MÔ HÌNH PHÂN ĐOẠN NGỮ NGHĨA HÌNH ẢNH ỨNG DỤNG DẪN ĐƯỜNG RÔ BỐT DI ĐỘNG VỚI CẤU HÌNH TỐI ƯU TÀI NGUYÊN

SEMANTIC SEGMENTATION MODEL FOR MOBILE ROBOT'S NAVIGATION WITH OPTIMAL ARCHITECTURE OF RESOURCE

Trần Đình Mạnh Cường[1], Đặng Thái Việt[1,*]

DOI: https://doi.org/10.57001/huih5804.2024.024

TÓM TẮT

Do hình ảnh kỹ thuật cung cấp đầy đủ thông tin cần thiết nên thị giác máy tính đóng một vai trò quan trọng trong điều hướng robot di động. Từ hình ảnh thu được, robot di động sẽ khoanh vùng và di chuyển đến đích đã định. Tùy theo sự phức tạp của môi trường, việc tránh chướng ngại vật vẫn đòi hỏi một hệ thống cảm biến phức tạp với yêu cầu hiệu quả tính toán cao. Nghiên cứu trong bài báo đưa ra một giải pháp thời gian thực cho vấn đề trích xuất cảnh hành lang từ một hình ảnh duy nhất. Sử dụng mô hình phân đoạn ngữ nghĩa cực nhanh tích hợp với kỹ thuật lượng tử hóa để giảm thiểu tham số đào tạo và chi phí tính toán của đào tạo mô hình phân đoạn ngữ nghĩa. Ngoài ra, mIoU đạt 89% và độ chính xác cao là 98%. Các kết quả mô phỏng được so sánh với các phương pháp gần đây để chứng minh tính khả thi của phương pháp đã thử. Cuối cùng, các tác giả áp dụng hiệu quả hình ảnh được phân đoạn để xây dựng chế độ xem trực diện của robot di động nhằm xác định quy hoạch đường đi của robot di động có sẵn.

Từ khóa: *Thị giác máy tính, phân đoạn ngữ nghĩa, rô bốt tự hành, điều hướng.*

ABSTRACT

Due to the wealth of information extracted from digital images, computer vision plays a significant role in mobile robot navigation. Based on the captured images, mobile robots localize and move to the intended destination. Due to the complexity of the environment, obstacle avoidance still requires a complex sensor system with a high computational efficiency requirement. This study offers a real-time solution to the problem of extracting corridor scenes from a single image. Using ultra fast semantic segmentation model integrating with quatization technique to reduce the numerous of training parameters and computational cost. Moreover, the mean Intersection over Union (mIoU) achieves 89% and high accuracy is 98%. The simulation results are compared with recent methods to prove the feasibility of the prosed method. Finally, the authors effectively apply the segmented image to construct the mobile robot frontal view to identify the available mobile robot path planning.

Keywords: *Computer vision, semantic segmentation, autonomous mobile robot, navigation.*

[1]Đại học Bách khoa Hà Nội

[*]Email: viet.dangthai@hust.edu.vn

Ngày nhận bài: 12/6/2023

Ngày nhận bài sửa sau phản biện: 15/9/2023

Ngày chấp nhận đăng: 20/01/2024

DANH MỤC VIẾT TẮT

MTCNN: Mạng tích chập xếp tầng đa tác vụ

DCNN: Mạng tích chập học sâu

CNN: Mạng tích chập

FPS: Số khung hình/giây

AI: Trí tuệ nhân tạo

IoT: Internet kết nối vạn vật

1. GIỚI THIỆU

Robot di động điều hướng môi trường của chúng một cách an toàn bằng cách nhận biết chướng ngại vật và vật thể di chuyển trong thời gian thực. Hệ thống cảm biến kết hợp với thuật toán điều hướng phát hiện chướng ngại vật bao gồm máy quét laze, cảm biến và camera [1]. Gần đây, việc sử dụng Lidar hoặc máy ảnh số trong cung cấp thông tin về môi trường di chuyển đã trở nên phổ biến. Mặc dù chi phí cao và số bước tính toán lớn [2]. Lidars cung cấp dữ liệu độ sâu theo mọi hướng, duy trì một thế giới gần đúng hoàn hảo. Ngoài ra, máy ảnh cung cấp dữ liệu cảnh không tốn kém để phát hiện bất kỳ đối tượng nào [3]. Do sự phổ biến của máy ảnh đơn có độ chính xác cao, giá cả phải chăng, những nhược điểm tồn tại trước đây đã được loại bỏ. Dựa trên các phương pháp nhận dạng môi trường bằng thị giác máy tính, các dạng mô hình phân đoạn hình ảnh thời gian thực và điều hướng tự động đã được thực hiện thành công.

