

Chương 1 TỔNG QUAN

1.1 Động lực nghiên cứu

Trong các hệ thống thông minh dựa vào video, việc phát triển khả năng nhận thức của hệ thống là một trong những nhiệm vụ quan trọng và thách thức nhất. Con người có thể hiểu được nội dung của video rất dễ dàng với sự thay đổi của không gian - thời gian và tri thức mà họ được trao dồi trong quá trình học tập, nghiên cứu và làm việc, nhưng để máy tính có thể hiểu được nội dung trong đoạn video là một vấn đề thách thức trong lĩnh vực TGMT. Mặc dù đã có nhiều nghiên cứu về phương pháp và thuật toán hướng tới phát triển và hoàn thiện khả năng nhận thức của máy nhưng vẫn chưa có thuật toán nào hiệu quả và bền vững để làm cho máy có khả năng suy nghĩ và hoạt động một cách hợp lý, linh hoạt để thích nghi với môi trường xung quanh chúng như là con người đã làm. Đối với khả năng nhận thức của máy, nhận dạng hành vi người được xem là một nhiệm vụ khó khăn và phức tạp nhất. Điều này bởi vì biểu diễn hành vi người là cực kỳ đa dạng và đáng diệu của người cũng có một số lượng lớn các bậc tự do. Ví dụ, cùng một hoạt động nhưng hai người khác nhau có thời gian hoàn thành là khác nhau hoặc trong hoạt động uống nước một người có thể sử dụng hai tay để bưng cốc nước nhưng người khác có thể không. Điều này giải thích tại sao nhiều phương pháp có kết quả tốt trên tập dữ liệu huấn luyện nhưng lại thất bại khi ứng dụng vào tập dữ liệu khác hoặc các ứng dụng thực tế. Vì vậy, phát triển và tăng độ chính xác của bài toán nhận dạng hành vi còn cần rất nhiều nỗ lực của các nhà nghiên cứu trên toàn thế giới. Mỗi đóng góp dù rất nhỏ cũng có ý nghĩa to lớn trong việc mang robot đến gần hơn với cuộc sống của con người và tăng cường khả năng hiểu của các hệ thống thông minh dựa vào video cũng như làm rõ hơn về hệ thống nhận thức của con người.

Giải quyết bài toán nhận dạng hành vi người ngày càng trở nên cấp thiết trong cả khía cạnh lý thuyết và thực tế. Từ khía cạnh thực tế, một hệ thống có thể phân tích hành vi có thể giúp xây dựng một chuyên gia trong một miền ứng dụng hoặc giúp phát hiện những mẫu hành vi trong các hệ thống giám sát, hệ thống tương tác người – máy hoặc robot. Từ khía cạnh lý thuyết, bài toán này có thể trả lời các câu hỏi: ***“Các khái niệm được trình bày như thế nào trong các hệ thống thông minh dựa vào video?”***. Ngoài ra, nó cũng giúp bộc lộ những hiểu nhầm của con người về hệ thống nhận thức của não người hoặc cũng có thể xác minh các giả thuyết về hoạt động của não người.

1.2 Mục tiêu

Mục tiêu của luận án là nghiên cứu, hệ thống hóa và cải tiến một số phương pháp cho bài toán nhận dạng hành vi người trong video. Để đạt được mục tiêu chung của luận án, các mục tiêu chính của luận án có thể được mô tả như sau:

- Khảo sát các nghiên cứu liên quan đến bài toán hành vi người trong video để cho thấy hiện trạng về sự tiến triển các nghiên cứu trên thế giới và ứng dụng của hành vi người trong video.
- Nghiên cứu và đề xuất mô hình nâng cao hiệu quả cho bài toán nhận dạng hành vi trong video.

1.3 Đóng góp của luận án

Dựa trên các kết quả nghiên cứu và các công trình đã được công bố. Luận án đã hệ thống hóa và cải tiến một số phương pháp cho bài toán hành vi người theo hướng độ phức tạp tăng dần với sự tiến hóa của các bài toán con là hành động và hoạt động. Các đóng góp chính của luận án có thể được tổng kết như sau:

Thứ nhất, đối với bài toán nhận dạng hành động, luận án tiến đề xuất mô hình nhận dạng hành động sử dụng thông tin từ 2 nguồn dữ liệu màu và độ sâu. Mô hình rút trích đặc trưng cục bộ dựa vào STIP và sử dụng các bộ miêu tả đặc trưng trên cả 2 nguồn dữ liệu. Mỗi hành động được biểu diễn bởi mô hình gán trọng số mềm GMM kết hợp Kmeans++ và kỹ thuật phân đoạn video để biểu diễn mối liên hệ thời gian trong video. Cuối cùng, mô hình học đa nhân được sử dụng để xây dựng bộ trọng số tổng hợp các nguồn đặc trưng khác nhau để tăng hiệu quả nhận dạng. Thực nghiệm các phương pháp đề xuất trên 2 tập dữ liệu chuẩn quốc tế 3D Action Pairs và UT-Kinect Action. Các đóng góp chi tiết trong mô hình nhận dạng hành động người được công bố trong các nghiên cứu [CT02], [CT03], [CT04], [CT06], [CT08], [CT09] và [CT10].

Thứ hai, đề xuất mô hình nhận dạng hoạt động trong môi trường sinh hoạt hàng ngày dựa trên biểu diễn chuỗi hoạt động con. Mô hình này tận dụng sức mạnh của màu, độ sâu và khung xương trong việc rút trích đặc trưng không gian – thời gian biểu diễn hình dáng, ngữ cảnh và tương tác trong và ngoài của người khi thực hiện hoạt động. Mô hình HMM được sử dụng để mô biểu diễn sự tiến hóa của hoạt động theo chuỗi thời gian và HCRF để tăng hiệu quả phân lớp hoạt động. Thực nghiệm các phương pháp đề xuất trên 2 tập dữ liệu chuẩn quốc tế CAD120 và MSRDailyActivity3D. Các đóng góp này được thể hiện trong các nghiên cứu đã được công bố [CT01], [CT02], [CT04] và [CT06].

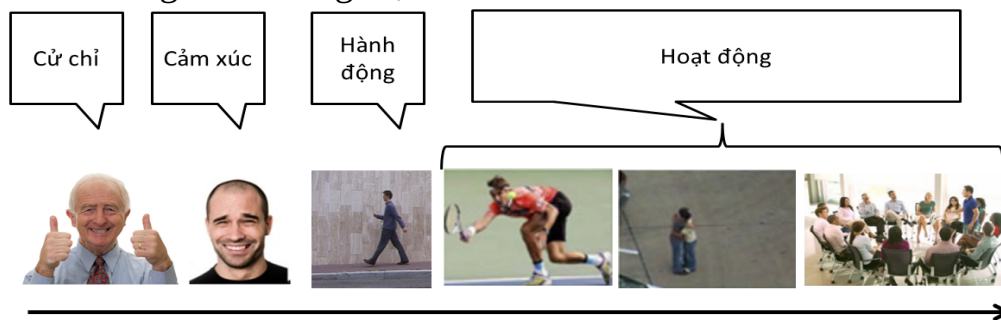
Thứ ba, đề xuất mô hình nhận dạng hoạt động trong môi trường không ràng buộc dựa trên bản đồ đặc trưng không gian – thời gian và học chuyển tiếp mô hình mạng

học sâu. Mỗi hoạt động được biểu diễn bởi một tập các đặc trưng tĩnh và động từ mạng VGG được học chuyển đổi từ bộ trọng của mạng VGG được học từ ImageNet. Cuối cùng, phương pháp Naïve Bayes Nearest Neighbor (NBNN) được sử dụng để tăng hiệu quả nhận dạng cũng như tính dễ dàng mở rộng khi lượng dữ liệu và số lớp gia tăng. Thực nghiệm các phương pháp đề xuất trên 2 tập dữ liệu chuẩn quốc tế UCF101 và HMDB51. Các đóng góp này được thể hiện trong các nghiên cứu đã được công bố [CT03], [CT05], và [CT07].

Chương 2 CÁC NGHIÊN CỨU LIÊN QUAN VÀ HƯỚNG TIẾP CẬN CỦA LUẬN ÁN

2.1 Bài toán nhận dạng hành vi người

Hành vi người được xem là mẫu chuyển động của người trong khoảng thời gian nhất định [2]. Việc xem xét hành vi của người dựa trên quá trình thực hiện của các bộ phận trên cơ thể người cũng như quá trình tương tác khi thực hiện hành vi. Bài toán hành vi người có thể được xem xét và tổng hợp gồm 4 đối tượng: cử chỉ, cảm xúc, hành động và hoạt động. Mỗi bài toán thể hiện một khía cạnh trên cơ thể người trong quá trình thực hiện các hành vi. Mức độ phức tạp của bài toán hành vi được phân cấp theo **hình 2-1**. Ngoài ra, bài toán nhận dạng hành vi được xem là bài toán nhận dạng mẫu trong lĩnh vực khoa học máy tính với đầu vào là video. Vì vậy, nó thừa hưởng những tính chất chung của mô hình nhận dạng mẫu và bổ sung các giai đoạn để thích hợp với bài toán hành vi người trong video. Luận án tập trung nghiên cứu và phát triển 2 bài toán con là hành động và hoạt động để làm trường hợp minh chứng cho các thuộc tính mang tính bản chất của bài toán hiểu hành vi người. Đồng thời, luận án cũng tiến hành thực nghiệm các mô hình khác nhau để đánh giá khả năng ứng dụng của các mô hình trong môi trường thực tế.



Hình 2-1. Các cấp độ của các bài toán trong nhận dạng hành vi người

2.2 Lịch sử bài toán nhận dạng hành vi người

Bài toán nhận dạng hành vi đã có một quá trình lịch sử phát triển lâu đời với nhiều công trình nghiên cứu đã được công bố cũng như những ứng dụng đã triển khai trong thực tế đã được làm rõ trong các nghiên cứu khảo sát [1, 2, 3, 56, 82, 110]. Bài toán nhận dạng hành vi người được hình thành vào những năm 1970 [56], khi nó thu hút sự quan tâm nghiên cứu của cộng đồng khoa học bởi vì nó có một mảng rộng các ứng dụng tiềm năng như y khoa, tương tác người – máy, robot trợ giúp, thành phố thông minh, nhà thông minh, trò chơi tương tác... Lược sử tóm tắt các sự kiện chính tác động đến bài toán nhận dạng vi người được tóm lược trong **bảng 2-1** mở rộng từ nghiên cứu [56].

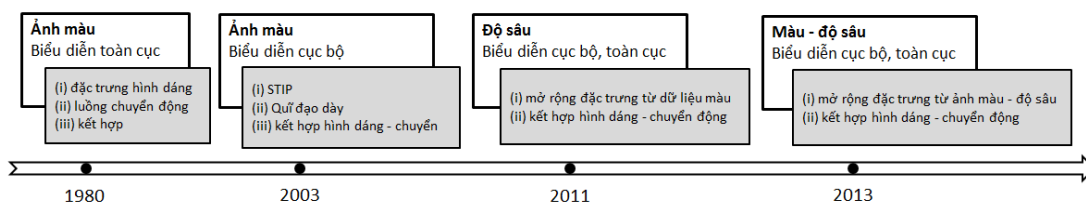
Năm	Tóm tắt các nghiên cứu	Đối tượng chính
1973	Thực nghiệm của Johansson gắn đèn chiếu vào các khớp của con người. Biểu diễn mỗi bộ phận của cơ thể bởi hai điểm và xác định cấu trúc theo phép chiếu trực giao.	Bộ phận cơ thể người
1982	Tập dữ liệu Hoffman người đàn ông di chuyển với 6 điểm. Đây là sự khởi đầu của việc ước lượng cấu trúc và nhận dạng hành vi của đối tượng liên kết.	Di chuyển
1982 đến 2010	Tập trung nghiên cứu phát triển khả năng nhận dạng hành vi người trong những điều kiện khác nhau trên dữ liệu ảnh màu và đặc trưng được thiết kế.	Tập trung vào cử chỉ, cảm xúc và hành động
2011 đến 2013	Các nghiên cứu dựa vào đặc trưng thiết kế và trên nhiều nguồn dữ liệu khác nhau như ảnh màu, độ sâu và khung xương. Một số nghiên cứu đã ứng dụng mạng học sâu vào bài toán nhận dạng hành vi để tăng cường độ chính xác của bài toán.	Hành động và hoạt động
2014 đến nay	Các nghiên cứu tập trung vào việc khai thác mạng học sâu CNN trong biểu diễn video. Đồng thời, kết hợp đặc trưng thiết kế và đặc trưng học sâu để tăng hiệu quả nhận dạng hành vi.	Hành động và hoạt động

Bảng 2-1. Tóm lược quá trình phát triển của bài toán nhận dạng hành vi người

2.3 Các nghiên cứu liên quan

2.3.1 Đặc trưng thiết kế

Quá trình phát triển các nghiên cứu đặc trưng thiết kế có thể được minh họa trong hình 2-2. Đối với đặc trưng thiết kế có thể chia thành 2 nhóm con: đặc trưng cục bộ và đặc trưng toàn cục.



