HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

TRẦN QUỐC ĐẠT

KỸ THUẬT HỌC SÂU CHO BÀI TOÁN THEO VẾT ĐA ĐỐI TƯỢNG

Chuyên ngành: HỆ THỐNG THÔNG TIN

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

(Theo định hướng ứng dụng)

TP. HỒ CHÍ MINH – NĂM 2021

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS.TS LÊ HOÀNG THÁI**

Phản biện 1: .....

Phản biện 2: .....

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn tại Học viện Công nghệ Bưu chính Viễn Thông

Vào lúc: ..... giờ ..... ngày ..... tháng ..... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu Chính Viễn Thông.

MỞ ĐẦU

Trong những năm gần đây, việc phát hiện và tái xác định đối tượng đã có nhiều tiến bộ đáng kể. Hai kỹ thuật này là thành phần cốt lõi để hình thành hệ thống theo dõi đa đối tượng. Tuy nhiên, việc hoàn thành hai nhiệm vụ trong một mạng duy nhất để cải thiện tốc độ suy luận chưa được quan tâm nhiều. Các nỗ lực ban đầu cho việc hợp nhất hai nhiệm vụ trên cho kết quả thấp. Nguyên nhân chủ yếu: là do kỹ thuật tái nhận dạng chưa được huấn luyện phù hợp. Trong luận văn, chúng tôi tìm hiểu những lý do cơ bản đằng sau sự thất bại; tiến tới, đề nghị một phương pháp cơ bản đơn giản để giải quyết các vấn đề.

Mục tiêu của hệ thống đề xuất là: dự đoán đường đi của nhiều vật thể được chú ý trong các video. Xây dựng một mô hình nhận dạng theo vết nhiều đối tượng tiến tới xa hơn có áp dụng mô hình hệ thống cho một số lĩnh vực thực tế như: an ninh quốc phòng, giao thông vận tải,…

Luận văn gồm 5 chương chính với các nội dung sau:

Chương 1: Giới thiệu về phương pháp dò tìm đối tượng, phân tích các vấn đề đang gặp phải của phương pháp và đề xuất các giải pháp, kỹ thuật có thể áp dụng vào đề tài.

Chương 2: Trình bày về các công trình nghiên cứu trong và ngoài nước liên quan mật thiết tới đề tài

Chương 3: Trình bày quy trình thực hiện dò tìm và tái định danh đối tượng, từ quá trình huấn luyện & nội suy đặc trưng đến việc theo vết online và đánh giá độ chính xác của mô hình

Chương 4: Trình bày chi tiết việc xây dựng bộ dữ liệu, quá trình cụ thể cài đặt mô hình cho thuật toán và đánh giá kết quả thực nghiệm trên bộ dữ liệu xây dựng với các bộ dữ liệu có sẵn khác.

Chương 5: Kết luận nội dung đã được trong đề tài, nêu những khó khăn, hạn chế trong quá trình nghiên cứu đã gặp phải và đề xuất hướng phát triển tiếp theo.

Đề tài: KỸ THUẬT HỌC SÂU CHO BÀI TOÁN THEO VẾT ĐA ĐỐI TƯỢNG

Tóm tắt luận văn

CHƯƠNG 1. CƠ SỞ LÝ THUYẾT

1.1. Các phương pháp dò tìm đối tượng

1.2. Phân tích vấn đề

- Neo không phù hợp với Re-ID

- Tổng hợp đặc trưng trên nhiều lớp

- Kích thước của các đặc trưng Re-ID

1.3. Giải pháp

1.3.1 Giới thiệu hướng tiếp cận mới

1.3.2 Mạng xương sống (Backbone Network)

1.3.3 Nhánh phát hiện vật thể

1.3.4 Nhánh định danh vật thể

1.4 Các kỹ thuật áp dụng

1.4.1 Hàm lỗi

- Kỹ thuật Focal Loss

- Heatmap Loss

- Offset and Size Loss

- Identity Embedding Loss

1.4.2 Online Tracking

- Network Inference

- Online Box Linking

1.5 Kết luận chương 1

CHƯƠNG 2. CÁC CÔNG TRÌNH LIÊN QUAN

2.1. Phương pháp Two-Step MOT

Các phương pháp MOT hiện đại như [2] [4] [6] [7] [5] thường coi việc phát hiện đối tượng và Re-ID là hai nhiệm vụ riêng biệt. Đầu tiên chúng áp dụng các bộ dò CNN như [17] [18] [19] để khoanh vùng tất cả các đối tượng được quan tâm trong hình ảnh theo các hộp (boxes). Sau đó, trong một bước riêng biệt, họ cắt hình ảnh theo các hộp và đưa chúng vào mạng nhúng nhận dạng để trích xuất các tính năng Re-ID và liên kết các hộp với tạo thành nhiều track. Các công trình thường tuân theo một phương pháp tiêu chuẩn để liên kết hộp, trước tiên tính toán ma trận chi phí theo các tính năng Re-ID và Intersection over Unions (IoU) (chỉ số đánh giá được sử dụng để đo độ chính xác của Object detector trên tập dữ liệu cụ thể) của các hộp giới hạn (bounding boxes), sau đó sử dụng Kalman Filter [20] và thuật toán Hungarian [21] để hoàn thành nhiệm vụ liên kết. Một số lượng nhỏ các công trình như [6] [5] [7] sử dụng các chiến lược liên kết phức tạp hơn như mô hình nhóm và RNNs.