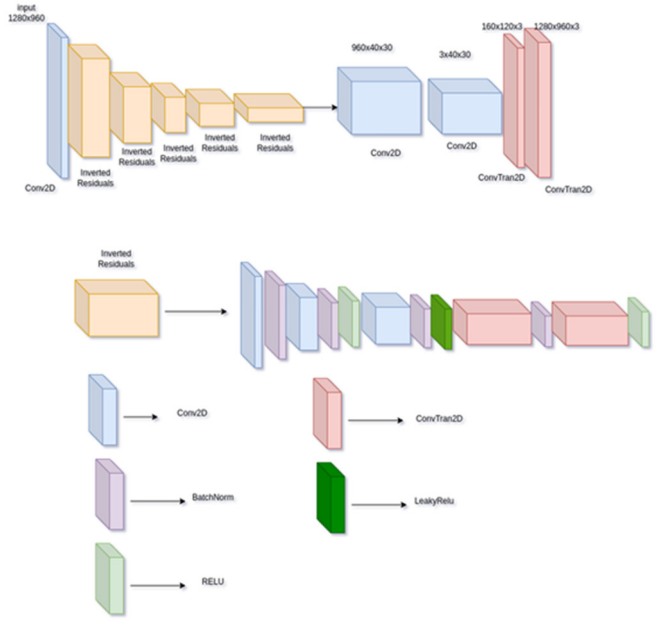
Phân đoạn ngữ nghĩa sử dụng học sâu (Deep learning) là một thách thức quan trọng trong nhiều ứng dụng dựa trên tầm nhìn [4-6], bao gồm trích xuất đặc trưng từ hình ảnh, nhận thức của rô bốt về môi trường di chuyển và các kích cỡ tối ưu hình ảnh. Rusli và các cộng sự [5] đã sử dụng thuật toán Canny edge và biến đổi Houghline để điều hướng robot di động. Ở đó, đường vạch trên đường và chướng ngại vật đã được phát hiện bởi các điểm mốc đường biên. Minaee và cộng sự [6] đã khảo sát các mô hình phân đoạn dựa trên học sâu đã thể hiện hiệu suất vượt trội trong các thử nghiệm phân đoạn hình ảnh trực quan nhằm giải quyết tình trạng khan hiếm bộ dữ liệu thông thường để đánh giá phân đoạn đối tượng. Mặc dù quy trình đạt được tốc độ xử lý cao nhưng

độ chính xác của nó vẫn khá thấp khi các yếu tố môi trường thay đổi. Shelhamer và cộng sự [7] đã chuyển đổi các mạng phân loại hiện đại (AlexNet, VGG và GoogLeNet) thành các mạng tích chập hoàn toàn (FCN) để nâng cao hiệu suất và độ chính xác của mô hình phân đoạn. Wang và cộng sự [8] đã giới thiệu một mạng tích chập, có tên là Adaptive Feature Fusion Unet (AFF-UNet) để cải thiện hiệu suất phân đoạn ngữ nghĩa. Sử dụng các mô hình VGG-FCN như [1, 2] hoặc Unet [8] để giải quyết vấn đề phân đoạn ảnh cho kết quả chính xác cao nhưng không phù hợp để triển khai cơ sở hạ tầng do chi phí tính toán cao.

