Nhận Dạng Hành Động Con Người Qua Camera

Quàng Minh Anh, **Nguyễn Anh Tuấn**, **Phạm Đình Tuấn**, **Nguyễn Tuấn Dũng**, **Nguyễn Thị Mai Lan**
Nhóm 8, CNTT 16-05, Khoa Công Nghệ Thông Tin, Trường Đại Học Đại Nam, Việt Nam
ThS Lê Trung Hiếu, **ThS Nguyễn Văn Nhân**
Giảng viên hướng dẫn, Khoa Công Nghệ Thông Tin, Trường Đại học Đại Nam, Việt Nam

TÓM TẮT NỘI DUNG

Nhận dạng hành động con người qua camera là một đề tài nhằm xây dựng hệ thống có thể phân loại các hành động khác nhau từ video bằng mô hình LSTM.

Quá trình thực hiện bao gồm bốn bước chính. Đầu tiên, dữ liệu được thu thập bằng cách ghi lại video cho từng hành động, mỗi hành động có 100 mẫu, sau đó chuyển đổi video thành tập hợp các khung hình (frame) và gán nhãn tương ứng.

Tiếp theo, dữ liệu được xử lý bằng cách tiền xử lý video (cắt, resize, chuyển sang grayscale), trích xuất đặc trưng từ từng frame bằng OpenCV, Mediapipe hoặc CNN, và tạo chuỗi dữ liệu đầu vào phù hợp cho LSTM.

Sau đó, mô hình LSTM được xây dựng và huấn luyện bằng cách chia dữ liệu thành tập train/test, tối ưu tham số và đánh giá độ chính xác.

Cuối cùng, mô hình được triển khai để nhận diện hành động theo thời gian thực từ camera và kiểm thử với dữ liệu mới.

Đề tài sử dụng ngôn ngữ lập trình Python cùng các thư viện OpenCV, TensorFlow/Keras, Mediapipe để xử lý dữ liệu và huấn luyện mô hình.

***Từ khóa*—Nhận dạng hành động, LSTM, AI, IoT.**

I. GIỚI THIỆU

Nhận dạng hành động con người là một chủ đề quan trọng trong lĩnh vực thị giác máy tính và trí tuệ nhân tạo, với nhiều ứng dụng trong giám sát an ninh, chăm sóc sức khỏe, tương tác người-máy và môi trường thông minh. Mục tiêu của nghiên cứu này là xây dựng một hệ thống có khả năng phân loại các hành động của con người từ chuỗi video bằng mô hình Long Short-Term Memory (LSTM).

Quy trình thực hiện gồm bốn giai đoạn chính. Đầu tiên, dữ liệu được thu thập bằng cách ghi lại video của các hành động khác nhau, mỗi hành động có 100 mẫu. Sau đó, các video này được chuyển đổi thành các khung hình (frame) và gán nhãn tương ứng. Tiếp theo, dữ liệu được tiền xử lý, bao gồm cắt video, thay đổi kích thước, chuyển đổi sang ảnh xám và trích xuất đặc trưng bằng OpenCV, Mediapipe hoặc mạng nơ-ron tích chập (CNN). Quá trình này giúp tạo ra chuỗi dữ liệu đầu vào phù hợp cho mô hình LSTM.

Sau khi hoàn thành tiền xử lý, mô hình dựa trên LSTM được xây dựng và huấn luyện. Dữ liệu được chia thành tập huấn luyện và kiểm thử, đồng thời tối ưu các siêu tham số để đạt độ chính xác cao trong phân loại hành động. Cuối cùng, mô hình đã huấn luyện được triển khai để nhận diện hành động theo thời gian thực từ camera, giúp hệ thống có thể xử lý luồng video trực tiếp và phân loại các hành động của con người một cách linh hoạt.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Nhận dạng hành động con người là một lĩnh vực nghiên cứu quan trọng trong thị giác máy tính, với nhiều ứng dụng trong giám sát an ninh, tương tác người - máy, phân tích thể thao và chăm sóc sức khỏe. Các nghiên cứu trước đây đã phát triển nhiều bộ dữ liệu nhằm hỗ trợ huấn luyện và đánh giá các mô hình học sâu trong nhận dạng hành động. Trong phần này, chúng tôi trình bày một số bộ dữ liệu tiêu biểu đã được xây dựng.