Hình 2-2. Các cột mốc chính của quá trình tiến hóa các phương pháp rút trích đặc trưng thiết kế và dữ liệu trong bài toán nhận dạng hành vi người

Các phương pháp tiếp cận đặc trưng toàn cục tập trung mô tả toàn bộ vùng dữ liệu ảnh hoặc thể tích không gian - thời gian[17, 112, 44, 95]. Vùng trọng yếu thường được xác định thông qua phương pháp trừ nền hoặc theo vết. Cách biểu diễn này thì hiệu quả bởi vì chúng mã hóa nhiều thông tin. Tuy nhiên, các cách tiếp cận này nhạy cảm với sự thay đổi của khung nhìn, sự xuất hiện của đối tượng, nhiễu và che khuất. Ngược lại, các tiếp cận dựa vào đặc trưng cục bộ dựa trên ý tưởng các vùng thị giác nổi bật. Những cách tiếp cận này biểu diễn quan sát như là một tập hợp các mô tả cục bộ hoặc mẫu. . Một số lượng lớn các nghiên cứu như HOG3D [9], HOG/HOF [43], SURF3D [33] mở rộng của SURF [16], 3D-SIFT[91] mở rộng của SIFT [26], LTP [75] và quỹ đạo dày với các bộ mô tả HOG/HOG/MHB[37, 38, 42] để rút trích các đặc trưng nổi bật như hình dáng, biên cạnh, góc và chuyển động. Cùng với đó là sự ra đời của các thiết bị thu nhận dữ liệu 3 chiều đã tạo thêm kênh dữ liệu giúp rút trích thông tin biểu diễn hiệu quả hình dáng và chuyển động của người [52, 69, 71, 73, 82, 92, 93, 100, 111, 115, 117]. Để tận dụng sức mạnh của từng loại đặc trưng một số nghiên cứu đã tiến hành kết hợp các dạng đặc trưng khác nhau trong quá trình rút trích và biểu diễn nội dung của video. Tuy nhiên, mỗi hướng tiếp cận đều tồn tại những vấn đề cần giải quyết để có thể biểu diễn hiệu quả hình dáng và chuyển động của hành vi người (bảng 2-2).

Đặc trưng	Ưu điểm	Khuyết điểm
Toàn cục	- Chứa nhiều thông tin - Mô tả được cấu trúc không gian - thời gian của hành vi trong video	- Phụ thuộc vào tiền xử lý như trừ nền, phát hiện đối tượng... - Nhạy cảm với sự biến động của hình nền và che khuất
Cục bộ	- Đơn giản, hiệu quả	- Lỗi lượng hóa giả từ

	- Bền vững sự biến động của hình nền và che khuất	- Phụ thuộc vào bộ phát hiện điểm trọng yếu - Không biểu diễn được tương quan không gian và thời gian trong cấu trúc hành vi
--	---	---

Bảng 2-2. Bảng so sánh đặc trưng toàn cục và cục bộ trong biểu diễn hành vi

2.3.2 Đặc trưng học

Khác với phương pháp dựa trên tri thức con người trong việc rút trích đặc trưng để biểu diễn hành vi người. Phương pháp này dựa trên việc thiết kế mạng và các bộ lọc để tự động trích chọn đặc trưng từ dữ liệu thô. Các phương pháp cập đặc trưng học có thể phân thành 2 nhóm chính: i) không dựa vào mạng học sâu; ii) dựa vào mạng học sâu. Các phương pháp không dựa vào mạng học sâu tập trung vào việc sử dụng học tự diễn, học dựa vào kỹ thuật lập trình tiến hóa để tìm kiếm đặc trưng biểu diễn hành vi một cách hiệu quả. Học từ điển là phương pháp dựa trên học biểu diễn dựa trên trình bày thừa dữ liệu đầu vào. Ngoài ra, lập trình tiến hóa là một kỹ thuật tiến hóa được lấy cảm hứng từ quy trình tiến hóa tự nhiên. Nó thường sử dụng để giải những bài toán mà không có tri thức biết trước của giải pháp cần hiện thực hóa. Các nghiên cứu điển hình như sử dụng giải thuật tiến hóa của Liu [73] và mô hình HMAX [40, 41] với các bộ lọc khác nhau kết hợp trên miền không gian và thời gian trong biểu diễn nội dung của video. Ngược lại, các phương pháp dựa vào mạng học sâu tập trung vào việc thiết kế cấu trúc mạng để chuyển đổi thông tin dữ liệu thô ban đầu thành ngữ nghĩa của hành vi người [66, 106, 78, 99, 21, 46, 39]. Một trong những điểm yếu của phương pháp học sâu này là cần rất nhiều dữ liệu học và tốn rất nhiều thời gian học để học mô hình đi kèm theo đó là một hệ thống máy tính được trang bị với chi phí cao. Ngoài ra, việc thiết kế kiến trúc mạng như thế nào để đạt được hiệu quả cao cũng là một trong những vấn đề cần thảo luận nhiều hơn [114]. Đến thời điểm hiện tại cũng chưa có một qui luật chung nào để chỉ ra rằng làm thế nào có thể thiết kế mạng học sâu hiệu quả. Đặc biệt là mạng học sâu vẫn được xem như là một hộp đen trong khi ứng dụng vào thực tế. Điều này làm cho việc rất khó giải thích kết quả trong trường hợp kiến trúc được thiết kế không đạt được kết quả như mong muốn của người thiết kế kiến trúc mạng. Đồng thời, việc tiến hành mở rộng số lớp hoặc bổ sung dữ liệu học thì phải tiến hành học lại toàn bộ kiến trúc mạng nhưng kết quả của việc học này có giúp tăng hiệu quả của mô hình hay không thì vẫn chưa có câu trả lời chính xác. Đây là một trong những điểm yếu quan trọng của phương pháp này mà cần được xem xét kỹ lưỡng khi ứng dụng vào một bài toán thực tế. Trong thời gian gần đây, đã có nghiên cứu tiến hành kết hợp đặc trưng học sâu và đặc trưng được thiết kế [77] để dụng sức mạnh hỗ trợ lẫn nhau nhằm giảm chi phí cho việc chỉ học đặc trưng từ dữ liệu thô đầu vào bằng cách tận dụng hiệu quả của tri thức biết trước trong đặc trưng thiết kế.

2.3.3 Mô hình phân lớp

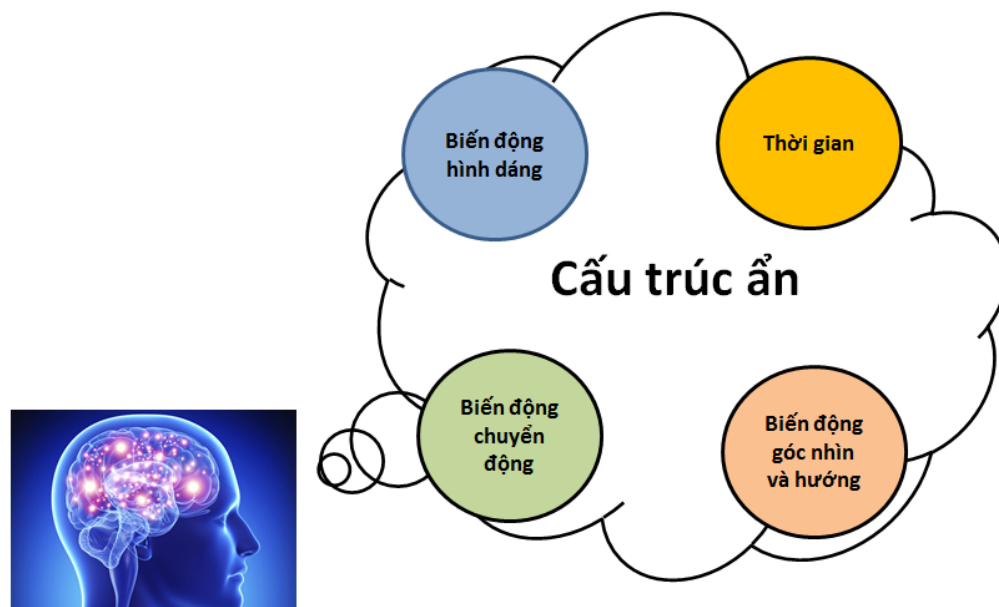
Trong bài toán nhận dạng mẫu tổng quát cũng như bài toán nhận dạng hành vi, bước cuối cùng toàn bộ quá trình trình chuyển từ thông tin thị giác thành ngữ nghĩa là một mô hình xác định được nhãn của mẫu dựa trên các véc tơ đặc trưng được rút trích trong các bước trung gian để đại diện cho hành vi. Các hướng nghiên cứu chính của bài toán phân lớp mẫu có thể được chia thành hai loại mô hình phân lớp [87]: i) mô hình học tạo sinh; ii) mô hình học dị biệt. Hướng tiếp cận học tạo sinh học mô hình phát sinh dữ liệu bởi ngữ cảnh bằng cách ước lượng dựa vào giả định và phân bố của các mô hình. Sau đó sử dụng mô hình này để dự đoán dữ liệu không nhìn thấy trước đó, bởi vì hướng tiếp cận này giả định mô hình đã được học nắm bắt được mô hình thực tế. Ngược lại, phương pháp học dị biệt cố gắng chỉ xây dựng một mô hình dựa trên dữ liệu được quan sát. Nó làm cho các giả định ít hơn trên các phân phối nhưng phụ thuộc rất nhiều vào chất lượng của dữ liệu. Hai câu hỏi quan trọng khi nghiên cứu sử dụng mô hình dị biệt là: i) Dữ liệu có đại diện cho bài toán?, ii) Có nhiều dữ liệu hay không?. Mỗi loại mô hình phân lớp có những điểm mạnh và điểm yếu riêng được tóm lược trong **bảng 2-3**, vì vậy việc sử dụng mô hình còn phải phụ thuộc vào điều kiện yêu cầu ứng dụng để có thể chọn một mô hình thích hợp nhất trong các mô hình trên.

Mô hình	Ưu điểm	Khuyết điểm	Ví dụ
Tạo sinh	- Học nhanh - Dễ dàng mở rộng số lớp	- Độ chính xác thấp hơn mô hình dị biệt - Không phụ thuộc quá nhiều vào dữ liệu - Duyệt qua các bộ phân lớp cho mỗi lần phân lớp mẫu mới	- GMM, Bayes, HMM, NBNN ...
Dị biệt	- Học chậm - Độ chính xác cao	- Học lại toàn bộ khi thêm mẫu hoặc thêm lớp mới - Phụ thuộc chất lượng dữ liệu	- ANN, SVM, hồi qui logic, CRF ...

