NHẬN DẠNG HÌNH TRẠNG BÀN TAY SỬ DỤNG THUẬT TOÁN YOLOv7

Nguyễn Thị Thanh Tâm*, Nguyễn Thị Tính+
*Học viện Công nghệ Bưu chính Viễn thông
+Trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên

Tóm tắt: Bên cạnh lời nói, cử chỉ nói chung và cử chỉ tay nói riêng là một trong những hình thức giao tiếp phổ biến nhất. Cử chỉ tay có thể truyền đạt nhiều nội dung một cách trực quan. Bài toán nhận dạng cử chỉ tay đã thu hút sự quan tâm nghiên cứu trong lĩnh vực thị giác máy tính những năm gần đây. Tuy nhiên, bài toán này vẫn còn một số thách thức bởi tương tác người-máy dùng cử chỉ tay cần tự nhiên, độ chính xác nhận dạng cao và thời gian đáp ứng nhanh. Trong bài báo này, chúng tôi đề xuất sử dụng thuật toán You Only Look Once phiên bản 7 (YOLOv7) cho bài toán nhận dạng hình trạng bàn tay (còn gọi là cử chỉ tĩnh). Thực nghiệm được tiến hành với tập hình trạng bàn tay trong trò chơi oẳn tù tì. Kết quả thực nghiệm cho thấy phương pháp nhận dạng hình trạng bàn tay sử dụng thuật toán YOLOv7 cho hiệu suất tốt hơn cả về tốc độ tính toán và độ chính xác so với hai phương pháp sử dụng thuật toán YOLOv5 và sử dụng Faster R-CNN.

Từ khóa: Tương tác người-máy, Thị giác máy tính, Nhận dạng cử chỉ, Học sâu, YOLO.

I. GIỚI THIỆU

Nhu cầu tương tác giữa con người và máy tính ngày càng được mở rộng trong những ngữ cảnh sử dụng đa dạng khác nhau với yêu cầu về sự thuận tiện, tự nhiên. Với nhu cầu này, nhiều phương pháp và kỹ thuật mới đã và đang được nghiên cứu, phát triển. Nhận dạng cử chỉ là một trong những bài toán được quan tâm nghiên cứu trong đó các phương pháp sử dụng công nghệ thị giác máy tính và trí tuệ nhân tạo đã đạt được nhiều thành tựu cả trên phương diện nghiên cứu lý thuyết và ứng dụng thực tiễn [1]. Các bài toán cụ thể của nhận dạng cử chỉ tay bao gồm nhận dạng ngôn ngữ ký hiệu [2], nhận dạng ngôn ngữ tín hiệu đặc biệt được sử dụng trong thể thao [3], nhận dạng hoạt động [4], phát hiện tư thế/hình trạng [5], theo dõi/đánh giá hoạt động tập thể dục [6], và điều khiển nhà thông minh/các ứng trong dụng trong đời sống hàng ngày sử dụng cử chỉ tay [7].

Trong những năm qua, các nhà khoa học đã sử dụng các thuật toán, các phương pháp khác nhau để giải quyết các bài toán nêu trên nhằm mang lại những ứng dụng hữu ích trong đời sống [8]. Việc sử dụng cử chỉ tay trong các ứng dụng khác nhau đã góp phần cải thiện việc tương tác giữa người-máy [9]. Trong đó, sự phát triển của các hệ thống nhận dạng cử chỉ đóng một vai trò quan trọng. Cử chỉ tay được sử dụng ngày càng rộng rãi trong các lĩnh vực khác nhau. Chúng ta có thể thấy cử chỉ tay được ứng dụng trong các ứng dụng đa phương tiện như trò chơi [10], thực tế ảo [11] và thực tế tăng cường [12], hỗ trợ sinh hoạt [13], đánh giá phát triển nhận thức [14], v.v. Gần đây, cử chỉ tay cũng được sử dụng trong tương tác giữa con người và robot trong môi trường sản xuất [15] và ứng dụng trong xe tự hành [16].

Nhận dạng cử chỉ tay là kỹ thuật trong đó chúng ta sử dụng các thuật toán, các phương pháp trong các lĩnh vực như xử lý hình ảnh, thị giác máy tính, học máy, học sâu để hiểu hình trạng và chuyển động của bàn tay [17]. Cử chỉ tay được chia làm hai loại, cử chỉ tĩnh và cử chỉ động. Trong bài báo này, chúng tôi tập trung nghiên cứu bài toán nhận dạng cử chỉ tĩnh (nhận dạng hình trạng bàn tay).

Đã có nhiều thuật toán nhận dạng hình trạng bàn tay được đề xuất [18]. Một trong những phương pháp gần đây cho kết quả tốt là phương pháp sử dụng thuật toán YOLOv5 cho bài toán phát hiện và nhận dạng ngôn ngữ ký hiệu Mỹ [19]. Có nhiều phiên bản cải tiến của thuật toán YOLO được giới thiệu gần đây trong đó mới nhất là YOLOv7 [20]. YOLOv7 là một mô hình mạng nơ-ron tích chập đã được chứng minh là đạt hiệu quả tốt trong bài toán phát hiện đối tượng thời gian thực [20]. Tuy nhiên, theo khảo sát của chúng tôi, chưa có nghiên cứu nào sử dụng thuật toán YOLOv7 cho bài toán nhận dạng hình trạng bàn tay. Vì vậy, chúng tôi đề xuất sử dụng thuật toán YOLOv7 cho bài toán này. Dưới đây là những đóng góp chính của nghiên cứu này:

1) Đề xuất phương pháp nhận dạng hình trạng bàn tay sử dụng thuật toán YOLOv7, một thuật toán phát hiện đối tượng hiệu quả được giới thiệu gần đây.

