

MAI_ARM: Robot Tay Máy Thông Minh sử dụng Trí tuệ nhân tạo Đa thể thức

Đinh Bảo Minh, Trần Đức Sơn, Nguyễn Thế Hường, Đỗ Nam, Lê Minh Hoàng, Hoàng Văn Xiêm

Bộ môn Kỹ thuật Robot, Khoa Điện tử-Viễn thông

Trường Đại học Công Nghệ, Đại học Quốc gia Hà Nội

Email: minhdingh@vnu.edu.vn, transonh12c5@gmail.com, huongrbe@gmail.com, donam.2801@gmail.com,

lhoang17062000@gmail.com, xiemhoang@vnu.edu.vn

Abstract— Các hệ thống robot tay máy đóng một vai trò rất quan trọng trong sản xuất công nghiệp hiện đại. Việc ứng dụng robot tay máy trong đời sống hàng ngày sẽ giúp ích con người rất nhiều, ví dụ như tự động hóa hoàn toàn quy trình phân phát thuốc và thực phẩm cho bệnh nhân nhiễm COVID-19. Ứng dụng robot tay máy trong đời sống đòi hỏi một cách thức giúp người dùng dễ dàng tương tác với robot. Trong bài báo này, một phương pháp tương tác với robot tay máy sử dụng trí tuệ nhân tạo đa thể thức kết hợp giọng nói-hình ảnh được xây dựng. Bài báo cũng đề xuất phương pháp Chessboard Calibration giúp nâng cao độ chính xác trong việc xác định vị trí thực thi của robot tay máy. Một robot tay máy 4 bậc tự do (4-DOF) chế tạo bằng phương pháp in 3D sẽ được sử dụng để thực thi và đánh giá mô hình trí tuệ nhân tạo đa thể thức xây dựng được. Kết quả thực thi hệ thống được ghi hình và trình bày tại: <https://youtu.be/RHgAvHXMH6I>

Keywords- Robot tay máy, Trí tuệ nhân tạo, Trí tuệ nhân tạo đa thể thức

I. GIỚI THIỆU

Robot tay máy là một loại robot gồm nhiều khớp chuyển động cơ học, có thể thực hiện được các chức năng tương tự như tay của con người. Với khả năng hoạt động liên tục không nghỉ, robot tay máy đóng một vai trò rất quan trọng trong sản xuất công nghiệp. Ngày nay, robot tay máy được nghiên cứu để ứng dụng nhiều hơn trong đời sống hàng ngày.

Phần lớn robot tay máy sử dụng trong công nghiệp thường được lập trình và điều khiển thông qua bộ điều khiển chuyên dụng. Điều này đòi hỏi người sử dụng phải được đào tạo chuyên môn để đảm bảo độ chính xác cũng như sự an toàn. Do đó, việc có một cơ chế tương tác dễ dàng hơn sẽ giúp robot tay máy có thể ứng dụng rộng rãi hơn trong đời sống hàng ngày. Một trong những giải pháp được nghiên cứu là ứng dụng trí tuệ nhân tạo đa thể thức.

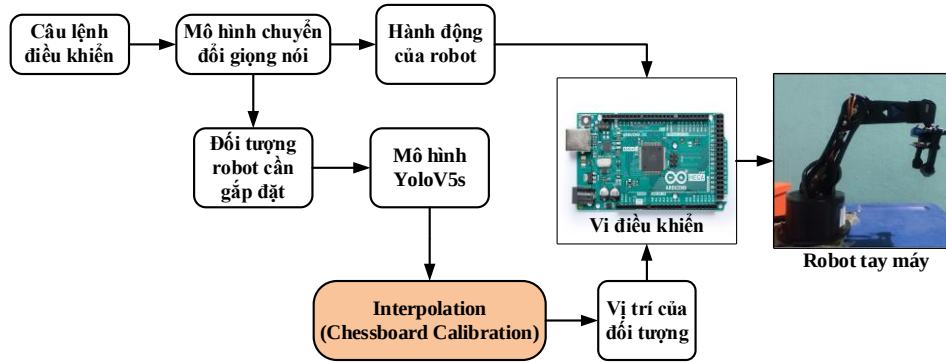
Trí tuệ nhân tạo đa thể thức (hay học đa phương thức) là một mô hình học máy có khả năng hiểu và xử lý thông tin từ nhiều nguồn thông tin không đồng nhất (như hình ảnh, giọng nói, chữ viết...) [1]. Việc học các đặc tính từ nhiều nguồn thông tin có thể mang lại khả năng tạo ra suy luận mạnh mẽ hơn và quyết định đưa ra linh hoạt hơn. Trong bài toán tương tác với robot tay máy,

các mô hình học đa phương thức thường kết hợp xử lý thông tin giọng nói-cử chỉ và giọng nói-hình ảnh.

