

ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC BÁCH KHOA

VÕ HOÀNG AN

**GÁN NHÃN ĐỐI TƯỢNG DI CHUYỂN
QUA NHIỀU CAMERA
(TO ASSIGN LABEL FOR MOVING OBJECTS
IN MULTIPLE CAMERAS)**

Chuyên ngành : KHOA HỌC MÁY TÍNH
Mã số: 60480101

LUẬN VĂN THẠC SĨ

TP. HỒ CHÍ MINH, tháng 6 năm 2018

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI
TRƯỜNG ĐẠI HỌC BÁCH KHOA –ĐHQG -HCM**

Cán bộ hướng dẫn khoa học : PGS TS. NGUYỄN THANH BÌNH

Cán bộ chấm nhận xét 1 :

Cán bộ chấm nhận xét 2 :

Luận văn thạc sĩ được bảo vệ tại Trường Đại học Bách Khoa, ĐHQG Tp. HCM
ngày tháng năm

Thành phần Hội đồng đánh giá luận văn thạc sĩ gồm:

(Ghi rõ họ, tên, học hàm, học vị của Hội đồng chấm bảo vệ luận văn thạc sĩ)

1.
2.
3.
4.
5.

Xác nhận của Chủ tịch Hội đồng đánh giá LV và Trưởng Khoa quản lý chuyên ngành sau khi luận văn đã được sửa chữa (nếu có).

CHỦ TỊCH HỘI ĐỒNG

TRƯỞNG KHOA

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: VÕ HOÀNG AN..... MSHV:1670211

Ngày, tháng, năm sinh: 15/06/1993..... Nơi sinh: Bình Định

Chuyên ngành: KHOA HỌC MÁY TÍNH..... Mã số : 60480101.....

I. TÊN ĐỀ TÀI: Gán nhãn đối tượng di chuyển qua nhiều camera (To assign label for moving objects in multiple cameras)

II. NHIỆM VỤ VÀ NỘI DUNG:

- Tìm hiểu các công trình nghiên cứu liên quan đến gán nhãn đối tượng.
- Đề xuất phương pháp gán nhãn cho đối tượng di chuyển qua nhiều camera.
- Hiện thực theo phương pháp đề xuất để đánh giá kết quả đạt được.

III. NGÀY GIAO NHIỆM VỤ : (Ghi theo trong QĐ giao đề tài).....

IV. NGÀY HOÀN THÀNH NHIỆM VỤ: (Ghi theo trong QĐ giao đề tài).....

V. CÁN BỘ HƯỚNG DẪN: PGS TS. NGUYỄN THANH BÌNH

Tp. HCM, ngày tháng năm 20....

CÁN BỘ HƯỚNG DẪN
(Họ tên và chữ ký)

CHỦ NHIỆM BỘ MÔN ĐÀO TẠO
(Họ tên và chữ ký)

TRƯỞNG KHOA.....

(Họ tên và chữ ký)

LỜI CẢM ƠN

Đầu tiên, tôi xin gửi lời cảm ơn sâu sắc đến **PGS TS. Nguyễn Thanh Bình**, thầy đã tận tình hướng dẫn, cung cấp tài liệu cũng như động viên, khích lệ, giúp đỡ và cho tôi những ý kiến đóng góp hết sức quý báu trong suốt thời gian tôi thực hiện đề tài này.

Tôi cũng xin gửi lời cảm ơn chân thành đến quý Thầy Cô đang công tác tại Khoa Khoa học và Kỹ thuật Máy Tính, trường Đại học Bách Khoa TP. Hồ Chí Minh, những người đã nhiệt tình truyền đạt kiến thức, kinh nghiệm trong suốt hai năm qua để tôi có những nền tảng vững chắc.

Cuối cùng, tôi xin gửi lời cảm ơn tới gia đình và bạn bè đã động viên, giúp đỡ tôi rất nhiều trong quá trình thực hiện đề tài này.

Một lần nữa, xin chân thành cảm ơn tất cả mọi người.

Thành phố Hồ Chí Minh, ngày 30 tháng 05 năm 2018

Võ Hoàng An

TÓM TẮT LUẬN VĂN THẠC SĨ

Trong những năm trở lại đây, lĩnh vực thị giác máy tính phát triển nhanh và dần chiếm một vị trí quan trọng trong sự phát triển của không chỉ ngành khoa học máy tính mà còn của các ngành kinh tế. Chính vì điều này, ngày càng có nhiều nghiên cứu về thị giác máy tính đặc biệt là nghiên cứu về truy vết đối tượng.

Trong luận văn này, tôi sẽ trình bày phương pháp để gán nhãn cho các đối tượng di chuyển qua nhiều camera. Để xây dựng được một phương pháp có thể gán nhãn được chính xác một đối tượng xuất hiện trong hệ thống camera có vùng không gian trùng lặp, trước tiên, tôi dùng phương pháp phát hiện đối tượng YOLO để nhận dạng các đối tượng và xác định vị trí của chúng trong camera, từ đó rút trích được các đặc trưng về màu sắc, hình dáng, SIFT...của chúng. Sau đó, sử dụng giải thuật GSA với giá trị đầu vào là các đặc trưng rút trích được để gán nhãn cho các đối tượng di chuyển trên những frame ảnh liên tiếp nhau. Bên cạnh đó, với mỗi đối tượng di chuyển, tôi cũng sẽ sử dụng kết hợp thêm một bộ lọc Kalman Filter để nâng cao tính chính xác của việc gán nhãn đồng thời giải quyết bài toán che phủ đối tượng. Cuối cùng, để có thể gán nhãn cho các đối tượng di chuyển qua nhiều camera, tôi đề xuất phương pháp xác định vùng không gian trùng lặp giữa các camera đó, từ đó dựa trên vị trí của các đối tượng trong vùng không gian trùng lặp, tôi sẽ gán nhãn cho chúng một cách nhất quán.

LỜI CAM KẾT

Tôi xin cam đoan rằng, đề tài luận văn thạc sĩ “Gán nhãn cho đối tượng di chuyển qua nhiều camera” là công trình nghiên cứu của tôi dưới sự hướng dẫn của **PGS TS. Nguyễn Thanh Bình**, xuất phát từ yêu cầu thực tiễn của đề tài và nguyện vọng tìm tòi, khám phá của bản thân tôi.

Những tài liệu tham khảo trong đề tài được trích dẫn hết sức rõ ràng, đúng theo quy tắc khoa học. Kết quả của đề tài chưa từng được công bố trước đây dưới bất cứ hình thức nào.

Thành phố Hồ Chí Minh, ngày 30 tháng 05 năm 2018

Tác giả luận văn

Võ Hoàng An

MỤC LỤC

MỤC LỤC	
Bảng ký tự viết tắt	
CHƯƠNG 1. GIỚI THIỆU	1
1.1. GIỚI THIỆU ĐỀ TÀI	1
1.2. MỤC TIÊU VÀ NỘI DUNG ĐỀ TÀI	2
1.2.1. Mục tiêu đề tài	2
1.2.2. Nội dung đề tài	3
1.3. GIỚI HẠN ĐỀ TÀI.....	3
1.4. ĐÓNG GÓP CỦA ĐỀ TÀI	4
1.4.1. Đóng góp về mặt khoa học.....	4
1.4.2. Đóng góp về mặt thực tiễn	4
1.5. PHƯƠNG PHÁP NGHIÊN CỨU	5
1.6. CẤU TRÚC LUẬN VĂN	5
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN	7
2.1 CƠ SỞ LÝ THUYẾT	7
2.1.1 Cắt frame từ đoạn video	7
2.1.2 Xác định vùng không gian trùng lặp	7
2.1.3 Phát hiện đối tượng di chuyển.....	8
2.1.4 Rút trích đặc trưng của đối tượng.....	9
2.1.5 Gán nhãn đối tượng trên từng camera.....	10
2.2 CÁC NGHIÊN CỨU LIÊN QUAN.....	10
CHƯƠNG 3. GÁN NHÃN ĐỐI TƯỢNG DI CHUYỂN QUA NHIỀU CAMERA	20
3.1. MÔ TẢ BÀI TOÁN.....	20
3.2. PHƯƠNG PHÁP ĐỀ XUẤT.....	20
3.2.1. Chuẩn bị dữ liệu	22
3.2.2. Xác định vùng không gian trùng lặp	22
3.2.3. Phát hiện đối tượng	24
3.2.4. Rút trích đặc trưng đối tượng	24
3.2.5. Gán nhãn cho đối tượng trên từng camera	27
3.2.6. Gán nhãn cho đối tượng xuất hiện giữa các camera	29
3.2.7. Dữ liệu đầu ra.....	31

3.3. PHƯƠNG PHÁP ĐÁNH GIÁ	31
CHƯƠNG 4. THÍ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ	32
4.1. TẬP DỮ LIỆU ĐÁNH GIÁ.....	32
4.2. KẾT QUẢ THÍ NGHIỆM	32
CHƯƠNG 5. KẾT LUẬN	37
5.1. KẾT QUẢ ĐẠT ĐƯỢC.....	37
5.2. ƯU ĐIỂM VÀ NHƯỢC ĐIỂM PHƯƠNG PHÁP GÁN NHÃN.....	37
5.2.1. Ưu điểm.....	37
5.2.2. Nhược điểm.....	38
5.3. HƯỚNG MỞ RỘNG.....	38
TÀI LIỆU THAM KHẢO	39

Bảng ký tự viết tắt

RGB : red-green-blue

HSV: Hue, Saturation, Value

IoU: Intersection over Union

GSA: Gale Shaply Algorithm

MTMC: Multiple Tracker Multiple Camera.

FOV: Field Of View

SIFT: the Scale Invariant Feature Transform.

CNN: Convolutional Neural Network.

R-CNN: Region-based Convolutional Neural Network.

VOC: Virtual Object Classes.

CL: Convolutional Layer.

FCL: Full Convolutional Layer.

HAAR, SIFT, HOG.

HIM: Hu's Invariant Moments.

LIDAR: Laser Imaging, Detection And Ranging.

BING: BInarized Normed Gradients

DBSCAN: Density-Based Spatial Clustering of Applications with Noise.

SVM: Support Vector Machine.

YOLO: You Only Look Once.

DPM: Deformable Part Model.

FPNN: Filter Pairing Neural Network.

CHƯƠNG 1. GIỚI THIỆU

1.1. GIỚI THIỆU ĐỀ TÀI

Ngày nay, với sự phát triển ồ ạt của nền kinh tế toàn cầu và xu thế dân số già đối với các nước phát triển đang là một dấu hỏi lớn chưa có lời giải đối với bài toán nhân lực. Dân số già dẫn đến nguồn nhân lực trẻ, năng động đang dần hiếm hoi khiến cho việc thiếu lao động con người trong các ngành công nghiệp và dịch vụ rất phổ biến. Điều này thúc đẩy cho việc làm sao có thể đưa máy tính vào thay thế người lao động, chính vì thế mà ngày càng có rất nhiều lĩnh vực trong đời sống cần đến ứng dụng của thị giác máy tính.

Trong lĩnh vực quảng cáo, để người chủ quảng cáo biết được số lượng khách hàng chú ý tới quảng cáo của họ, cách thông thường là họ sẽ phải thuê người theo dõi và tính số người xem quảng cáo đó của họ. Tuy nhiên, với sự giúp đỡ của thị giác máy tính, người quảng cáo có thể đặt một hệ thống camera trước quảng cáo và tiến hành theo dõi lượng người ngược mặt lên nhìn vào màn hình quảng cáo. Điều này giúp cho người quảng cáo có thể tính được số lượt xem đối với quảng cáo đó là bao nhiêu mà không bị giới hạn về thời, bởi vì hệ thống camera có thể hoạt động xuyên suốt trong thời gian dài và đưa ra kết quả thống kê liên tục. Từ kết quả này, người quảng cáo có thể thấy được mức độ cuốn hút của quảng cáo đối với người xem và đưa ra chiến lược và ý tưởng quảng cáo thu hút người theo dõi.

Trong lĩnh vực an ninh, các camera chống trộm từ các điểm đặt camera có thể thu được hình ảnh của những kẻ trộm trước, trong và sau khi thực hiện hành vi trái pháp luật. Một hệ thống camera có thể xác định được đối tượng di chuyển trong video ghi hình và thu được đặc trưng, thông tin của đối tượng đó, thậm chí trước khi thực hiện hành vi trái pháp luật, một hệ thống camera thông minh có thể phát hiện được kẻ khả nghi có khả năng thực hiện hành vi trái pháp luật để đưa ra cảnh báo. Từ đó, thay vì dùng mắt thường để xác định đối tượng có hành vi sai trái, ta có thể dùng hệ thống thị giác máy tính để

đưa ra thông tin của đối tượng, giúp người điều tra có thể truy vết đối tượng một cách dễ dàng.

Trong lĩnh vực dịch vụ, thị giác máy tính có thể áp dụng một cách hiệu quả cho việc tìm kiếm người đi lạc. Khi một gia đình vào trong nhà ga, trường hợp nhà ga đông thì rất có thể một thành viên nào đó bị lạc khỏi mọi người, đặc biệt là trẻ nhỏ. Phương pháp thông thường có thể dùng loa để thông báo đến người lạc đường và cử người đi tìm kiếm. Tuy nhiên, đối với tình huống một trẻ nhỏ bị lạc và đang bấn loạn, sợ hãi thì phương pháp này đôi khi không cho kết quả khả quan. Áp dụng thị giác máy tính chúng ta chỉ cần xác định được đối tượng thất lạc dựa trên camera ghi nhận thời gian họ vào nhà ga, sau đó hệ thống sẽ tự động theo dõi đối tượng này di chuyển từ camera này đến camera khác, đến khi xác định được vị trí người đó đứng cuối cùng, từ đó có thể tìm lại được người thất lạc.

Chính từ việc nhận thấy được những ứng dụng quan trọng đó của thị giác máy tính đối với các lĩnh vực xung quanh mà tôi đã chọn thực hiện đề tài “*gán nhãn đối tượng di chuyển qua nhiều camera*” này.

1.2. MỤC TIÊU VÀ NỘI DUNG ĐỀ TÀI

1.2.1. Mục tiêu đề tài

Chúng ta có thể thấy được tầm quan trọng của thị giác máy tính trong việc thay thế mắt người to lớn như thế nào. Từ việc giám sát, thống kê đến việc vận hành tự động của máy móc, thiết bị, dây chuyền sản xuất, các thiết bị tự động ... Chính vì thế mà ngày càng có nhiều công trình nghiên cứu nhằm cải thiện và phát triển thị giác máy tính. Với đề tài này, tôi đặt mục tiêu xây dựng một hệ thống có thể phát hiện đối tượng di chuyển thông qua hai camera quan sát. Một hệ thống như vậy rất có ích cho việc giám sát và truy vết đối tượng, có thể được áp dụng phổ biến trong các hệ thống giám sát sân bay, nhà ga giúp cho việc tìm người bị thất lạc nhanh hơn, ít tổn nhân lực hơn và không gây hoang mang cho những người khác, hay cũng có thể áp dụng cho việc giữ an ninh đối với các cơ quan, tổ chức và các nơi công cộng. Khi có bất kì đối tượng trộm cắp, cướp bóc hay thậm chí là đối tượng tình nghi có khả năng gây ra một hành động phạm tội nào đó, bằng việc phân tích thêm hành vi của đối tượng, ta cũng có thể phát hiện sớm và đưa ra cảnh báo cho các cá nhân, tổ chức có liên quan

đến việc đảm bảo an ninh của khu vực đó. Mục tiêu của đề tài này vẫn chỉ ở mức làm sao có thể phát hiện đối tượng di chuyển qua nhiều camera nhằm truy vết đối tượng.