Ưu điểm của phương pháp two-step là chúng có thể sử dụng mô hình phù hợp nhất cho từng nhiệm vụ tương ứng mà không cần thỏa hiệp. Ngoài ra, chúng có thể cắt hình ảnh theo các hộp giới hạn đã phát hiện và thay đổi kích thước của chúng thành cùng một kích thước trước khi dự đoán các tính năng Re-ID. Điều này giúp xử lý các biến thể tỷ lệ của đối tượng. Kết quả là, những cách tiếp cận này [4] đã đạt được hiệu suất tốt nhất trên các bộ dữ liệu công khai. Tuy nhiên, chúng thường rất chậm vì cả tính năng phát hiện đối tượng và nhúng tính năng Re-ID đều cần nhiều tính toán mà không cần chia sẻ giữa chúng. Vì vậy, thật khó để đạt được suy luận về tốc độ video vốn được yêu cầu trong nhiều ứng dụng.

2.2 Phương pháp One-Shot MOT

Với sự phát triển vượt bậc của tính năng học đa tác vụ [8] [22] [23] trong học sâu, one-shot MOT đã bắt đầu thu hút nhiều sự chú ý của công cuộc nghiên cứu. Ý tưởng cốt lõi là thực hiện đồng thời việc phát hiện đối tượng và nhúng danh tính (các tính năng Re-ID) trong một mạng duy nhất để giảm thời gian suy luận thông qua việc chia sẻ hầu hết các tính toán. Ví dụ: Track-RCNN [8] thêm đầu Re-ID ở trên cùng của Mask-RCNN [17] và hồi quy của một hộp giới hạn và tính năng Re-ID cho mỗi

đề xuất. JDE [10] được giới thiệu trên cùng của YOLOv3 [18] framework giúp đạt được suy luận tốc độ video gần bằng nhau.

Tuy nhiên, độ chính xác theo dõi của phương pháp one-shot thường thấp hơn so với phương pháp two-step. Điều này là do các tính năng Re-ID đã học không tối ưu, dẫn đến số lượng lớn các công tắc ID. Để giải quyết vấn đề, chúng tôi đề xuất sử dụng các phương pháp tiếp cận không có mỏ neo cho cả phát hiện đối tượng và nhúng danh tính để cải thiện đáng kể độ chính xác theo dõi trên tất cả các điểm chuẩn.

2.3 Các công trình khác

- “TransMOT: Spatial-Temporal Graph Transformer for Multiple Object Tracking”

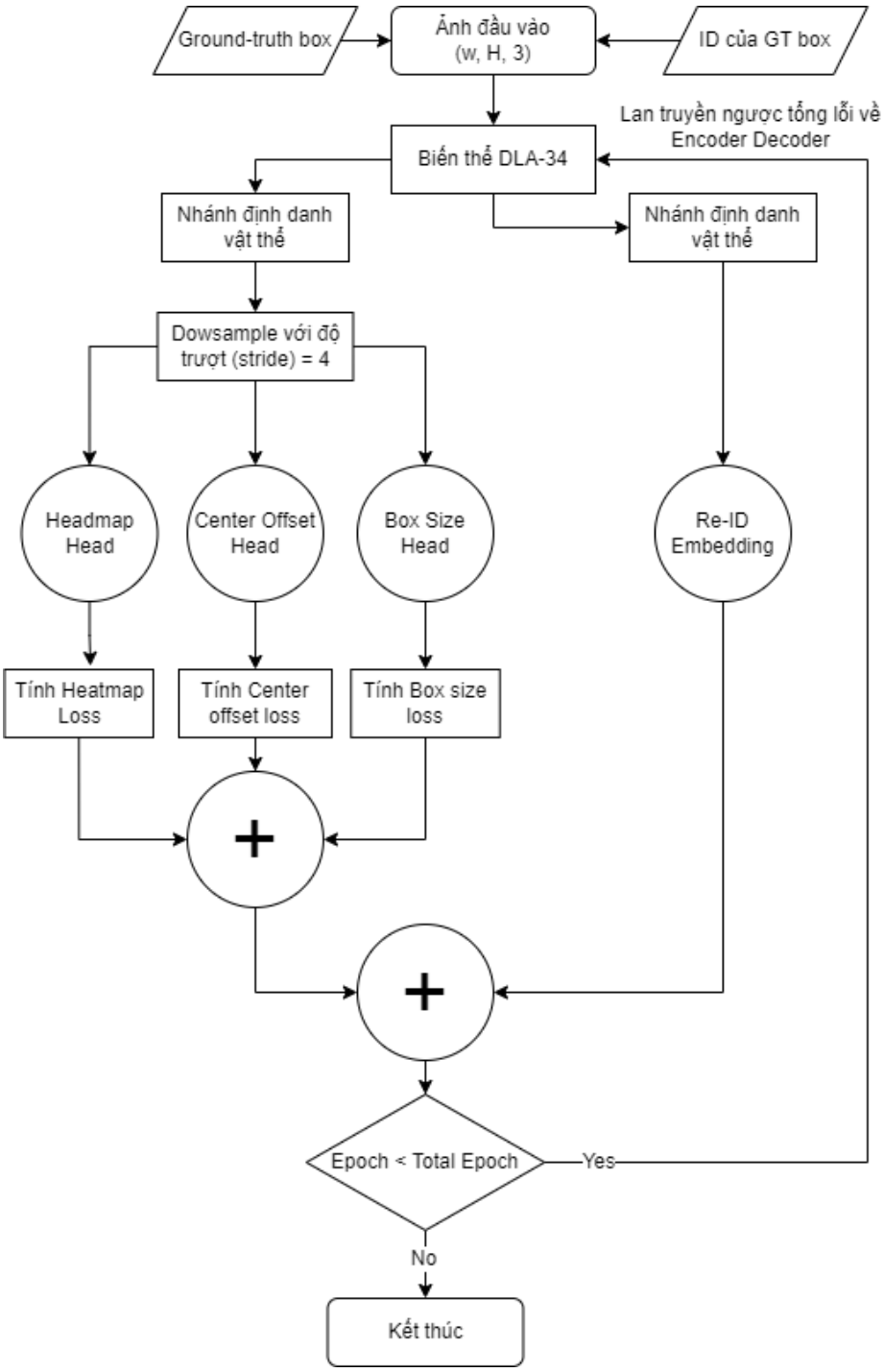
- “ByteTrack: Multi-Object Tracking by Associating Every Detection Box”

