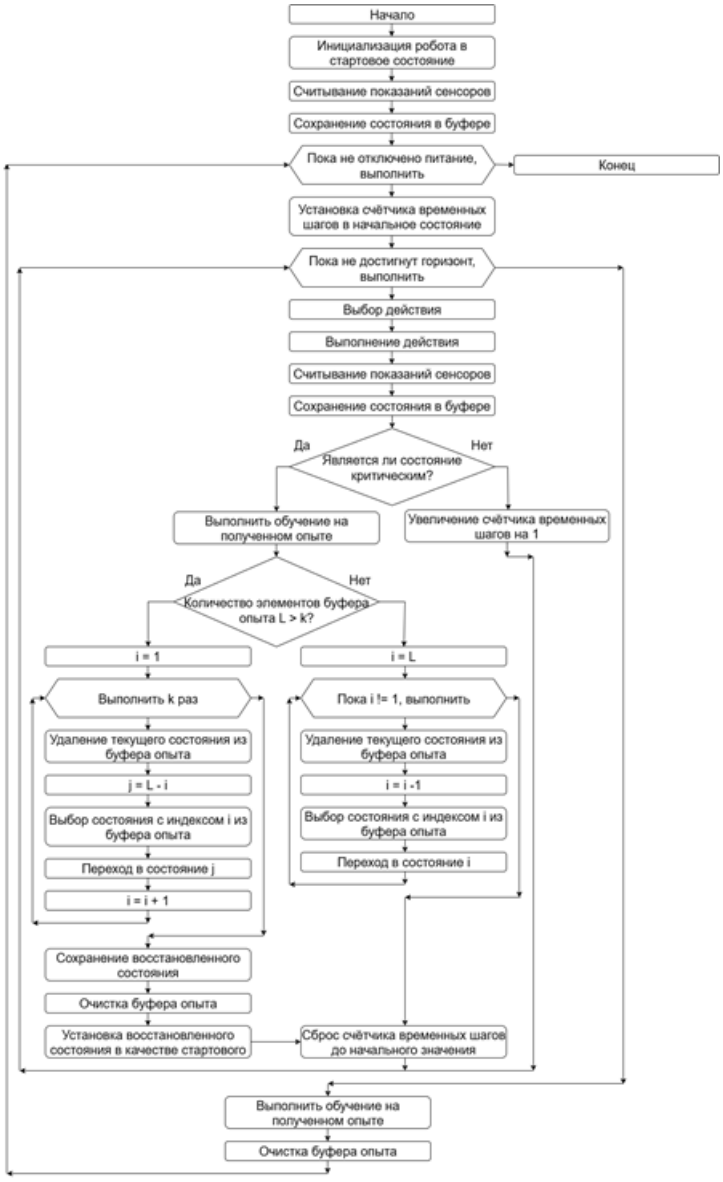

Рисунок 1 – Общий алгоритм ликвидации критических состояний в процессе решения задачи управления локомоцией шагающего робота

Алгоритм является общим, поскольку в нём не отображены детали процесса обучения и реализации операции выбора действия, которые осуществляются с использованием глубокого алгоритма обучения с подкреплением. В качестве стартового состояния выступает стабильная и безопасная конфигурация углов сочленений робота, из которого производится дальнейший выбор альтернативы с последующим выполнением перехода в новое состояние. В случае возникновения критического состояния, если количество элементов, содержащихся в буфере на текущий момент времени t меньше k, то переход осуществляется в стартовое, характерное для соответствующего эпизода. Иначе, производится реконфигурация до состояния d_{t-k}. Важно отметить, что переход на k шагов назад реализуется постепенно, путём перебора всех предыдущих состояний, начиная с пограничного. Необходимо обязательно выполнить обучение до момента реконфигурации, поскольку алгоритму ОП следует запомнить все действия, приведшие к

возникновению критического состояния. После, буфер передаёт данные в блок обработки действий по принципу стека, удаляя при этом информацию, до момента достижения d_{t-k}. Производится сохранение стабильного состояния d_{t-k}. Очищается буфер с установкой d_{t-k} в качестве новой стартовой конфигурации робота.