Bảng 2-3. Bảng so sánh phương pháp phân lớp tạo sinh và dị biệt

2.4 Hướng tiếp cận của luận án

Dựa vào các phân tích và đánh giá trên, luận án đi đến đề xuất hướng tiếp cận bằng cách nghiên cứu để làm rõ bản chất bài toán hành vi người dựa trên nguyên lý chia đề trị và giả thuyết 2 dòng vỏ não thị giác trong việc biểu diễn hình dáng và chuyển động [40]. Cụ thể hóa các lý thuyết này bằng các mô hình cụ thể cho hai bài toán con là hành động và hoạt động trong bài toán hành vi người như sau:

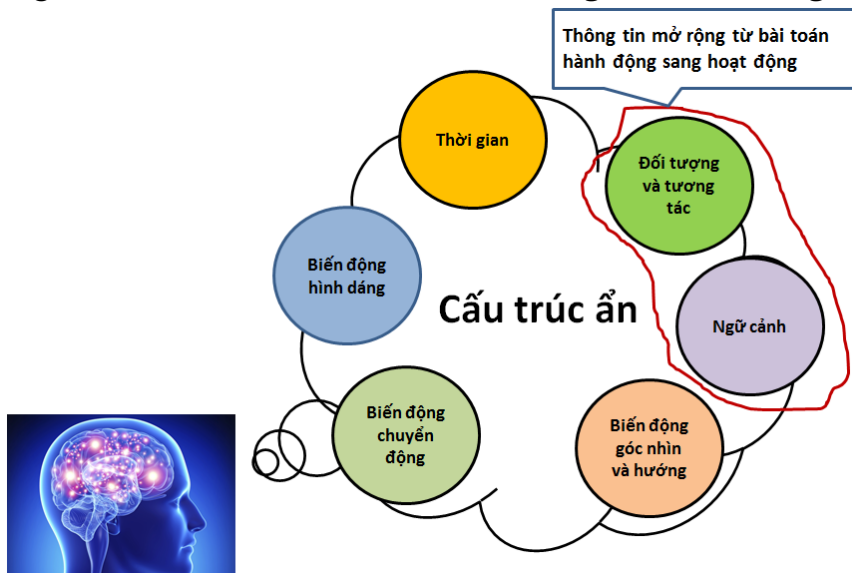


Hình 2-3. Cấu trúc nhận thức hành động ở người dựa vào hệ thống thông tin thị giác trong video

- **Mô hình 1:** Nhận dạng hành động người dựa trên dữ liệu màu – độ sâu. Mô hình này áp dụng lý thuyết đặc trưng cục bộ để giải quyết cho cả bài toán hành động và hoạt động. Luận án rút trích đặc trưng cục bộ biểu diễn hình dáng và chuyển động của hành động từ dữ liệu màu, độ sâu và kết hợp màu – độ sâu. Mô hình học phân lớp dựa trên việc học đa nhân để xây dựng bộ trong số tin cậy từ các nguồn đặc trưng khác nhau để tăng hiệu quả của nhận dạng. Bộ trọng số kết hợp đặc trưng này được xem như là phần mô tả trong cấu trúc ẩn của vỏ não trong quá trình nhận thức thông tin thị giác. Mô hình này giúp giải thích được cơ chế hiểu hành động trong video sử dụng thông tin thị giác ở não người được tổng hợp như **hình 2-3**.
- **Mô hình 2:** Nhận dạng hành động người trong sinh hoạt hàng ngày. Mô hình này dựa trên việc bổ sung các tính chất của mô hình nhận dạng hoạt động người (hình 2-2) để đáp ứng được tính mở rộng cũng như đặc thù của bài toán hoạt động người trong môi trường sinh hoạt hàng ngày. Luận án đề xuất xác định đối tượng thực hiện hoạt động dựa trên dữ liệu khung xương. Trên cơ sở này rút trích đặc trưng hình dáng, chuyển động, ngữ cảnh cũng như tương tác cục bộ - toàn cục của người và đối tượng tương tác dựa vào thông tin rút trích từ các khớp nối khung xương. Đồng thời, một hoạt động sẽ được xem xét như là một tập hợp các hoạt động con có mối quan hệ thời gian trong quá trình thực hiện hoạt động. Một mô hình phân lớp kết hợp bởi mô hình tạo sinh và dị biệt để tăng hiệu quả của nhận dạng. Mô hình này

giúp giải thích được cơ chế hiểu hoạt động trong video sử dụng thông tin thị giác ở não người được tổng hợp như **hình 2-4**.

- **Mô hình 3:** Nhận dạng hoạt động trong môi trường không ràng buộc. Đây là bài toán thách thức nhất trong bài toán phân tích và hiểu hành vi người. Trong bài toán này, việc làm sao để biểu diễn được hoạt động chứa các biến động đa dạng không tuân theo các mẫu cấu trúc không gian cũng như biến động theo thời gian. Vì vậy, việc áp dụng mô hình 2 vào dạng dữ liệu không ràng buộc là không thích hợp. Để vượt qua thách thức to lớn này, luận án dựa trên giả thuyết 2 dòng vỏ não thị giác [40] làm nền tảng trong quá trình rút trích đặc trưng để nhận thức được nội dung trong video. Hai luồng nhận thức này được mô hình hóa bằng đặc trưng tĩnh dựa trên khung hình chính và đặc trưng động dựa trên mẫu chuyển động dày và mẫu lịch sử chuyển động. Các biến động lớn trong bài toán này được rút trích và biểu diễn bởi mạng học sâu đã chứng minh hiệu quả trong các bài toán nhận dạng ảnh và video. Cũng như cách tiếp cận hoạt động con độc lập được phân lớp bởi mô hình NBNN để có thể vượt qua các thách thức về tính hỗn độn của dữ liệu thu nhận được. Mô hình này giúp giải thích được cơ chế hiểu hoạt động trong video sử dụng thông tin thị giác ở não người được tổng hợp như hình 2-4 tương tự như mô hình nhận thức hoạt động sinh hoạt hàng ngày.



Hình 2-4. Cấu trúc nhận thức hoạt động ở người dựa vào hệ thống thông tin thị giác trong video

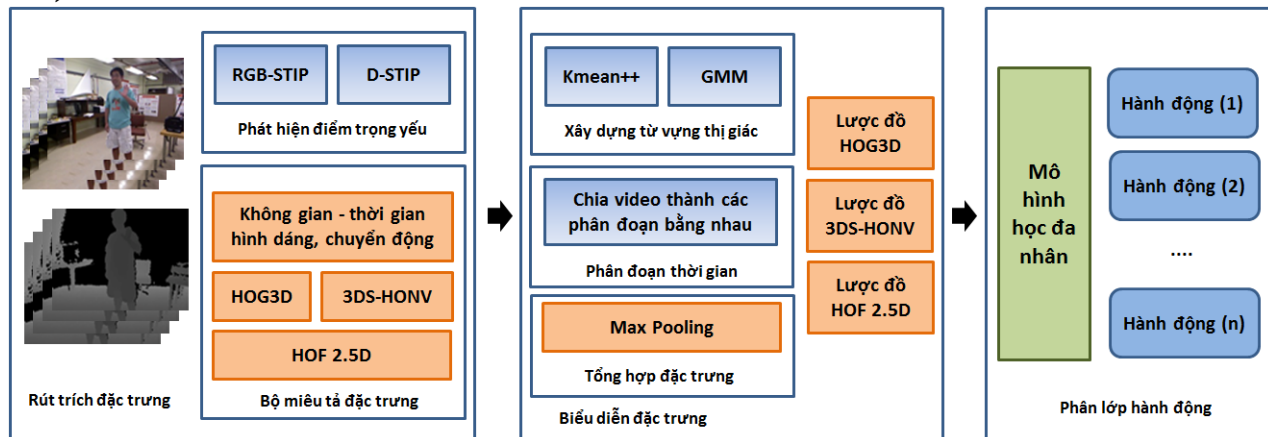
Chương 3 MÔ HÌNH NHẬN DẠNG HÀNH ĐỘNG

3.1 Giới thiệu

Nhận dạng hành động người là bài toán được xem như là bước trung gian để làm rõ yếu tố tác động đến biểu diễn hiệu quả hình dáng và chuyển động của người trong video. Các kết quả nghiên cứu trên mô hình này có ý nghĩa to lớn trong việc tiến đến nhận dạng mức cao hơn trong bài toán nhận dạng hành vi là hoạt động người. Trong nghiên cứu này, luận án đề xuất mô hình hình nhận dạng dựa trên việc tổng hợp các dạng đặc trưng cục bộ khác nhau trong việc biểu diễn hành động từ ảnh màu và độ sâu. Mô hình học đa nhận được ứng dụng để tạo bộ trong số kết hợp các đặc trưng này để nâng cao hiệu quả nhận dạng của hệ thống.

3.2 Mô hình đề xuất

Các nghiên cứu trong bài toán nhận dạng hành động tập trung vào việc làm sao biểu diễn được thay đổi của hình dáng và chuyển động trong video theo không gian và thời gian. Trong nghiên cứu này, luận án đề xuất mô hình dựa trên việc biểu diễn đặc trưng hình dáng và chuyển động của các đối tượng trong video dựa trên biểu diễn cục bộ không gian – thời gian tại các điểm nổi bật theo không gian – thời gian (**hình 3-1**).



Hình 3-1. Kiến trúc tổng quát của mô hình nhận dạng hành động người

- Giai đoạn 1: Phân video thành các đoạn bằng nhau để mã hóa thông tin chuyển đổi cục bộ - toàn cục trong video.
- Giai đoạn 2: Rút trích điểm đặc trưng nổi bật từ cả 2 kênh màu và độ sâu sử dụng thuật toán STIP là phương pháp mở rộng của Harris2D hay còn gọi là Harris3D.
- Giai đoạn 3: Tổng hợp các bộ miêu tả thành véc tơ đặc trưng biểu diễn cho mỗi phân đoạn trong video sử dụng mô hình GMM cải tiến với thuật toán khởi tạo ban đầu bằng Kmeans++.

- Giai đoạn 4: Sử dụng kỹ thuật max pooling để tổng hợp đặc trưng từ các phân đoạn trong video thành các véc tơ đặc trưng cuối cùng đại diện cho hành động.
- Giai đoạn 5: Huấn luyện mô hình học đa nhân để tổng hợp các đặc trưng từ các nguồn dữ liệu khác nhau để tăng hiệu quả phân lớp các đặc trưng được rút trích.

3.2.1 Rút trích đặc trưng

Phát hiện điểm trọng yếu là bước đầu tiên có ý nghĩa quan trọng trong các bài toán biểu diễn nội dung video dựa vào mẫu cục bộ. Toàn bộ video được phát hiện để thu hẹp vùng không gian tìm kiếm và trích chọn được những đặc trưng nổi bật nhất chứa nhiều thông tin về sự biến đổi về biên cạnh và sự chuyển động trong không gian-thời gian, bộ dò tìm điểm trọng yếu STIP là phương pháp mở rộng của Harris2D (hay còn gọi là Harris3D) được đề xuất bởi Laptev [44] sẽ được áp dụng trên cả 2 kênh dữ liệu màu và độ sâu.

Biểu diễn điểm trọng yếu là bước mô tả nội dung thị giác tại các điểm trọng yếu ở bước trước đó. Quá trình này sẽ sử dụng các phương pháp rút trích đặc trưng để mô tả được tính chất người trong quá trình thực hiện hành động bằng nội dung thị giác. Nó có ý nghĩa to lớn trong việc biểu diễn hành động mô tả nội dung thị giác của một hành động trong video. Luận án sử dụng 3 bộ miêu tả để biểu diễn cho thông tin của hoạt động từ kênh màu với HOG3D[9], độ sâu với 3DS-HONV [92] và kết hợp màu – độ sâu với HOF2.5D [92]. Các bộ miêu tả biểu diễn thông tin hình dáng và chuyển động của các vùng cục bộ trong video theo từng kênh dữ liệu và tổng hợp các kênh dữ liệu.

3.2.2 Biểu diễn hành động

Để biểu diễn hành động dựa vào đặc trưng cục bộ thì mô hình túi từ. Mô hình này gồm 2 bước chính: i) học các từ vựng thị giác bằng thuật toán gom nhóm; ii) biểu diễn đặc trưng cho hành động bằng cách xây dựng lược đồ từ vựng thị giác từ tham số mô hình ở bước trước. Các nghiên cứu trước đó mô hình BOW sử dụng thuật toán gom nhóm K-means để xây dựng từ điển thị giác, khi đó mỗi mẫu cục bộ chỉ được gán vào một mẫu đại diện. Việc gán trọng số nhị phân cho lược đồ xuất hiện của từ bằng cách cộng vào lược đồ 1 hoặc 0 cho sự xuất hiện hay vắng của một từ. Phương thức gán trọng số này dựa trên việc tìm kiếm trong từ điển từ gần nhất với đặc trưng của điểm trọng yếu. Nhiều nghiên cứu cho rằng việc gán trọng số chỉ dựa vào từ gần nhất không phải là lựa chọn tối ưu vì hai từ được gom nhóm gần nhau có thể thuộc vào 2 nhóm khác nhau khi kích thước từ điển tăng. Hơn nữa, hai điểm trọng yếu được

gán vào cùng một nhóm thì không nhất thiết tương đồng với nhóm được gán. Đồng thời, để tránh việc gán hai từ có cùng đóng góp như nhau trong biểu diễn video, nó không thể hiện được tầm quan trọng của các từ trong biểu diễn video. Luận án sử dụng mô hình GMM giúp đánh giá mức độ đóng góp của mẫu cục bộ vào quá trình biểu diễn dựa trên trọng số là xác suất ước lượng của dữ liệu trong quá trình học từ vựng thị giác qua mô hình GMM. Điều này giúp tăng hiệu quả biểu diễn hành động của mô hình đề xuất. Trong nghiên cứu này, để đánh khởi tạo ngẫu nhiên tham số ban đầu của mô hình GMM, luận án sử dụng thuật toán gom nhóm Kmeans++ [6] đã chứng minh được hiệu quả cho quá trình gom nhóm dữ liệu. Ưu điểm của phương pháp này tập trung vào giúp quá trình khởi tạo ban đầu hiệu quả hơn so với việc khởi tạo ngẫu nhiên các nhóm ban đầu bằng các tính khoảng cách khởi tạo nhóm để tránh rơi vào các mẫu dị biệt hoặc nhiều làm cho quá trình gom nhóm không hội tụ hoặc kết quả gom nhóm không tối ưu.

