http://doi.org/10.37550/tdmu.VJS/2023.04.451

KẾT HỢP XỬ LÝ ẢNH THUẦN VÀ HỌC SÂU TRONG VẤN ĐỀ NHẬN DIỆN KẸT XE

Nguyễn Văn Bình[(1)]

(1) Trường Đại học Thủ Dầu Một

Ngày nhận bài 26/6/2023; Ngày gửi phản biện 27/06/2023; Chấp nhận đăng 15/7/2023

Liên hệ email: binhnv@tdmu.edu.vn

https://doi.org/10.37550/tdmu.VJS/2023.04.451

Tóm tắt

Trong cuộc sống đô thị hiện nay kẹt xe là một trong những tình trạng nhức nhối đối với những người tham gia giao thông nói chung và những cánh tài xế nói riêng. Những năm gần đây đã có nhiều phương pháp khoa học công nghệ được đề xuất nhằm cải thiện vấn đề trên trong giờ cao điểm. Trong bài báo này giải pháp nhận diện kẹt xe được tác giả đề xuất với phương pháp sử dụng góc nhìn của nhiều camera tại các giao lộ trên địa bàn thành phố để nhận biết số lượng và lưu lượng xe di chuyển tại thời điểm và từ đó đưa ra dự đoán có kẹt xe hay không. Phương pháp đề xuất sự kết hợp giữa phương pháp học sâu và các phương pháp trích xuất đặc trưng truyền thống như SIFT (Scale-Invari ant Feature Transform- SIFT). Trong đó ScaledYOLOv4 (You Only Look Once) được sử dụng để phát hiện các phương tiện như xe máy, xe con. Sau đó phương pháp theo dõi tính tốc độ di chuyển và lưu lượng của các xe trong đám đông sử dụng feature matching như SURF, SIFT được áp dụng. Kết quả cho thấy phương pháp đề xuất cho ra một kết quả tốt trong việc nhận diện kẹt xe.

Từ khóa: *học sâu, kỹ thuật tương xứng đặc trưng, nhận diện phương tiện giao thông, ScaledYOLOv4, SIFT*

Abstract

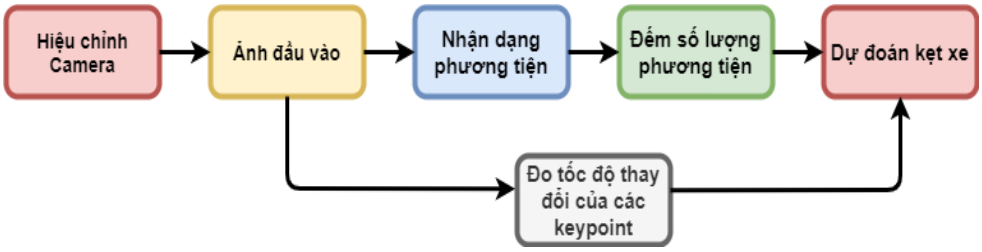
ENHANCED TRAFFIC JAM RECOGNITION BASES ON IMAGE PROCESSING AND DEEP LEARNING FUSION

In today's urban life, traffic jam is one of the painful situations for traffic participants in general and drivers in particular. In recent years, many scientific and technological methods have been proposed to improve this problem during peak hours. In this paper, the author proposes a solution to identify traffic jams with the method of using the view of many cameras at intersections in the city to recognize the number and volume of traffic moving at the time and from that predicts traffic jams or not. The method proposes a combination of deep learning and traditional feature extraction methods such as SIFT (Scale-Invariant Feature Transform- SIFT). In which ScaledYOLOv4 (You Only

Look Once) is used to detect vehicles such as motorbikes and cars. Then the tracking method calculates the moving speed of the vehicles in the crowd using feature matching such as SURF, SIFT is applied. The results show that the proposed method gives a good result in detecting traffic jams.

1. Giới thiệu

Trong xã hội ngày nay, với sự gia tăng về số lượng phương tiện tham gia giao thông ngày càng đông, các công nghệ giám sát giao thông thông minh, hệ thống giao thông thông minh (ITS: Intelligent Transportation Systems) ngày càng được sử dụng rộng rãi để trích xuất được nhiều loại thông tin quan trọng. Hệ thống giao thông thông minh- ITS là một hệ thống lớn với nhiều công nghệ cao trong nhiều phương diện điện tử, viễn thông, tin học được sử dụng nhằm mục đích giảm bớt tắc nghẽn giao thông, các tác động xấu đến môi trường, đảm bảo an toàn. Các nghiên cứu về lĩnh vực nhận diện kẹt xe hiện nay tập trung chủ yếu vào 2 hướng chính đó là nhận diện số lượng xe đang di chuyển trên dường và tốc độ của chúng bởi vì chỉ nhận diện và đếm số lượng xe là không đủ bởi vì trong các trường hợp trên các đường quốc lộ số lượng xe đông nhưng các xe máy nhỏ có thể chạy với tốc độ lên tới 40km/h thì cũng sẽ không xét vào trường hợp kẹt xe. Về công nghệ nhận dạng và đếm số lượng xe tĩnh như xe hơi, xe máy trong số lượng nhiều đã được quan tâm từ rất lâu, gần đây đã phát triển rất nhanh và cho ra kết quả khá tốt. Từ các phương pháp truyền thống như nhận diện đèn nền với độ chính xác lên tới 99% vào năm 2014 (Nidhal và cs., 2014). Ngoài ra còn các phương pháp sử dụng mạng nơ ron tích chập như Faster R-CNN (Xu và cs., 2017) được áp dụng trong việc nhận diện các phương tiện gia thông. Cũng như các phương pháp ra mới ra đời trong những năm gần đây như CenterNet (Duan và cs., 2019), YoloV4 (Bochkovskiy và cs., 2020), YoloV4-CSP (Wang và cs., 2020) đã cho thấy sự vượt trội về nhiều phương diện như độ chính xác cũng như là tốc độ thực thi.



