TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

BỘ MÔN CÔNG NGHỆ PHẦN MỀM

Nguyễn Đại Phát – 19110425

Lê Hải Dương– 19110341

Đề Tài

TÌM HIỂU CÁC MÔ HÌNH HỌC SÂU TRONG NHẬN DẠNG HÀNH ĐỘNG NGƯỜI VÀ VIẾT ỨNG DỤNG MINH HỌA

TIỂU LUẬN CHUYÊN NGÀNH

GIÁO VIÊN HƯỚNG DẪN

ThS. TRẦN CÔNG TÚ

TP. HCM, Tháng 12 năm 2022

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

BỘ MÔN CÔNG NGHỆ PHẦN MỀM

Nguyễn Đại Phát – 19110425

Lê Hải Dương– 19110341

Đề Tài

TÌM HIỂU CÁC MÔ HÌNH HỌC SÂU TRONG NHẬN DẠNG HÀNH ĐỘNG NGƯỜI VÀ VIẾT ỨNG DỤNG MINH HỌA

TIỂU LUẬN CHUYÊN NGÀNH

GIÁO VIÊN HƯỚNG DẪN

ThS. TRẦN CÔNG TÚ

TP. HCM, Tháng 12 năm 2022

TRƯỜNG ĐH SƯ PHẠM KỸ THUẬT TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN

****

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh Phúc

****

PHIẾU NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

Họ và tên Sinh viên 1: NGUYỄN ĐẠI PHÁT MSSV : 19110425

Họ và tên Sinh viên 2: LÊ HẢI DƯƠNG MSSV : 19110341

Ngành: Công Nghệ Thông Tin

Tên đề tài: TÌM HIỂU CÁC MÔ HÌNH HỌC SÂU TRONG NHẬN DẠNG HÀNH ĐỘNG NGƯỜI VÀ VIẾT ỨNG DỤNG MINH HỌA.

Họ và tên Giáo viên hướng dẫn: ThS. Trần Công Tú

NHẬN XÉT

1. Về nội dung đề tài & khối lượng thực hiện: …….....
   ……
2. Ưu điểm:……...…..
   ……
3. Khuyết điểm:…..…….……...
   ……
4. Đề nghị cho bảo vệ hay không ? ……..
5. Đánh giá loại :……..
6. Điểm :……..

Tp. Hồ Chí Minh, ngày tháng năm 2022

Giáo viên hướng dẫn

(Ký & ghi rõ họ tên)

TRƯỜNG ĐH SƯ PHẠM KỸ THUẬT TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN

****

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh Phúc

****

PHIẾU NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

Họ và tên Sinh viên 1: NGUYỄN ĐẠI PHÁT MSSV : 19110425

Họ và tên Sinh viên 2: LÊ HẢI DƯƠNG MSSV : 19110341

Ngành: Công Nghệ Thông Tin

Tên đề tài: TÌM HIỂU CÁC MÔ HÌNH HỌC SÂU TRONG NHẬN DẠNG HÀNH ĐỘNG NGƯỜI VÀ VIẾT ỨNG DỤNG MINH HỌA.

Họ và tên Giáo viên phản biện: ThS. Lê Vĩnh Thịnh

NHẬN XÉT

1. Về nội dung đề tài & khối lượng thực hiện: …….....
   ……
2. Ưu điểm:……...……..
   ……
3. Khuyết điểm:…..…….……...…….
   ……
4. Đề nghị cho bảo vệ hay không ? ……..
5. Đánh giá loại :……..
6. Điểm :……..

Tp. Hồ Chí Minh, ngày tháng năm 2022

Giáo viên phản biện

(Ký & ghi rõ họ tên)

LỜI CẢM ƠN

“TÌM HIỂU CÁC MÔ HÌNH HỌC SÂU TRONG NHẬN DẠNG HÀNH ĐỘNG NGƯỜI VÀ VIẾT ỨNG DỤNG MINH HỌA” là đề tài mà chúng em đã chọn để thực hiện Tiểu luận chuyên ngành, chuyên ngành Công nghệ Phần mềm lần này. Để hoàn thành tốt được bài tiểu luận này lần, nhóm chúng em xin gửi lời cảm ơn chân thành nhất đến thầy Trần Công Tú, người Thầy đã tận tình chỉ bảo, hướng dẫn, chúng em trong suốt quá trình nghiên cứu, thực hiện đề tài lần này.

Trong quá trình thực hiện bài luận, nhóm chúng em đã nhận được sự hỗ trợ, giúp đỡ từ các anh chị em, bạn bè cùng khoa, cùng trường để giúp hoàn thành việc thu thập và gán nhãn bộ dữ liệu, phục vụ cho nghiên cứu của mình, nhóm chúng em xin được gửi lời cảm ơn đến mọi người.

Một lần nữa, xin chân thành cảm ơn!

Sinh viên thực hiện đề tài

Nguyễn Đại Phát.

Lê Hải Dương.

Trường ĐH Sư phạm Kỹ thuật TP.HCM

Khoa Công Nghệ Thông Tin

ĐỀ CƯƠNG TIỂU LUẬN CHUYÊN NGÀNH

Họ và tên Sinh viên 1: NGUYỄN ĐẠI PHÁT MSSV : 19110425

Họ và tên Sinh viên 2: LÊ HẢI DƯƠNG MSSV : 19110341

Ngành: Công Nghệ Thông Tin

Tên đề tài: TÌM HIỂU CÁC MÔ HÌNH HỌC SÂU TRONG NHẬN DẠNG HÀNH ĐỘNG NGƯỜI VÀ VIẾT ỨNG DỤNG MINH HỌA.

Giáo viên hướng dẫn: ThS. Trần Công Tú

Nhiệm Vụ Của Luận Văn:

1. Hệ thống kiến thức đã học Deep Learning.
2. Tìm hiểu các thuật toán nhận dạng hành động người sử dụng Deep Learning.
3. Thu thập và xây dựng tập dữ liệu về hành động tập thể hình.
4. Xử lý dữ liệu, huấn luyện và đánh giá model học sâu trên tập dữ liệu đã thu thập.

Đề cương viết luận văn:

1. **Phần MỞ ĐẦU**
    1. Đặt vấn đề
    2. Giới hạn đề tài
    3. Mục tiêu của đề tài
2. **Phần NỘI DUNG**
    1. Chương 1: TỔNG QUAN VỀ DEEP LEARNING
    2. Chương 2: CÁC PHƯƠNG PHÁP HỌC SÂU TRONG BÀI TOÁN NHẬN DẠNG HÀNH ĐỘNG
    3. Chương 3: ỨNG DỤNG PHƯƠNG PHÁP NHẬN DẠNG HÀNH ĐỘNG CON NGƯỜI VÀO BÀI TOÁN ĐẾM HÀNH ĐỘNG TRONG QUÁ TRÌNH TẬP THỂ HÌNH
3. **Phần KẾT LUẬN**

1. Kết quả đạt được
2. Những hạn chế của đề tài
3. Hướng phát triển

4. Phần TÀI LIỆU THAM KHẢO

KẾ HOẠCH THỰC HIỆN ĐỀ TÀI

STT	Thời gian	Công việc	Ghi chú
1	Từ 31/08/2022 Đền 07/09/2022	Hệ thống các kiến thức đã học về Deep Learning và các khái niệm liên quan.	Bao gồm: AI, ML, DL
2	Từ 07/09/2022 Đến 19/19/2022	Tìm hiểu xem human action recognition thường được áp dụng, giải quyết những vấn đề gì trong đời sống.	Tìm kiếm những vấn đề thường được giải quyết trong đời sống được giải quyết bởi human action recognition.
3	Từ 07/09/2022 Đến 14/09/2022	Tìm kiếm một số ứng dụng của Human action recognition	Tìm kiếm các ứng dụng thực tế liên quan đến các hành động tập thể thao, tập yoga
4	Từ 20/09/2022 Đến 23/09/2022	Các công nghệ thường được áp dụng vào các ứng dụng này là gì ?	Tìm hiểu chi tiết các công nghệ này có phải các công nghệ mới dễ nâng cấp, phát triển và nguồn data dữ liệu có dễ tìm kiếm hay không.
5	Từ 24/09/2022 Đến 25/09/2022	Tìm hiểu các phương pháp dùng để đo đạc đánh giá các công nghệ thường được áp dụng.	
6	Từ 26/09/2022 Đến 29/09/2022	Bắt đầu thu thập các dữ liệu liên quan và cần thiết cho việc thử nghiệm	Cần quan tâm đến Kích thước tập dữ liệu, đặc trưng của tập dữ liệu là gì ?
7	Từ 30/09/2022 Đến 02/10/2022	Tìm kiếm các phương pháp phân loại hành động đang được sử dụng phổ biến hiện nay bao	Hai phương pháp phù hợp với đồ án được tìm thấy là CNN và 3D

		gồm cả các phương pháp truyền thống và nâng cao,..	
8	Từ 03/10/2022 Đến 09/10/2022	Tìm hiểu sâu hơn về các đặc trưng, thuộc tính của phương pháp vừa tìm được.	Tìm hiểu về mediapipe pose, yolov7 pose, các phương pháp đếm hành động
9	Từ 10/10/2022 Đến 01/12/2022	Bắt đầu thử nghiệm các công nghệ và data tìm được và so sánh tìm ra phương pháp phù hợp và tốt nhất	
10	02/12/2022 Đến 10/12/2022	Chọn model và hoàn thiện	Kết thúc thử nghiệm model, lựa chọn model phù hợp nhất.
11	11/12/2022 Đến 16/12/2022	Đánh giá lại toàn bộ mô hình	Nhìn nhận lại toàn bộ mô hình.
112	Từ 17/12/2022 Đến 25/12/2022	Tổng hợp dữ liệu đã nghiên cứu, viết báo cáo tiểu luận, thiết kế Powerpoint	File word, pptx từ 26/12/2022 Đến 31/12/2022

Ý kiến giáo viên hướng dẫn: …….

Ngày tháng năm 2022 Ngày tháng năm 2022

(ký & ghi rõ họ tên) **Người viết đề cương**

Nguyễn Đại Phát

Lê Hải Dương

MỤC LỤC

DANH MỤC ẢNH MINH HỌA

DANH MỤC BẢNG BIỂU

DANH MỤC THUẬT NGỮ VÀ TỪ VIẾT TẮT

Thuật ngữ	Tên đầy đủ	Nghĩa tạm dịch
AI	Artificial Intelligence	Trí tuệ nhân tạo
CNNs	Convolutional Neural Networks	Mạng nơ-ron tích chập
DL	Deep Learning	Học sâu
DNN	Deep Neural Network	Mạng nơ-ron sâu
GPU	Graphics Processing Unit	Bộ xử lý liên quan đến đồ hoạ
ML	Machine Learning	Học máy
RAM	Random Access Memory	bộ nhớ khả biến
RNN	Recurrent Neural Networks	Mạng nơ-ron hồi quy
YOLO	You Ony Look One	Mô hình cục bộ hoá hành động trong không thời gian

TÓM TẮT

Dựa nên nền tảng kiến thức đã có về Deep Learning, Machine Learning trong các môn học bổ trợ trước, trong đề tài "TÌM HIỂU CÁC MÔ HÌNH HỌC SÂU TRONG NHẬN DẠNG HÀNH ĐỘNG NGƯỜI VÀ VIẾT ỨNG DỤNG MINH HỌA", nhóm đã tiếp tục tìm hiểu và vận dụng những kiến thức để giải quyết bài toán đếm số hành động trong các bài tập thể hình có tính chu kỳ.

YOLOv7 Pose – You Only Look Once Version 7 là mô hình học sâu được sử dụng trong việc biểu diễn tư thế của người tập, từ đó nhóm áp dụng phương pháp so sánh giữa các tư thế được nhóm đề xuất để giải quyết bài toán đếm số hàng động. Trong tiểu luận này sẽ trình bày chi tiết về phương pháp mà nhóm thực hiện cũng như giới thiệu các kiến thức cơ bản về lĩnh vực nhận dạng hành động con người.

PHẦN 1: MỞ ĐẦU

1. Đặt vấn đề

Từ thời điểm Covid-19 bắt đầu bùng nổ toàn cầu, để chống chọi với Covid-19 và các biến chủng của nó rất nhiều trong số chúng ta khi trong thời gian cách ly toàn xã hội đã xây dựng những thói quen nâng cao sức khỏe bằng việc tập thể dục thể thao. Nhu cầu này bùng nổ mạnh mẽ trong thời kỳ đại dịch dẫn đến những bài toán mới mở ra cho các nhà phát triển các ứng dụng phục vụ người dùng. Nhu cầu của người dùng cuối có thể từ đơn giản là các bài tập cardio, body weight (các bài tập chủ yếu sử dụng chính sức nặng của cơ thể để đốt calo) đến các bài tập sử dụng các tạ hay thiết bị tại phòng tập chuyên dụng (phòng GYM).