Thuật toán 3.1: Thuật toán GMM kết hợp Kmeans++ xây dựng từ vựng thị giác

Mục tiêu: xây dựng các vector trung bình, trọng số và ma trận hiệp phương sai đại diện cho mỗi nhóm được gom cụm

Đầu vào: Tập dữ liệu D , K là số nhóm cần gom

Đầu ra: μ_i, Σ_i, π_j đại diện cho mỗi nhóm thứ $i = 1 \dots K$

Begin

1// Áp dụng thuật toán Kmeans++ trên tập dữ liệu D

2// Tính μ_i, Σ_i, π_j dựa trên các nhóm gom được ở bước trên.

3// Áp dụng thuật toán GMM với trọng tâm khởi tạo ở bước trên

End

Một nhược điểm của mô hình giỏ từ là không mô tả được cấu trúc không thời gian trong quá trình thực hiện hoạt động. Để vượt qua thách thức này, luận án đề xuất sử dụng kỹ thuật phân đoạn thời gian trong video. Mỗi video sẽ được phân thành các đoạn có chiều dài bằng nhau. Mô hình tổng hợp đặc trưng GMM sẽ được sử dụng cho mỗi phân đoạn video tạo tập vector đặc trưng $F = \{f_i, i = 1 \dots n\}$ đại diện cho một hoạt động. Cuối cùng, kỹ thuật max pooling được sử dụng để tạo vector cuối cùng đại diện cho mỗi hoạt động:

$$f_v[j] = \max(f_i[j]), \text{ với } j = 1 \dots m, i = 1 \dots n \quad (3.1)$$

trong đó $f_i[j]$ là thành phần thứ j của vector đặc trưng f_i có chiều dài m , f_v là vector đặc trưng đại diện cho video.

Như vậy với việc rút trích đặc trưng từ kênh màu – độ sâu với các đặc trưng HOG3D, 3DS-HONV và HOF2.5D đại diện cho mỗi video. Một hành động sẽ được

đại diện bởi tập gồm 3 vec tơ đặc trưng $F = \{f_{HOG3D}, f_{3DS-HONV}, f_{HOF2.5D}\}$, các vec tơ đặc trưng này sẽ được đưa vào mô hình học phân lớp để xác định nhãn hoạt động.

3.2.3 Học đa nhân

Mô hình học đa nhân (MKL) [13] đã thu hút nhiều sự quan tâm nghiên cứu của cộng đồng TGMT và nhận dạng mẫu trong nhiều nghiên cứu trong bài toán ảnh và video. Ý tưởng chính của mô hình MKL là tận dụng sức mạnh của nhiều nhân và nhiều đặc trưng được kết hợp một cách tối ưu để tăng hiệu quả nhận dạng thay vì chỉ sử dụng một nhân trong mô phân lớp dựa vào học hàm nhân. Trong nghiên cứu này, luận án sử dụng mô hình SimpleMKL [13] dựa trên mô hình SVM [22] để kết hợp nhiều loại đặc trưng cùng với nhiều hàm nhân khác nhau tạo ra bộ trọng số kết hợp và tổng quát hóa các bộ phân lớp SVM. Thuật toán SimpleMKL được đề xuất dựa trên mô hình SVM bằng cách giải quyết bài toán tối ưu lồi sau:

$$\min_w \max_{\beta} L(W, \beta) = -\frac{1}{2} \beta^T \left(\sum_{m=1}^M w_m K_m \right) \beta + y^T \beta \quad (3.2)$$

$$\sum_{i=1}^N \beta_i = 0, \quad 0 \leq \beta_i y_i \leq C, \quad i = 1, 2, \dots, N \quad (3.3)$$

$$\sum_{m=1}^M w_m = 1, \quad w_m \geq 0, \quad i = 1, 2, \dots, M \quad (3.4)$$

Với $y = (y_1, y_2, \dots, y_N)^T$ và $\beta = (\beta_1, \beta_2, \dots, \beta_N)^T$ là các vec tơ hệ số Lagrange trong mô hình SVM, K_m là hàm nhân thứ m , M là số lượng hàm nhân và $W = (w_1, w_2, \dots, w_M)^T$ là trọng số của các hàm nhân.

Trong luận án này, nghiên cứu đề xuất sử dụng hàm nhân để kết hợp ba vec tơ đặc trưng được rút trích từ HOG3D, 3DSHONV và HOF2.5D tại mỗi cấp độ của hàm nhân, trong mô hình SimpleMKL được định nghĩa như sau:

$$K_{i,j} = \sum_{m=1}^M [w_m K_m(x_i^{HOG3D}, x_j^{HOG3D}) + w_{m+M} K_m(x_i^{3DSHONV}, x_j^{3DSHONV}) + w_{m+2M} K_m(x_i^{HOG2.5D}, x_j^{HOG2.5D})] \quad (3.5)$$

Trong thuật toán SimpleMKL (thuật toán 3.5), vòng lặp ngoài để giải bài toán tối ưu trong phương trình hàm mục tiêu lồi, vòng lặp con là để tối ưu bộ phân lớp và trọng số kết hợp các hàm nhân. Một mẫu mới $\{x_0^{HOG3D}, x_0^{3DSHONV}, x_0^{HOG2.5D}\}$, nhãn dự đoán y_0 có thể được xác định theo công thức sau:

$$y_0 = \text{sgn} \left[\sum_{i=1}^N \sum_{m=1}^M \beta_i (w_m K_m(x_i^{HOG3D}, x_0^{HOG3D}) + w_{m+M} K_m(x_i^{3DSHONV}, x_0^{3DSHONV}) + w_{m+2M} K_m(x_i^{HOG2.5D}, x_0^{HOG2.5D})) \right] \quad (3.6)$$

Thuật toán 3.5: Thuật toán SimpleMKL

Mục tiêu: xây dựng bộ trọng số phân lớp tương ứng với từng hàm nhân của bộ phân lớp SVM

Đầu vào: Tập dữ liệu D , M là số hàm nhân

Đầu ra: bộ trọng số w tương ứng cho từng nhân của bộ phân lớp SVM

Begin

1// Khởi tạo trọng số $w_m = \frac{1}{M}$

2// **Lặp** cho đến khi hội tụ

3// Tính toán hàm mục tiêu $J(w, \beta)$ (theo công thức 3.16) sử dụng hàm giải SVM trên tập dữ liệu D với $K = \sum_m w_m K_m$

4// Tính đạo hàm $\frac{\partial J}{\partial w_m}$ và hướng giảm bộ trọng W theo thuật toán rút gọn đạo hàm [13]

5// $J^* = 0, w^* = w, W^* = W, \mu = \text{argmax}_m w_m$

6// **Lặp** $J^* < J(w, \beta)$

7// $w = w^*, W = W^*$

8// $v = \text{argmin}_{\{m|w_m < 0\}} -w_m/W_m, \gamma_{max} = -w_v/W_v$

9// $w^* = w + \gamma_{max} W, W_\mu^* = W_\mu - W_v, W_v^* = 0$

10// Tính toán hàm mục tiêu J^* sử dụng hàm giải SVM với $K = \sum_m w_m^* K_m$

11// **Kết thúc lặp**

12// Duyệt dọc theo đường thẳng theo W với $\gamma \in [0, \gamma_{max}]$ (gọi hàm giải SVM cho mỗi trị thử γ)

13// $w = w + \gamma W$

14// **Kết thúc lặp**

End

3.3 Thực nghiệm

Luận án tiến hành thực nghiệm đánh giá mô hình trên 2 tập dữ liệu chuẩn quốc tế UTKinect-Action [79] và 3D Action Pairs [53]. Luận án sử dụng phương pháp học đa nhân với 2 hàm nhân là Gauss và hàm đa thực với nhiều mức độ tham số khác nhau và chiến lược một và tất cả được sử dụng để phân nhiều lớp. Kết quả nhận dạng đối với tập dữ liệu UTKinect-Action mô hình cao hơn so trước đó là 5.1% và đạt được kết quả tối ưu tương đương với nghiên cứu trước đó trên tập dữ liệu 3D Action Pair.

Ngoài ra, thực nghiệm đánh giá kỹ thuật phân đoạn trong video làm nền tảng cho việc ứng dụng kỹ thuật này vào các mô hình trong các chương tiếp theo.

Chương 4 MÔ HÌNH NHẬN DẠNG HOẠT ĐỘNG NGƯỜI TRONG SINH HOẠT HÀNG NGÀY

4.1 Giới thiệu

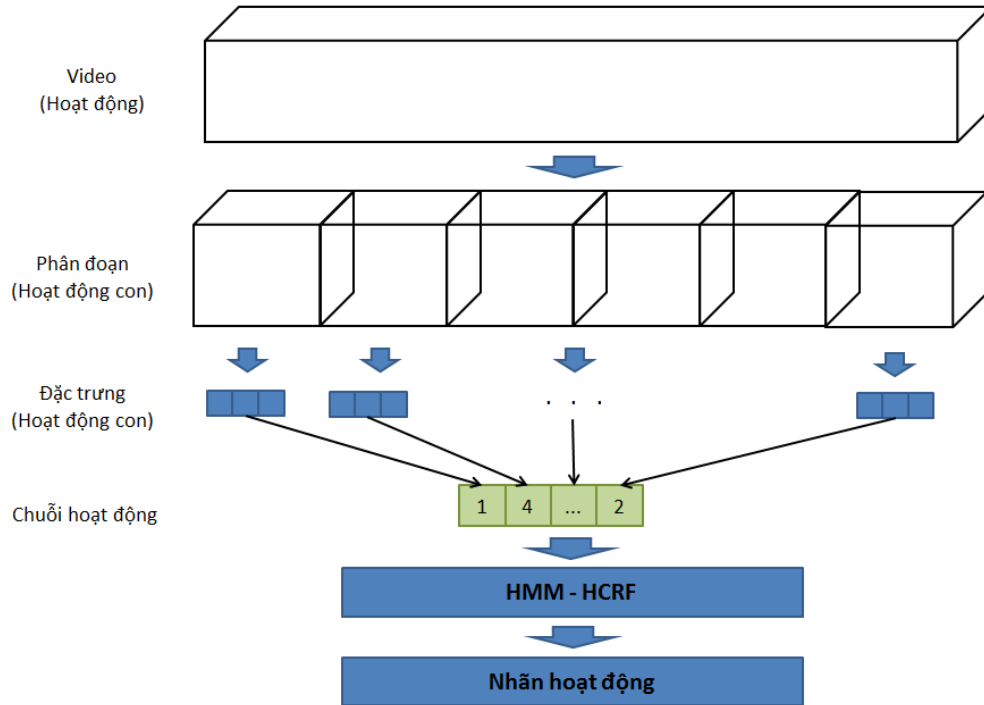
Nhận dạng hoạt động người được xem là mức cao nhất trong bài toán nhận dạng hành vi người với nhiều biến động phức tạp trong quá trình thực hiện hoạt động. Luận án đề xuất mô hình nhận dạng hoạt động người trong mô trường sinh hoạt hằng ngày. Với dạng hoạt động này, quá trình thực hiện hoạt động có sự phụ thuộc thời gian giữa các nội dung thể hiện trong video. Như vậy, việc biểu diễn quan hệ thời gian có ý nghĩa to lớn trong việc nâng cao hiệu quả của việc nhận dạng. Trong mô hình này, luận án xem mỗi hoạt động bao gồm một chuỗi các hoạt động con. Mỗi hoạt động con biểu diễn quá trình thay đổi hình dáng, chuyển động và ngữ cảnh tương tác trong một khoảng thời gian. Mỗi liên hệ theo thời gian giữa các hoạt động con này sẽ được mô hình hóa bởi các mô hình máy học được xem như là một cấu trúc ẩn của vỏ não thị giác khi tiếp nhận thông tin.

4.2 Mô hình đề xuất

Các nghiên cứu trong bài toán nhận dạng hoạt động trong sinh hoạt hàng ngày tập trung vào việc làm sao biểu diễn được sự tiến hóa của nội dung trong video. Nghiên cứu đề xuất chuyển nội dung video thành một chuỗi dữ liệu thời gian bằng cách phân video thành các đoạn bằng nhau được gọi là hoạt động con. Một hoạt động con là mô tả sự thay đổi về không gian thời gian của hình dáng, chuyển động và tương tác của các đối tượng trong phạm vi cục bộ. Tận dụng sức mạnh của camera 3D trong việc phát hiện vị trí người và các bộ phận với độ chính xác cao. Nghiên cứu đề xuất rút trích đặc trưng cả cơ thể lẫn các bộ phận để mô tả từ tổng quan đến chi tiết cũng như cấu trúc của cơ thể người trong quá trình thực hiện hoạt động. Ngoài ra, để mã hóa cấu trúc ẩn trong quá trình nhận thức hoạt động, nghiên cứu đề xuất sử dụng HMM và HCRF để tăng hiệu quả nhận dạng của hệ thống. Mô hình hệ thống đề xuất gồm các giai đoạn được thực hiện như **hình 4-1**:

- Giai đoạn 1: Phân video thành các đoạn bằng nhau thể hiện một nội dung của hoạt động trong một phân đoạn không gian - thời gian được gọi là hoạt động con.