Hình 1. *Tổng quan hệ thống*

Phần còn lại thể hiện sự đóng góp của bài báo:

- Cải tiến độ chính xác của YOLOV4-CSP
- Kết hợp giữa các phương pháp xử lý ảnh truyền thống với các phương pháp SOTA trong mảng học sâu.
- So sánh các phương pháp thị giác máy tính truyền thống trong vấn đề đo dòng lưu lượng.

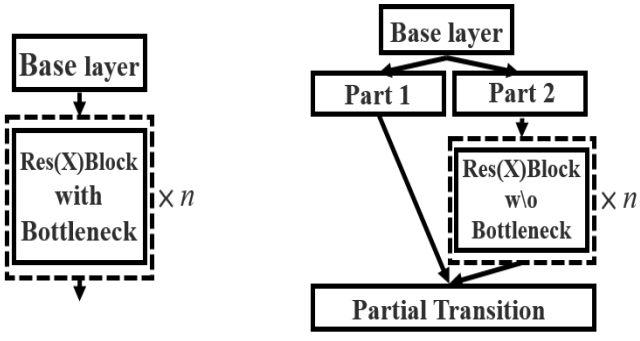
http://doi.org/10.37550/tdmu.VJS/2023.04.451

2. Nhận diện số lượng và lưu lượng trên đường

2.1. Nhận diện và đếm số lượng phương tiện tham gia thông

2.1.1. Cấu trúc CSP (Cross Stage Partial)

Cấu trúc CSP được giới thiệu trong bài báo CSPNet (Wangvà cs., 2019) như là một xương sống giúp làm giảm số lượng tham số trong một mô hình mà không đánh mất quá nhiều độ chính xác của mô hình.

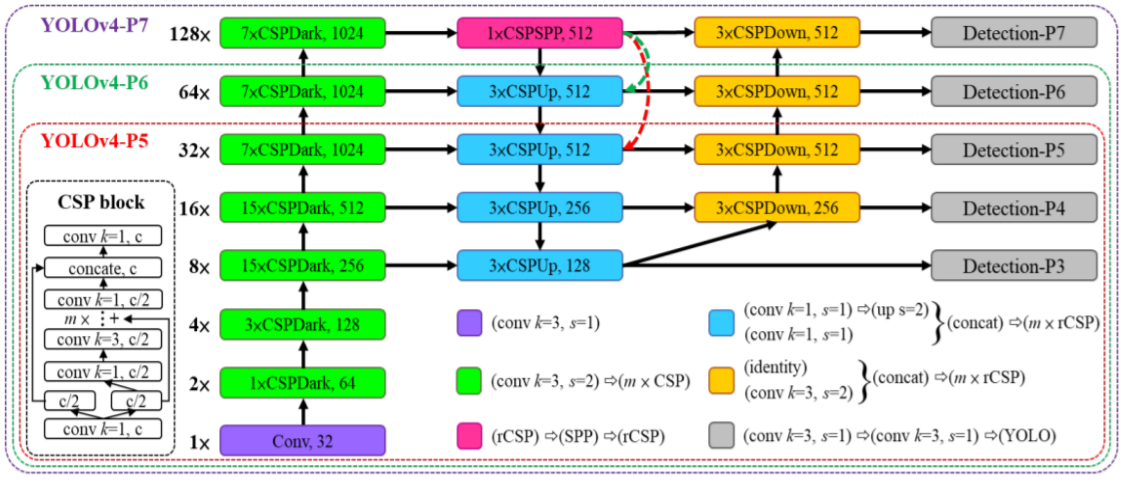


Hình 2 diễn tả cấu trúc CSP trong mô hình ReNext cho thấy so với mô hình gốc thì số lượng tính toán có thể giảm đi khoảng phân nữa. Bởi vì theo cấu trúc csp đặc trưng sẽ được chia ra làm hai phần bằng nhau một sẽ đi xử lý và một phần đi tắt và cuối cùng sẽ ghép 2 phần trên lại.

Hình 2. *Khối CSP trong mô hình ResNeXt.*
a) Không có khối CSP, b) đã được thêm vào khối csp

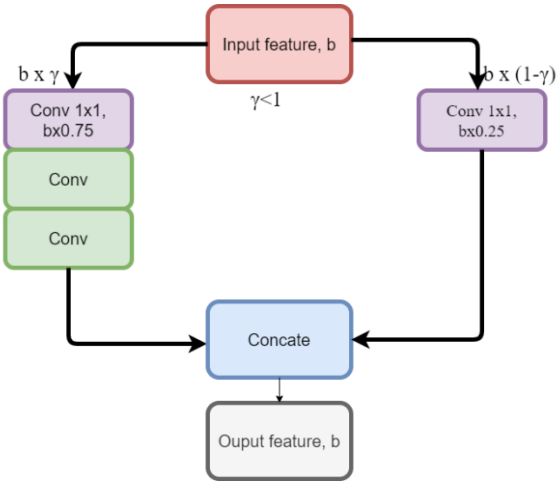
2.1.2. Yolov4-CSP trong nhận dạng đối tượng

YOLOV4 (Bochkovskiyvà cs., 2020) với sự kết hợp với khối CSP được giới thiệu trong bài báo (Wangvà cs., 2019) là một trong những mô hình SOTA (State Of The Art) tốt nhất với 55%AP (Average Precision) trên tập dữ liệu COCO (Lin và cs., 2014).