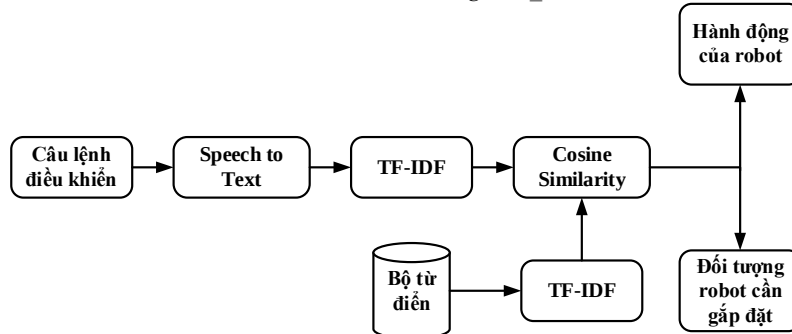
Các nghiên cứu [2-4] mô tả giải pháp kết hợp xử lý thông tin giọng nói-cử chỉ để tương tác với robot tay máy. Nghiên cứu [2] trình bày một phương pháp dạy robot trực tuyến (online teaching) thông qua cử chỉ và giọng nói. Đồng thời, nghiên cứu đề xuất phương pháp AVFT cho phép trích xuất tối đa thông tin từ lời nói và cử chỉ bằng cách chuyển hóa chúng thành dạng văn bản. Giải pháp đề xuất giúp người ít chuyên môn có thể tương tác với robot một cách nhanh chóng và hiệu quả. Trong nghiên cứu [3] phương pháp điều khiển robot tay máy 3 bậc tự do thông qua giọng nói và cử chỉ được đề xuất. Theo đó, điều khiển thông qua cử chỉ sẽ được sử dụng nếu robot không nghe được câu lệnh điều khiển bằng giọng nói. Yongda và các cộng sự đề xuất một phương pháp [4] tương tác với robot tay máy kết hợp giọng nói và cử chỉ. Kết quả thực nghiệm chỉ ra phương pháp đề xuất cho hiệu suất tốt hơn các phương pháp đã được đề xuất khác.

Các nghiên cứu [5, 6] đề xuất giải pháp tương tác với robot tay máy kết hợp xử lý thông tin giọng nói-hình ảnh. Công bố [5] phần mềm nhận dạng giọng nói bằng tiếng Thổ Nhĩ Kỳ được sử dụng để điều khiển tương tác với robot tay máy công nghiệp. Thông tin điều khiển bằng giọng nói sẽ được kết hợp với thông tin vị trí của vật có được từ phương pháp xử lý ảnh sẽ được sử dụng để vận hành robot. Lin và các cộng sự qua nghiên cứu [6] đề xuất hệ thống phân loại đối tượng sử dụng robot tay máy. Hệ thống có chức năng nhận dạng và tổng hợp ngôn ngữ, kết hợp nhận thức môi trường point cloud 3D thông qua camera RGB-D. Từ đó có khả năng suy luận để tương tác với con người hiệu quả hơn.

Trong nghiên cứu này, một mô hình trí tuệ nhân tạo đa thể thức kết hợp giọng nói-hình ảnh sẽ được xây dựng nhằm điều khiển robot tay máy thực hiện tác vụ gấp đặt vật. Các bộ thư viện có sẵn sẽ được sử dụng để xây dựng một mô hình nêu trên. Ngoài ra, phương pháp nâng cao độ chính xác vị trí vật với Chessboard Calibration sẽ được đề xuất để tăng hiệu suất của tác vụ gấp đặt vật. Robot tay máy 4 bậc tự do (4-DOF) được thiết kế và chế tạo bằng phương pháp in 3D sẽ được sử dụng để kiểm



Hình 1: Kiến trúc hệ thống MAI_ARM



Hình 2. Mô hình chuyển đổi giọng nói

chúng mô hình trí tuệ nhân tạo đa thể thức xây dựng được.

