NGHIÊN CỨU TỐI ƯU BÀI TOÁN ĐỊNH VỊ BẢN ĐỒ CHO ROBOT DI ĐỘNG TRONG MÔI TRƯỜNG KHÔNG XÁC ĐỊNH SỬ DỤNG PHƯƠNG PHÁP HỌC TĂNG CƯỜNG

NAVIGATION FOR MOBILE ROBOT IN UNKNOWN ENVIRONMENT USING REINFORCEMENT LEARNING METHODS

Nguyễn Anh Tú[1,*], Nguyễn Hồng Sơn[1], Bùi Huy Anh[1], Trần Quốc Hoàn[2]

DOI: https://doi.org/10.57001/huih5804.2023.081

TÓM TẮT

Robot di động ngày càng được sử dụng rộng rãi trong nhiều lĩnh vực của nền công nghiệp 4.0. Ứng dụng robot di động có thể mang lại những hiệu quả về thời gian, tối ưu sản xuất, lợi ích kinh tế,... Tuy nhiên, để robot di động ổn định, linh hoạt trong môi trường thay đổi luôn là vấn đề được nhiều nhà khoa học quan tâm. Trong đó, bài toán điều hướng robot di động được xem là bài toán quan trọng trong lĩnh vực robot. Nhiều nghiên cứu và giải pháp kỹ thuật đã được đề xuất và thử nghiệm nhằm giải quyết vấn đề này, trong đó phương pháp học tăng cường (RL) đã thu hút được nhiều sự quan tâm vì những ưu điểm như giúp robot có khả năng tự học và ngày càng nâng cao khả năng của robot. Bài báo này đề xuất một giải pháp điều hướng cho robot di động dựa trên thuật toán Q-Learning và phương pháp tránh vật cản tự động. Bên cạnh đó, mối quan hệ và đặc điểm giữa các hành vi của robot và điều kiện môi trường cũng được phân tích. Kết quả của mô phỏng trên nền tảng Gazebo được so sánh với kết quả khi ứng dụng thuật toán SARSA để chứng minh tính hiệu quả của phương pháp đề xuất.

Từ khóa: *Mobile robot, điều hướng, học tăng cường, Q-Learning.*

ABSTRACT

Mobile robots are used in a wide range of fields in the 4.0 industry. Using mobile robots can provide such advantages as time efficiency, production optimization, and so on. However, operating mobile robots with stability and flexibility is always a big challenge and receives numerous researchers' attention. In particular, mobile robot navigation is considered a key technique in the robotics field. Many research studies and technical methods have been proposed and implemented to solve this issue, in which the reinforcement learning (RL) method has attracted considerable attention because of its ability to learn from experience and powerful adaptability. This paper proposes a navigation framework for mobile robots based on a Q-Learning algorithm and obstacle avoidance method. The relationship and characteristics between robot behaviors and environmental conditions are also analyzed. The results of the Gazebo simulation are then compared to those of SARSA to prove the efficiency of the proposed approach.

Keywords**:** *Mobile robot, navigation, reinforcement learning, Q-Learning.*

[1]Trường Đại học Công nghiệp Hà Nội

[2]Phòng Kỹ thuật Quang học và Cơ khí nghiệp vụ, Viện Khoa học và Công nghệ

[*]Email: tuna@haui.edu.vn

Ngày nhận bài: 02/3/2023

Ngày nhận bài sửa sau phản biện: 22/4/2023

Ngày chấp nhận đăng: 26/4/2023

1. GIỚI THIỆU

Điều hướng chuyển động là một nhiệm vụ quan trọng trong lĩnh vực robot di động [1 - 5]. Bài toán điều hướng nhằm xác định các đường dẫn tối ưu cho robot, tránh các loại vật cản khác nhau trong suốt quá trình di chuyển để đảm bảo robot chuyển động từ điểm xuất phát đến điểm đích trong môi trường 2D hoặc 3D. Trong hai thập kỷ qua, nhiều nghiên về bài toán điều hướng cho robot di động đã được các nhà khoa học công bố, trong đó giải pháp kết hợp các các thuật toán khác nhau giúp robot hoạt động ổn định và đạt được độ chính xác cao. Mohseni và cộng sự [6] đề xuất bộ định vị cho robot di dộng dựa trên thuật toán tối ưu hóa Cuckoo đột biến (EMCOA) và giải thuật di truyền (GA). Ánh xạ dựa trên lưới bản đồ cũng được sử dụng để tính điểm đường đi một cách hiệu quả, cho phép xác định quỹ đạo không va chạm từ vị trí ban đầu đến vị trí đích trong suốt quá trình robot di chuyển. Theo [7], một thuật toán điều hướng tích hợp trên robot di động dựa vào Template matching VO/IMU/UWB được đề xuất. Điểm nổi bật của nghiên cứu là sử dụng bộ lọc đồng thời bộ lọc Kalman và hàm sai số để giảm thiểu lỗi vị trí. Các thử nghiệm cho thấy phương pháp điều hướng tích hợp được đề xuất có thể cải thiện đáng kể độ chính xác định vị. Trong một nghiên cứu gần đây [8], nhóm tác giả đã thiết kế một thuật toán tránh chướng ngại vật được nhúng trong một bộ điều khiển cho hệ robot song phương (gồm UAV và UGV) để thực hiện nhiệm vụ tìm đường dẫn tối ưu. Bộ điều khiển này được xây dựng dựa trên mô hình cấu trúc ảo để hướng dẫn đội hình UAV-UGV tránh chướng ngại vật đồng thời nâng cao khả năng xử lý tình huống cho UGV. Để thực hiện được mục tiêu đó, các thông tin từ môi trường được thu thập từ tất cả robot và xử lý đồng thời theo thời gian thực. Bên cạnh đó, nguyên lý đồng bộ bù cũng được áp dụng để giảm lỗi theo dõi vận tốc cho hệ thống. Điều này cho phép thiết lập vận tốc hoạt động một cách linh hoạt mà vẫn khắc phục được sự xuất hiện các điểm kỳ dị về vận tốc trong suốt quá trình chuyển động của robot, kể cả việc dừng đột ngột đội hình robot khi cần thiết. Duszak và cộng