Hình 3. *Cấu trúc YoloV4-CSP*

Hình 3 cho thấy Yolov4-CSP có 3 sự lựa chọn cho cấu hình P5, P6 và P7 thì số lượng parameter và khối lượng thực thi sẽ khác nhau. Ở khối CSP của bài YoloV4-CSP tác giả đã sử dụng khối CSP bằng cách chia đặc trưng theo tỉ lệ đồng đều cho 2 nhánh.

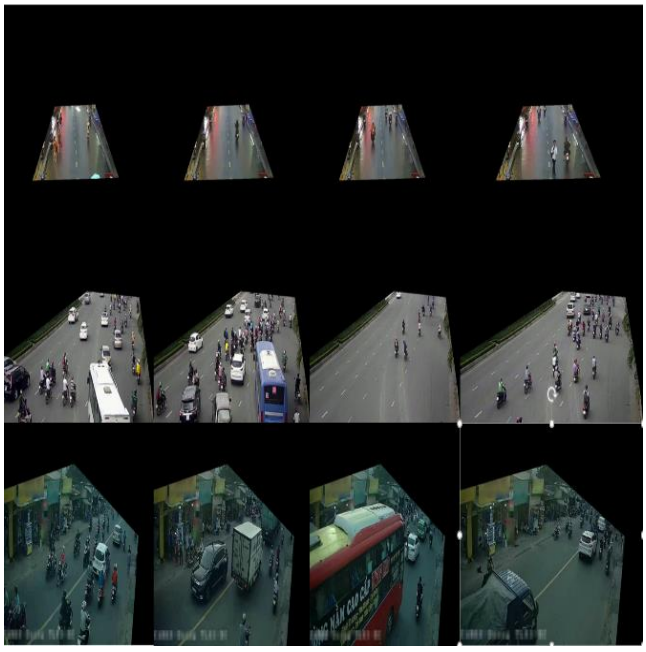


Hình 4. *Khối CSP được chia theo tỉ lệ 75-25*

Trong bài báo này tác giả đề xuất sự thay đổi ở khối CSP đó là lấy 75% đặc trưng đi tiếp tục qua các lớp tích chập và 25% còn lại sẽ đi trực tiếp xuống để ghép với những đặc trưng đã được xử lý, điều đó được diễn tả ở hình 4.

2.1.3. Tập dữ liệu huấn luyện

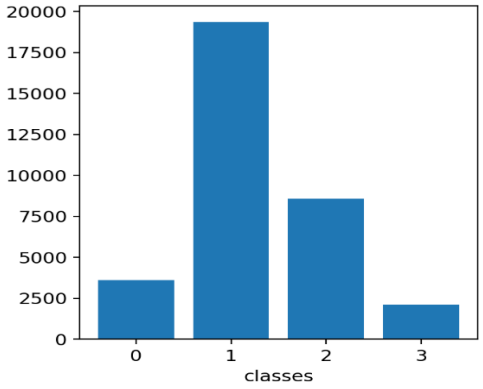
Trong bài báo này tác giả sử dụng bộ dữ liệu được huấn luyện trên tập dữ liệu COCO (Lin và cs., 2014). Sau đó mô hình được huấn luyện lại trên bộ dữ liệu tự gán nhãn bao gồm khoảng 18000 ảnh trên các đoạn phim được quay lại bởi cuộc thi Hồ Chí Minh AI Challenge (2020).



Hình 5. *Bộ dữ liệu huấn luyện trích xuất từ cuộc thi HCM AI Challenge*

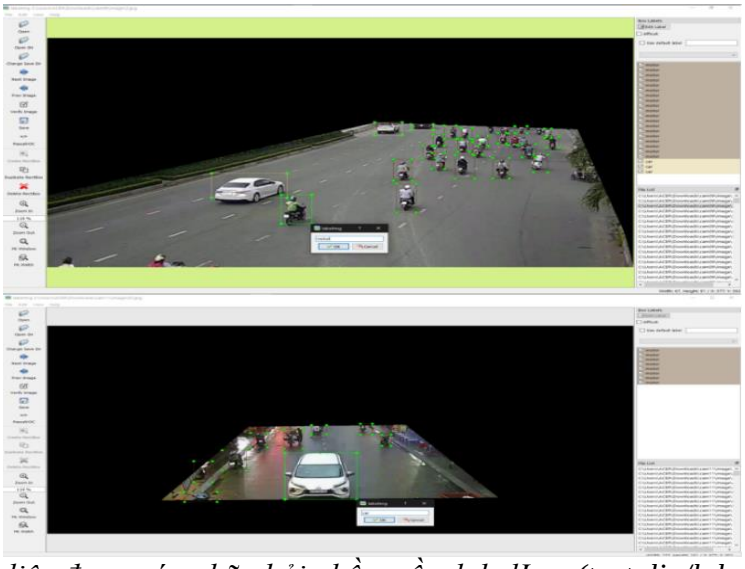
Do sự mất cân bằng giữa các nhóm trong tập dữ liệu nên tác giả đã sử dụng kĩ thuật làm giàu dữ liệu bằng cách xoay, cắt ảnh và sau đó trộn lẫn chúng lại với mục đích cho

ra một bộ dữ liệu tốt hơn cho việc huấn luyện. Nhưng có vấn đề xảy ra là nếu lạm dụng việc trộn lẫn dữ liệu quá nhiều thì mô hình sẽ cho ra kết quả không kém hơn. Sau khi qua quá trình thực nghiệm tác giả chỉ lấy 20% dữ liệu trên 18000 tấm ảnh đi biến đổi (xoay và cắt ảnh) và trộn lẫn chúng để đưa vào bộ dữ liệu huấn luyện.