2) Nghiên cứu khảo sát so sánh hiệu quả của phương pháp nhận dạng hình trạng bàn tay sử dụng thuật toán YOLOv7 với hai phương pháp sử dụng YOLOv5 và sử dụng Faster R-CNN.

Nội dung còn lại của bài báo được chia làm 4 phần. Phần II trình bày tóm lược một số nghiên cứu liên quan. Phần III trình bày phương pháp nhận dạng hình trạng bàn tay sử dụng thuật toán YOLOv7. Phần IV trình bày kết quả thực nghiệm. Cuối cùng, kết luận và hướng nghiên cứu tiếp theo được trình bày ở phần V.

II. MỘT SỐ NGHIÊN CỨU LIÊN QUAN

Đã có nhiều phương pháp được đề xuất cho bài toán nhận dạng hình trạng bàn tay [18]. Trong các hướng tiếp cận, gần đây các phương pháp sử dụng học sâu thu được kết quả ấn tượng. Trong phần này, chúng tôi chủ yếu tập trung điểm qua một số phương pháp dựa trên học sâu để phát hiện đối tượng nói chung và nhận dạng hình trạng bàn tay nói riêng.

Tác giả liên hệ: Nguyễn Thị Thanh Tâm,
Email: ntttam@ptit.edu.vn
Đến tòa soạn: 9/2022, chỉnh sửa: 10/2022, chấp nhận đăng: 10/2022.

Một số mạng học sâu có kết quả tốt trong bài toán phát hiện các đối tượng, chẳng hạn như AlexNet [21], VGGNet [22], ResNet [23], v.v. Ngoài ra còn có một số các mô hình dựa trên việc đề xuất vùng như R-CNN [24], Fast R-CNN [25] và Faster R-CNN [26]. Các thuật toán này bao gồm hai giai đoạn: đề xuất vùng có khả năng xuất hiện đối tượng và sau đó xác định xem có thực sự có đối tượng trong vùng đó hay không. Có một số phương pháp áp dụng những mô hình CNN này để nhận dạng cử chỉ tay. [27] sử dụng kiến trúc AlexNet cho bài toán nhận dạng ngôn ngữ ký hiệu Mỹ (ASL). [28] sử dụng mô hình VGG-19 cho bài toán phân lớp hình trạng bàn tay dựa trên dữ liệu ảnh độ sâu.

Trong bài báo [29], các tác giả trình bày một phương pháp nhận dạng hình trạng tay trong ngôn ngữ cử chỉ Nhật Bản trên dữ liệu ảnh RGB. Phương pháp này sử dụng sử dụng mạng ResNet. Rahaf Abdulaziz Alawwad và cộng sự [30] giới thiệu một hệ thống nhận dạng Ngôn ngữ ký hiệu Ả Rập sử dụng Faster R-CNN. Phương pháp này đạt độ chính xác 93% trên tập dữ liệu nhóm nghiên cứu tự thu thập với nền phức tạp.

Bên cạnh những phương pháp nêu kể trên, gần đây, các biến thể mới của thuật toán YOLO liên tục được giới thiệu với hiệu năng ngày càng được cải thiện. Từ đó, các phương pháp nhận dạng hình trạng bàn tay cũng nhanh chóng sử dụng các phiên bản thuật toán YOLO mới nhất. Tại hội thảo Chinese Control Conference lần thứ 41 diễn ra từ 25-27 tháng 7 năm 2022 vừa qua, Guangxiang Li và cộng sự [31] đã trình bày phương pháp nhận dạng cử chỉ tay sử dụng thuật toán YOLOv5. Phương pháp này đạt mAP (0.5) bằng 0.995, mAP (0.5) value reaches 0.995, mAP (0.5:0.95) bằng 0.865, và F1-score là 0.96, một kết quả ấn tượng trên tập cử chỉ tay trong một hệ thống tự checkout không tiếp xúc trong ngữ cảnh cần giữ khoảng cách vì bệnh dịch Covid-19.

Chưa dừng lại ở đó, gần đây Chien-Yao Wang và công sự [20] giới thiệu thuật toán YOLOv7 với những cải tiến và hiệu năng tốt hơn những phiên bản trước. Tuy nhiên, chúng tôi chưa thấy có nghiên cứu nào sử dụng thuật toán YOLOv7 cho bài toán nhận dạng hình trạng bàn tay. Đây là động lực để chúng tôi tiến hành nghiên cứu này.

III. HỆ THỐNG NHẬN DẠNG HÌNH TRẠNG BÀN TAY SỬ DỤNG THUẬT TOÁN YOLOv7

Hình 1 trình bày sơ đồ tổng quát hệ thống nhận dạng hình trạng bàn tay sử dụng thuật toán YOLOv7. Chúng tôi sử dụng thuật toán YOLOv7 [20] bởi đây là biến thể mới nhất của họ thuật toán YOLO và được báo cáo là có hiệu năng tốt nhất cả về thời gian tính toán cũng như độ chính xác.

