PHÂN LOẠI PHƯƠNG TIỆN GIAOTHÔNG VIỆT NAM TRONG KHÔNG ẢNH

Trịnh Thị Thanh Trúc, Võ Duy Nguyên, Nguyễn Tấn Trần Minh Khang
Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

***Abstract*—**Với thực trạng giao thông đô thị Việt Nam đang gặp rất nhiều vấn đề bức thiết như: mật độ tham gia giao thông dày đặc, cơ sở hạ tầng chưa đáp ứng được lưu lượng phương tiện, thì việc đưa ra các phương án bao quát mang tính ổn định về lâu về dài luôn nhận được rất nhiều sự quan tâm từ cộng đồng. Trong đó, các ứng dụng về công nghệ thông tin luôn được xem là một trong những giải pháp được ưu tiên hàng đầu vì những lợi thế về chi phí, thời gian cũng như độ chính xác mà nó đạt được. Nhận thấy tính thời sự của vấn đề trên, nhóm chúng tôi tiến hành nghiên cứu và khảo sát các phương pháp máy học để phân lớp phương tiện giao thông trong không ảnh, mặt khác, chúng tôi cũng xây dựng một bộ dữ liệu UIT-CVID21 (Classifying Vehicle In Image From Drone) gồm 10K ảnh, cho 4 lớp đối tượng bus, car, truck, van phản ánh tình hình giao thông Việt Nam nhằm tạo ra những tiền đề trên cơ sở thực nghiệm cho các nghiên cứu về sau trong ứng dụng quản lý lưu lượng phương tiện giao thông, phân luồng giao thông, khắc phục ùn tắc.

***Keywords*—** Drone, kNN, Logistic, SVM, Vehicle classification, Vietnamese traffic, aerial images.

I. GIỚI THIỆU

Kinh tế ngày càng phát triển nhu cầu di chuyển và vận tải ngày càng tăng. Số lượng phương tiện giao thông vì thế cũng tăng theo. Lượng xe lớn tạo nhiều áp lực cho các cơ quan quản lý, giám sát giao thông. Các vấn đề như tắc đường, tai nạn giao thông, thống kê phương tiện, quy hoạch các tuyến đường và bãi đỗ xe cũng đang hiển nhiên rất cần những giải pháp kịp thời và mang tính ổn định cao. Một lượng lớn các thiết bị camera, cảm biến, radar được lắp đặt để giám sát xe cộ và thu thập thông tin giao thông, giúp các cơ quan nắm tình hình lưu lượng giao thông, mật độ phương tiện và tình trạng đỗ xe. Tuy nhiên, những phương pháp này không cung cấp đủ cái nhìn tổng quan về tình hình giao thông, mà đây lại là thông tin quan trọng để đưa ra các giải pháp phát triển và giải quyết tình trạng hiện tại.

Gần đây, hình ảnh chụp từ UAVs được ưu tiên hơn do khả năng bao quát cả khu vực và độ phân giải không gian cao hơn từ 0.1 đến 0.5m [1] và dễ thu thập hơn. Với độ phân giải cao, các phương tiện giao thông dễ dàng được phát hiện, kể cả đối tượng nhỏ như ôtô, xe máy. Việc phát hiện phương tiện giao thông với ảnh chụp từ UAVs tồn tại nhiều thách thức do sự xuất hiện của nhiều đối tượng khác như dây điện, cây xanh, tòa nhà, máy điều hòa không khí, bảng hiệu, thùng rác có thể gây nhiễu và đưa ra các cảnh báo sai. Đặc biệt là điều kiện ánh sáng, khi mà việc xuất hiện bóng của các tòa nhà, phương tiện giao thông, cây cối gây ra rất nhiều khó khăn trong công tác nhận dạng và phát hiện đối tượng phương tiện (Hình 1).

Bài toán tồn tại nhiều thách thức do sự đa dạng về góc nhìn, hình dáng, kích thước của đối tượng. Đặc biệt, trong không ảnh các hình ảnh được thu từ trên không nên xuất hiện nhiều thách thức mới [2]–[5] như : (1) Hình ảnh có thể chụp ở nhiều độ cao khác nhau, ảnh độ phân giải cao, các đối tượng chiếm tỉ lệ đa dạng, phân bố thưa hoặc tập trung dày đặc; (2) Sự mất cân bằng giữa đối tượng trong các lớp và giữa foreground và background do độ cao khi chụp; (3) Tính di động của Drone/Flycam, cũng như camera gắn kèm, xuất hiện nhiều góc nhìn làm cho cùng một đối tượng (phía trước, sau, bên hông,.. hay hướng các đối tượng xuất hiện tùy ý) lại tạo nên đa dạng các thể hiện; (4) Điều kiện thời tiết, ánh sáng (trời nắng, nhiều mây, sương mù, ban ngày, ban đêm, ngược sáng, ...) ảnh hưởng đến khả băng hiển thị đối tượng trong ảnh. Một đối tượng có thể chụp từ phía trước, sau, bên hông, top-down, ở các thời điểm khác nhau, ánh sáng khác nhau tạo nên sự đa dạng và phức tạp cho bài toán phân lớp.

Hình 1: *Không ảnh về giao thông ở các nước trên thế giới.*

Để giải quyết những tồn tại trên, chúng tôi xây dựng bộ dữ liệu chụp từ trên không thông qua Drone cho việc phân lớp các phương tiện giao thông thu thập tại Việt Nam với quy mô 10k ảnh, bao gồm 4 lớp bus, car, truck, van. Sau đó, chúng tôi tiến hành đánh giá 3 phương pháp máy học phổ biến với các vector đặc trưng được rút trích từ 5 kiến trúc mạng và 2 đặc trưng truyền thống (cụ thể là HOG và LBP), cung cấp thống kê toàn diện làm cơ sở nền tảng cho các nghiên cứu tiếp theo.

