Định danh và nhận dạng vận động trong thời gian thực sử dụng mô hình học sâu

[1]ThS. Lê Trung Hiếu*
[1]Khoa Công nghệ Thông tin, Trường Đại học Đại Nam
Email: hieult@dainam.edu.vn

[2]Nguyễn Hữu Huy
[2]Khoa Công nghệ Thông tin, Trường Đại học Đại Nam

[3]Nguyễn Thanh Bình
[3]Khoa Công nghệ Thông tin, Trường Đại học Đại Nam

[4]Phạm Đình Nghĩa
[4]Khoa Công nghệ Thông tin, Trường Đại học Đại Nam

[5]Nguyễn Văn Nhân
[5]Khoa Công nghệ Thông tin, Trường Đại học Đại Nam

Ngày nhận bài: 15/09/2025
Ngày chấp nhận đăng: 29/09/2025

Tóm tắt **- Bài báo này giới thiệu một hệ thống tích hợp trí tuệ nhân tạo cho bài toán định danh và nhận dạng vận động trong thời gian thực, hướng đến tự động hóa quá trình đánh giá thể lực. Hệ thống thực hiện hai chức năng chính: (1) xác thực danh tính người dùng dựa trên đặc trưng khuôn mặt, và (2) phân tích chuỗi chuyển động cơ thể để nhận dạng cũng như đếm số lần thực hiện các bài tập phổ biến như chống đẩy và gập bụng. Trong nghiên cứu này, đặc trưng khuôn mặt được trích xuất bằng phương pháp học sâu trên nền tảng CNN, trong khi dữ liệu vận động được biểu diễn qua các điểm khớp cơ thể thu được từ video và xử lý bằng mô hình LSTM nhằm phân loại trạng thái tư thế theo thời gian. Kết quả thực nghiệm cho thấy hệ thống đạt độ chính xác cao, hoạt động ổn định và có khả năng triển khai trong các môi trường giáo dục, thể thao và giám sát thể chất quy mô lớn. Nghiên cứu góp phần mở rộng ứng dụng của học sâu trong lĩnh vực nhận diện hành vi người và đánh giá thể chất tự động.**

Từ khóa - Định danh khuôn mặt, nhận dạng vận động, thị giác máy tính, học sâu, Mediapipe, LSTM, gập bụng, chống đẩy.

I. GIỚI THIỆU

Trong những năm gần đây, nhu cầu ứng dụng các công nghệ trí tuệ nhân tạo (AI) và học sâu (deep learning) trong lĩnh vực giám sát, đào tạo và đánh giá thể chất ngày càng gia tăng. Việc định danh và theo dõi vận động viên, học sinh hoặc người tập luyện thể thao đóng vai trò quan trọng trong nhiều bối cảnh, từ giáo dục thể chất, huấn luyện chuyên nghiệp, đến các hệ thống giám sát an ninh và chăm sóc sức khỏe thông minh. Tuy nhiên, các phương pháp truyền thống dựa vào quan sát thủ công của giám khảo hoặc thiết bị đo lường đơn lẻ thường gặp phải những hạn chế đáng kể, bao gồm tính chủ quan, sai lệch do yếu tố con người, thiếu tính khách quan và không đảm bảo độ ổn định khi số lượng đối tượng cần đánh giá lớn [1], [2]. Điều này đặt ra thách thức trong việc tìm kiếm các giải pháp khoa học và công nghệ có khả năng tự động hóa, nâng cao độ tin cậy và mở rộng quy mô.

Cùng với sự phát triển mạnh mẽ của thị giác máy tính (computer vision) và khả năng xử lý chuỗi dữ liệu của các mô hình học sâu, đặc biệt là mạng nơ ron tích chập (CNN) và mạng nơ ron hồi tiếp (RNN), việc nghiên cứu các hệ thống vừa có khả năng định danh (identification) vừa có thể nhận dạng vận động (action recognition) trong thời gian thực đã trở thành một hướng tiếp cận tiềm năng. Vấn đề này có ý nghĩa quan trọng không chỉ trong giáo dục và thể thao, mà còn trong nhiều lĩnh vực liên quan như y sinh học, phục hồi chức năng, tương tác người -máy, và các hệ thống an ninh thông minh.

Trong bối cảnh đó, nghiên cứu này đề xuất một hệ thống tích hợp dựa trên học sâu, với hai chức năng chính: (1) định danh cá nhân thông qua đặc trưng khuôn mặt nhằm đảm bảo tính minh bạch, cá nhân hóa và an toàn dữ liệu, và (2) nhận dạng và đếm số lần thực hiện các động tác vận động phổ biến như chống đẩy (push-up) và gập bụng (sit-up) trong thời gian thực. Cách tiếp cận này vừa giải quyết được bài toán xác thực danh tính, vừa cung cấp công cụ đánh giá thể chất khách quan và đáng tin cậy, có thể ứng dụng trực tiếp trong môi trường học đường, phòng tập gym hoặc hệ thống thi đấu chuyên nghiệp.