Tuy nhiên khi nhu cầu tập tại nhà tăng cao lại mở ra những vấn đề khó khăn cho người dùng. Khó khăn có thể kể đến là nhu cầu đảm bảo chất lượng của buổi tập (đảm bảo về cả số lượng bài tập, số lượng hành động, độ chính xác của thực hiện), nhu cầu theo dõi hiệu quả (calo tiêu đốt). Thông thường, các khó khăn này không là vấn đề lớn với các người tập lâu năm hay tập tại phòng GYM nơi có các thiết bị đo chuyên dụng hay có huấn luyện viên cá nhân, tuyên lại là các khó khăn đáng kể cho người mới, người bận rộn có nhu cầu thể thao nhanh chóng nhưng vẫn hiệu quả.

Chính vì những điều này, chúng đã thúc đẩy những nhà phát triển xây dựng các ứng dụng các phương pháp, ứng dụng phục vụ các nhu cầu khác nhau của người dùng. Mục tiêu của bàn luận này là giới thiệu một phương pháp ứng dụng Pose Estimation cho việc đếm các hành động tập luyện.

Bài toán Pose Estimation là một bài toán nhánh xử lý hình ảnh mà ở đó nhiệm vụ là nhận diện và phân tích tư thế của người. Thành phần chính của bài toán này là việc có thể chuẩn hoá được toàn bộ cơ thể người cần dự đoán. Có ba hướng tiếp cận chính cho bài toán này gồm hướng khung xương (skeleton-based model), hướng viền (contour-based model) và hướng hình khối (volume-based model).

2. Giới hạn đề tài

Vì bài luận này nằm trong phạm vi giới hạn thời gian của một tiểu luận chuyên ngành, nên việc nghiên cứu và triển khai cũng năm trong thời gian hạn chế của kỳ học và một số hạn chế khác sau.

- Thứ nhất: Vì thời gian có hạn nhóm chỉ tập trung tìm hiểu một số mô hình tiêu biểu và đã được công bố trên các tạp chí quốc tế.
- Thứ hai: Bài toán yêu cầu cần phải xử lý với tốc độ tính toán cao nên yêu cầu cơ sở triển khai phải có GPU để đáp ứng yêu cầu về tốc độ.
- Thứ ba: Do giới hạn về tài nguyên nên nhóm chỉ thực hiện training trên pre-train model với 1 epoch trên toàn tập dữ liệu với mục tiêu tìm hiểu về dữ liệu và kiến trúc model.

3. Mục tiêu của đề tài

Đề tài đặt ra một số mục tiêu nghiên cứu chính sau:

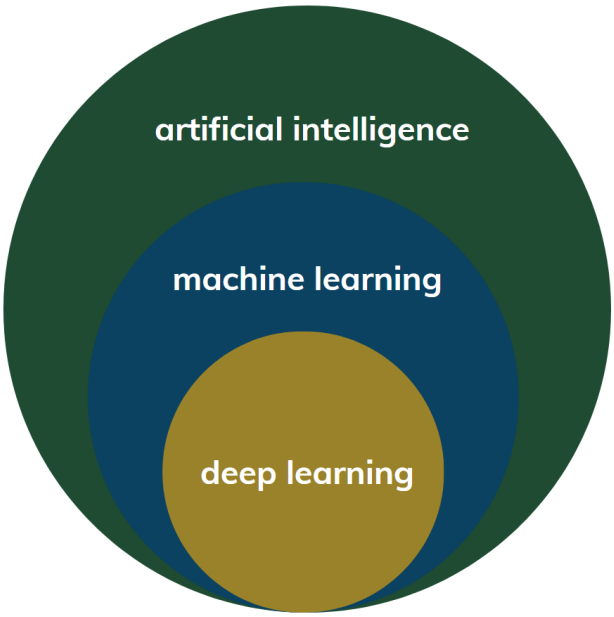
- Trình bài một số kiến thức chính về Deep learning và các khái niệm cơ bản cho việc hiểu thuật toán nhận dạng hành động con người trong đề tài.
- Trình bài chi tiết về mô hình học sâu mà tiểu luận sử dụng để nhận dạng hành động con người
- Áp dụng mô hình học sâu vào bài toán đếm hành động tập thể hình

PHẦN 2: NỘI DUNG

CHƯƠNG 1: TỔNG QUAN VỀ DEEP LEARNING

1.1. DEEP LEARNING

1.1.1. Khái niệm Deep Learning



Hình 1.1 Hình minh họa quan hệ giữa AI, ML, DL

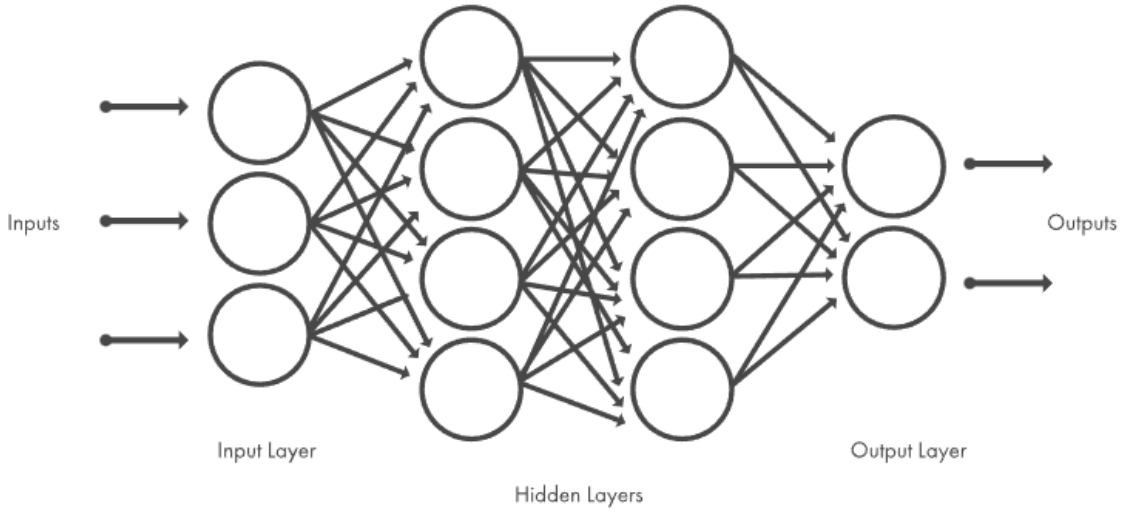
Deep learning là một nhánh của Machine learning còn Machine learning là một nhánh của AI. DL sử dụng các mạng thành kinh với nhiều lớp để đưa ra các dự đoán. Ý tưởng DL là cố gắng tái hiện lại cách thức hoạt động của bộ nào của con người.

DL được sử dụng trong nhiều ứng dụng như là nhận dạng tiếng nói, chữ viết, ô tô tự hành, sản xuất tự động … Thật không quá khi nói DL được ứng dụng trong hầu hết các lĩnh vực trong thời đại công nghệ 4.0.

Vì thế DL đang nhận được nhiều sự quan tâm chú ý gần đây. Và những kết quả mà nó mang lại trong một số nhiệm vụ đã mở ra một kỷ nguyên mới cho sự phát triển của AI.

1.1.2. Cách thức hoạt động của Deep Learning

Các phương pháp học sâu đa số đề sử dụng kiến trúc mạng thần kinh, nên đó là lý do các mô hình thường được gọi với cái tên khác là deep neural networks. Từ “deep” (sâu) được hiểu là số lớp hidden layers trong neural network.



Hình 1.2 Thành phần cơ bản của một mô hình DL

Input layer là lớp sẽ nhận dữ liệu đầu vào và chuyển nó vào phần tiếp theo của mạng

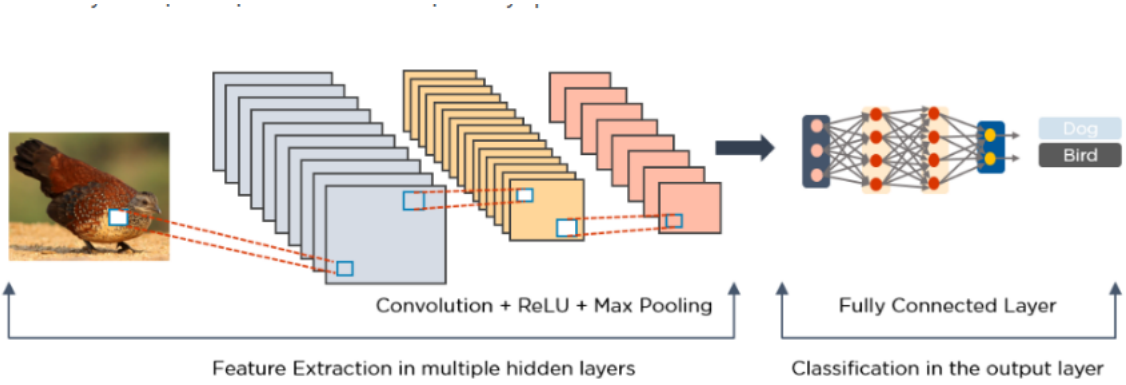
Hidden layers các lớp ẩn là nơi đưa ra các dự đoán thể hiện hiệu suất và độ phức tạp của mạng lưới. Nó giống với phần xử lý tính toán của bộ não con người.

Output layer là lớp đưa ra kết quả từ những phân tích dự đoán trước đó của lớp Hidden layers.

1.2. Một số mô hình Deep Learning phổ biến

1.2.1. Convolutional Neural Networks (CNNs)

CNN 's, còn được gọi là ConvNets, bao gồm nhiều lớp và chủ yếu được sử dụng để xử lý hình ảnh và phát hiện đối tượng. Yann LeCun đã phát triển CNN đầu tiên vào năm 1988 khi nó được gọi là LeNet. Nó được sử dụng để nhận dạng các ký tự như mã ZIP và chữ số. CNN được sử dụng rộng rãi để xác định hình ảnh vệ tinh, xử lý hình ảnh y tế, chuỗi thời gian dự báo và phát hiện dị thường.



Hình 1.3 Hình minh họa về cách hoạt động của một mô hình CNN

Mạng CNN là một tập hợp các lớp Convolution chồng lên nhau và sử dụng các hàm nonlinear activation như ReLU và tanh để kích hoạt các trọng số trong các node. Mỗi một lớp sau khi thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho các lớp tiếp theo.

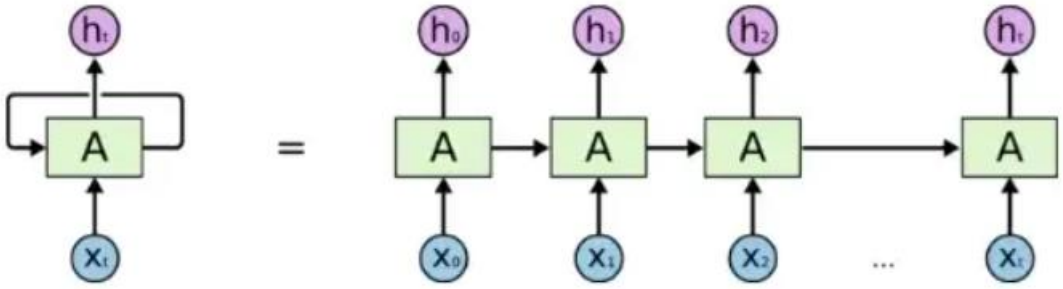
Theo thuật ngữ toán học thuần túy, một Convolution (tích chập) đại diện cho sự kết hợp của hai hàm f(x) và g(x) , khi hàm này trượt lên hàm kia. Đối với mỗi độ dịch chuyển trượt nhỏ (dx) , các điểm tương ứng của hàm thứ nhất f(x) và ảnh phản chiếu của hàm thứ hai g(t − x) được nhân với nhau rồi cộng lại. Kết quả là tích chập của hai hàm, được biểu diễn bằng biểu thức [f * g](t).

1.2.2. Recurrent neural networks (RNN)

Recurrent neural networks (RNN) là một loại mạng thần kinh mạnh mẽ để mô hình hóa dữ liệu trình tự, chẳng hạn như chuỗi thời gian hoặc ngôn ngữ tự nhiên.

Các thuật toán học sâu này thường được sử dụng cho các vấn đề thông thường hoặc thời gian, chẳng hạn như dịch ngôn ngữ, xử lý ngôn ngữ tự nhiên (nlp), nhận dạng giọng nói và chú thích hình ảnh; chúng được tích hợp vào các ứng dụng phổ biến như Siri, tìm kiếm bằng giọng nói và Google Dịch. Giống như các mạng nơ ron chuyển tiếp

và tích chập (CNN), các mạng nơ ron hồi quy sử dụng dữ liệu huấn luyện để học. Chúng được phân biệt bởi "bộ nhớ" của chúng khi chúng lấy thông tin từ các đầu vào trước đó để tác động đến đầu vào và đầu ra hiện tại. Trong khi các mạng nơ-ron sâu truyền thống giả định rằng đầu vào và đầu ra độc lập với nhau, thì đầu ra của mạng nơ-ron tái phát phụ thuộc vào các phần tử trước đó trong chuỗi.