Hình 6. *Phân bố các nhóm trong data. 0 xe tải, 1 xe máy, 2 xe con, 3 xe bus*

Hình 7. *Dữ liệu sau khi được biến đổi*



Hình 8. *Dữ liệu được gán nhãn bởi phần mềm labelImg (tzutalin/labelImg, 2021)*

2.2. Ước lượng lưu lượng giao thông bằng kỹ thuật đối sánh đặc trưng

Trong bài toán đối sánh đặc trưng ảnh có rất nhiều phương pháp được đưa ra để lấy đặc trưng của ảnh tĩnh bằng các mô tả đặc trưng (Feature Descriptor) để ra các điểm quan trọng (KeyPoint) và đem các điểm đó so sánh với một frame bất kì sau đó để quan sát được sự chuyển động của các vật thể qua một chuỗi bức ảnh tĩnh. Trong bài báo này, tác giả đã thực hiện 2 phương pháp kinh điển trong thị giác máy tính để trích xuất đặc trưng là SIFT (Lowe, 2004) (Scale-invariant feature transform) và SURF (Bay và cs., 2006) (Speed-up robust feature).

2.2.1. Kỹ thuật trích xuất đặc trưng SIFT:

Được giới thiệu bởi David Lowe (1999), SIFT được biết đến với điểm mạnh là ít bị nhiễu bởi sự thay đổi cường độ ánh sáng, nhiễu, góc xoay của ảnh và được ứng dụng trong các lĩnh vực như dựng hình 3D, nhận diện vật thể cũng như là một số lĩnh vực khác. Thuật toán của SIFT (Lowe, 2004) bao gồm 4 giai đoạn.

- Phát hiện cực trị theo tỷ lệ không gian (Scale-space extrema detection).
- Tìm các điểm đăc trưng (Keypoint).
- Xác định hướng.
- Bộ mô tả điểm đặc trưng.

Phát hiện cực trị theo tỷ lệ

Thuật toán được bắt đầu bằng việc tìm kiếm những điểm đặc biệt (Keypoint) bằng cách nhân chập ảnh với các bộ lọc Gaussian ở nhiều kích thước khác nhau và sau đó tấm ảnh được liên tiếp làm mờ Gaussian (Gaussian-blurred) được thực hiện. Các điểm đặc biệt ở đây là các cực đại hoặc cực tiểu của phương pháp Difference of Gaussians (DoG) ở nhiều kích cỡ khác nhau.

$$D(a, b, \alpha) + L(a, b, k_j\alpha) = L(a, b, k_i\alpha) \quad (1)$$

Trong đó:

• $D(a, b, \alpha)$ là DoG của bức ảnh.

• $I(a, b)$ là bức ảnh với chiều ngang là a chiều cao là b.

• $L(a, b, k_j\alpha)$ là kết quả của phép nhân chập giữa tấm ảnh và bộ làm mở Gauss với kích cỡ là $k\alpha$.

• $k_i\alpha, k_j\alpha$ kích cỡ của bộ làm mờ Gauss.

Sau khi tính được DoG của toàn ảnh, xét trên từng pixel so sánh với 8 điểm lân cận và so sánh với 18 điểm tương ứng của kích cỡ ảnh trên dưới, điểm ảnh đó sẽ được coi là keypoint tiềm năng nếu điểm ảnh đó là cực trị địa phương.

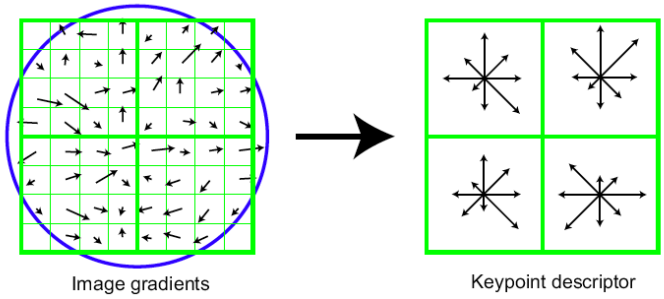
Xác định vị trí Keypoint

Khi có tất cả các keypoint tiềm năng thì bước tiếp theo sẽ là lọc để chọn các keypoint tốt bằng cách sử dụng Hessian 2×2 và dùng một ngưỡng EV (Eigen value) để lọc các keypoint không tốt.

Xác định hướng

Keypoint sẽ được gắn với một số hướng tương ứng, điều đó phụ thuộc vào Image Gradient. Từ đó sẽ đưa ra những phép tính phù hợp cho mỗi keypoint.

Bộ miêu tả đặc trưng



Hình 9. *Miêu tả đặc trưng của SIFT (Lowe, 2004)*

Các bộ miêu tả đặc trưng của SIFT được tính bằng cách lấy 16×16 điểm liền kề của keypoint đó rồi chia thành 4 khối với kích mỗi khối là 4×4. Sẽ có 8 ngăn (bin) để chứa hướng được tạo ra, vì vậy mỗi bộ miêu tả sẽ có 128 ngăn tạo thành một vector đặc trưng.

Hình 10. *Đặc trưng SIFT trên ảnh xám*

2.2.2 Kỹ thuật trích xuất đặc trưng SURF:

Kỹ thuật trích xuất đặc trưng nhanh và bền vững được Bay, Ess, T. Tuytelaars, and Van Gool giới thiệu vào năm 2009. Thuật toán trích xuất đặc trưng nhanh và bền vững cũng dựa vào những bước cơ bản như SIFT nhưng cụ thể nội dung cơ bản ở từng bước là khác nhau. SURF sử dụng các bộ lọc vuông như một phép xấp sỉ của Gaussian không giống như SIFT, không giống như thuật toán SIFT sử dụng các bộ lọc phân tầng để phát hiện những điểm đặc trưng bất biến theo tỷ lệ trong đó chúng ta có thể tính toán các DoG (Difference of Gaussian) dần dần trên các hình ảnh có kích cỡ thay đổi. Bởi vì sử dụng những bộ lọc vuông nên SURF cho tốc độ nhanh hơn so với SIFT (sử dụng DoG trên nhiều kích cỡ hình ảnh). Chi tiết thuật toán SURF (Bay và cs., 2006) cũng chia ra làm 2 phần:

- Phát hiện điểm quan tâm (Interest point detection)
- Bộ mô tả vùng lân cận cục bộ (Local neighborhood description)

Phát hiện điểm quan tâm.