Đối với bài toán định danh, hệ thống khai thác sức mạnh của học sâu trong trích xuất đặc trưng khuôn mặt, dựa trên CNN và kỹ thuật nhúng vector (embedding). Phương pháp này cho phép ánh xạ khuôn mặt thành không gian đặc trưng có tính phân biệt cao, từ đó đối chiếu chính xác với cơ sở dữ liệu danh tính đã được gán nhãn trước đó [3]. Quá trình này đảm bảo rằng chỉ những người dùng đã được xác thực mới có thể tham gia vào hệ thống, đồng thời tăng cường tính an toàn, minh bạch và khả năng truy vết.

Song song, mô - đun nhận dạng vận động được triển khai dựa trên sự kết hợp giữa công cụ trích xuất khung xương người từ video (pose estimation) và mô hình LSTM (Long Short - Term Memory) để xử lý chuỗi chuyển động [4], [5]. Trong đó, khung xương cơ thể được biểu diễn bằng các keypoints theo thời gian, phản ánh tư thế và sự thay đổi trạng thái vận động của người tập. LSTM, với khả năng lưu giữ và khai thác thông tin dài hạn, cho phép phân tích chính xác các chuỗi hành vi, từ đó phân loại động tác và tính toán số lần lặp lại một cách tự động. Sự kết hợp này mang lại độ chính xác cao hơn so với các phương pháp truyền thống dựa trên ngưỡng hoặc quy tắc hình học đơn giản.

Một điểm nhấn quan trọng của nghiên cứu là việc tích hợp hai thành phần định danh khuôn mặt và nhận dạng vận động vào trong cùng một hệ thống thời gian thực. Nhờ

đó, hệ thống không chỉ dừng lại ở việc giám sát hành vi vận động, mà còn gắn liền dữ liệu kết quả với đúng đối tượng người dùng, đảm bảo tính toàn vẹn dữ liệu và hỗ trợ phân tích dài hạn. Đây là yếu tố có giá trị đặc biệt trong các nghiên cứu giáo dục thể chất, quản lý huấn luyện viên - vận động viên, cũng như trong xây dựng các nền tảng hỗ trợ ra quyết định dựa trên dữ liệu.

Những đóng góp khoa học chính của bài báo có thể được tóm lược như sau:

- Đề xuất một hệ thống học sâu tích hợp giữa định danh cá nhân bằng khuôn mặt và nhận dạng vận động trong thời gian thực, hướng đến giải pháp toàn diện cho đánh giá thể chất.
- Áp dụng kết hợp kỹ thuật trích xuất khung xương cơ thể từ video và mô hình LSTM để xử lý chuỗi hành vi vận động, giúp nâng cao độ chính xác và tính ổn định trong điều kiện thực tế.
- Xây dựng bộ dữ liệu thử nghiệm có gán nhãn thủ công và quy trình huấn luyện - đánh giá mô hình từ đầu đến cuối, đóng góp cho cộng đồng nghiên cứu trong lĩnh vực action recognition.

Cấu trúc của bài báo được tổ chức như sau: Mục II Các nghiên cứu và công trình liên qua; Mục III mô tả chi tiết kiến trúc và thành phần hệ thống; Mục IV quá trình thu thập dữ liệu; Mục V Thực nghiệp và đánh giá; cuối cùng, Mục VI đưa ra kết luận

II. CÁC NGHIÊN CỨU VÀ CÔNG TRÌNH LIÊN QUAN

Hệ thống nhận diện khuôn mặt và đếm động tác thể chất là sự kết hợp giữa hai lĩnh vực chính của thị giác máy tính: xác thực sinh trắc học và phân tích chuyển động thời gian thực. Phần này tổng hợp các nghiên cứu và công trình liên quan đã đóng góp nền tảng lý thuyết và công nghệ cho hệ thống.

A. Xác thực khuôn mặt trong thị giác máy tính

Nhận diện khuôn mặt là một trong những bài toán nổi bật trong thị giác máy tính và đã có nhiều hướng tiếp cận được phát triển. Phương pháp kinh điển như Viola-Jones [9] sử dụng đặc trưng Haar và bộ phân loại Cascade cho khả năng phát hiện thời gian thực, nhưng hiệu quả kém trong điều kiện ánh sáng hoặc góc chụp phức tạp. Hướng tiếp cận hiện đại hơn sử dụng Histogram of Oriented Gradients (HOG) kết hợp với SVM [10], được King hiện thực hóa trong thư viện Dlib [8], tuy nhiên vẫn chịu ảnh hưởng lớn từ yếu tố môi trường.

Trong những năm gần đây, các mạng nơ ron tích chập (CNN) đã được ứng dụng mạnh mẽ để trích xuất đặc trưng khuôn mặt và ánh xạ vào không gian vector. Điển hình là FaceNet [11], sử dụng Triplet Loss để tối ưu khoảng cách giữa các khuôn mặt cùng người và khác người. Ngoài ra, các mô hình như ArcFace [12] và DeepFace [13] đã cải tiến thêm các hàm mất mát, giúp hệ thống phân biệt tốt hơn các khuôn mặt trong môi trường thực tế.

Các thư viện mã nguồn mở như OpenCV [15], dlib [8] và face_recognition [14] được sử dụng rộng rãi trong hệ thống thực thi do khả năng triển khai nhanh và độ chính xác tương đối cao. Một số nghiên cứu mới cũng đã thử nghiệm kết hợp Transformer và cơ chế Attention trong nhận diện khuôn mặt để tăng độ chính xác trong môi trường phức tạp [16].