Phần còn lại của bài báo được triển khai như sau: Phần II mô tả quy trình xây dựng mô hình MAI_ARM. Tiếp đó, phần III trình bày các kết quả thực nghiệm và đánh giá. Cuối cùng, kết luận của bài báo được trình bày tại phần IV.

II. MAI_ARM: ROBOT TAY MÁY THÔNG MINH SỬ DỤNG TRÍ TUỆ NHÂN TẠO ĐA THỂ THỨC

Kiến trúc hệ thống MAI_ARM được mô tả như trong Hình 1. Theo đó, câu lệnh điều khiển bằng giọng nói sẽ được phân tích để nhận biết hành động và đối tượng của hành động. Tiếp đó, mô hình mạng YoloV5s [7] sẽ nhận diện và xác định vị trí theo tọa độ pixel của đối tượng xác định được qua câu lệnh điều khiển bằng giọng nói. Thông tin vị trí theo tọa độ pixel của đối tượng được đi qua khối Interpolation sử dụng phương pháp đề xuất Chessboard Calibration giúp tối ưu sai số có thể có trong quá trình chuyển đổi từ tọa độ pixel sang tọa độ thực gây ra bởi camera. Cuối cùng, vi điều khiển

sẽ xử lý thông tin hành động và vị trí làm việc và tiến hành điều khiển robot tay máy 4 bậc tự do thực hiện tác vụ được yêu cầu.

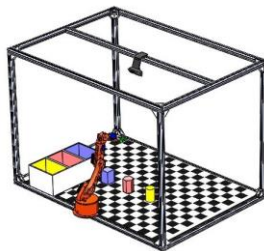
A. Mô hình chuyển đổi giọng nói

Công nghệ chuyển đổi giọng nói [8] là quá trình nhận dạng mẫu, với mục đích là phân lớp thông tin đầu vào là tín hiệu tiếng nói thành một dãy tuần tự các mẫu đã được học trước đó và lưu trữ chúng trong bộ nhớ. Công nghệ chuyển giọng nói thành văn bản giúp cho con người có thể giao tiếp với máy móc một cách dễ dàng hơn, thay vì phải nhập văn bản từ bàn phím thì sử dụng giọng nói.

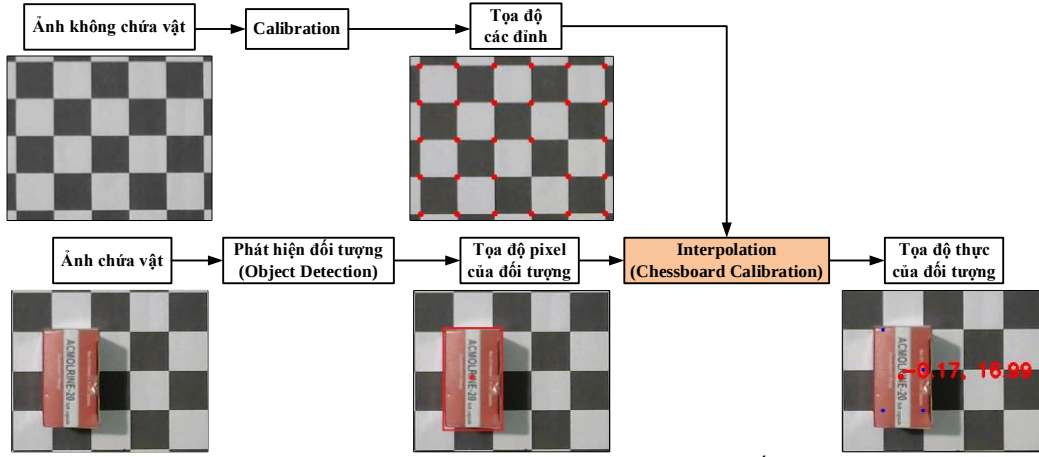
Bài báo trình bày một mô hình chuyển đổi giọng nói thành văn bản có ý nghĩa với robot tay máy. Nghiên cứu tạo ra một bộ từ điển bao gồm các câu nói có thể đặt ra, hoặc các hành vi thể hiện chức năng của robot có thể thực hiện được. Dựa trên bộ từ điển này mô hình chuyển đổi giọng nói sẽ đưa ra thông tin về hành động của robot và đối tượng mà robot cần gấp đặt. Quy trình của thuật toán như sau:

B1: Câu lệnh điều khiển sẽ được đi qua khối Speech to Text. Tại đây, thư viện chuyển đổi giọng nói thành văn bản SpeechRecognition [9] sẽ được sử dụng. Một văn bản thô mô tả câu lệnh điều khiển sẽ được tạo ra sau khi đi qua khối Speech to Text.