sự [9] đề xuất phương pháp hoạch định đường dẫn và xây dựng bản đồ môi trường sử dụng lưới lục giác. Dữ liệu 3D thu được bằng cách tích hợp hệ thống cảm biến và hệ thống Atena trên robot. Dữ liệu được thu thập trong phạm vi một hình chữ nhật có diện tích khoảng 400m^2. Kết quả thực nghiệm cho thấy độ chính xác của hệ thống được nâng cao trong các tác vụ điều hướng ngoài trời và cải thiện được quy hoạch đường đi cục bộ trên các địa hình gồ ghề trong thời gian thực. Qifei và cộng sự [10] nghiên cứu phương pháp nhận dạng đường dẫn điều hướng cho robot di động dựa trên thuật toán phân cụm K-mean và xử lý ảnh trên hệ thống định vị trực quan. Để giải quyết vấn đề nhiễu trên hình ảnh từ các hiệu ứng ánh sáng khác nhau, không gian màu HIS đã được áp dụng. Nhằm cải thiện độ chính xác của việc nhận dạng đường dẫn cho robot, giá trị K được khảo sát liên tục và tính chọn hợp lý để trích xuất đặc trưng từ hình ảnh. Từ đó, thông tin cần thiết được tách hoàn toàn khỏi nền của ảnh và các tham số dữ liệu được chuyển đổi nhanh chóng để thực hiện điều hướng robot.

Bài báo này đề xuất phương pháp định vị và tránh vật cản cho robot di động hoạt động trong môi trường đa vật thể dựa trên thuật toán học tăng cường. Mô hình robot di động gồm đầy đủ các thông số hình học và thông số vật lý được xây dựng trên nền tảng phần mềm Gazebo [11]. Các hoạt động huấn luyện cho mô hình để robot tự tìm đường di chuyển được thực hiện cho cả thuật toán Q-Learning [11, 12] và thuật toán SARSA [13]. Kết quả thử nghiệm được so sánh giữa hai thuật toán để đánh giá hiệu quả và chất lượng của các hoạt động huấn luyện. Phần còn lại của bài báo được cấu trúc như sau: Phần 2 trình bày cơ sở lý thuyết của các thuật toán đề xuất. Phần 3 xây dựng thông số mô phỏng và tương tác vật lý của robot trên nền tảng Gazebo. Đánh giá kết quả được chỉ ra chi tiết trong phần 4. Phần 5 là kết luận chung của bài báo.

2. CƠ SỞ LÝ THUYẾT

Phương pháp học tăng cường tập trung vào việc học hướng tới mục tiêu từ sự tương tác khác nhau. Thực thể thực hiện quá trình học tập sẽ không biết trước hành động cần phải thực hiện, thay vào đó phải tự khám phá ra hành động nào mang lại phần thưởng lớn nhất bằng cách kiểm tra các hành động này thông qua phương pháp thử sai. Các thành phần cơ bản trong học tăng cường bao gồm:

- Tác nhân (Agent): đóng vai trò trong việc giải quyết các vấn đề ra quyết định, tác động dưới sự không chắc chắn.
- Môi trường (Environment): là những gì tồn tại bên ngoài tác nhân, tiếp nhận các tác độc từ tác nhân và tạo ra phần thưởng và những quan sát.
- Hành động (Actions): tập hợp các phương thức hành động mà tác nhân tác động đến môi trường.
- Trạng thái (State): trạng thái của tác nhân sau khi tác động qua lại với môi trường.
- Phần thưởng (Reward): là giá trị thu được tương ứng với mỗi cặp Trạng thái - Hành động của tác nhân nhận được khi thực hiện tương tác với môi trường.
- Tập (Episode): một chu kỳ bao gồm các tương tác giữa tác nhân và môi trường từ thời điểm bắt đầu đến kết thúc.
- Chính sách (Policy): là hàm biểu diễn sự tương quan giữa những quan sát thu được từ môi trường và hành động cần thực hiện.