Trong sơ đồ này, backbone trích xuất đặc trưng của ảnh đầu vào. Đặc trưng được trích xuất bởi backbone là đầu vào của neck. Neck nhận đầu đầu vào là đặc trưng do Backbone trích xuất và tạo ra các đặc trưng kim tự tháp. Cuối cùng, head là các lớp đầu ra.

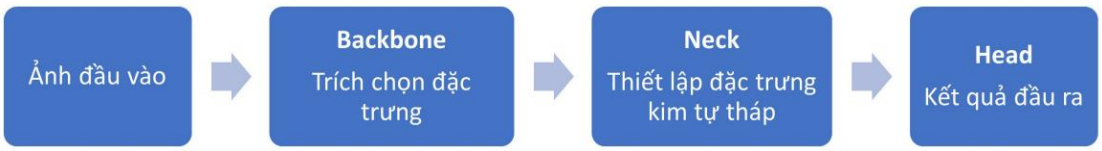
YOLOv7 cải thiện tốc độ và độ chính xác bằng cách cải tiến kiến trúc mạng. Tương tự như Scaled YOLOv4 [32], YOLOv7 không sử dụng các ImageNet pre-trained backbone mà huấn luyện trên bộ dữ liệu COCO [33]. Về kiến trúc, YOLOv7 sử dụng E-ELAN (Extended Efficient Layer Aggregation Network), Hình 2 [34] và kĩ thuật Model Scaling [35] để khuếch đại độ lớn của model giúp đạt được hiệu năng tốt hơn.

E-ELAN không thay đổi con đường lan truyền gradient của kiến trúc ban đầu mà sử dụng phép tích chập nhóm để làm giầu đặc trưng bằng cách thêm các đặc trưng và kết hợp các đặc trưng của các nhóm khác nhau theo cách xáo trộn và hợp nhất. Phương pháp này có thể tăng cường các đặc trưng được học bởi các bản đồ đặc trưng và cải thiện việc sử dụng các tham số và tính toán.

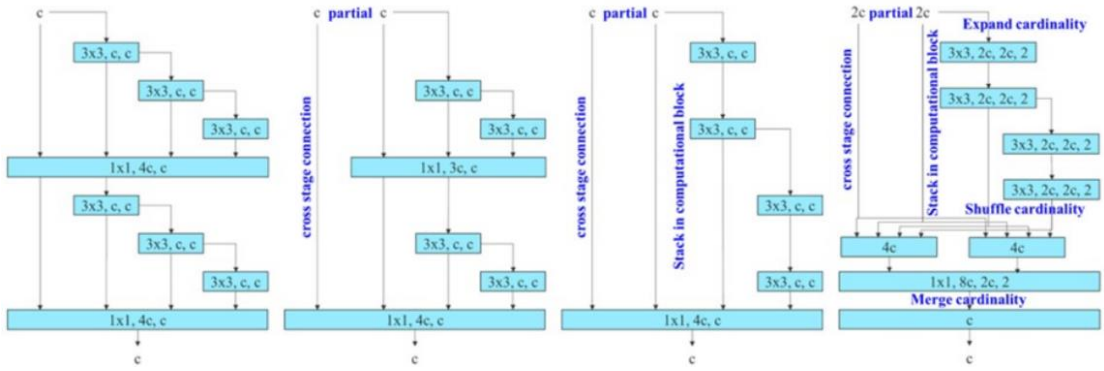
Các ứng dụng khác nhau yêu cầu các mô hình khác nhau. Trong khi một số ứng dụng cần các mô hình có độ chính xác cao, một số ứng dụng khác lại ưu tiên tốc độ. Kỹ thuật Model Scaling được thực hiện để điều chỉnh mô hình phù hợp với các yêu cầu này và phù hợp với các thiết bị tính toán khác nhau. Trong khi điều chỉnh kích thước mô hình, các tham số sau được xem xét bao gồm: Resolution (độ phân giải, kích thước của hình ảnh đầu vào), Width (số kênh), Depth (số lớp), Stage (số đặc trưng kim tự tháp). NAS (Network Architecture Search) [36] là phương pháp Model Scaling được sử dụng phổ biến. Trong YOLOv7, cách tiếp cận compound model scaling được sử dụng để đạt được kết quả tối ưu hơn, Hình 3.

Ngoài ra, YOLOv7 còn sử dụng kỹ thuật BoF (Bag of Freebies) để nâng cao hiệu suất của mô hình trong khi không tăng chi phí huấn luyện. Bag of freebies là những phương pháp chủ yếu thay đổi chiến lược huấn luyện hoặc do đó chủ yếu làm tăng chi phí huấn luyện trong khi hầu như không làm tăng chi phí dự đoán nhưng lại cải thiện đáng kể độ chính xác. Cụ thể, YOLOv7 áp dụng Label Assignment với việc sử dụng ground truth và prediction từ Head để tạo ra target, từ đó tính loss giữa target và prediction của model. Thay vì sử dụng prediciton của Auxiliary head để tạo ra target cho nó, trong YOLOv7, Auxiliary head sử dụng prediction của Lead Head để tạo ra target (Hình 4).

