PHÁT HIỆN VÀ PHÂN LOẠI NGƯỜI ĐI BỘ SỬ DỤNG PHƯƠNG PHÁP HỌC SÂU

PEDESTRIAN DETECTION AND CLASSIFICATION USING DEEP LEARNING

LÊ QUYẾT TIẾN[1*], NGUYỄN VĂN HÙNG[2], TRẦN THỊ HƯƠNG[1], NGUYỄN HỮU TUÂN[1]

[1]Khoa Công nghệ thông tin, Trường Đại học Hàng hải Việt Nam

[2]Học viên cao học ngành Công nghệ thông tin - Khóa 2020.1, Trường Đại học Hàng hải Việt Nam

**Email liên hệ: tienlqcnt@vimaru.edu.vn*

Tóm tắt

Trong nghiên cứu này, đóng góp chính của nhóm tác giả tập trung vào giải quyết bài toán phát hiện và phân loại người đi bộ (người trưởng thành hay trẻ em) trong hình ảnh dựa trên phương pháp học sâu theo hai hướng tiếp cận. Ở hướng thứ nhất, bài toán được chia thành hai bài toán thành phần: phát hiện người đi bộ và phân loại người đi bộ. Hình ảnh người đi bộ sẽ được tách ra từ hình ảnh đầu vào và đưa qua bộ phân loại để xác định người đi bộ đó là người lớn hay trẻ em. Cụ thể, bài toán phát hiện người đi bộ được nghiên cứu dựa trên mô hình phát hiện đối tượng YOLO trong khi bài toán phân loại hình ảnh người đi bộ được nghiên cứu trên mô hình VGG, Inception, ResNet và EfficientNet. Ở hướng tiếp cận thứ hai, bài toán được nghiên cứu theo hướng phát hiện và phân loại người đi bộ sử dụng duy nhất một mô hình cụ thể là mô hình phát hiện đối tượng YOLO. Kết quả thu được của nghiên cứu tương đối tốt với cả hai hướng tiếp cận. Hướng tiếp cận thứ nhất cho độ chính xác trung bình phát hiện người đi bộ đạt 0.797 và độ chính xác phân loại người đi bộ đạt 0.955. Tuy nhiên hướng tiếp cận thứ hai thể hiện sự vượt trội khi cho độ chính xác cao hơn đạt 0.851 đồng thời có thời gian thực thi tốt hơn nhiều so với hướng tiếp cận thứ nhất.

***Từ khóa**: Phát hiện đối tượng, phân loại hình ảnh, người đi bộ, người lớn, trẻ em, học sâu.*

Abstract

In this study, the main contribution is to solve the task of pedestrian detection and adult / kid classification by using two approaches. In the first one, the task is divided into two sub-tasks: pedestrian detection and adult / kid classification. Pedestrian image regions are cropped from input images and passed through a classifier to determine if they are adult images or kid images. Specifically, the pedestrian detection task is studied by using an object detection model YOLO while the classification task is studied by using typical deep models: VGG, Inception, ResNet and EfficientNet. In the second approach, only one object detection model, YOLO is used to detect and classify pedestrians. The obtained results are quite good for both approaches. The first one has a good mean average precision of the pedestrian detection task at 0.797 and the classification accuracy is 0.955. However, the second approach has much better results with a higher mean average precision 0.851 and a much better performing time compared to the first approach.

***Keywords**: Object detection, image classification, pedestrian, adult, kid, deep learning*

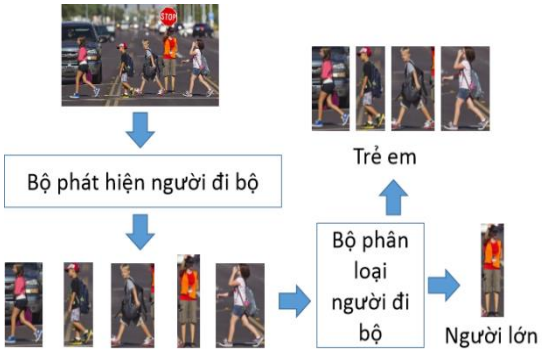
1. Giới thiệu

Ngày nay, tai nạn giao thông đã và vẫn đang là một vấn đề nổi cộm của xã hội. Theo thông tin từ Cục Cảnh sát giao thông - Bộ Công an, 6 tháng đầu năm 2021, toàn quốc xảy ra 6.278 vụ tai nạn giao thông, làm chết 3.147 người, bị thương 4.465 người. Nguyên nhân chủ yếu là do các lỗi vi phạm giao thông và thực trạng trên phản ánh tính phức tạp cũng như mức độ nguy hiểm trong việc tham gia giao thông tại Việt Nam. Các biện pháp hỗ trợ người tham gia giao thông đã và đang trở thành một nhu cầu cấp thiết nhằm giảm thiểu rủi ro tai nạn. Việc ứng dụng khoa học công nghệ để giải quyết vấn đề này hiện đang là hướng giải quyết có tiềm năng lớn.