Hình 1.4 Mô tả cách hoạt động của mô hình RNN

RNN có xu hướng gặp phải hai vấn đề, được gọi là exploding gradients và vanishing gradients. Những vấn đề này được xác định bởi kích thước của gradient, là độ dốc của hàm mất mát dọc theo đường cong lỗi. Khi độ dốc quá nhỏ, nó tiếp tục trở nên nhỏ hơn, cập nhật các tham số trọng số cho đến khi chúng trở nên không đáng kể.

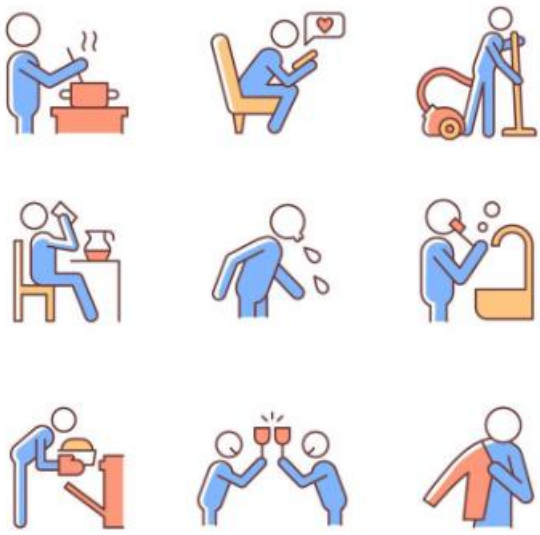
CHƯƠNG 2: CÁC PHƯƠNG PHÁP HỌC SÂU TRONG BÀI TOÁN NHẬN DẠNG HÀNH ĐỘNG

2.1. Khái quát về bài toán nhận dạng hành động con người

Phát hiện hoạt động của con người đóng một vai trò quan trọng trong các tương tác giữa người với người và các mối quan hệ giữa các cá nhân. Rất khó để trích xuất vì nó cung cấp thông tin về danh tính, tính cách và trạng thái tâm lý của một cá nhân. Khả năng con người nhận thức được các hoạt động của người khác là một trong những chủ đề nghiên cứu chính trong khoa học về thị giác máy tính và học máy.

Trong số các kỹ thuật phân loại khác nhau, có hai câu hỏi chính được đặt ra: "Hành động nào?" (chẳng hạn như vấn đề nhận dạng) và "Ở đâu trong video?" (tức là các vấn đề nội địa hóa). Để phát hiện hoạt động của con người, máy tính phải xác định trạng thái chuyển động của người đó để phát hiện hoạt động đó một cách hiệu quả. Các hoạt động của con người như "đi bộ" và "chạy" diễn ra tự nhiên trong cuộc sống hàng ngày và tương đối dễ nhận ra. Mặt khác, các quy trình phức tạp hơn như "gọt vỏ táo" khó xác định hơn.

Cử chỉ được coi là chuyển động nguyên thủy của các bộ phận cơ thể của một người có thể tương ứng với một hành động cụ thể của người này. Các hành động nguyên tử là các chuyển động của một người mô tả một chuyển động nhất định có thể là một phần của các hoạt động phức tạp hơn. Tương tác giữa người với đồ vật hoặc người với người là các hoạt động của con người có sự tham gia của hai hoặc nhiều người hoặc đồ vật. Hành động nhóm là các hoạt động được thực hiện bởi một nhóm hoặc nhiều người. Hành vi của con người là những hành động thể chất gắn liền với cảm xúc, tính cách, trạng thái tâm lý của cá nhân. Cuối cùng, các sự kiện là các hoạt động cấp cao mô tả các hành động xã hội giữa các cá nhân và chỉ ra ý định hoặc vai trò xã hội của một người.

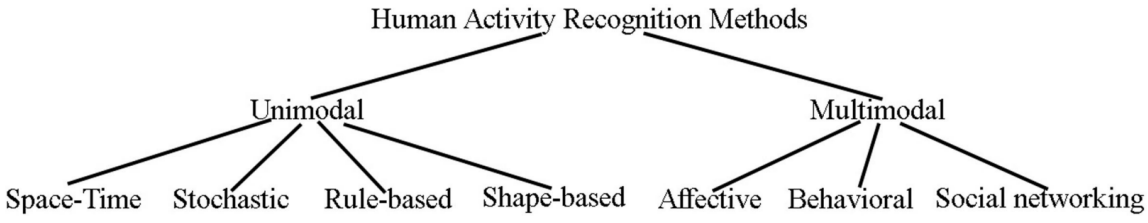


Hình 2.1 Minh họa về một số hành động của con người

Còn đối với nhóm bài toán phân loại hành động yêu cầu đầu ra là xác định loại hành động diễn ra trong một đoạn video. Trong nhánh này lại được phân ra hai bài toán riêng biệt là biểu diễn hành động và nhận dạng tương tác người máy.

2.2. Các thuật toán nhận dạng hành động sử dụng học sâu

Có thể chia các phương pháp nhận dạng học sâu theo sơ đồ sau được tham khảo tại [1].



Hình 2.2 Proposed hierarchical categorization of human activity recognition methods

2.2.1. Phương pháp Unimodal Methods

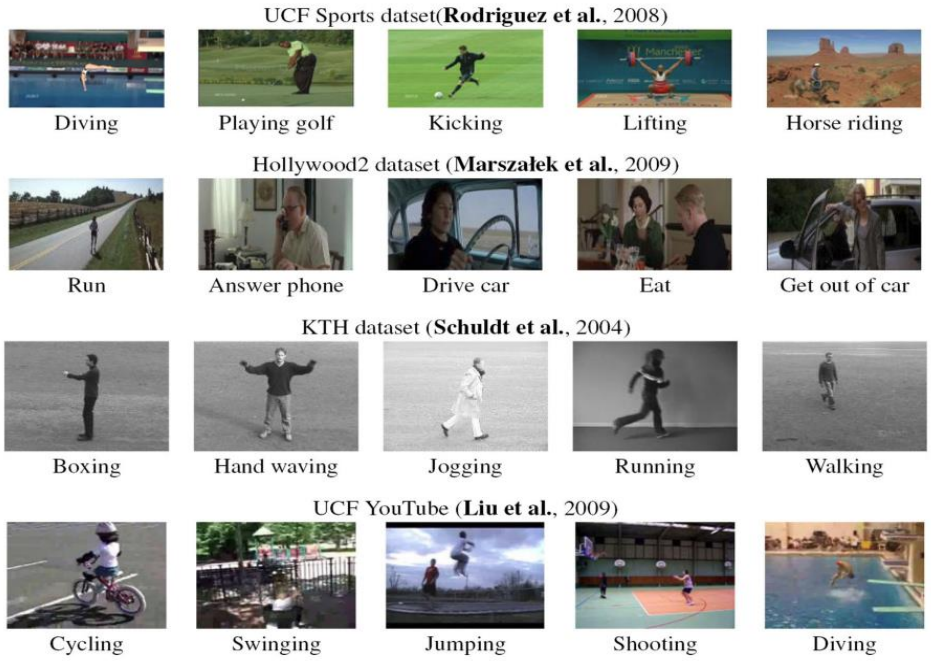
Unimodal human activity recognition methods là nhận dạng và dự đoán hành động dựa trên một phương thức nhất định [1]. Hầu hết các phương pháp hiện có thể hiện các hoạt động của con người dưới dạng một tập hợp các tính năng trực quan được trích xuất từ các chuỗi video hoặc hình ảnh tĩnh và nhận dạng nhãn hoạt động cơ bản bằng một số mô hình phân loại. Unimodal phù hợp để nhận biết các hoạt động của con người dựa trên các đặc điểm chuyển động. Tuy nhiên, khả năng nhận ra lớp cơ bản chỉ từ chuyển động là một nhiệm vụ đầy thách thức. Vấn đề chính là làm thế nào chúng ta có

thể đảm bảo tính liên tục của chuyển động theo thời gian khi một hành động xảy ra đồng nhất hoặc không đồng nhất trong một chuỗi video. Một số phương pháp sử dụng các đoạn quỹ đạo chuyển động , trong khi các phương pháp khác sử dụng toàn bộ chiều dài của đường cong chuyển động bằng cách theo dõi các tính năng dòng quang học. Có thể chia nhỏ phương pháp này thành bốn loại sau:

- space-time
- stochastic
- rule-based
- shapebased approaches.

2.2.2. Phương pháp Space-Time Methods

Các phương pháp tiếp cận Space-Time tập trung vào việc nhận biết các hoạt động dựa trên các đặc điểm Space-Time hoặc dựa trên quỹ đạo phù hợp. Họ xem xét một hoạt động trong khối Space-Time 3D, bao gồm sự nối các không gian 2D trong thời gian. Một hoạt động được thể hiện bằng một tập hợp các tính năng hoặc quỹ đạo không thời gian được trích xuất từ chuỗi video.



Hình 2.3 Một số ví dụ về cách tiếp cận Space-Time dựa trên quỹ đạo dày đặc và bộ mô tả chuyển động

Một nhóm phương pháp chính dựa trên dòng quang học, đã được chứng minh là một gợi ý quan trọng. các hành động của con người được nhận dạng từ các chuỗi video thể thao có độ phân giải thấp bằng cách sử dụng bộ phân loại hàng xóm gần nhất, trong đó con người được thể hiện bằng các cửa sổ có chiều cao 30 pixel.

2.2.3. Stochastic Methods

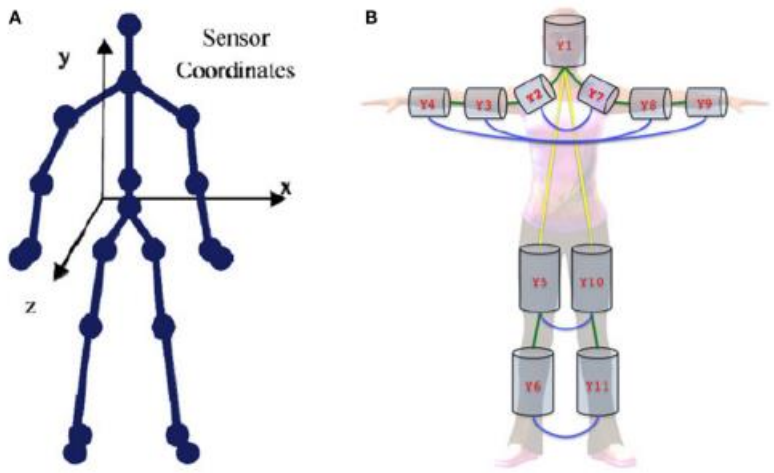
Các nhà nghiên cứu đã nghĩ ra và sử dụng nhiều kỹ thuật ngẫu nhiên, chẳng hạn hidden Markov model (HMM)[8]và hidden conditional random fields (HCRF[9], để suy ra các kết quả hữu ích cho việc nhận dạng hoạt động của con người.

2.2.4. Rule-Based Methods

Rule-based là phương pháp tiếp cận dựa trên quy tắc xác định các sự kiện đang diễn ra bằng cách mô hình hóa một hoạt động bằng cách sử dụng các quy tắc hoặc bộ thuộc tính mô tả một sự kiện [10]. Mỗi hoạt động được coi là một tập hợp các quy tắc/thuộc tính nguyên thủy, cho phép xây dựng mô hình mô tả để ghi nhận hoạt động của con người. Nhận dạng hành động của các cảnh phức tạp với nhiều đối tượng. Mỗi chủ thể phải tuân theo một tập hợp các quy tắc nhất định trong khi thực hiện một hành động.

2.2.5. Shape-Based Methods

Mô hình hóa tư thế và ngoại hình của con người đã nhận được phản hồi tích cực từ các nhà nghiên cứu trong những thập kỷ qua. Các bộ phận của cơ thể người được mô tả trong không gian 2D dưới dạng các mảng hình chữ nhật và dưới dạng hình thể tích trong không gian 3D [11]. Ai cũng biết rằng thuật toán nhận dạng hoạt động dựa trên bóng người đóng vai trò quan trọng trong việc nhận biết hành động của con người. Vì hình bóng người bao gồm các chi được nối với nhau, điều quan trọng là phải lấy được các bộ phận cơ thể người chính xác từ video. Vấn đề này được coi là một phần của quá trình nhận dạng hành động. Nhiều thuật toán truyền tải rất nhiều thông tin về việc giải quyết vấn đề này.



Hình 2.4 Ví dụ về phương pháp Shape-base

2.2.6. Multimodal Methods

Một sự kiện có thể được mô tả bằng các loại tính năng khác nhau cung cấp nhiều thông tin hữu ích hơn. Trong bối cảnh này, một số multimodal methods dựa trên sự hợp nhất tính năng, có thể được thể hiện bằng hai chiến lược khác nhau: hợp nhất sớm và hợp nhất muộn [1]. Cách dễ nhất để đạt được lợi ích của nhiều tính năng là ghép nối trực tiếp các tính năng trong một vectơ tính năng lớn hơn và sau đó tìm hiểu hành động cơ bản. Kỹ thuật kết hợp đặc trưng này có thể cải thiện hiệu suất nhận dạng, nhưng vectơ đặc trưng mới có kích thước lớn hơn nhiều.

Multimodal methods được phân thành ba loại:

- affective methods
- behavioral methods
- methods based on social networking

Multimodal methods mô tả các hành động hoặc tương tác nguyên tử có thể tương ứng với trạng thái tình cảm của một người mà người đó giao tiếp và phụ thuộc vào cảm xúc và/hoặc chuyển động cơ thể.