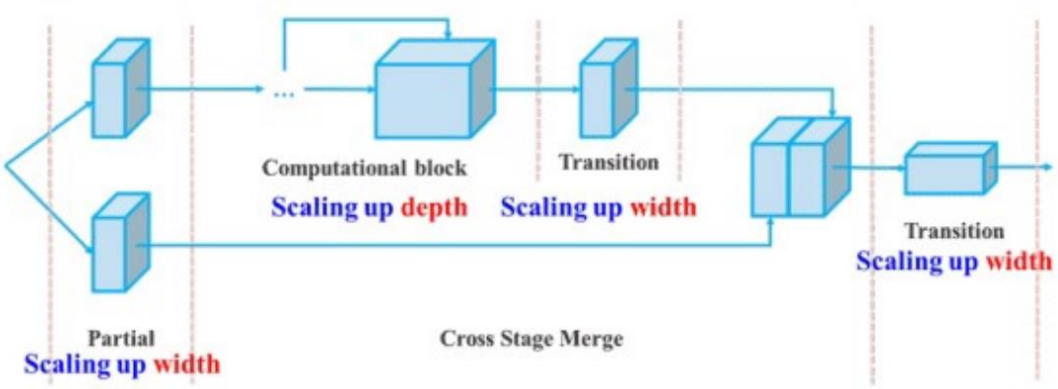
Bên cạnh đó, thay vì sử dụng cùng một target được tạo ra từ Lead Head, YOLOv7 tạo ra 2 target khác nhau từ Lead Head, một cho bản thân Lead Head (fine label), một cho Aux Head (coarse label) (Hình 4). Tất cả các Aux Head của YOLOv7 đều được áp dụng ở trong Neck.



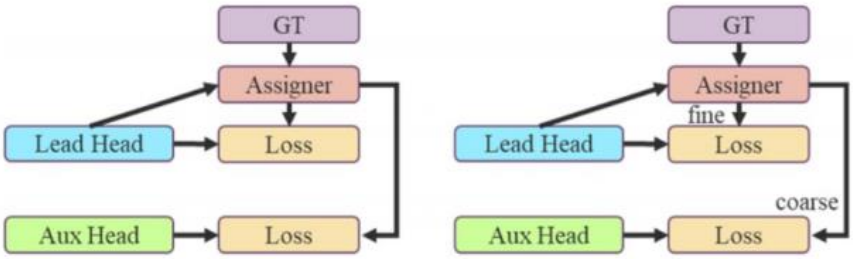
Hình 1. Sơ đồ tổng quát hệ thống nhận dạng hình trạng bàn tay sử dụng thuật toán YOLOv7



Hình 2. Backbone E-ELAN trong YOLOv7 [20]



Hình 3. Compound scaling trong YOLOv7 [20]



Hình 4. Compound scaling trong YOLOv7 [20]

IV. THỰC NGHIỆM

A. Cơ sở dữ liệu và độ đo đánh giá

Chúng tôi sử dụng một cơ sở dữ liệu mở Rock Paper Scissors trên Roboflow Universe, [37]. Cơ sở dữ liệu này bao gồm 4 hình trạng bàn tay (Love, Paper, Rock, và Scissors, Hình 5) trong trò chơi Oẳn tù tì (Paper Rock Scissors), một ví dụ điển hình cho hệ thống tương tác người-máy sử dụng cử chỉ tay.

Cơ sở dữ liệu Rock Paper Scissors có tổng số 928 ảnh, được chia thành 3 phần dành để train, validate, và test. Chi tiết thông tin được trình bày trong Bảng I. Chúng tôi lựa chọn tiến hành thực nghiệm trên cơ sở dữ liệu với số lượng ảnh không lớn do hạn chế về tài nguyên sử dụng cho việc huấn luyện mô hình.

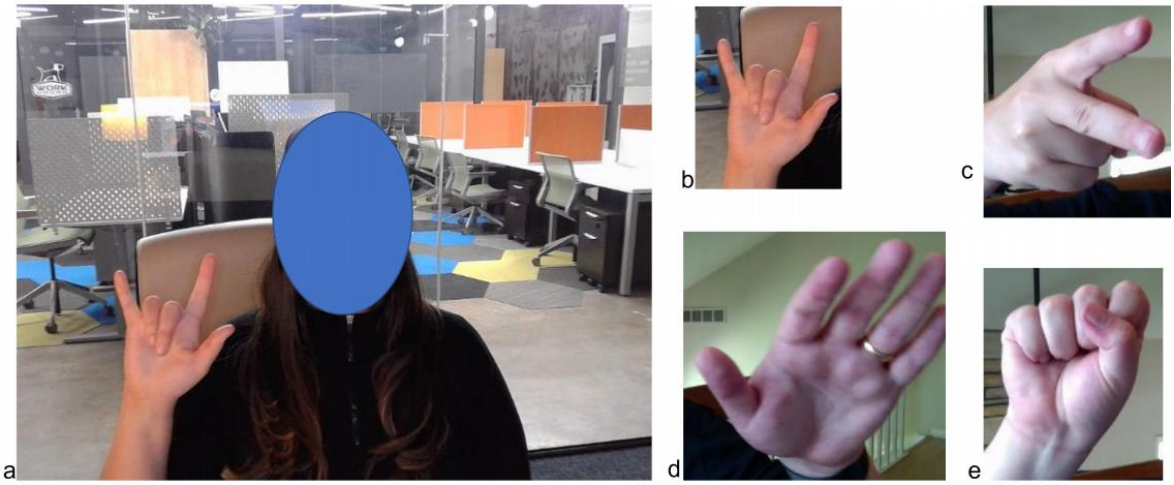
Để đánh giá phương pháp, chúng tôi sử dụng các độ đo đánh giá IoU (Intersection over Union), Precision, Recall, F-score, và mAP.