Chi tiết về các đặc trưng và mô hình cùng những nghiên cứu liên quan xin được trình bày ở chương 2. Phần còn lại của bài báo được tổ chức như sau: trong chương 3 chúng

Tác giả liên hệ: Nguyễn Tấn Trần Minh Khang
Email: khangnttm@uit.edu.vn
Đến tòa soạn: 18/6/2021, chỉnh sửa: 15/8/2021, chấp nhận đăng: 31/8/2021

tôi đi sâu vào các phương pháp và quá trình thực nghiệm, chương 4 sẽ trình bày và đánh giá kết quả, cuối cùng chương 5 chúng tôi sẽ đưa ra kết luận và hướng nghiên cứu tiếp theo.

II. NGHIÊN CỨU LIÊN QUAN

A. Bộ dữ liệu về phương tiện giao thông đã công bố:

Năm 2016, Huynh và cộng sự [6] đã giới thiệu một bộ dữ liệu phương tiện giao thông Việt Nam được thu thập từ hai nguồn dữ liệu khác nhau. Một trong số đó được cắt từ video [7]. Với góc quay cố định, thời điểm ban ngày, hướng từ trên xuống và chếch về phía bên phải của phương tiện, video dài năm phút đã thu thập được tổng cộng 1600 frame trong một phạm vi tương đối nhỏ nhưng mật độ xe lưu thông lại khá dày đặt. Mặt khác, nhóm tác giả cũng thu thập thêm một nguồn dữ liệu thứ hai từ bài báo "Learning Bag of Visual Words for Motorbike Detection".

Trong "Learning Bag of Visual Words for Motorbike Detection" [8], Thai và nhóm nghiên cứu đã đề xuất một phương pháp phát hiện xe gắn máy (motor) trong khung cảnh, đồng thời, nhóm nghiên cứu cũng xây dựng riêng một bộ dữ liệu để tiến hành thực nghiệm và đánh giá phương pháp được đề xuất. Bộ dữ liệu chỉ thu thập một loại đối tượng duy nhất là xe motor bằng một camera đặt cố định một góc 15 độ; thu thập trên nhiều thời điểm khác nhau trong một ngày và trên một giao lộ.

Năm 2016, Dinh [9] giới thiệu một bộ dữ liệu phương tiện giao thông bao gồm xe gắn máy và ô tô mà nhóm tác giả đã tiến hành thu thập để giải quyết bài toán phát hiện hai đối tượng phương tiện này. Họ cũng sử dụng một camera cố định, để quay 2 video ghi lại luồng phương tiện di chuyển trên 2 con đường khác nhau. Dữ liệu được tác giả thu thập hoàn toàn vào ban ngày.

Ho và cộng sự [10] cũng như các nhóm nghiên cứu trước, sử dụng một camera cố định để ghi lại hình ảnh các phương tiện đang lưu thông trên đường phố Việt Nam. Nhưng điểm nổi bật của bộ dữ liệu IVC-U20 của tác giả là đã linh động trong việc thu thập dữ liệu khi số video được ghi cũng nhiều gấp đôi so với các nghiên cứu trước. Mặt khác, điều kiện hình ảnh được ghi lại cũng đa dạng hơn khi dữ liệu thu thập được vào cả ban ngày, ban đêm và khi trời có mưa.

B. Hướng tiếp cận:

Trong khảo sát này, chúng tôi đi theo một trong những hướng tiếp cận cơ bản nhất cho các phương pháp phân lớp đối tượng. Đó là đầu tiên, chúng tôi đưa ảnh đầu vào qua các bước xử lý cắt chọn, sau đó, đem đi rút trích đặc trưng rồi đưa vào các mô hình phân lớp, cuối cùng, đầu ra chúng tôi thu được chính là nhãn phù hợp cho loại đối tượng trong ảnh.

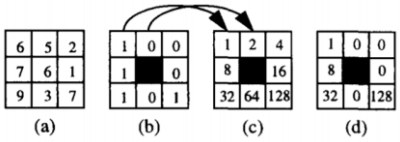
Hình 2: *Kiến trúc phân loại tổng quan.*

1) Các phương pháp rút trích đặc trưng:

HOG: Phương pháp rút trích đặc trưng hình ảnh HOG [11] lần đầu tiên được giới thiệu bởi hai tác giả Navneet Dalal và Bill Triggs tại hội nghị CVPR 2005. Đây là phương pháp sử dụng thông tin về sự phân bố của những thay đổi về mức sáng trong toàn bộ ảnh hay còn được gọi là intensity gradient của ảnh. Cụ thể, ảnh được chia thành các ô nhỏ (cells), sau đó, tác giả tính toán một đồ thị tần suất của sự biến thiên về mức sáng (histogram of oriented gradients) trong từng ô nhỏ đó. Tiếp theo, tiến hành chuẩn hóa các ô nhỏ theo một khối lớn hơn gọi là các blocks. Cuối cùng ta thu được một vector đặc trưng từ các khối đã được chuẩn hóa.

Hình 3: *Quy trình rút trích đặc trưng HOG [11].*

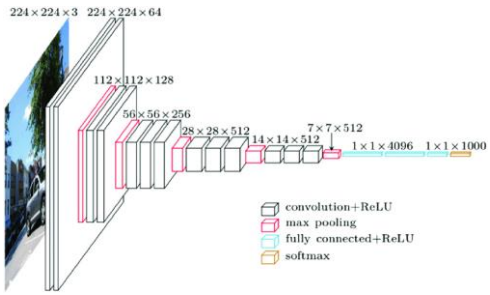
LBP: LBP [12] –Local Binary Pattern hay còn được biết đến là mẫu nhị phân địa phương được Ojala giới thiệu lần đầu tiên vào năm 1996 cho bài toán nhận dạng mặt người. Đây là một phương pháp lấy thông tin ảnh từ việc đo độ tương phản cục bộ và xét trên từng pixel của ảnh. Đầu tiên chọn ra n pixel xung quanh pixel đang xét, khởi tạo chuỗi nhị phân n bit, xét lần lược n pixel cục bộ theo một thứ tự nhất định. Cuối cùng thu được một vector đặc trưng có số lượng phần tử bằng với số lượng pixel của ảnh ban đầu. Điểm nhấn của thuật toán trích rút đặc trưng LBP là việc cài đặt khá đơn giản, thời gian tính toán giá trị đặc trưng lại nhanh vì nó làm việc với giá trị số nguyên.