- Giai đoạn 2: Rút trích đặc trưng cho các hoạt động con. Mỗi hoạt động con sẽ được rút trích để biểu diễn được 2 luồng nhận thức của vỏ não thị giác: i) luồng hình dáng; ii) luồng chuyển động.
- Giai đoạn 3: Biểu diễn hoạt động dưới dạng chuỗi thời gian sử dụng mô hình gom nhóm GMM cải tiến quá trình khởi tạo sử dụng thuật toán Kmeans++.
- Giai đoạn 4: Huấn luyện mô hình HMM và mô hình HCRF để nhận dạng hoạt động của video mới.



Hình 4-1. Kiến trúc tổng quát của mô hình đề xuất

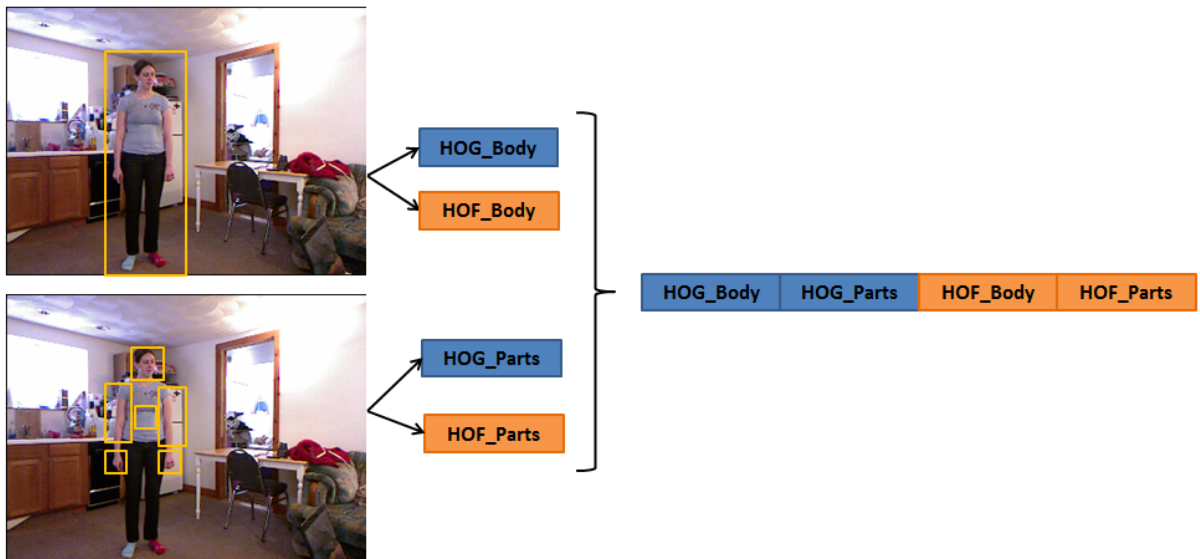
4.2.1 Rút trích đặc trưng

Đặc trưng từ dữ liệu khung xương: dữ liệu khung xương chứa 15 khớp nối trên toàn bộ cơ thể người: đầu, cổ, thắt lưng, vai trái, khuỷu tay trái, vai phải, khuỷu tay phải, hông trái, gối phải, hông phải, gối phải, tay trái, tay phải, chân trái và chân phải. Mỗi khớp nối gồm tọa độ x, y và z trong không gian 3 chiều và thông về hướng chuẩn hóa theo nghiên cứu [98]. Các đặc trưng được rút trích thể hiện các tính chất: 1) hình dáng cơ thể dựa trên sai khác vị trí về hướng của các lớp nối trên cơ thể với thắt lưng và độ lệch giữa 2 bàn chân với thắt lưng để biết người đang đứng hay ngồi; 2) đặc trưng vị trí tay được tính bằng cách tính độ lệch vị trí khớp nối trái và phải với vị trí tương đối với thắt lưng và đầu; 3) chuyển động của khung xương được xác định chuyển động bằng cách tính độ lệch các khớp nối tại khung hiện tại với 9 khung hình trước đó $S = \{T - 1, T - 3, T - 5, T - 7, T - 9, T - 11, T - 15, T - 17, T - 19\}$. Với mỗi khớp nối luận án tính sự khác biệt giữa vị trí và thông tin hướng theo 9 khung hình trong tập S và góc của khớp nối.

Đặc trưng	Số chiều
Hình dáng cơ thể	46
Vị trí tay	14
Chuyển động của khung xương	396
Đặc trưng kết hợp	456

Bảng 4-1. Bảng tổng hợp kết quả rút trích đặc trưng khung xương

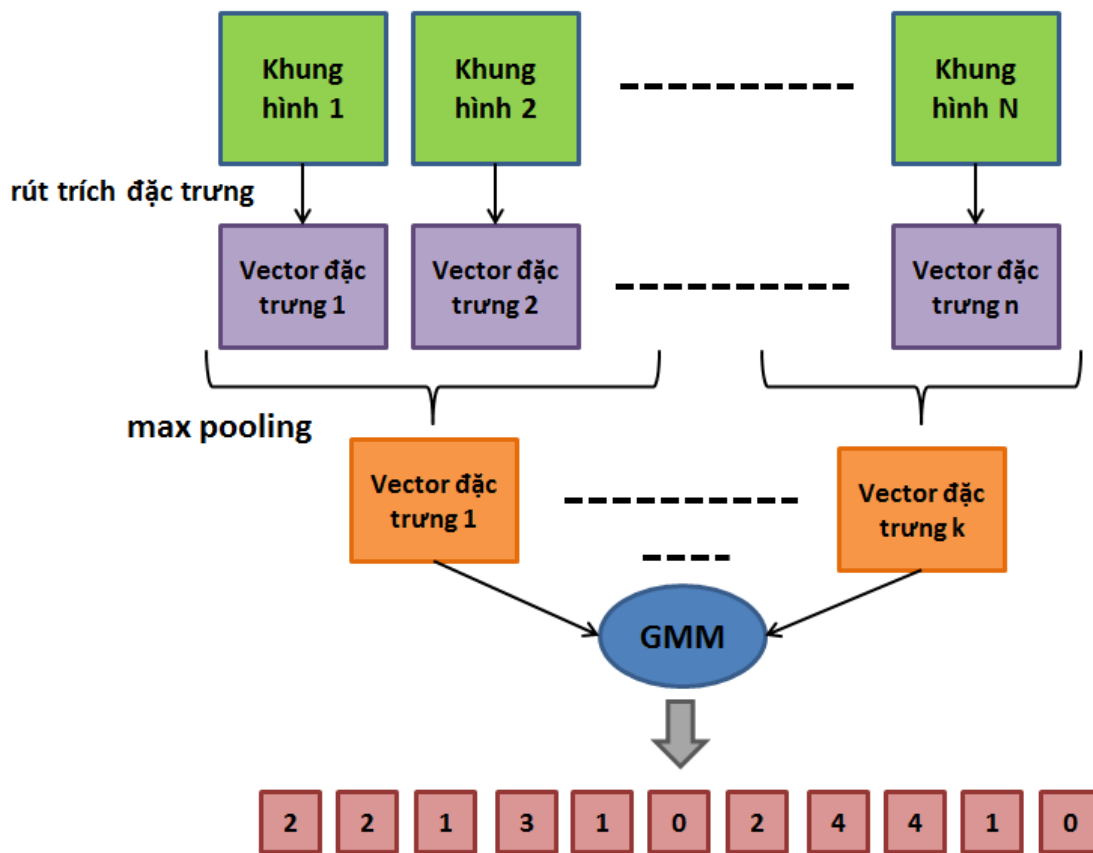
Đặc trưng màu - độ sâu: mỗi khung hình, luận án rút trích 7 vùng ảnh đại diện như **hình 4-2** gồm một vùng toàn bộ cơ thể và 6 vùng bộ phận của cơ thể cánh tay trái, cánh tay phải, tay trái, tay phải, đầu và bụng. Các vùng đại diện này được rút trích thông tin hình dáng và chuyển động trên cả ảnh màu và ảnh độ sâu để biểu diễn cho một nội dung tại một khung hình. Như vậy sẽ có 14 vùng ảnh đại diện được rút trích cho ảnh màu và 14 vùng ảnh được rút trích cho ảnh độ sâu cho hình dáng và chuyển động. Với việc biểu diễn cả toàn cục và cục bộ các bộ phận của người, các đặc trưng rút được tại mỗi khung hình không chỉ biểu diễn toàn cục hình dáng, chuyển động của người mà còn biểu diễn được mối tương tác giữa các bộ phận và toàn bộ cơ thể cũng như đối tượng trong quá trình thực hiện hoạt động. Ngoài ra việc lấy vùng cơ thể rộng thêm 20% chiều rộng và chiều cao khi rút trích đặc trưng để mã hóa được thông tin ngữ cảnh trong quá trình thực hiện hoạt động. Phương pháp tính đặc trưng HOG [85] cho ảnh bằng phiên bản điều chỉnh so với phương pháp gốc: các vùng ảnh được chia thành các khối 3x3, số hướng lượng hóa là 9 và quá trình tổng hợp các lược đồ được tính gộp đầu nhau 50%. Kết quả mỗi vùng sẽ được đại diện bởi một véc tơ đặc trưng $2 \times 2 \times 9 = 36$ chiều. Mỗi khung hình rút trích đặc trưng từ kênh màu và độ sâu tạo thành một véc tơ đặc trưng với số chiều là $2 \times 14 \times 36 = 1008$.



Hình 4-2. Minh họa rút trích đặc trưng của cơ thể và bộ phận trên dữ liệu màu

4.2.2 Phát sinh chuỗi hoạt động

Hoạt động người có thể được xem như là một tập của các hoạt động con và một hoạt động con có thể được định nghĩa như là một kết hợp của hình dáng cơ thể, bộ phận cơ thể, chuyển động và tương tác trong một khoảng thời gian. Luận án áp dụng kỹ thuật phân đoạn vào video, mỗi video sẽ được chia thành một tập các đoạn video con có chiều dài bằng nhau (15 khung hình theo thực nghiệm trong chương 3). Mỗi đoạn video này có thể được xem như là một hoạt động con như trong nghiên cứu trước đó [54]. Thay vì mô hình hóa các hoạt động con thành các khối chứa nhiều thông tin và đánh nhãn các hoạt động con có thể có. Nghiên cứu áp dụng kỹ thuật gom nhóm các hoạt động con thành các nhóm riêng biệt. Các hoạt động con tương đồng nhau sẽ được gom thành một nhóm dựa vào các đặc trưng thì giác được rút trích. Bằng cách này, mỗi hoạt động sẽ được đại diện bởi một chuỗi các số có thứ tự mà mỗi số là chỉ mục của nhóm tương ứng với hoạt động con thuộc về nhóm đó. Trong nghiên cứu này, luận án sử dụng thuật toán GMM để tiến hành gom nhóm các đặc trưng đại diện của các hoạt động con thành các nhóm đại diện bằng thuật toán EM như đã được đề xuất trong **chương 3** khi thực hiện biểu diễn hành động.



Hình 4-3. Minh họa quá trình chuyển đổi một hoạt động từ một chuỗi khung hình thành chuỗi số tự nhiên với số nhóm là 5

4.2.3 Mô hình Markov ẩn

Sau quá trình phát sinh chuỗi hoạt động con thì mỗi video được đại diện bởi một chuỗi thời gian. Bài toán nhận dạng hoạt động trở thành bài toán xác định nhãn của chuỗi thời gian. Mô hình Markov ẩn (HMM) là mô hình được xây dựng bằng phương pháp thống kê [67], hiệu quả trong việc giải quyết các suy luận không chắc chắn của dữ liệu có mối quan hệ theo diễn tiến thời gian. Điều này thích hợp với bài toán nhận dạng hoạt động trong môi trường sinh hoạt hằng ngày vốn có sự thay đổi quá trình thực hiện theo cấu trúc thời gian. Trong mô hình HMM, các xác suất quan sát có mối liên hệ phụ thuộc với nhau làm tăng hiệu quả trong suy luận. Ngoài ra, HMM có khả năng huấn luyện dựa trên dữ liệu học để có thể rút ra các qui luật dựa trên xác suất mà không cần nhiều sự giám sát của con người [67].