2.2.7. Affective Methods

Cốt lõi của trí tuệ cảm xúc là hiểu được bản đồ giữa trạng thái cảm xúc của một người và các hoạt động tương ứng, có liên quan chặt chẽ đến trạng thái cảm xúc và giao tiếp của một người với người khác [12]. Các nghiên cứu về điện toán tình cảm mô hình hóa khả năng của một người thể hiện, nhận biết và kiểm soát các trạng thái tình cảm của họ về cử chỉ tay, nét mặt, thay đổi sinh lý, lời nói và nhận dạng hoạt động. Lĩnh vực

nghiên cứu này thường được coi là sự kết hợp của thị giác máy tính, nhận dạng mẫu, trí tuệ nhân tạo, tâm lý học và khoa học nhận thức.

Một vấn đề quan trọng trong tính toán tình cảm là dữ liệu được chú thích chính xác. Xếp hạng là một trong những công cụ chú thích phổ biến nhất. Tuy nhiên, đây là một thách thức để đạt được trong các tình huống trong thế giới thực, vì các sự kiện tình cảm được thể hiện theo một cách khác bởi những người khác nhau hoặc xảy ra đồng thời với các hoạt động và cảm xúc khác. Tiền xử lý các chú thích tình cảm có thể gây bất lợi cho việc tạo các mô hình tình cảm chính xác và mơ hồ do các biểu diễn sai lệch của chú thích ảnh hưởng.

2.2.8. Behavioral Methods

Recognizing human behaviors [13] từ các chuỗi video là một nhiệm vụ đầy thách thức đối với cộng đồng thị giác máy tính. Một hệ thống nhận dạng hành vi có thể cung cấp thông tin về tính cách và trạng thái tâm lý của một người, và các ứng dụng của nó thay đổi từ giám sát bằng video cho đến tương tác giữa người với máy tính. Phương pháp tiếp cận hành vi nhằm mục đích nhận ra các thuộc tính hành vi, tín hiệu đa phương thức phi ngôn ngữ, chẳng hạn như cử chỉ, nét mặt và tín hiệu thính giác. Các yếu tố có thể ảnh hưởng đến hành vi của con người có thể được chia thành nhiều thành phần, bao gồm cảm xúc, tâm trạng, hành động và tương tác với người khác. Do đó, việc nhận ra các hành động phức tạp có thể rất quan trọng để hiểu hành vi của con người. Một khía cạnh quan trọng của nhận dạng hành vi con người là lựa chọn các tính năng phù hợp, có thể được sử dụng để nhận dạng hành vi trong các ứng dụng, chẳng hạn như chơi game và sinh lý học.

2.2.9. Methods Based on Social Networking

Tương tác xã hội là một phần quan trọng trong cuộc sống hàng ngày. Một thành phần cơ bản của hành vi con người là khả năng tương tác với người khác thông qua hành động của họ [1]. Tương tác xã hội có thể được coi là một loại hoạt động đặc biệt trong đó một người điều chỉnh hành vi của mình tùy theo nhóm người xung quanh. Hầu hết các hệ thống mạng xã hội ảnh hưởng đến hành vi của mọi người, chẳng hạn như Facebook, Twitter và YouTube, đo lường các tương tác xã hội và suy ra cách các trang web đó có thể liên quan đến các vấn đề về danh tính, quyền riêng tư, vốn xã hội, văn hóa giới trẻ và giáo dục. Hơn nữa, lĩnh vực tâm lý học đã thu hút sự quan tâm lớn trong

việc nghiên cứu các tương tác xã hội, vì các nhà khoa học có thể suy ra thông tin hữu ích về hành vi của con người.

2.3. Phương pháp Shape-Based với YOLOv7

2.3.1. Lý do chọn phương pháp Shape-Base với YOLOv7 cho bài tiểu luận

Trong các phương pháp được giới thiệu ở trên thì Shape-Base trong những năm gần đây là một trong những phương pháp được chú ý và nghiên cứu nhiều nhất, bởi những kết quả đầy hướng hẹn trong việc nhận dạng cử chỉ hành động của con người. Đó cũng là lý do tại sao bàn tiểu luận này tiếp cận bằng phương pháp Shape-Base.

Với phương pháp Shape-Based nhóm đã thử hiện trên một số mô hình được đánh giá cao về mặt hiệu suất hoặc về mặt thời gian. Sau khi thử một số mô hình như Mediapipe Pose [6], ViTPose-G [14], YOLOv7 [4]thì nhóm nhận thấy Mediapipe Pose có thời gian dự đoán nhanh nhất trên cả thiết bị GPU và CPU, tuy nhiên độ chính xác của mô hình này lại thấp nhất so với hai mô hình còn lại. Với ViTPose-G thì nhóm nhận thấy mô hình đạt độ chính xác cao nhất trong ba mô hình tuy nhiên ViTPose-G là mô hình transformer nên có độ phức tạp cao và thời gian chạy là lâu nhất. Với YOLOv7 thì là mô hình cân bằng giữa thời gian chạy và độ chính xác so với hai mô hình còn lại.

2.3.2. Giới thiệu YOLOv7

Trong bài tiểu luận này nhóm sử dụng mô hình YOLOv7 cho nhiệm vụ nhận dạng các keypoint trên cơ thể con người. Lý do mà nhóm chọn YOLOv7 là vì sau khi tìm hiểu các phương pháp sử dụng deep learning khác thì YOLOv7 đã cân bằng được giữa tốc độ tính toán và độ chính xác. Một vài mô hình khác mà nhóm đã tìm hiểu như là MediaPipe pose do google phát triển có tốc độ xử lý nhanh tụy nhiên độ chính xác thấp hơn. Còn với ViTPose-G là mô hình transformer có độ chính xác cao nhất trên tập dữ liệu COCO tuy nhiên lại tốn nhiều thời gian tính toán.

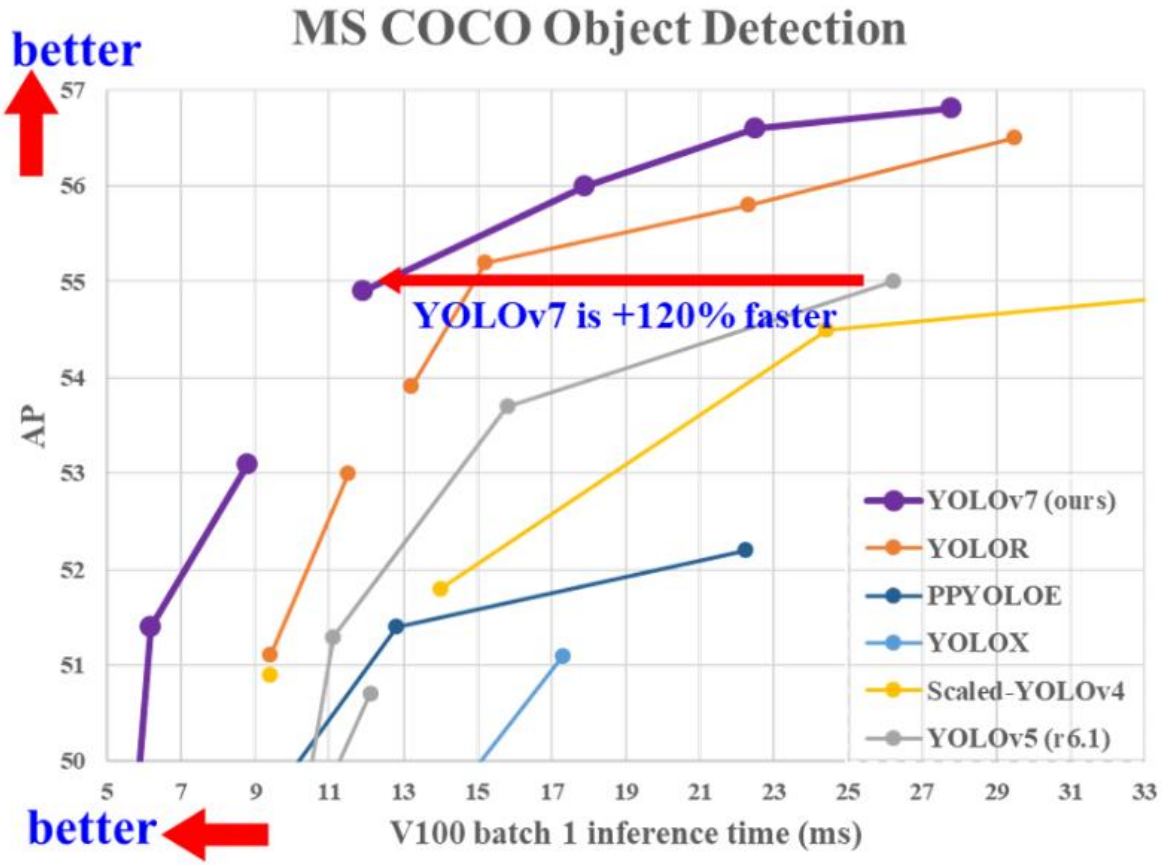
You Only Look Once (YOLO) là một trong những kiến trúc mô hình và thuật toán phát hiện đối tượng phổ biến nhất . Nó sử dụng một trong những kiến trúc mạng thần kinh tốt nhất để tạo ra độ chính xác cao và tốc độ xử lý tổng thể, đó là lý do chính khiến nó trở nên phổ biến. Nếu chúng ta tìm kiếm thuật toán phát hiện đối tượng trên Google, kết quả đầu tiên sẽ liên quan đến mô hình YOLO.

Thuật toán YOLO nhằm mục đích dự đoán một lớp của đối tượng và bounding box xác định vị trí đối tượng trên ảnh đầu vào. Nó nhận ra từng hộp giới hạn bằng bốn thông số sau:

- Trung tâm của hộp giới hạn (x,y)
- Chiều rộng của hộp (w)
- Chiều cao của hộp (h)
- Độ tin cậy của dự đoán (confident)

YOLOv7 là mô hình YOLO version 7 với thông số được trích dẫn sau:

"YOLOv7 vượt qua mọi model Object Detection trong cả tốc độ và độ chính xác từ 5 FPS tới 160 FPS và đạt độ chính xác cao nhất với 56.8% AP trong số toàn bộ các model Object Detection real-time, có tốc độ 30 FPS hoặc hơn trên GPU V100. YOLOv7-E6 (56 FPS trên V100, 55.9% AP) vượt qua cả backbone nhà Transformer là SWIN-L Cascade-Mask R-CNN (9.2 FPS trên A100, 53.9% AP) với 509% về tốc độ và 2% về AP, hay là cả các backbone CNN cao cấp như ConvNeXt-XL Cascade-Mask R-CNN (8.6 FPS trên A100, 55.2% AP) với 551% về tốc độ và 0.7% về AP. Và đương nhiên là YOLOv7 cũng vượt qua cả: YOLOR, YOLOX, Scaled-YOLOv4, YOLOv5, DETR, Deformable DETR, DINO-5scale-R50, ViT-Adapter-B cũng như là rất nhiều các mạng Object Detection khác cả về mặt tốc độ cũng như là độ chính xác. Hơn nữa, YOLOv7 được train trên COCO từ đầu mà không sử dụng bất kì pretrained nào."[4]



Hình 2.5: So sánh các mô hình real-time object detectors

YOLOv7 pose được nhóm tác giả YOLOv7 thực hiện dựa trên một phương pháp phát hiện pose trước đó cho mô hình YOLOv5, với paper tên "YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object" [5].

2.4. Một số bộ dữ liệu tiêu chuẩn trong nhận dạng hành động người

Nhiều bộ dữ liệu hiện có cho bài toán human activity recognition đã được tạo ra trong môi trường được kiểm soát, với những người tham gia thực hiện các hành động cụ thể. Hơn nữa, một số bộ dữ liệu không phải là chung chung, mà bao gồm một nhóm hoạt động cụ thể, chẳng hạn như thể thao và các hành động đơn giản, thường được thực hiện bởi một diễn viên. những bộ dữ liệu này vẫn còn phổ biến để phân loại hoạt động của con người, vì chúng cung cấp một tiêu chí đánh giá tốt cho nhiều phương pháp mới. Một vấn đề quan trọng trong việc xây dựng bộ dữ liệu nhận dạng hoạt động của con người phù hợp là chú thích của từng hành động, thường được người dùng thực hiện thủ công, khiến nhiệm vụ bị sai lệch.

Các tập dữ liệu được nhóm thực hiện khảo sát là KTH [15], Weizman [16] , UCF Sports [17] , MuHAVi [18] , UCF50 [19], UCF101 [20].

Dataset name and category	Year	classes	actors	videos	Resolution
KTH (Schuldt et al., 2004)	2004	6	25	2,391	160 × 120
Weizman (Blank et al., 2005)	2005	10	9	90	180 × 144
UCF Sports (Rodriguez et al., 2008)	2008	9		200	720 × 480
MuHAVi (Singh et al., 2010)	2010	17	14	6,676	720 × 576
UCF50 (Reddy and Shah, 2013)	2013	50			
UCF101 (Soomro et al., 2012)	2012	101		13,320	320 × 240

Bảng 1:Thống kê một số bộ dữ liệu cho bài toán Single action recognition

Các tập dữ liệu được nhóm thực hiện khảo sát là UCF YouTube [21], Hollywood2 [22] , HMDB51 [23], TVHI [24].