Hình 4: *Hai phiên bản không nhân trọng số (b) và có nhân trọng số (d) của LBP [12].*

VGG: Dù không phải là quán quân cuộc thi ILSVRC 2014 nhưng VGGNet [13] vẫn là một trong những mạng học sâu gây được nhiều sự chú ý nhất do tính hiệu quả –về thời gian cũng như độ chính xác –đã cải thiện rất đáng kể so với các mạng ZFNet [14] và AlexNet [15]. Cụ thể, thay vì sử dụng bộ lọc có kích thước lớn như 11x11 trong AlexNet hay 7x7 của ZFNet thì VGGNet dùng bộ lọc kích thước nhỏ 3x3. Điều này dẫn đến ít tham số mà mô hình phải học hơn nên việc tính toán cũng bớt phức tạp hơn, các trọng số của mạng sẽ hội tụ nhanh hơn và mô hình cũng giảm được khả năng bị overfitting.

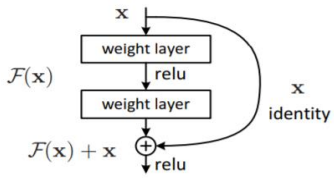
VGGNet có nhiều biến thể hay còn được gọi là những mạng thuộc họ VGG, chủ yếu, những biến thể này khác nhau về số lượng các layers trong kiến trúc mạng. Ví dụ như đối với VGG16, thì block 1 và block 2 bao gồm 2 lớp tích chập và 1 lớp MaxPool; block 3,4,5 gồm 3 lớp tích chập và 1 lớp Max Pool; cuối cùng là một lớp kết nối đầy đủ và 1 lớp SoftMax.



Hình 5: *Kiến trúc VGG16*[1].

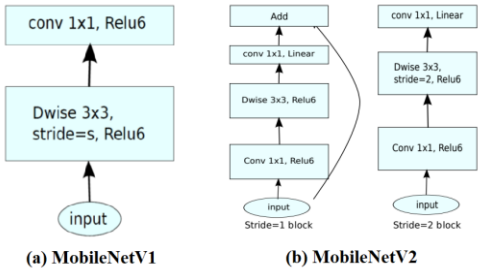
Tương tự như VGG16 cũng có 2 lớp tích chập và 1 lớp MaxPool ở hai block đầu tiên nhưng VGG19 có sự thay đổi trong kiến trúc mạng ở block thứ 3, thứ 4 và thứ 5. Cụ thể, thay vì có 3 lớp tích chập như VGG16 thì ở VGG19, tác giả thay thế thành 4 lớp tích chập và sau đó là một lớp MaxPool; cuối cùng vẫn là một lớp kết nối đầy đủ và 1 lớp SoftMax. Chính nhờ điểm cải thiện này, mà mạng VGG19 học sâu hơn nói cách khác mạng học được nhiều đặc trưng của ảnh hơn.

ResNet50: Những kiến trúc trước đây thường cải thiện độ chính xác nhờ gia tăng chiều sâu của mạng. Nhưng thực nghiệm cho thấy rằng đến một ngưỡng độ sâu nhất định thì độ chính xác của mô hình sẽ bị bão hòa, thậm chí còn xảy ra trường hợp mô hình kém chính xác hơn do hiện tượng vanishing và exploding gradient [16]. Để giải quyết vấn đề trên, nhóm nghiên cứu của Microsoft đã áp dụng kỹ thuật batch normalization hay còn được biết đến là một phương pháp chuẩn hóa dữ liệu về dạng phân phối với kỳ vọng bằng 0 và độ lệch chuẩn bằng 1; kỹ thuật thứ hai là sử dụng kết nối tắt (skip connection)–bỏ qua một vài lớp trung gian để gradient không bị triệt tiêu và có thể lan truyền được đến những layers cuối cùng. Kiến trúc Resnet được xây dựng theo các khối tích chập (Conv Block) và khối xác định (Identity Block) sử dụng bộ lọc kích thước 3x3 gọi chung là các residual blocks. Những số hậu tố của các phiên bản của mạng Resnet chỉ ra số lớp trong kiến trúc ResNet đó, đơn cử với ResNet 50, mạng sẽ có 50 lớp nằm trong các khối tích chập và khối xác định nối tiếp nhau một cách liên tục.



Hình 6: *Kiến trúc residual block với kết nối tắt của mô hình ResNet [16].*

MobileNet V2: Điểm chung của các mô hình họ MobileNet [17] là sử dụng một cách tính tích chập mới có tên là Separable Convolution để giảm kích thước mô hình và giảm độ phức tạp tính toán. Nhờ vào điểm cải thiện này mà mô hình có thể chạy được trên ứng dụng trên di động, các thiết bị nhúng hoặc đảm nhiệm cả những nhiệm vụ chạy trên thời gian thực (real–time). Năm 2017, nhóm nghiên cứu Google công bố một phiên bản mới đó là MobileNetV2 [18] với một số điểm cải tiến mang lại kết quả khá ấn tượng. Cụ thể, kiến trúc mạng cũng sử dụng các kết nối tắt như của ResNet nhưng thay vì giữ nguyên kết cấu residual block cũ –số lượng kênh ở input và output của một block lớn hơn so với các layer trung gian –thì ở phiên bản cải tiến này, MobileNetV2 được điều chỉnh ngược lại. Vì tác giả cho rằng các layer trung gian trong một block sẽ làm nhiệm vụ biến đổi phi tuyến nên cần dày hơn để tạo ra nhiều phép biến đổi hơn. Mặt khác, thực nghiệm cho thấy việc sử dụng các biến đổi phi tuyến tại input và output của các residual blocks sẽ làm cho thông tin bị mất mát gây ảnh hưởng đến kết quả cuối cùng, do đó, tác giả cũng đã thay thế hàm phi tuyến tại layer input và output bằng các phép chiếu tuyến tính. Những điểm cải tiến nêu trên đã giúp cho mô hình MobileNetV2 giảm được 30% lượng tham số đầu vào, nhanh hơn 30–40%, và độ chính xác cũng cải thiện đáng kể so với phiên bản MobileNet cũ.