1.2.2. Nội dung đề tài

Để đạt được mục tiêu trên, tôi sẽ thực hiện các công việc sau:

- (i) Tìm hiểu các công trình nghiên cứu liên quan để có cái nhìn tổng quát và các kiến thức cơ bản đối với lĩnh vực thị giác máy tính nói chung và đề tài mà mình đang thực hiện nói riêng. Cũng như tìm hiểu được nhược điểm cần khắc phục và thế mạnh của từng nghiên cứu trước để đề ra hướng xây dựng và phương pháp đề xuất của mình.
- (ii) Đề xuất ra phương pháp gán nhãn cho đối tượng di chuyển qua nhiều camera.
- (iii) Hiện thực theo phương pháp đề xuất để từ đó đánh giá kết quả đạt được và tính chính xác của phương pháp đề xuất.

1.3. GIỚI HẠN ĐỀ TÀI

Trong vấn đề nhận dạng đối tượng di chuyển qua nhiều camera thông qua việc gán nhãn đối với đối tượng, chúng ta có rất nhiều hướng để phát triển mà mở rộng. Với đề tài này của mình, tôi muốn xoay quanh việc giải quyết mục tiêu mà mình đặt ra:

- Chỉ hiện thực trên hai camera được thiết kế theo nội dung của đề tài.
- Từ hai đoạn video thu được từ hai camera, tôi có thể gán nhãn các đối tượng di chuyển trong mỗi đoạn video, đồng thời có thể đảm bảo được tính nhất quán, chính xác của việc gán nhãn.
- Đảm bảo tính nhất quán trong việc gán nhãn đối tượng khi đối tượng đó di chuyển qua hai camera. Có nghĩa là khi một đối tượng di chuyển từ camera này sang camera khác, hệ thống mà tôi đề xuất phải đảm bảo gán cùng một nhãn cho đối tượng đó khi đối tượng xuất hiện trên cả hai camera.
- Thực hiện đo đạt kết quả của hệ thống đối với cả hai trường hợp đặt camera trùng lắp song song và trùng lắp không song song, từ đó so sánh đâu là cách đặt camera cho độ chính xác cao hơn.

1.4. ĐÓNG GÓP CỦA ĐỀ TÀI

1.4.1. Đóng góp về mặt khoa học

(i) Đề tài nhằm giải quyết bài toán gán nhãn cho đối tượng di chuyển qua nhiều camera cũng là một bài toán nhỏ trong bài toán lớn truy vết đối tượng di chuyển qua nhiều camera. Đây là một trong những đề tài hấp dẫn và thách thức đối với thị giác máy tính. Tuy chỉ dừng lại ở việc gán nhãn đối với hai camera có vùng không gian trùng lấp, chưa thể giải quyết được bài toán khi hai camera không có khoảng không gian trùng lấp, nhưng nó cũng đóng góp một phần lớn trong việc đa dạng hóa các phương pháp truy vết đối tượng di chuyển qua nhiều camera có vùng không gian trùng lấp. Từ đó, tạo cơ sở so sánh cho những phương pháp được đề xuất sau này nhằm cải thiện về tính chính xác cũng như tốc độ xử lý bài toán.

(ii) Làm cơ sở, tài liệu tham khảo cho các nghiên cứu sau này trong lĩnh vực thị giác máy tính.

1.4.2. Đóng góp về mặt thực tiễn

Việc truy vết đối tượng có một vai trò quan trọng trong thực tiễn khi mà ngày nay, thị giác máy tính đóng một vai trò quan trọng trong các lĩnh vực đời sống, đặc biệt là an ninh và dịch vụ. Tuy nhiên, với chỉ một camera ta không thể xây dựng được một vùng quan sát rộng lớn. Tầm nhìn của mỗi camera có một phạm vi hẹp, mỗi camera sẽ chỉ quan sát được một phần không gian nhỏ. Do đó, để có thể xây dựng được một vùng không gian quan sát rộng lớn, ta cần phải kết hợp nhiều camera lại với nhau. Các camera này sẽ chia sẻ vùng không gian mà nó quan sát được, tạo thành một hệ thống nhằm mô hình hóa không gian rộng lớn hơn, giải quyết được bài toán về che phủ khi truy vết đối tượng trên một camera duy nhất.

Với phương pháp gán nhãn đối tượng di chuyển qua nhiều camera này, đề tài đóng góp giải pháp truy vết đối tượng trong một vùng không gian rộng lớn. Trong môi trường thực tiễn, nó sẽ góp một phần rất quan trọng trong việc truy tìm vị trí của một người khi họ di chuyển trong hệ thống camera đã được thiết lập trước hay truy vết đối tượng vi phạm pháp luật, trộm cắp trong tòa nhà, văn phòng, chung cư...

1.5. PHƯƠNG PHÁP NGHIÊN CỨU

Trong lĩnh vực khoa học, có hai phương pháp nghiên cứu cơ bản được sử dụng để định hướng cho việc nghiên cứu của mỗi đề tài đó là nghiên cứu định tính và nghiên cứu định lượng.

Nghiên cứu định tính là phương pháp tiếp cận với mục tiêu thăm dò, mô tả và đưa ra lời giải thích phù hợp dựa trên các phương pháp khảo sát có liên quan đến đối tượng mà mình đang nghiên cứu về mặt nhận thức, kinh nghiệm, dự định, động cơ thúc đẩy, hành vi và thái độ...

Nghiên cứu định lượng có một cách tiếp cận khác, cũng tìm hiểu thông tin từ các nghiên cứu khác nhau nhưng thông tin tìm hiểu ở đây là những con số cụ thể đã được lượng hóa, đo lường nhằm phản ánh và diễn giải các mối quan hệ giữa các nhân tố với nhau.

Đối với đề tài này, tôi sử dụng phương pháp nghiên cứu định lượng. Cách tiếp cận của tôi sử dụng nguồn tài liệu từ các nghiên cứu liên quan đến thị giác máy tính nói chung và truy vết đối tượng nói riêng để có được một cách nhìn tổng quan về phương pháp mà các nghiên cứu trước đó đã và đang ứng dụng nhằm giải quyết các bài toán tương tự như tôi đề xuất. Với một lượng nghiên cứu to lớn như vậy, tôi hiểu và có thể xây dựng cho mình một cách tiếp cận mới nhằm giải quyết bài toán “*gán nhãn đối tượng di chuyển qua nhiều camera*” này. Phần quan trọng nhất của phương pháp nghiên cứu này là xây dựng được mô hình đề xuất và thống kê dữ liệu thu thập được để chứng minh mức độ hiệu quả của mô hình mà mình đề xuất.

1.6. CẤU TRÚC LUẬN VĂN

Luận văn được tổ chức thành 5 chương có cấu trúc như sau:

- **Chương 1: Giới thiệu.** Trong chương này, tôi sẽ giới thiệu sơ qua về đề tài, mục tiêu và nội dung, những giới hạn khi thực hiện đề tài, phương pháp nghiên cứu cũng như những đóng góp của đề tài về mặt khoa học và thực tiễn;
- **Chương 2: Cơ sở lý thuyết và các nghiên cứu liên quan.** Giới thiệu cơ sở lý thuyết và các nghiên cứu liên quan mà tôi đã tìm hiểu để thực hiện đề tài;
- **Chương 3: *Gán nhãn đối tượng di chuyển qua nhiều camera*.** Trong chương này, tôi sẽ mô tả về những yêu cầu của bài toán, phương pháp mà tôi đề xuất để giải

quyết các bài toán đó và phân phương pháp đánh giá để xác định được phương pháp đề xuất này hiệu quả hay không trên các ngữ cảnh mà tôi đã đặt ra trong đề tài này;

- **Chương 4: Thí nghiệm và đánh giá kết quả.** Giới thiệu về tập dữ liệu bao gồm nguồn thu dữ liệu và các thông số kỹ thuật mà tôi sử dụng đồng thời thực hiện thí nghiệm trên tập dữ liệu này để thu được kết quả của phương pháp được đề xuất.
- **Chương 5: Kết luận.** Trong chương này, dựa trên kết quả đạt được từ thí nghiệm tôi sẽ đưa ra ưu nhược điểm cũng như những nguyên nhân dẫn đến các ưu nhược điểm này, đồng thời có những bình luận về chúng để đưa ra hướng mở rộng cho phương pháp nhằm cải tiến phương pháp đề xuất để thu được kết quả tốt hơn.

CHƯƠNG 2.

CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

2.1 CƠ SỞ LÝ THUYẾT

2.1.1 Cắt frame từ đoạn video

Đối với thị giác máy tính, chúng ta không thể xử lý trực tiếp trên dữ liệu là một đoạn video được, mà chúng ta cần phải chuyển các đoạn video này sang tập các frame ảnh. Đoạn video cơ bản được tạo thành từ một tập hợp các frame ảnh liên tiếp nhau. các frame này là các ảnh chụp được tại một thời điểm cụ thể. Do đó mà việc xử lý video sẽ được ánh xạ sang việc xử lý từng frame ảnh. Việc chuyển từ video sang frame ảnh ta có thể sử dụng các chương trình hỗ trợ như matlab 7.12.0, FFMPEG, OpenCV, EmguCV...Tất cả các chương trình này được áp dụng rất phổ biến trong lĩnh vực xử lý ảnh và yêu cầu cụ thể ở đây là chuyển video sang các frame ảnh.

FFMPEG là một thư viện có nhiều tiện ích hỗ trợ cho việc xử lý video. Tính năng nổi bật nhất là khả năng encode/decode nhiều định dạng video khác nhau, giúp chuyển video qua từ định dạng này sang định dạng khác. Ngoài ra, FFMPEG còn hỗ trợ cắt đoạn video để chuyển sang dạng frame và từ frame chuyển sang các file ảnh.

OpenCV có lẽ là một bộ thư viện khá phổ biến đối với hầu hết những người làm việc trong lĩnh vực xử lý ảnh. OpenCV là một thư viện mã nguồn mở cung cấp các interface C/C++, Java, Python và hỗ trợ cho Windows, Linux, Mac OS, iOS và cả Android. Các đặc trưng nổi bật có thể kể đến rút trích đặc trưng thông qua các giải thuật như PCA..., phát hiện đối tượng như khuôn mặt, người, xe hơi, xử lý đoạn video và chuyển sang các frame...

EmguCV cơ bản là một OpenCV nhưng nó được tạo ra để hỗ trợ phát triển trên ngôn ngữ C#, vì vậy nên nó có đầy đủ các tính năng nổi bật của OpenCV.

2.1.2 Xác định vùng không gian trùng lặp

Trong bài toán truy vết đối tượng qua nhiều camera có vùng không gian trùng lặp, việc xác định vùng không gian trùng lặp là một bước hết sức quan trọng và cần thiết. Vùng không gian mà cả hai camera này quan sát được là một trong những yếu tố

giúp xác định được vị trí của đối tượng trong không gian thực tế, nhằm đảm bảo quá trình gán nhãn cho đối tượng được nhất quán. Phương pháp xác định vùng không gian trùng lắp có thể chia ra làm hai loại dựa vào vị trí đặt của camera.

Trường hợp hai camera được đặt song song để quan sát cùng một hướng như camera giao thông quan sát được đặt hai bên đường hoặc tổng quát hơn là camera được đặt để quan sát địa hình theo chiều dài...Ta có thể rút trích các điểm đặc biệt để tìm sự tương đồng giữa hai frame ảnh thu được từ hai camera. Từ các điểm tương đồng đó, ta sẽ xác định được vùng không gian trùng lắp giữa chúng. Trường hợp hai camera được đặt không song song, ta không thể sử dụng phương pháp rút trích điểm tương đồng được mà thay vào đó, ta phải thực nghiệm đo đạc để tìm vùng không gian trùng lắp thích hợp. Phương pháp sẽ được trình bày rõ hơn trong phần 3.2.2.

2.1.3 Phát hiện đối tượng di chuyển

Có rất nhiều phương pháp đã được đưa ra để phát hiện đối tượng di chuyển trong một đoạn video. Ta có thể chia ra thành các lớp: Point detectors, background subtraction, segmentation, supervised learning. Đơn giản nhất và cổ điển nhất là background subtraction. Tuy nhiên, phương pháp đem lại hiệu quả cao nhất và có tính khoa học lại phải kể đến phương pháp supervised learning.

Phương pháp supervised learning gọi theo tiếng việt là học có giám sát, đây là một phương pháp khá là phổ biến trong thời gian gần đây và có khuynh hướng sẽ trở thành phương pháp chính hỗ trợ không chỉ trong việc phát hiện đối tượng mà còn cả trong các mặt ứng dụng khác của thị giác máy tính nói riêng và trí tuệ nhân tạo nói chung. Phương pháp này mang tính khoa học ở chỗ nó mô phỏng quá trình học của con người để áp dụng cho máy tính. Con người thông qua quá trình sống, học tập và làm việc mà ghi nhận được những kiến thức khoa học thì ở đây máy tính cũng được học từ những kiến thức khoa học mà con người thu nhận và truyền tải cho nó để biến nó trở thành tri thức mà máy có thể hiểu và thực thi.

Một cách cụ thể hơn bằng ngôn ngữ tự nhiên, con người học được cách nhận dạng một chiếc xe máy bằng việc tiếp nhận vô số các hình ảnh, âm thanh cũng như các tính chất của xe máy để từ đó lọc ra được những đặc trưng mà một chiếc xe máy có thể có và như thế ta học được cách nhận dạng đối tượng nào là xe máy. Còn đối với máy tính, ta cũng cung cấp tập dữ liệu gọi là tập huấn luyện. Tập dữ liệu này cũng

tương tự là những hình ảnh, âm thanh, các tính chất của một chiếc xe máy. Với tập dữ liệu này, chúng ta sẽ xây dựng phương pháp để máy tính tính toán, xây dựng mẫu đại diện cho đối tượng xe máy. Cuối cùng máy tính có thể đưa ra quyết định một đối tượng có phải là xe máy hay không dựa trên mức độ tương đồng của đối tượng đó với mẫu đã xây dựng.