B. Kết quả

Chúng tôi sử dụng Google Colaboratory miễn phí để tiến hành transfer learning với mô hình pre-train trên tập dữ liệu COCO. Mô hình tốt nhất được lưu lại từ việc huấn luyện qua 55 epochs. Kết quả thử nghiệm trên tập test được trình bày trong các hình 6, 7, 8.

Hình 6 là đường cong thể hiện mối quan hệ giữa Precision và Recall của từng lớp và của tất cả các lớp. Hình 7 cho thấy biến thiên của F1- score tương ứng với sự thay đổi của ngưỡng độ tin cậy (confidence). F1-score tính trên tất cả các lớp đạt giá trị tốt nhất là 0.91 ứng với ngưỡng độ tin cậy bằng 0.484. Ở bên phải giá trị cực đại này, khi ngưỡng độ tin cậy yêu cầu đối với kết quả phát hiện càng cao thì Recall càng giảm còn Precision càng tăng nhưng F1-score giảm. Tương tự, ở bên trái giá trị cực đại này, ngưỡng độ tin cậy càng giảm thì Recall càng tăng và Precision giảm, kết quả là F1-score giảm.

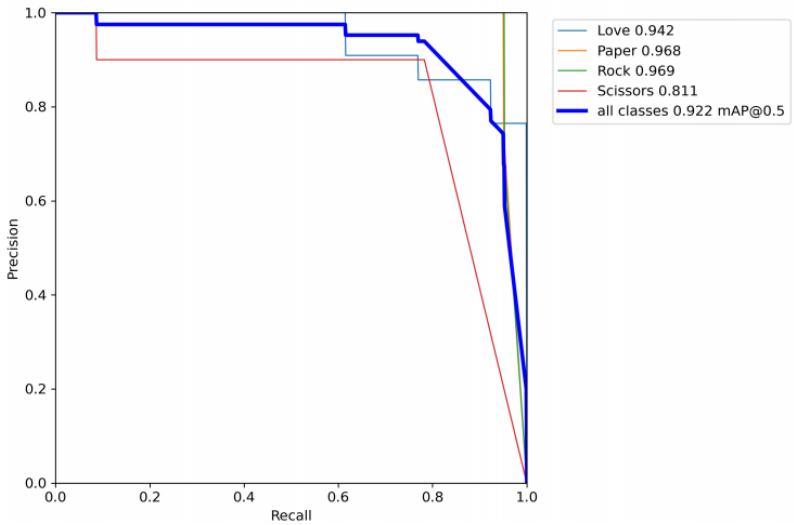
Hình 8 trình bày chi tiết ma trận nhầm lẫn. Nhìn vào hình này ta thấy mô hình nhầm lẫn nhiều nhất là việc phát hiện nhầm vùng nền thành hình trạng *Paper*. Điều này có thể là do hình trạng *"Paper"* thực hiện bằng cách xòe phẳng bàn tay ra nên có ít đặc trưng để phân biệt.



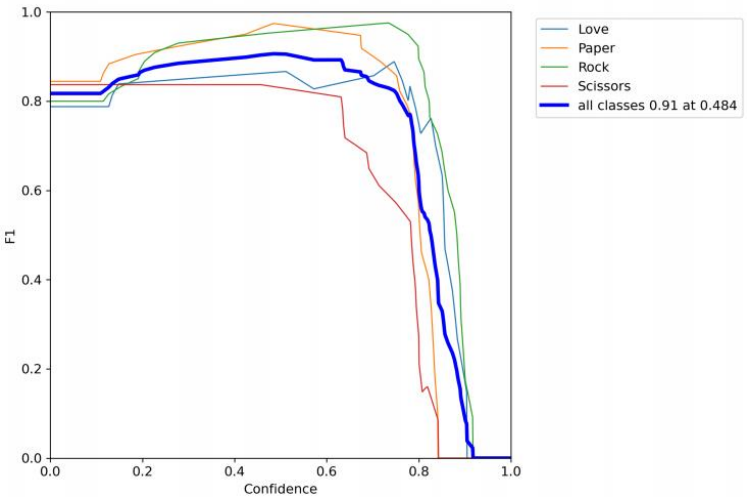
Hình 5. Cơ sở dữ liệu thử nghiệm, [37]. (a) Ảnh ví dụ. (b) Love. (c) Scissors. (d) Paper. (e) Rock.