B2: Văn bản thu được ở **B1** sẽ được đi qua khối TF-IDF [10, 11]. TF-IDF là sự kết hợp của tần số xuất hiện của một từ trong một mẫu (TF) và nghịch đảo của tần số của từ đó trong toàn bộ tập dữ liệu (IDF), cho phép mỗi mẫu trong tập dữ liệu có trọng số khác nhau. Tập trọng số này



Hình 3: Mô phỏng môi trường làm việc của robot



Hình 4: Quy trình xác định vị trí theo tọa độ thực của đối tượng

được lưu thành dạng vector biểu thị đặc trưng của văn bản. TF-IDF được tính như sau:

$$TF-IDF = TF(w, D) \cdot \log \left(\frac{C}{DF(w)} \right) \quad (1)$$

Trong đó:

- $TF(w, D)$ là đại diện cho tần số của từ w xuất hiện trong văn bản D ;
- C là tổng số văn bản trong toàn tập dữ liệu;
- $DF(w)$ là số văn bản có chứa từ w trong toàn tập dữ liệu.

B3: Bộ từ điển gồm 20-30 câu lệnh điều khiển robot tay máy sau khi đi qua khối TF-IDF tương tự khối TF-IDF ở **B2** sẽ cho ra một vector truy vấn biểu thị đặc trưng của bộ từ điển. Vector đặc trưng của văn bản (x) sẽ được so sánh với vector truy vấn (y) bởi phép tính Cosine Similarity [12] theo công thức (2). Kết quả của phép tính nằm trong đoạn từ 0 đến 1 cho chúng ta số liệu biểu thị góc giữa vector đặc trưng của văn bản và vector truy vấn. Góc này càng nhỏ (giá trị $sim(x,y)$ càng lớn) thì mức độ tương đương càng lớn.

$$sim(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|} \quad (2)$$

B4: Sau khi đi qua khối Cosine Similarity, hệ thống sẽ xác định được thông tin về hành động của robot và đối tượng mà robot cần gặp đặt.

B. Mô hình xác định đối tượng và vị trí theo tọa độ pixel của đối tượng

Hiện nay, có rất nhiều bài toán đòi hỏi ta phải xác định tên đối tượng và tọa độ của đối tượng ở trên ảnh. Kết hợp hai vấn đề nêu trên, ta có bài toán phát hiện đối tượng (object detection). Các mô hình để giải quyết bài toán phát hiện đối tượng hiện nay thường được chia làm hai loại chính đó là one-stage và two-stage [13]:

- One-stage: Mô hình tìm kiếm đối tượng bằng cách chạy một lần lấy mẫu dày đặc các vị trí. Vì vậy tốc độ của loại mô hình này rất cao;
- Two-stages: Có thêm bước tìm kiếm vùng, sau đó tìm kiếm vật thể trong vùng xác định được. Loại mô hình này cho độ chính xác rất cao, nhưng phải đánh đổi bằng tốc độ.

Mục tiêu của nghiên cứu là hỗ trợ tương tác với robot tay máy, ví vậy nghiên cứu lựa chọn sử dụng một mô hình one-stage với ưu thế về tốc độ mà loại mô hình này mang lại.

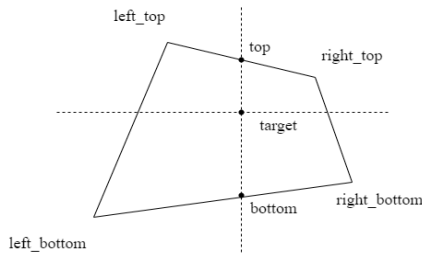
Yolo (You only look once) [14] là mô hình one-stage được sử dụng rộng rãi những năm gần đây trong bài toán phát hiện đối tượng vì tốc độ xử lý nhanh, độ chính xác tương đối cao và dễ triển khai. Bài báo sử dụng mô hình YoloV5s [7] cho nhiệm vụ xác định đối tượng và vị trí theo tọa độ pixel của đối tượng.