Dataset name and category	Year	classes	actors	videos	Resolution
UCF YouTube (Liu et al., 2009)	2009	11		> 1.100	
Hollywood2 (Marszałek et al., 2009)	2009	12		3,669	
HMDB51 (Kuehne et al., 2011)	2011	51		6,849	320 × 240
TVHI (Patron-Perez et al., 2012)	2012	4	20	300	320 × 240

Bảng 2 :Thống kê một số bộ có dữ liệu là Movie

Các tập dữ liệu được nhóm thực hiện khảo sát là PETS 2004 [25], PETS 2007 [26], VIRAT [27].

Dataset name and category	Year	classes	actors	videos	Resolution
PETS 2004 (CAVIAR) (Fisher, 2004)	2004	6		28	384 × 288
PETS 2007 (Fisher, 2007b)	2007	3		7	768 × 576
VIRAT (Oh et al., 2011)	2011	23		17	1,920 × 1,080

Bảng 3: Thống kê một số bộ dữ liệu Surveillance

Các tập dữ liệu được nhóm thực hiện khảo sát là TUM Kitchen [28], Tow-person interaction [29], MSRC-12 Kinect gesture [30], J-HMDB [31] ,UPCV Action [32].

Dataset name and category	Year	classes	actors	videos	Resolution
TUM Kitchen (Tenorth et al., 2009)	2009	10	4	20	324 × 288
Two-person interaction (Yun et al., 2012)	2012	8	7	≈300	640 × 480
MSRC-12 Kinect gesture (Fothergill et al., 2012)	2012	12	30	594	
J-HMDB (Jhuang et al., 2013)	2013	21	1	928	240 × 320
UPCV Action (Theodorakopoulos et al., 2014)	2014	10	20	≈200	

Bảng 4: Thống kê một số bộ dữ liệu Pose

Các tập dữ liệu được nhóm thực hiện khảo sát là URADL [33], ADL [34], MPII [35], Breakfast [36].

Dataset name and category	Year	classes	actors	videos	Resolution
URADL (Messing et al., 2009)	2009	17	5	150	1,280 × 720
ADL (Pirsiavash and Ramanan, 2012)	2012	18	20	10 h	1,280 × 960
MPII Cooking (Rohrbach et al., 2012)	2012	65	12	44	1,624 × 1,224
Breakfast (Kuehne et al., 2014)	2014	10	52	≈77 h	320 × 240

Bảng 5: Thống kê một số bộ dữ liệu Daily living

Các tập dữ liệu được nhóm thực hiện khảo sát là CCV [37], FPSI [38], Broadcast field hockey [39], USAA [40], Sport-1M [41], ActivityNet [42], WWW Crowd [43].

Dataset name and category	Year	classes	actors	videos	Resolution
CCV (Jiang et al., 2011)	2001	20		9.317	
FPSI (Fathi et al., 2012)	2012	6	8	≈42 h	1,280 × 720
Broadcast field hockey (Lan et al., 2012b)	2012	11		58	
USAA (Fu et al., 2012)	2012	8		≈200	
Sports-1M (Karpathy et al., 2014)	2014	487		1 M	
ActivityNet (Heilbron et al., 2015)	2015	203		27,801	1,280 × 720
WWW Crowd (Shao et al., 2015)	2015	94		10,000	640 × 360

Bảng 6: Thống kê một số bộ dữ liệu Social Networking

Các tập dữ liệu được nhóm thực hiện khảo sát là BEHAVE [44], Canal9 [45], USC Createive IT [46], Parliament [47].

Dataset name and category	Year	classes	actors	videos	Resolution
BEHAVE (Fisher, 2007a)	2007	8		321	640 × 480
Canal9 (Vinciarelli et al., 2009)	2009	2	190	≈42 h	720 × 576
USC Creative IT (Metallinou et al., 2010)	2010	50	16	100	
Parliament (Vrigkas et al., 2014b)	2014	3	20	228	320 × 240

Bảng 7: Thống kê một số bộ dữ liệu về Behavior

2.5. Các thách thức của bài toán nhận dạng hành động con người

Việc phát triển một hệ thống nhận dạng hoạt động của con người hoàn toàn tự động, có khả năng phân loại các hoạt động của con người với sai số thấp, là một nhiệm vụ đầy thách thức do các vấn đề, chẳng hạn như lộn xộn nền, che khuất một phần, thay đổi tỷ lệ, góc nhìn, ánh sáng và hình thức cũng như độ phân giải khung hình . Ngoài ra, việc chú thích vai trò hành vi tốn nhiều thời gian và yêu cầu kiến thức về sự kiện cụ thể. Hơn nữa, những điểm tương đồng giữa các lớp và giữa các lớp làm cho vấn đề trở nên khó khăn hơn. Nghĩa là, các hành động trong cùng một lớp có thể được thể hiện bởi những người khác nhau với các chuyển động cơ thể khác nhau và các hành động giữa các lớp khác nhau có thể khó phân biệt vì chúng có thể được biểu thị bằng thông tin tương tự. Cách thức mà con người thực hiện một hoạt động phụ thuộc vào thói quen của họ và điều này làm cho vấn đề xác định hoạt động cơ bản trở nên khá khó xác định. Ngoài ra, việc xây dựng một mô hình trực quan để học và phân tích chuyển động của con người trong thời gian thực với bộ dữ liệu chuẩn không đầy đủ để đánh giá là một.

CHƯƠNG 3: ỨNG DỤNG PHƯƠNG PHÁP NHẬN DẠNG HÀNH ĐỘNG CON NGƯỜI VÀO BÀI TOÁN ĐẾM HÀNH ĐỘNG TRONG QUÁ TRÌNH TẬP THỂ HÌNH

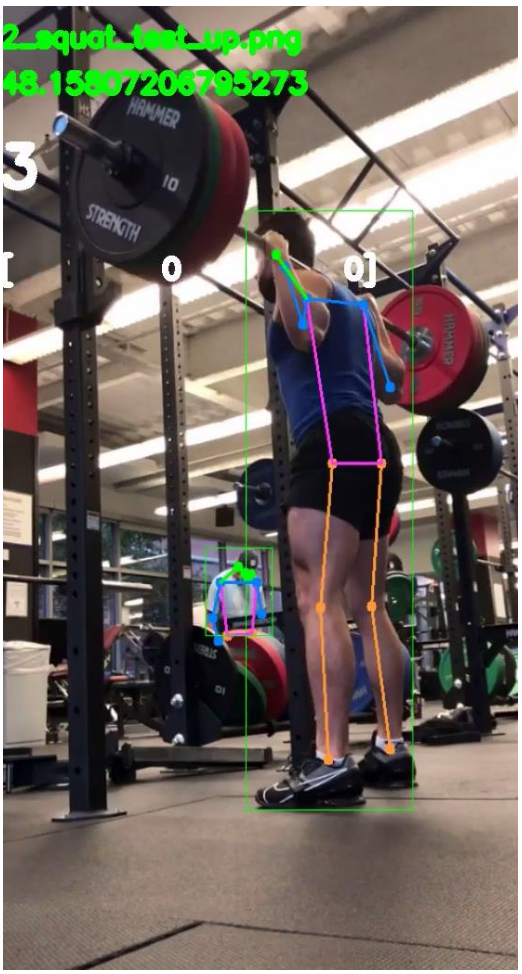
3.1. Bài toán đếm hành động trong quá trình tập thể hình

Artificial intelligence technology đã thực sự cần thiết trong nhiều ngành công nghiệp bao gồm cả ngành thể dục. Ước tính tư thế con người là một trong những nghiên cứu quan trọng trong lĩnh vực Thị giác máy tính trong vài năm qua. Trong dự án này, các kỹ thuật ước tính tư thế và học máy sâu được kết hợp để phân tích hiệu suất và báo cáo phản hồi về số lần lặp lại của các bài tập đã thực hiện trong thời gian thực. Sử dụng công nghệ máy học trong ngành thể dục có thể giúp trọng tài đếm số lần lặp lại của bất kỳ bài tập nào trong các cuộc thi Cử tạ hoặc CrossFit .

Dự án cung cấp giải pháp đếm số lần lặp lại một bài tập thể chất trong thời gian thực. Phương pháp này sử dụng ước tính tư thế để theo dõi các vận động viên, nhận ra các bài tập đã thực hiện của họ, đếm số lần lặp lại và phân tích hiệu suất của các lần lặp lại.

Trong các cuộc thi Crossfit, mỗi vận động viên được bắt cặp với một giám khảo riêng. Giám khảo đếm và đánh giá số lần lặp lại của vận động viên. Trong các cuộc thi lớn, các giám khảo bị giới hạn trong việc đếm số lần lặp lại mà không đánh giá. Việc sử dụng công nghệ học sâu trong lĩnh vực này có thể giảm bớt số lượng giám khảo tham gia và nâng cao tính khách quan cho các cuộc thi.

Hay nếu bạn là một người tập thể hình tự do tại nhà và muốn theo dõi thông số của bạn thân thì việc sử dụng ứng dụng đếm số lần lặp lại của từng động tác sẽ giúp cho bạn có thể tập trung hoàn toàn vào việc tập luyện trong khi đã có AI đếm cho bạn.



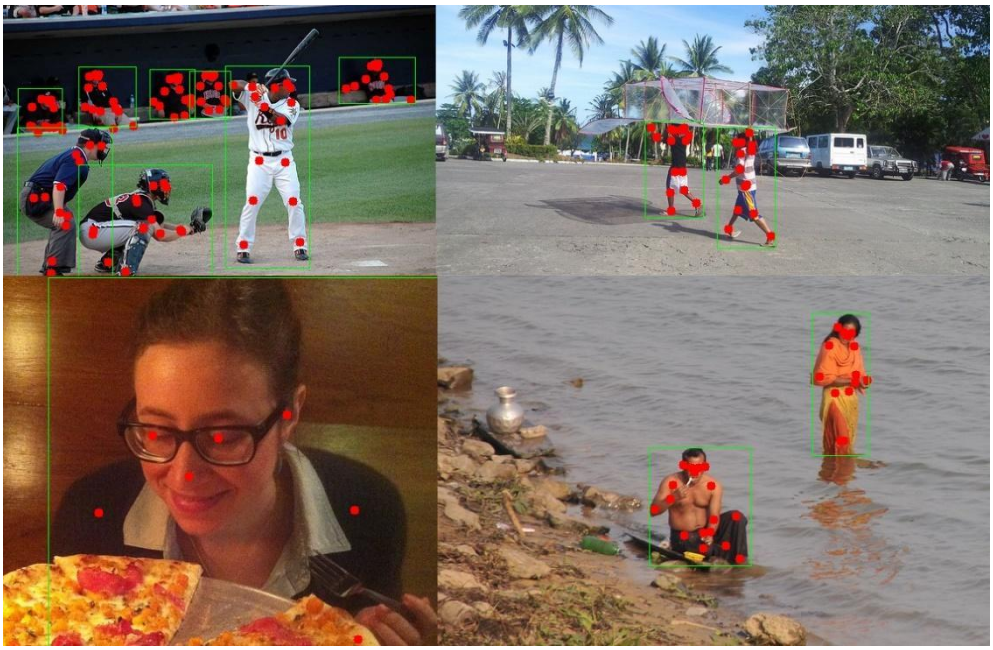
Hình 3.1: Ứng dụng đếm số lần squat của người tập

3.2. Bộ dữ liệu sử dụng

Tập dữ liệu được sử dụng để training mô hình là tập coco 2017, với nhãn được nhóm tác giả YOLOv7 cung cấp trên github. [2]Có tất cả 2.346 file annotation cho tập validation và 56,599 file annotation cho tập Training. Trong mỗi file annotation có n (n là số người trong bức ảnh kèm theo) dòng với mỗi dòng gồm 56 số viết liên tiếp cách nhau bởi dấu cách " ".

- Số đầu tiên (anno[0]) là class của pose, tuy nhiên thì tất cả các pose đề có class là 0. Vì YOLO là họ mô hình được phát triển cho bài toán object detection nên khi chuyển sang detection pose thì vẫn giữ lại thành phần class.
- Bốn số tiếp theo (anno[1:5]) lần lượt là x,y,w,h của bbox của người trong khung hình.
- Các số còn lại (anno[5:]) lần lượt là các cặp điểm (x1,y1,confident1) …. (x17,y17,confident17).

Chú ý là tất cả các tọa độ x,y được chuyển về dưới dạng tỷ về với chiều cao và chiều rộng của bức ảnh.



Hình 3.2: Biểu diễn nhãn của dữ liệu lên hình ảnh

Trong dữ liệu hay paper không đề cập tới nhãn của từng điểm, tuy nhiên thứ tự của các điểm trong file annotation có thứ tự từ trên đỉnh đầu xuống thân và xuống chân.