- “Track to Detect and Segment: An Online Multi-Object Tracker”

CHƯƠNG 3. QUY TRÌNH THỰC HIỆN DÒ TÌM VÀ TÁI ĐỊNH DANH ĐỐI TƯỢNG

3.1 Huấn luyện và nội suy ra đặc trưng

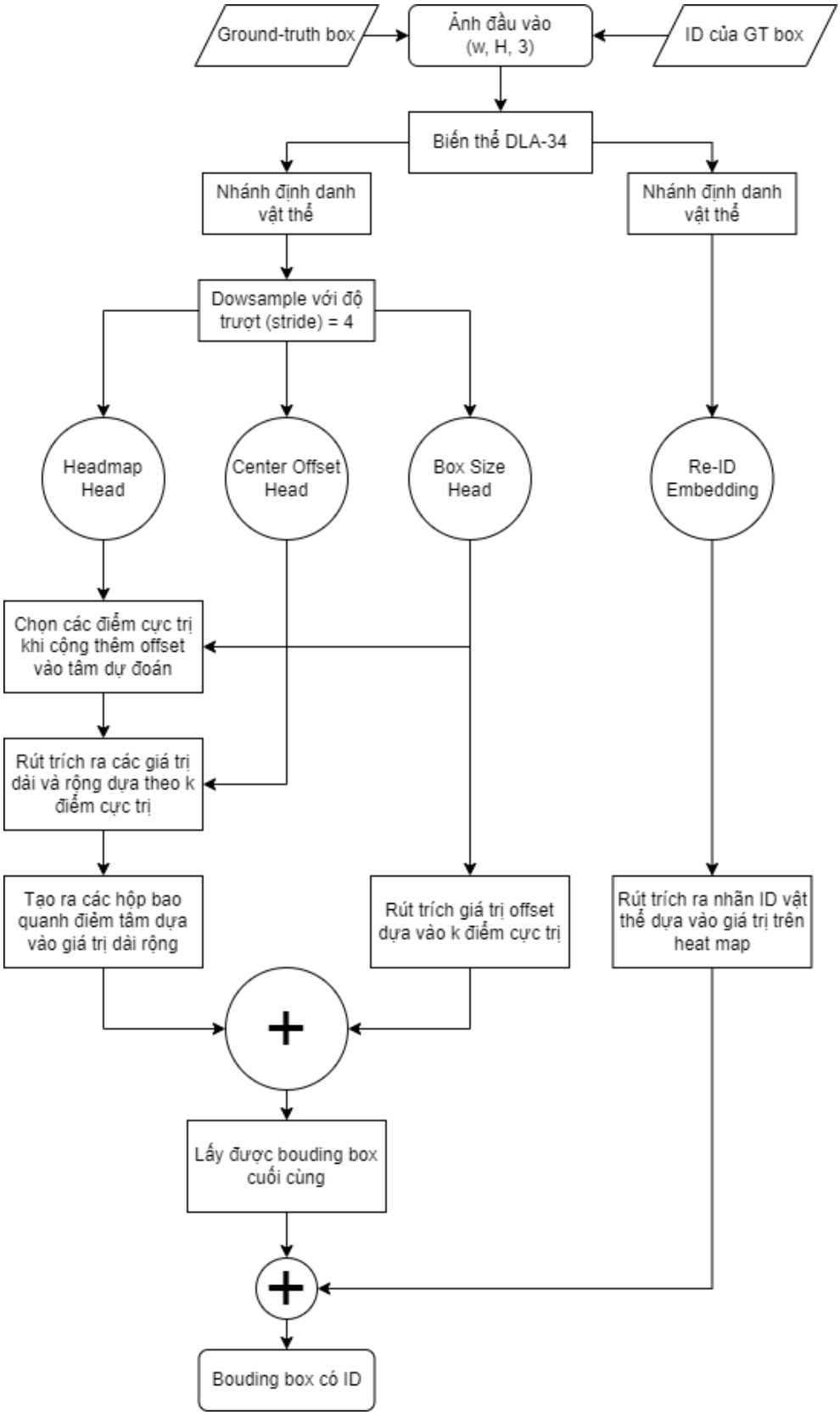
3.1.1. Huấn luyện



*Hình **Error! No text of specified style in document.**.1 Flowchart huấn luyện*

3.1.2. Nội suy đặc trưng

Đặc trưng ở đây bao gồm bounding box và ID của bounding box đó.