C. Nâng cao độ chính xác vị trí theo tọa độ thực đối tượng với Chessboard Calibration

Để robot tay máy có thể hiểu được thông tin vị trí của đối tượng, tọa độ pixel xác định được bằng mô hình YoloV5s cần phải được chuyển đổi qua tọa độ thực. Trong quá trình chuyển đổi, do ảnh hưởng của ống kính camera, việc xuất hiện sai số là điều không thể tránh khỏi. Do đó, nghiên cứu đề xuất giải pháp nâng cao độ chính xác vị trí đối tượng với Chessboard Calibration. Kết quả kiểm thử ở phần III cho thấy giải pháp giúp giảm đáng kể sai số trong quá trình chuyển đổi. Tuy nhiên có một nhược điểm của phương pháp đó là phương pháp chỉ hoạt động trong một môi trường với sàn có dạng một bàn cờ. Do đó, một môi trường làm việc như Hình 3 được xây dựng gồm một robot tay máy, sàn là bàn cờ với mỗi ô cờ có cạnh 3cm, và một camera cố định song song và cách sàn 60cm. Môi trường làm việc này cũng sẽ được sử dụng cho việc kiểm thử và đánh giá hệ thống được thực hiện ở phần III của bài báo. Hình 4 mô tả tổng quan quy trình xác định vị trí theo tọa độ thực của đối tượng. Trong quy trình này, phương pháp đề xuất được tích hợp trong bước Interpolation.

Với bước Calibration, đầu tiên chúng tôi sử dụng Biến đổi Hough (Hough Transformation) [15] để tìm các đường thẳng trên bàn cờ. Tiếp theo tìm giao của chúng và lọc dựa trên khoảng cách giữa các điểm giao này để tìm ra các đỉnh ô bàn cờ.

Mô hình YoloV5s mô tả ở phần II.B được sử dụng tại bước Phát hiện đối tượng (Object Detection). Vị trí



Hình 5: Bi-linear Interpolation

theo tọa độ pixel của đối tượng được xác định là tâm của hình chữ nhật bao quanh đối tượng.

Ở bước Interpolation, nhằm xác định vị của đối tượng theo tọa độ thực, đầu tiên nghiên cứu thử phương pháp Scaling theo tỉ lệ giữa độ dài cạnh ô bàn cờ thực tế và độ dài cạnh ô bàn cờ theo số pixel. Tuy nhiên sai số của phương pháp này là rất lớn và đồng thời không ổn định do trên các vùng ảnh khác nhau thì tỉ lệ scale thay đổi rất nhiều phụ thuộc vào ống kính camera. Qua quan sát chúng tôi đưa ra hai nhận xét như sau:

- Nội suy tọa độ vật dựa vào các đỉnh bàn cờ ở gần vật sẽ cho kết quả cao hơn do trong cùng một vùng ảnh sự sai số tỉ lệ khoảng cách do distortion là không nhiều;
- Robot tay máy đặt ở vị trí (x_0, y_0) giữa một ô bàn cờ, do camera và bàn cờ được lắp đặt sao cho trục tọa độ của bàn cờ gần cùng hướng với trục tọa độ ảnh nên tọa độ thực của ô bàn cờ (x, y) có thể tính gần đúng bằng:

$$\left(\frac{\left\lceil \frac{dx \cdot 2}{l} \right\rceil \cdot r}{2}, \frac{\left\lceil \frac{dy \cdot 2}{l} \right\rceil \cdot r}{2} \right) \quad (3)$$

Với:

- $dx = x - x_0$;
- $dy = y - y_0$;
- l là chiều dài ước lượng một ô bàn cờ trên ảnh;
- r là chiều dài thực của ô bàn cờ;
- Dấu $\lceil \rceil$ biểu thị phần nguyên.

Với những quan sát trên nghiên cứu đề xuất phương pháp làm tăng độ chính xác trong xác định vị của đối tượng theo tọa độ thực theo quy trình sau:

B1: Tính tham số l tương đối. Nghiên cứu thực hiện lấy ngẫu nhiên 3 đỉnh trong danh sách các đỉnh ô bàn cờ có được ở bước Calibration, sau đó chọn ra điểm gần nó nhất và tính khoảng cách giữa chúng. Quy trình này