B. Định danh và nhận dạng các hoạt động thể thao

Nhận diện động tác người dùng thông qua webcam hoặc camera IP là một lĩnh vực ứng dụng phổ biến trong thể thao, chăm sóc sức khỏe và an ninh [17]. Các phương pháp truyền thống như HMM [18] hoặc DTW [19] từng được dùng để nhận diện chuỗi động tác, nhưng khả năng xử lý thời gian thực và độ chính xác chưa đáp ứng tốt yêu cầu thực tế.

Với sự phát triển của học sâu, các mô hình Recurrent Neural Network (RNN), đặc biệt là LSTM [5], đã được áp dụng rộng rãi để phân tích chuỗi dữ liệu keypoints. CNN thường được dùng để trích xuất đặc trưng không gian từ hình ảnh đầu vào, sau đó kết hợp với LSTM để học đặc trưng thời gian [20]. Ngoài ra, Bi-LSTM [21] và GRU cũng được ứng dụng trong nhiều nghiên cứu đếm động tác như squat, push-up với độ chính xác trên 95

Một số công trình sử dụng mạng GNN [22] hoặc mô hình dựa trên Transformer [23] để học mối quan hệ không gian-thời gian từ dữ liệu keypoints. Trong khi đó, công cụ Mediapipe của Google [4] và OpenPose [24] đóng vai trò như front-end xử lý dữ liệu chuyển động, giúp trích xuất khung xương người dùng với tốc độ cao và tài nguyên thấp.

C. Tích hợp hệ thống

Hệ thống nhận diện động tác thường hoạt động theo pipeline gồm: thu thập video từ webcam, trích xuất keypoints qua Mediapipe/OpenPose, phân tích chuỗi chuyển động bằng mô hình học sâu (CNN-LSTM) và cập nhật kết quả vào cơ sở dữ liệu. Ví dụ, một hệ thống đếm chống đẩy sử dụng Mediapipe

+ LSTM đã đạt độ chính xác trên 95%. [25].

Dữ liệu kết quả bài tập được lưu trữ vào cơ sở dữ liệu quan hệ như MySQL [7], đảm bảo tính toàn vẹn và truy xuất hiệu quả. Các công trình khác cũng triển khai API bằng Flask [6] hoặc FastAPI để kết nối giữa mô hình AI và giao diện người dùng. Nhằm đảm bảo khả năng cập nhật thời gian thực, giao tiếp WebSocket [26] thường được tích hợp để giảm độ trễ và nâng cao trải nghiệm tương tác.

Ở tầng giao diện, các framework như Bootstrap, ReactJS, hoặc VueJS giúp hiển thị trực quan tiến trình bài tập và thông báo đến người dùng. Các nghiên cứu như [2] cho thấy việc tối ưu truyền dữ liệu và phản hồi bằng WebSocket giúp giảm thời gian phản hồi hệ thống, đặc biệt trong các ứng dụng real-time như thể thao hoặc y tế.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Hệ thống đề xuất bao gồm hai giai đoạn chính: *(i) giai đoạn huấn luyện* và *(ii) giai đoạn triển khai*.