SURF sử dụng một bộ phát hiện các điểm dựa trên Hessian ma trận để tìm những điểm mong muốn. Bên cạnh đó định thức của ma trận Hessian cũng được sử dụng làm thước đo để đo sự thay đổi cục của các điểm cục bộ xung quanh và các điểm được chọn để quyết định xem điểm đó có phải là cực đại hay không và đồng thời định thức của ma trận Hessian còn được dùng để chọn kích cỡ bộ lọc.

$$H(d,\delta) = \begin{pmatrix} L_{ww}(d,\delta) & L_{wh}(d,\delta) \\ L_{hw}(d,\delta) & L_{hh}(d,\delta) \end{pmatrix}$$

Trong đó:

- $d\ (w, h)$ là một điểm trên ảnh gốc
- $H(d,\delta)$ là ma trận Hessian tại điểm d và kích cỡ δ.
- $L_{ww}(d,\delta)$ là đạo hàm bậc hai của Gaussian với ảnh gốc tại điểm d

Biểu diễn kích cỡ và không gian của các điểm cần quan tâm.

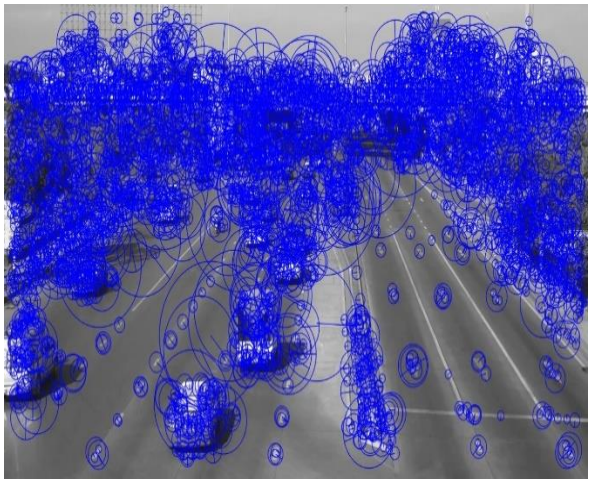
Việc phát hiện các keypoint thường được diễn ra ở các kích cỡ khác nhau, bởi vì một phần việc tìm kiếm các điểm tương ứng đòi hỏi sự so sánh các tấm ảnh ở nhiều kích cỡ khác nhau. Ở đây SURF sử dụng kỹ thuật kim tự tháp đa độ phân giải (Multi-resolution pyramid), bức ảnh được dịch thành tọa độ để tái tạo bức ảnh gốc với hình dạng Kim tự tháp Gaussian hoặc Laplacian Pyramidal, dẫn đến một bức ảnh có cùng kích thước nhưng băng thông thấp hơn. Điều này tạo ra hiệu ứng làm mờ Scale-Space trên hình ảnh gốc, đảm bảo rằng các điểm ưa thích vẫn bất biến về kích thước.

Bộ miêu tả đặc trưng.

SURF (Bay và cs., 2006) sử dụng bộ miêu tả đặc trưng dùng để cung cấp mô tả về các đặc điểm của các điểm keypoint và những điểm hàng xóm của nó. Trong bộ miêu tả của SURF quá trình được chia làm 2 gia đoạn.

- Giai đoạn đầu tiên là thiết lập định hướng có thể lặp lại bằng cách sử dụng dữ liệu từ một vùng hình tròn xung quanh điểm quan tâm. Sau đó, bộ mô tả SURF được trích xuất từ một vùng hình vuông được căn chỉnh theo hướng được chỉ định.

- Giai đoạn sau đó sẽ xây dựng bộ miêu tả dựa trên tổng số phản hồi của Wavelet Haar (2021).