Bên cạnh đó, cuộc cách mạng khoa học công nghệ đang diễn ra mạnh mẽ ở Việt Nam cũng như trên toàn thế giới. Việc triển khai các hệ thống camera hành trình trong tham gia giao thông và việc ứng dụng trí tuệ nhân tạo, thị giác máy tính vào cuộc sống đang ngày càng phổ biến hơn. Các camera hành trình thông thường chỉ có chức năng ghi lại hình ảnh mà chưa tận dụng được vào việc hỗ trợ người điều khiển phương tiện giao thông. Việc phát hiện người đi bộ và phân loại người đi bộ là người trưởng thành hay trẻ em là tiền đề cho nhiều giải pháp hỗ trợ giảm thiểu rủi ro tai nạn (cảnh báo người sang đường, cảnh báo trẻ em chạy phía trước,…).

TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

Bài toán phát hiện người đi bộ không phải một bài toán mới nhưng bài toán phân loại người đi bộ là người trưởng thành hay trẻ em hiện vẫn chưa được các nghiên cứu đi sâu. Trong bài báo này, một vấn đề chưa có câu trả lời được đưa ra: Việc tổng quát hóa các đặc trưng của người đi bộ nói chung (bao gồm cả người lớn và trẻ em) hay phân biệt hóa các đặc trưng của trẻ em và các đặc trưng của người lớn riêng rẽ hiệu quả hơn trong bài toán phát hiện người đi bộ? Nói cách khác, việc phân định riêng biệt trẻ em và người lớn có làm phức tạp hóa bài toán phát hiện người đi bộ và liệu có hiệu quả hơn khi tách biệt bài toán phát hiện người đi bộ và bài toán phân loại người đi bộ? Xuất phát từ vấn đề được nêu ra, bài toán phát hiện và phân loại người đi bộ được nghiên cứu theo hai hướng tiếp cận. Ở hướng tiếp cận thứ nhất, bài toán được chia thành bài toán phát hiện người đi bộ và bài toán phân loại người đi bộ (minh họa trong Hình 1). Cụ thể, khuôn hình người đi bộ sẽ được xác định và trích xuất ra từ hình ảnh đầu vào ở bước thứ nhất thông qua đặc trưng của người đi bộ nói chung (bao gồm cả người lớn và trẻ em). Ở bước thứ hai, khuôn hình trích xuất được sẽ được phân loại là người lớn hay trẻ em (thông qua các đặc trưng phân loại người lớn và trẻ em).



Hình 1. Hướng tiếp cận sử dụng bộ phát hiện và bộ phân loại người đi bộ riêng biệt

Trái ngược lại, trong hướng tiếp cận thứ hai, các đối tượng người lớn đi bộ và trẻ em đi bộ sẽ được phát hiện và phân loại trong một bước thực hiện thông qua đặc trưng người lớn đi bộ và trẻ em đi bộ như được minh họa trong Hình 2.

Hình 2. Hướng tiếp cận sử dụng bộ phát hiện và phân loại người đi bộ tích hợp

Trong nghiên cứu này, bài toán phát hiện và phân loại người đi bộ được tập trung nghiên cứu giải quyết. Đồng thời, các ưu nhược điểm của hai hướng tiếp cận bài toán trên cũng được nghiên cứu, đánh giá và so sánh để trả lời câu hỏi được đặt ra ban đầu.

2. Bối cảnh nghiên cứu

2.1. Bài toán phát hiện người đi bộ

Bài toán phát hiện người đi bộ là một bài toán thuộc họ các bài toán phát hiện đối tượng. Trong đó, phát hiện đối tượng là sự kết hợp của bài toán định vị đối tượng và phân loại đối tượng khi xác định khung bao quanh từng đối tượng trong hình đồng thời xác định lớp (nhãn) của đối tượng.

Các hướng giải quyết tiêu biểu trước đây cho bài toán phát hiện đối tượng có thể kể đến như "các biến thể Viola & Jones" [1], biểu đồ định hướng gradient (Histogram of Oriented Gradients - HOG) [2], bộ phát hiện phần biến dạng (Deformable Part Detectors - DPM) [3]. Ngày nay, các hướng tiếp cận mạng học sâu sử dụng mô hình CNN [8] đã và vẫn đang là hướng tiếp cận hiệu quả nhất cho bài toán phát hiện đối tượng nói chung và bài toán phát hiện người đi bộ nói riêng. Với hướng tiếp cận này có thể kể đến hai họ mô hình tiêu biểu là các mô hình R-CNN (Regions with Convolutional Neural Network - mạng nơ ron tích chập vùng) [4], [6], [7] và họ mô hình YOLO (You Only Look Once - bạn chỉ nhìn một lần) [9], [10], [11], [12].