Các tác giả thiết kế một mô hình phân đoạn với FCN là bộ giải mã và Mobilenet V2 là bộ mã hóa để giải quyết nhiệm vụ phân đoạn ngữ nghĩa [1, 3, 9] và đạt được khả năng hiểu cảnh hiệu quả cho điều hướng tự trị. Các tác giả sử dụng một mô hình được đào tạo trước đó từ các mạng tiền thân cho bộ mã hóa. Sau đó, các tác giả sử dụng phản ứng tổng hợp đa tỷ lệ để tạo dự đoán phân đoạn trong khối giải mã. Độ chính xác đạt được mô hình phân đoạn ngữ nghĩa đề xuất là 98%. Kết quả đầu ra của mô hình phân đoạn ngữ nghĩa hình ảnh giúp xây dựng chế độ xem trực diện của khu vực chung, phát hiện các đối tượng trong môi trường di chuyển của rô bốt. Trên cơ sở đầu ra của mô hình phân đoạn, chiến lược dẫn đường cho rô bốt được xây dựng đảm bảo khả năng tránh các vật cản di động. Cuối cùng bộ điều khiển PID đồng thời cho hai động cơ trái và phải giúp rô bốt di chuyển bám theo quỹ đạo với độ sai lệch góc lái nhỏ cho phép.

2. PHƯƠNG PHÁP ĐỀ XUẤT

2.1. Cấu trúc mô hình mạng phân đoạn ngữ nghĩa



Hình 1. Cấu trúc đề xuất về mô hình FCN phân đoạn ngữ nghĩa dựa trên MobilenetV2

Mô hình kiến trúc trên đã truyền cảm hứng cho nhiều biến thể. Các mạng trung tính tích chập (CNN) đã được chứng minh là có khả năng tạo ra các kết quả tiên tiến nhất về các vấn đề phân loại hình ảnh bằng các mô hình như LeNet và AlexNet [1, 2, 4]. Hiệu quả và hiệu suất đã được nâng cao nhờ các cải tiến tiếp theo như VGGNet, GoogleNet và ResNet [1, 2, 4, 9]. Các mạng nơ ron tích chập cuối cùng đã được phát triển thành một mạng tích chập hoàn chỉnh. Để đạt được việc ghi nhãn pixelwise ngay lập tức trong khi vẫn duy trì kết quả phân đoạn đáng tin cậy, các tác giả tạo một mạng dựa trên mạng nơ ron tích chập đầy đủ (FCN) [4, 9]. VGG được ưa thích hơn AlexNet do mô hình trước nổi tiếng hơn và dự báo kém tin cậy hơn mô hình tiếp sau [1, 2, 9]. Phân đoạn ngữ nghĩa là một thuật toán trong lĩnh vực thị giác máy tính và xử lý hình ảnh với mục tiêu là phân loại và phân đoạn từng pixel trong một hình ảnh thành các lớp tương ứng với nội dung ngữ nghĩa riêng biệt. Các tác giả triển khai mạng phân đoạn để ghi nhãn pixel cụ thể ngay lập tức chủ yếu dựa trên khái niệm về mô hình FCN. Mô hình mạng FCN phân đoạn ngữ nghĩa dựa trên MobilenetV2 được xây dựng như sau trong hình 1.

2.2. Huấn luyện mô hình

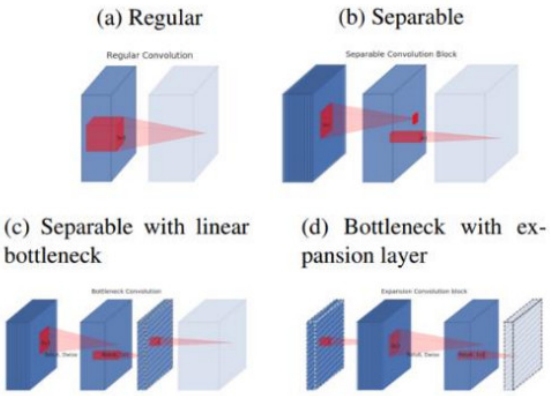
Các thực nghiệm của nghiên cứu được tiến hành trên một máy chủ có các thông số kỹ thuật sau sử dụng Python 3.11.0 với khung Tensorflow 1.4: chíp máy tính Corel7 11th 2,50GHz - 4,90GHz, Nivida 2080TI 12GB VRAM, RAM 32GB, hệ điều hành 64 bit và Windows 10 Home tiếng Anh. Các tham số đã cài đặt được thiết lập như sau: batchsize là 4, tốc độ học (learning rate) là 0,001 và epochs bằng 100. Quá trình đào tạo được thực hiện bằng cách tối ưu hóa hàm tổn thất Entropy chéo Blanced (BCE) được biểu thị trong công thức số (1):

$$L_{Balance-CE}(y,\hat{y})=\{\beta\times y\log(\hat{y})+(1-\beta)\times(1-y)\log(1-\hat{y})\} \quad (1)$$