Hình 11. *Đặc trưng SIFT trên ảnh xám*

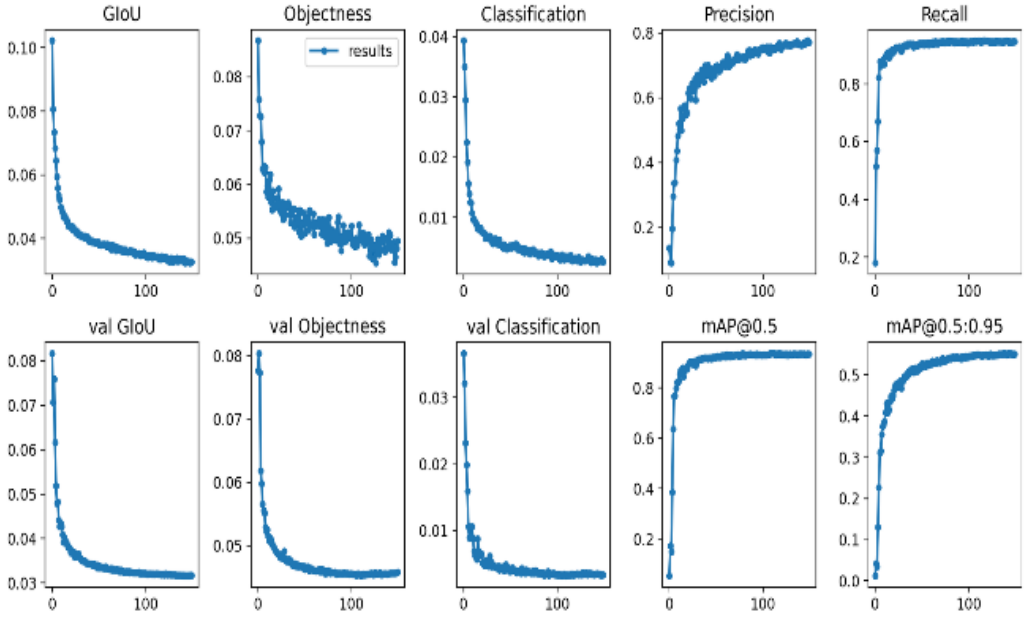
2.2.3 Kỹ thuật đối sánh đặc trưng

Đối sánh đặc trưng trên ảnh là một trong những kỹ thuật được sử dụng trong rất nhiều bài toán thị giác máy tính như hiệu chỉnh camera, nhận dạng đối tượng, đo dòng lưu lượng thay đổi của các điểm ảnh. Việc đối sánh của đặc trưng của ảnh được tiếp cận phổ biến bằng cách tìm cách điểm đặc trưng cần quan tâm giữa nhiều tấm ảnh. Trong bài báo này tác giả đã sử dụng kỹ thuật SIFT và SURF như được nêu trên để tìm ra các keypoints trên từng tấm ảnh và đối sánh chúng với nhau thông qua các bộ mô tả và kết hợp với các điểm keypoints. Trong bài báo này tác giả sử dụng bộ đối sánh Brute-Force là một bộ đối sánh đơn giản. Nó lấy mô tả của một đối tượng trong bộ đầu tiên và được khớp với tất cả các đối tượng khác trong bộ thứ hai bằng cách sử dụng một số tính toán khoảng cách. Sau đó sử dụng KNN như một bộ đối sánh cuối cùng ta sẽ nhận được k cặp điểm phù hợp nhất với k được chỉ định bởi người sử dụng.

3. Kết quả thực nghiệm

Chuẩn bị cho việc thực nghiệm việc nhận dạng các phương tiện giao thông tác giả đã sử bộ dữ liệu COCO và tập dữ liệu tự gán nhãn trích xuất từ camera của cuộc thi Hồ Chí Minh AI Challenge. Sau khi huấn luyện tác giả đã đi thu thập dữ liệu ùn tắc giao thông bằng cách tìm kiếm trên internet. Cấu hình máy để train bộ dữ liệu:

– Thông tin hệ thống: Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz, GPU Tesla T4 16 GB VRAM.



Hình 12. *Đồ thị quá trình huấn luyện mô hình YOLOV4-CSP (75-25)*

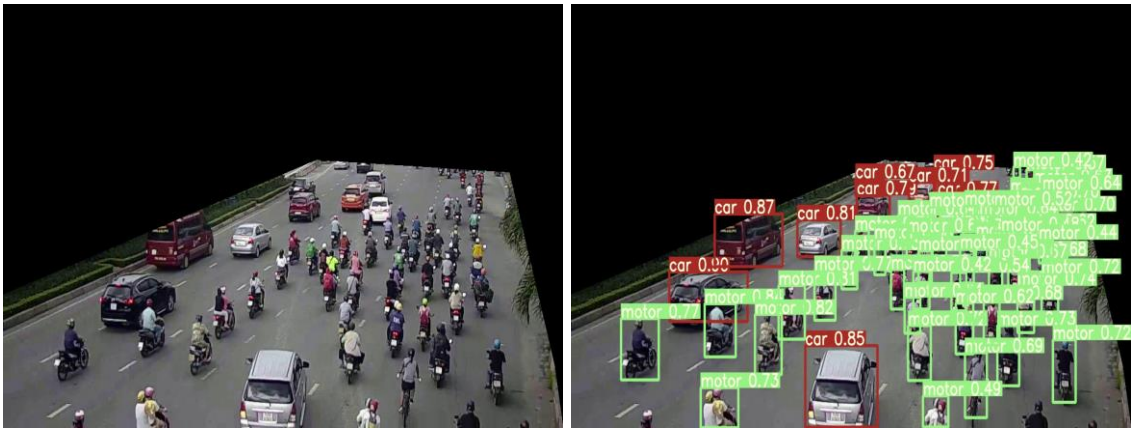
Vì trong bài toán này tác giả tập trung nhiều về recall và cả độ chính xác nên mô hình đã đạt đến khoảng 90% recall và độ chính xác lên đến khoảng 80%.

(a) (b)

(c) (d)

***Hình 13**. Kết quả của mô hình*

(a) Bộ dữ liệu tự gán nhãn và cấu trúc CSP 75-25 confidence 0.4 (b) Bộ dữ liệu tự gán nhãn và cấu trúc CSP 50-50 confidence 0.5, (c) Bộ dữ liệu COCO, (d)Bộ dữ liệu tự gán nhãn và cấu trúc CSP 75-25 confidence 0.5.



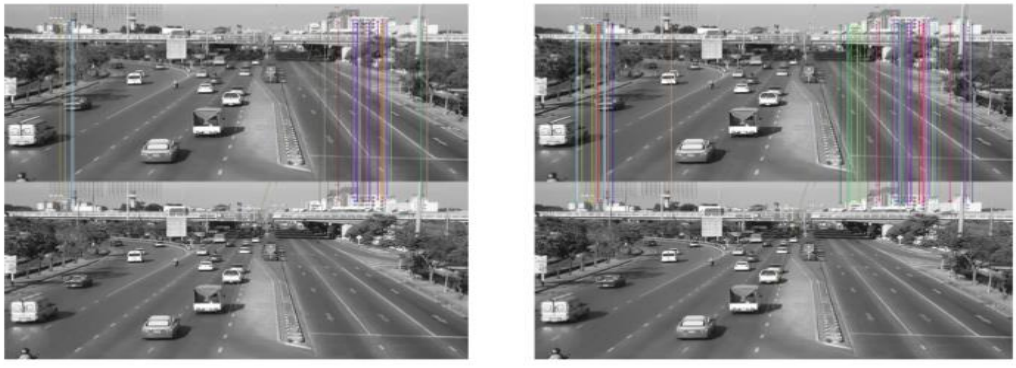
(a) (b)