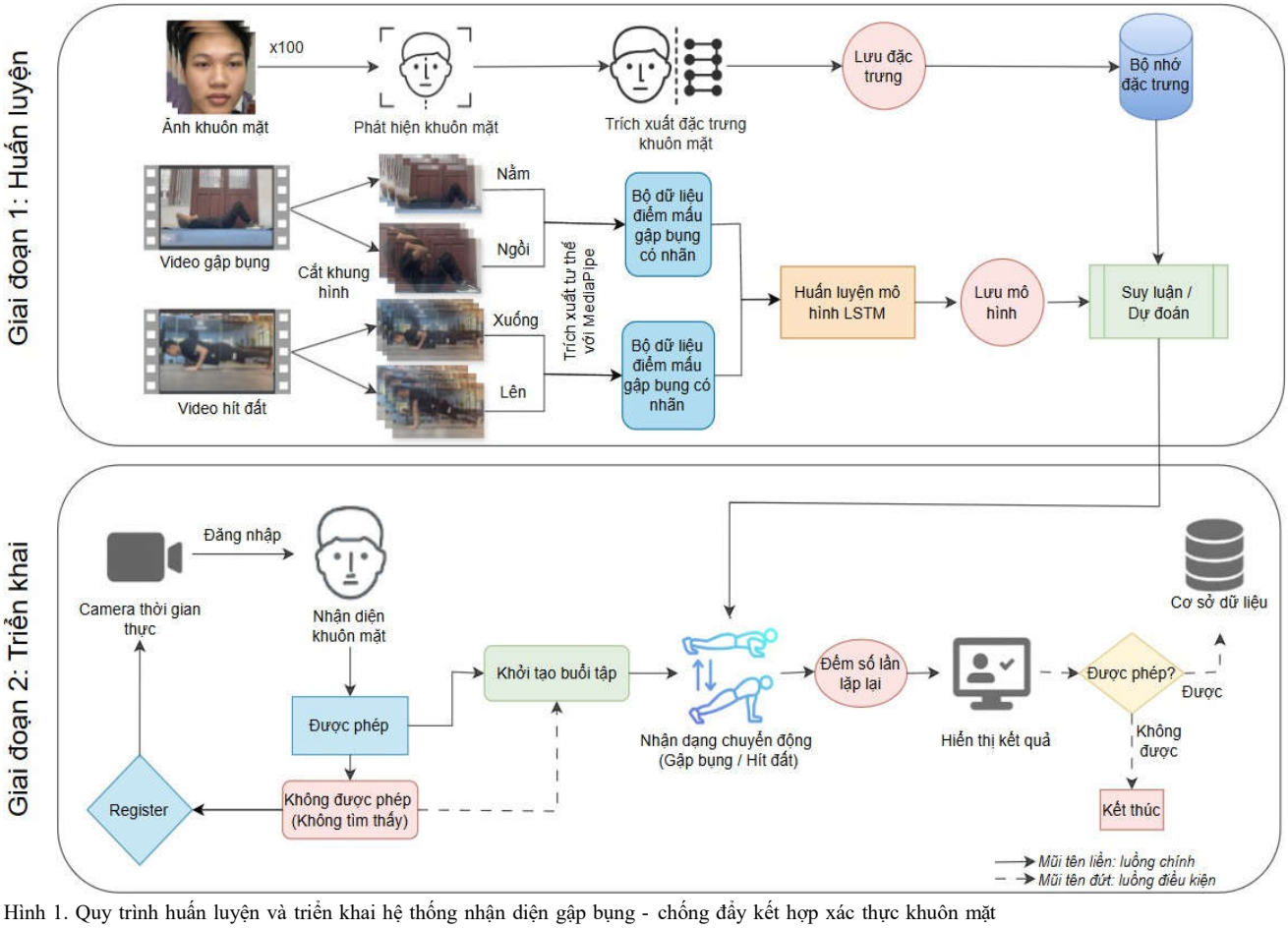
Trong giai đoạn huấn luyện, dữ liệu khuôn mặt và dữ liệu động tác được thu thập, tiền xử lý và đưa vào các mô hình học sâu để xây dựng *bộ nhớ đặc trưng khuôn mặt* và mô hình phân loại động tác dựa trên LSTM.

Trong giai đoạn triển khai, hệ thống tương tác trực tiếp với người dùng thông qua camera. Đầu tiên, người dùng

được xác thực bằng nhận diện khuôn mặt. Sau đó, mô hình LSTM xử lý chuỗi dữ liệu keypoints thu từ video thời gian thực để nhận diện và đếm số lần động tác. Thông tin luyện tập chỉ được lưu vào cơ sở dữ liệu khi danh tính người dùng đã được xác thực. Hai giai đoạn này được tích hợp thành một quy trình khép kín, như minh họa ở Hình 1, nhằm đảm bảo hệ thống vừa cá nhân hóa vừa có độ chính xác cao.

A. Giai Đoạn 1: Huấn Luyện

Trong giai đoạn huấn luyện, hệ thống được thiết kế và xây dựng theo hai luồng xử lý song song: *nhận diện khuôn mặt* và *nhận diện hoạt động thể chất*.



Hình 1. Quy trình huấn luyện và triển khai hệ thống nhận diện gập bụng - chống đẩy kết hợp xác thực khuôn mặt

Nhận diện khuôn mặt: Để đảm bảo tính cá nhân hóa cho từng người dùng, mỗi cá nhân cung cấp khoảng 100 ảnh khuôn mặt ở các góc chụp và điều kiện ánh sáng khác nhau. Trước tiên, hệ thống áp dụng thuật toán phát hiện khuôn mặt (ví dụ MTCNN hoặc Haar Cascade) để tách vùng mặt ra khỏi nền ảnh. Sau đó, các đặc trưng hình học và hình thái học của khuôn mặt được trích xuất bằng mô hình học sâu chuyên biệt cho nhận dạng (ví dụ CNN hoặc FaceNet). Kết quả của bước này là một vector đặc trưng có khả năng phân biệt rõ ràng giữa các cá nhân. Các vector đặc trưng được lưu trữ trong bộ nhớ đặc trưng, phục vụ cho giai đoạn suy luận và nhận diện trong thời gian thực.

Nhận diện hoạt động thể chất: Đối với bài toán nhận dạng động tác, hệ thống thu thập dữ liệu dưới dạng video gập bụng và hít đất. Các video này được tách thành các khung hình riêng lẻ, sau đó tiến hành gán nhãn trạng thái cụ thể của từng khung hình (ví dụ: gập bụng - nằm/ngồi; hít đất - lên/xuống). Tiếp theo, hệ thống sử dụng công cụ MediaPipe để trích xuất bộ điểm mấu (*keypoints*) trên cơ thể, bao gồm các khớp quan trọng như vai, hông, gối và khuỷu tay. Chuỗi dữ liệu keypoints thu được phản ánh rõ ràng chuyển động qua thời gian.