2.1.4 Rút trích đặc trưng của đối tượng

Đối với các phương pháp rút trích đặc trưng của đối tượng, mục tiêu là tìm ra được các đặc trưng thể hiện sự khác biệt giữa đối tượng này với các đối tượng khác. Một trong số đó có thể kể đến như:

a) Màu

Màu là một trong những đặc trưng quan trọng nhất giúp cho việc phân biệt đối tượng. Màu rất dễ để phân tích thông qua frame ảnh cũng như ý tưởng khá đơn giản. Chất lượng của đặc trưng màu phụ thuộc lớn vào không gian màu sử dụng để biểu diễn đối tượng. Một số phương pháp rút trích đặc trưng màu của đối tượng phổ biến là moment màu, moment màu mờ, biểu đồ màu...Chính vì có nhiều phương pháp rút trích đặc trưng màu mà nó trở nên rất phổ biến trong việc nhận diện đối tượng trong ảnh.

b) Hình dáng

Hình ảnh cũng khá là quan trọng trong rút trích đặc trưng đối tượng. Khi một đối tượng quay lưng lại hoặc thay đổi trang phục thì đặc trưng màu sắc không còn có thể giúp phân biệt các đối tượng được nữa, nhưng hình dáng của đối tượng thì vẫn sẽ không thay đổi nhiều, một người béo thì không thể gầy ngay trong khi di chuyển qua camera được và ngược lại. Một đặc trưng về hình dáng tốt là một đặc trưng mà không bị ảnh hưởng bởi sự thay đổi hình dáng do quá trình di chuyển, xoay hay thu phóng kích thước của đối tượng. Phương pháp hiệu quả nhất là sử dụng moment bất biến của hình dáng. Phương pháp này sẽ giúp rút trích ra các vector đặc trưng không thay đổi dựa trên hình dáng của đối tượng.

2.1.5 Gán nhãn đối tượng trên từng camera

Gán nhãn thực chất là quá trình truy vết đối tượng từ frame ảnh này sang frame ảnh khác. Việc xác định đối tượng trong frame ảnh tiếp theo có thể dựa vào việc phân tích đặc trưng của đối tượng như hình dáng, màu sắc, vị trí của đối tượng trong frame ảnh hiện tại với frame ảnh trước.

Ta có thể gán nhãn cho đối tượng thông qua ba lớp giải thuật:

- Giải thuật phân lớp, phân cụm: sử dụng các đặc trưng thu được của đối tượng trong các frame ảnh, ta xây dựng mô hình mẫu đại diện cho đối tượng đó. Khi đó, khi một đối tượng mới phát hiện trong frame ảnh, ta so sánh với tập mẫu để xác định được đối tượng thuộc mẫu nào và gán nhãn cho mẫu đó.
- Giải thuật chuỗi thời gian: sử dụng các phép tính toán, ước lượng để từ các thông số hiện tại, phỏng đoán giá trị tương lai của đối tượng. Giá trị phỏng đoán ở đây thường là vị trí, kích thước của đối tượng.

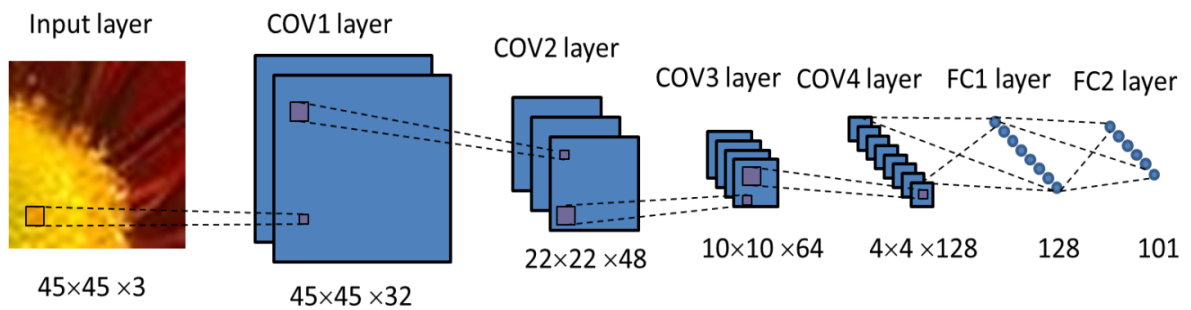
2.2 CÁC NGHIÊN CỨU LIÊN QUAN

Lĩnh vực thị giác máy tính phát triển rất nhanh và dần phổ biến mà ngày nay hầu hết các lĩnh vực đều cần đến như truy vết tội phạm, phát hiện người vi phạm giao thông, tính toán mật độ xe để điều tiết giao thông...tất cả đều xoay quanh bài toán cốt yếu nhất, đó là bài toán nhận dạng đối tượng.

Joseph [1, 2] đã đề xuất một hệ thống phục vụ cho việc phát hiện đối tượng trong frame ảnh có tên YOLO. Tác giả xây dựng hệ thống YOLO là một mạng CNN. Cũng giống các phương pháp phát hiện đối tượng khác như DPM, R-CNN, Fast R-CNN... YOLO có khả năng dự đoán được vị trí của các đối tượng đồng thời phân lớp cho các đối tượng đó dựa trên việc học các đặc trưng của đối tượng như HAAR, SIFT [3], HOG [4] từ các ảnh trong tập huấn luyện có kích thước đầy đủ. Tác giả chứng minh được rằng YOLO tính toán nhanh với khả năng xử lý 45 frame ảnh trên một giây với bản đầy đủ và 155 frame ảnh trên một giây với bản thu nhỏ. Do đó, YOLO có thể được sử dụng trong phát hiện đối tượng đối với các ứng dụng đòi hỏi tính toán nhanh và đáp ứng thời gian thực. Tuy nhiên, YOLO vẫn chưa thể đáp ứng được độ chính xác cao khi so sánh với các phương pháp phát hiện đối tượng hiện đại như Fast R-CNN...

Một cách tiếp cận để giải quyết bài toán phát hiện đối tượng là phát hiện viền. Phát hiện viền của đối tượng được coi là một phương thức cơ bản trong phân mảng, nhận

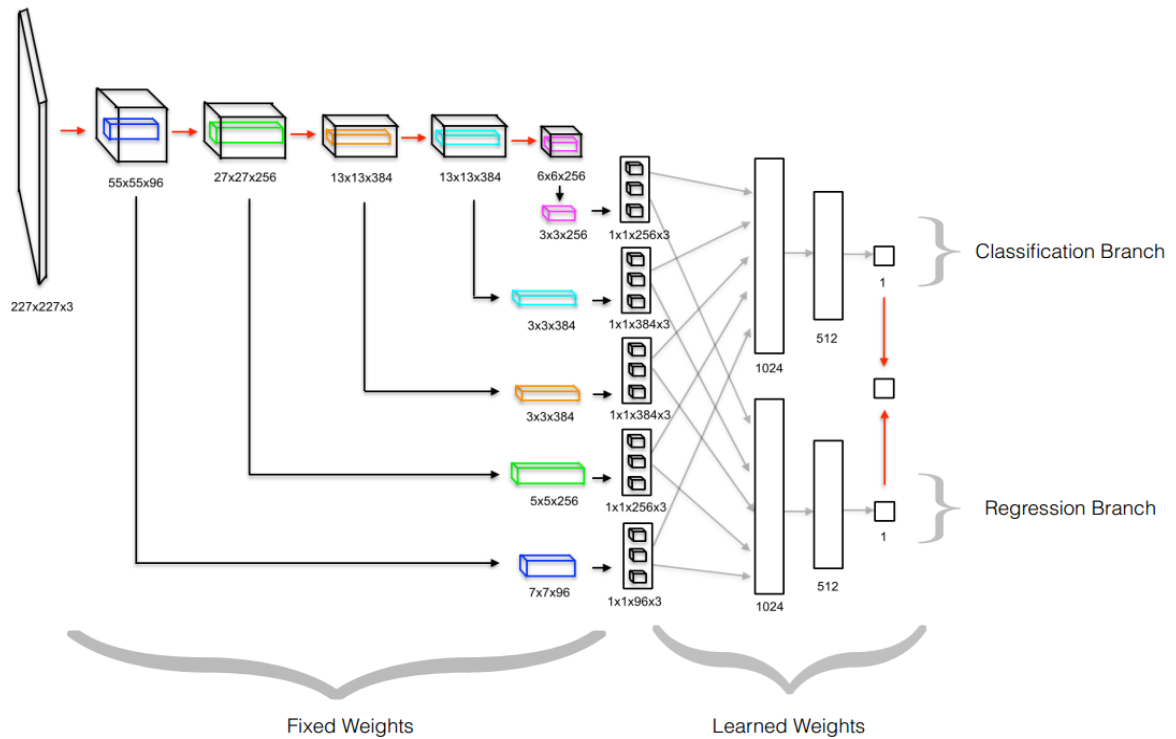
dạng ảnh ảnh và hệ thống phát hiện đối tượng. Phát hiện viền thường sử dụng các đặc trưng như SIFT và HOG của từng pixel trong ảnh để xác định pixel ảnh đang xét có thuộc đường viền hay không. Cách tiếp cận này được sử dụng khá rộng rãi và hỗ trợ các giải thuật hàng đầu trong bài toán phát hiện viền của đối tượng. Tuy nhiên, không thể phủ nhận rằng với cách tiếp cận trên giải thuật phân lớp không đem lại tính tách biệt cao giữa pixel ảnh thuộc và không thuộc viền. Chính vì vậy mà nhiều nhà khoa học đã sử dụng đặc trưng học sâu (deep features) để giải quyết vấn đề phân lớp giữa pixel thuộc và không thuộc viền. Gedas [5] và Wei [6] đề xuất xây dựng mạng nơ ron để rút trích đặc trưng học sâu. Wei [6] xây dựng mạng nơ ron với sáu lớp, bốn lớp đầu là lớp CL và hai lớp cuối là lớp liên kết đầy đủ FCL. Giá trị đầu vào của mạng CNN mà tác giả Wei đề xuất là một ảnh trong không gian màu RGB với mỗi mảng được chia nhỏ kích thước 45×45 và giá trị đầu ra là một vector 128 chiều được coi như là đặc trưng học sâu sử dụng cho các phương pháp phát hiện viền. Hình 2.1 mô tả kiến trúc mạng nơ ron mà Wei đề xuất:



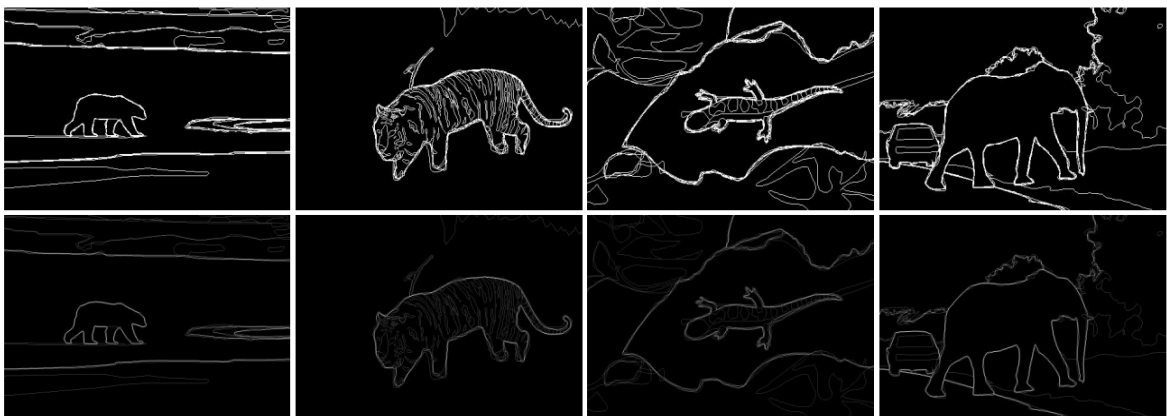
Hình 2.1: Cấu trúc CNN mà Wei đề xuất [6].

Gedas [5] lại đề xuất một cách tiếp cận khác cũng dựa vào việc xây dựng CNN. Giá trị đầu vào của cách tiếp cận mà tác giả đề xuất là ảnh cần phát hiện viền. Sử dụng phương pháp phát hiện cạnh của Canny (Canny edge detector) để chọn ra các điểm có khả năng nằm trên cạnh của đối tượng và rút trích ra mảng tương ứng với mỗi điểm được chọn nằm ở trung tâm của mảng đó. Tập các mảng thu được chuyển sang kích thước $227 \times 227 \times 3$ để đưa vào mạng KNet [7] rút trích đặc trưng là các mảng có chứa các điểm có khả năng thuộc viền của đối tượng. Các đặc trưng này sau đó được đưa vào mạng con phân nhánh với hai nhánh và mỗi nhánh gồm hai lớp liên kết đầy đủ. Nhánh thứ nhất được huấn luyện để thực hiện phân lớp viền và nhánh thứ hai được huấn luyện để học được sự khác nhau giữa các viền được nhận dạng bởi các phần khác nhau. Hình 2.2

mô tả kiến trúc của mạng CNN và hình 2.3 mô tả giá trị đầu ra của hai nhánh trong mạng CNN mà Gedas đề xuất:



Hình 2.2: Cấu trúc CNN Gedas Bertasius đề xuất [5].



Hình 2.3: Kết quả đầu ra của nhánh phân lớp (trên) và của nhánh hồi quy (dưới) [5].

Dumitru [8] đã đề xuất phương pháp phát hiện nhiều đối tượng trong một frame ảnh được gọi là “DeepMultiBox”. Đối với bài toán phát hiện đối tượng trong frame ảnh, ta có rất nhiều cách như là background subtraction và motion detection, supervisor

learning... Phương pháp đề xuất này cũng sử dụng DNNs là một trong các giải thuật thuộc lớp supervisor learning. Mục tiêu của nghiên cứu là xây dựng được phương pháp dự đoán được tập các vùng chứa đối tượng gọi là bounding box, bounding box là một hình chữ nhật bao quanh đối tượng trong không gian 2D. Dữ liệu đầu ra của phương pháp phát hiện đối tượng được đề xuất bởi Dumitru ở đây bao gồm một tập các bounding box với các điểm tọa độ thể hiện vị trí của bounding box trong frame ảnh và giá trị cho biết độ tin cậy (tính chính xác) của việc xác định nhãn của đối tượng tương ứng với bounding box đó. Đóng góp chính của nghiên cứu này là xây dựng được một mạng neuron học sâu để phát hiện được đối tượng và thu được dữ liệu đầu ra như mô tả trên.

Shipra [9] đã thực hiện một cuộc khảo sát tập trung vào bài toán truy vết đối tượng trong đoạn video quan sát. Với bài nghiên cứu đó, tác giả đã làm rõ nhiều phương thức truy vết thuộc nhiều lớp khác nhau cũng như các chiến lược nhằm giải quyết bài toán truy vết như dựa vào vùng, viền của đối tượng... Đồng thời chỉ ra được điểm tích cực và tiêu cực của các chiến lược tiếp cận đó. Bài nghiên cứu cũng giới thiệu khá tổng quan về các kiến thức tuy cơ bản nhưng lại hữu ích cho những nghiên cứu về sau tham khảo và đặc biệt là chỉ ra điểm mạnh, điểm yếu của những phương pháp được sử dụng trong truy vết, điều này rất quan trọng cho những nhà nghiên cứu mới tìm hiểu về lĩnh vực thị giác máy tính nói chung và truy vết đối tượng nói riêng.