Bảng I. THÔNG TIN CƠ SỞ DỮ LIỆU ROCK PAPER SCISSORS [37]

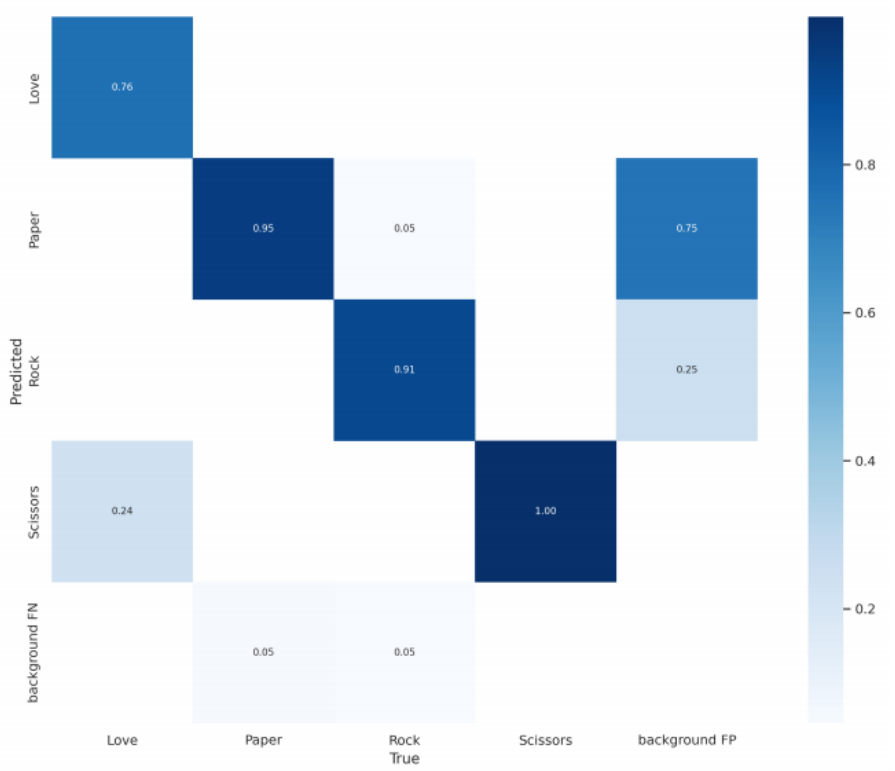
	Tổng số ảnh	*Số ảnh trong tập train*	*Số ảnh trong tập validation*	*Số ảnh trong tập test*
Số lượng	*928*	*665*	*174*	*89*
Tỉ lệ %	*100*	*71*	*19*	*10*



Hình 6. Đường cong Precision-Recall.



Hình 7. Đường cong F1-score.



Hình 8. Ma trận nhầm lẫn.

C. So sánh với phương pháp khác

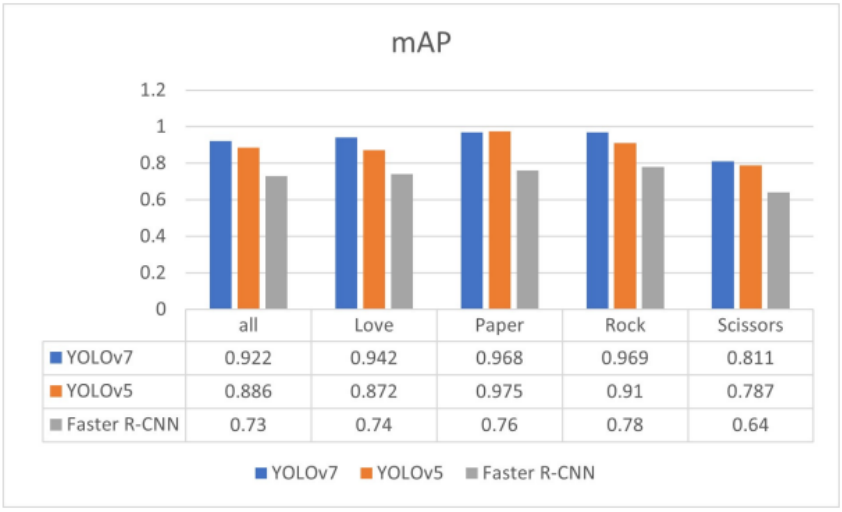
Để so sánh sự cải tiến của việc sử dụng thuật toán YOLOv7 so với các thuật toán trước đây, chúng tôi tiến hành huấn luyện hai thuật toán YOLOv5 và Faster R-CNN [26] với các lựa chọn tham số tốt nhất. Chúng tôi lựa chọn so sánh với phương pháp sử dụng thuật toán YOLOv5 [19] và sử dụng Faster R-CNN [30] bởi vì đây là hai trong số các thuật toán phát hiện đối tượng tốt nhất thuộc hai hướng tiếp cận: one-stage object detection (phương pháp một giai đoạn) và two-stage object detection (phương pháp hai giai đoạn), [38], [31], [30].

1) So sánh về độ chính xác

Hình 9 so sánh kết quả tốt nhất theo chỉ số mAP của phương pháp đề xuất sử dụng YOLOv7 với phương pháp sử dụng YOLOv5 và Faster R-CNN trên từng lớp và trên tất cả các lớp. Ta thấy phương pháp sử dụng YOLOv7 cho kết quả tốt hơn với mAP trung bình tăng từ 0.73 và 0.886 lên 0.922 so với phương pháp sử dụng Faster R-CNN và YOLOv5.

2) So sánh về thời gian tính toán

Về khía cạnh thời gian tính toán khi nhận dạng, phương pháp sử dụng YOLOv7 có tốc độ xử lý nhanh hơn phương pháp sử dụng YOLOv5. Cụ thể, phương pháp sử dụng YOLOv7 cần 15.1ms trong khi phương pháp sử dụng YOLOv5 cần 231.2ms để xử lý một ảnh kích thước 416x416. So với phương pháp sử dụng Faster R-CNN, phương pháp đề xuất nhanh hơn khoảng 17 lần.