Mô hình LSTM (Long Short-Term Memory) sau đó được huấn luyện trên chuỗi dữ liệu keypoints này để học cách phân biệt và dự đoán loại động tác. Khác với các mô hình phân loại tĩnh, LSTM tận dụng khả năng ghi nhớ phụ thuộc theo chuỗi, do đó có thể mô hình hóa chính xác đặc điểm động học của các bài tập thể chất. Sau khi huấn luyện, mô hình được lưu lại và sẵn sàng triển khai trong môi trường thực tế.

B. Giai Đoạn 2: Triển Khai

Trong giai đoạn triển khai, hệ thống hoạt động trực tiếp với camera thời gian thực và tương tác cùng người dùng.

Đầu tiên, người dùng cần thực hiện đăng nhập bằng khuôn mặt. Camera thu nhận hình ảnh và hệ thống trích xuất vector đặc trưng, sau đó so khớp với bộ nhớ đặc trưng. Nếu tìm thấy sự tương đồng với dữ liệu đã đăng ký, người dùng được cấp quyền truy cập và bắt đầu buổi tập. Trường hợp không có kết quả khớp, người dùng cần tiến hành đăng ký mới để bổ sung dữ liệu khuôn mặt vào hệ thống.

Sau khi xác thực thành công, hệ thống khởi tạo buổi tập luyện. Camera tiếp tục ghi nhận chuyển động cơ thể và dữ liệu keypoints được đưa vào mô hình LSTM để phân loại động tác (gập bụng hoặc hít đất). Mô hình đồng thời đếm số lần lặp lại động tác và hiển thị trực tiếp kết quả trên màn hình giao diện người dùng. Thông tin kết quả luyện tập, bao gồm số lần thực hiện và loại bài tập, được lưu trữ vào cơ sở dữ liệu nếu người dùng đã được xác thực danh tính.

Ngược lại, nếu danh tính không được xác thực (ví dụ khuôn mặt không khớp), hệ thống vẫn có thể đếm số lần động tác nhằm hỗ trợ việc tập luyện, nhưng kết quả sẽ không được ghi nhận vào cơ sở dữ liệu để đảm bảo tính riêng tư và chính xác trong theo dõi tiến trình luyện tập.

C. Tích Hợp Hai Giai Đoạn

Phương pháp đề xuất kết hợp chặt chẽ hai giai đoạn huấn luyện và triển khai để đảm bảo tính toàn vẹn, chính xác và tin cậy của hệ thống. Trong đó, nhận diện khuôn mặt đóng vai trò bảo mật và gắn kết dữ liệu luyện tập với đúng cá nhân, trong khi mô hình LSTM xử lý và đánh giá các hoạt động thể chất một cách chính xác.

Cách tiếp cận này cho phép hệ thống vừa hoạt động như một công cụ theo dõi luyện tập thông minh, vừa đóng vai trò như một nền tảng quản lý dữ liệu cá nhân hóa. Nhờ đó, người dùng có thể theo dõi tiến trình luyện tập của bản thân qua thời gian, đồng thời hệ thống đảm bảo an toàn và độ tin cậy cao trong lưu trữ dữ liệu.

IV. THU THẬP DỮ LIỆU

Dữ liệu gồm hai nhóm: **khuôn mặt** (xác thực danh tính) và **hoạt động vận động** (gập bụng, chống đẩy). Việc thu thập được thực hiện trong môi trường có kiểm soát, với sự tham gia của 7 người (3 thành viên nhóm và 4 tình nguyện viên), tất cả đều ký cam kết đồng thuận và được đảm bảo quyền riêng tư.

A. Dữ Liệu Khuôn Mặt

Ảnh khuôn mặt được ghi bằng camera HD trong điều kiện ánh sáng đồng đều. Mỗi người cung cấp khoảng 100 ảnh chính diện, loại bỏ ảnh mờ hoặc lệch góc. Bộ dữ liệu sau chuẩn hóa được dùng để huấn luyện và đánh giá mô hình xác thực khuôn mặt dựa trên vector embedding 128 chiều.

Hình 2. Một số ảnh mẫu trong tập dữ liệu khuôn mặt

B. Dữ Liệu Hoạt Động Vận Động

Nghiên cứu tập trung vào hai bài tập phổ biến: **gập bụng** và **chống đẩy**. Video được ghi lại ở nhiều góc quay (0°, 45°, 90°) để tăng tính đa dạng hình thái và khả năng tổng quát của mô hình. Từ video, frame được trích xuất và gán nhãn thủ công theo *hai trạng thái nhị phân*:

- Gập bụng: *ngồi* / *nằm*.
- Chống đẩy: *xuống* / *lên*.

Để giảm sai lệch gán nhãn, dữ liệu được hai người annotator độc lập đánh dấu, sau đó thống nhất theo nguyên tắc majority voting.

Hình 3. Ví dụ frame gán nhãn trong dữ liệu chống đẩy [25].

C. Thống Kê Dữ Liệu

Dữ liệu thu thập gồm ảnh khuôn mặt và video gập bụng, chống đẩy, đã được gán nhãn trong (Bảng I).