Hình 6. Robot tay máy 4 bậc tự do

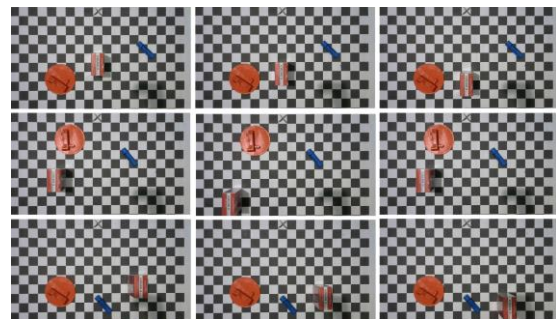
được lặp lại 5 lần, lấy trung bình tất cả các khoảng cách tính được trong 5 lần này sẽ cho ta giá trị tham số l ;

B2: Với một điểm ảnh target (mục tiêu) $(real_x, real_y)$, nghiên cứu xác định 4 đỉnh gần điểm đó nhất ở 4 hướng left_top, right_top, left_bottom, right_bottom. Sau đó dựa vào tọa độ theo pixel của các điểm này, theo (3) tính được tọa độ thực gần đúng của 4 điểm này;

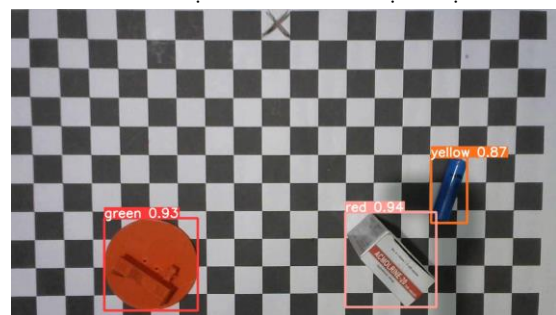
B3: Với tọa độ thực của 4 điểm xác định được ở **B2**, phương pháp Bi-linear Interpolation được sử dụng để nội suy ra tọa độ thực của điểm ảnh mục tiêu $(inter_x, inter_y)$.

BẢNG 1: CÁC CÂU LỆNH ĐIỀU KHIỂN VÀ ĐỐI TƯỢNG SỬ DỤNG

Câu lệnh điều khiển	Đối tượng	Nhãn sử dụng cho mô hình YoloV5s
Lấy hình tròn		green
Lấy cái hộp		red
Lấy cục pin		yellow



Hình 7: Một số hình ảnh của bộ dữ liệu



Hình 8: Kết quả nhận diện của mô hình YoloV5s

```

YOLOv5 2021-9-10 torch 1.9.1-cu102 CPU
Fusing layers...
Model Summary: 224 layers, 7059304 parameters, 0 gradients
Adding AutoShape...
image 1/1: 720x1280 1 green, 1 red, 1 yellow
Speed: 7.2ms pre-process, 100.9ms inference, 0.8ms NMS per image at shape (1,
Saved 1 image to runs/detect/exp5

```

	xcenter	ycenter	width	height	confidence	class	name
0	885.401062	584.307434	211.873657	220.723145	0.935299	1	red
1	326.371277	594.716125	218.509171	215.401733	0.934376	0	green
2	1021.444214	422.114410	84.081848	153.319153	0.871027	2	yellow

Hình 9: Độ tin cậy khi nhận diện của mô hình YoloV5

BẢNG II: KẾT QUẢ SO SÁNH SAI SỐ XÁC ĐỊNH VỊ TRÍ VẬT GIỮA CÁC PHƯƠNG PHÁP

Lần đo PP	1	2	3	4	5	6	7	8	9	10	TB
Scaling	14.49	21.50	27.47	25.16	26.63	14.01	18.86	24.65	29.17	32.90	23.48
Scaling + Undistort	15.00	22.37	28.89	27.30	29.89	16.51	22.39	27.53	32.23	35.83	25.79
Our	5.00	10.26	15.21	14.22	16.46	8.67	13.73	16.24	18.78	21.18	13.97

^a Đơn vị đo (mm)

BẢNG III: SỐ LẦN THỰC HIỆN CHÍNH XÁC KHI NÓI NHANH

Lấy cái hộp	Lấy hình tròn	Lấy cục pin
46/50	42/50	43/50

BẢNG IV: SỐ LẦN THỰC HIỆN CHÍNH XÁC KHI NÓI CHẬM

Lấy cái hộp	Lấy hình tròn	Lấy cục pin
49/50	48/50	50/50

D. Robot tay máy 4 bậc tự do

Robot tay máy 4 bậc tự do như Hình 6 được chế tạo bằng phương pháp in 3D sẽ được sử dụng để thực hiện tác vụ đầu ra của mô hình trí tuệ nhân tạo đa thể thức. Thông tin về lệnh điều khiển và vị trí điểm đích có được từ mô hình trí tuệ nhân tạo đa thể thức sẽ được truyền tới vi điều khiển Arduino Mega 2560 Rev3 [16] thông qua giao tiếp Serial (UART). Tại đây, vi điều khiển sẽ thực hiện các bài toán động học thuận và nghịch để di chuyển tay máy. Robot tay máy sau khi thực hiện xong câu lệnh điều khiển sẽ tự động trở về vị trí (0, 0) và chờ câu lệnh tiếp theo.

III. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Các điều kiện đánh giá

Toàn bộ quá trình thực nghiệm và đánh giá kết quả được tiến hành trong môi trường làm việc mô tả trong Hình 3. Bảng I mô tả thông tin về các câu lệnh điều khiển và đối tượng được sử dụng cho quá trình thực nghiệm. Bộ dữ liệu gồm 20 ảnh kích cỡ (1240×720) (Hình 7) chứa đối tượng được mô tả trong Bảng I sẽ được sử dụng để kiểm thử và đánh giá của hệ thống MAI_ARM. Khoảng cách Euclid được sử dụng để tính toán sai số giữa tọa độ nội suy được từ phương pháp Chessboard Calibration với tọa độ thực tế của đối tượng. Khoảng cách Euclid được tính theo công thức sau:

$$err_e = \sqrt{err_x^2 + err_y^2} \quad (4)$$

Trong đó:

- $err_x = inter_x - real_x$ là sai số giữa tọa độ nội suy và tọa độ thực theo trục x
- $err_y = inter_y - real_y$ là sai số giữa tọa độ nội suy và tọa độ thực theo trục y

A. Kết quả phát hiện đối tượng của mô hình YoloV5s

Với tập dữ liệu gồm 20 ảnh chứa 3 đối tượng như mô tả ở phần III.A, mô hình YoloV5s cho kết quả rất chính xác. Kết quả mô tả ở Hình 8 và Hình 9 cho thấy mô hình nhận diện chính xác vật với độ tin cậy (confidence) > 87.18%. Như vậy, đối với tập dữ liệu và số lượng đối tượng cần nhận diện ít, mô hình YoloV5s cho hiệu quả rất tốt. Trong tương lai, chúng tôi sẽ tiến hành kiểm thử

với tập dữ liệu cũng như số lượng đối tượng cần nhận diện nhiều hơn để kiểm chứng hiệu suất của mô hình với tập dữ liệu lớn.

B. Kết quả nâng cao độ chính xác vị trí với Chessboard Calibration

Chúng tôi thực hiện kiểm thử và so sánh thuật toán đề xuất với phương pháp Scaling và phương pháp kết hợp Scaling với Undistort. Các phương pháp so sánh được thực hiện bằng việc sử dụng libcbdetect [17] và OpenCV [18]. Kết quả so sánh được mô tả trong Bảng II.

Kết quả Bảng II cho thấy, phương pháp đề xuất giảm sai số trung bình so với thuật toán Scaling 9.51mm (từ 23.48mm xuống 13.97mm). So với kết hợp kỹ thuật Undistort và Scaling, phương pháp đề xuất giúp giảm 11.82mm sai số trung bình (từ 25.79mm xuống 13.97mm). Tuy nhiên sai số vẫn còn là khá lớn vì hiện tại chúng tôi đang sử dụng tâm hình chữ nhật bao quanh đối tượng để xác định vị trí của đối tượng. Điều này là không đúng đặc biệt là trong trường hợp vật đứng vì khi đó tâm đáy và tâm hộp bao có thể lệch nhau tới vài mm, gây ra sai số rất lớn

C. Kết quả tổng thể của hệ thống

Kiểm thử hiệu suất của hệ thống MAI_ARM, nghiên cứu tiến hành thực nghiệm điều khiển robot tay máy với 3 câu lệnh mô tả tại Bảng I, mỗi câu lệnh thực hiện 50 lần. Bảng III mô tả kết quả khi chúng tôi nói câu lệnh điều khiển với tốc độ nhanh (2.40 - 4.03 từ/giây). Bảng IV mô tả kết quả khi chúng tôi giảm tốc độ nói xuống (2.03 - 3.48 từ/giây). Dựa vào kết quả Bảng III và Bảng IV có thể thấy, hệ thống nhận diện và thực hiện câu lệnh điều khiển khá chính xác khi nói nhanh và rất chính xác khi nói chậm.