Các tham số của mô hình HMM

N : số lượng trạng thái ẩn

ω : tập các trạng thái $\omega = \{1, 2, \dots, N\}$, tương ứng với cấu trúc ẩn trong vỏ não thị giác trong việc hiểu thông tin từ chuỗi hoạt động con

M : số lượng ký hiệu quan sát

V : tập các ký hiệu $V = \{1, 2, \dots, M\}$, tương ứng với các hoạt động con được gom nhóm trong bài toán nhận dạng hoạt động

A : ma trận xác suất chuyển đổi trạng thái

$$a_{ij} = P(\omega_{t+1} = j | \omega_t = i) \quad (4.1)$$

trong đó $i = 1..N, j = 1..N$

với điều kiện:

$$\sum_{j=1}^N a_{ij} = 1 \quad (4.2)$$

B : ma trận phân phối xác suất quan sát

$$b_j(k) = P(o_t = k | \omega_t = j) \quad (4.3)$$

trong đó $k = 1..M$

với điều kiện:

$$\sum_{k=1}^M b_j(k) = 1 \quad (4.4)$$
$$j = 1..N$$

π : xác suất khởi tạo của các trạng thái

λ : mô hình $\lambda(\pi, A, B)$

Đối với bài toán nhận dạng hoạt động sử dụng mô hình HMM. Mỗi lớp hoạt động sẽ được mô hình hóa thành một mô hình HMM theo **thuật toán 4.1**. Hoạt động mới cần nhận dạng sẽ được xác định bằng cách tính ước lượng tham số với từng mô hình và chọn ra mô hình có kết quả ước lượng là cao nhất theo **thuật toán 4.2**.

Thuật toán 4.1: Thuật toán học mô hình phân lớp HMM

Mục tiêu: xây dựng bộ tham số HMM cho mỗi nhãn hoạt động

Đầu vào: Tập dữ liệu $C = \{C_1, C_2, \dots, C_L\}$ của nhãn phân lớp

Đầu ra: Mô hình tham số cho mỗi lớp hoạt động $\lambda_i(\pi_i, A_i, B_i)$, $i = 1 \dots L$

Begin

1// Lặp mỗi lớp hoạt động C_i

2// Ước lượng mô hình $\lambda_i(\pi_i, A_i, B_i)$ theo thuật toán Baum-Welch [67] (phụ lục B).

3// Kết thúc lặp

4// Tập các mô hình $\lambda_i(\pi_i, A_i, B_i)$

End

Thuật toán 4.2: Thuật toán phân lớp mô hình HMM

Mục tiêu: xác định nhãn hoạt động cho một chuỗi đầu vào mới

Đầu vào: Chuỗi dữ liệu hoạt động, mô hình tham số cho mỗi lớp hoạt động

$\lambda_i(\pi_i, A_i, B_i)$, $i = 1 \dots L$

Đầu ra: Nhãn của hoạt động

Begin

1// Lặp mỗi lớp hoạt động λ_i

2// Tính xác suất $P_i(O|\lambda_i)$ theo thuật toán Forward [67] (phụ lục B).

3// Kết thúc lặp

4// Xác định nhãn bằng công thức: $\lambda_m = \underset{i=1..L}{\operatorname{argmax}} P(O|\lambda_i)$

End

4.2.4 Mô hình HCRF

Khi giải quyết các bài toán dữ liệu chuỗi thời gian trong lĩnh vực TGMT thì HMM là một phương pháp được sử dụng rộng rãi. Nhưng vì HMM là mô hình tạo sinh nên khi sử dụng nó yêu cầu giả định về sự độc lập của dữ liệu quan sát. Các ràng buộc này tạo ra các khó khăn hoặc không thể chấp nhận sự phụ thuộc tầm xa của các quan sát hoặc giữa các đặc trưng bị chồng chéo nhau trong các quan sát [103, 108]. Để vượt qua các vấn đề trên một giải pháp có thể được sử dụng đó là sử dụng mô hình di biệt. Các mô hình di biệt cho phép mô hình hóa sự độc lập giữa các quan sát và sự phụ thuộc cục bộ giữa trạng thái và các quan sát. Trường ngẫu nhiên có điều kiện (CRF) [50] là một biến thể mô hình di biệt của HMM cung cấp một phương pháp để ước lượng xác suất của nhãn cho toàn bộ một chuỗi dữ liệu.

Công thức tính xác suất của một chuỗi trạng thái $y = (y_0, y_1, \dots, y_T)$ với chuỗi quan sát cho trước $x = (x_0, x_1, \dots, x_T)$ và bộ tham số θ như sau:

$$p(y|x, \theta) = \frac{1}{Z(x)} \prod_{C_p \in C} \prod_{\psi_c \in C_p} \psi_c(x_c, y_c; \theta_p) \quad (4.5)$$

trong đó:

$$\psi_c(x_c, y_c; \theta_p) = \exp \left\{ \sum_{k=1}^{K(p)} \lambda_{pk} f_{pk}(x_c, y_c) \right\} \quad (4.6)$$

$$Z(x) = \sum_y \prod_{C_p \in C} \prod_{\psi_c \in C_p} \Psi_c(x_c, y_c; \theta) \quad (4.7)$$

f_{pk} là véc tơ đặc trưng, một hàm của trạng thái y_c và quan sát đầu vào x_c ; λ_{pk} là véc tơ tham số của thành phần trọng số thứ k trong véc tơ đặc trưng f_{pk} ; θ bao gồm λ_{pk} và λ_{pk} ; C là tập mẫu các clique của đồ thị trong CRF [50].

Hằng số chuẩn hóa Z được gọi là hàm phân vùng của tổng các trường hợp có thể có của Y , hàm phân vùng này chiếm phần lớn chi phí tính toán trong giai đoạn học.

Để có thực hiện bài toán phân lớp, chúng ta sẽ thêm nhiều hơn một biến phân lớp vào mô hình CRF bằng cách mở rộng công thức và được đặt tên là biến phân loại ω , ta có công thức sau:

$$p(y, \omega | x, \theta) = \frac{1}{Z(x)} \prod_{C_p \in C} \prod_{\psi_c \in C_p} \psi_c(x_c, y_c, \omega_c; \theta_p) \quad (4.8)$$

trong đó:

$$\psi_c(x_c, y_c, \omega_c; \theta_p) = \exp \left\{ \sum_{k=1}^{K(p)} \lambda_{pk} f_{pk}(x_c, y_c, \omega_c; \theta_p) \right\} \quad (4.9)$$

Công thức tính HCRF có thể được xác định bởi một tập các trạng thái ẩn như sau:

$$\begin{aligned} p(\omega|x, \theta) &= \sum_y p(y, \omega | x, \theta) \\ &= \sum_y \frac{1}{Z(x)} \prod_{C_p \in C} \prod_{\psi_c \in C_p} \psi_c(x_c, y_c, \omega_c; \theta_p) \end{aligned} \quad (4.10)$$

Chúng ta có thể thấy sự khác biệt chính giữa HMM và HCRF là trong khi HMM là mô hình đồ thị có hướng với biến quan sát là xác định của biến trạng thái, HCRF mô hình hóa phân phối $P(y, \omega|x)$ một cách trực tiếp với ω là biến phân loại, y là biến ẩn được mô hình như là trường Markov toàn cục với điều kiện của quan sát x .

Sau quá trình rút trích đặc trưng và phát sinh chuỗi đại diện cho mỗi hoạt động trong video. Vấn đề trở thành bài toán phân lớp chuỗi với kỳ vọng phân lớp hành vi trong video dựa trên các chuỗi này. Chúng ta có thể tiến hành học phân phối trạng thái của các hành vi khác nhau theo mô hình đệ biệt. Theo các nghiên cứu [51, 102], luận

án sử dụng hàm mục tiêu trong bước huấn luyện các tham số của mô hình HCRF theo công thức:

$$L(\theta) = \sum_{i=1}^n \log P(y_i|x_i, \theta) - \frac{1}{2\sigma^2} \|\theta\|^2 \quad (4.11)$$

với n là số lượng mẫu huấn luyện. Tham số tối ưu của mô hình HCRF được sử dụng theo công thức:

$$\theta^* = \operatorname{argmax}_{\theta} L(\theta) \quad (4.12)$$

Đến thời điểm hiện tại chúng ta đã biết cách như thế nào để tạo ra được một mô hình HCRF. Như vậy, để tìm nhãn hoạt động ω một mẫu dữ liệu với x với tham số tối ưu θ^* của mô hình HCRF theo công thức:

$$\omega^* = \operatorname{argmax}_{\omega} P(\omega|x, \theta^*) \quad (4.13)$$

Thuật toán 4.3: Thuật toán học HCRF tổng quát

Mục tiêu: xây dựng bộ tham số HCRF cho mỗi nhãn hoạt động

Đầu vào: Tập dữ liệu D , siêu tham số xác định độ lớn tổng quát hoá λ , số lần lặp tối đa T , hệ số học $\eta_t (t = 1 \dots T)$

Đầu ra: Bộ tham số θ của mô hình

Begin

1// Khởi tạo bộ trọng số $\theta^1 = 0$

2// Lặp cho đến khi hội tụ hoặc đến số vòng lặp tối đa T

3// Lặp $m = 1$ đến M

4// Lấy mẫu huấn luyện (x^m, y^m) từ tập D và tính xác suất có điều kiện với mô hình hiện tại θ^t cho mỗi $\bar{y}$

$$P_{\theta^t}(\bar{y}|\bar{x}) = \frac{1}{Z(\theta^t, \bar{x})} \exp(\sum_i \theta_V^t \cdot f_V(\bar{x}, y_i) + \sum_i \theta_E^t \cdot f_E(\bar{x}, y_i, y_{i-1}))$$

5// Tính kỳ vọng của véc tơ đặc trưng

$$E_{\theta^t}[f(\bar{x}^m, \bar{y}^m)] = \sum_{\bar{y}} P_{\theta^t}(\bar{y}|\bar{x}^m) f(\bar{x}^m, \bar{y}^m)$$

6// Kết thúc lặp

7// Cập nhật mô hình:

$$\theta^{t+1} \leftarrow -(1 - \lambda\eta_t)\theta^t + \eta_t M^{-1} \sum_{m=1}^M (f(\bar{x}^m, \bar{y}^m) - E_{\theta^t}[f(\bar{x}^m, \bar{y}^m)])$$

8// Kết thúc lặp

9// Mô hình đầu ra: $\hat{\theta} \leftarrow \theta^{T+1}$

End

Trong nghiên cứu này, luận án thay bước khởi tạo trong **thuật toán 4.3** với trọng số khởi tạo của hàm thể hiện quan hệ trạng thái và quan sát đầu vào trong mô hình CRF là trọng số của các biến trạng thái bằng các tham số của mô hình học phân lớp HMM được học trước đó trong bài toán phân lớp hành vi bằng HMM.

4.3 Thực nghiệm

Luận án đánh giá mô hình thực nghiệm trên 2 tập dữ liệu chuẩn quốc tế là CAD-120 và MSR-DailyActivity3D. Trong đó tập dữ liệu CAD-120 có thời gian thực hiện dài và chứa nhiều thách thức trong việc biểu diễn tương tác hoạt động hơn so với MSR-DailyActivity3D. Trong nghiên cứu này, luận án sử dụng chiều dài của mỗi phân đoạn là 15 khung hình đã thực nghiệm đánh giá trên các tập dữ liệu trong các nghiên cứu của bài toán nhận dạng hành động và hoạt động trong chương 3. Nghiên cứu cũng đánh giá tác động của số lượng nhóm đến kết quả của quá trình phân lớp của mô hình HMM và chọn ra số nhóm tốt nhất để đánh giá kết quả phân lớp với số lượng trạng thái ẩn khác nhau trong mô hình HMM và HCRF.

Chương 5 MÔ HÌNH NHẬN DẠNG HOẠT ĐỘNG TRONG MÔI TRƯỜNG KHÔNG RÀNG BUỘC

5.1 Gợi thiệu

Hoạt động trong môi trường không ràng buộc là bài toán thách thức nhất và có nhiều ứng dụng nhất trong bài toán nhận dạng hành vi người. Các hoạt động được thực hiện với sự biến động lớn trong toàn bộ quá trình thực hiện hoạt động bên trong một lớp hoạt động cũng như giữa các lớp hoạt động. Để vượt qua các thách thức này, luận án đề xuất việc sử dụng mô hình học chuyển tiếp từ các mạng học sâu để nâng cao hiệu quả của việc rút trích đặc trưng. Đồng thời, xem hoạt động được hình thành từ một tập mẫu các hoạt động con và NBNN được sử dụng để phân lớp nhãn hoạt động dựa trên hàm tính khoảng cách mẫu hoạt động – lớp hoạt động thay cho hàm tính khoảng cách mẫu hoạt động – mẫu hoạt động.