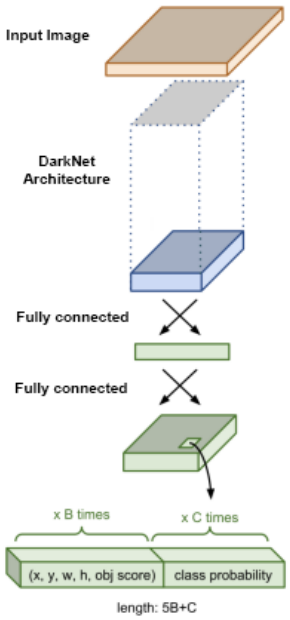
Họ mô hình R-CNN được đề cập tới với ba mô hình tiêu biểu là R-CNN [4], Fast R-CNN [6] và Faster R-CNN [7]. Mô hình R-CNN [4] bao gồm ba thành phần chính: Bộ đề xuất vùng (region proposal), bộ trích xuất đặc trưng (feature extractor) và bộ phân loại và điều chỉnh hồi quy (classifier and regressor). Trong đó, bộ đề xuất vùng chịu trách nhiệm đề xuất các vùng có thể chứa vật thể, các vùng này được giới hạn bởi các các hình chữ nhật gọi là hộp giới hạn (bounding box). Bộ trích xuất đặc trưng làm nhiệm vụ tính toán trích xuất các đặc trưng từ các vùng được đề xuất thông qua các mạng nơ ron tích chập. Cuối cùng bộ phân loại và điều chỉnh hồi quy sẽ phân loại hình ảnh chứa trong vùng đề xuất về đúng nhãn và điều chỉnh lại hộp giới hạn dựa trên các đặc trưng được trích xuất.

Mô hình Fast R-CNN [6] sau đó được phát triển lên từ mô hình R-CNN với sự thay đổi là bản đồ đặc trưng (feature map) được tính toán cho toàn bộ hình ảnh từ trước sau đó bản đồ đặc trưng cục bộ cho từng vùng đề xuất sẽ được trích xuất ra từ bản đồ đặc trưng toàn cục thông qua phép gộp vùng quan tâm (regions of interest pooling).

Mô hình Faster R-CNN [7] là sự nâng cấp từ mô hình Fast R-CNN khi sử dụng bản đồ đặc trưng toàn cục (được trích xuất cho toàn bộ hình ảnh) để đề xuất vùng ảnh thay vì sử dụng phương pháp tìm kiếm có chọn lọc (selective search) để đề xuất vùng ảnh như R-CNN và Fast R-CNN.

Nếu họ mô hình R-CNN thực hiện phát hiện đối tượng qua hai giai đoạn: Đề xuất vùng và phân loại vùng thì họ mô hình YOLO chỉ thực hiện công việc này qua một giai đoạn duy nhất. Có thể họ mô hình R-CNN trong một số trường hợp có thể có độ chính xác cao hơn nhưng xét về thời gian thực thi thì họ mô hình YOLO đang cho thấy sự khác biệt đáng kể khi các mô hình YOLO có thời gian thực thi nhỏ hơn nhiều so với họ mô hình R-CNN nhưng vẫn đảm bảo sự cân bằng với độ chính xác cao.

Mô hình YOLOV1 [9] hoạt động dựa trên ý tưởng như sau: Ảnh đầu vào được phân chia thành một lưới gồm nhiều ô, mỗi ô đảm nhận việc dự đoán các tọa độ và nhãn của hộp giới hạn có tâm nằm trong ô đó. Mô hình sử dụng một mạng học sâu để tính toán các bản đồ đặc trưng sau đó kết nối với các lớp kết nối đầy đủ (fully connected layer) để đưa ra nhãn, tọa độ và kích thước của các hộp giới hạn như trong Hình 3.

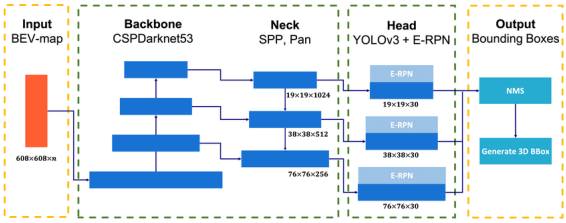


Hình 3. Kiến trúc mô hình YOLOV1

Mô hình YOLOV2 [10] được nâng cấp từ mô hình YOLOV1 với sự khác biệt cơ bản là sử dụng các lớp chuẩn hóa (normalization layers) và việc thay thế các lớp kết nối đầy đủ dự đoán trực tiếp ra tọa độ, kích thước các hộp giới hạn bởi các lớp hộp neo (anchor box layer) điều chỉnh tọa độ, kích thước của các hộp giới hạn.

Mô hình YOLOV3 [11] đưa ra một số thay đổi về kiến trúc của mạng tích chập so với YOLOV2 đồng thời việc phát hiện đối tượng trong hình ảnh sẽ được thực hiện nhiều lần, mỗi lần sử dụng kích thước khác nhau nhằm phát hiện đối tượng ở các tỷ lệ ảnh khác nhau.