Một số bộ dữ liệu đã được phát triển để nhận dạng hành động con người qua camera:

- **[J. Wang, 2021]**: Xây dựng một bộ dữ liệu quy mô lớn tập trung vào các hoạt động hàng ngày như đi bộ, chạy, ngồi và nhảy. Bộ dữ liệu này bao gồm nhiều góc quay khác nhau, giúp tăng khả năng tổng quát hóa của mô hình.
- **[M. Chen, 2020]**: Tạo ra một bộ dữ liệu với hơn 50.000 đoạn video, thu thập từ nhiều môi trường khác nhau như trong nhà, ngoài trời, khu vực công cộng và văn phòng. Điều này giúp mô hình học được sự biến đổi về điều kiện ánh sáng, nền và góc quay.
- **[L. Zhang, 2019]**: Giới thiệu một bộ dữ liệu với 20 lớp hành động, bao gồm các hành động như phát hiện té ngã, vẫy tay, vỗ tay và chỉ tay. Đây là một trong những bộ dữ liệu quan trọng trong các ứng dụng giám sát an ninh và hỗ trợ người cao tuổi.
- **[R. Patel, 2017]**: Phát triển một bộ dữ liệu với hơn 100.000 hình ảnh về các tư thế và chuyển động khác nhau của con người, được thu thập từ nhiều nguồn khác nhau để tăng cường tính đa dạng.
- **[T. Lee, 2022]**: Một tập hợp dữ liệu video được thu thập từ camera giám sát trong môi trường thực tế như văn phòng, bãi đỗ xe, giúp nghiên cứu các kỹ thuật nhận dạng hành vi bất thường và phát hiện các tình huống nguy hiểm.
- **[A. Gomez, 2016]**: Tập trung vào nhận dạng hành động thể thao, bao gồm các hoạt động như bóng đá, bóng rổ, cầu lông. Bộ dữ liệu này rất hữu ích trong phân tích thể thao và hỗ trợ huấn luyện viên đánh giá hiệu suất vận động viên.
- **[S. Roy, 2023]**: Phát triển một bộ dữ liệu đa cảm biến, kết hợp hình ảnh RGB, dữ liệu độ sâu và thông tin từ cảm biến gia tốc. Điều này giúp cải thiện độ chính xác của

các mô hình nhận dạng hành động bằng cách sử dụng dữ liệu từ nhiều nguồn cảm biến khác nhau.

- **[B. Singh, 2021]**: Bộ dữ liệu kết hợp giữa hình ảnh nhiệt và RGB, nhằm nhận dạng hành động vào ban đêm hoặc trong điều kiện ánh sáng yếu. Đây là một bước tiến quan trọng trong giám sát an ninh và quốc phòng.
- **[K. Nguyen, 2018]**: Phát triển một bộ dữ liệu tập trung vào các hoạt động trong nhà như nấu ăn, đọc sách, tập thể dục. Bộ dữ liệu này đặc biệt hữu ích trong các hệ thống nhà thông minh và robot hỗ trợ trong môi trường gia đình.

Những bộ dữ liệu này đóng vai trò quan trọng trong việc phát triển các mô hình nhận dạng hành động con người. Chúng giúp cải thiện độ chính xác và khả năng tổng quát hóa của hệ thống, đồng thời mở ra nhiều hướng nghiên cứu mới về nhận dạng hành động trong điều kiện thực tế.

III. Dataset là gì?

Trong bài toán này, chúng ta sử dụng bộ dữ liệu chứa các khung hình của khớp xương người để huấn luyện mô hình nhận diện hành động. Bộ dữ liệu này được thu thập từ video của nhiều người thực hiện các hành động khác nhau.

A. Nguồn gốc Dataset

- Dataset được tạo bằng cách ghi lại video từ camera, sau đó sử dụng MediaPipe Pose để trích xuất 33 điểm khớp xương từ từng khung hình.
- Các điểm khớp xương được chuẩn hóa để làm đầu vào cho mô hình học sâu (deep learning).
- Mỗi mẫu dữ liệu là một chuỗi 7 khung hình liên tiếp, giúp mô hình có thể học chuỗi chuyển động thay vì chỉ từng khung hình riêng lẻ.