5.2 Học chuyển tiếp

Định nghĩa của học chuyển tiếp được dựa trên bài khảo sát của Pan và Yang [105]. Các khái niệm của học chuyển tiếp (transfer Learning) bao gồm miền (domain) và nhiệm vụ (task):

- Miền $\mathcal{D}$: gồm không gian véc tơ đặc trưng $\mathcal{X}$ và hàm phân phối xác suất biên $P(X)$ trên không gian đặc trưng đó, với $X = \{x_1, \dots, x_n\}$, $x_i \in \mathcal{X}$ là các mẫu trong tập huấn luyện. Ta có: $\mathcal{D} = \{\mathcal{X}, P(X)\}$.
- Bài toán $\mathcal{T}$: gồm không gian nhãn $\mathcal{Y}$ và hàm phân phối xác suất có điều kiện $P(Y|X)$ có được từ tập huấn luyện chứa các cặp $\{x_i, y_i\}$ với $x_i \in \mathcal{X}$ và $y_i \in \mathcal{Y}$. Ta có: $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$

- Cho trước $\mathcal{D}_S, \mathcal{T}_S$ là miền và bài toán của dữ liệu nguồn, $\mathcal{D}_T, \mathcal{T}_T$ là miền và bài toán của dữ liệu đích, X_T là tập dữ liệu huấn luyện của bài toán đích, Y_T là tập nhãn đích. Nhiệm vụ của học chuyển tiếp là phải học được hàm $P(Y_T|X_T)$ của $\mathcal{D}_T$ từ thông tin có được từ $\mathcal{D}_S, \mathcal{T}_S$ với $\mathcal{D}_S \neq \mathcal{D}_T$ hoặc $\mathcal{T}_S \neq \mathcal{T}_T$.

Từ định nghĩa trên, đã có rất nhiều kỹ thuật khác nhau trong phương pháp học chuyển tiếp được nghiên cứu từ trước đến nay và được thống kê trong nghiên cứu của Pan và Yang [105]. Trong đó, 2 kỹ thuật học chuyển tiếp được sử dụng rộng rãi trong mô hình mạng học sâu hiện nay:

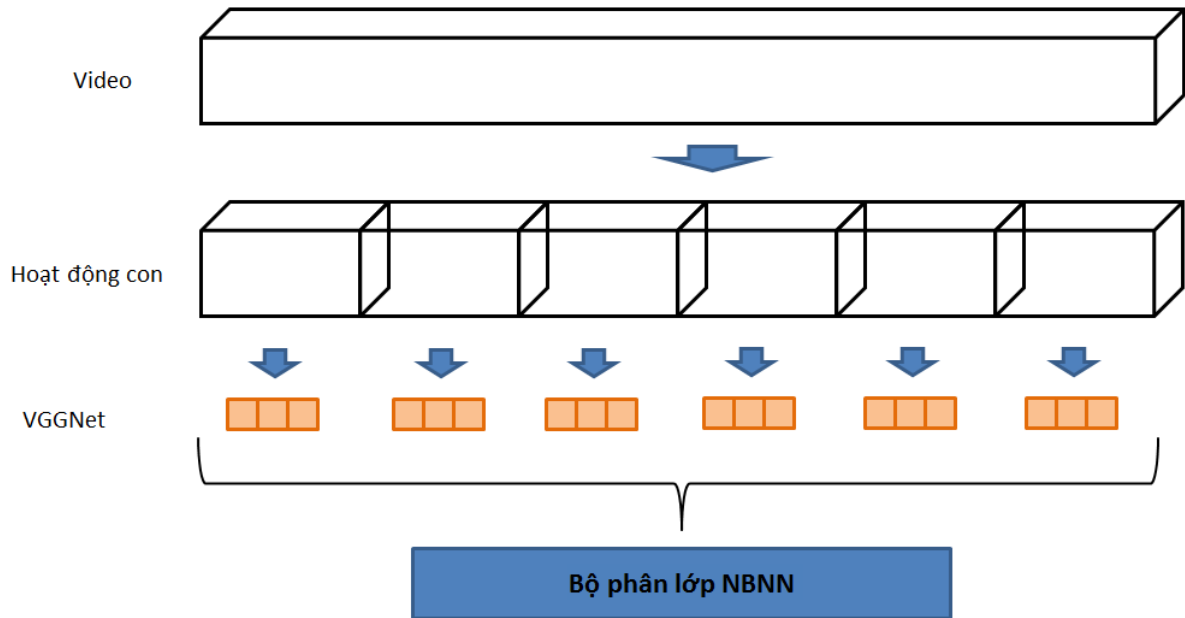
Sử dụng đặc trưng của mạng học sâu. Giả sử bài toán cần giải quyết là A. Chúng ta chọn một mô hình học sâu đã được huấn luyện trên dữ liệu của bài toán là B (bài toán này có sự tương đồng với A), sau đó cho dữ liệu của bài toán A chạy qua mô hình học sâu của B và lấy kết quả đầu ra ở một tầng nào đó của mô hình học sâu này làm đặc trưng học sâu rút được từ tập dữ liệu của A. Sau khi có các đặc trưng mới đại diện cho tập dữ liệu của bài toán A, chúng ta dùng nó làm dữ liệu đầu vào của một mô hình phân lớp (SVM, Softmax...) và học ra mô hình phân lớp cho bài toán A.

Sử dụng kỹ thuật fine-tuning với các mô hình đã được huấn luyện sẵn. Đối với kỹ thuật này, chúng ta cũng chọn một mô hình học sâu đã được huấn luyện trên dữ liệu của bài toán B (có sự tương đồng với bài toán cần giải quyết A), sau đó dùng mô hình này làm mô hình để học tiếp trên dữ liệu của bài toán A. Chúng ta sẽ dùng toàn bộ các trọng số của mô hình B làm trọng số khởi tạo cho bài toán và tiếp tục điều chỉnh (fine-tune) các trọng số này cho phù hợp với bài toán A. Chúng ta cũng có thể giữ nguyên trọng số ở một số tầng đầu của mô hình không thay đổi và chỉ điều chỉnh trọng số ở một số tầng cuối của mô hình cho phù hợp với bài toán A, hoặc học lại trọng số ở một số tầng cuối của mô hình với trọng số được khởi tạo ngẫu nhiên.

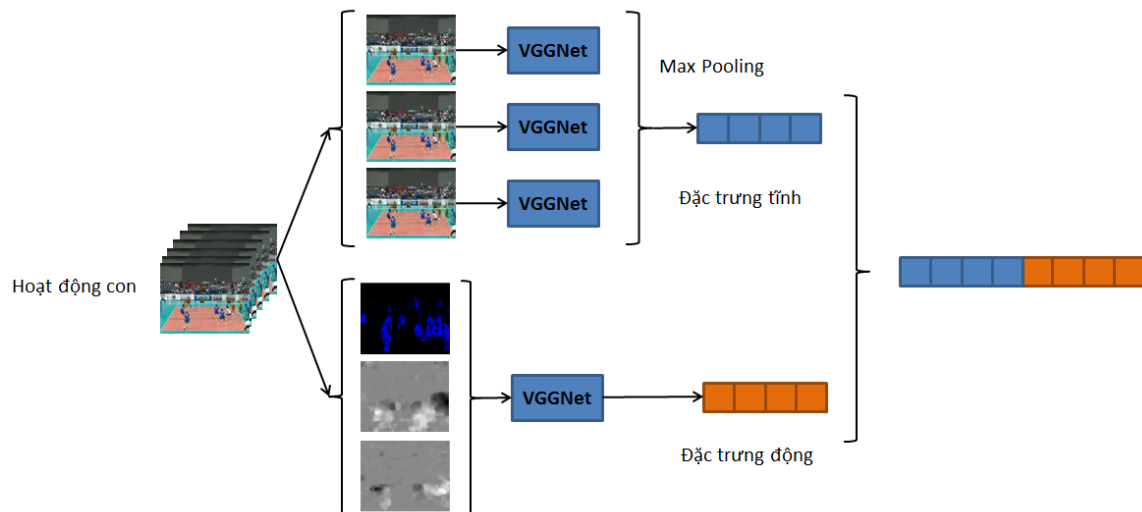
Trong mô hình đề xuất luận án đã tiến hành thực nghiệm đánh giá kết quả trên cả 2 hướng tiếp cận trên. Kết quả thực nghiệm đã cho thấy phương pháp học lại mạng từ bộ trọng số chuyển đổi đạt được kết quả cao hơn và phù hợp với bộ dữ liệu có số lượng mẫu lớn.

5.3 Mô hình đề xuất

Luận án tập trung vào khai thác đặc trưng cả tĩnh và động kết hợp với mô hình học sâu để tăng hiệu quả biểu diễn hoạt động dựa trên hình dáng, chuyển động, tương tác và ngữ cảnh được rút trích từ quá trình học và chuyển đổi tri thức từ mạng học sâu trước đó. Mô hình đề xuất được thực hiện theo trình tự các phân hệ như **hình 5-1**:



Hình 5-1. Kiến trúc mô hình nhận dạng hoạt động người trong điều kiện không ràng buộc



Hình 5-2. Minh họa quá trình rút trích đặc trưng tại mỗi hoạt động con

- Giai đoạn 1: Phân video thành các đoạn bằng nhau thể hiện một nội dung của hoạt động trong một phân đoạn không gian - thời gian được gọi là hoạt động con.
- Giai đoạn 2: Rút trích đặc trưng tĩnh và động cho các hoạt động con (**hình 5-2**). **Đặc trưng tĩnh** được trích chọn các khung hình chính theo chiều dài nội dung của hoạt động con, sử dụng mạng VGG để rút trích đặc trưng chuyển đổi của các khung hình này. Cuối cùng, sử dụng kỹ thuật max pooling để tổng hợp thành một đặc trưng duy nhất đại diện cho hoạt động con. **Đặc trưng động** được trích chọn bằng cách tính đặc trưng MHI, chuyển động theo hướng x và y cho mỗi hoạt động con. Các đặc trưng này được truyền

qua mạng học sâu tích chập VGG để tạo thành một đặc trưng động được học chuyển đổi cuối cùng đại diện cho một hoạt động con.

- Giai đoạn 3: Các dữ liệu từ tập học được rút trích từ bước 2 và lưu lại trong cơ sở dữ liệu. Khi mẫu cần xác định nhãn hoạt động, thuật toán NBNN được ứng dụng để xác định nhãn cho video mới. Để tăng hiệu quả của quá trình tìm kiếm láng giềng gần nhất trong quá trình nhận dạng luận án sử dụng các phương pháp đánh chỉ mục LSH [72].

5.3.1 Mô hình VGG

VGG [57] là dạng mở rộng mạng học sâu tích chập được đề xuất bởi Karen Simonyan và Andrew Zisserman. VGG được sử dụng để phân 1000 lớp ảnh trên tập dữ liệu ảnh ImageNet. Kiến trúc mạng này đề xuất cấu hình mạng tích chập mới với kích thước nhân tích chập nhỏ là 3×3 và 1×1 , kích thước cửa sổ max pooling là 2×2 , độ sâu của kiến trúc mạng lên đến 19 lớp. Kiến trúc mạng này đánh giá tác động của độ sâu đến hiệu quả nhận dạng bằng cách xây dựng các mạng lớn hơn từ việc huấn luyện mở rộng dựa trên các mạng có kiến trúc nông hơn. Trong nghiên cứu này, VGG được sử dụng để tận dụng sức mạnh trong việc rút trích đặc trưng tĩnh từ các khung hình chính của các phân đoạn và được huấn luyện để rút trích đặc trưng động từ các bản đồ đặc trưng của luồng chuyển động và các mẫu chuyển động theo thời gian. Như vậy, một phân đoạn sẽ được sử dụng VGG để rút trích 2 luồng đặc trưng là hình dáng và chuyển động tương thích với giả thuyết 2 dòng vỏ não thị giác [40]. Trong nghiên cứu này, luận án sử dụng VGG16 để huấn luyện mạng với bộ trọng số khởi tạo được chuyển đổi từ mô hình học trên ImageNet với việc thay lớp soft-max bằng số lớp đầu ra của của tập dữ liệu huấn luyện.