với $\beta=1-\frac{y}{H\times W}$, và HxW là tổng số pixel trong ảnh. $\hat{y}$ là xác suất phân phối của hàm loss sofmax, y là giá trị chân lý phát triển. Bên cạnh đó β được sử dụng để điều chỉnh số lượng tiêu cực giả (false negative) và tích cực giả (positive negative). Nếu bạn muốn giảm số lượng tiêu cực giả thì $\beta > 1$, để giảm số lượng tích cực giả thì < 1. Hơn nữa, chức năng tối ưu hóa Adam được sử dụng để tăng tốc độ tìm trọng số và độ lệch phù hợp cho mô hình. Cuối cùng, sử dụng kỹ thuật truyền thông số học để tăng tốc độ học tập. Các tác giả khởi tạo bộ tham số sử dụng bằng với tham số mô hình MobileNetV2 đã được đào tạo trước đó. Các tính năng cơ bản của MobilenetV2 bao gồm tối ưu hóa hàm mất mát và việc sử dụng kiến trúc mô hình nâng cao từ mô hình FCN. Nút cổ chai tuyến tính và khối dư đảo ngược (shortcut links between bottlenecks) cũng được khuyến nghị bổ sung cho các kết cấu có thể tách rời theo chiều sâu của MobileNet V2 [9].

Do đầu vào và đầu ra của một khối trong kiến trúc phần dư thông thường thường có nhiều kênh hơn so với các lớp trung gian, khối phần dư của MobileNet v2 là nghịch đảo của thiết kế này. Để giảm số lượng tham số mô hình, chúng tôi sử dụng biến đổi tích chập phân tách theo chiều sâu và khối dư đảo ngược giữa các lớp. Giải pháp được đề xuất cho phép giảm nhẹ kích thước của mô hình MobileNet. Hình 2 mô tả

tổ chức bên trong của tích chập nút cổ chai được phân tách theo chiều sâu với phần dư.



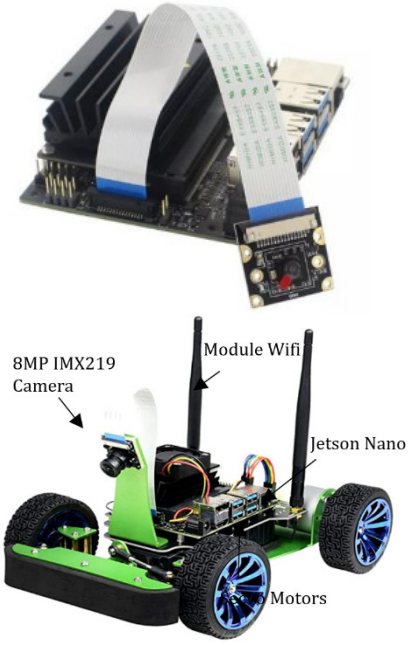
Hình 2. Mô hình khối tích chập tách rời trong MobileNetV2

3. KẾT QUẢ VÀ THẢO LUẬN

- Khối thu nhận hình ảnh: 8MP IMX219 Camera for NVIDIA Jetson Nano 77 Degree 1080P CSI Camera Module với góc nhìn rộng phía trước 160 độ. Hỗ trợ kết nối trực tiếp với Jetson nano 4G.