BẢNG 1. THỐNG KÊ DỮ LIỆU THU THẬP ĐƯỢC

Loại dữ liệu	Số lượng	Ghi chú
Khuôn mặt (ảnh)	400 ảnh	4 người × 100 ảnh
Video gập bụng	30 video	~ 30,000 frame trích xuất
Video chống đẩy	27 video	~ 10,000 frame trích xuất
Tổng frame vận động	40,000+	Đã gán nhãn thủ công

D. Tiền Xử Lý Dữ Liệu

Trong giai đoạn tiền xử lý, dữ liệu khuôn mặt và dữ liệu vận động được xử lý theo các bước khác nhau để đảm bảo tính đồng nhất và phù hợp với mô hình học sâu. Đối với dữ

liệu khuôn mặt, các ảnh được căn chỉnh dựa trên vị trí mắt và mũi, sau đó chuẩn hóa về kích thước cố định 128 × 128 pixel nhằm giảm nhiễu và duy trì đặc trưng ổn định.

Với dữ liệu vận động, video được trích xuất khung hình ở tốc độ 15 - 30 fps, sau đó sử dụng MediaPipe để nhận diện và trích xuất 33 *keypoints* của khớp cơ thể, hình thành chuỗi dữ liệu thời gian phục vụ cho các mô hình tuần tự như LSTM.

Ngoài ra, để tăng tính đa dạng của tập dữ liệu và hạn chế hiện tượng overfitting, kỹ thuật tăng cường dữ liệu (*data augmentation*) được áp dụng. Các phương pháp chính bao gồm xoay ảnh một góc nhỏ (khoảng $\pm 10^{\circ}$), điều chỉnh độ sáng để mô phỏng các điều kiện ánh sáng khác nhau, và lật ngang ảnh

nhằm tăng khả năng tổng quát hóa của mô hình.

V. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Phần này trình bày cấu hình thực nghiệm, tiêu chí đánh giá, hiệu năng và kết quả thí nghiệm của hệ thống nhận diện, bao gồm hai nhiệm vụ: nhận diện gập bụng và nhận diện chống đẩy.

***A.** Thiết Lập Thực Nghiệm và Tiêu Chí Đánh Giá*

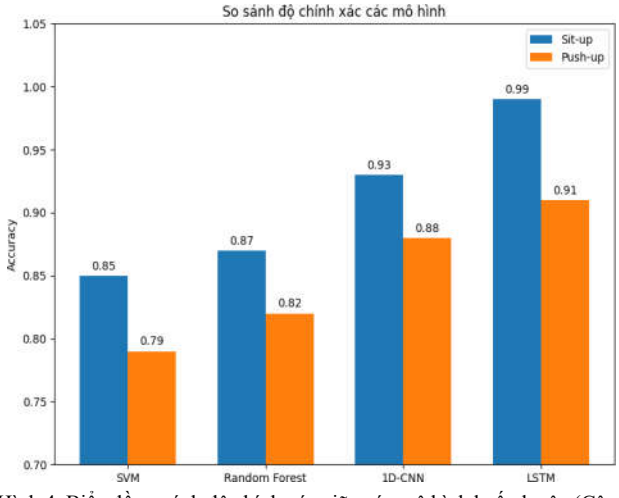
Dữ liệu: Bộ dữ liệu gồm ~ 30,000 mẫu gập bụng và 10,000 mẫu chống đẩy, được trích xuất từ video và gán nhãn theo hai trạng thái (gập bụng: nằm/ngồi; chống

đẩy: lên/xuống). Dữ liệu chuẩn hóa về [0,1] và chia 80% huấn luyện, 20% kiểm tra.

Tiêu chí đánh giá: Các chỉ số sử dụng gồm Accuracy, Pre- cision, Recall, F1-score, kèm ma trận nhầm lẫn và AUC-ROC để đánh giá khả năng phân biệt của mô hình.

B. Kết Quả Thực Nghiệm

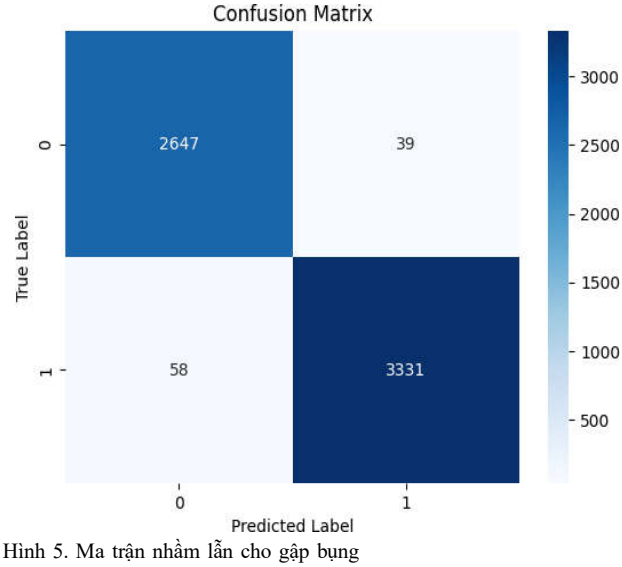
1. So Sánh Mô Hình Cho Nhận Diện Động Tác: Các mô hình thử nghiệm: SVM, Random Forest, 1D-CNN, LSTM. Hình 4 cho thấy LSTM đạt độ chính xác cao nhất ở cả hai bài tập, vượt trội so với các mô hình truyền thống.