*Hình **Error! No text of specified style in document.**.2 Flowchart mô tả cách nội suy đặc trưng*

3.2. Theo vết online (Online Tracking)

Các phương pháp trước đây hiện thực Trình theo dõi thường qua 4 bước chính:

1. Khung ban đầu của đoạn phim
2. Phát hiện đối tượng: Sử dụng các mạng phát hiện đối để trích xuất ô chứa đối tượng.
3. Định danh đối tương: Tính độ tương đồng giữa các ô chứa đối tượng ở khung trước đó và khung hiện tại (đối tượng giống nhau có khoảng cách nhỏ, khác nhau có khoảng cách lớn).
4. Liên kết đối tượng với ID được tạo từ khung trước, tạo mới nếu chưa từng phát hiện

Điểm khác biệt chính là ở bước số 2 và 3 thay vì thực hiện từng bước riêng biệt, nhóm tác giả sử dụng Deep Layer Aggregation để tạo ra bộ đặc trưng có thể dùng cho cả bước Phát hiện đối tượng và Định danh đối tượng, từ đó cho phép trình theo dõi hoạt động với tốc độ cao hơn so với phương pháp truyền thống.

3.3. Đánh giá độ chính xác của mô hình

Trong vài năm gần đây, cộng đồng theo dõi đa đối tượng đã phát triển mạnh mẽ một phần là do đầu tư lớn từ ngành công nghiệp xe tự hành. Điều này đã dẫn đến một số lượng lớn các MOT benchmark mới được đề xuất. Nhiều trình theo dõi xếp hạng này đã sử dụng chỉ số đánh giá MOTA (Multiple Object Tracking Accuracy) [9], [27],... Chỉ số này đo lường độ chính xác tổng thể của cả trình theo dõi và phát hiện. Nó xử lý cả đầu ra theo dõi và đầu ra phát hiện.

Thước đo thứ hai được sử dụng gần đây được MOT benchmark áp dụng là IFD1 [28]. Chỉ số này được đề xuất đặc biệt để theo dõi các đối tượng sử dụng bởi ‘multi-camera MOT’. IDF1 gần đây cũng đã được triển khai như một chỉ số phụ trên tiêu chuẩn MOTChallenge và đã được ưu tiên hơn MOTA để đánh giá bởi một số phương pháp theo dõi camera đơn. Chỉ số IDF1 là tỷ lệ giữa các phát hiện được xác định chính xác trên số lượng trung bình của các phát hiện xác thực và được tính toán.

Tiếp theo chỉ số HOTA (Higher Order Tracking Accuracy) được định nghĩa bởi [29] có thể đánh giá tất cả các khía cạnh của việc theo dõi. Chỉ số HOTA đo lường rõ ràng cả hai loại lỗi (nhấn mạnh quá mức đến việc phát hiện và liên kết) và kết hợp

chúng một cách cân bằng. HOTA cũng tích hợp tỷ lệ đo lường độ chính xác bản địa hóa của các kết quả theo dõi không có trong MOTA hoặc IDF1. HOTA có thể được sử dụng như một chỉ số thống nhất duy nhất để xếp hạng các trình theo dõi, đồng thời phân tách thành một nhóm các chỉ số phụ có thể đánh giá các khía cạnh khác nhau của việc theo dõi riêng biệt và cho phép các trình theo dõi được điều chỉnh cho các yêu cầu khác nhau.

3.3. Kết luận chương 3

CHƯƠNG 4. CÀI ĐẶT VÀ THỰC NGHIỆM

4.1. Tập dữ liệu thực nghiệm

4.1.1. Tập dữ liệu đã công bố: Multiple Object Tracking Benchmark

Bộ dữ liệu chứa các chuỗi video trong môi trường không bị giới hạn được quay bằng cả máy ảnh tĩnh và máy ảnh chuyển động. Theo dõi và đánh giá được thực hiện trong các tọa độ hình ảnh. Trong luận văn này sẽ lấy tập dữ liệu training của mỗi bộ MOT gồm MOT15, MOT16, MOT17, MOT20.

4.1.2. Tập dữ liệu xây dựng

Bộ dữ liệu này được xây dựng ở các khu vực công cộng ở Thành phố Hồ Chí Minh và Tây Ninh kết hợp với một video tìm kiếm được. Bộ video tập trung chủ yếu vào việc ghi lại cảnh hoạt động, di chuyển của người dân trong khu vực.

Thông tin chi tiết của bộ dữ liệu được trình bày trong bảng dưới đây:

*Bảng **Error! No text of specified style in document.**.1 Thông tin của tập dữ liệu MOT25*

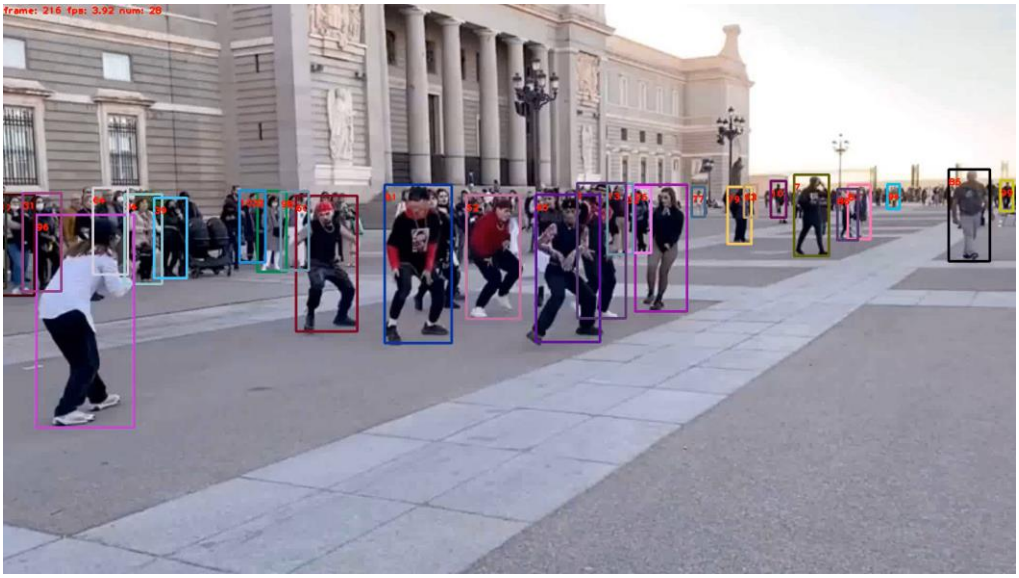
Tên video	FPS	Độ phân giải	Độ dài (số frame, số giây)
MOT25-01	30	1920x1080	1800 (01:00)
MOT25-03	24	1920x1080	370 (00:15)
MOT25-05	25	1920x1080	220 (00:08)
MOT25-07	30	1920x1080	1762 (00:59)
MOT25-09	30	1920x1080	3846 (02:09)
MOT25-11	30	1920x1080	1410 (00:47)
MOT25-13	30	1920x1080	3936 (02:12)
MOT25-15	30	1920x1080	991 (00:33)