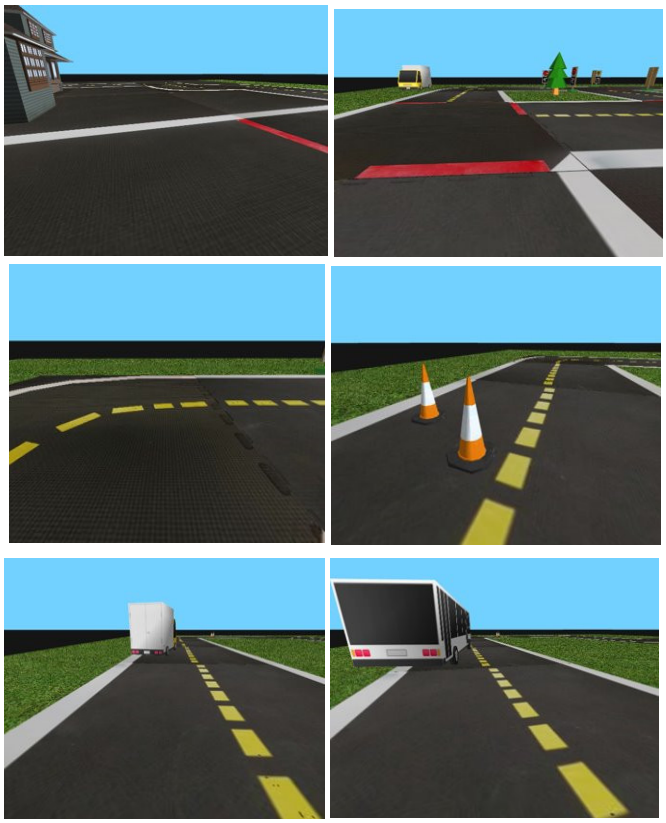
- Khối xử lý trung tâm gồm Jetson nano 4G có chức năng xử lý hình ảnh, tạo đường dẫn từ vùng cho phép di chuyển từ hình ảnh thu được và tạo tín hiệu điều khiển xuống board vi điều khiển Arduino. Vi điều khiển Arduino kết nối driver điều khiển, khối công suất cấp nguồn cho 2 động cơ secvo điều khiển trực tiếp bánh lái rô bốt tự hành trong hình 3.

- Bộ dữ liệu dùng cho huấn luyện mô hình phân đoạn ngữ nghĩa là 1200 bức ảnh từ thư viện của phần mềm mô phỏng Ducktown [10].



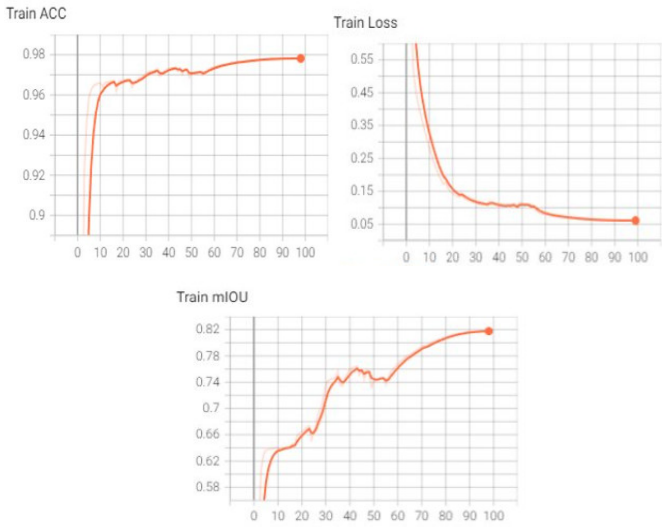
Hình 3. Mô hình rô bốt tự hành với máy tính nhúng Jetson nano và camera 8MP IMX219

Trong phần mềm mô phỏng Ducktown, có đầy đủ các đối tượng như nhà, xe, người, vật và đường đi,… như trong hình 4. Dựa trên bộ dữ liệu công bố công khai Ducktown, mô hình mạng nơ ron phân đoạn ngữ nghĩa được huấn luyện đảm bảo độ chính xác và chất lượng yêu cầu.



Hình 4. Đường di chuyển của rô bốt với các vật cản tĩnh và động

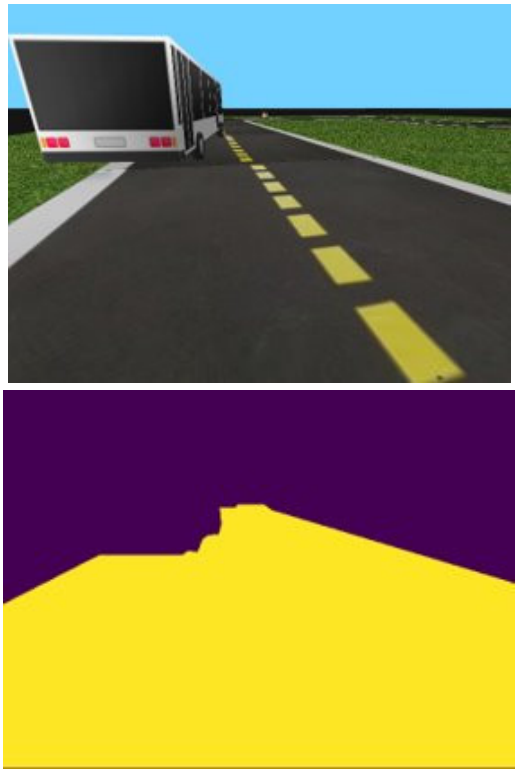
Với cấu trúc siêu nhanh của mô hình phân đoạn ngữ nghĩa đề xuất, các thông số huấn luyện mô hình được biểu diễn như hình 5.