Yan [10] đã đề xuất sử dụng đặc trưng ORB (Oriented FAST and Rotated BRIEF) để cải thiện hiệu suất của phương pháp truy vết đối tượng sử dụng Mean Shift. Giải thuật Mean Shift thông thường sử dụng đặc trưng về màu sắc của đối tượng để truy vết. Các đặc trưng màu ở đây được thu nhận từ không gian màu RGB và chuyển sang không gian màu HSV nhằm giảm bớt sự tác động từ các yếu tố ngoại như ánh sáng... Nhưng với nghiên cứu [10], tác giả sử dụng đặc trưng ORB là một sự cải tiến dựa trên phát hiện đặc trưng FAST [11] và mô tả đặc trưng BRIEF [12]. So với SIFT và SURF, ORB cải tiến hơn về tốc độ tính toán cũng như đảm bảo tính bất biến của đặc trưng trong các trường hợp các đối tượng bị thay đổi vì xoay, thu phóng hay sự chiếu sáng từ bên ngoài.

Jeong [13] đã đưa ra giải pháp giải quyết vấn đề phủ lấp giữa các đối tượng di chuyển trong camera. Rõ ràng trong thực tế, khi các đối tượng di chuyển qua lại trong camera ngẫu nhiên không theo một hướng nhất định thì việc hai đối tượng che phủ lẫn nhau trong camera là rất thường xuyên. Khi các đối tượng chồng lấp lên nhau như vậy, ta không thể sử dụng các phương pháp phát hiện đối tượng như background subtraction

và motion information, supervisor learning để xác định đối tượng bị che phủ đằng sau được mà chỉ có thể sử dụng các phương pháp ước lượng, phỏng đoán. Do đó, trong nghiên cứu [13] tác giả đã sử dụng kalman filter và đề xuất phương pháp của mình nhằm giải quyết bài toán che phủ giữa các đối tượng. Đầu tiên, tác giả sử dụng background subtraction và motion information để phát hiện nhiều đối tượng di chuyển trong camera. Sau đó, xác định được số lượng các đối tượng di chuyển trong frame. Bước thứ hai, tác giả sử dụng Kalman Filter cho mỗi đối tượng ghi nhận được. Tuy nhiên, việc sử dụng một Kalman Filter cho một đối tượng ghi nhận được sẽ dẫn đến tình trạng ở frame ảnh tiếp theo họ không thể biết chính xác được đối tượng nào sẽ tương ứng với bộ Kalman Filter nào trước đó. Chính vì thế, tác giả đề xuất giải thuật xác định đối tượng ghi nhận và bộ Kalman Filter đúng của nó sử dụng hàm chi phí bao gồm các đặc trưng cũng như là xác định được hai đối tượng che phủ hợp nhất lại với nhau hay tách rời nhau. Hình 2.4 thể hiện các bước trong phương pháp đề xuất của tác giả.

Hai bước quan trọng mà tác giả đề xuất để giải quyết được bài toán truy vết các đối tượng bị che phủ lẫn nhau là bước xác định các đối tượng che phủ đang hợp nhất lại với nhau hay đang tách ra và bước gán đối tượng phát hiện được trong frame ảnh tiếp theo đúng với bộ Kalman Filter của nó trong frame ảnh trước. Để phát hiện được các đối tượng đang hợp nhất hay tách rời nhau trong vùng che phủ, tác giả sử dụng tỉ lệ giữ chiều cao và chiều rộng của đối tượng phát hiện được so sánh với ngưỡng đề xuất. Cụ thể:

$$Hợp = \begin{cases} R_k^i > \tau_{ratioUp}, i=1,...,m \\ R_k^i < \tau_{ratioDown}, i=1,...,m \end{cases} \quad (2.1)$$

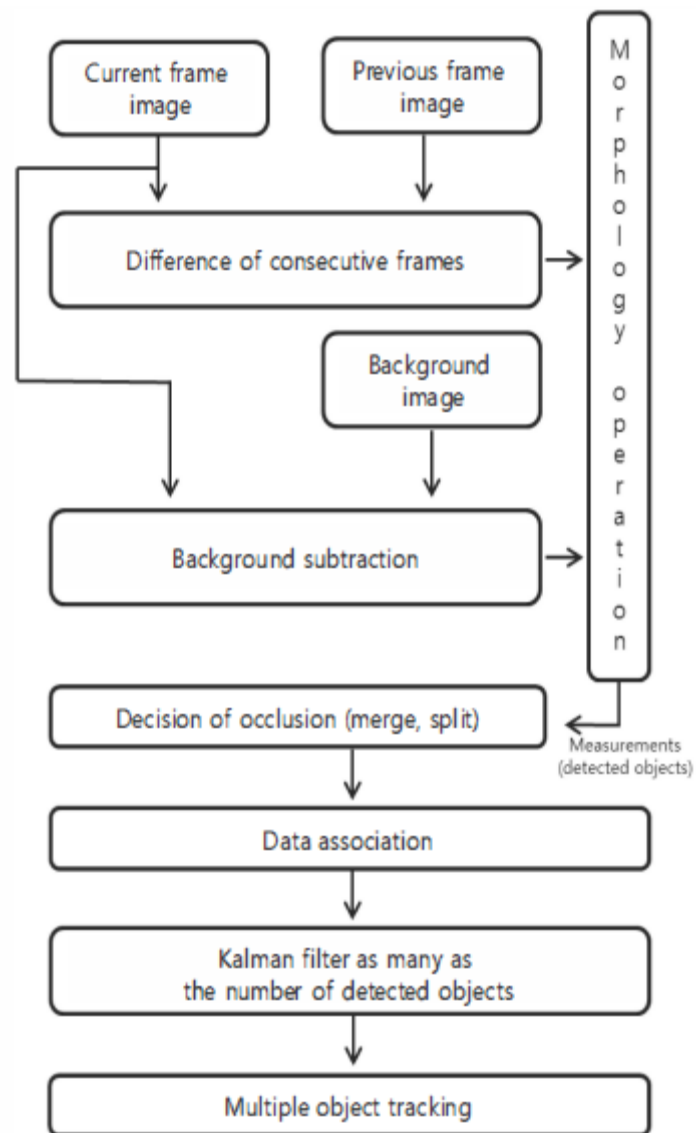
$$Tách = \tau_{ratioDown} < R_k^i < \tau_{ratioUp}$$

Với m : số lượng các đối tượng phát hiện được trong frame ảnh thứ k .

k : frame ảnh thứ k .

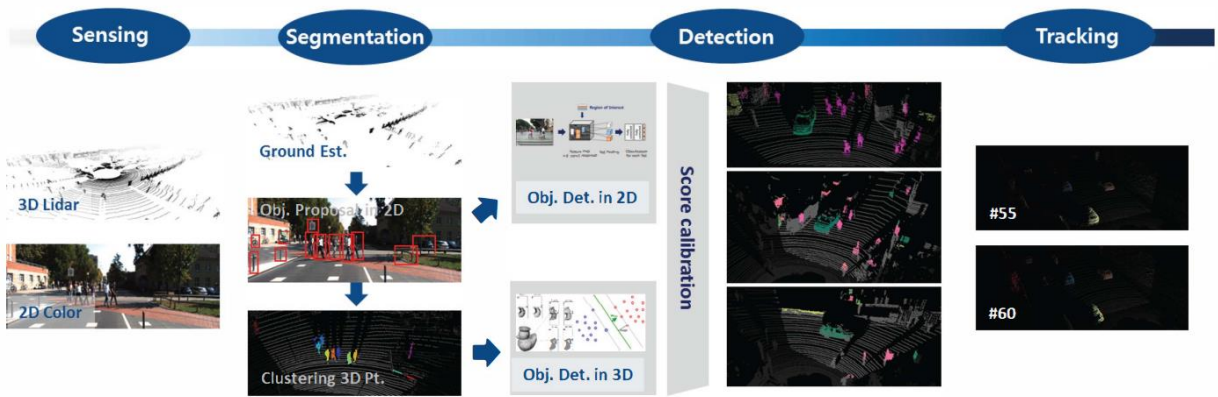
$\tau_{ratioUp}$ và $\tau_{ratioDown}$ là ngưỡng trên và ngưỡng dưới của tỉ lệ giữa chiều cao và chiều rộng của đối tượng phát hiện được trong frame ảnh.

Bước quan trọng thứ hai là bước làm thế nào để xác định đúng đối tượng ghi nhận được với bộ Kalman Filter tương ứng của nó. Để thực hiện được điều này, tác giả sử dụng hai yếu tố là yếu tố về khoảng cách giữa giá trị dự đoán với giá trị của đối tượng ghi nhận được và yếu tố về diện tích của đối tượng giữa các frame ảnh với nhau.



Hình 2.4: Sơ đồ khối phương pháp đề xuất [13].

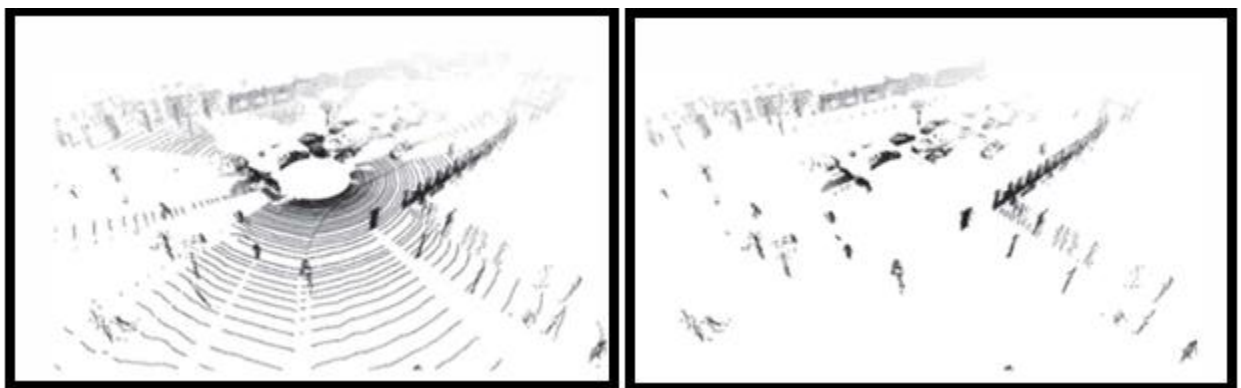
Soonmin [14] đã đề xuất một thư viện cho việc nhận diện và truy vết nhiều đối tượng dựa trên các thông tin từ camera màu và 3D LIDAR (một công nghệ quét laze). Thư viện mà tác giả đề xuất có tốc độ xử lý nhanh, phản hồi trong thời gian thực nên có thể sử dụng trong các thiết bị thông minh, rô bốt trong lĩnh vực giao thông. Hình 2.5 giới thiệu các bước trong thư viện mà tác giả đã đề xuất:



Hình 2.5: Tổng quan về các bước trong phương pháp mà Soonmin đề xuất [14].

Theo đó, thư viện đề xuất xử lý dữ liệu thông qua bốn bước.

- **Bước 1: Sensing.** Lấy dữ liệu đầu vào là các ảnh thu được từ camera màu và điểm ảnh 3D thu được từ 3D LIDAR.
- **Bước 2: Segmentation.**
 - o Phỏng đoán sơ đồ địa hình: từ các điểm ảnh 3D, sử dụng các phương pháp phỏng đoán sơ đồ địa hình để loại bỏ các điểm địa hình như hình 2.6:



(a) Input point cloud

(b) After removing ground Points

Hình 2.6: Kết quả đầu vào và đầu ra của quá trình phỏng đoán sơ đồ địa hình [14].

- o Đề xuất các đối tượng xuất hiện trong ảnh: với đầu vào là ảnh thu được của camera màu, sử dụng các phương pháp đề xuất đối tượng xuất hiện trong ảnh như BING [15], EdgeBox [16], Geodesic [17], Selective Search [18] để thu được tọa độ và kích thước của các đối tượng có thể có trong ảnh.
- o Gom cụm các điểm ảnh 3D: lấy dữ liệu từ bước phỏng đoán sơ đồ địa hình và bước đề xuất đối tượng xuất hiện trong ảnh, bước này sử dụng giải

thuật gom cụm DBSCAN để gom các điểm ảnh 3D từ danh sách các điểm ảnh thu được từ giai đoạn phỏng đoán sơ đồ địa hình để loại bỏ các điểm nhiễu và chỉ giữ lại các điểm ảnh có khả năng là điểm ảnh thuộc đối tượng. Để tăng độ chính xác của DBSCAN, họ sử dụng thêm kết quả từ giai đoạn đề xuất đối tượng để giữ lại các điểm ảnh có xác suất cao thuộc đối tượng.

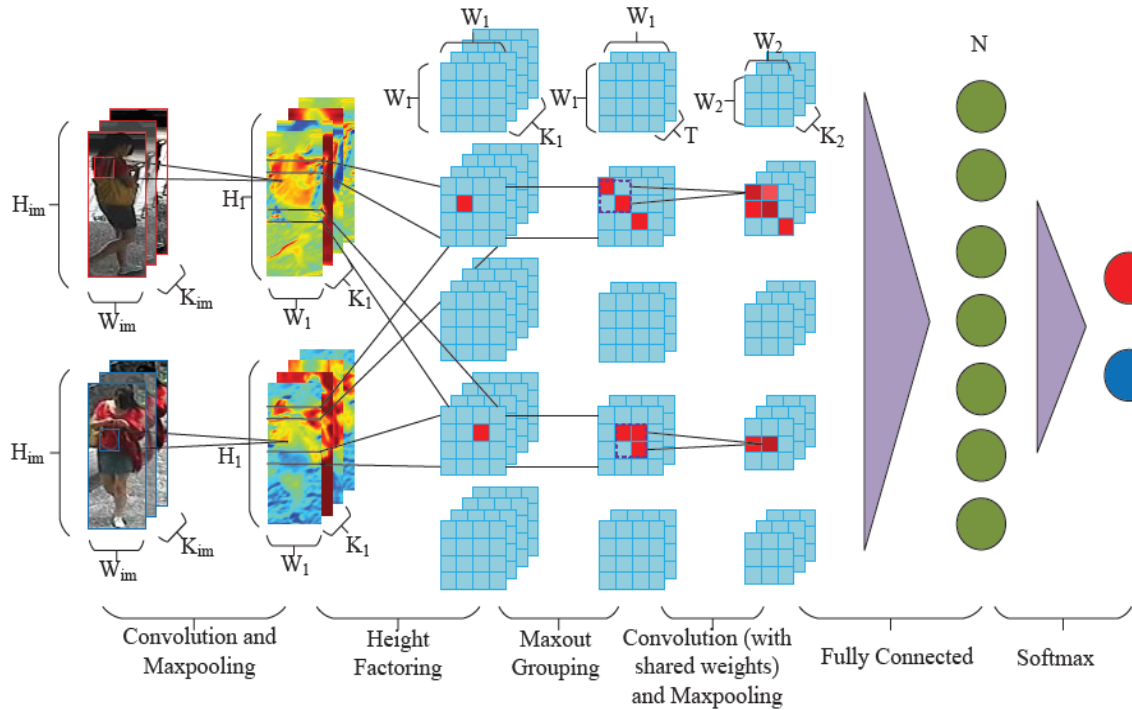
- **Bước 3: detection.** Sử dụng Fast R-CNN [19] để xác định đối tượng trong ảnh 2D và sử dụng linear SVN để huấn luyện các điểm ảnh 3D. Từ hai kết quả rời rạc, họ tính giá trị tin cậy cho quá trình phát hiện đối tượng của mình.
- **Bước 4: Tracking.** rút trích năm đặc trưng như giá trị trung bình tọa độ điểm ảnh 3D, phương sai, sơ đồ màu, kích thước của mảng và số lượng điểm ảnh 3D chứa trong đối tượng. Sử dụng khoảng cách Euclid để tính được độ tương đồng của các đối tượng giữa hai frame ảnh liên tiếp nhau.