Hình 4. Biểu đồ so sánh độ chính xác giữa các mô hình huấn luyện (Gập bụng và Chống đẩy)

C. Đánh Giá Hiệu Năng Mô Hình

1. Nhận Diện Gập Bụng: Mô hình LSTM phân loại hai trạng thái *gập* và *không gập* từ chuỗi keypoints, đạt Accuracy ≈ 0.99 và F1-score ≈ 0.98 trên 6,075 mẫu kiểm thử. Sai sót chủ yếu xảy ra ở pha chuyển đổi giữa hai trạng thái.

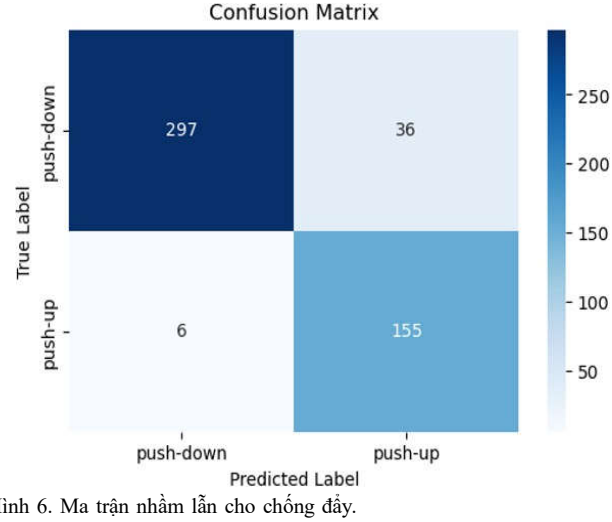


Hình 5. Ma trận nhầm lẫn cho gập bụng

BẢNG 2. KẾT QUẢ THỰC NGHIỆM PHÂN LOẠI CHO GẬP BỤNG

Trạng thái	Precision	Recall	F1-score	Accuracy	Số mẫu
Nằm	0.95	0.96	0.96	0.99	2647
Ngồi	0.99	0.98	0.98	0.99	3331
Trung bình	**0.97**	**0.97**	**0.97**	**0.99**	**5978**

2. Nhận Diện Chống Đẩy: Mô hình LSTM phân loại hai trạng thái *push-down* và *push-up*, đạt Accuracy ≈ 0.91 trên 494 mẫu kiểm thử, với F1-score lần lượt 0.93 và 0.88. Nhầm lẫn chủ yếu ở các pha chuyển tiếp nhanh.



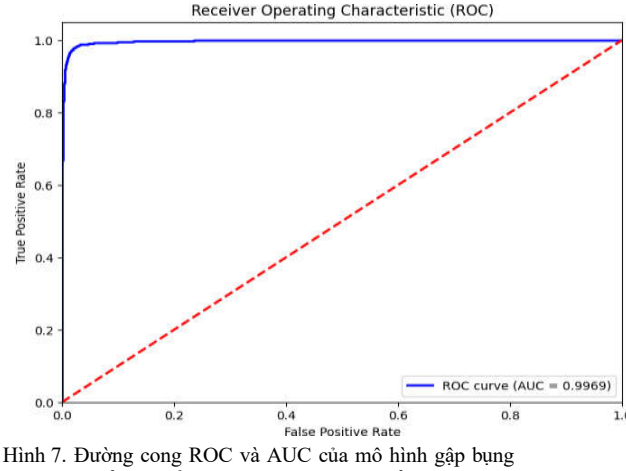
Hình 6. Ma trận nhầm lẫn cho chống đẩy.

BẢNG 3. KẾT QUẢ THỰC NGHIỆM PHÂN LOẠI CHO CHỐNG ĐẨY

Trạng thái	Precision	Recall	F1-score	Accuracy	Số mẫu
Xuống	0.98	0.89	0.93	0.91	297
Lên	0.81	0.96	0.88	0.91	155
Trung bình	**0.90**	**0.93**	**0.91**	**0.91**	**452**

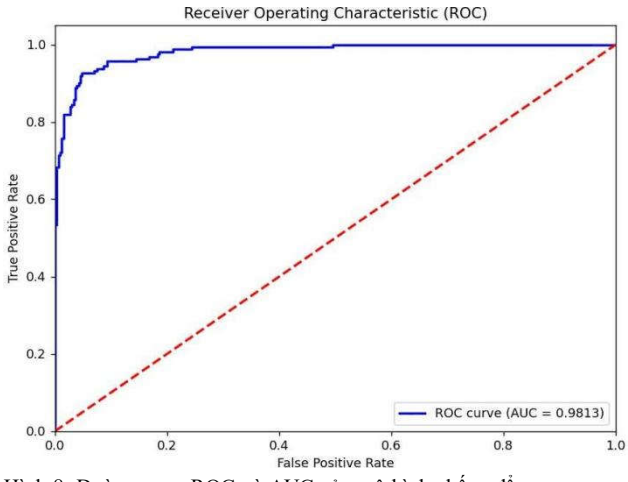
D. Đánh Giá Hiệu Năng Mô Hình

1. Gập Bụng: Mô hình hội tụ tốt, đạt AUC ≈ 0.997, cho thấy khả năng phân loại gần như hoàn hảo giữa 2 trạng thái (Nằm và ngồi) (Hình 7).