Phần dữ liệu hình ảnh có thể download từ trang chủ của COCO dataset.[3]COCO dataset là một tập dữ liệu có độ khó cao trong nhiệm vụ nhận dạng keypoint trên cơ thể người, vì tính phức tạp của dữ liệu, có những con người ở xa khung ảnh, hoặc có những bộ phận thì khuất trong khung hình, hay cơ thể không được biểu diễn liên tục trong khung hình.

3.3. Training model YOLOv7

3.3.1. Thông số thiết bị training mô hình

Để training model YOLOv7 nhóm đã thực hiện trên cơ sở hạ tầng của Google Colab pro với chế độ GPU. Với thông tin của máy là:

- GPU: Tesla T4
- RAM: ~12.6 GB Available
- Disk: ~33 GB Available

Do giới hạn về tài nguyên sử dụng cũng như một vài hạn chế của đề tài đã nêu trên nhóm thực hiện training 1 epoch với toàn bộ tập train (56.599 tấm ảnh). Với mục tiêu chính là hiểu dữ liệu, hiểu mô hình, cài đặt và ứng dụng mô hình.

3.3.2. Các bước chuẩn bị training

```
[1] !nvidia-smi -L

GPU 0: Tesla T4 (UUID: GPU-e7847a6a-073a-5b46-aa22-85cbbd4e6c21)

[2] from google.colab import drive
    drive.mount('/content/drive')

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

Setup

```
[3] import shutil
```

```
!mkdir coco_kpts
%cd coco_kpts
shutil.copy("/content/drive/MyDrive/TLCN/coco2017labels-keypoints.zip",'/content/coco_kpts')
!unzip /content/coco_kpts/coco2017labels-keypoints.zip -d /content/coco_kpts
!mkdir images
%cd images
!wget http://images.cocodataset.org/zips/train2017.zip
!wget http://images.cocodataset.org/zips/val2017.zip

!unzip train2017.zip
!unzip val2017.zip
```

Hình 3.3: Setup dữ liệu chuẩn bị cho quá trình training

Đầu tiên thực hiện download dữ liệu về workspace của colab và unzip. File label được copy ra từ drive của nhóm.

```
[ ] import cv2
    path_anno = '/content/coco_kpts/labels/val2017/000000006471.txt'
    path_img = path_anno.replace('labels', 'images').replace('txt', 'jpg')
    img = cv2.imread(path_img)
```

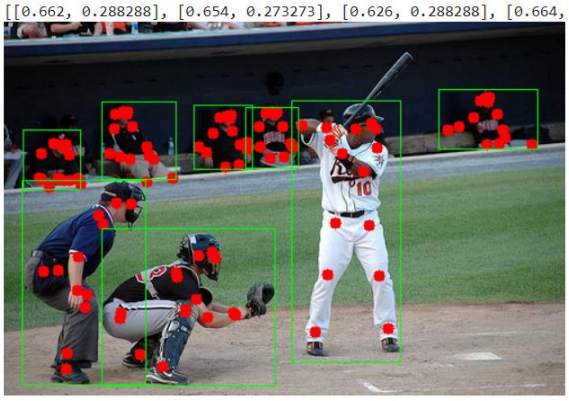
```
import cv2
from google.colab.patches import cv2_imshow
import numpy as np
import os

def take_point(path_anno):
    with open(path_anno, 'r') as f:
        lines = f.readlines()
    points = []
    class_ids = []
    bbox = []
    for line in lines:
        line = line.split(' ')
        class_id = int(line[0])
        for i in range(2, len(line), 3):
            if float(line[i+2]) > 0.9:
              point = [float(line[i]), float(line[i+1])]
              points.append(point)
              class_ids.append(class_id)
        x1,y1 = float(line[1])-float(line[3])/2, float(line[2])-float(line[4])/2
        x2,y2 = float(line[1])+float(line[3])/2, float(line[2])+float(line[4])/2
        bbox.append([x1,y1,x2,y2])
        # break
    return points , class_ids, bbox
```

Hình 3.4: Đoạn mã visualize dữ liệu của từ file annotation

Tiếp theo thực hiện visualize dữ liệu để kiểm tra.

```
/content/coco_kpts/images/val2017/000000006471.jpg
333 500
[[0.662, 0.288288], [0.654, 0.273273], [0.626, 0.288288], [0.664, 0.339339], [0.58, 0.318318], [0.64, 0.402402], [0.53, 0.279279], [0.602,
```



Hình 3.5: Dữ liệu trong file annotation được được visualize

```
!git clone https://github.com/WongKinYiu/yolov7.git
```

```
Cloning into 'yolov7'...
remote: Enumerating objects: 1094, done.
remote: Total 1094 (delta 0), reused 0 (delta 0), pack-reused 1094
Receiving objects: 100% (1094/1094), 69.89 MiB | 16.17 MiB/s, done.
Resolving deltas: 100% (518/518), done.
```

```
%cd yolov7
!git checkout pose
```

```
/content/yolov7
```

```
[ ] !mkdir /content/yolov7/weights
shutil.copy("/content/drive/MyDrive/TLCN/yolov7-w6-person.pt",'/content/yolov7/weights/yolov7-w6-person.pt')
shutil.copy("/content/drive/MyDrive/TLCN/yolov7-w6-pose.pt",'/content/yolov7-w6-pose.pt')
```

```
'/content/yolov7-w6-pose.pt'
```

```
[ ] !pip install -r requirements.txt
```

Hình 3.6: Setup thư viện YOLOv7 và copy pretrain model ra workspace colab

3.3.3. Hyperparameter cho quá trình training

Cuối cùng là lựa chọn Hyperparameter cho quá trình training mô hình. Dưới đây là các Hyperparameter chính mà nhóm truyền vào cho quá trình training.

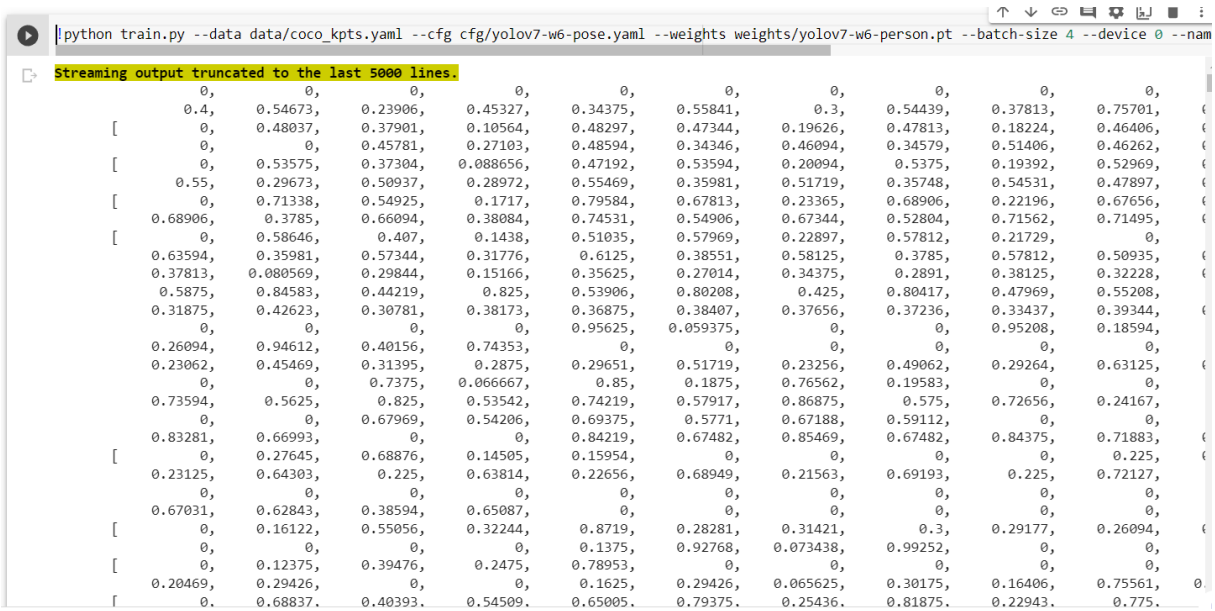
Ý nghĩa của các Hyperparameter:

- data : truyền vào đường dẫn tới file config cho tập dữ liệu.
- cfg : truyền vào đường dẫn tới file config cho mô hình training.
- weights : truyền vào đường dẫn tới file model pre-train.
- batch-size : số lượng ảnh truyền vào trong một lần cập nhật.

- device : truyền vào tài nguyên training model (GPU hoặc CPU).
- name : tên mô hình.
- hyp : đường dẫn đến file config cho các Hyperparameter khác.
- kpt-label : biểu thị kiểu dữ liệu truyền vào là pose.
- epochs : số lần học trên toàn bộ tập dữ liệu.

  Config của các Hyperparameter:
- data data/coco_kpts.yaml
- cfg cfg/yolov7-w6-pose.yaml
- weights weights/yolov7-w6-person.pt
- batch-size 4
- device 0
- name yolov7-w6-pose
- hyp data/hyp.pose.yaml
- kpt-label
- epochs 1

3.3.4. Kết quả của quá trình training model YOLOv7



Hình 3.7: Kết quả sau khi kết thúc training 1 epoch

Kết quả đạt được khi kiểm tra trên tập validation là với 2346 tấm ảnh có 6352 nhãn thì

- Precision là 0.898
- Recall là 8.67
- mAP với ngưỡng 0.5 là 0.938

- mAP với ngưỡng 0.95 là 0.726

```
!python test.py --data data/coco_kpts.yaml --conf 0.001 --iou 0.65 --weights /content/yolov7-w6-pose.pt --kpt-label

Namespace(augment=False, batch_size=32, conf_thres=0.001, data='data/coco_kpts.yaml', device='', dump_img=False, exist_ok=False, flip_test=False,
YOLOv5 🚀 cad7aca torch 1.13.0+cu116 CUDA:0 (Tesla T4, 15109.75MB)

Fusing layers...
/usr/local/lib/python3.8/dist-packages/torch/functional.py:504: UserWarning: torch.meshgrid: in an upcoming release, it will be required to pass
  return _VF.meshgrid(tensors, **kwargs)  # type: ignore[attr-defined]
Model Summary: 494 layers, 80178356 parameters, 80178356 gradients, 101.6 GFLOPS
/usr/local/lib/python3.8/dist-packages/torch/nn/modules/module.py:673: UserWarning: The .grad attribute of a Tensor that is not a leaf Tensor is
  if param.grad is not None:
val: Scanning '../coco_kpts/val2017' images and labels... 2346 found, 0 missing, 0 empty, 0 corrupted: 100% 2346/2346 [00:01<00:00, 1875.20it/s]
val: New cache created: ../coco_kpts/val2017.cache
               Class      Images      Labels           P           R      mAP@.5  mAP@.5:.95: 100% 74/74 [01:15<00:00,  1.02s/it]
                 all        2346        6352       0.898       0.867       0.938       0.726
Speed: 11.0/1.2/12.3 ms inference/NMS/total per 640x640 image at batch-size 32