Thư viện mà tác giả xây dựng có thể giải quyết các trường hợp thách thức ví dụ như bài toán che phủ mà một camera hoặc 3D LIDAR không thể giải quyết độc lập được. Đồng thời với độ chính xác cao và khả năng tính toán trong thời gian thực, thư viện rất kì vọng có thể đưa vào ứng dụng trong các phương tiện di chuyển thông minh.

Latha [20] đã thực hiện một cuộc khảo sát tập trung vào việc giải quyết bài toán che phủ các đối tượng trong truy vết. Tác giả đã phân loại khá đầy đủ các trường hợp có thể xảy ra trong quá trình truy vết đối tượng như là đối tượng không bị che phủ, bị che phủ một phần, bị che phủ hoàn toàn và bị che phủ hoàn toàn trong một khoảng thời gian dài, hay đối tượng rời khỏi vùng quan sát trong hệ thống nhiều camera cũng được coi là một trạng thái bị che phủ. Theo những nghiên cứu của mình Ms. Latha xét thấy rằng để giải quyết bài toán che phủ thì việc sử dụng camera đơn không mang lại hiệu quả cao bởi lẽ các cách tiếp cận hiện tại trong việc truy vết đối tượng hầu hết đều dựa vào kết quả của quá trình phát hiện đối tượng mà trong trường hợp che phủ ta không thể thu được. Do đó, việc sử dụng hệ thống nhiều camera sẽ mang lại kết quả tốt hơn rõ rệt. Với việc sử dụng nhiều camera cùng tham gia theo dõi, ta có thể thu được các đặc trưng về độ sâu, vị trí, kết cấu và màu sắc của đối tượng, ngay cả trong trường hợp bị che phủ, ta cũng có thể sử dụng đặc trưng về không gian, thời gian và góc nhìn chia sẻ được giữa các camera.

Wei [21] đề xuất xây dựng một mạng nơ ron học sâu nhằm xác định định danh của người xuất hiện trong đoạn camera. Ở nghiên cứu [21] tác giả đề xuất mạng nơ ron FPNN để xác định định danh của đối tượng di chuyển thường được sử dụng trong các hệ

thông camera phân tán không có vùng trùng lặp hay bất cứ thông tin liên hệ trực tiếp với nhau dựa trên các đặc trưng về màu sắc của đối tượng. Điểm đóng góp của nghiên cứu này đó là thay vì sử dụng các đặc trưng thủ công bằng thao tác trực tiếp thì tác giả đề xuất phương pháp thu và học các đặc trưng từ dữ liệu có được một cách tự động. mạng nơ ron FPNN được mô tả như hình 2.7:



Hình 2.7: Filter pairing neural network [21].

Trong môi trường thực tiễn việc nhận dạng đối tượng thông qua camera không chỉ đơn giản là nhận dạng trên duy nhất một camera, vì phạm vi theo dõi của một camera rất giới hạn nên không thể đáp ứng được yêu cầu thiết yếu của thực tiễn khi mà việc theo dõi đối tượng có thể là xuyên suốt qua một vùng không gian rộng lớn từ không gian giữa hai căn phòng đến không gian giữa các căn nhà hay thậm chí là không gian giữa các tuyến đường..., Ergys [22] đã làm một cuộc nghiên cứu về vấn đề độ hiệu quả và tập dữ liệu cho nhiều đối tượng truy vết qua nhiều camera. Trong nghiên cứu này, Ristani định nghĩa những biện pháp mới để tính toán hiệu suất theo dõi của nhiều đối tượng trong nhiều camera MTMC. Tác giả cũng giới thiệu về tập dữ liệu đã được hiệu chỉnh lớn nhất cho tới thời điểm thực hiện nghiên cứu đó để so sánh các phương pháp theo dõi MTMC.

Tóm lại, từ việc tìm hiểu những nghiên cứu liên quan, tôi thấy được điểm mạnh và điểm yếu của các nghiên cứu trong lĩnh vực thị giác máy tính. Từ đó, tôi tìm ra cho mình phương pháp gán nhãn đối tượng qua nhiều camera.

CHƯƠNG 3.

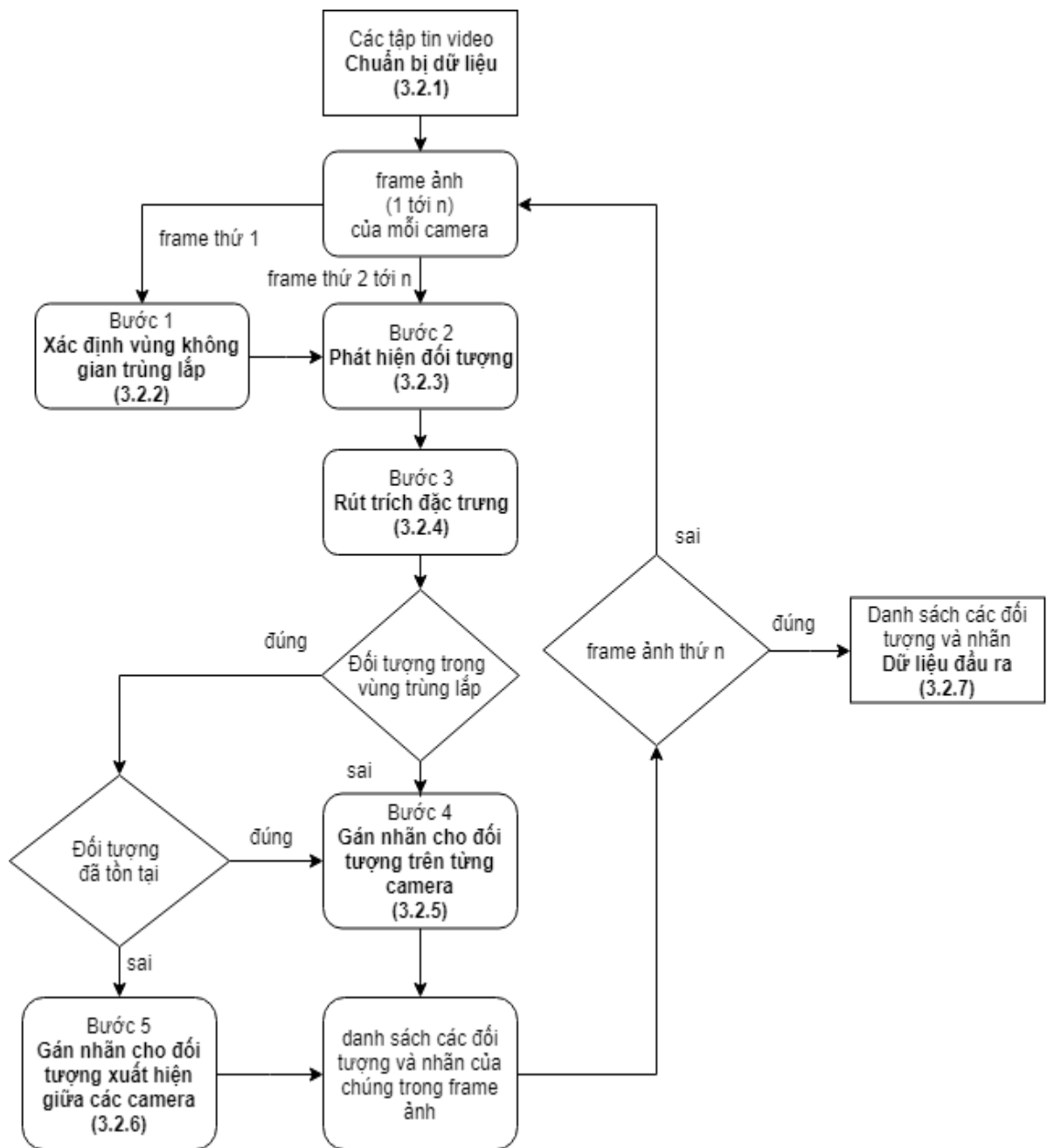
GÁN NHÃN ĐỐI TƯỢNG DI CHUYỂN QUA NHIỀU CAMERA

3.1. MÔ TẢ VÀ YÊU CẦU BÀI TOÁN

Như đã trình bày ở chương 1, mục tiêu của đề tài là làm sao gán nhãn được cho các đối tượng di chuyển qua nhiều camera khác nhau. Đầu vào của bài toán là một tập các đoạn video thu được từ hệ thống camera và đầu ra mong đợi là danh sách các đối tượng di chuyển trong hệ thống camera đã được gán nhãn trên từng frame ảnh liên tiếp nhau của các đoạn video đầu vào. Như vậy, dựa trên mục tiêu đề ra, tôi phải giải quyết bài toán thuộc lớp MTMC. Với những bài toán thuộc lớp này, tôi phải chia thành các bài toán nhỏ và giải quyết từng bài toán nhỏ cụ thể đó. Trước tiên tôi phải phát hiện được các đối tượng di chuyển trong từng camera. Việc phát hiện các đối tượng này có ý nghĩa quan trọng, tác động lớn đối với tính chính xác của việc gán nhãn bước sau. Phát hiện đối tượng càng chính xác bao nhiêu thì việc gán nhãn sẽ cho độ tin cậy cao bấy nhiêu. Bài toán nhỏ thứ hai là phải gán nhãn được cho các đối tượng vừa nhận được từ bài toán trước và đảm bảo tính nhất quán trong quá trình gán nhãn khi mà các đối tượng có thể di chuyển theo nhiều hướng khác nhau, có thể di chuyển gần hoặc xa camera hay hình dáng, kích thước cũng sẽ thay đổi theo góc nhìn... Và bài toán cuối cùng mà tôi phải giải quyết đó là làm sao đảm bảo tính nhất quán trong việc gán nhãn cho đối tượng di chuyển qua nhiều camera. Các camera này có thể được thiết lập trong hai trường hợp: các camera được đặt song song và không song song.

3.2. PHƯƠNG PHÁP ĐỀ XUẤT

Với bài toán đã được mô tả ở trên, thách thức lớn nhất đó là phải chia nhỏ thành nhiều bài toán con. Như vậy, tôi không thể giải quyết thông qua một bước duy nhất, mà phải chia ra từng phần nhỏ để giải quyết từng bài toán cụ thể. Sơ đồ khối hình 3.1 giúp cho tôi có thể giới thiệu khái quát về phương pháp mà tôi đề xuất để giải quyết các bài toán đặt ra trên:



Hình 3.1. Sơ đồ khối quy trình gán nhãn cho các đối tượng di chuyển.

Theo sơ đồ khối, quy trình gán nhãn gồm 7 bước:

- **Bước 1:** Chuẩn bị dữ liệu.
- **Bước 2:** Xác định vùng không gian trùng lắp.
- **Bước 3:** Phát hiện đối tượng.
- **Bước 4:** Rút trích đặc trưng.
- **Bước 5:** Gán nhãn cho đối tượng trên từng camera.

- **Bước 6:** Gán nhãn cho đối tượng xuất hiện giữa các camera.
- **Bước 7:** Xuất dữ liệu đầu ra.

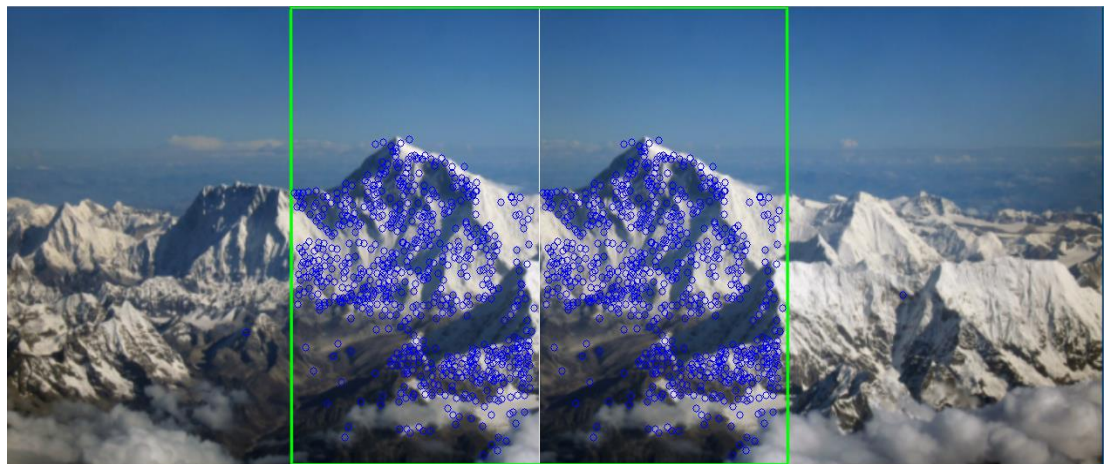
Các bước được mô tả cụ thể trong các phần sau.

3.2.1. Chuẩn bị dữ liệu

Với một hệ thống camera được thiết kế theo mô tả trong chương giới thiệu đề tài tôi sẽ thu được hai đoạn video định dạng dav tương ứng với từng camera. Tôi sẽ đặt tên file lần lượt là camera1.avi và camera2.avi. Chi tiết về các thông số và cách thu thập dữ liệu sẽ được giới thiệu chi tiết ở phần 4.1.

3.2.2. Xác định vùng không gian trùng lặp

a) Hai camera đặt song song với nhau



Hình 3.2: Tìm vùng không gian trùng lặp sử dụng đặc trưng SIFT[24].