4. Экспериментальная часть

Для апробации предложенного алгоритма было произведено сравнение качества его функционирования с качеством стандартного способа применения метода глубокого ОП на реальном роботе. Под стандартным способом подразумевается классическая последовательность выбора и реализации действия, которая завершается в момент падения (повреждения) устройства, либо при достижении конца эпизода. В качестве робототехнической платформы была выбрана двуногая конфигурация, состоящая из небольшого торса и нижних конечностей, как наиболее неустойчивая из всех существующих шагающих платформ [25]. Для реализации сравнения, в роли метода ОП, был использован алгоритм проксимальной оптимизации стратегии (PPO) [2, 3]. В качестве локомоторной программы выступало прямолинейное движение по ровной поверхности. Для ликвидации повреждений при использовании классического подхода, производилась моментальная остановка перехода при соприкосновении конечностей между собой, как и для случая критических состояний. Такое допущение оправдано стоимостью оборудования и обеспечивает равные условия для каждого из рассматриваемых случаев. Эксперименты проводились на 6000 эпизодах, размерностью в 300 шагов каждый. Скорость обучения исполнителя и критика - 0,0001. Количество эпох на итерацию - 4. Константная добавочная величина приращения угла $\Delta = 10^{\circ}$. При падении или соприкосновении соседних конечностей (или частей конечности) агент получает отрицательное вознаграждение в размере -100. Для чистоты эксперимента логика награды распространяется и на критические состояния. За каждый стабильный и безопасный шаг вперёд генерируется положительное вознаграждение в размере 1. Направление контролируется показаниями датчика гироскопа-акселерометра. Максимальная возможная награда за эпизод равна 300 единиц. Для оценки были выбраны следующие метрики качества: *процент эпизодов без фактического падения, процент полностью завершённых эпизодов, максимальная длина эпизода*. В таблице 1 представлены полученные экспериментальные результаты.

При использовании стандартного подхода к обучению агента, применённого напрямую к аппаратной платформе робота, без предварительной настройки, возникает необходимость непосредственного вмешательства человека, поскольку требуется возвращать механизм в устойчивое состояние после факта падения. Велик риск повреждения корпуса робота и подвижных механизмов, вследствие генерации недопустимых значений углов сочленений. Предлагаемый подход обеспечивает повышение автономности, поскольку позволяет не допустить возникновение неравновесных состояний и ликвидирует возможность повреждения, поскольку в случае фиксации касания между конечностями производится моментальная остановка выполнения перехода. При его применении предполагается начало работы из стабильной и безопасной конфигурации шарниров. При этом механизм восстановления состояния на k шагов назад, по траектории принятых решений, позволяет агенту рассматривать выбор альтернатив методом проб и ошибок в контексте безопасного мониторинга.

Таблица 1 – Показатели метрик качества функционирования

Метрика	Классический	Предлагаемый
Процент эпизодов без фактического падения	0,83%	85%
Процент полностью завершённых эпизодов	0,83%	0,83%
Максимальная длина эпизода	300	300

Ранее описанный способ организации множества действий, позволяет агенту незначительно изменять конфигурацию механизма. В результате, получается «осторожный» агент, который способен фактически выполнять планирование на практике. К тому же, возврат на k шагов назад позволяет лучше произвести изучение критической в данный момент времени области графа состояний и закрепить «правильный» выбор с минимальным ущербом для устройства. Применение классического подхода предполагает завершение эпизода при возникновении непоправимого состояния, следствием которого является падение. В отличие от него, предлагаемый подход позволяет в значительной степени сузить пространство исследуемых состояний до области безопасных без фактического прерывания процесса взаимодействия робота с окружающей средой. Важно отметить, что данный алгоритм не предназначен для повышения устойчивости и производительности агента, а необходим для формирования безопасного и более автономного процесса обучения на реальном механизме.

Выводы

В контексте настоящего исследования был разработан общий алгоритм ликвидации критических состояний для решения задачи управления локомоцией шагающего робота на базе методов ОП, обеспечивающий обучение агента напрямую «в железе» без предварительной настройки стратегии в среде симуляции. Представлены его основные идеи, применяемые для предотвращения падения механизма и обеспечения стабильной и безопасной работы, на протяжении всего периода функционирования. Рассмотрены преимущества предлагаемого алгоритма и подхода к управлению физическим шагающим роботом в целом, по сравнению с классическим применением ОП напрямую в том же контексте. Сравнение было произведено на реальном двуногом шагающем роботе при выполнении прямолинейного движения по ровной поверхности. Интеллектуальный агент строился на базе проксимальной оптимизации стратегии PPO. В результате сравнительного анализа предлагаемый алгоритм продемонстрировал сто кратный прирост устойчивости механизма.