Một phần của quá trình thực thi của hệ thống được ghi hình và trình bày tại địa chỉ: <https://youtu.be/RHgAyHXMHG6I>. Qua video mô tả, có thể nhận thấy, trường hợp tọa độ nội suy bị sai khác quá lớn so với tọa độ thực tế của vật (robot tay máy gấp huyệt vật) xảy ra rất ít. Hệ thống chủ yếu phụ thuộc vào việc

phân tích câu lệnh điều khiển. Như vậy, với việc nói chậm câu lệnh điều khiển, hệ thống MAI_ARM mô tả trong bài báo sẽ giúp người dùng có thể tương tác với robot tay máy một cách hiệu quả và dễ dàng.

IV. KẾT LUẬN

Trong bài báo này, một phương pháp tương tác giữa người và robot tay máy sử dụng trí tuệ nhân tạo đa thể thức kết hợp giọng nói-hình ảnh đã được trình bày. Đồng thời, để cải thiện độ chính xác trong quá trình chuyển đổi từ tọa độ pixel qua tọa độ thực, nghiên cứu đề xuất phương pháp Chessboard Calibration. Kết quả thực nghiệm với robot tay máy 4 bậc tự do cho thấy mô hình đề xuất giúp người dùng tương tác dễ dàng với robot, và robot thực hiện đúng công việc được yêu cầu với độ chính xác cao.

Kết quả thực nghiệm cũng cho thấy, để đạt được độ chính xác cao đòi hỏi người sử dụng phải nói chậm. Do đó, trong tương lai chúng tôi sẽ tích hợp thêm mô tả bằng cử chỉ để cải thiện độ chính xác của mô tả bằng lời nói khi người sử dụng nói nhanh.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội theo đề tài mã số CN20.42.

TÀI LIỆU THAM KHẢO

- [1] T. Baltrušaitis, C. Ahuja and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423 - 443, Jan. 2018.
- [2] G. Du, M. Chen, C. Liu, B. Zhang and P. Zhang, "Online Robot Teaching with Natural Human-robot Interaction," IEEE Transactions on Industrial Electronics, vol. 65, no. 12, pp. 9571 - 9581, Dec. 2018.
- [3] N. Kandalafi, P. S. Kalidindi, S. Narra and H. N. Saha, "Robotic arm using voice and Gesture recognition," in 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Nov. 2018.
- [4] D. Yongda, L. Fang and X. Huang, "Research on multimodal human-robot interaction based on speech and gesture," Computers and Electrical Engineering, vol. 72, p. 443-454, 2018.
- [5] M. C. Bingol and O. Aydogmus, "Performing predefined tasks using the human-robot interaction on speech recognition for an industrial robot," Engineering Applications of Artificial Intelligence, vol. 95, 2020.
- [6] Y. Lin, H. Min, H. Zhou and F. Pei, "A Human-Robot-Environment Interactive Reasoning Mechanism for Object Sorting Robot," IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 3, pp. 611 - 623, Sept. 2018.
- [7] "yolov5," Online Available: <https://github.com/ultralytics/yolov5>.
- [8] G. Potamianos, C. N. Neti, J. Luetin and L. M. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview," Visual and Audio-Visual Speech Processing, 2004.
- [9] "SpeechRecognition 3.8.1," Online Available: <https://pypi.org/project/SpeechRecognition/>.
- [10] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513-523, 1988.
- [11] "TfidfVectorizer," Online Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.
- [12] G. Salton, Automatic text processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley Longman Publishing Co., Inc., 1989.
- [13] L. Weng, "Object Detection Part 4: Fast Detection Models," Online Available: <https://lilianweng.github.io/lil-log/2018/12/27/object-detection-part-4.html>, Dec. 2018.
- [14] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [15] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," Communications of the ACM, vol. 15, no. 1, pp. 11-15, Jan. 1972.
- [16] "Arduino Mega 2560 Rev3," Online Available: <https://store.arduino.cc/products/arduino-mega-2560-rev3>.
- [17] A. Geiger, F. Moosmann, Ö. Car and B. Schuster, "Automatic camera and range sensor calibration using a single shot," in 2012 IEEE International Conference on Robotics and Automation, Jun. 2012.
- [18] G. Bradski, "The openCV library," Dr. Dobb's Journal: Software Tools for the Professional Programmer, vol. 25, no. 11, pp. 120-123, 2000.