Evaluating xtcocotools mAP... saving runs/test/exp/yolov7-w6-pose_predictions.json...
xtcocotools unable to run: No module named 'xtcocotools'
/usr/local/lib/python3.8/dist-packages/torch/nn/modules/module.py:673: UserWarning: The .grad attribute of a Tensor that is not a leaf Tensor is
  if param.grad is not None:
Results saved to runs/test/exp
```

Hình 3.8: Hình kết quả khi kiểm tra trên tập validation

Hình 3.9: Hình ảnh dự đoán so với nhãn của dữ liệu

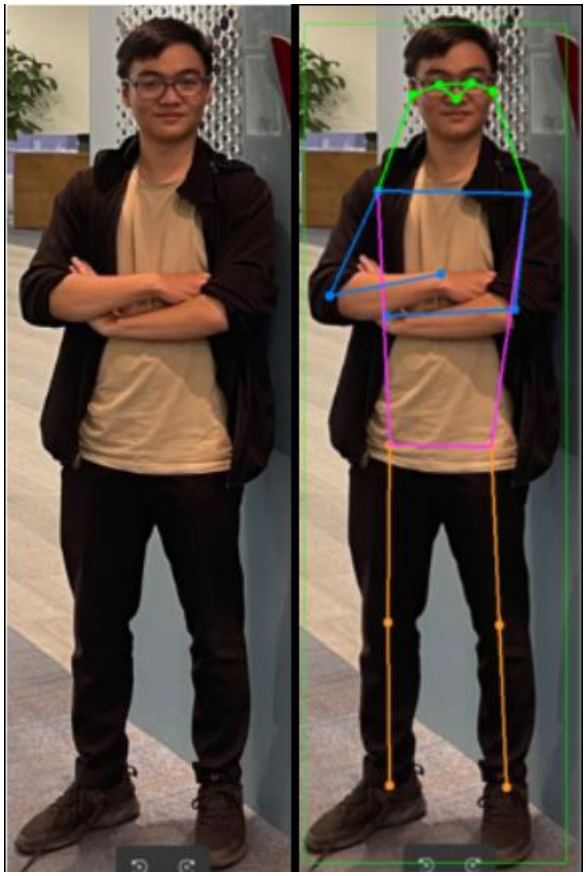
Vì nhóm chỉ thực hiện training trên 1 epoch nên kết quả đạt được không có thay đổi hay cải tiến gì so với mô hình được phát hành trước đó.

3.4. Phương pháp phát hiện hành động từ kết quả biểu diễn hành động con người của yolov7

3.4.1. Giới thiệu phương pháp Encoder pose do nhóm đề xuất

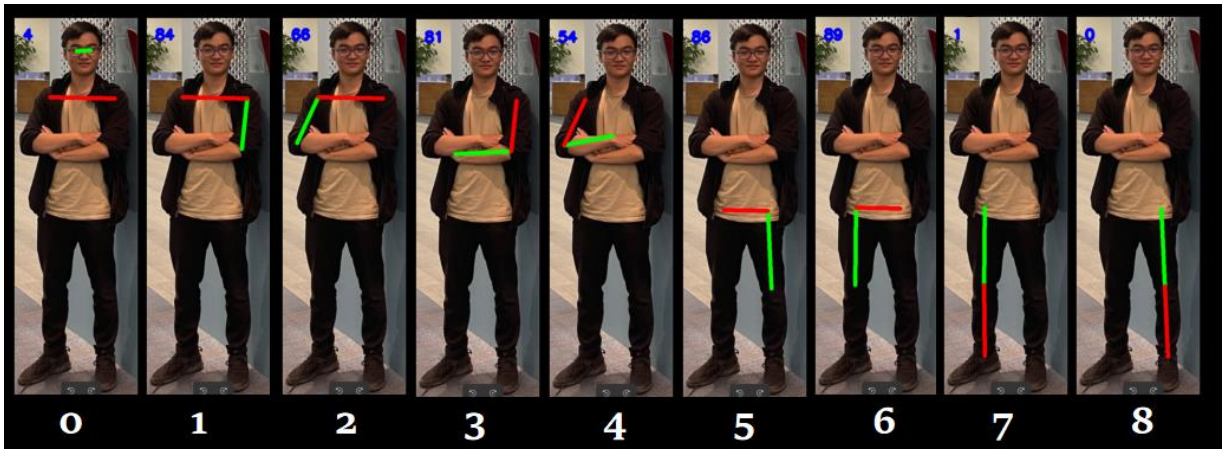
Sau khi kết thúc quá trình training mô hình YOLOv7 pose, thì kết quả nhận được của quá trình predict hình ảnh từ model sẻ cho ra: một mảng numpy array có shape là (n,58). Với n là số nhân vật trong bức ảnh và 58 bảo gồm :

- Hai số đầu: class id, confident class.
- Năm số tiếp theo: x_bbox, y_bbox, w_bbox, h_bbox, confident bbox.
- Các số còn lại: x1, y1, confident1,…., x17, y17, confident17



Hình 3.10: Visualize thông tin predict của model YOLOv7 pose

Sau khi có thông tin human pose thì để so sánh các pose với nhau nhóm thực hiện chuyển đổi biểu diễn pose về biểu diễn góc giữa những khớp nối trên cơ thể. Tất cả 17 điểm keypoints sẽ được chuyển thành một vector chứa thông tin 9 cái góc của các bộ phận trên cơ thể.



Hình 3.11: Các góc được biểu diễn lại từ 17 keypoints

Đây là phương pháp với ý tưởng đã được thực hiện trước đây với một vài mô hình khác và với các khớp khác trên cơ thể. Nhóm đã thực hiện thay đổi để phù hợp với kết quả trả ra của mô hình YOLOv7 và với mục đích của ứng dụng của bài tiểu luận.

3.4.2. Phương pháp so sánh độ tương đồng giữa hai pose

Việc so sánh giữa hai pose sẽ chuyển thành so sánh biểu diễn của hai vectơ. Nhóm thực hiện đo sự sai khác giữa từng góc của 2 pose, sau đó tính tổng lỗi sai khác giữa hai pose. Tuy nhiên tùy thuộc vào mỗi hành động thì có những bộ phận (góc) sẻ đóng vai trò phân loại lớn và số khác sẽ ngược lại, vì thế nhóm thực hiện nhân thêm một mặt nạ cho vector sai khác của pose sau đó mới tính tổng lại.

```
def compare_two_poses(vector1, vector2, mask=np.ones(9)):
  error =  np.sum(np.abs(vector1 - vector2)*mask)
  return error
```

Hình 3.12: Mã code tính sự sai khác giữa 2 pose

Threshold để phân biệt giữa 2 pose sẽ được tính bằng công thức sau đây. Tính số điểm mà mặt nạ khác 0, sau đó nhân với 15. Ý tưởng của phương pháp này là cho phép mỗi bộ phận có thể có sự sai khác trong khoảng 15 độ.

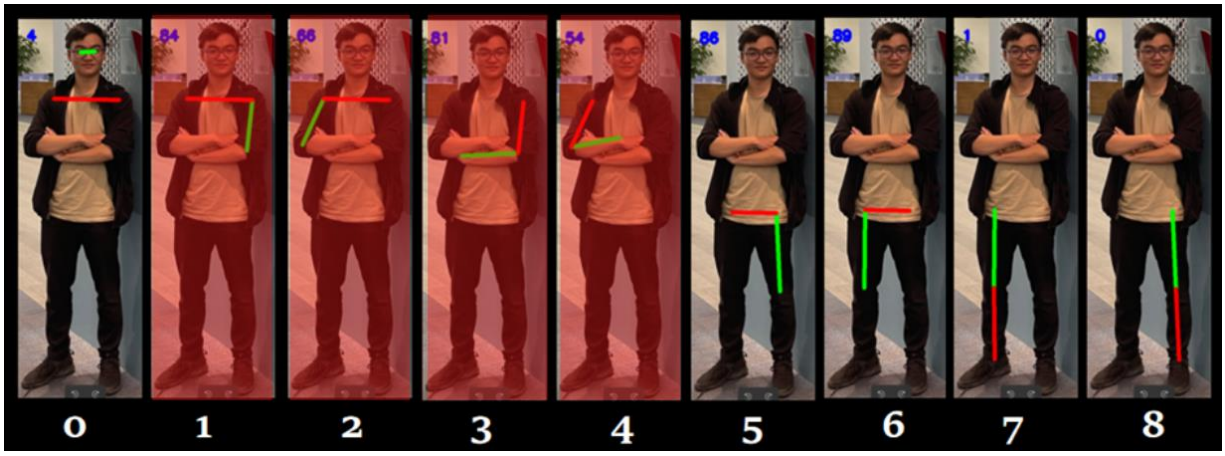
```
squat_mask = np.array([0.5,0,0,0,0,1,1,1,1])
threshold = np.count_nonzero(squat_mask)*15
threshold
```

```
75
```

```
compare_with_list_action(squat_test_up, squat_action_vector, squat_action_name, squat_mask)
```