***Hình 14.** Nhận diện đông phương tiện giao thông trên xa lộ Hà Nội với độ chính xác cao*

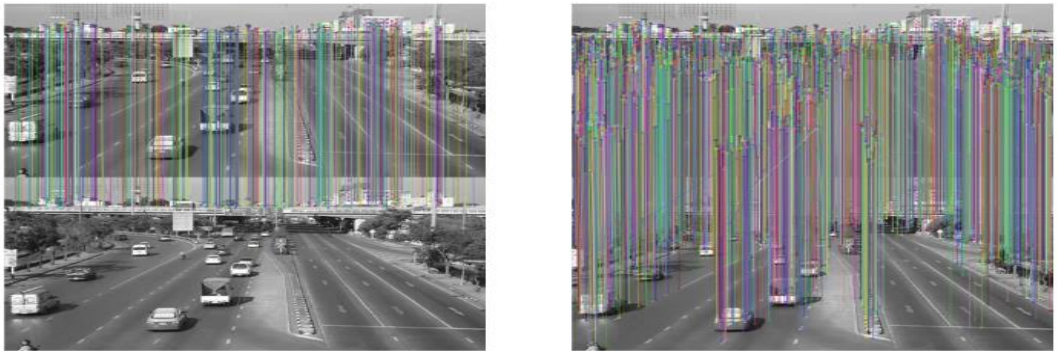
Sau khi đã ước lượng được số lượng phương tiện tham gia giao thông thì tác giả sẽ ước lượng dòng lưu lượng di chuyển của các phương tiện là nhanh hay chậm. Bằng cách đối sánh đặc trưng giữa 2 frame giống nhau và tính hiệu vị trí của chúng. Nếu biết thông số của camera thì chúng ta có thể dễ dàng tính thương giữa khoảng hiệu 2 vị trí và thời gian khung hình trả về thì có thể đo được tốc độ của dòng lưu thông. Nhưng do hạn chế về mặt thu thập dữ liệu. Tác giả dùng những bộ dữ liệu có sẵn, nên trong bài báo này tác giả sẽ chỉ đưa ra hiệu giữa 2 đặc trưng của các frame.

Đối sánh đặc trưng SIFT, SURF

Hình 15. *Trích xuất SIFT keypoint.*

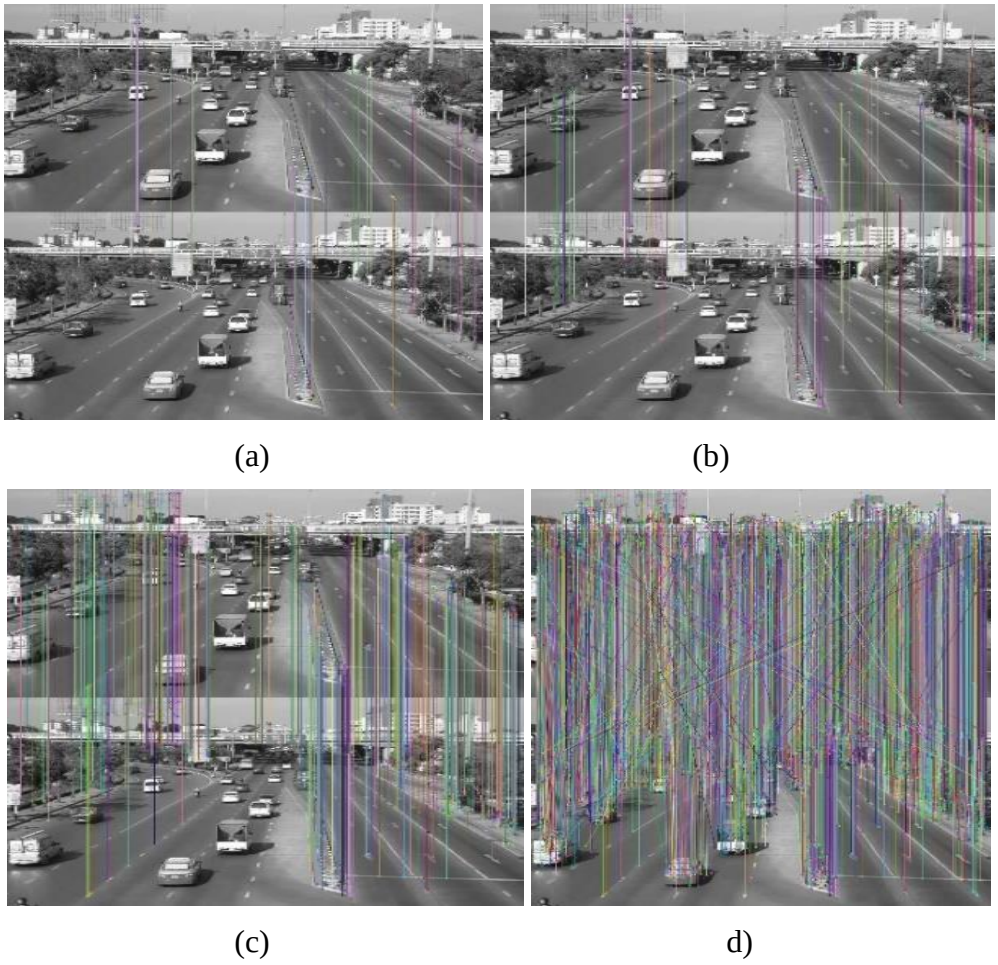


(a) (b)