Mô hình YOLOV4 [12] có những sự thay đổi đáng kể so với mô hình YOLOV3. Cụ thể mô hình YOLOV4 được chia thành ba thành phần chính bao gồm: Xương sống (backbone), cổ (neck) và đầu (head). Trong đó phần xương sống dùng để trích chọn đặc trưng, phần cổ dùng để trộn các bản đồ đặc trưng đã học được. Phần đầu trong YOLOV4 được chia thành hai phần bộ dự đoán dày đặc (dense prediction) và bộ dự đoán thưa thớt (sparse prediction). Trong đó bộ dự đoán dày đặc sử dụng các bộ phát hiện một giai đoạn và bộ dự đoán thưa thớt sử dụng các bộ dự đoán hai giai đoạn. Kiến trúc YOLOV4 được thể hiện trong Hình 4.



Hình 4. Kiến trúc mô hình YOLOV4

Hiện mô hình YOLOV5 đã được phát hành và đưa vào sử dụng. Mặc dù chưa có nhiều tài liệu chính thức về chi tiết mô hình nhưng YOLOV5 được đánh giá đem lại hiệu suất tốt cũng như đảm bảo về tốc độ.

2.2. Bài toán phân loại người đi bộ

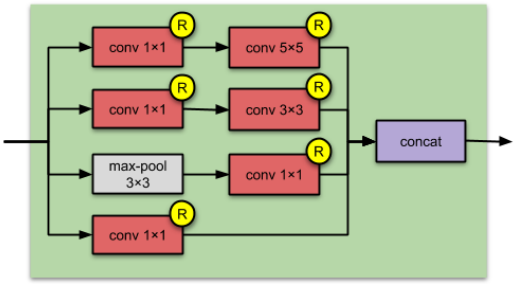
Bài toán phân loại người đi bộ thuộc vào dạng bài toán phân loại hình ảnh (image classification). Trong quá khứ, các bài toán phân loại hình ảnh chủ yếu được dựa trên các đặc trưng thủ công (handcrafted features) và có các kết quả không thật sự ấn tượng nhưng với sự ra đời của phương pháp học sâu, bài toán phân loại hình ảnh đang được giải quyết rất tốt với hiệu quả cao vượt trội [5]. Đã có rất nhiều các mô hình mạng học sâu được đưa ra và có thể kể đến một số mô hình tiêu biểu như LeNet, AlexNet [5], VGG [13], GoogLeNet [14], ResNet [15], EfficientNet [16].

Các mô hình LeNet, AlexNet [5] hay VGG [13] có kiến trúc chủ yếu bao gồm các lớp tích chập đơn thuần chịu trách nhiệm học các đặc trưng từ hình ảnh. Đầu ra của các lớp này được kết nối với các lớp kết nối đầy đủ để thực hiện các tác vụ (phân loại, hồi quy,…). Kiến trúc các mô hình này được đại diện bởi kiến trúc VGG16 được thể hiện trong Hình 5.

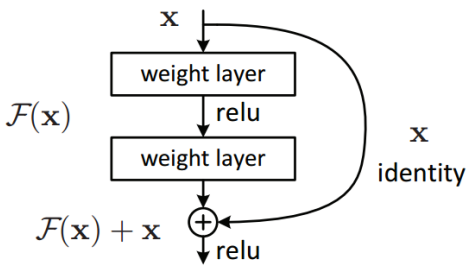
TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

Hình 5. Kiến trúc mô hình VGG16

Mô hình GoogLeNet hay Inception [14] được đưa ra với ý tưởng mở rộng mô hình mạng theo chiều rộng sử dụng các lớp tích chập thông thường kết hợp với các khối inception (hấp thu). Khối inception bao gồm các nhánh song song chứa các lớp tích chập với kích thước khác nhau. Kết quả tính toán từ các nhánh song song được ghép lại thành một đầu ra duy nhất (Hình 6).



Hình 6. Kiến trúc khối inception (hấp thu)



Hình 7. Kết nối tắt (skip connection)

Kiến trúc ResNet [15] được đưa ra kế thừa một số điểm trong kiến trúc khối của GoogLeNet nhưng sử dụng các kết nối tắt (skip connection). Kết nối tắt giữ thông tin không bị mất đi sau nhiều phép biến đổi bằng cách kết nối lớp phía trước với lớp phía sau không thông qua một vài lớp trung gian (Hình 7).

Mô hình EfficientNet [16] được tiếp cận theo hướng mới so với các mô hình trước đó. Mô hình tập trung vào việc mở rộng tham số theo cả ba chiều bao gồm độ sâu, độ rộng và độ phân giải của mạng. Mô hình cho phép giảm chi phí tính toán mà vẫn đảm bảo tính hiệu quả. Kiến trúc EfficientNet B0 được thể hiện trong Hình 8.