Hình 7: *Kiến trúc khối tích chập của MobileNetV1 (a) và MobileNetV2 (b)[18].*

EfficientNet B0: EfficientNet [19] được giới thiệu bởi Mingxing Tan và Quoc V. Le như một cách tiếp cận mới cho việc thay đổi các thông số của mạng hay còn được biết đến với thuật ngữ model scaling –thu phóng mô hình. Với những mô hình trước đây, các tác giả thường chỉ tập trung thay đổi một trong những thông số: độ sâu, độ rộng, độ phân giải một cách riêng lẻ, rời rạc, nhưng thực nghiệm đã chứng minh, việc điều chỉnh chỉ đạt được đến một ngưỡng nhất định nào đó thì độ chính xác của mô hình sẽ bị bão hòa, thậm chí có thể làm mô hình kém hiệu quả hơn so với mô hình ban đầu. Ý tưởng của nhóm tác giả hướng đến là việc phối hợp và cân bằng các thông số mạng một cách có hệ thống để mang đến hiệu suất tốt hơn. Thực tế, EfficientNetB0 –phiên bản đơn giản nhất – đã trả về một kết quả ấn tượng với độ chính xác lên đến 77% khi xét trên bộ dữ liệu ImageNet.

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	28×28	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: *Các thông số trên mạng EfficientNetB0 [19].*

[1] https://www.researchgate.net/figure/An-overview-of-the-VGG-16-model-architecture-this-model-uses-simple-convolutional-blocks_fig2_328966158

2) Các phương pháp phân lớp:

k–Nearest Neighbours: k–Nearest Neighbor là một trong những thuật toán học có giám sát được áp dụng cho cả hai bài toán phân lớp và hồi quy. Đây cũng được xem là một thuật toán học máy đơn giản nhất vì khi huấn luyện, mô hình không thực sự học được bất cứ điều gì từ dữ liệu đưa vào mà tất cả các kết quả trả về chỉ dựa trực tiếp trên nhãn của k điểm dữ liệu lân cận mà thôi. Việc chọn siêu tham số k cũng ảnh hưởng rất nhiều để độ chính xác của mô hình đầu ra. Công thức tổng quát Minkowski dùng để tính khoảng cách giữa các điểm dữ liệu được trình bày như sau:

$$L = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \quad (1)$$

Với L là khoảng cách Manhattan khi p = 1 và là Euclidean khi p = 2.

Support Vector Machine (SVM): Đầu tiên, trong thuật toán SVM có một khái niệm là margin –thuật ngữ chỉ khoảng cách gần nhất từ một điểm dữ liệu tới mặt phân cách giữa các lớp. Bài toán SVM là bài toán đi tìm một siêu phẳng tối ưu mà tại đó margin của các lớp dữ liệu là lớn nhất và bằng nhau. SVM có thể được áp dụng cho cả hai bài toán phân lớp và hồi quy. Margin được tính theo công thức sau:

$$margin = \min_n \frac{y_n(\boldsymbol{w}^T \boldsymbol{x}_n + b)}{\|\boldsymbol{w}\|_2} \quad (2)$$

Sau đó tối ưu hóa bằng cách cập nhật lại w, b sao cho margin là lớn nhất:

$$(\boldsymbol{w}, b) = \arg\max_{\boldsymbol{w},b} \left\{ \min_n \frac{y_n(\boldsymbol{w}^T \boldsymbol{x}_n + b)}{\|\boldsymbol{w}\|_2} \right\} = \arg\max_{\boldsymbol{w},b} \left\{ \frac{1}{\|\boldsymbol{w}\|_2} \min_n y_n(\boldsymbol{w}^T \boldsymbol{x}_n + b) \right\} \quad (3)$$

Logistic regression: Hồi quy logistic là một mô hình hồi quy nhằm dự đoán giá trị đầu ra rời rạc. Trong khi đó, hồi quy tuyến tính là một mô hình hồi quy dự đoán giá trị đầu ra liên tục. Nguyên lý hoạt động của hồi quy logistic là đưa đầu ra của mô hình hồi quy tuyến tính đi qua một hàm kích hoạt có tên là sigmoid, để tất cả giá trị output của hàm giả định sẽ nằm trong khoảng [0,1] và xem giá trị này chính là xác suất biến cố input thuộc một lớp xác định nào đó trong bộ dữ liệu huấn luyện. Hàm sigmoid được định nghĩa như sau:

$$S(x) = \frac{1}{1 + e^{-x}} \quad (4)$$

III. BỘ DỮ LIỆU UIT-CVID21

1) Tổng quan về bộ dữ liệu:

Bộ Dữ liệu UIT-CVID21[2] được ghi lại bằng thiết bị không người lái (drone) với độ phân giải chủ yếu ở mức 960×720 pixel. Dữ liệu bao gồm 3,982 ảnh được chọn lọc từ các frame trong video. Các video được quay vào thời điểm ban ngày, trong điều kiện thời tiết quang tạnh và thời tiết có sương mù; mật độ phương tiện tham gia giao thông là vừa phải, phân bố trung bình 50 bbox, đã qua chọn lọc thủ công, cho một ảnh; loại phương tiện được thu thập là xe buýt, xe ô tô, xe tải, xe van tương ứng với 4 lớp bus, car, truck, van được quy định trong bộ dữ liệu.