Hình 9. So sánh mAP giữa phương pháp nhận dạng hình trạng bàn tay sử dụng YOPOv7 với phương pháp sử dụng YOPOv5 và Faster R-CNN.

V. KẾT LUẬN VÀ HƯỚNG NGHIÊN CỨU TIẾP THEO

Trong bài báo này, chúng tôi đã đề xuất sử dụng thuật toán YOLOv7 cho bài toán nhận dạng hình trạng bàn tay. Chúng tôi đã tiến hành thực nghiệm trên một tập dữ liệu công khai với những phân tích so sánh hiệu năng của phương pháp nhận dạng hình trạng bàn tay sử dụng YOLOv7 với hai phương pháp sử dụng YOLOv5 và sử dụng Faster R-CNN. Kết quả thực nghiệm cho thấy phương pháp sử dụng thuật toán YOLOv7 cho hiệu suất tốt hơn cả về tốc độ tính toán và độ chính xác.

Trong tương lai, chúng tôi sẽ tiếp tục nghiên cứu, khảo sát trên nhiều tập dữ liệu với số lượng dữ liệu lớn hơn đồng thời so sánh với nhiều phương pháp khác nhau.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Học viện Công nghệ Bưu chính Viễn thông cơ sở tại Hà Nội trong đề tài có mã số 06-2022-HV-ĐPT-PT.

TÀI LIỆU THAM KHẢO

[1] T. Wang, Y. Li, J. Hu, A. Khan, L. Liu, C. Li, A. Hashmi, and M. Ran, "A survey on vision-based hand gesture recognition," in *International Conference on Smart Multimedia*. Springer, 2018, pp. 219–231.

[2] S. Sharma and S. Singh, "Vision-based hand gesture recognition using deep learning for the interpretation of sign language," *Expert Systems with Applications*, vol. 182, p. 115657, 2021.

[3] J. ˇZemgulys, V. Raudonis, R.Maskeli unas, and R. Damasevi ˇ ˇcius, "Recognition of basketball referee signals from real-time videos," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 3, pp. 979–991, 2020.

[4] C. Pham, L. Nguyen, A. Nguyen, N. Nguyen, and V.-T. Nguyen, "Combining skeleton and accelerometer data for human fine-grained activity recognition and abnormal behaviour detection with deep temporal convolutional networks," *Multimedia Tools and Applications*, vol. 80, no. 19, pp. 28 919–28 940, 2021.

[5] T.-H. Tran, D. T. Nguyen, and T. P. Nguyen, "Human posture classification from multiple viewpoints and application for fall detection," in *2020 IEEE Eighth International Conference on Communications and Electronics (ICCE)*. IEEE, 2021, pp. 262–267.

[6] P. Movva, H. Pasupuleti, and H. Sarma, "A self learning yoga monitoring system based on pose estimation," in *International Conference on Human-Computer Interaction*. Springer, 2022, pp. 81–91.

[7] H.-G. Doan, T.-H. Tran, H. Vu, T.-L. Le, V.-T. Nguyen, S. V. Dinh, T.-O. Nguyen, T.-T. Nguyen, and D.-C. Nguyen, "Multi-view discriminant analysis for dynamic hand gesture recognition," in *Asian Conference on Pattern Recognition*. Springer, 2019, pp. 196–210.

[8] P. N. Huu, Q. T. Minh *et al.*, "An ann-based gesture recognition algorithm for smart-home applications," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 14, no. 5, pp. 1967–1983, 2020.

[9] B. K. Chakraborty, D. Sarma, M. K. Bhuyan, and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human–computer interaction," *IET Computer Vision*, vol. 12, no. 1, pp. 3–15, 2018.

[10] K. Aggarwal and A. Arora, "Hand gesture recognition for real-time game play using background elimination and deep convolution neural network," in *Virtual and Augmented Reality for Automobile Industry: Innovation Vision and Applications*. Springer, 2022, pp. 145–160.

[11] T. Wang, X. Qian, F. He, X. Hu, Y. Cao, and K. Ramani, "Gesturar: An authoring system for creating freehand interactive augmented reality applications," in *the 34th Annual ACM Symposium on User Interface Software and Technology*, 2021, pp. 552–567.

[12] L. T. De Paolis, S. T. Vite, M. Á. P. Casta˜neda, C. F. Dominguez Velasco, S. Muscatello, and A. F. Hernández Valencia, "An augmented reality platform with hand gestures-based navigation for applications in image-guided surgery: prospective concept evaluation by surgeons," *International Journal of Human– Computer Interaction*, vol. 38, no. 2, pp. 131–143, 2022.

[13] D. O. Anderez, A. Lotfi, and C. Langensiepen, "A hierarchical approach in food and drink intake recognition using wearable inertial sensors," in *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, 2018, pp. 552–557.

[14] F. Negin, P. Rodriguez, M. Koperski, A. Kerboua, J. Gonzàlez, J. Bourgeois, E. Chapoulie, P. Robert, and F. Bremond, "Praxis: Towards automatic cognitive assessment using gesture recognition," *Expert systems with applications*, vol. 106, pp. 21–35, 2018.

[15] J. Berg, A. Lottermoser, C. Richter, and G. Reinhart, "Human-robot-interaction for mobile industrial robot teams," *Procedia CIRP*, vol. 79, pp. 614–619, 2019.