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224 × 224	32	1
2	MBConv1, k3x3	112 × 112	16	1
3	MBConv6, k3x3	112 × 112	24	2
4	MBConv6, k5x5	56 × 56	40	2
5	MBConv6, k3x3	28 × 28	80	3
6	MBConv6, k5x5	14 × 14	112	3
7	MBConv6, k5x5	14 × 14	192	4
8	MBConv6, k3x3	7 × 7	320	1
9	Conv1x1 & Pooling & FC	7 × 7	1280	1

Hình 8. Kiến trúc mô hình EfficientNet B0

Bài toán phân loại hình ảnh người đi bộ là người lớn hay trẻ em chưa được đưa ra nhiều trong các nghiên cứu nhiều trước đây. Trong [17], bài toán được thực hiện dựa trên việc tính toán tỷ lệ kích thước của khung giới hạn toàn bộ người và khung giới hạn phần mặt. Khung giới hạn cơ thể được xác định dựa trên biểu đồ định hướng gradient (HOG) [2] và khung giới hạn phần mặt được xác định dựa trên phương pháp Viola & Jones [1]. Tuy nhiên ý tưởng của mô hình đưa ra không thật sự tốt khi kích thước khung giới hạn cơ thể sẽ thay đổi tùy theo tư thế người chứ không cố định như người đứng thẳng. Trong bài báo này, bài toán sẽ được tập trung giải quyết dựa trên các mô hình học sâu.

3. Nghiên cứu bài toán phát hiện và phân loại người đi bộ

3.1. Hướng tiếp cận bài toán

3.1.1. Hướng sử dụng bộ phát hiện người đi bộ và bộ phân loại người đi bộ riêng biệt

Xuất phát từ câu hỏi liệu việc tổng quát hóa các đặc trưng cho người đi bộ nói chung có đơn giản và hiệu quả hơn phân biệt hóa đặc trưng cho người lớn đi bộ và trẻ em đi bộ, hướng tiếp cận thứ nhất (Hình 1) sử dụng mô hình phát hiện đối tượng YOLOV5 (một trong các mô hình điển hình nhất ở thời điểm hiện tại về phát hiện đối tượng thời gian thực) để phát hiện người đi bộ. Hình ảnh người đi bộ được trích xuất và

đưa qua bộ phân loại nhị phân để xác định đó là người lớn hay trẻ em sử dụng một bộ đặc trưng khác. Ở giai đoạn này, các mô hình điển hình cho tác vụ phân loại hình ảnh bao gồm mô hình VGG16 [13], ResNet50 [15], InceptionV3 [14] và EfficientNetB0 [16] được xem xét để thực hiện việc phân loại. Các mô hình trên được thay thế các lớp cuối cùng bởi ba lớp kết nối đầy đủ với số nơ ron lần lượt là 16, 16 và 1 để kết hợp các đặc trưng học được và thực hiện việc phân loại hình ảnh người lớn và trẻ em. Trong đó lớp cuối cùng chỉ có một đầu ra để thực hiện bài toán nhị phân trong khi số đầu ra của hai lớp trước đó không quá lớn để tránh hiện tượng overfitting.

3.1.2. Hướng sử dụng bộ phát hiện và phân loại người đi bộ tích hợp

Khác với hướng tiếp cận thứ nhất sử dụng các đặc trưng người đi bộ nói chung để tách vùng ảnh người đi bộ ra và sử dụng các đặc trưng phân biệt người lớn và trẻ em để phân biệt hình ảnh người đi bộ, cách tiếp cận thứ hai học trực tiếp các đặc trưng phát hiện người lớn đi bộ và trẻ em đi bộ (Hình 2). Do nghiên cứu hướng tới các giải pháp chạy thời gian thực nên mô hình YOLOV5 được lựa chọn để thực hiện công việc này.

3.2. Cài đặt, thực nghiệm và kết quả

3.2.1. Cài đặt và thực nghiệm

Chương trình thử nghiệm được cài đặt trên môi trường Google Colab với bộ xử lý đồ họa Nvidia K80 với bộ nhớ 12GB và tốc độ 0,82GHz sử dụng ngôn ngữ Python.

Bộ dữ liệu sử dụng trong thí nghiệm là Cityscapes [18] với hơn 2.700 hình ảnh chụp đường phố (chứa các phương tiện giao thông, người đi bộ,…) của hơn 20 thành phố khác nhau kết hợp với bộ dữ liệu do nhóm tác giả thu thập bao gồm khoảng 4000 hình ảnh người đi bộ. Tất cả người đi bộ trong hình đều được khoanh vùng và gán nhãn người lớn hoặc trẻ em trong đó tỷ lệ người đi bộ trẻ em và người lớn lần lượt là 44,8% và 55,2% (trên tổng số hơn 28.000 nhãn được gán).