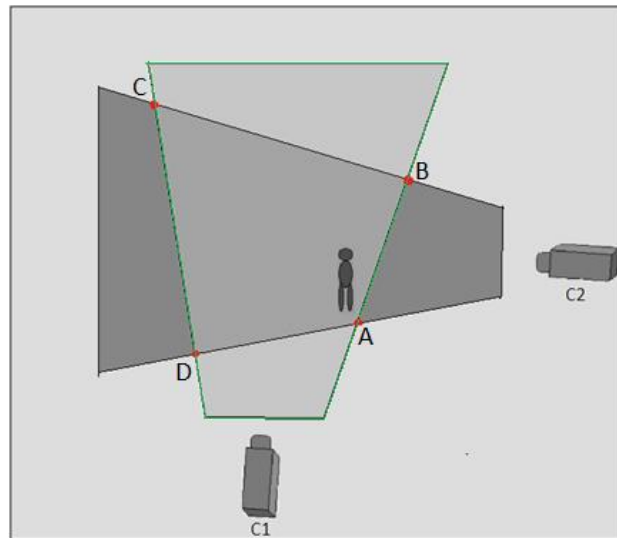
Đối với hai camera đặt song song với nhau ta có thể dễ dàng xác định được vùng không gian trùng lặp thông qua việc rút trích đặc trưng SIFT giữa hai frame ảnh cắt được từ hai camera. Để thực hiện được điều đó, tôi sẽ lần lượt thực hiện các bước sau:

- + **Bước 1:** cắt frame ảnh đầu tiên từ hai đoạn video thu được từ camera.
- + **Bước 2:** rút trích đặc trưng SIFT để tìm được các điểm đặc trưng (landmark) của mỗi frame ảnh.
- + **Bước 3:** tìm tất cả các điểm đặc trưng khớp nhau giữa hai frame ảnh

+ **Bước 4:** dựa trên các điểm đặc trưng khớp nhau đó, ta có thể tìm được phép chiếu giữa hai frame ảnh và khoanh vùng được vùng không gian trùng lắp. Hình 3.2 thể hiện kết quả của quá trình rút trích các điểm tương đồng và thực hiện phép chiếu dựa trên các điểm tương đồng đó của hai bức ảnh:

b) Hai camera đặt chéo nhau

Để xác định được vùng không gian trùng lắp của hai camera trong trường hợp này ta không thể sử dụng đặc trưng SIFT được mà phải dùng phương pháp tìm điểm nằm trên đường biên.



Hình 3.3: Hai camera cắt nhau.

Như ví dụ bên trên, để xác định được vùng không gian mà hai camera có thể cùng quan sát được, tôi sử dụng một người di chuyển theo đường biên (đường biên là đường mà ở đó đối tượng đang ở ranh giới giữa trong và ngoài vùng quan sát được của camera) của camera một. Người đó sẽ lần lượt di chuyển qua các vị trí A, B, C, D theo chiều ngược kim đồng hồ. Đặc điểm để nhận biết vị trí nào là A, B, C, D đó là các điểm này là điểm mà người di chuyển xuất hiện đồng thời trên cả hai camera một và hai. Cụ thể:

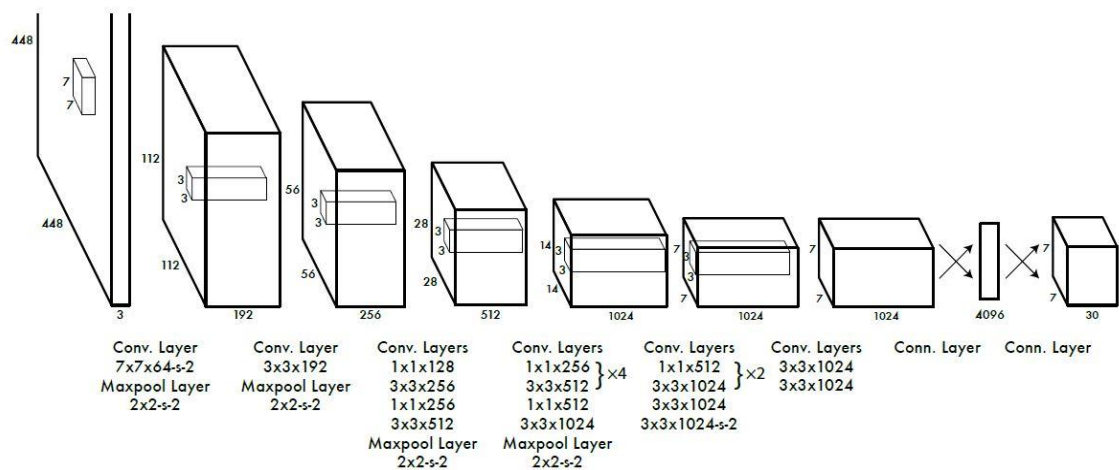
- Điểm A là điểm mà người đó lần đầu tiên xuất hiện trên camera hai.
- Điểm B là điểm cuối cùng mà người đó xuất hiện trên camera hai.
- Điểm C là điểm mà người đó lần đầu tiên xuất hiện trở lại camera hai.
- Điểm D là điểm mà người đó xuất hiện cuối cùng trên camera hai.

Từ bốn điểm A, B, C, D này ta xác định được vùng không gian trùng lắp mà cả hai camera đều thấy được. Chi tiết về vị trí các điểm A, B, C và D xác định được thể hiện như hình 3.3:

3.2.3. Phát hiện đối tượng

Đối với bài toán nhận dạng đối tượng di chuyển trong đoạn video sẽ có nhiều cách để giải quyết. Ở đây tôi sử dụng phương pháp mạng nơ ron để xây dựng mô hình đối tượng. Mạng nơ ron mà tôi sử dụng ở đây là mạng nơ ron YOLO [1].

Joseph [1] đã xây dựng một mạng nơ ron CNN dựa trên tập dữ liệu VOC 2007 và 2012. Lớp CL của mạng nơ ron dùng để rút trích các đặc trưng từ ảnh bao gồm các đặc trưng HAAR, SIFT, HOG còn các lớp FCL dự đoán xác suất đầu ra và tọa độ của đối tượng. Mạng YOLO được lấy cảm hứng từ mạng GoogLeNet phục vụ cho việc phân loại ảnh. Mạng YOLO có 24 lớp CL và sau cùng là 2 lớp FCL được mô tả như hình 3.4:



Hình 3.4: Kiến trúc mạng YOLO với 24 CL và 2 FCL [1].

3.2.4. Rút trích đặc trưng đối tượng

Đầu ra của quá trình nhận dạng đối tượng giúp ta tìm được vị trí thực của đối tượng trong frame ảnh, từ đó ta tiến hành thu thập các đặc trưng của đối tượng bao gồm:

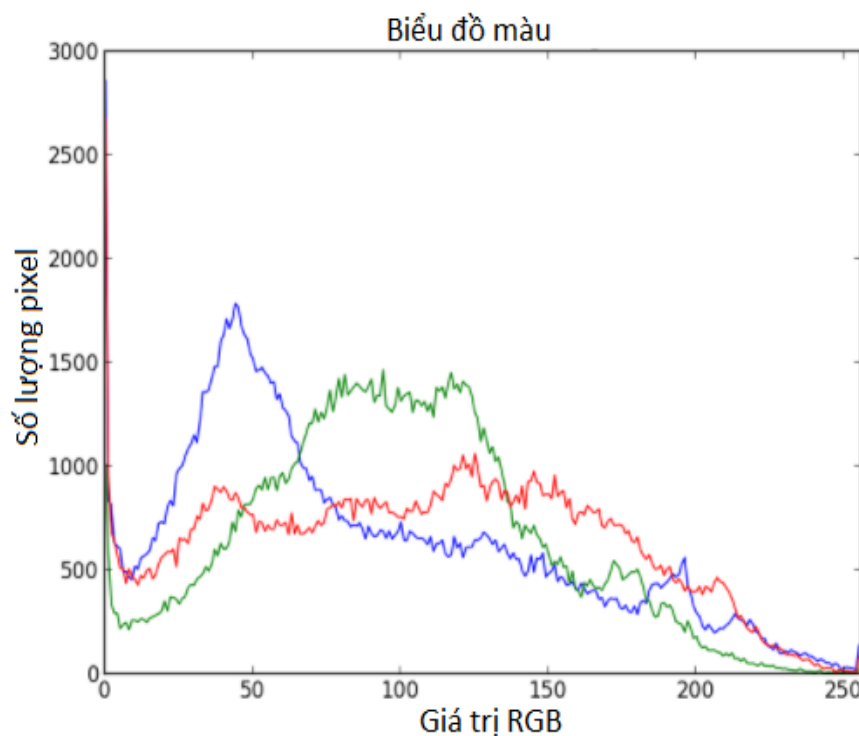
a) Màu

Ở đây, rút trích đặc trưng biểu đồ màu để đại diện cho đối tượng. Biểu đồ màu thể hiện cho sự phân tán của màu trong frame ảnh. Trong ảnh kỹ thuật số, ta sẽ có

rất nhiều khoảng màu khác nhau và biểu đồ màu sẽ là số lượng của các pixel nằm trong từng khoảng màu nhất định đó. Biểu đồ màu có thể được sử dụng trong không gian màu RGB hay HSV do đó nó rất được phổ biến trong thị giác máy tính nói chung và xử lý ảnh nói riêng. Hình 3.5 và 3.6 minh họa cho biểu đồ màu:



Hình 3.5: Ảnh của một chú chó [25].



Hình 3.6: Biểu đồ màu của hình 3.5.

Trong hình 3.6, trục hoành thể hiện vùng giá trị của các màu trong không gian màu RGB và trục tung thể hiện số lượng pixel có cùng giá trị màu tương ứng.

b) Hình dáng

Để tăng tính chính xác của việc phân lớp, tôi sử dụng thêm phương pháp các moment bất biến của Hu (HIM) để rút trích đặc trưng hình dáng của đối tượng. Điểm mạnh của phương pháp này là loại bỏ được rào cản về sự thay đổi hình dáng của đối tượng bởi việc xoay, thay đổi kích thước, góc nhìn đối với camera. HIM là đặc trưng rất có lợi đối với ảnh 2 chiều, nếu chúng ta đại diện đối tượng R cho một khung ảnh, moment trung tâm của thứ tự $(p + q)$ của R được định nghĩa như sau:

$$\mu_{p,q} = \sum_{(x,y) \in R} (x - x_c)^p (y - y_c)^q \quad (3.1)$$

Trong đó, (x_c, y_c) là trung tâm của đối tượng.

Ta chuẩn hóa moment trung tâm theo công thức:

$$\eta_{p,q} = \frac{\mu_{p,q}}{\mu_{0,0}^\gamma}, \quad \gamma = \frac{p+q+2}{2} \quad (3.2)$$

Dựa vào các moment trung tâm đã được chuẩn hóa, Hu giới thiệu bảy moment bất biến:

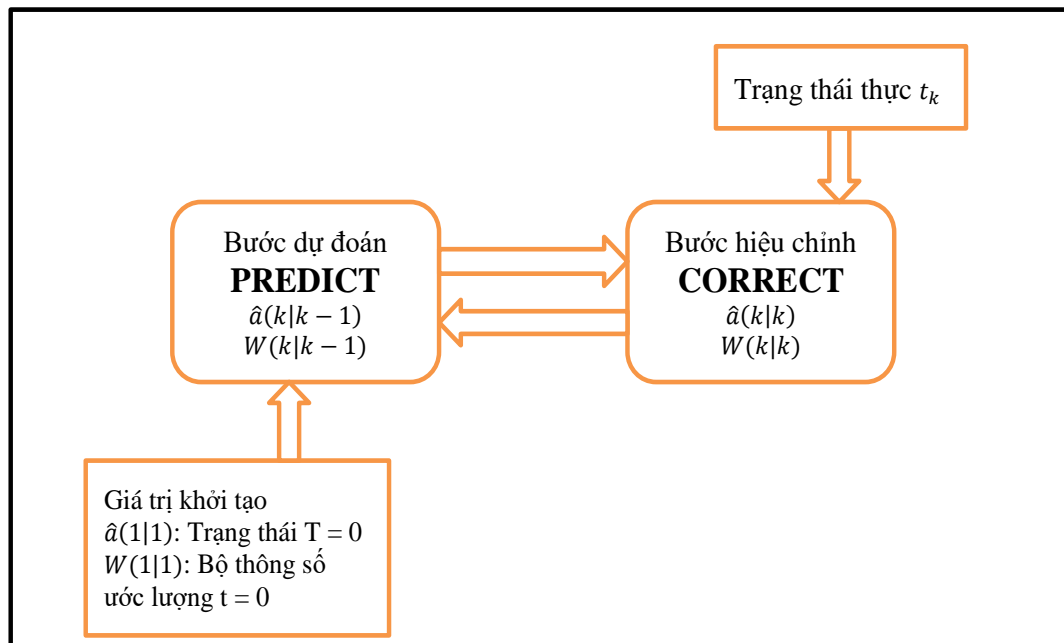
$$\begin{aligned} \emptyset_1 &= \eta_{2,0} + \eta_{0,2} \\ \emptyset_2 &= (\eta_{2,0} - \eta_{0,2})^2 + 4\eta_{1,1}^2 \\ \emptyset_3 &= (\eta_{3,0} - 3\eta_{1,2})^2 + (\eta_{0,3} - 3\eta_{2,1})^2 \\ \emptyset_4 &= (\eta_{3,0} + \eta_{1,2})^2 + (\eta_{0,3} + \eta_{2,1})^2 \\ \emptyset_5 &= (\eta_{3,0} - 3\eta_{1,2})(\eta_{3,0} + \eta_{1,2})[(\eta_{3,0} + \eta_{1,2})^2 - 3(\eta_{0,3} + \eta_{2,1})^2] + \\ &\quad + (\eta_{0,3} - 3\eta_{2,1})(\eta_{0,3} + \eta_{2,1})[(\eta_{0,3} + \eta_{2,1})^2 - 3(\eta_{3,0} + \eta_{1,2})^2] \\ \emptyset_6 &= (\eta_{2,0} - \eta_{0,2})[(\eta_{3,0} + \eta_{1,2})^2(\eta_{0,3} + \eta_{2,1})^2 + 4\eta_{1,1}(\eta_{3,0} + \eta_{1,2})(\eta_{0,3} + \eta_{2,1})] \\ \emptyset_7 &= (3\eta_{2,1} - \eta_{0,3})(\eta_{3,0} + \eta_{1,2})[(\eta_{3,0} + \eta_{1,2})^2 - 3(\eta_{0,3} + \eta_{2,1})^2] \end{aligned} \quad (3.3)$$

Bảy moment bất biến này là những đặc trưng cực kì hữu dụng khi mà nó không bị thay đổi cho dù đối tượng trong ảnh có bị thay đổi kích thước, xoay hoặc di chuyển theo các chiều khác nhau. Điều này giúp cho việc phân lớp đối tượng chính xác hơn trong môi trường thực tế đối với camera quan sát. Đối tượng có thể di chuyển qua lại theo nhiều hướng khác nhau hoặc di chuyển lại gần hay ra xa camera, tất cả những thay đổi đó đều làm cho hình dáng của đối tượng không còn giống như ban đầu nữa và khiến cho việc phân lớp đối tượng gặp khó khăn.

3.2.5. Gán nhãn cho đối tượng trên từng camera

Để thực hiện việc gán nhãn cho đối tượng trên từng frame ảnh nối tiếp nhau, tôi sử dụng giải thuật Kalman filter.