4.2. Xây dựng bộ dữ liệu MOT25 Chi tiết quá trình huấn luyện

4.2.1. Xây dựng tracker

Từ tập dữ liệu MOT25, tiến hành chạy tuần tự tập lệnh (được công bố bởi nghiên cứu của tác giả [32]) sau trên môi trường của Google Colaboratory [33] (bật chế độ sử dụng bộ tăng tốc phần cứng GPU) để xây dựng bộ dữ liệu tracker cho bài

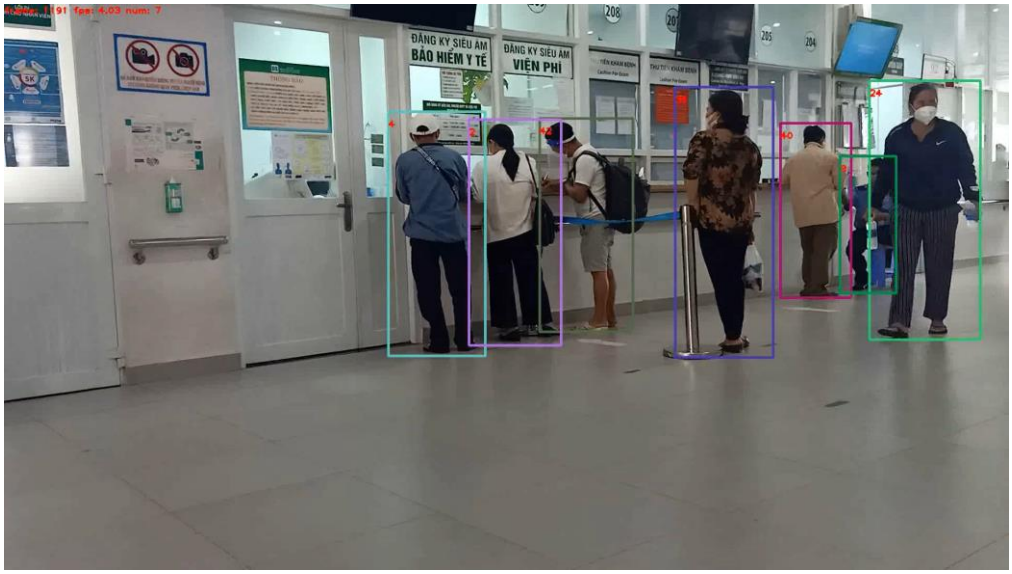
nghiên cứu. Tiến hành chạy mô hình để phát hiện các đối tượng trên các bộ dữ liệu đã chuẩn bị (chi tiết trong luận văn) đạt kết quả như sau:



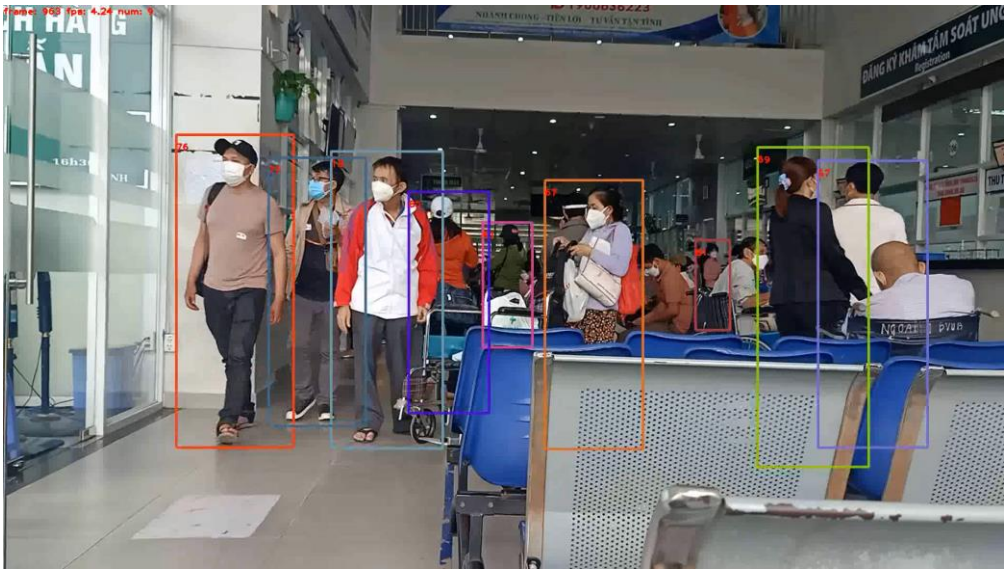
*Hình **Error! No text of specified style in document.**.3 Detect người đi bộ trên đường phố ở video nhảy múa đường phố*