Trong đó, tác nhân và môi trường là hai thành phần cốt lõi của một mô hình học tăng cường. Hai thành phần này tương tác liên tục với nhau theo trình tự: Tác nhân thực hiện các tương tác tới môi trường thông qua các hành động, từ đó môi trường tác động lại các hành động của tác nhân. Môi trường lưu trữ các luồng thông tin khác nhau và phản hồi cho tác nhân một "giá trị khen thưởng" sau mỗi hành động của tác nhân. Giá trị này biểu hiện mức độ hiệu quả từng hành động của tác nhân trong quá trình hoàn thành nhiệm vụ. Mục đích của phương pháp học tăng cường là tác nhân tìm ra được chính sách tối đa hoá giá trị phần thưởng tích luỹ trong thời gian dài. Trong hướng tiếp cận của bài báo, tác giả chỉ ra tính hiệu quả của phương thức triển khai mô hình đề xuất dựa trên hai thuật toán học tăng cường Q-Learning và SARSA.

2.1. Thuật toán Q-Learning

Q-Learning là một thuật toán học tăng cường thực hiện phương thức cập nhật giá trị (values-based) dựa trên cập nhật hàm giá trị từ phương trình Bellman [14]. Phương trình Bellman tính toán giá trị kỳ vọng của trạng thái như sau:

$$V^*(s_t, a_t) = \max_a Q^\pi(s_t, a_t) \quad (1)$$

Trong đó: $V^*(s_t)$ là giá trị tối ưu trả về từ giá trị kỳ vọng theo trạng thái s_t theo chính sách thực hiện π; $\max Q^\pi$ là giá trị Q lớn nhất thể hiện hành động a_t tại trạng thái s_t theo chính sách π.

Phương trình tính toán giá trị Q kỳ vọng thực hiện một hành động a_t tại trạng thái s_t dựa trên phương trình Bellman:

$$Q^*(s_t, a_t) = r_t + \gamma \max_a Q^*(s_{t+1}, a) \quad (2)$$

Trong đó: $Q^*(s_t, a_t)$ là giá trị kỳ vọng của phần thưởng mà phương trình hướng đến nhằm tối ưu cho mỗi cặp trạng thái s_t và hành động a_t tại thời điểm t; r_t là phần thưởng tức thời nhận lại được tại thời điểm t; γ là hằng số chiết khấu xác định mức độ quan trọng được trao cho phần thưởng hiện tại và phần thưởng trong tương lai; $\max_a Q^*(s_{t+1}, a)$ là giá trị kỳ vọng lớn nhất có thể xảy ra của Q tại trạng thái s_{t+1} với mọi hành động a.

Q-Learning là một thuật toán Off-policy, quá trình học của mô hình chủ yếu dựa trên giá trị của chính sách tối ưu và độc lập với các hành động của chủ thể. Off-policy được định nghĩa là tác nhân tuân theo một chính sách quyết định cho việc lựa chọn hành động để đạt trạng thái s_{t+1} từ trạng thái s_t. Kể từ trạng thái s_{t+1}, tác nhân sử dụng một chính sách khác cho khâu quyết định này. Phương trình của thuật toán Q-Learning được trình bày như sau:

$$Q_{st,at} = Q_{st,at} + \alpha * [r_t + \gamma * \max Q(s_{t+1}, a) - Q_{st,at}] \quad (3)$$

$Q^*(s, a)$ trong (3) là giá trị kỳ vọng (phần thưởng của chiết khấu tích lũy trong việc thực hiện hành động a ở trạng thái s và sau đó tuân theo chính sách tối ưu. Hành động từ mỗi trạng thái thu được của thuật toán Q-Learning được xác định bởi quy trình ra quyết định Markov (MDP) [15, 16]. Các bước triển khai thuật toán được trình bày như trong bảng 1.

Bảng 1. Thuật toán cập nhật Q-Learning.

Đầu vào:

Tập trạng thái $S = \{1, 2, \ldots, s_n\}$;

Tập hành động $A = \{1, 2, \ldots, a_n\}$;

Hàm phần thưởng: $S \times A \rightarrow [\ ]$

Khởi tạo các siêu tham số của thuật toán: $\alpha, \gamma \in [0; 1]$;

Phương thức:

Khởi tạo Q: $S \times A \rightarrow [\ ]$ ngẫu nhiên;

for giá trị Q chưa hội tụ:

do: đặt trạng thái $s_t \in S$;

for s không phải là trạng thái cuối:

do:

Lựa chọn hành động a mới (dựa vào chính sách tối ưu);

Thực hiện hành động a;

Quan sát trạng thái s mới và thu nhận phần thưởng R;

Cập nhật;

$$Q_{st,at} = Q_{st,at} + \alpha^*[r_t + \gamma^* \max Q(st+1,a) - Q_{st,at}]$$

Cập nhật trạng thái $s' \leftarrow s$

End for

End for

Đầu ra:

Hành động tốt nhất được lựa chọn a'.