[16] G. Young, H. Milne, D. Griffiths, E. Padfield, R. Blenkinsopp, and O. Georgiou, "Designing mid-air haptic gesture controlled user interfaces for cars," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. EICS, pp. 1–23, 2020.

[17] L. Zulpukharkyzy Zholshiyeva, T. Kokenovna Zhukabayeva, S. Turaev, M. Aimambetovna Berdiyeva, and D. Tokhtasynovna Jambulova, "Hand gesture recognition methods and applications: A literature survey," in t*he 7th International Conference on Engineering & MIS 2021*, 2021, pp. 1–8.

[18] S. Anwar, S. K. Sinha, S. Vivek, and V. Ashank, "Hand gesture recognition: a survey," in *Nanoelectronics, Circuits and Communication Systems*. Springer, 2019, pp. 365–371.

[19] T. F. Dima and M. E. Ahmed, "Using yolov5 algorithm to detect and recognize american sign language," in *2021 International Conference on Information Technology (ICIT)*. IEEE, 2021, pp. 603–607.

[20] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-ofthe-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[25] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[27] A. A. Barbhuiya, R. K. Karsh, and S. Dutta, "Alexnetcnn based feature extraction and classification of multiclass asl hand gestures," in *Proceeding of Fifth International Conference on Microelectronics, Computing and Communication Systems*. Springer, 2021, pp. 77–89.

[28] S. Amir *et al.*, "Hand posture classification with convolutional neural networks on vgg-19 net architecture," in *IOP Conference Series: Earth and Environmental Science*, vol. 575, no. 1. IOP Publishing, 2020, p. 012186.

[29] J. Qi, K. Xu, and X. Ding, "Approach to hand posture recognition based on hand shape features for human–robot interaction," *Complex & Intelligent Systems*, vol. 8, no. 4, pp. 2825–2842, 2022.

[30] R. A. Alawwad, O. Bchir, and M. M. B. Ismail, "Arabic sign language recognition using faster r-cnn," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 3, 2021.

[31] G. Li, D. Li, and A. Yang, "Real-time hand gesture detection based on yolov5s," in *2022 41st Chinese Control Conference (CCC)*. IEEE, 2022, pp. 7047– 7052.

[32] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[34] Y. LeCun *et al.*, "Generalization and network design strategies," *Connectionism in perspective*, vol. 19, no. 143-155, p. 18, 1989.

[35] P. Dollár, M. Singh, and R. Girshick, "Fast and accurate model scaling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 924–932.

[36] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 1997–2017, 2019.

[37] T. Roboflow, "Rock paper scissors dataset," https://universe.roboflow.com/teamroboflow/rock-paper-scissors-detection, dec 2021, visited on 2022-11-02. [Online]. https://universe.roboflow.com/teamroboflow/rock-paper-scissors-detection.

[38] S. Saxena, A. Paygude, P. Jain, A. Memon, and V. Naik, "Hand gesture recognition using yolo models for hearing and speech impaired people," in *2022 IEEE Students Conference on Engineering and Systems (SCES)*. IEEE, 2022, pp. 1–6.

HAND POSTURE RECOGNITION USING YOLOv7 ALGORITHM

Abstract: Besides spoken words, body language, including hand gestures, is one of the most common communication modalities. Hand gestures can convey much content visually. The problem of hand gesture recognition has attracted research interest in computer vision in recent years. However, this problem still has challenges because human-machine interaction needs natural hand gestures, high recognition accuracy, and fast response time. In this paper, we propose to use the You Only Look Once algorithm version 7 (YOLOv7) for the hand posture recognition problem (also known as static hand gestures). The experiment was conducted with a set of hand postures Rock-Paper-Scissors game. Experimental results show that the hand posture recognition method using the YOLOv7 algorithm obtains better calculation speed and accuracy performance than the two methods using the YOLOv5 and Faster R-CNN algorithms.

Keywords: Human-Machine Interaction, Com puter Vision, Hand Posture Recognition, Deep Learning, YOLO.

Nguyễn Thị Thanh Tâm tốt nghiệp kỹ sư Công nghệ thông tin tại Đại học Thái Nguyên năm 2009. Năm 2017, nhận bằng Thạc sĩ khoa học ngành Hệ thống thông tin tại trường Đại học Công nghệ, Đại học Quốc gia Hà Nội. ThS Nguyễn Thị Thanh Tâm là giảng viên tại Khoa Đa phương tiện - Học viện Công nghệ Bưu chính Viễn thông (PTIT). Lĩnh vực nghiên cứu: Học máy, Phát triển Ứng dụng Đa phương tiện.

Email: ntttam@ptit.edu.vn

Nguyễn Thị Tính là giảng viên ngành Công nghệ thông tin tại Đại học Thái Nguyên (ICTU). Cô nhận bằng Kỹ sư Công nghệ Thông tin Đại học Thái Nguyên vào năm 2008. Sau đó, nhận bằng Thạc sĩ Công nghệ Thông tin tại Manuel S. Enverga University Foundation - Lucena City - Philippines vào năm 2010. Lĩnh vực nghiên cứu: trí tuệ nhân tạo và nhận dạng cử chỉ, máy học/học sâu.

Email: nttinh@ictu.edu.vn