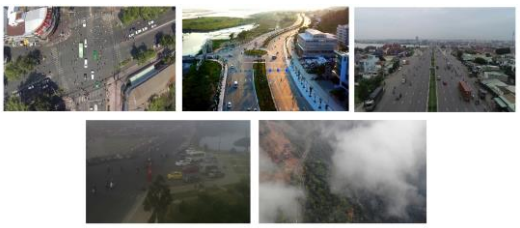
Bảng 1: *So sánh các bộ dữ liệu về phương tiện giao thông được thu thập tại Việt Nam.*

Bộ dữ liệu	Thiết bị	Số video	Điều kiện	Năm
[7]	Camera cố định	1 video	Ngày	2013
[6]	Camera cố định	1 video	Ngày/ Đêm	2014
[9]	Camera cố định	2 video	Ngày	2016
[10]	Camera cố định	5 video	Ngày / Đêm/ Mưa	2020
UIT-CVID21	Drone	18 video	Sáng/ Sương	2021

UIT-CVID21 là một bộ dữ khá hiếm hoi trong thời điểm hiện tại mà có thể thu được các góc quay đa dạng nhất, mà lại mang rõ đặc thù đường phố giao thông Việt Nam nhờ vào sự linh động của thiết bị không người lái, cụ thể là drone, như chúng tôi đã đề cập ở phần trước. Những bộ dữ liệu trước đây hầu như được quay bằng camera tĩnh với góc quay, độ cao cố định, và điều này mang lại những thách thực hoàn toàn khác biệt so với một bộ dữ liệu được quay bằng drone như của chúng tôi.

Những khác biệt này tạo ra nhiều thách thức như sau:

- Thiết bị quay: Việc quay bằng drone so với camera tĩnh sẽ thu thập được hình ảnh với góc quay và độ cao khác nhau. Do đó khi xét một đối tượng, nó có những thể hiện về góc nhìn, tỷ lệ, kích thước hoàn toàn khác nhau.
- Số video sử dụng: UIT-CVID21 có phong phú và đa dạng các cảnh quan khi số cảnh quay được thu thập lên đến 18 video riêng biệt. Điều này tạo cho bộ dữ liệu chúng tôi một cái nhìn bao quát về đường phố Việt Nam từ đồng bằng đến đồi núi, từ nông thôn đến các trung tâm thành phố lớn –những vị trí mà sự khác biệt về mật độ giao thông là khá lớn.
- Ánh sáng/Điều kiện thời tiết: Điểm mới của bộ dữ liệu chúng tôi là thu thập những thước dữ liệu bị ảnh hưởng bởi sương mù –một thách thức rất quen thuộc trong các bài toán xử lý hình ảnh phương tiện giao thông.
- Năm: Với UIT-CVID21 chúng tôi tự tin nhận định đây là bộ dữ liệu về phương tiện giao thông Việt Nam từ Drone mới nhất trong thời điểm hiện tại, nhờ đó mà bộ dữ liệu của chúng tôi có thể nắm được phần nào các điểm đổi mới trong xu hướng thay đổi về luồng xe của năm 2021.



Hình 9: *Ảnh được chụp từ drone chưa qua bước tiền xử lý.*

[2] UIT-CVID21 published at https://uit-together.github.io/datasets/

2) Xử lý dữ liệu:

a) Gán nhãn và Cắt chọn đối tượng trong ảnh:

Gán nhãn: Trong lĩnh vực thị giác máy tính có khá nhiều cách để chú thích cho một đối tượng trong ảnh, ví dụ với COCO, bộ dữ liệu quy ước hộp giới giới hạn bao gồm (x–top left, y–top left, w, h) trong đó (x–top left, y–top left) là tọa độ điểm phía trên–bên trái và (w, h) là chiều rộng và chiều cao tương ứng của hộp giới hạn. Khác với điểm "top–left" của COCO, ta có định dạng YOLO (x–center, y–center, w, h) sử dụng điểm trung tâm với (x–center, y–center) là tọa độ tâm điểm của hộp giới hạn. Ngoài ra còn có bộ dữ liệu Visdron (x–center, y–center, w, h, θ) không chỉ sử dụng thuộc tính điểm trung tâm (x–center, y–center) như YOLO hay (w, h) tương ứng với chiều rộng và chiều cao của hộp giới hạn như COCO, mà bộ dữ liệu này còn sử dụng thêm thuộc tính θ để diễn tả góc nghiêng của hộp giới hạn so với phương ngang.

Riêng đối với bộ dữ liệu UIT-CVID21, chúng tôi đơn giản sử dụng phương pháp chú thích tương tự như định dạng của Pascal VOC. Từ 3,982 ảnh gốc, chúng tôi dùng công cụ chú thích các hộp giới hạn hình chữ nhật một cách thủ công, cho từng đối tượng trên ảnh. Những hộp giới hạn này được quy định theo định dạng (x left, y top, x right, y bottom), trong đó (x left, y top) là tọa độ điểm phía trên–bên trái và (x right, y bottom) là tọa độ điểm phía dưới–bên phải.