Hình 5. Các tham số huấn luyện của mô hình phân đoạn ngữ nghĩa với độ chính xác gần 98,7%, loss: gần 0,06 và mIoU lên tới 82%

Từ việc huấn luyện thành công mô hình phân đoạn ngữ nghĩa, hệ thống thị giác rô bốt sẽ tạo ra kết quả đầu ra sau khi phân đoạn hình ảnh thu được để tạo ra 2 vùng phân biệt, vùng cho phép di chuyển (màu vàng) và vùng vật cản (màu

tím) như hình 6. Áp dụng phương pháp chuyển đổi tỷ lệ từ kích thước ảnh chụp môi trường sang kích thước thực tế của môi trường rô bốt di chuyển các tác giả có thể áp dụng mô hình điều khiển PID tích hợp với điều khiển bám quỹ đạo di chuyển từ kế hoạch dẫn đường tạo ra từ vùng kết quả xử lý hình của mô hình phân đoạn ngữ nghĩa.



Hình 6. Kết quả hình ảnh đầu ra của mô hình phân đoạn ngữ nghĩa từ hình ảnh thu nhận từ góc nhìn phía trước

Mô hình biểu diễn rô bốt di động dạng bánh được biểu diễn như hình 7, với 2 bánh lái được điều khiển đồng thời bởi bộ điều khiển PID cho 2 động cơ secvo dẫn động bánh lái trái và phải [11]. Điểm gốc hình học giữa đường nối trục 2 bánh O_m. Vận tốc lái bánh trái v_L và bánh phải v_R được biểu diễn như phương trình (2):

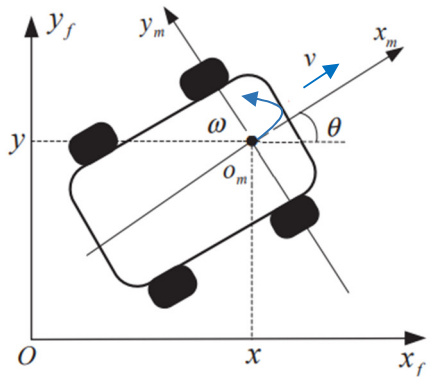
$$\upsilon_L = R\omega_L; \quad \upsilon_R = R\omega_R \tag{2}$$

với R: đường kính bánh xe, ω_L và ω_R là vân tốc góc bánh trái và phải tương ứng, θ góc lái phía trước của rô bốt. Phương trình động lực học của rô bốt di động được biểu diễn như phương trình (3):

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \theta \end{bmatrix} = \begin{bmatrix} \cos\theta & 0 \\ \sin\theta & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \upsilon \\ \omega \end{bmatrix} \tag{3}$$

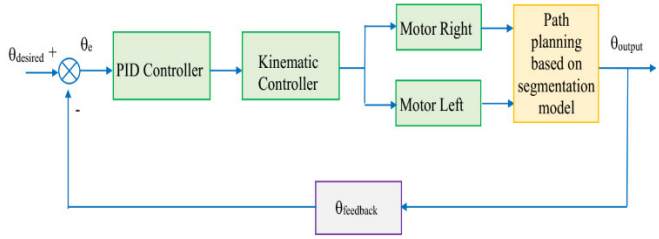
Với việc áp dụng bộ điều khiển PID đồng thời cho động cơ truyền động bánh trái và bánh phải của phương trình (2) vào (3) ta có phương trình biểu diễn cho tín hiệu điều khiển đầu vào (ω_L, ω_R) như sau:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \theta \end{bmatrix} = \begin{bmatrix} \cos\theta & 0 \\ \sin\theta & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \upsilon \\ \omega \end{bmatrix} = \begin{bmatrix} \cos\theta & 0 \\ \sin\theta & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{R\omega_L + R\omega_R}{2} \\ \frac{-R\omega_L + R\omega_R}{2} \end{bmatrix} = \begin{bmatrix} \cos\theta & 0 \\ \sin\theta & 0 \\ 0 & 1 \end{bmatrix} r \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \omega_L \\ \omega_R \end{bmatrix} \tag{4}$$