Thí nghiệm thứ nhất được thực hiện để đánh giá hướng tiếp cận sử dụng bộ phát hiện và bộ phân loại người đi bộ riêng biệt (Hình 1), mô hình YOLOV5 được huấn luyện để phát hiện người đi bộ trên bộ dữ liệu gồm 5.464 hình ảnh và được đánh giá trên bộ dữ liệu gồm 1.193 hình ảnh (các hình ảnh này có kích thước 640x640) và được đánh dấu khoanh vùng và gán nhãn người đi bộ. Mô hình được huấn luyện trong 50 vòng (epoch) và dừng khi bị hiện tượng overfitting (khớp quá mức). Trong hướng tiếp cận này, các mô hình phân loại được huấn luyện với 22.660 hình ảnh và đánh giá trên 5.660 hình ảnh. Các hình ảnh này là hình ảnh trẻ em và người lớn được trích xuất ra từ các hình ảnh thuộc tập dữ liệu nêu trên và đưa về kích thước 128x128. Tương tự mô hình được huấn luyện trong 100 vòng với tỷ lệ học (learning rate) là 0,001 và thực tế được dừng lại sớm hơn nếu bị overfitting. Các mô hình phân loại được đánh giá bởi độ chính xác (accuracy - công thức (1)) được tính bằng tỷ lệ giữa số mẫu phân loại đúng (correct prediction number) trên tổng số mẫu phân loại (sample number).

$$accuracy = \frac{correct\ prediction\ number}{sample\ number} \quad (1)$$

Thí nghiệm thứ hai được thực hiện để đánh giá hướng tiếp cận sử dụng bộ phát hiện và phân loại người đi bộ tích hợp (Hình 2), mô hình YOLOV5 được huấn luyện để phát hiện và phân loại người lớn đi bộ và trẻ em đi bộ trên bộ dữ liệu gồm 5.464 hình ảnh và được đánh giá trên bộ dữ liệu gồm 1.193 hình ảnh (các hình ảnh này có kích thước 640x640) và được đánh dấu khoanh vùng người đi bộ đồng thời đánh nhãn là trẻ em hay người lớn. Tương tự như thí nghiệm thứ nhất, mô hình được huấn luyện trong 50 vòng và dừng khi bị hiện tượng overfitting. Các mô hình phát hiện đối tượng trong hai thí nghiệm được đánh giá bởi bộ ba giá trị bao gồm độ chính xác (precision - công thức (2)), chỉ số gợi nhớ (recall - công thức (3)) và giá trị chính xác trung bình (mAP - công thức (4)).

$$precsion = {TP}/{TP + FP} \quad (2)$$

$$recall = {TP}/{TP + FN} \quad (3)$$

Với TP (True Positive) là số dự đoán vùng đối tượng chính xác trong khi FP (False Positive) và FN (False Negative) là số dự đoán vùng đối tượng sai và số dự đoán vùng đối tượng bị sót.

Với mỗi phân lớp, một đường cong dựa trên chỉ số precision và recall được xác định và phần diện tích nằm dưới đường cong đó được đại diện bởi chỉ số AP (Average Precision). Chỉ số mAP được tính bằng trung bình giá trị AP của tất cả các phân lớp.

$$mAP = {1}/{N} \sum_{i=1}^{N} AP_i \quad (4)$$

TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

3.2.2. Kết quả

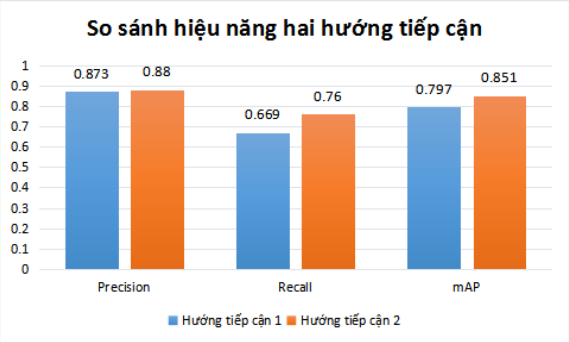
Bảng 1. Kết quả thực nghiệm bộ phát hiện và phân loại người đi bộ tách biệt

Mô hình YOLOv5 phát hiện người đi bộ	
Độ chính xác (precision)	0,873
Chỉ số gợi nhớ (recall)	0,669
Giá trị chính xác trung bình (mAP)	0,797
Các mô hình phân loại người đi bộ	
Mô hình	Độ chính xác (accuracy)
VGG16	0,943
ResNet50	0,955
InceptionV3	0,922
EfficientNetB0	0,728
Thời gian thực thi: phát hiện người đi bộ: 0,023 giây/ hình ảnh phân loại người đi bộ: (0,008 giây/ người đi bộ) x (số người đi bộ/ hình ảnh)	

Bảng 2. Kết quả thực nghiệm bộ phát hiện và phân loại người đi bộ tích hợp

Độ chính xác (precision)	0,880
Chỉ số gợi nhớ (recall)	0,760
Giá trị chính xác trung bình (mAP)	0,851
Thời gian thực thi: Phát hiện và phân loại người đi bộ: 0,024 giây/ hình ảnh	