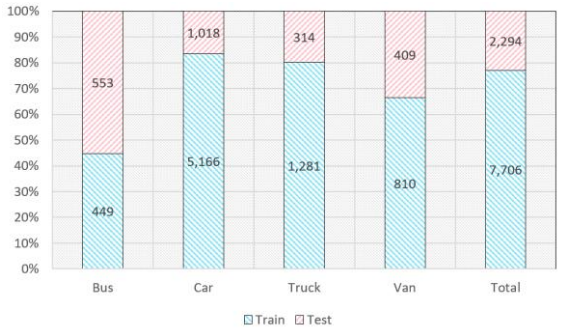
Cắt chọn đối tượng trong ảnh: Sau giai đoạn gán nhãn, chúng tôi tiến hành cắt chọn đối tượng thành từng ảnh đơn lẻ, rồi phân chúng vào 4 lớp (bus, car, truck, van). Sở dĩ chọn 4 lớp xe buýt, xe ô tô, xe tải và xe van là vì trong quá trình quan sát, chúng tôi nhận thấy rằng những hình ảnh của đối tượng thu được từ thiết bị không người lái từ độ cao trên dưới 100m, là khá nhỏ, dễ gây nhầm lẫn. Cụ thể, sự nhầm lẫn về hình thái dễ xảy ra đối với cặp đối tượng xe buýt–xe tải và xe ô tô–xe van.

Hình 10: *Ảnh thu được sau bước cắt chọn đối tượng trong ảnh.*

Mặt khác, bởi vì sử dụng hộp giới hạn theo dạng hình chữ nhật, cho nên, sau giai đoạn cắt chọn đối tượng trong ảnh theo tọa độ và chú thích ở bước gán nhãn, bộ dữ liệu thu được sẽ xuất hiện những đối tượng bị cắt xém, không đầy đủ bộ phận, bị che khuất, chồng các đối tượng vào nhau. Việc cắt ảnh kèm với những yếu tố về góc quay, ánh sáng, bối cảnh chính là những thách thức trong bộ dữ liệu mà chúng tôi muốn đề cập.

b) Thống kê:

Bộ dữ liệu của chúng tôi bao gồm 4 lớp: bus, car, truck, van; được quay bằng thiết bị không người lái (drone); trong điều kiện thời tiết quang tạnh và thời tiết có sương mù; và toàn bộ ảnh được thu thập là vào thời điểm ban ngày.

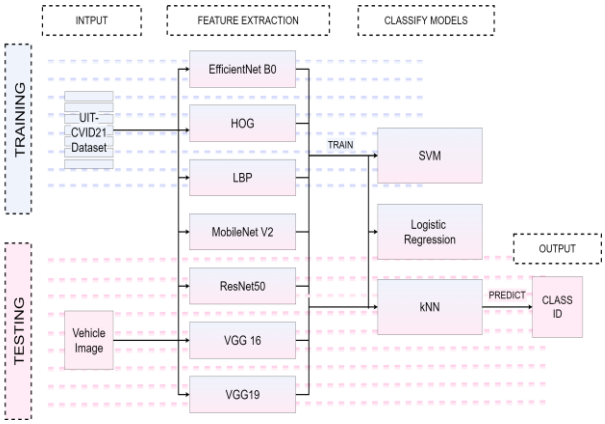


Hình 11: *Thống kê phân phối ảnh giữa tập train và tập test.*

Bộ dữ liệu UIT-CVID21 bao gồm tổng cộng 10,000 ảnh, trong đó, có 7,706 ảnh được dùng để huấn luyện và 2,294 ảnh dùng để kiểm tra kết quả mô hình, được phân chia một cách ngẫu nhiên. Kích thước ảnh trung bình của bộ dữ liệu là 1751 pixel, trong đó, ảnh nhỏ nhất có kích thước 13×17 pixel và ảnh có kích thước lớn nhất là 409×473 pixel.

IV. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Sau khi thu thập và xử lý dữ liệu, chúng tôi chia tập dữ liệu thành hai phần train và test với tỉ lệ 7:3 và cách chia là ngẫu nhiên cho từng lớp; trong đó, tập train chứa hình ảnh dùng cho quá trình huấn luyện, còn tập test thì được sử dụng để đánh giá kết quả mô hình. Tiếp theo, chúng tôi tiến hành rút trích lần lượt bảy đặc trưng EfficientNet B0, HOG, LBP, MobileNet V2, ResNet50, VGG16, VGG19, sau đó, đưa những đặc trưng này đi qua các mô hình phân lớp và huấn luyện độc lập với nhau. Ở đây chúng tôi dùng ba phương pháp phân lớp là SVM, hồi quy Logistic và k–Nearest Neighbors. Với phương pháp k–Nearest Neighbors, chúng tôi khảo sát trên giá trị k = 5 được đặt mặc định trong thư viện Scikit–learn, đồng thời, độ đo Minkowski cũng được sử dùng trong quá trình thực nghiệm để xét khoảng cách giữa các điểm dữ liệu trên tập train và điểm dữ liệu cần xét trong tập test.



Hình 12: *Lưu đồ nghiên cứu thực nghiệm.*

Sau giai đoạn huấn luyện, chúng tôi tiến hành đánh giá kết quả thực nghiệm từ 2,294 ảnh trong tập test. Kết quả chi tiết xin được trình bày chi tiết bên dưới.

A. Độ đo:

Accuracy: Độ chính xác của mô hình, được tính bằng tỉ

lệ giữa số mẫu được dự đoán đúng trên tổng số mẫu dữ liệu được đưa vào đánh giá.

$$Accuracy = \frac{\sum_{i=1}^{C} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{C}$$

Precision: Tỉ lệ số mẫu được dự đoán đúng là Positive trên tổng số mẫu được dự đoán là Positive.

$$Precision = \frac{\sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i}}{C}$$

Recall: Tỉ lệ số mẫu được dự đoán đúng là Positive trên tổng số mẫu thật sự là Positive.

$$Recall = \frac{\sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i}}{C}$$

$F_1 - score$: Là độ đo biểu diễn giá trị trung bình điều hòa giữa precision và recall.

$$F_1 - score = 2\frac{Precision \times Recall}{Precision + Recall}$$

Trong đó:

- TP: True Positive
- FP: False Positive
- TN: True Negative
- FN: False Negative
- C: Số lớp dữ liệu đầu vào

B. Phân tích kết quả thực nghiệm:

Sau khi đưa dữ liệu đã xử lý vào thực nghiệm theo các bước dựa trên lưu đồ như Hình 12, chúng tôi tiến hành đánh giá tất cả các kết quả theo độ đo đã được trình bày ở phần IV.A.