Kalman filter là một trong những giải thuật khá nổi tiếng trong lớp giải thuật chuỗi thời gian. Kalman filter là giải thuật ước lượng đệ quy giữa hai trạng thái dự đoán (prediction) và hiệu chỉnh (correction) nhằm xác định trạng thái của một quá trình tuyến tính. Trạng thái thứ nhất là trạng thái dự đoán, ở trạng thái này giải thuật kalman filter sẽ dự đoán giá trị trạng thái tiếp theo của quá trình dựa trên các thông số đã được tính toán. Tới giai đoạn thứ hai là giai đoạn hiệu chỉnh, khi ta có được giá trị thực của trạng thái dự đoán trước đó, các thông số dự đoán sẽ được cập nhật lại để chuẩn bị cho giai đoạn dự đoán tiếp theo. Các bước của giải thuật kalman filter mô tả theo hình 3.7:



Hình 3.7: Chu trình trong giải thuật Kalman Filter.

Với việc sử dụng giải thuật Kalman filter để dự đoán vị trí của đối tượng trong frame ảnh tiếp theo, ta có thể gán nhãn cho đối tượng di chuyển trong đoạn video.


Tuy nhiên, trong việc gán nhãn đối tượng di chuyển trên từng frame ảnh nếu sử dụng giải thuật kalman filter thông thường sẽ không cho kết quả tốt khi các đối tượng di chuyển gần nhau. Do đó, tôi sử dụng phương pháp tính toán IoU để cải thiện hiệu

suất gắn nhãn dựa trên kalman filter. Phương pháp này thường được sử dụng trong các giải thuật phát hiện đối tượng có giá trị dự đoán. Thực tế, giá trị dự đoán và giá trị thu nhận được bao giờ cũng có một khoảng chênh lệch như hình 3.8:



Hình 3.8: Giá trị dự đoán và giá trị thực của quá trình phát hiện đối tượng.

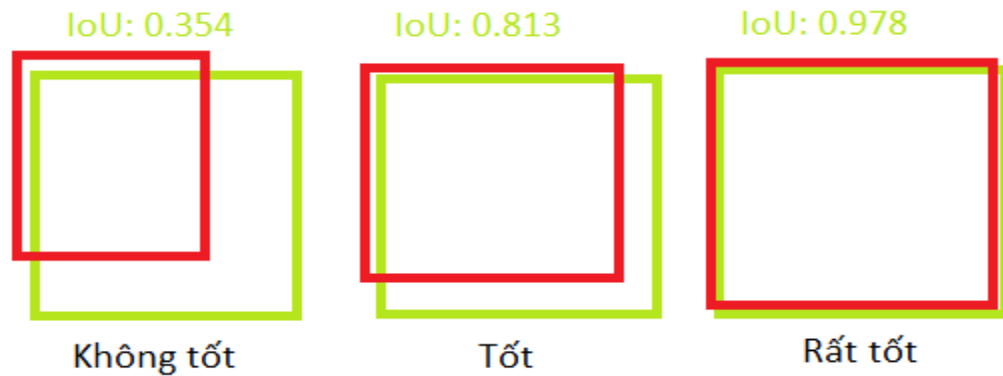
Khi đó, giá trị IoU sẽ được tính dựa trên tỉ số giữa vùng giao và vùng hợp của giá trị thực thu được và giá trị dự đoán theo hình 3.9:

$$\text{IoU} = \frac{\text{Vùng giao}}{\text{Vùng hợp}}$$


Hình 3.9: Cách tính giá trị IoU.

Giá trị IoU này được sử dụng như là một thước đo xác định tính đúng đắn của giá trị dự đoán so với giá trị thu được. Miền giá trị của $\text{IoU} \in [0,1]$, khi giá trị gần về 1 nghĩa là vùng giao sẽ gần bằng vùng hợp và giá trị dự đoán sẽ gần bằng với giá trị

thu được. Ý nghĩa giá trị IoU trong quá trình gán nhãn đối tượng được thể hiện trong hình 3.10:

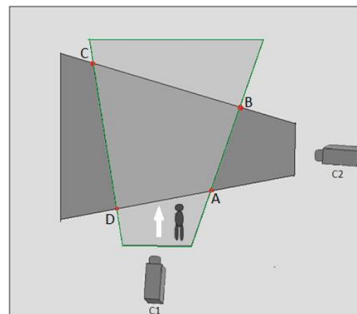


Hình 3.10: Ý nghĩa của giá trị IoU.

3.2.6. Gán nhãn cho đối tượng xuất hiện giữa các camera

Dựa trên việc thiết lập hệ thống camera, khi một đối tượng di chuyển từ camera này (C1) sang camera kia (C2), đối tượng sẽ di chuyển qua vùng không gian trùng lấp giữa hai camera. Lúc này, đối tượng khi ở camera C1 đã được gán nhãn và như vậy, để đảm bảo tính nhất quán trong việc gán nhãn giữa các camera, tôi sẽ gán nhãn cho đối tượng đó trong camera C2 bằng nhãn của nó trong camera C1.

Để có thể xác định chính xác đối tượng vừa xuất hiện trong camera C2 này là đối tượng nào trong camera C1, tôi sẽ sử dụng vị trí của đối tượng trong vùng không gian trùng lấp. Trong bước 3.2.2, tôi đã xác định được vùng không gian trùng lấp giữa hai camera. Khi đó trong camera C1 và C2, tôi sẽ lần lượt có được tọa độ của 4 đỉnh A, B, C, D như hình 3.3. Việc xác định được đối tượng vừa xuất hiện trong C2 là đối tượng nào trong C1 ta tính toán khoảng cách từ các đối tượng trong C1 đến các cạnh tương ứng của tứ giác ABCD, được biểu diễn như hình 3.11:



Hình 3.11: Đối tượng di chuyển từ camera C1 sang camera C2.

Trong hình 3.11 trên, một đối tượng di chuyển từ camera C1 sang camera C2 theo hướng mũi tên, như vậy vị trí đầu tiên mà đối tượng xuất hiện trong camera C2 sẽ gần cạnh AD của tứ giác ABCD. Khi phát hiện được đối tượng trong camera C2, tôi sẽ tính khoảng cách của tất cả các đối tượng xuất hiện trong vùng không gian trùng lấp ABCD của camera C1 đến cạnh AD. Đối tượng nào có khoảng cách gần với AD nhất rất có khả năng là đối tượng ta cần xét. Tuy nhiên, chỉ với một tiêu chí như vậy, tôi chưa thể kết luận đó chính là đối tượng mà ta quan tâm vì rất có thể một đối tượng khác đã xuất hiện trong camera C2 vô tình đứng gần AD hơn cả đối tượng mà ta quan tâm. Chính vì thế, tôi sẽ sử dụng giải thuật GSA.

Giải thuật GSA (hay Stable Matching) là một giải thuật rất phổ biến sử dụng để xác định các cặp tương đồng giữa hai nhóm đối tượng. Áp dụng vào bài toán của tôi, tôi sẽ có hai nhóm đối tượng cần ghép cặp với nhau. Nhóm thứ nhất là nhóm các đối tượng di chuyển xuất hiện trong vùng không gian trùng lấp của camera C1 ký hiệu tập M và nhóm thứ hai là nhóm các đối tượng di chuyển xuất hiện trong vùng không gian trùng lấp của camera C2 ký hiệu tập N. Giải thuật GSA sẽ tìm các cặp đối tượng tương ứng cùng xuất hiện trong cả hai camera cụ thể như hình 3.12 sau:

```
function GSA {
    # khởi tạo các phần tử  $m \in M$  và  $n \in N$  là các phần tử tự do (chưa có cặp)
    while  $\exists$  m tự do mà vẫn có một phần tử n nào đó thích hợp {
        n = phần tử thích hợp nhất trong các phần tử thích hợp bắt cặp với m
        if n tự do
            (m, n)          # m và n trở thành một cặp
        else  $\exists$  (m', n)    # n đã có cặp với m'
            if n tương đồng với m hơn m'
                m' trở thành phần tử tự do
                (m, n)      # m và n trở thành một cặp
            else
                (m', n)     # m' và n vẫn được giữ là một cặp
        n bị loại khỏi danh sách các phần tử thích hợp với m
    }
}
```

Hình 3.12: Mô tả giải thuật GSA.

Sau khi thực hiện giải thuật GSA, ta sẽ có được một tập các đối tượng ở vùng không gian trùng lấp của camera C1 tương ứng với các đối tượng xuất hiện ở vùng không gian trùng lấp của camera C2. Sau đó ta sẽ có được đối tượng tương ứng với đối tượng vừa mới xuất hiện trong camera C2 và thực hiện gán nhãn cho nó.

3.2.7. Dữ liệu đầu ra

Sau khi gán nhãn xong tôi thu được đối tượng đã được gán nhãn với các đặc trưng của nó. Dữ liệu này tôi sẽ lưu vào kho dữ liệu. Mỗi lần xử lý một frame, tôi sẽ có một danh sách các đối tượng di chuyển trong foreground frame với đầy đủ thông tin về đặc trưng màu sắc, hình dáng và nhãn của nó.

3.3. PHƯƠNG PHÁP ĐÁNH GIÁ

Với bài toán này, tôi sẽ có 2 tiêu chí để đánh giá độ chính xác của việc gán nhãn: tính nhất quán, chính xác (là tính đúng đắn trong việc gán nhãn, cùng một đối tượng phải được gán cùng một nhãn) của việc gán nhãn trên từng camera và tính nhất quán, chính xác của việc gán nhãn trên hai camera.

Khi một đối tượng xuất hiện trong camera, tôi sẽ xác định đối tượng đó và gán nhãn cho nó. Khi đối tượng di chuyển, nếu việc gán nhãn đối tượng đó qua các frame ảnh nhất quán, tôi sẽ cho đó là một gán nhãn đúng ngược lại sẽ là gán nhãn sai. Tương tự như vậy, khi đối tượng di chuyển qua hai camera khác nhau, nếu việc gán nhãn ở hai camera nhất quán trên đối tượng, tôi sẽ xác định đó là một gán nhãn đúng, ngược lại là gán nhãn sai.

Tiến hành quá trình trên cho từng đối tượng di chuyển trong hai đoạn camera, tôi sẽ thu được số lượng các đối tượng được xác định là gán nhãn đúng và sai. Từ đó sẽ thống kê trên số liệu thu thập đó để xác định hiệu suất và đưa ra nhận xét cho phương pháp mà tôi đề xuất.

CHƯƠNG 4.

THÍ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

4.1. TẬP DỮ LIỆU ĐÁNH GIÁ

Với mục tiêu đã đặt ra trong chương 1, tôi sẽ xây dựng một hệ thống gán nhãn đối tượng di chuyển qua hai camera được thiết lập trong cả hai trường hợp: có vùng không gian trùng lấp song song và không song song. Kết quả dự kiến sau khi xây dựng hệ thống:

- Hệ thống có khả năng phát hiện đối tượng di chuyển trong hai đoạn video.
- Gán nhãn thành công và đảm bảo được tính nhất quán đối với việc gán nhãn. Một đối tượng di chuyển qua nhiều vị trí khác nhau trong đoạn video đều được gán nhãn giống nhau.
- Đảm bảo tính nhất quán trong việc gán nhãn đối với một đối tượng di chuyển qua cả hai camera.

Tập dữ liệu dùng để kiểm chứng kết quả dự kiến được thu từ hai nguồn:

- Tập dữ liệu do tác giả Francois [23] xây dựng trong khuôn viên trường đại học của họ gồm 4 video. Các camera được đặt ghi hình ở độ cao 1.8m với các góc nhìn chéo nhau và có vùng không gian trùng lấp. Độ phân giải của video thu được từ camera quan sát là 360x288 và tốc độ ghi hình 25 frames/giây.
- Tập dữ liệu do tôi xây dựng trong khuôn viên trường Đại học Bách Khoa Tp.Hồ Chí Minh gồm 6 video. Các camera được đặt ghi hình ở độ cao 1.4m với góc nhìn song song và có vùng không gian trùng lấp. Độ phân giải của video thu được từ camera quan sát là 960x720 và tốc độ ghi hình 30 frames/giây.

Kết quả thí nghiệm trên tập dữ liệu trên được trình bày ở phần tiếp theo.

4.2. KẾT QUẢ THÍ NGHIỆM

Trong thí nghiệm, tôi sẽ lần lượt chạy các đoạn video thu được từ camera như mô tả dữ liệu tập đánh giá bao gồm cả đặt chéo nhau và đặt song song bằng hệ thống mình đề xuất. Tuy nhiên, trong khuôn khổ trình bày kết quả đánh giá này, tôi chỉ trích xuất kết quả của 10 đoạn video. Việc đánh giá kết quả sẽ được tính dựa trên việc gán nhãn trên

từng camera và trên cả hệ thống. Kết quả lần lượt sẽ được biểu diễn trong bảng 4.1, 4.2, 4.3:

STT	Số đối tượng xuất hiện	Số đối tượng gán nhãn đúng	Số đối tượng gán nhãn sai	Chính xác(%)
Tập dữ liệu [23], hai camera đặt chéo nhau				
1	10	8	2	80
2	12	9	3	75
3	13	9	4	69
4	10	8	2	80
Tập dữ liệu tự xây dựng, hai camera đặt song song				
1	5	5	0	100
2	4	4	0	100
3	8	8	0	100
4	15	14	1	93.3
5	12	11	1	91.6
6	11	10	1	90.9

Bảng 4.1: Kết quả thí nghiệm trên camera 1.

STT	Số đối tượng xuất hiện	Số đối tượng gán nhãn đúng	Số đối tượng gán nhãn sai	Chính xác(%)
Tập dữ liệu [23], hai camera đặt chéo nhau				
1	10	8	2	80
2	12	9	3	75
3	13	9	4	69
4	10	8	2	80
Tập dữ liệu tự xây dựng, hai camera đặt song song				
1	5	5	0	100
2	4	4	0	100
3	8	7	1	87.5
4	13	11	2	84.6
5	11	11	0	100
6	9	9	0	100

Bảng 4.2: Kết quả thí nghiệm trên camera 2.