```
(11.610577874558274, '2_squat_test_up.png')
```

Khi so sánh 2 động tác squat down, thành phần các góc cánh tay được loại bỏ vì động tác squat thường có nhiều bài tập khác nhau về tay đi kèm.



Hình 3.13: Hình minh họa các hành động bị loại bỏ trong quá trình so sánh pose

Phương pháp so sánh độ tương đồng giữa hai pose này được nhóm đề xuất dựa trên ý tưởng của hàm mean absolute error [7].

3.4.3. Phương pháp đếm sự lặp lại của động tác thể hình

Để xây dựng được ứng dụng đếm số lần lặp lại của các hoạt động thể dục, thì nhóm sẽ tạo ra một folder chứa các hành động chính của một bài tập, ví dụ như hành động squat thì sẽ có hành động lên và hành động xuống. Hình ảnh về hai hành động đó sẽ được lưu trong folder squat. Khi truyền vào một video cần được xử lý thì từng frame ảnh trong video sẽ được so sánh với tất các các hành động trong folder và khi xuất hiện 1 hành động lên tiếp sau một hành động xuống thì bài toán sẽ được đếm là 1 lần thực hiện hành động squat.

Hình 3.14: Thông tin tư thế người tập giống nhất so với tổ hợp động tác của bài tập

Thông tin trên cùng là tên hành động mà tư thế người tập giống nhất so với tổ hợp các động tác của bài tập.

Thông tin tiếp theo là sai số của tư thế người tập và động tác có tên trên đó. Hình ảnh màu đỏ trên hai tham số này có nghĩa là lỗi đang lớn hơn ngưỡng và không được tính là một lần thực hiện hành động, ngược lại với màu xanh.

Chữ số màu trắng to nhất biểu thị số lần đã hoàn thành một chu kỳ động tác. Vector màu trắng ở cuối biểu thị tiến độ tập của người thực hiện trong một chu trình tập. Nếu vector toàn là số 0 thì là lúc người tập đang bắt đầu một chu kỳ mới. Kết thúc sẽ là lúc người tập đạt tất cả là số 1 và reset lại vector thành một vector 0.

3.5. Kết quả đạt được của dự án.

3.5.1. Dữ liệu đánh giá bài toán đếm hành động

Dữ liệu được thu thập là dữ liệu thực tế được lấy trên các trang mạng xã hội như Youtube, Facebook,.. và được chuyển đổi sang file MP4 để thuận tiện cho việc cắt ghép chỉnh sửa video. Sau đó video này sẽ được chuyển đổi sang file MP4 và tải xuống để thuận tiện cho việc cắt ghép và chỉnh sửa video. Dữ liệu sẽ được cắt ghép chỉnh sửa bởi phần CapCut để có những góc quay khung hình tối ưu nhất cho việc nhận dạng hành động. Sau đó dữ liệu sẽ được đẩy lên Google Drive để thuận tiện cho việc training và đánh giá.

	Tổng số video	Số lượng hành động	Số người tham gia
Squat	33	528	14
Push up	11	178	10

Bảng 8: Bảng thống kê dữ liệu đánh giá bài toán đếm hành động

3.5.2. Kết quả đánh giá với động tác Squat và động tác Push up

Sau khi chạy code dự đoán hành động người tập nhóm thực hiện kiểm tra và đánh giá độ chính xác dựa trên từng động tác tập. Từ đó nhóm thực hiện tính sự sai khác giữa mỗi hành động và mỗi video. Với mỗi video (full video) nếu có 1 động tác thiếu hoặc dư khi đếm sẽ tính là video đó đã sai.

		Squat		Push up	
Full video	True	25 (video)	75.76%	6 (video)	54.54%
	False	8 (video)	24.24%	5 (video)	45.45%
Each action cycle	True	497 (action)	94.13%	164 (action)	92.14%
	False	31 (action)	5.87%	14 (action)	7.86%

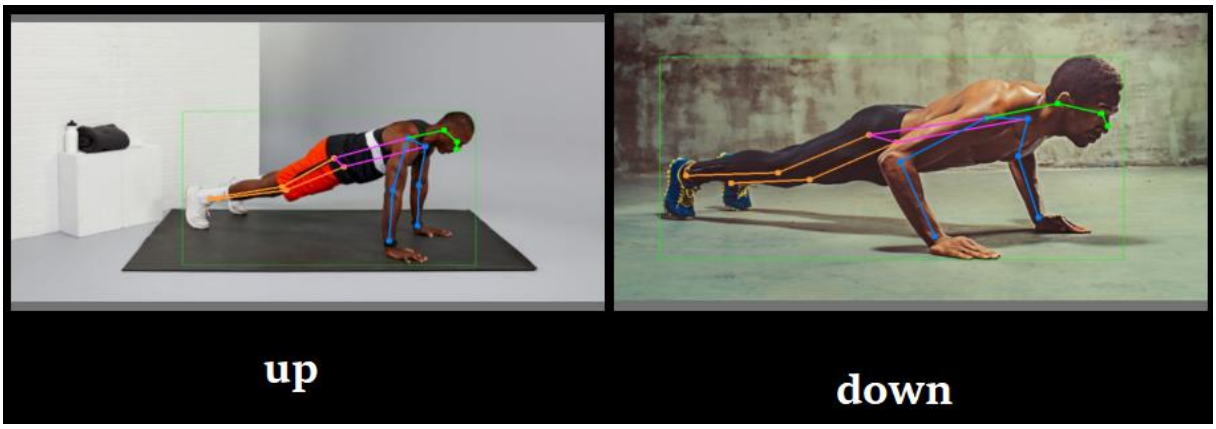
***Bảng** 9: Bảng kết quả đánh giá trên tập dữ liệu test.*

3.5.3. Giải thích kết quả đạt được

Kết quả nhận dạng trên mỗi hành động (each action cycle) đạt trên 92% cho thấy phương pháp nhận dạng chu kỳ hành động do nhóm đề xuất đạt hiệu quả cao trong việc nhận dạng chu kỳ hành động, tuy nhiên khi đánh giá trên mỗi video (full video) thì tỉ lệ giảm xuống chỉ đạt 75% với Squat và 54% với Push up.

Lý do là khi tập một trong các động tác của người tập chưa đạt được sự chính xác mà động tác tập yêu cầu, có thể lên chưa đủ độ cao hoặc xuống chưa đủ sâu.

Ngoài ra có sự chân lệch lớn giữa sai số của hai động tác là Push up và Squat, cho thấy động tác của hành động tập ảnh hưởng tới kết quả của bài toán, vì với hành động Push up, cơ thể của người tập giữa hai trạng thái up và down chỉ khác nhau ở góc giữa hai cổ tay, do đó sai số tăng lên khi số góc biểu diễn của hành động giảm xuống.



Hình 3.15: Hình biểu diễn hai trạng thái của bài tập Push up

PHẦN 3: KẾT LUẬN

1. Ý nghĩa đạt được

Sau khi hoàn thành đề tại "TÌM HIỂU CÁC MÔ HÌNH HỌC SÂU TRONG NHẬN DẠNG HÀNH ĐỘNG NGƯỜI VÀ VIẾT ỨNG DỤNG MINH HỌA" dưới sự chỉ bảo và hướng dẫn của thầy Trần Công Tú, nhóm chúng em đã hoàn thành được các mục tiêu đề ra, cũng như đã áp dụng được kiến thức học được để tạo ra một sản phẩm demo.

1.1. Ý nghĩa khoa học

Bài tiểu luận đã khái quát lại các cơ sở lý thuyết, các khái niệm cơ bản trong Deep Learning, giới thiệu một số mô hình DL phổ biến giúp người đọc có cái nhìn khác quát về lý thuyết, tứ đó hiểu được các phương pháp nhận dạng hành động được trình bày phía sau đó.

Thông qua việc thực hiện đề tài, nhóm chúng em đã kiểm chứng được hiệu suất của mô hình YOLOv7, hiểu được tâm dữ liệu và các tham số training mô hình. Việc thử nghiệm mô hình trên sản phẩm đề mô giúp nhóm phát hiện những hạn chế và ưu điểm của mô hình so với các kiến trúc DL khác.

1.2. Ý nghĩa thực tiễn

Quan việc thực hiện đề tài nhóm đã học thêm được nhiều kiến thức mới, đặc biệt là các thuật toán DL và các phương pháp nhận dạng hành động con người. Nhận dạng hành động con người là một nhánh của computer vision và là nhánh còn nhiều thách thức do sự phức tạp của hành vi, cử chỉ, cảm xúc của con người đưa đến những ý nghĩa về hành động là khác nhau.

Áp dụng được kiến thức được học từ những môn học khác đã giúp nhóm có thể đi đến cuối tiểu luận. Làm việc với cái tài liệu, paper đã giúp khả năng đọc của nhóm tăng cao. Cảm ơn vì những kiến thức quý giá đã học và những thời gian thực hiện dự án đã giúp nhóm trưởng thành hơn rất nhiều.

Mặc dù ứng dụng chỉ dừng lại ở mức demo và chưa hoàn thiện những nhóm đã giới thiệu một phương pháp so sánh độ tương đồng giữa hai pose từ 17 kepoint trên cơ thể và đưa nó vào ứng dụng thực tiễn. Từ đó làm nền móng để có thể phát triển nghiên cứu thêm khi vào đồ án tốt nghiệp.

2. Hạn chế của đề tài

Phương pháp tiếp cận của nhóm yêu cầu hệ thống có GPU để xử lý với tốc độ cao điều này sẽ tăng chi phí khi triển khai thực tiễn.

Hạn chế tiếp theo là độ chính xác khi so sánh hai pose phụ thuộc nhiều vào độ chính xác của bước nhận diện human pose, và các khung hình không có tình liên kết của thời gian và không có áp dụng được biểu diễn của toa độ, điều này là một thiếu sót cho việc nhận dạng hành hành động.

Một số hành động thể dục đặc thù sẽ gây nhầm lẫn cho mô hình nhận dạng pose.

3. Hướng phát triển

Thứ nhất, thu tập và tạo một tập dữ liệu keypoints cho các hành động tập thể dục, sau đó training thêm để cải thiện độ chính xác của mô hình

Thứ hai, có thể tạo một hình DL để Decoder pose thành một vector thay vì chuyển về góc, ngoài ra có thể thêm các mạng RNN hay LSTM để nhận dạng thêm thông tin thời gian và thêm thông tin biểu diễn không gian của các keypoint vào.

Thứ ba deploy backend lên aws và hoàn thiện phần front-end để hoàn thiện hơn sản phẩm.

TÀI LIỆU THAM KHẢO

[1] Vrigkas, M., Nikou, C., & Kakadiaris, I. A. (2015). A review of human activity recognition methods. Frontiers in Robotics and AI, 2, 28.

[2] GitHub - WongKinYiu - WongKinYiu/yolov7 at pose. (2022).Retrieved 25 December 2022, from https://github.com/WongKinYiu/yolov7/tree/pose.

[3] COCO - Common Objects in Context. (2022). Retrieved 25 December 2022, from https://cocodataset.org/#home

[4] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696.

[5] Maji, D., Nagori, S., Mathew, M., & Poddar, D. (2022). YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2637-2646).

[6] Pose. (2022). Retrieved 28 December 2022, from https://google.github.io/mediapipe/solutions/pose.html

[7] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific Model Development Discussions, 7(1), 1525-1534.

[8] Eddy, S. R. (1996). Hidden markov models. Current opinion in structural biology, 6(3), 361-365.

[9] Wang, S. B., Quattoni, A., Morency, L. P., Demirdjian, D., & Darrell, T. (2006, June). Hidden conditional random fields for gesture recognition. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 2, pp. 1521-1527). IEEE.

[10] Sargano, A. B., Gu, X., Angelov, P., & Habib, Z. (2020). Human action recognition using deep rule-based classifier. Multimedia Tools and Applications, 79(41), 30653-30667.

[11] Uddin, M., Lee, J. J., & Kim, T. S. (2008, June). Shape-based human activity recognition using independent component analysis and hidden Markov

model. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 245-254). Springer, Berlin, Heidelberg.

[12] Liang, J., Xu, C., Feng, Z., & Ma, X. (2016). Affective interaction recognition using spatio-temporal features and context. Computer Vision and Image Understanding, 144, 155-165.

[13] Pentland, A., & Liu, A. (1999). Modeling and prediction of human behavior. Neural computation, 11(1), 229-242.

[14] Xu, Y., Zhang, J., Zhang, Q., & Tao, D. (2022). ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. arXiv preprint arXiv:2204.12484.

[15] KTH, Schuldt, C., Laptev, I., and Caputo, B. (2004). “Recognizing human actions: a local SVM approach,” in Proc. International Conference on Pattern Recognition,(Cambridge), 32–36

[16] Weizman, Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). “Actions as space-time shapes,” in Proc. IEEE International Conference on Computer Vision, (Beijing), 1395–1402.

[17] UCF Sports, Rodriguez, M. D., Ahmed, J., and Shah, M. (2008). “Action MACH: a spatio-temporal maximum average correlation height filter for action recognition,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Anchorage, AK), 1–8.

[18] MuHAVi, Singh, S., Velastin, S. A., and Ragheb, H. (2010). “Muhavi: a multicamera human action video dataset for the evaluation of action recognition methods,” in Proc.IEEE International Conference on Advanced Video and Signal Based Surveillance, (Boston, MA), 48–55.

[19] UCF50, Reddy, K. K., and Shah, M. (2013). Recognizing 50 human action categories of web videos. Mach. Vis. Appl. 24, 971–981. doi:10.1007/s00138-012-0450-4

[20] UCF101, Soomro, K., Zamir, A. R., and Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. Cornell University Library. CoRR, abs/1212.0402.

[21] UCF YouTube, Liu, J., Luo, J., and Shah, M. (2009). "Recognizing realistic actions from videos in the wild," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Miami Beach, FL), 1–8.

[22] Hollywood2, Marszałek, M., Laptev, I., and Schmid, C. (2009). "Actions in context," in Proc.IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (Miami Beach, FL), 2929–2936.

[23] HMDB51, Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). "HMDB: a large video database for human motion recognition," in Proc. IEEE International, Conference on Computer Vision (Barcelona), 2556–2563.

[24] TVHI, Patron-Perez, A., Marszalek, M., Reid, I., and Zisserman, A. (2012). Structured learning of human interactions in TV shows. IEEE Trans. Pattern Anal. Mach.Intell. 34, 2441–2453. doi:10.1109/TPAMI.2012.24

[25] PETS 2004, Fisher, R. B. (2004). PETS04 Surveillance Ground Truth Dataset. Available at: http://www-prima.inrialpes.fr/PETS04/.

[26] PETS 2007, Fisher, R. B. (2007b). PETS07 Benchmark Dataset. Available at: http://www.cvg.reading.ac.uk/PETS2007/data.html

[27] VIRAT, Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., et al. (2011). "Real-time human pose recognition in parts from single depth images," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern, Recognition (Colorado Springs, CO), 1297–1304.

[28] TUM Kitchen, Tenorth, M., Bandouch, J., and Beetz, M. (2009). "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition," in Proc. IEEE International Workshop on Tracking Humans for the Evaluation of Their Motion in Image Sequences (THEMIS) (Kyoto), 1089–1096.

[29] Tow-person interaction, Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L., and Samaras, D. (2012). “Twoperson interaction detection using body-pose features and multiple instance learning,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (Providence, RI), 28–35.

[30] MSRC-12 Kinect gesture, Fothergill, S., Mentis, H. M., Kohli, P., and Nowozin, S. (2012). “Instructing people for training gestural interactive systems,” in Proc. C.

[31] J-HMDB, Jhuang, H., Gall, J., Zuffi, S., Schmid, C., and Black, M. J. (2013). “Towards understanding action recognition,” in Proc. IEEE International Conference on Computer Vision (Sydney, NSW), 3192–3199.

[32] UPCV Action, Theodorakopoulos, I., Kastaniotis, D., Economou, G., and Fotopoulos, S. (2014).Pose-based human action recognition via sparse representation in dissimilarity space. J. Vis. Commun. Image Represent. 25, 12–23. doi:10.1016/j.jvcir.2013.03.008.

[33] URADL, Messing, R., Pal, C. J., and Kautz, H. A. (2009). “Activity recognition using the velocity histories of tracked keypoints,” in Proc. IEEE International Conference on Computer Vision (Kyoto), 104–111.

[34] ADL, Pirsiavash, H., and Ramanan, D. (2012). “Detecting activities of daily living in firstperson camera views,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 2847–2854.

[35] MPII, Rohrbach, M., Amin, S., Mykhaylo, A., and Schiele, B. (2012). “A database for fine grained activity detection of cooking activities,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1194–1201.

[36] Breakfast, Kuehne, H., Arslan, A., and Serre, T. (2014). “The language of actions: recovering the syntax and semantics of goal-directed human activities,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 780–787.

[37] CCV, Jiang, Y. G., Ye, G., Chang, S. F., Ellis, D. P. W., and Loui, A. C. (2011). “Consumer video understanding: a benchmark database and an evaluation of human and machine performance,” in Proc. International Conference on Multimedia Retrieval (Trento), 29–36.

[38] FPSI, Fathi, A., Hodgins, J. K., and Rehg, J. M. (2012). “Social interactions: a first-person perspective,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1226–1233.

[39] Broadcast field hockey, Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., and Mori, G. (2012b). Discriminative latent models for recognizing contextual group activities. IEEE Trans.Pattern Anal. Mach. Intell. 34, 1549–1562. doi:10.1109/TPAMI.2011.228

[40] USAA, Fu, Y., Hospedales, T. M., Xiang, T., and Gong, S. (2012). “Attribute learning for understanding unstructured social activity,” in Proc. European Conference on Computer Vision, Lecture Notes in Computer Science, Vol. 7575 (Florence),530–543.

[41] Sport-1M, Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L.(2014). “Large-scale video classification with convolutional neural networks,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 1725–1732.

[42] ActivityNet, Heilbron, F. C., Escorcia, V., Ghanem, B., and Niebles, J. C. (2015). “ActivityNet: a large-scale video benchmark for human activity understanding,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 961–970.

[43] WWW Crowd, Shao, J., Kang, K., Loy, C. C., and Wang, X. (2015). “Deeply learned attributes for crowded scene understanding,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 4657–4666.

[44] BEHAVE, Fisher, R. B. (2007a). Behave: Computer-Assisted Prescreening of Video Streams for Unusual Activities. Available http://homepages.inf.ed.ac.uk/rbf/BEHAVE/.

[45] Canal9, Vinciarelli, A., Dielmann, A., Favre, S., and Salamin, H. (2009). “Canal9: a database of political debates for analysis of social interactions,” in Proc. International Conference on Affective Computing and Intelligent Interaction and Workshops (Amsterdam: De Rode Hoed), 1–4.

[46] USC Createive IT, Metallinou, A., Lee, C. C., Busso, C., Carnicke, S. M., and Narayanan, S. (2010). “The USC creative IT database: a multimodal database of theatrical improvisation,” in Proc. Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (Malta: Springer), 1–4.

[47] Parliament, Vrigkas, M., Nikou, C., and Kakadiaris, I. A. (2014b). “Classifying behavioral attributes using conditional random fields,” in Proc. 8th Hellenic Conference on Artificial Intelligence, Lecture Notes in Computer Science, Vol. 8445 (Ioannina), 95–104.