Библиография

1. Саттон, Р. С. Обучение с подкреплением: Введение. 2-е изд. : Пер. с англ. / Р. Саттон, Э. Барто. – Москва : ДМК Пресс, 2020. – 552 с. : ил. – ISBN 978-5-97060-097-9.
2. Моралес, Мигель. Грокаем глубокое обучение с подкреплением : учебное пособие / М. Моралес. – Санкт-Петербург : Питер, 2023. – 464 с. : ил. – (Серия "Библиотека программиста"). – ISBN 978-5-4461-3944-6.
3. Уиндер, Ф. Обучение с подкреплением для реальных задач / пер. с англ. – СПб.: БХВ-Петербург, 2023. – 400 с. : ил. – ISBN 978-5-9775-6885-2.
4. Ришал Харбанс. Грокаем алгоритмы искусственного интеллекта. – СПб.: Питер, 2023. –

368 с.: ил. – (Серия "Библиотека программиста"). – ISBN 978-5-4461-2924-9.
5. Кашко, В. В. Применение методов обучения с подкреплением для реализации движения шагающих роботов / В. В. Кашко, С. А. Олейникова // Современные информационные технологии. Теория и практика. – 2024. – С. 256-262. – EDN: GRDVBI.
6. Кашко, В. В. Анализ методов обучения с подкреплением для управления роботизированными системами / В. В. Кашко, С. А. Олейникова // Инновационные технологии: теория, инструменты, практика. – 2024. – Т. 1. – С. 133-140. – EDN: LTXEUX.
7. Юревич, Е. И. Основы робототехники – 4-е изд., перераб. и доп.: учебное пособие / Е. Юревич. – СПб.: БХВ-Петербург, 2017. – 304 с.: ил. – (Учебная литература для вузов). – ISBN 978-5-9775-3851-0.
8. Y. Shao, Y. Jin, X. Liu, W. He, H. Wang, and W. Yang, "Learning free gait transition for quadruped robots via phase-guided controller," IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 1230-1237, 2021.
9. X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 3803-3810.
10. Gangapurwala, S., Mitchell, A., and Hacoutis, I. Guided constrained policy optimization for dynamic quadrupedal robot locomotion. IEEE Robot. Autom. Lett. 5, 3642-3649, 2020. doi: 10.1109/LRA.2020.2979656. – EDN: ZSVETN.
11. Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., and Hutter, M. Learning agile and dynamic motor skills for legged robots. Science Robotics. 4, eaau5872, 2019. 10.1126/scirobotics.aau5872.
12. F. Zhang, J. Leitner, M. Milford, and P. Corke, "Modular deep Q networks for sim-to-real transfer of visuo-motor policies," arXiv preprint arXiv:1610.06781, 2016.
13. J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017, pp. 23-30.
14. K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman, "Quantifying generalization in reinforcement learning," arXiv preprint arXiv:1812.02341, 2018.
15. Smith, L., Kew, J., Li, T., Luu, L., Peng, X., Ha, S., Tan, J., and Levine, S. Learning and Adapting Agile Locomotion Skills by Transferring Experience. 2023. 10.48550/arXiv.2304.09834.
16. L. Han, Q. Zhu, J. Sheng, C. Zhang, T. Li, Y. Zhang, H. Zhang, Y. Liu, C. Zhou, R. Zhao et al., "Lifelike agility and play on quadrupedal robots using reinforcement learning and generative pretrained models," arXiv preprint arXiv:2308.15143, 2023.
17. Кашко, В. В. Математическая модель универсальной системы управления шагающим роботом на основе методов обучения с подкреплением / В. В. Кашко, С. А. Олейникова // Моделирование, оптимизация и информационные технологии. – 2024. – Т. 12. – № 1(44). – С. 12. – DOI: 10.26102/2310-6018/2024.44.1.025. – EDN: HBSQBB.
18. Kashko, V. V. Formalization of the task of controlling the movement of a walking robot / V. V. Kashko, S. A. Oleinikova // Anthropocentric sciences in education: challenges, transformations, resources. – 2024. – P. 342-345. – EDN: ASVCIB.
19. Кашко, В. В. Формализация задачи управления шагающим роботом на основе алгоритмов обучения с подкреплением / В. В. Кашко, С. А. Олейникова // Интеллектуальные информационные системы. Труды Международной научно-практической конференции. Воронеж. – 2025. – С. 243-247.
20. Кашко, В. В. Обобщённый алгоритм решения задачи управления шагающим роботом на базе интеллектуального агента с использованием методов глубокого обучения с подкреплением / В. В. Кашко, С. А. Олейникова // Научная опора Воронежской области.