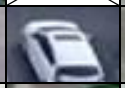
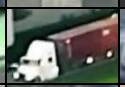
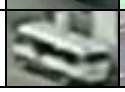
Bảng 2: Kết quả thực nghiệm trên bộ dữ liệu UIT-CVID21.

Phân lớp \ Đặc trưng		LBP	HOG	EfficientNetB0	MobileNetV2	ResNet50	VGG16	VGG19
Logistic	**Precision**	39.00	40.98	59.00	47.21	45.62	50.69	36.51
	Recall	43.42	46.21	60.55	50.87	50.57	52.01	48.74
	F1–score	35.55	38.33	58.14	42.7	40.39	42.61	37.18
	Accuracy	**43.42**	46.21	**60.55**	50.87	50.57	52.01	48.74
SVM	**Precision**	25.58	60.28	64.65	51.83	63.35	46.99	51.67
	Recall	44.64	56.63	55.36	49.96	52.48	51.13	50.52
	F1–score	28.04	53.31	50.05	43.05	47.68	41.88	41.69
	Accuracy	**44.64**	**56.63**	55.36	49.96	52.48	51.13	50.52
k–NN	**Precision**	27.23	53.05	29.09	46.2	55.44	46.87	41.63
	Recall	39.19	54.71	44.16	48.82	49.26	42.28	40.98
	F1–score	31.51	53.36	33.66	39.52	46.14	34.63	33.31
	Accuracy	**39.19**	**54.71**	44.16	48.82	49.26	42.28	40.98

Cụ thể, với bộ phân loại SVM, việc ảnh được đưa vào rút trích đặc trưng HOG đã mang lại độ chính xác cao nhất tại 56,63% so với các phương pháp rút trích khác trên cùng bộ phân lớp. Mặt khác, phương pháp hồi quy Logistic đã thể hiện tính hiệu quả đáng kể của nó khi kết quả phân lớp sau khi dữ liệu đầu vào được đưa qua mạng học sâu EfficientNetB0 để rút trích đặc trưng. Với độ chính xác Accuracy lên đến 60,55% và F1–score là 58.14%, hồi quy Logistic-EfficientNetB0 trở thành phương pháp mang lại kết quả cao nhất trong cả quá trình thực nghiệm của chúng tôi. Trái lại, thuật toán phân lớp k–NN lại thể hiện kém hiệu quả nhất khi áp dụng cùng với phương pháp rút trích đặc trưng LBP, cụ thể chỉ có 39,18% dự đoán là trùng khớp với tập test. Dù độ đo F1–score tương ứng của phương pháp k–NN-LBP chỉ đạt 31.51% nhưng lại cao hơn khá nhiều khi so sánh với phương pháp SVM có cùng kỹ thuật tiền xử lý LBP là 28.04% (Bảng 2).

Qua kết quả thu được, chúng tôi nhận thấy rằng cần có sự suy xét kỹ lưỡng trong quá trình lựa chọn đặc trưng cho bước tiền xử lý ảnh trước khi đưa vào mô hình phân lớp. Đã có nhiều minh chứng cho việc mạng học sâu thể hiện rất tốt trong các nhiệm vụ trích xuất thông tin và với nghiên cứu của chúng tôi thì nhận định này cũng không là ngoại lệ. Khi mà, các kết quả thu được từ mạng học sâu hầu hết mang lại độ chính xác cao hơn từ 2–14% so với các phương pháp khác. Mặt khác, ta cũng không nên bỏ qua các phương pháp rút trích truyền thống như HOG, khi đặc trưng này xử lý khá tốt các thách thức về chất lượng ảnh mà chúng tôi đã từng đề cập qua trong phần xử lý dữ liệu.

Bảng 3: Minh họa việc phân lớp ảnh vào các class.

Nhãn / Dự đoán	Bus	Car	Truck	Van
Bus				
Car				
Truck				
Van				

Khi trực quan hóa kết quả (Bảng 3), chúng tôi nhận thấy việc nhầm lẫn trong phân lớp xảy ra ở hầu hết các lớp. Nhưng do dữ liệu về lớp car khá đa dạng và đầy đủ cho nên việc dự đoán lớp car cũng chính xác hơn tất cả những lớp còn lại. Cụ thể, theo tập test 2292 ảnh của chúng tôi, thì không có bất kì trường hợp car nào bị nhầm thành bus. Mặt khác, với những trường hợp xe ô tô được quay theo góc máy chéo, xếch về bên hông sẽ dễ khiến cho đối tượng vô tình bị kéo dài ra hơn so với tỉ lệ thực dẫn đến sự nhầm lẫn về hình thái với các đối tượng thuộc lớp van. Ngoài ra, các ô tô được thu thập dưới góc máy top–view cũng làm cho

đối tượng thể hiện một dạng hình chữ nhật khá giống với các thùng hàng của container/truck gây ra sự nhầm lẫn giữa những lớp này. Còn đối với góc quay đối diện, thì các đối tượng thường thể hiện một hình dáng tương đối khá giống nhau với một đầu xe hình vuông, cửa sổ và phần đèn xe bên dưới dễ làm cho mô hình phân loại bị nhầm lẫn với ô tô và đưa tất cả về lớp car. Tóm lại, những đối tượng vô tình bị kéo dài ra do góc quay như đã đề cập ở trên thì sự nhầm lẫn thường quy về hai trường hợp chính: một là van –khi các phương tiện này có màu trắng xám; hai là bus – khi các đối tượng này có màu sắc đa dạng hoặc bị kéo dãn ra quá dài so với tỉ lệ thực.