Kết quả của các thí nghiệm theo hướng tiếp cận thứ nhất được thể hiện trong Bảng 1. Có thể thấy hiệu suất phát hiện người đi bộ của mô hình ở mức khá tốt với độ chính xác 0,873, chỉ số gợi nhớ 0,669 và độ chính xác trung bình 0,797. Việc hiệu suất chỉ dừng ở mức khá tốt có thể giải thích do trong bộ dữ liệu của bài toán bao gồm cả người lớn và trẻ em trong khi các bài toán phát hiện người đi bộ trong đa số các nghiên cứu trước đây được thực hiện trên tập dữ liệu gồm hình ảnh người lớn. Nói cách khác bộ dữ liệu này có độ phức tạp cao hơn và việc sử dụng một bộ đặc trưng đại diện cho cả người lớn và trẻ em đang cho thấy sự hiệu quả chưa thật sự tốt. Bên cạnh đó, hiệu suất của việc phân loại hình ảnh người đi bộ khá tốt với mô hình VGG16, ResNet50 và InceptionV3 (độ chính xác lần lượt là 0,943, 0,955 và 0,922). Mô hình EfficientNetB0 đang tỏ ra không thật sự phù hợp với bài toán khi độ chính xác ở mức thấp (0,760) và tình trạng overfitting diễn ra nhanh chỉ sau khoảng 20 vòng huấn luyện (độ chính xác trên tập huấn luyện hơn 0,9 trong khi độ chính xác trên tập đánh giá chỉ hơn 0,7). Ngoài ra, thời gian thực thi cho cách tiếp cận thứ nhất cũng khá cao với mức thời gian xử lý khoảng 0,1 giây cho một khung hình với 10 người đi bộ.



Hình 9. So sánh hiệu năng của hướng tiếp cận tách biệt bộ phát hiện và và bộ phân loại (hướng tiếp cận 1) và hướng tiếp cận tích hợp bộ phát hiện và phân loại người đi bộ (hướng tiếp cận 2)

Quan sát Bảng 2 và Hình 9, hướng tiếp cận sử dụng bộ phát hiện và phân loại người đi bộ tích hợp đem lại hiệu quả vượt trội so với hướng tiếp cận thứ nhất. Các giá trị bao gồm độ chính xác, chỉ số gợi nhớ và giá trị chính xác trung bình của mô hình đều cao hơn so với mô hình phát hiện người đi bộ với các giá trị lần lượt 0,880, 0,760 và 0,851.

Có thể thấy việc tổng quát hóa hình ảnh trẻ em và người lớn vào cùng một lớp hình ảnh người đi bộ để phát hiện không hiệu quả bằng việc phân biệt hóa hình ảnh trẻ em và hình ảnh người lớn vào hai lớp khác biệt. Điều này có thể lý giải bởi sự khác nhau giữa các đặc trưng của hình ảnh trẻ em và hình ảnh người lớn. Mặc dù hình ảnh người lớn và trẻ em đều có những đặc điểm chung của hình ảnh con người nhưng vẫn tồn tại những sự khác biệt trong tỷ lệ giữa các phần cơ thể. Việc cố ép hai lớp hình ảnh người lớn và trẻ em vào một lớp hình ảnh con người nói chung đã tạo ra sự mất mát các đặc trưng mô tả riêng cho từng lớp. Những đặc trưng mất đi này có thể là những đặc trưng tốt cho việc phát hiện hình ảnh người lớn hoặc hình ảnh trẻ nhỏ nói riêng điều đó dẫn đến việc sử dụng bộ phát hiện người đi bộ nói chung có hiệu suất thấp hơn bộ phát hiện người lớn và trẻ em. Nói cách khác bài toán phát hiện người đi bộ không đơn giản hơn bài toán phát hiện người lớn đi bộ và trẻ em đi bộ. Ngoài ra, nếu xét về thời gian thực thi, hướng tiếp cận thứ hai cũng đem lại hiệu quả vượt trội khi nhanh gấp hơn bốn lần khi cùng xem xét một hình ảnh có chứa 10 người đi bộ so với hướng tiếp cận thứ nhất. Sự khác biệt này xuất phát từ việc hướng tiếp cận thứ nhất sử dụng hai mô hình (mô hình phát hiện và mô hình phân loại) và thực hiện công việc qua hai giai đoạn trong khi hướng

tiếp cận thứ hai chỉ sử dụng một mô hình duy nhất và thực hiện công việc trong một giai đoạn. Kết quả thực nghiệm đã chứng minh rằng việc sử dụng mô hình phức tạp không phải lúc nào cũng đưa ra được kết quả chính xác hơn. Bên cạnh đó, mô hình phức tạp cùng số bước thực hiện lớn cũng ảnh hưởng tới thời gian huấn luyện cũng như tốc độ thực thi. Từ đó có thể kết luận hướng tiếp cận sử dụng bộ phát hiện và phân loại người đi bộ tích hợp hiệu quả vượt trội so với hướng tiếp cận sử dụng bộ phát hiện và phân loại riêng biệt.