Сборник трудов победителей конкурса научно-исследовательских работ студентов и аспирантов ВГТУ по приоритетным направлениям развития науки и технологий. Воронеж. – 2025. – С. 155-158. – EDN: OOTOMR.
21. Pestell, N., Griffith, T., Lepora, N. F. Artificial SA-I and RA-I afferents for tactile sensing of ridges and gratings. J. R. Soc. Interface. 19: 20210822, 2022. https://doi.org/10.1098/rsif.2021.0822. – EDN: QHNGNT.
22. Юревич, Е. И. Сенсорные системы в робототехнике : учеб. пособие / Е. И. Юревич. – СПб. : Изд-во Политехн. ун-та, 2013. – 100 с.
23. Lecture 5: Совместное развитие сенсорики и робототехники. [Электронный ресурс]: издание официальное. Москва : Интернет-Университет Информационных Технологий (ИНТУИТ), 2024. URL : https://intuit.ru/en/studies/courses/22789/1324/lecture/33070?page=5 – Дата публикации: 07.10.2024.
24. Самойлова, А. С. Система управления шагающим роботом, адаптивным к изменению кинематической схемы / А. С. Самойлова, С. А. Воротников // Мехатроника, автоматизация, управление. – Москва : Новые технологии, 2021. – Т. 22 : Роботы, мехатроника и робототехнические системы – № 11. – С. 601-609. – DOI: 10.17587/mau.22.601-609. – EDN: RHGNTJ.
25. Сиволобов, С. В. Математическое моделирование походки человека на основе пятизвенной модели антропоморфного механизма с использованием методов оптимизации / С. В. Сиволобов // Математическая физика и компьютерное моделирование. – 2024. – Т. 27. – № 1. – С. 62-85. – doi:10.15688/mpcm.jvolsu.2024.1.5. – EDN: AUNGTZ.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Статья посвящена разработке общего алгоритма ликвидации критических состояний в процессе управления локомоцией реального шагающего робота с применением методов глубокого обучения с подкреплением. Предметом исследования выступает проблема обеспечения устойчивости и предотвращения повреждений робототехнических механизмов при обучении агентов напрямую «в железе», без предварительной настройки в имитационной среде. Авторы формулируют задачу в строгих математических терминах, определяют класс критических состояний и предлагают алгоритм, который позволяет оперативно возвращать систему в безопасное состояние, обеспечивая непрерывность процесса обучения.
Методология исследования сочетает теоретическую формализацию задачи с экспериментальной проверкой. В работе разработан алгоритм, основанный на анализе состояний и возврате на несколько шагов назад по траектории решений, что позволяет избежать разрушительных сценариев. Экспериментальная часть выполнена на двуногом роботе с использованием метода проксимальной оптимизации стратегии, где были сопоставлены результаты классического подхода и предложенного метода. Такой комплексный подход обеспечивает научную обоснованность и практическую достоверность полученных результатов.
Актуальность исследования обусловлена ростом интереса к созданию автономных робототехнических систем, способных функционировать в реальных условиях без длительной предварительной настройки в симуляционных средах. Предлагаемый алгоритм позволяет существенно повысить уровень безопасности и автономности при обучении шагающих роботов, что делает работу значимой для развития современной

робототехники и интеллектуальных систем управления.

Научная новизна статьи заключается в предложении универсального подхода к ликвидации критических состояний, который обеспечивает безопасное взаимодействие агента с реальной аппаратной платформой. Авторы подчеркивают, что алгоритм не направлен на повышение производительности агента в классическом смысле, но создает условия для более надежного и устойчивого обучения без риска повреждения механизма. Такой взгляд расширяет существующие представления о применении методов глубокого обучения с подкреплением в робототехнике.

Стиль изложения отличается ясностью и академической строгостью. Работа имеет четко выстроенную структуру: от введения и постановки задачи к формализации, описанию алгоритма, экспериментальной проверке и выводам. Текст логичен, аргументация последовательна, результаты представлены убедительно. Библиография обширна и охватывает как фундаментальные труды в области обучения с подкреплением, так и современные исследования в робототехнике, что свидетельствует о глубокой проработке темы и опоре на актуальные источники.

Содержание статьи представляет интерес как для специалистов в области искусственного интеллекта и робототехники, так и для более широкой аудитории исследователей, работающих над проблемами автономных систем и управления сложными объектами. Авторский подход демонстрирует не только высокую исследовательскую культуру, но и стремление к практической применимости разработок.

В заключение отмечу, что статья отличается высокой степенью актуальности, оригинальностью предложенного решения и качеством его экспериментальной проверки. Работа вносит заметный вклад в развитие методов безопасного применения глубокого обучения с подкреплением в управлении реальными робототехническими системами и может быть рекомендована к публикации без доработки.