2.2. Thuật toán SARSA

Tương tự Q-Learning, SARSA là một thuật toán học tăng cường tuân thủ theo phương thức cập nhật Value-based và được tính toán dựa trên phương trình Bellman. Tuy nhiên, SARSA là một thuật toán On-policy. Thuật toán On-policy là thuật toán đánh giá và cải thiện cùng một chính sách *π*, hay nói cách khác tác nhân học và tuân theo một chính sách duy nhất xuyên suốt quá trình đào tạo.

SARSA là một thuật toán chỉ định rằng tại trạng thái thời điểm s_t, thực hiện hành động a_t, tiếp đó phần thưởng r_t được nhận lại và kết thúc với trạng thái s_{t+1}, đồng thời thực hiện hành động a_{t+1}. Do đó, chuỗi giá trị (s_t, a_t, r_t, s_{t+1}, a_{t+1}) đại diện cho chính tên gọi của thuật toán. Điểm khác biệt duy nhất là thành phần Q (s_{t+1}, a_{t+1}), thay vì tối đa hoá cập nhật dựa trên giá trị Q kỳ vọng cao nhất $\max Q(s_{t+1}, a)$ trong bảng giá trị kinh nghiệm như Q-Learning. SARSA được thiết lập thêm bước cập nhật hành động tại thời điểm kế tiếp. Phương trình cơ bản của thuật toán SARSA được trình bày như trong (4) và các bước triển khai thuật toán được mô tả trong bảng 2.

$$Q_{st,at} = Q_{st,at} + \alpha[r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q_{st,at}] \quad (4)$$

Bảng 2. Thuật toán cập nhật SARSA

Đầu vào:

Tập trạng thái $S = \{1, 2, \ldots, s_n\}$;

Tập hành động $A = \{1, 2, \ldots, a_n\}$;

Hàm phần thưởng: $S \times A \rightarrow [\ ]$

Khởi tạo các siêu tham số của thuật toán: $\alpha, \gamma \in [0; 1]$

Phương thức:

Khởi tạo Q: $S \times A \rightarrow [\ ]$ ngẫu nhiên;

for giá trị Q chưa hội tụ, **do:**

Đặt trạng thái $s_t \in S$;

Thực hiện hành động $a_t \in A$ (dựa vào chính sách tối ưu);

for s không phải là trạng thái cuối, **do:**

Lựa chọn hành động $a_{t+1} \in A$ (dựa vào chính sách tối ưu);

Thực hiện hành động a_{t+1};

Quan sát trạng thái s mới và thu nhận phần thưởng R;

Cập nhật;

$$Q_{st,at} = Q_{st,at} + \alpha[r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q_{st,at}]$$

Cập nhật trạng thái $s_{t+1} \leftarrow s_t$; $a_{t+1} \leftarrow a_t$;

End for

End for

Đầu ra:

Hành động tốt nhất được lựa chọn a_{t+1}.

3. XÂY DỰNG MÔ HÌNH HUẤN LUYỆN VÀ MÔ PHỎNG CHO ROBOT DI ĐỘNG

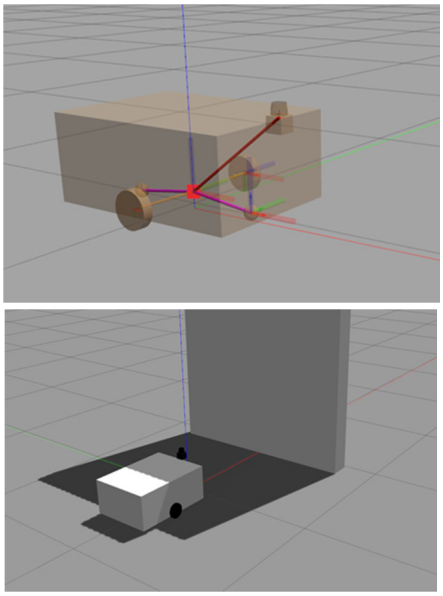
Để tiến hành thử nghiệm và đánh giá thuật toán, mô hình mô phỏng robot di dộng được xây dựng theo ba bước như sau:

Bước 1: Thiết lập môi trường trên Gazebo

Tác nhân của mô hình huấn luyện là robot di động dạng hai bánh vi sai với nhiệm vụ là thu thập dữ liệu của môi trường bằng các quan sát (observation) từ cảm biến Lidar SICK Scan NAV. Do việc tiền thiết lập cho quá trình đào tạo robot là một công đoạn hết sức quan trọng, nhóm tác giả thiết kế các bước tiền xử lý này trên môi trường mô phỏng. Môi trường mô phỏng được xây dựng trên nền tảng 3D-Gazebo. Trong đó các đặc tính và tương tác vật lý giữa robot, vật cản và điều kiện môi trường đều được định nghĩa trước.