*Hình **Error! No text of specified style in document.**.4 Detect người đi bộ ở khu vực Thánh thất Tây Ninh*

*Hình **Error! No text of specified style in document.**.5 Detect người đi bộ trước cửa bệnh viện Ung Bướu*



*Hình **Error! No text of specified style in document.**.6 Detect người đi bộ khu vực khám bệnh của bệnh viện*



*Hình **Error! No text of specified style in document.**.7 Detect người đi bộ khu khám bệnh của bệnh viện*

*Hình **Error! No text of specified style in document.**.8 Detect người đi bộ khu vực mua sắm ở siêu thị*

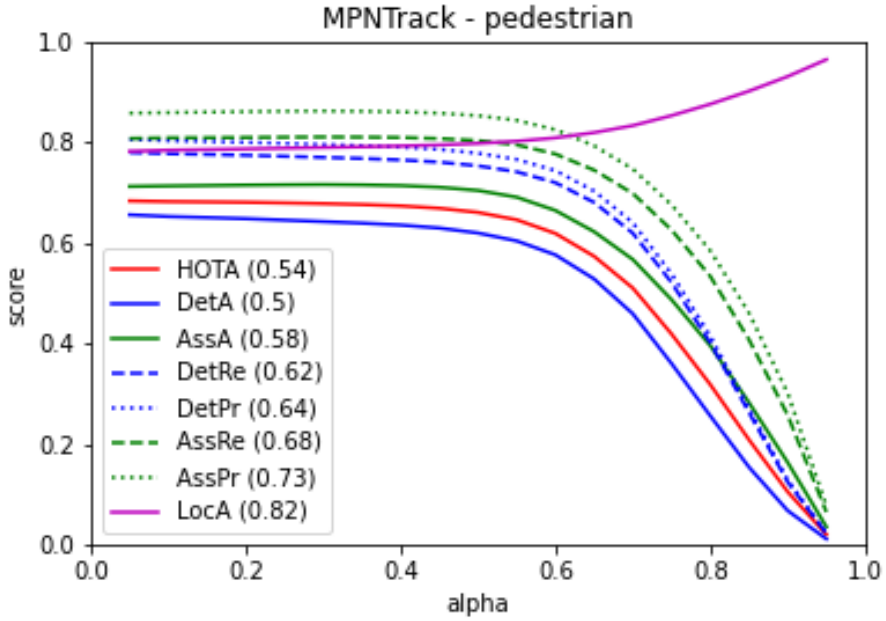
4.2.1. Xây dựng ground true

Với bộ data tracker thu được ở bước phía trên, từ các frame được xuất từ video, tiến hành đánh nhãn cho các frame này. Từ đó, ta thu được bộ ground true.

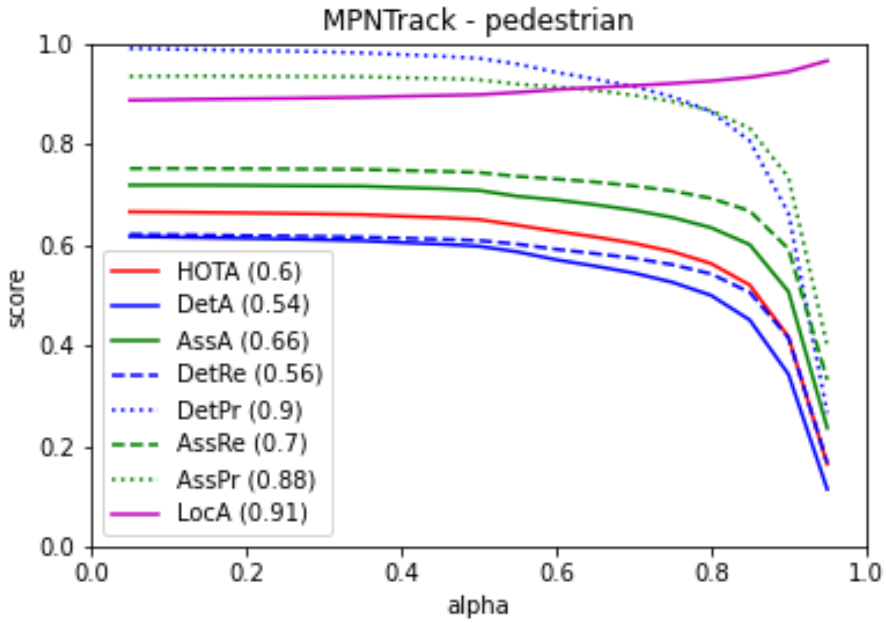
4.3. Đánh giá và so sánh các bộ dữ liệu với TrackEval

Để tiến hành so sánh và đánh giá các bộ dữ liệu, ta cũng cần thực thi tập lệnh với Google Colaboratory. Chi tiết tập lệnh được trình bày cụ thể trong luận văn. Sau