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi thực hiện khảo sát các phương pháp máy học áp dụng rút trích đặc truyền thống, đặc trưng học sâu trên các lớp bus/car/truck/van thông qua bộ dữ liệu UIT-CVID21. Với mục tiêu xây dựng một bộ dữ liệu sát với bối cảnh giao thông Việt Nam đi cùng những công nghệ thu thập dữ liệu linh động, không người lái, UIT-CVID21 đã trở nên khá thách thức khi trả về kết quả chỉ ở mức tương đối trong lần khảo sát này. Ở những nghiên cứu kế tiếp, chúng tôi hy vọng sẽ cải thiện kết quả, mặt khác, tiến hành tiếp cận các lớp phương tiện thách thức hơn như xe đạp và xe máy nhằm mang lại những đóng góp trên cơ sở thực nghiệm, góp phần giải quyết các thực trạng giao thông đô thị tại Việt Nam.

LỜI CẢM ƠN

Nghiên cứu được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (ĐHQG-HCM) trong khuôn khổ Đề tài mã số DS2021-26-01. Chúng tôi xin chân thành cảm ơn phòng thí nghiệm Truyền thông Đa phương tiện (MMLab), trường Đại học Công nghệ Thông Tin, ĐHQG-HCM đã hỗ trợ chúng tôi trong quá trình nghiên cứu và thực nghiệm.

TÀI LIỆU THAM KHẢO

[1] A. Kembhavi, D. Harwood, and L. S. Davis, "Vehicle detection using partial least squares," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 6, pp. 1250–1265, 2011.

[2] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3974–3983.

[3] H. Fan, D. Du, L. Wen, P. Zhu, Q. Hu, H. Ling, M. Shah, J. Pan, A. Schumann, B. Dong, et al., "Visdrone-mot2020: The vision meets drone multiple object tracking challenge results," in European Conference on Computer Vision, Springer, 2020, pp. 713–727.

[4] QM Chung, TD Le, TV Dang, ND Vo, TV Nguyen, K Nguyen, "Data augmentation analysis in vehicle detection from aerial videos". In: 2020 RIVF International Conference on Computing and Communication Technologies (RIVF). IEEE, 2020. p. 1-3.

[5] K Nguyen, NT Huynh, PC Nguyen, KD Nguyen, ND Vo, TV Nguyen, "Detecting objects from space: An evaluation of deep-learning modern approaches", Electronics 2020, 9 (4), 583.

[6] C.-K. Huynh, T.-S. Le, and K. Hamamoto, "Convolutional neural network for motorbike detection in dense traffic," in 2016 IEEE Sixth International Conference on Communications and Electronics (ICCE), IEEE, 2016, pp. 369–374.

[7] Rob Whitworth. Ho Chi Minh City (Saigon), "Vietnam Rush Hour Traffic in Real Time,". 2013. URL: https://www.youtube.com/watch?v=Op1hdgzmhXM. (accessed: 22.08.2020).

[8] Ngoc Dung Thai et al. "Learning bag of visual words for motorbike detection,". In: 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV). IEEE. 2014, pp. 1045–1050.

[9] V.-T. Dinh, N.-D. Luu, and H.-H. Trinh, "Vehicle classification and detection based coarse data for warning traffic jam in vietnam," in 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS), IEEE, 2016, pp. 223–228.

[10] N. Ho, M. Pham, N. D. Vo, and K. Nguyen, "Phát hiện phương tiên giao thông trong các trung tâm thành phố lớn với phương pháp yolov4," in 2020 The 23nd National Symposium of Selected ICT Problems (@), 2020, pp. 344–349.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Ieee, vol. 1, 2005, pp. 886– 893.

[12] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," Pattern recognition, vol. 29, no. 1, pp. 51–59, 1996.

[13] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 10, pp. 1943–1955, 2015.

[14] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision, Springer, 2014, pp. 818–833.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, pp. 1097–1105, 2012.

[16] 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.

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient onvolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[19] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

CLASSIFYING VEHICLE IN IMAGE FROM DRONE

Abstract: With the current situation of traffic in Vietnam, we are facing many outstanding problems such as high traffic density or the infrastructures could not being adapt to the dramatic increase in the number of vehicles. Therefore, proposing comprehensive plans to bring

stability in the long term has always received a lot of attention from the community. Meanwhile, information technology applications are always considered to be one of the most priority solutions because of their advantages of cost, efficient and accuracy. In this project, our team conducted research and surveyed machine learning methods to classify vehicle in aerial images. On the other hand, we also built a dataset named UIT-CVID21 (Classifying Vehicle In Image From Drone) consisting of 10K images for four classes of bus, car, truck, van, which can reflect the reality of Vietnam traffic. Our project aims to create premises for later studies and address problems such as traffic density management, traffic separation and traffic congestion.

Keywords: Drone, kNN, Logistic, SVM, Vehicle classification, Vietnamese traffic, aerial images.

Trịnh Thị Thanh Trúc hiện đang là sinh viên năm 2, khoa Hệ Thống Thông Tin, Trường Đại học Công nghệ Thông tin, ĐHQG-HCM.

Võ Duy Nguyên nhận bằng Cử nhân Công nghệ Thông tin năm 2013, Thạc sĩ Khoa học máy tính năm 2018 tại Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM. Hiện đang công tác tại Phòng thí nghiệm Truyền thông Đa phương tiện, Trường ĐH Công nghệ Thông Tin, ĐHQG-HCM. Nghiên cứu trong lĩnh vực Thị giác máy tính, mạng học sâu.

Nguyễn Tấn Trần Minh Khang tốt nghiệp Đại học Tổng hợp năm 1996. Anh nhận bằng Thạc sĩ năm 2002, nhận bằng Tiến sĩ Công nghệ Thông tin năm 2014 tại Trường ĐH Khoa học Tự nhiên TPHCM. Nghiên cứu trong lĩnh vực Thị giác máy tính, mạng học sâu, khai thác dữ liệu, các giải thuật metaheuristic, bài toán thời khóa biểu.