Quá trình đào tạo tác nhân là quá trình thu thập dữ liệu và huấn luyện robot cách thực hiện nhiệm vụ nhất định nào đó với mức hiệu quả tăng dần, từ đó sử dụng bộ thông số dữ liệu thu được để đánh giá hiệu năng của các thuật

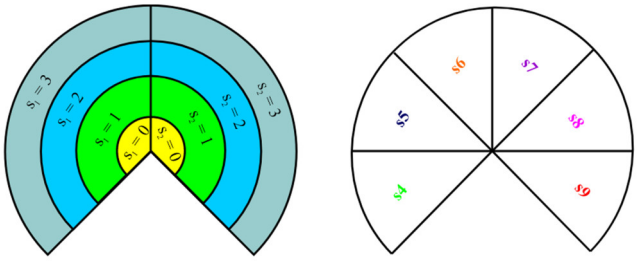
toán học tăng cường khác nhau. Trước khi thực hiện quá trình thu thập dữ liệu từ môi trường, cần khởi tạo một số thông số liên như: vị trí khởi tạo; góc quét giới hạn, độ rộng dải quét, các vùng thiết lập của lidar; vị trí bắt đầu hành động của robot…



Hình 1. Xây dựng mô hình robot trên nền tảng Gazebo

Bước 2: Thiết lập không gian trạng thái - hành động cho mô hình:

Không gian trạng thái của robot được thiết lập dựa trên dữ liệu đo từ cảm biến laser Sick NAV245. Trạng thái không an toàn được định nghĩa khi khoảng cách từ robot đến vật cản là 1m và góc quét của cảm biến laser trong phạm vi [-90°,90°] theo hướng của robot đang di chuyển. Các vật thể nằm ngoài vùng va chạm sẽ được tính là đang ở vùng an toàn. Hình 2 mô tả các khu vực khác nhau dựa trên phạm vi quét của cảm biến trên robot.



Hình 2. Không gian trạng thái của mô hình

Hình 2a mô tả không gian trạng thái về khoảng giới hạn va chạm của tác nhân. Không gian này gồm 3 mảng, trong đó: s_1 biểu thị không gian giới hạn bên trái, s_2 biểu thị không gian giới hạn bên phải; s_3 là phần tử thể hiện tổ hợp từ s_1 và s_2.

$$s_i = \begin{cases} 0, \text{ va cham} \\ 1, \text{ gan va cham} \\ 2, \text{ xuat hien vat can} \\ 3, \text{ khoang an toan} \end{cases}; \quad i = (1;2;3) \qquad (5)$$

Hình 2b mô tả không gian trạng thái giới hạn về vùng xác định vị trí vật cản theo góc, bao gồm sáu thành phần chính $s_4, s_5, s_6, s_7, s_8, s_9$ như công thức (6):

$$s_i = \begin{cases} 0, \text{ khong xuat hien vat can} \\ 1, \text{ xuat hien vat can} \end{cases}; \quad i = (4,5,6,7,8,9) \qquad (6)$$

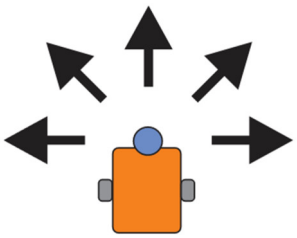
Do đó, số trường hợp trạng thái mà tác nhân quan sát từ môi trường với 9 biến trạng thái đề cập bên trên được tính như sau:

$$M = \prod_{i=1}^{k} n_{s_i} = n_{s_1}.n_{s_2}.n_{s_3}.n_{s_4}.n_{s_5}.n_{s_6}.n_{s_7}.n_{s_8}.n_{s_9} \qquad (7)$$

$$= 4.4.3.2.2.2.2.2.2 = 3072$$

Không gian hành động của tác nhân mô hình được tính toán dựa trên hai loại vận tốc chính là vận tốc tịnh tiến của robot $v \in \{0;0,35\}$ (m/s) và vận tốc quay $\omega \in \{0;0,8\}$ (rad/s). Mỗi cặp vận tốc [v, ω] tương ứng với năm đầu ra của các ước tính với giá trị Q cho từng hành động khác nhau như trong công thức (8) và hình 3.

$$action[5] = [\text{turnleft}, \text{skewleft}, \text{moveforward}, \text{skewright}, \text{turn right}] \qquad (8)$$