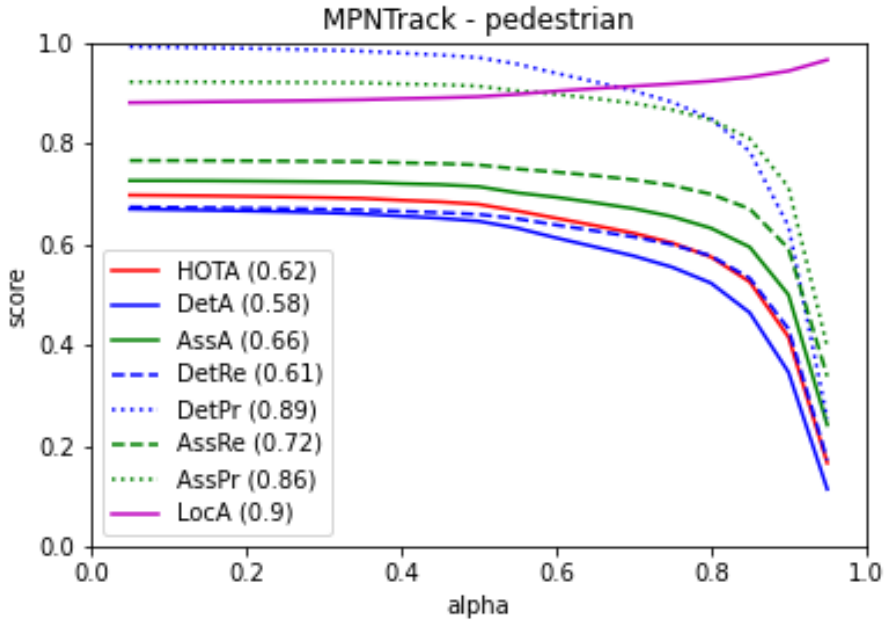
đó tiến hành chạy đánh giá lần lượt với MOT15, MOT16, MOT17, MOT20 và MOT25 ta thu được kết quả như sau:



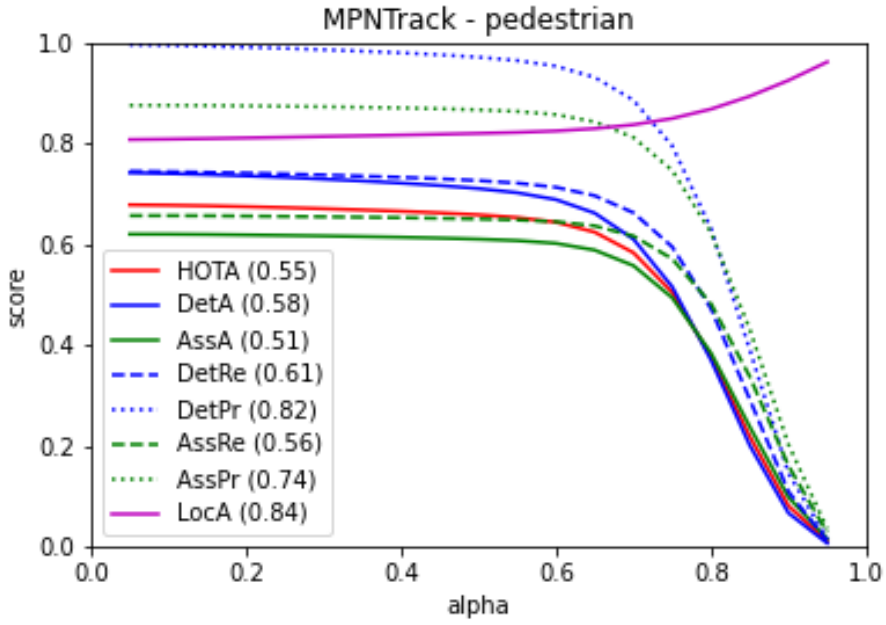
*Hình **Error! No text of specified style in document.**.9 Kết quả chạy TrackEval của bộ MOT15*



*Hình **Error! No text of specified style in document.**.10 Kết quả chạy TrackEval của bộ MOT16*



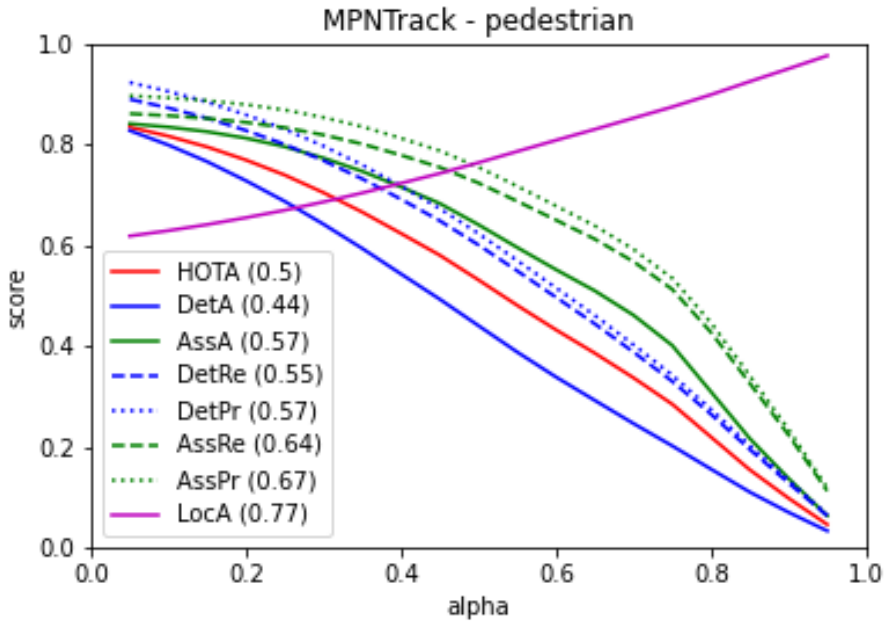
*Hình **Error! No text of specified style in document.**.11 Kết quả chạy TrackEval của bộ MOT17*



*Hình **Error! No text of specified style in document.**.12 Kết quả chạy TrackEval của bộ MOT20*

Từ biểu đồ của các bộ MOT 16 đến 20 trên ta thấy các thông đố, độ đo (metrics) đều trên 50%. Trong đó, tất cả các bộ dữ liệu đều đạt cao nhất là LocA dao động từ 82% đến 91%. Đối với MOT15,16,17 thấp nhất là chỉ số DetA với giá trị lần

lượt là 50%, 54%, 58%, MOT20 là chỉ số AssA với 51%. Các chỉ số còn lại khá tương đồng dao động từ 55% đến 89% với tất cả các bộ MOT.