STT	Số đối tượng xuất hiện	Số đối tượng gán nhãn đúng	Số đối tượng gán nhãn sai	Chính xác(%)
Tập dữ liệu [23], hai camera đặt chéo nhau				
1	10	8	2	80
2	12	9	3	75
3	13	9	4	69
4	10	8	2	80
Tập dữ liệu tự xây dựng, hai camera đặt song song				
1	5	4	1	80
2	4	4	0	100
3	8	6	2	75
4	16	14	2	87.5
5	12	11	1	91.6
6	11	10	1	90.9

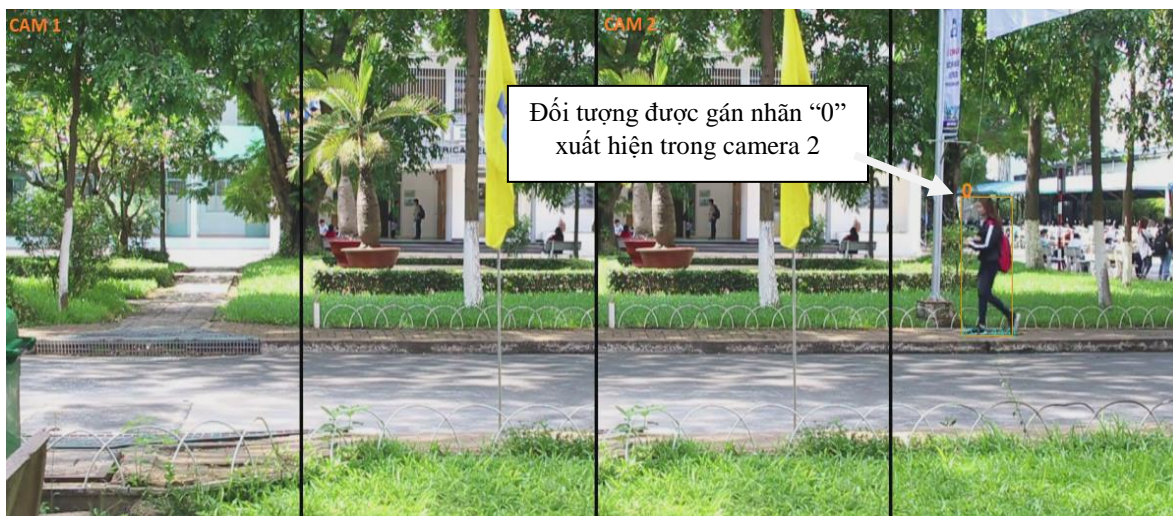
Bảng 4.3: Kết quả thí nghiệm trên hệ thống hai camera.

Bảng 4.1, 4.2 thể hiện kết quả thí nghiệm trên từng camera đơn và bảng 4.3 là kết quả thí nghiệm trên hệ thống hai camera mà tôi xây dựng.

Để tính toán được bảng 4.1, 4.2, với mỗi đoạn video, tôi sẽ tiến hành gán nhãn cho từng đối tượng di chuyển trong đó, còn bảng 4.3 tôi sẽ tiến hành gán nhãn cho từng đối tượng khi đối tượng xuất hiện trong hệ thống đến khi đối tượng rời khỏi hệ thống. Hình 4.1, 4.2, 4.3 mô tả các trạng thái gán nhãn cơ bản của đối tượng:

- Xuất hiện trong một camera.
- Vào vùng không gian trùng lắp của hai camera.
- Xuất hiện trong camera còn lại.

Mỗi đối tượng xuất hiện trong hệ thống sẽ được tính là 1. Trong trường hợp đối tượng rời khỏi hệ thống và quay trở lại sau đó sẽ được coi là một đối tượng mới vì hệ thống không giải quyết bài toán gán nhãn trên vùng không gian không trùng lắp. Một đối tượng được gán cùng một nhãn từ khi đối tượng đó xuất hiện trong hệ thống đến khi đối tượng rời khỏi hệ thống sẽ được tính là 1 gán nhãn đúng. Độ chính xác của việc gán nhãn sẽ dựa trên tỉ số giữa số đối tượng gán nhãn đúng trên tổng số các đối tượng xuất hiện trong hệ thống.



Hình 4.1: Đối tượng xuất hiện trong camera 2.



Hình 4.2: Đối tượng xuất hiện trong vùng không gian trùng lặp của hai camera.



Hình 4.3: Đối tượng xuất hiện trong camera 1.

Từ kết quả của những thí nghiệm, tôi thấy rằng quá trình gán nhãn chịu ảnh hưởng từ các bước rút trích đặc trưng của đối tượng (màu, hình dáng, vị trí), truy vết đối tượng dựa trên giải thuật Kalman Filter hay phát hiện đối tượng dựa trên YOLO. Trong đó, có ảnh hưởng trực tiếp và lớn nhất tới kết quả của hệ thống là phát hiện đối tượng. Việc phát hiện đối tượng sai lệch và thiếu chính xác dẫn đến việc rút trích đặc trưng đối tượng không mang lại giá trị tối ưu nhằm phân biệt và truy vết đối tượng của phương pháp Kalman Filter. Để cải thiện hệ thống, việc quan trọng nhất vẫn là tìm cách nâng cao hiệu suất của giai đoạn phát hiện đối tượng.

Có thể nhìn thấy rõ các tác động trên ảnh hưởng như thế nào đối với tính đúng đắn của việc gán nhãn thông qua kết quả từ các bảng 4.1, 4.2, 4.3. Độ chính xác của các đoạn video từ tập dữ liệu [23] thấp hơn nhiều so với tập dữ liệu mà tôi xây dựng vì các đoạn video trong tập dữ liệu [23] thu được từ các camera thường, không có rõ nét và màu bị mờ gần như là video trắng đen nên việc rút trích đặc trưng không thu được các đặc trưng có độ phân biệt cao giữa các đối tượng dẫn đến quá trình gán nhãn sai lệch. Bên cạnh đó, ngoài tập dữ liệu [23], tôi cũng xây dựng tập dữ liệu với nhiều ngữ cảnh từ đơn giản đến phức tạp để làm phong phú thêm nguồn kết quả và tạo sự rõ ràng trong việc đánh giá điểm mạnh, điểm yếu cũng như môi trường và ngữ cảnh mà hệ thống có thể gán nhãn với độ chính xác cao. Độ phức tạp của ngữ cảnh thể hiện ở việc chồng lấp giữa các đối tượng cũng như giữa các vật cản với các đối tượng, khiến cho việc phát hiện đối tượng bằng mạng nơ ron YOLO không mang lại hiệu suất cao, từ đó làm giảm tính chính xác của hệ thống gán nhãn.

CHƯƠNG 5.

KẾT LUẬN

5.1. KẾT QUẢ ĐẠT ĐƯỢC

Dựa trên những nội dung đã được đề cập tới trong chương 1, từ bài nghiên cứu này:

- Tôi đã thu thập được nhiều những kiến thức liên quan đến thị giác máy tính nói chung và truy vết đối tượng nói riêng. Việc tham khảo nhiều những nghiên cứu liên quan đã giúp tôi có được cái nhìn rõ hơn về các phương pháp truy vết đối tượng, hiểu được ưu, nhược điểm của các phương pháp đó để từ đó vận dụng vào giải quyết bài toán mà tôi tìm hiểu trong nghiên cứu này.
- Đề xuất được phương pháp giải quyết cho bài toán gán nhãn đối tượng di chuyển qua nhiều camera và hiểu rõ hơn về ưu, nhược điểm của phương pháp mà tôi đề xuất để từ đó rút ra những kinh nghiệm để có thể mở rộng hướng phát triển cho nghiên cứu của tôi được sâu hơn, rộng hơn và có tính thực tiễn cao hơn.
- Hiện thực được phương pháp mà tôi đề xuất để kiểm chứng và chứng minh những gì mình tìm hiểu và những gì mình xây dựng có thật sự giải quyết được bài toán đã mô tả ở trên và đáp ứng được mục tiêu đề ra hay không.

Kết quả thu được tuy chưa được chính xác hoàn toàn, nhưng nó cũng đảm bảo được mục tiêu đề ra của đề tài và thực hiện đúng với hướng phát triển của nó. Với kết quả này, tôi có thể đầu tư và cải tiến bằng cách cải thiện trước tiên là phương pháp phát hiện đối tượng, sau đó là giải thuật kalma filter cũng như cải thiện các đặc trưng rút trích được của đối tượng để tăng độ chính xác cho quá trình gán nhãn, mở rộng thêm cho việc gán nhãn đối tượng di chuyển qua nhiều camera không trùng lặp.

5.2. ƯU ĐIỂM VÀ NHƯỢC ĐIỂM PHƯƠNG PHÁP GÁN NHÃN

5.2.1. Ưu điểm

- Tận dụng được phương pháp phát hiện đối tượng di chuyển trong đoạn camera đã được hiện thực trước đó với độ xử lý nhanh đáp ứng được các tác vụ đòi hỏi kết quả trả về trong thời gian thực.

- Giải thuật Kalman Filter đơn giản và dễ hiện thực nhưng vẫn đáp ứng được những yêu cầu cơ bản trong việc giải quyết bài toán truy vết đối tượng di chuyển trong camera.
- Phương pháp xác định vùng không gian trùng lấp dễ hiện thực và có tính chính xác cao, có thể áp dụng rộng rãi vì tính đơn giản của nó.
- Giải thuật GSA đưa ra được tập đối tượng tương đồng trên hai camera hiệu quả và có độ chính xác cao, hiện thực đơn giản và xử lý nhanh.
- Phương pháp đề xuất dễ hiểu.

5.2.2. Nhược điểm

- Giải thuật Kalman Filter rất dễ bị nhiễu khi hai đối tượng di chuyển chồng lấp lên nhau. Khi đó, phương pháp phát hiện đối tượng rất khó để có thể phát hiện được đối tượng bị che khuất đằng sau. Do đó, trong trường hợp các đối tượng bị che phủ, cần phải cải tiến Kalman Filter để mang lại hiệu quả cao
- Các đặc trưng rút trích được chưa thật sự kết hợp với nhau một cách hiệu quả để quá trình so trùng đạt kết quả cao.

5.3. HƯỚNG MỞ RỘNG

Đề tài này chỉ dừng lại ở việc gán nhãn cho đối tượng di chuyển qua nhiều camera có vùng không gian trùng lấp vì các giới hạn về thời gian hiện thực, phương pháp rút trích và kết hợp các đặc trưng của đối tượng chưa mang lại hiệu quả cao. Mục tiêu mà đề tài muốn hướng tới là giải quyết được bài toán truy vết đối tượng di chuyển qua nhiều camera có vùng không gian trùng lấp và không có vùng không gian trùng lấp nên tiếp theo sau tôi sẽ cải thiện việc rút trích đặc trưng sinh trắc của đối tượng di chuyển đảm bảo tính định danh cho đối tượng khi đối tượng xuất hiện ở những camera khác nhau. Đồng thời có thể giải quyết được bài toán trong thời gian thực. Tìm kiếm phương pháp phát hiện đối tượng di chuyển vừa có tính chính xác cao, vừa có thời gian tính toán và thực thi nhanh để cải thiện tốc độ xử lý của phương pháp đề xuất.

TÀI LIỆU THAM KHẢO

1. Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “*You Only Look Once: Unified, Real-Time Object Detection*”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp: 779-788, 2016.
2. Joseph Redmon, Ai Farhadi, “*YOLO9000: Better, Faster, Stronger*”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 6517-6525, 2017.
3. Lowe, David G, “*Distinctive image features from scale-invariant keypoints*”, International journal of computer vision, Vol 60, number 2, pp. 91-110, 2004.
4. Dipen Narendra Dalal, Bill Triggs, “*Histograms of oriented gradients for human detection*”, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), IEEE Conferences, Vol 1, pp. 886-893, 2005.
5. Gedas Bertasius, Jianbo Shi, Lorenzo Torresani, “*DeepEdge: A multi-sccale bifurcated deep network for top-down contour detection*”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 4380-4389, 2015.
6. Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, Zhijiang Zhang, “*DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection*”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 3982-3991, 2015.
7. A. Krizhevsky, I. Sutskever, G. E. Hinton, “*Imagenet classification with deep convolutional neural networks*”, Advances in neural information processing system, pp. 1097-1105, 2012.
8. Dumitru Erhan, Christian Szegedy, Alexander Toshev, Dragomir Anguelov, “*Scalable Object Detection Using Deep Neural Networks*”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 2155-2162, 2014.

9. Shipra Ojha, Sachin Sakhare, “*Image processing techniques for object tracking in video surveillance – A survey*”, International Conference on Pervasive Computing (ICPC), IEEE Conferences, pp. 1-6, 2015.
10. Yan Yang, Xiaodong Wang, Jiande Wu, Haitang Chen, Zhaoyuan Han, “*An improved mean shift object tracking algorithm based on ORB feature matching*”, The 27th Chinese Control and Decision Conference (CCDC), IEEE Conferences, pp. 4996-4999, 2015.
11. Rosten E, Drummond T, “*Fusing points and lines for high performance tracking*”, Tenth IEEE International Conference on Computer Vision (ICCV’05), IEEE Conferences, Vol 2, pp.1508-1515, 2005.
12. Michael Calonder, Vincent Lepetit, Christoph Strecha, Pascal Fua, “*Brief: binary robust independent elementary features*”, European Conference on Computer Vision (ECCV), Springer, pp. 778-792, 2010.
13. Jong-Min Jeong, Tae-Sung Yoon, Jin-Bae Park, “*Kalman filter based multiple objects detection-tracking algorithm robust to occlusion*”, SICE Annual Conference (SICE), IEEE Conferences, pp. 941-946, 2014.
14. Soonmin Hwang, et al, “*Fast multiple objects detection and tracking fusing color camera and 3D LIDAR for intelligent vehicles*”, Ubiquitous Robots and Ambient Intelligence (URAI), IEEE Conferences, pp. 234-239, 2016.
15. Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, Philip Torr, “*BING: Binarized Normed Gradients for Objectness Estimation at 300fps*”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 3286-3293, 2014.
16. C. Lawrence Zitnick, Piotr Dollr, “*Edge boxes: Locating object proposals from edges*”, European Conference on Computer Vision (ECCV), Springer, pp. 391-405, 2014.
17. Philipp Krhenbhl, Vladlen Koltun, “*Geodesic object proposals*”, European Conference on Computer Vision (ECCV), Springer, pp. 725-739, 2014.

18. Jasper RR Uijlings, Koen E. A. van de Sande, Theo Gevers, Arnold W. M. Smeulders, “*Selective search for object recognition*”, International Journal of Computer Vision (IJCV), Springer, Vol 104, Number 2, pp. 154-171, 2013.
19. Ross Girshick, “*Fast R-CNN*”, IEEE International Conference on Computer Vision (ICCV), IEEE Conferences, pp. 1440-1448, 2015.
20. Latha Anuj, M T Gopala Krishna, “*Multiple camera based multiple object tracking under occlusion: A survey*”, 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE Conferences, pp. 432-437, 2017.
21. Wei Li, Rui Zhao, Tong Xiao, Xiaogang Wang, “*DeepReID: Deep Filter Pairing Neural Network for Person Re-identification*”, Computer Vision and Pattern Recognition (CVPR), IEEE Conferences, pp. 152-259, 2014.
22. Ristani, Ergys, et al. “*Performance measures and a data set for multi-target, multi-camera tracking.*”, European Conference on Computer Vision (ECCV), Springer, pp. 17-35, 2016.
23. Francois Fleuret, Jerome Berclaz, Richard Lengagne, Pascal Fua, “*Multicamera People Tracking with a Probabilistic Occupancy Map*”, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Journals and Magazines, Vol. 30, Number 2, pp. 267-282, 2008.
24. <http://www.lalung.vn/upload/images/1-toan-canh-ngon-nui.jpg>. (Last accessed 01 june 2018)
25. <https://www.hund-und-herrchen.de/bilder/webseite/berichte/hund-am-strand-k.jpg>. (Last accessed 01 june 2018)