Hình 3. Không gian hành động của mô hình

Bước 3: Xây dựng hàm phần thưởng

Hàm phần thưởng cho mục tiêu hành động của tác nhân được định nghĩa dựa trên hai hàm phần thưởng con r_1 và r_2 như sau:

$r_1 = 200 * \text{distance\_rate}$ nếu robot đến điểm đích; (9)

$r_1 = -8,0$ nếu robot đi xa điểm đích;

$r_2 = +500$ nếu robot dừng ở điểm đích; (10)

Từ đó, giá trị tổng phần thưởng r_{sum} được tính như sau:

$r_{sum} = r_1 + r_2$ khi robot di chuyển hướng đến điểm đích; (11)

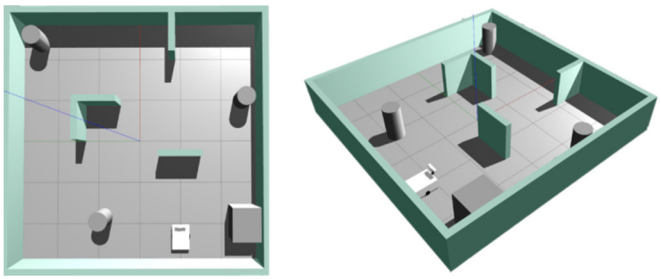
$r_{sum} = -\infty$ khi robot va chạm vật cản.

4. KẾT QUẢ MÔ PHỎNG VÀ BÀN LUẬN

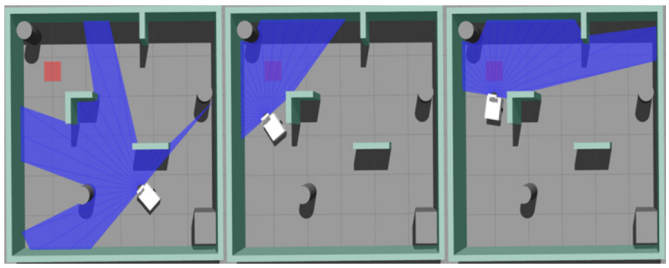
Mục tiêu của quá trình mô phỏng nhằm đánh giá mô hình và so sánh hiệu năng của hai thuật toán học tăng cường Q-Learning (Off-policy) và SARSA (On-policy) trên tác nhân là robot di động dạng hai bánh vi sai.

Tại thời điểm ban đầu, chiến lược đào tạo sẽ đặt cho robot một điểm đích bất kỳ trên bản đồ. Nhiệm vụ của robot là tuân thủ theo thuật toán học tăng cường được lựa chọn để tìm cách tới điểm đích với quãng đường ngắn nhất mà không bị va chạm vào vật cản trong quá trình di chuyển. Sau mỗi lần va chạm vật cản, robot kết thúc chu kỳ

học tập hiện thời và chương trình được tái thiết lập lại trạng thái ban đầu với các thông số học tập được cập nhật. Kinh nghiệm được tích luỹ giúp robot di chuyển ổn định hơn so với các vòng lặp trước đó.