4. Kết luận

Với mục tiêu xây dựng một hệ thống phát hiện và phân loại người đi bộ trong hình ảnh, nghiên cứu đã đề xuất hai hướng tiếp cận cho bài toán: Hướng sử dụng bộ phát hiện, bộ phân loại riêng biệt và hướng sử dụng bộ phát hiện và phân loại tích hợp. Các hướng tiếp cận được nghiên cứu và đánh giá chặt chẽ trên bộ dữ liệu lớn được kết hợp từ bộ dữ liệu Cityscapes và bộ dữ liệu xây dựng bởi nhóm tác giả. Kết quả thực nghiệm cho thấy hướng sử dụng bộ phát hiện và phân loại người đi bộ tích hợp có hiệu quả vượt trội với độ chính xác trung bình 0,851 và thời gian thực thi 0,024 giây/ hình ảnh. Điều đó thể hiện việc tổng quát hóa các người đi bộ (bao gồm cả người lớn và trẻ em) không hiệu quả bằng việc phân biệt hóa người lớn và trẻ em trong bài toán phát hiện người đi bộ. Ngoài ra một bài toán chưa được đi sâu là bài toán phân loại hình ảnh người trưởng thành và trẻ em cũng được giải quyết trong nghiên cứu này. Các mô hình học sâu được xem xét đã cho các kết quả phân loại với độ chính xác ấn tượng. Mô hình InceptionV3, VGG16 và ResNet50 lần lượt có độ chính xác: 0,922, 0,943 và 0,955.

Trong tương lai các hệ thống cảnh báo rủi ro khi lái xe và các hệ thống hỗ trợ lái xe tự động sử dụng camera hành trình là mục tiêu mà nhóm tác giả đang hướng tới để mở rộng nghiên cứu.

TÀI LIỆU THAM KHẢO

[1] Viola, Paul, and Michael Jones. *Rapid object detection using a boosted cascade of simple features*. CVPR 2001. Vol.1, 2001.
doi: 10.1109/CVPR.2001.990517.

[2] Dalal, N., Triggs, B., *Histograms of oriented gradients for human detection*. CVPR (2005),
doi: 10.1109/CVPR.2005.177.

[3] Cho, Hyunggi, et al., *Real-time pedestrian detection with deformable part models*. IEEE Intelligent Vehicles Symposium, 2012.
doi: 10.1109/IVS.2012.6232264.

[4] Girshick, Ross, et al. *Rich feature hierarchies for accurate object detection and semantic segmentation*. CVPR, pp.580-587, 2014.
doi: 10.1109/CVPR.2014.81

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *ImageNet classification with deep convolutional neural networks*, Commun. ACM, Vol.60, No.6, pp.84-90, May 2017.
doi: 10.1145/3065386.

[6] Girshick, Ross. *Fast r-cnn*. Proceedings of the IEEE international conference on computer vision. 2015.
doi: 10.1109/ICCV.2015.169.

[7] Ren, Shaoqing, et al. *Faster r-cnn: Towards real-time object detection with region proposal networks*. Advances in neural information processing systems 28. pp.91-99, 2015.
doi: 10.1109/TPAMI.2016.2577031.

[8] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition,* arXiv:1409.1556 [cs], Apr. 2015, Accessed: Apr. 22, 2021.[Online]
Available: http://arxiv.org/abs/1409.1556.

[9]Redmon, Joseph, et al. *You only look once: Unified, real-time object detection*. CVPR. 2016. doi: 10.1109/CVPR.2016.91.

[10] Redmon, Joseph, and Ali Farhadi. *YOLO9000: better, faster, stronger.* CVPR. 2017.
doi: 10.1109/CVPR.2017.690

[11] Redmon, Joseph, and Ali Farhadi. *Yolov3: An incremental improvement. a*rXiv preprint arXiv:1804.02767 (2018).
Available at: http://arxiv.org/abs/1804.02767.

[12] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. *Yolov4: Optimal speed and accuracy of object detection.* arXiv preprint arXiv:2004.10934 (2020).
Available at: http://arxiv.org/abs/2004.10934.

[13] Simonyan, Karen, and Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556 (2014).
Available at: http://arxiv.org/abs/1409.1556.

[14] Szegedy, Christian, et al. *Rethinking the inception architecture for computer vision*. CVPR. 2016.
doi: 10.1109/CVPR.2016.308.

TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

[15] He, Kaiming, et al. *Deep residual learning for image recognition.* CVPR. 2016.
doi: 10.1109/CVPR.2016.90.

[16] M. Tan and Q. V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, arXiv:1905.11946 [cs, stat], Sep. 2020, Accessed: Apr. 22, 2021. [Online].
Available at: http://arxiv.org/abs/1905.11946.

[17] Ince, Omer F., et al. *Child and adult classification using ratio of head and body heights in images*. International Journal of Computer and Communication Engineering *3.2* (2014).
doi: 10.7763/IJCCE.2014.V3.304.

[18] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B. *The cityscapes dataset for semantic urban scene understanding*. Proceedings of the IEEE conference on computer vision and pattern recognition, pp.3213-3223, 2016.

Ngày nhận bài: 11/03/2022
Ngày nhận bản sửa: 23/03/2022
Ngày duyệt đăng: 29/03/2022