*Hình **Error! No text of specified style in document.**.13 Kết quả chạy TrackEval của bộ MOT25*

Từ biểu đồ của bộ MOT25 ta thấy chỉ số LocA vẫn đạt giá trị cao với 77%, giá trị DetA thấp nhất với 44%, các chỉ số còn lại dao động từ 50% đến 67%. Vì lý do ở Việt nam ít người đi bộ, thường di chuyển trên xe máy và tình trạng dịch bệnh COVID đang diễn biến phức tạp nên chưa thể quay được các video với mật độ người đi bộ cao. Cùng với đó bộ dữ liệu huấn luyện của mô hình chưa có huấn luyện các hình ảnh này (ngồi xe máy, đi xe máy,...) cho nên kết quả phát hiện đối tượng thấp hơn rất nhiều so với bộ dữ liệu công bố trên trang MOT Benchmark.

Kết quả chi tiết của các biểu đồ được mô tả trong bảng sau:

*Bảng **Error! No text of specified style in document.**.2 Kết quả các chỉ số đánh giá của bộ data MOT25*

Video	HOTA	MOTA	IDF1	MT	ML	IDs
MOT25-01	72.3%	49.5%	47.4%	84	72	371
MOT25-03	86.0%	83.2%	91.0%	9	10	13
MOT25-05	71.9%	15.7%	31.9%	8	12	34
MOT25-07	80.8%	21.6%	59.1%	34	20	115
MOT25-09	91.6%	32.9%	64.5%	33	11	145

MOT25-11	92.9%	48.5%	73.1%	11	3	39
MOT25-13	92.5%	51.2%	74.7%	70	21	210
MOT25-15	95.4%	53.9%	76.6%	11	3	40

*Bảng **Error! No text of specified style in document.**.3 Kết quả tổng hợp các chỉ số đánh giá của các bộ data*

Dataset	HOTA	MOTA	IDF1	MT	ML	IDs
MOT15	68.3%	53.9%	68.3%	237	138	413
MOT16	66.5%	59%	68.4%	160	131	404
MOT17	69.7%	64.4%	71.2%	649	360	1344
MOT20	67.7%	72.1%	67.8%	1070	236	2674
MOT25*	83.3%	23.4%	59.9%	252	142	967

Sau khi chạy thử chatbot với 3 trường hợp nêu trên, với các từ khóa hợp lý, mô hình đã đưa ra dự đoán và trả lời chính xác câu hỏi từ phía người dùng. Nhưng vẫn không loại trừ khả năng mô hình có thể đưa ra dự đoán sai với những từ khóa chưa có trong bộ dữ liệu. Tuy nhiên, với kết quả này, nhận thấy mô hình đã có hoạt động hiệu quả và chính xác, có thể ứng dụng vào thực tế.

4.5. Nhận xét

Nhìn chung tất cả các bộ dữ liệu đều nhận diện được đều trên 50%, tuy nhiên bộ dữ liệu ở Việt Nam còn khá yếu kém so với các bộ dữ liệu còn lại vì các lý do được liệt kê ở trên. Từ đó cũng nhận ra phương pháp luận văn xây dựng so sánh với bộ dữ liệu có sẵn thì kết quả khá ổn định, còn bộ dữ liệu tự xây dựng thì chủ quan. Chỉ số MOTA thấp cho thấy độ chính xác của hệ thống khi detech trên bộ dữ liệu MOT25 kém so với các bộ dữ liệu khác.

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết quả nghiên cứu của đề tài

Với đề tài luận văn này, tác giả đã tập trung nghiên cứu các cách thức và ứng dụng học sâu để tiến hành theo vết đa đối tượng trên bộ dữ liệu của mình, cụ thể là xây dựng được bộ data tracker và ground true có tên MOT25 gồm tổng cộng 8 video với độ dài từ 20 đến 90 giây trên mỗi video tập trung ở khu vực công cộng Thành phố Hồ Chí Minh và Tây Ninh. Sau đó, chạy TrackEval để đánh giá với kết quả thu được trên bộ chỉ số đánh giá HOTA, MOTA, IDF1, MT, ML , IDs lần lượt là 83.3%, 23.4%, 59.9%, 252, 142, 967. Tuy nhiên với kết quả này thì luận văn vẫn còn có thể cải thiện thêm để có thể đạt được hiệu năng tốt hơn trên bộ data tốt hơn nữa

5.2. Đề xuất phương pháp và thuật toán xử lý

Trong quá trình thực hiện bài luận văn cũng không tránh khỏi thiếu sót:

- Bộ dữ liệu chưa tối ưu: số lượng người đi bộ trong mỗi video khá ít, chưa quay được video với tần số người di chuyển cao.
- Mô hình (với bộ dữ liệu training được sử dụng) chưa có nhiều tư thế và hành động của người như tư thế ngồi, ngồi xe máy, lái xe, cúi nhặt đồ,.. vốn phổ biến ở Việt Nam, chỉ phát hiện được số ít người với tư thế này.

5.3. Hướng phát triển của đề tài

Vấn đề kiến nghị và hướng đi tiếp theo của nghiên cứu:

- Xây dựng thêm nhiều bộ data mới với số lượng người nhiều hơn, khu vực quay video đa dạng hơn.
- Sử dụng thêm nhiều mô hình khác hoặc tự xây dựng một mô hình mới để thực hiện việc tracker phong phú hơn.