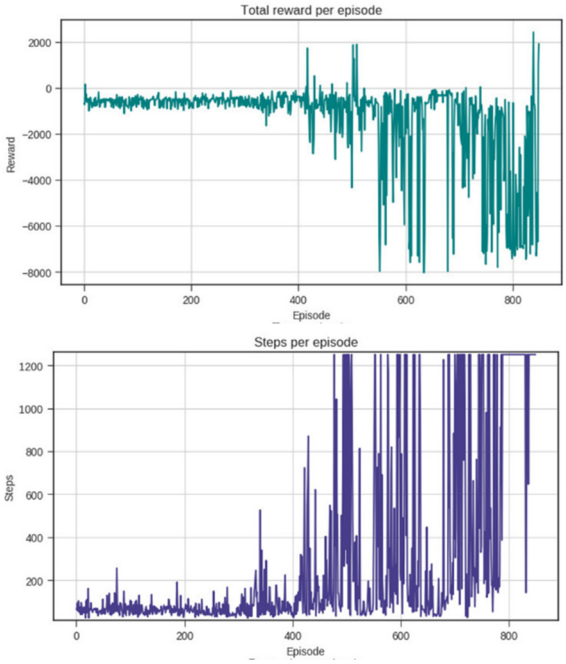
Hình 4. Môi trường mô phỏng trên Gazebo



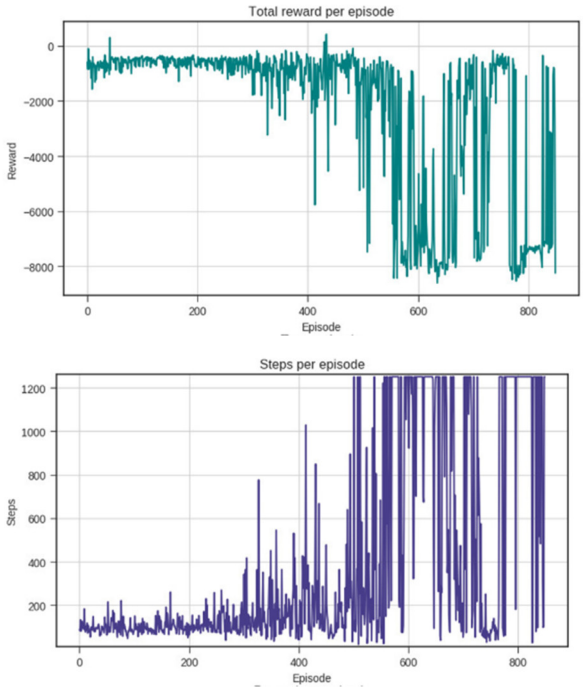
Hình 5. Robot mở rộng dần khả năng tìm được điểm đích trong quá trình huấn luyện

Sau nhiều vòng lặp chu kỳ học tập, robot tránh được vật cản trong quá trình di chuyển và tiến được đến điểm đích, thể hiện tính hội tụ của thuật toán.

Kết quả của quá trình đào tạo mô hình học tăng cường trên robot di động được thể hiện qua các thông số tổng phần thưởng tích lũy, số bước tối đa robot di chuyển trên mỗi vòng huấn luyện và phần thưởng của từng vòng huấn luyện.

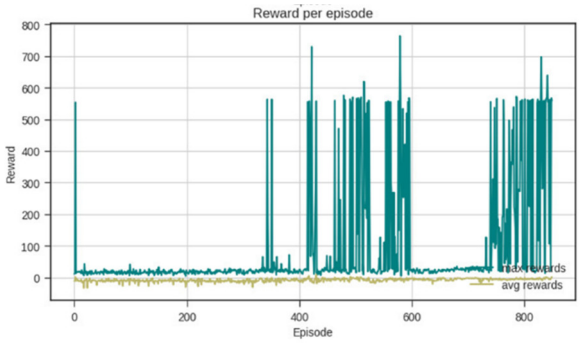


Hình 6. Kết quả mô phỏng bằng thuật toán Q-Learning

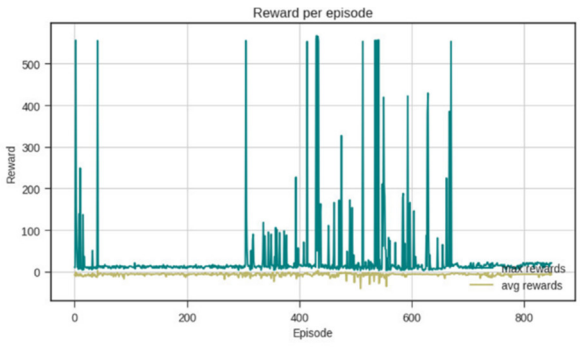


Hình 7. Kết quả mô phỏng bằng thuật toán SARSA.

Dựa vào hình 6 và 7, có thể thấy thuật toán Q-Learning cho kết quả hội tụ của chức năng tránh vật cản sau khoảng 400 episodes với số bước tăng dần, tuy nhiên giá trị phần thưởng tích luỹ chỉ ở mức tương đối ổn định. Trong khi đó, thuật toán SARSA cho kết quả tránh vật cản kém hơn với độ dao động của giá trị bước mỗi vòng huấn luyện trên đồ thị lớn.



Hình 8. Kết quả mô phỏng bằng thuật toán Q-Learning



Hình 9. Kết quả mô phỏng bằng thuật toán SARSA

Hình 8 và 9 cho thấy giá trị phần thưởng mà tác nhân robot đạt được sau mỗi vòng huấn luyện của thuật toán Q-Learning tốt hơn so với thuật toán SARSA. Giá trị phần thưởng tối đa mà Q-Learning đạt được là gần 800, cao hơn so với giá trị phần thưởng từ SARSA (giá trị đạt ở mức ngưỡng 500). Bên canh đó, đồ thị thống kê chỉ ra rằng số lần robot di chuyển đến điểm đích trong quá trình huấn luyện hay nói cách khác thuật toán đạt được giá trị phần thưởng cao khi áp dụng Q-Learning nhiều hơn so với khi áp dụng thuật toán SARSA. Điều này cho thấy thuật toán Q-Learning đáp ứng tốt hơn với các hành động rời rạc trên robot di động.