5.3.2 Đặc trưng tĩnh

Một video bao gồm một tập các khung hình diễn tả sự thay đổi của nội dung, trong một số hoạt động chỉ cần nhìn vào một số khung hình thì con người có thể dễ dàng xác định được hoạt động được thực hiện trong video. Luận án tiến hành lấy 3 khung hình chính tại mỗi hoạt động con (15 khung hình) theo thứ tự diễn tiến của thời gian 3, 8 và 13 theo kết quả đánh giá phân đoạn **chương 3**. Điều này giúp mô tả một cách hiệu quả sự thay đổi của hình dáng và ngữ cảnh khi thực hiện hoạt động trong khoảng thời gian nhất định. Các khung hình chính này sẽ được truyền qua mạng VGGNet để mã hóa thông tin hình dáng, tương tác và ngữ cảnh của một phân đoạn video. Sau quá trình này một phân đoạn sẽ gồm 3 véc tơ đặc trưng mô tả nội dung. Cuối cùng, kỹ thuật max pooling được sử dụng lên 3 véc tơ đặc trưng này để tạo véc tơ đặc trưng cuối cùng đại diện cho một phân đoạn hay một hoạt động con trong video.

5.3.3 Đặc trưng động

Chuyển động là một đặc tính quan trọng được sử dụng trong các bài toán nhận dạng dựa vào video. Nó chứa các tính chất về sự biến đổi trong quá trình thực hiện của đối tượng như hướng chuyển động, tốc độ di chuyển, sự thay đổi của chuyển động và hình dáng theo không gian – thời gian. Các đặc tính của chuyển động giúp phân biệt giữa các hoạt động có hình dáng tương đồng khi thực hiện hoạt động nhưng tốc độ di chuyển của đối tượng thực hiện hành động khác nhau. Ngoài ra, luồng chuyển động cũng thể hiện được sự xuất hiện của đối tượng giữa các khung hình liên kế giúp phân biệt hiệu quả hơn giữa các hoạt động có tính chất tương đồng nhưng khác nhau về cách thức thực hiện hành động ví dụ chạy (run), đi bộ (walk), nhảy bước (jogging). Để biểu diễn chuyển động trong một hoạt động con, nghiên cứu sử dụng 2 đặc trưng là ảnh lịch sử chuyển động (MHI) và luồng chuyển động. Như vậy mỗi hoạt động con trong không gian – thời gian (3 chiều) đã được chuyển về thành không gian 2 chiều trên ảnh ánh xạ. Điều này có ý nghĩa to lớn trong việc tận dụng sức mạnh của mô hình học chuyển đổi trong ảnh không gian từ các mạng học sâu trên ImageNet. Các mô hình này làm giảm đáng kể chi phí tính toán cũng như thời gian học và tận dụng được các mô hình đã được huấn luyện từ các nghiên cứu với lượng dữ liệu lớn.

MHI [17] là đặc trưng mô tả quá trình chuyển đổi chuyển động với 2 tính chất: i) như thế nào chuyển động được xảy ra; ii) chuyển động xảy ra ở những nơi nào. MHI được định nghĩa là một hàm lịch sử theo thời gian của chuyển động tại mỗi điểm được định nghĩa theo công thức:

$$H_{\tau}(x, y, t) = \begin{cases} \tau & \text{nếu } D(x, y, t) = 1 \\ \max(0, H_{\tau}(x, y, t - 1) - 1) & \text{ngược lại} \end{cases} \quad (5.1)$$

Trong đó:

- x, y lần lượt là tọa độ của điểm ảnh trong khung hình
- $D(x, y, t)$ là ảnh nhị phân của vùng chuyển động tại thời điểm t
- τ là tham số định nghĩa khoảng thời gian của chuyển động
- $H_{\tau}(x, y, t)$ là MHI tại thời điểm t tại điểm (x, y) của khung hình

Luồng chuyển động quang học biểu diễn chuyển động trong video theo hướng và biên độ dịch chuyển. Nó biểu diễn mỗi điểm trong khung hình hiện tại chuyển động như thế nào trong khung hình tiếp theo. Vì vậy luồng chuyển động là một đặc trưng cấp thấp có vai trò quan trọng trong việc phân tích chuyển động của đối tượng trong video. Để có thể phân tích được hoạt động của đối tượng trong video thì luồng chuyển động thể hiện được sự di chuyển của các đối tượng trong quá trình thực hiện hoạt động cũng như quá trình tương tác giữa chúng. Trong nghiên cứu này, luận án sử dụng thuật toán luồng chuyển động dày của G. Farneback [32]. Mỗi phân đoạn chúng

tôi sử dụng kỹ thuật pooling để tổng hợp từng luồng chuyển động theo lướng x và y thành 2 ánh xạ chuyển động cuối cùng đại diện cho hoạt động con.

5.3.4 Mô hình phân lớp NBNN

NBNN lần đầu tiên được đề xuất bởi Boiman [89] ứng dụng trong bài toán phân lớp ảnh đã đạt được hiệu quả cao hơn cả phương pháp SVM. NBNN là mô hình phi tham số với một số ưu điểm trong bài toán phân lớp: i) giải quyết số lượng lớp lớn; ii) tránh vấn đề quá khớp trong học mẫu; iii) không yêu cầu giai đoạn học; iv) dễ dàng mở rộng khi dữ liệu học hoặc số lớp tăng lên. Trong nghiên cứu của mình, Boiman [89] đã chỉ ra hiệu quả của phương pháp láng giềng gần nhất bị giảm khi thực hiện bị lỗi lượng hóa đặc trưng và hàm tính khoảng cách ảnh - ảnh. Trong các thực nghiệm của mình, tác giả cũng đã chỉ ra rằng các mẫu dữ liệu tần suất lớn có lỗi lượng hóa thấp nhưng mẫu hiếm thì lỗi lượng hóa rất lớn. Thực tế, các điểm mẫu dữ liệu đặc biệt là rất hiếm. Vì vậy, mà việc lượng hóa làm giảm đáng kể khả năng dị biệt của các mẫu dữ liệu. Hơn nữa, nghiên cứu còn chỉ ra rằng hàm ảnh - lớp sử dụng toàn bộ phân phối của một lớp cung cấp khả năng tổng quát hóa tốt hơn so với hàm khoảng cách ảnh - ảnh.

Trong mô hình này, luận án mở rộng khái niệm NBNN cho bài toán phân lớp ảnh dựa vào đặc trưng cục bộ sang bài toán phân lớp hoạt động trong video dựa vào hoạt động con. Nghiên cứu trực tiếp tính toán hàm khoảng cách video – lớp hoạt động thay cho hàm tính khoảng cách video – video. Trong ngữ cảnh phân lớp hoạt động các đặc trưng cục bộ được xem như là các hoạt động con trong video hoạt động (**thuật toán 5.1**).

Thuật toán 5.1: Thuật toán NBNN phân hoạt động dựa vào hoạt động con Mục tiêu: xác định nhãn hoạt động cho một video mới Đầu vào: Video cần phân lớp V, tập dữ liệu $C = \{C_1, C_2 \dots, C_n\}$ của nhãn phân lớp, n là số láng giềng gần nhất (NN) Đầu ra: Nhãn hoạt động của video đầu vào L
Begin 1// Tính các đặc trưng biểu diễn hoạt động con $v_1, v_2 \dots, v_m$ của video V 2// $\forall v_i \forall C_j$ tìm n láng giềng gần nhất (NN) của v_i trong C_j: $NN_{C_j}(v_i)$ 3// $\hat{C} = \arg \min_c \sum_{i=1}^n \ v_i - NN_c(v_i)\ ^2$ End

5.4 Thực nghiệm

Trong chương này, luận án thực nghiệm đánh giá các phương pháp đề xuất trên 2 tập dữ liệu là UCF101 và HMDB51. Các kết quả thực nghiệm đã cho thấy đặc trưng học chuyển tiếp từ các mô hình học sâu đạt được kết quả khả quan trên các tập dữ liệu

thách thức. Trong đó, đặc trưng chuyển được học chuyển tiếp từ mạng học sâu đạt hiệu quả cao hơn so với đặc trưng hình dáng. Ngoài ra, việc kết hợp hai lại đặc trưng này giúp đạt được hiệu quả tối ưu cho bài toán nhận dạng hoạt động trong môi trường không ràng buộc trên các tập dữ liệu này. Các kết quả thực nghiệm đạt độ chính xác lần lượt là 94.80% và 72.70% cho UCF101 và HMDB51.

Chương 6 KẾT LUẬN

Luận án đã nghiên cứu và phát triển một số phương pháp trong bài toán nhận dạng hành vi người cụ thể là bài toán nhận dạng hành động và hoạt động trong video. Các thành quả trình bày trong luận án có thể được tổng kết như sau:

Thứ nhất, luận án đã mô tả một cách tổng quát về bài toán mang tính thời sự với tính ứng dụng cao và thu hút được nhiều sự quan tâm của cộng đồng nghiên cứu trong lĩnh vực TGMT trong thời gian gần đây: “nhận dạng hành vi người trong video”. Bài toán này được xem xét ở nhiều khía cạnh và dựa trên nền tảng của bốn bài toán con là nhận dạng cử chỉ, hành động, hoạt động và cảm xúc. Luận án tập trung vào việc nghiên cứu và cải tiến mô hình cho 2 bài toán chính nhận dạng hành động và hoạt động. Nội dung này được trình bày chính trong chương 2 và chi tiết hiện thực hóa các mô hình trong các chương 3, 4, 5 và trải rộng trên các công trình đã công bố của luận án.

Thứ hai, đối với mô hình nhận dạng hành động người, nghiên cứu đề xuất mô hình rút trích các điểm đặc trưng từ ảnh màu – độ sâu dựa trên đặc trưng cục bộ với bộ trích chọn điểm trọng yếu không gian – thời gian STIP và các bộ miêu tả HOG3D, 3DSHONV và HOF2.5D. Sử dụng mô hình BOW với từ vựng thị giác được gom nhóm với thuật toán GMM với Kmeans++ để tăng hiệu quả phát sinh giỏ từ thị giác và đánh trọng số mềm trong biểu diễn hành động. Để mô tả sự thay đổi theo thời gian kỹ thuật phân video thành các đoạn có chiều dài bằng nhau và áp dụng kỹ thuật max pooling để tổng hợp thành các véc tơ đặc trưng cuối cùng đại diện cho hành động trong video. Mô hình học đa nhân được sử dụng để tăng hiệu quả kết hợp đặc trưng từ các nguồn khác nhau của mô hình đề xuất. Mô hình được thực nghiệm đánh giá trên 2 tập dữ liệu chuẩn quốc tế (UTKinect-Action và 3D Action Pairs) để minh họa khả năng tổng quát hóa cũng như hiệu quả của phương pháp đề xuất. Nội dung này được trình bày chính trong chương 3 và trong các công trình nghiên cứu [CT02], [CT03], [CT04], [CT06], [CT08], [CT09] và [CT10].

Thứ ba, đối với mô hình nhận dạng hoạt động người trong cuộc sống hàng ngày, luận án mở rộng kết quả nghiên cứu ở mô hình nhận dạng hành động người. Nghiên cứu tiến hành tận dụng sức mạnh của dữ liệu khung xương trong việc rút trích đặc trưng. Điều này giúp cho việc xác định đối tượng thực hiện hành động và tương tác trở nên hiệu quả hơn. Đồng thời, đặc trưng được rút trích từ trên toàn bộ cơ thể cũng như các bộ phận trên cơ thể giúp biểu diễn được chính xác những tương tác cục bộ của người trong quá trình thực hiện hoạt động. Luận án khai thác hiệu quả của đặc trưng chuỗi thời gian trong việc biểu diễn hoạt động từ video dựa vào kỹ thuật phân đoạn hoạt động con và mô hình gom nhóm GMM kết hợp Kmeans++. Cuối cùng, mô hình HMM và HCRF được sử dụng để phân lớp chuỗi thời gian của quá trình chuyển đổi nội dung trong quá trình thực hiện hoạt động. Các nghiên cứu này được đánh giá một cách có hệ thống trên hai tập dữ liệu chuẩn quốc tế: CAD 120 và MSRDailyActivity 3D. Nội dung này được trình bày trong chương 4 và trong các công trình nghiên cứu [CT01], [CT02], [CT04] và [CT06].

Cuối cùng, đối với hoạt động người trong môi trường không ràng buộc, đây là một trong những bài toán thách thức nhất trong nghiên cứu về hành vi người. Trong bài toán này, luận án tập trung vào việc khai thác sức mạnh của mô hình mạng học sâu trong việc tự học đặc trưng hình dáng và chuyển động cho việc biểu diễn hoạt động con. Nghiên cứu đề xuất mô hình dựa trên việc phân đoạn video tương tự như trong mô hình nghiên cứu trước đó trong chương 3 và 4. Các nghiên cứu này được đánh giá một cách có hệ thống trên hai tập dữ liệu chuẩn quốc tế: UCF101 và HMDB 51. Nội dung này được trình bày trong chương 5 và trong các công trình nghiên cứu [CT03], [CT05], [CT07] và [CT09].

TÀI LIỆU THAM KHẢO

Chi tiết về tài liệu tham khảo được trình bày trong toàn văn của luận án