Hình 7. Đường cong ROC và AUC của mô hình gập bụng

2. Chống Đẩy: Mô hình hội tụ tốt, đạt AUC ≈ 0.981, cho thấy khả năng phân loại khá chính xác giữa hai trạng thái (Lên và xuống) (Hình 8).



Hình 8. Đường cong ROC và AUC của mô hình chống đẩy

E. Đánh Giá

Hệ thống đạt được kết quả khả quan nhưng vẫn tồn tại một số hạn chế. Tập dữ liệu còn nhỏ và thiếu đa dạng về người tham gia, bối cảnh và điều kiện ánh sáng, khiến khả năng khái quát chưa cao. Các yếu tố môi trường và chuyển động nhanh có thể làm giảm độ chính xác của keypoints, trong khi tình huống nhiều người xuất hiện đồng thời chưa được xử lý hiệu quả. Ngoài ra, nghiên cứu hiện mới giới hạn ở hai động tác cơ bản (gập bụng, chống đẩy), chưa kiểm chứng trên nhiều loại vận động khác.

Trong tương lai, việc mở rộng tập dữ liệu với nhiều người, điều kiện ánh sáng và góc quay đa dạng hơn là cần thiết để nâng cao tính tổng quát. Đồng thời, ứng dụng các mô hình không gian - thời gian hiện đại (Transformer, ST-GCN) và bổ sung dữ liệu từ cảm biến IMU có thể giúp cải thiện khả năng nhận dạng động học. Hướng phát triển tiếp theo còn bao gồm xây dựng cơ chế theo dõi đa người kèm ID, cũng như tối ưu tiền xử lý và augmentation để tăng tính ổn định và chống nhiễu cho hệ thống.

VI. KẾT LUẬN

Bài báo đã đề xuất và triển khai một hệ thống nhận diện vận động thể chất dựa trên sự kết hợp giữa xác thực khuôn mặt và phân tích chuỗi chuyển động từ keypoints cơ thể. Kết quả thực nghiệm cho thấy mô hình đạt độ chính xác cao trong các tác vụ đã khảo sát, với hiệu quả đặc biệt ở nhận diện gập bụng và chống đẩy, qua đó khẳng định tiềm năng ứng dụng của hướng tiếp cận này trong các bài toán theo dõi và đánh giá vận động.

Hệ thống đồng thời chỉ ra những hạn chế cần khắc phục, bao gồm sự đa dạng của dữ liệu, độ nhạy với điều kiện môi trường và khả năng mở rộng sang các động tác phức tạp hơn. Trong tương lai, việc mở rộng tập dữ liệu, áp dụng các mô hình học sâu không gian - thời gian tiên tiến, cũng như tích hợp thêm nguồn dữ liệu cảm biến sẽ là những hướng đi quan trọng nhằm nâng cao độ chính xác và khả năng ứng dụng trong môi trường thực tế.

MỤC LỤC

[1] R. M. Kaplan and D. P. Saccuzzo, *Psychological Testing: Principles, Applications, and Issues*, 9th ed., Cengage Learning, 2017.

[2] J. Sulla-Torres, B. Santos-Pamo, F. Cárdenas-Rodríguez, J. Angulo-Osorio, R. Gómez-Campos, and M. Cossio-Bolan˜os, "Multiplatform Computer Vision System to Support Physical Fitness Assessments in Schoolchildren," *Applied Sciences*, vol. 14, no. 16, art. 7140, 2024. doi:10.3390/app14167140.

[3] A. Geitgey, "face_recognition: Python library for face recognition," GitHub repository, https://github.com/ageitgey/face_recognition, accessed Jan. 2025.

[4] Google, "MediaPipe: Cross-platform, customizable ML solutions for live and streaming media," https://mediapipe.dev/, accessed Jan. 2025.

[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[6] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*, 2nd ed., O'Reilly Media, 2018.

[7] P. DuBois, *MySQL Cookbook*, 3rd ed., O'Reilly Media, 2014.

[8] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755-1758, 2009.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, 2001.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015.

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019.

[13] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. CVPR*, 2014.

[14] "face_recognition," Open source Python library by Adam Geitgey. [Online]. Available: https://github.com/ageitgey/face_recognition

[15] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[16] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[17] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976-990, 2010.

[18] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[19] M. Mu¨ller, *Dynamic Time Warping*, Springer, 2007.

[20] Y. Zhang et al., "Action recognition based on two-stream convolutional neural network and LSTM," *Multimedia Tools and Applications*, vol. 79,

[21] pp. 11383-11403, 2020.

[22] Q. Nguyen et al., "Human activity recognition using Bi-LSTM network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 811-823, 2021.

[23] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI*, 2018.

[24] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. ICML*, 2021.

[25] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi- person 2D pose estimation using part affinity fields," in *Proc. CVPR*, 2017.

[26] G. H. Samaan, A. R. Wadie, A. K. Attia, A. M. Asaad, A. E. Kamel, S.

[27] O. Slim, M. S. Abdallah, and Y.-I. Cho, "MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition," *Electronics*, vol. 11, no. 19, art. 3228, 2022. doi:10.3390/electronics11193228.

[28] I. Fette and A. Melnikov, "The WebSocket Protocol," RFC 6455, 2011. Available: https://www.rfc-editor.org/info/rfc6455.