5. KẾT LUẬN

Bài báo trình bày giải pháp định vị và tránh vật cản cho robot di động trong môi trường không xác định sử dụng phương pháp học tăng cường. Phương pháp thiết lập mô hình robot di động trên phần mềm Gazebo được đề xuất cho phép thực hiện quá trình huấn luyện, mô phỏng và đánh giá hiệu quả hoạt động của các thuật toán được đề xuất. Kết quả mô phỏng cho thấy cả hai thuật toán Q-Learning và SARSA đều có khả năng giúp robot tránh được vật cản không xác định trong quá trình chuyển động và tìm được đường đi đến điểm đích, tuy nhiên hiệu quả của thuật toán Q-Learning cao hơn và độ dao động của giá trị bước mỗi vòng huấn luyện nhỏ hơn. Nghiên cứu tiếp theo nên tập trung vào việc thử nghiệm và đánh giá trên mô hình vật lý để hoàn thiện mô hình huấn luyện và nâng cao hiệu quả và độ ổn định cho robot.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Trường Đại học Công nghiệp Hà Nội trong đề tài mã số 04-2022-RD/HĐ-ĐHCN.

TÀI LIỆU THAM KHẢO

[1]. T. Kim, S. Lim, G. Shin, G. Sim, D. Yun, 2022. *An Open-Source Low-Cost Mobile Robot System With an RGB-D Camera and Efficient Real-Time Navigation Algorithm*. in IEEE Access, vol. 10, pp. 127871-127881, doi: 10.1109/ACCESS.2022.3226784.

[2]. Nguyen X. T., Bui T. L., Bui H. A., Pham D. A., Miura N., 2022. *Control the Movement of Mobile Robot Using Fingers Gestures Based on Fuzzy Logic*. In Proceedings of the International Conference on Advanced Mechanical Engineering, Automation, and Sustainable Development 2021 (AMAS2021) (pp. 799-804).

[3]. Bui T. L., Nguyen T. H., Nguyen X. T., 2023. *A Controller for Delta Parallel Robot Based on Hedge Algebras Method.* Journal of Robotics.

[4]. Nguyen D. Q., Ho V. A., 2022. *Anguilliform swimming performance of an eel-inspired soft robot.* Soft Robotics, 9(3), 425-439.

[5]. J. Lee, et al., 2022. *ODS-Bot: Mobile Robot Navigation for Outdoor Delivery Services*. in IEEE Access, vol. 10, pp. 107250-107258, doi: 10.1109/ACCESS.2022.3212768.

[6]. M. Alireza, D. Vincent, W. Tony, 2021. *Experimental study of path planning problem using EMCOA for a holonomic mobile robot*. in Journal of Systems Engineering and Electronics, vol. 32, no. 6, pp. 1450-1462, doi: 10.23919/JSEE.2021.000123.

[7]. Y. Zheng, Q. Zeng, C. Lv, H. Yu, B. Ou, 2021. *Mobile Robot Integrated Navigation Algorithm Based on Template Matching VO/IMU/UWB.* in IEEE Sensors Journal, vol. 21, no. 24, pp. 27957-27966, doi: 10.1109/JSEN.2021.3122947.

[8]. V. P. Bacheti, A. S. Brandã, M. Sarcinelli-Filho, 2021. *A Path-Following Controller for a UAV-UGV Formation Performing the Final Step of Last-Mile-Delivery*. in IEEE Access, vol. 9, pp. 142218-142231, doi: 10.1109/ACCESS.2021.3120347.

[9]. Duszak P., Siemiątkowska B., Więckowski R., 2021. *Hexagonal Grid-Based Framework for Mobile Robot Navigation*. Remote Sensing, 13(21), 4216.

[10]. Du Q., Wang D., Sha L., 2020. *Recognition of mobile robot navigation path based on K-means algorithm.* International Journal of Pattern Recognition and Artificial Intelligence, 34(08), 2059028.

[11]. Uslu E., Cakmak F., Altuntaş N., Marangoz S., Amasyalı M. F., Yavuz S., 2017. *An architecture for multi-robot localization and mapping in the Gazebo/Robot Operating System simulation environment.* Simulation, 93(9), 771-780.

[12]. Zhao W., Fang Z., Yang Z., 2020. *Four-dimensional trajectory generation for UAVs based on multi-agent Q learning*. The Journal of Navigation, 73(4), 874-891.

[13]. Wenxia X., Yu B., Cheng L., Li Y., Cao X., 2021. *Multi-fuzzy Sarsa learning-based sit-to-stand motion control for walking-support assistive robot*. International Journal of Advanced Robotic Systems, 18(5), 17298814211050190.

[14]. Fischer P., 1974. *On Bellman's functional equation.* J. Math. Anal. Appl, 46(197), 212-227.

[15]. Puterman M. L., 1990. *Markov decision processes*. Handbooks in operations research and management science, 2, 331-434.

[16]. Steimle L. N., Kaufman D. L., Denton B. T., 2021. *Multi-model Markov decision processes*. IISE Transactions, 53(10), 1124-1139.

AUTHORS INFORMATION

Nguyen Anh Tu[1], Nguyen Hong Son[1], Bui Huy Anh[1], Tran Quoc Hoan[2]

[1]Hanoi University of Industry

[2]Deparment of Otical and Mechanical Engineering, Public Security Institute of Science and Technology