(c) (d)

Hình 16. *Đối sánh đặc trưng SIFT. a) lấy 100 keypoints, b) lấy 500 keypoints,c) lấy 1000 keypoints, d) lấy hết tất cả keypoints*



(a) (b)

(c) d)

Hình 17. *Đối sánh đặc trưng SIFT. a) lấy 100 keypoints, b) lấy 500 keypoints, c) lấy 1000 keypoints, d) lấy hết tất cả keypoints*

Bảng 1. *Đo hiệu khoảng cách các keypoint giữa các khung ảnh*

Số thứ tự khung ảnh so với khung ảnh đầu tiên	SIFT (pixel)	SURF (pixel)
Khung ảnh 1	0.118074276	0.10093031
Khung ảnh 2	0.13510878	0.13959445
Khung ảnh 3	0.14665028	0.07892276
Khung ảnh 4	0.2821357	0.04215052
Khung ảnh 5	0.43080702	0.023025826
Khung ảnh 6	0.28849754	0.18322277
Khung ảnh 7	0.4476639	0.22903006
Khung ảnh 8	0.46450144	0.029168062
Khung ảnh 9	0.5708234	0.12783214
Khung ảnh 10	0.84563005	0.20793006
Khung ảnh trên giây	1.5	4.5

Bảng 1 thể hiện khoảng cách thay đổi so với bức ảnh gốc qua từng khung ảnh kế tiếp. Các keypoint ở tấm ảnh gốc và lần lượt các tấm ảnh kế tiếp. Cuối cùng trung bình cộng của

hiệu các khoảng cách giữa 2 khung ảnh liên tiếp sẽ được tính để dự đoán tốc độ của dòng lưu lượng. Từ đó kết hợp với số lượng phương tiện đã được đếm từ phần nhận diện của YOLOV4-CSP và cho ra quyết định cuối cùng. Theo đánh giá cho thấy nếu như đo lưu lượng sử dụng SIFT thì sẽ ổn định hơn nhưng bù lại thì thời gian xử của SIFT sẽ dài hơn.

4. Kết luận

Trong bài báo này với mục đích được nhận diện được các đoạn đường kẹt xe nhờ vào dữ liệu được thu thập bởi các camera giao thông. Về mặt thuật toán YOLOV4-CSP đã làm tốt trong việc đếm số lượng các phương tiện giao thông trong mật độ đông nhờ vào các sự thay đổi ở cấu trúc CSP và việc cân bằng dữ liệu bằng cách trộn lẫn dữ liệu (mixup data). Về việc đo lưu lượng dòng di chuyển phương pháp SURF đã thể hiển được ưu điểm của nó là tốc độ xử lý nhưng bù lại bằng độ chính xác bị mất rất nhiều trong môi trường cảnh quang phức tạp.

Trong tương lai tác giả dự định sử dụng thêm thuật toán RANSAC (Random Sample Consensus) để giải quyết các keypoint bị nhiễu để tăng độ chính xác cho phần đo dòng lưu lượng.

TÀI LIỆU THAM KHẢO

[1] Bochkovskiy, C. Y. Wang, and H. Y. M. Liao (2021). YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv. arXiv*. Available: https://github.com/AlexeyAB/darknet.

[2] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao (2020). *Scaled-YOLOv4: Scaling Cross Stage Partial Network*. Available: http://arxiv.org/abs/2011.08036.

[3] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh (2020). *CSPNet: A New Backbone that can Enhance Learning Capability of CNN*. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work, pp. 1571-1580, Available: http://arxiv.org/abs/1911.11929.

[4] D. G. Lowe (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, *60*(2), pp. 91-110. doi: 10.1023/B:VISI.0000029664.99615.94.

[5] Difference of Gaussians - Wikipedia. https://en.wikipedia.org/wiki/Difference_of_Gaussians (accessed May 28, 2021).

[6] H. Bay, T. Tuytelaars, and L. Van Gool (2006). SURF: Speeded up robust features, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3951 LNCS, pp. 404-417, doi: 10.1007/11744023_32.

[7] Haar wavelet - Wikipedia. https://en.wikipedia.org/wiki/Haar_wavelet (accessed May 28,

[8] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian (2019). CenterNet: Keypoint Triplets for Object Detection. Proc. *IEEE Int. Conf. Comput. Vis*, vol. 2019, pp. 6568–6577. Accessed: May 28, 2021. [Online]. Available: http://arxiv.org/abs/1904.08189.

[9] K. Hasan Talukder and K. Harada. Haar Wavelet Based Approach for Image Compression and Quality Assessment of Compressed Image.

[10] Nidhal, U. K. Ngah, and W. Ismail (2014). *Real time traffic congestion detection system*. doi: 10.1109/ICIAS.2014.6869538.

[11] Scale-invariant feature transform - Wikipedia. https://en.wikipedia.org/wiki/Scale-invariant_feature_transform (accessed May 28, 2021).

[12] T. Y. Lin et al (2014). Microsoft COCO: Common objects in context, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8693 LNCS, no. PART 5, pp. 740-755, doi: 10.1007/978-3-319-10602-1_48.

[13] TRANG CHỦ - Hội thi giải pháp ứng dụng Trí tuệ Nhân tạo (AI) TP.HCM 2020. http://ai.icti-hcm.gov.vn/ (accessed May 28, 2021).

[14] tzutalin/labelImg: LabelImg is a graphical image annotation tool and label object bounding boxes in images. https://github.com/tzutalin/labelImg (accessed May 28, 2021).

[15] Y. Xu, G. Yu, Y. Wang, X. Wu, and Y. Ma (2017). "Car detection from low-altitude UAV imagery with the faster R-CNN. *J. Adv. Transp*. doi: 10.1155/2017/2823617.