Hình 7. Mô hình biểu diễn rô bốt với 2 bánh lái đồng thời

Với x, y và θ là vị trí và phương hướng của rô bốt tham chiếu, v_R và ω_R là vận tốc và vận tốc góc tương ứng cho động cơ truyền động bánh phải, v_L và ω_L là vận tốc và vận tốc góc tương ứng cho động cơ truyền động bánh trái. Sơ đồ điều khiển độc lập cho rô bốt di động được thiết kế như hình 8.



Hình 8. Sơ đồ Matlab-Simulink của hệ thống điều khiển rô bốt di động

4. KẾT LUẬN

Bài báo đề xuất mô hình phân đoạn ngữ nghĩa cấu hình tối ưu tài nguyên sử dụng máy ảnh số đơn thấu kính trong quá trình xử lý ảnh môi trường tạo kế hoạch dẫn đường cho rô bốt di động. Kết quả hình ảnh phân đoạn đảm bảo chất lượng, độ chính xác với tốc độ xử lý nhanh. Dựa trên kết quả thu được, một hệ thống điều khiển PID đồng thời cho 2 bánh lái của rô bốt di động được thiết kế. Các kết quả thực nghiệm trên mô phỏng minh chứng rô bốt có khả năng bám quỹ đạo dẫn đường với sai lệch thay đổi góc lái dưới 0,5rad. Mô hình mạng nơ ron phân đoạn ngữ nghĩa có ý nghĩa trong ứng dụng thực tế cho các hệ thống dẫn đường rô bốt tự hành thực tế.

LỜI CẢM ƠN

Nhóm tác giả trân trọng cảm ơn Trường Cơ khí và Viện VJIIST, Đại học Bách khoa Hà Nội đã hỗ trợ trong nghiên cứu này.

TÀI LIỆU THAM KHẢO

[1]. Dang T. V.. Bui N. T., "Multi-scale Fully Convolutional Network based Semantic Segmentation for Mobile Robot Navigation," *Electronics*, 12(3), 533, 2022.

[2]. Dang T. V., Bui N. T., "Obstacle Avoidance Strategy for Mobile Robot based on Monocular Camera," *Electronics*, 12(8), 1932, 2023.

[3]. Maulana I., Rasdina A., Priramadhi R.A., "Lidar applications for Mapping and Robot Navigation on Closed Environment," *J. Meas. Electron. Commun. Syst*. 4, 767–782, 2018.

[4]. W. Sun, R. Wang, "Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM," in *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 3, pp. 474-478, 2018.

[5]. Rusli L., Nurhalim B., Rusyadi R., "Vision-based vanishing point detection of autonomous navigation of mobile robot for outdoor applications," *J. Mechatron. Elect. Power Veh. Technol.* 12, 117–125, 2021.

[6]. Minaee S., Boykov Y.Y., Porikli F., Plaza A.J., Kehtarnavaz N., Terzopoulos D., "Image Segmentation Using Deep Learning: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 3523–3542, 2022.

[7]. Shelhamer V., Long J., Darrell T., "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell*. 1–12, 2016.

[8]. Wang C., Zhao Z., Ren Q., Xu Y., Yu Y., "Dense U-Net based on patch-based learning for retinal vessel segmentation," *Entropy*, 21, 168, 2019.

[9]. Dang T. V., Tran D. M. C., Tan P. X., "IRDC-Net: Lightweight Semantic Segmentation Network Based on Monocular Camera for Mobile Robot Navigation," *Sensors,* 23(15), 6907, 2023..

[10]. Kirill K., Konstantin C., Anton F., Artyom F., "Autonomous Wheels And Camera Calibration In Duckietown Project," *Procedia Comput. Sci.*, 186, 169-176, 2021.

[11]. VT Nguyen, QC Nguyen, DH Phan, TL Bui, VX Hoang, "Optimized PID Controller for Two-Wheeled Self-balancing Robot Based on Genetic Algorithm," *Lecture Notes in Networks and Systems* 752, 60–65, 2023

AUTHORS INFORMATION

Tran Dinh Manh Cuong, Dang Thai Viet

Hanoi University of Science and Technology, Vietnam