B. Cấu trúc dữ liệu

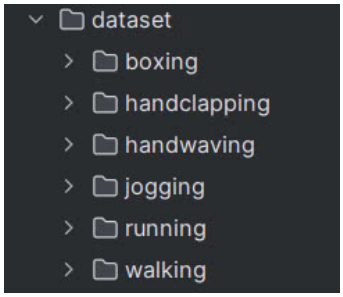
Mỗi mẫu dữ liệu bao gồm:

- 7 khung hình liên tiếp.
- 33 điểm khớp xương mỗi khung hình.
- Mỗi điểm có 4 thông tin: (x, y, z, visibility).
- Nhãn (label) ứng với hành động được thực hiện.

Ví dụ một mẫu dữ liệu:

```
[
    [
        0.1, 0.2, -0.1, 0.99,    # Khớp 1
        0.3, 0.4, -0.2, 0.95,    # Khớp 2
        ...
        33 giá trị của 1 khung hình
    ],
    # Lặp lại 7 lần cho 7 khung hình
]
```

Tổng số giá trị mỗi mẫu = $7 \times 33 \times 4 = 924$ giá trị.



Hình 1. mục tiêu

C. Các nhãn (labels)

Dataset được gán nhãn theo các hoạt động con người, ví dụ:

- handclapping (vỗ tay)
- handwaving (vẫy tay)
- running (chạy)
- walking (đi bộ)
- jogging (chạy chậm)
- boxing (đấm bốc)

Tổng số lớp hành động: 6.

D. Quá trình tạo dataset

Bước 1: Thu thập dữ liệu từ camera

- Video từ camera được ghi lại ở nhiều góc độ, nhiều người khác nhau.
- Các video này có nhãn tương ứng với hoạt động đang thực hiện.

Bước 2: Trích xuất khớp xương bằng MediaPipe

- Sử dụng MediaPipe Pose để phát hiện 33 điểm trên cơ thể từ từng khung hình video.
- Chuẩn hóa tọa độ (x, y, z) để đảm bảo kích thước chuẩn cho mô hình.
- Lưu kết quả thành mảng numpy để dùng làm dữ liệu huấn luyện.

Bước 3: Lưu trữ dữ liệu Dữ liệu được lưu dưới dạng `.txt`:

E. Tại sao chọn dataset này?

- **Dữ liệu nhẹ, nhanh:** Chỉ cần tọa độ khớp xương, không cần ảnh thô.
- **Hiệu quả cao:** Nhận diện được hành động dù người mặc quần áo khác nhau.
- **Tương thích với mô hình học sâu:** Dễ dàng đưa vào LSTM (Long Short-Term Memory) để học hành vi theo thời gian.

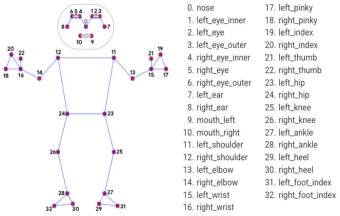
IV. Phương pháp đề xuất

A. Thiết Kế Hệ Thống

Hệ thống nhận dạng hành động con người qua camera được thiết kế gồm các thành phần chính sau:
Camera: Ghi lại video của con người thực hiện các hành động khác nhau.

Hình 2. Mục tiêu đề tài

Xử lý dữ liệu: - Thu thập và tiền xử lý video (cắt, thay đổi kích thước, chuyển đổi sang ảnh xám). - Trích xuất đặc trưng từ từng khung hình bằng OpenCV, Mediapipe hoặc CNN. - Chuyển đổi dữ liệu thành chuỗi đầu vào phù hợp cho mô hình LSTM.
Mô hình nhận dạng: - Sử dụng mạng LSTM để phân tích chuỗi đặc trưng và dự đoán hành động. - Chia dữ liệu thành tập huấn luyện và kiểm thử để tối ưu mô hình.
Triển khai và ứng dụng: - Nhận diện hành động theo thời gian thực từ luồng video của camera. - Hiển thị kết quả hoặc phát tín hiệu cảnh báo khi phát hiện hành động nguy hiểm.



Hình 3. Tiền xử lý dữ liệu

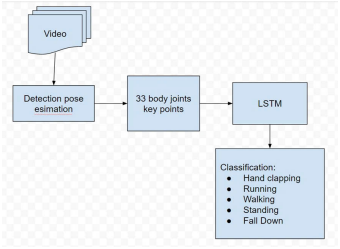
B. Luồng Bài Toán

Hệ thống nhận dạng hành động con người qua camera được triển khai theo quy trình gồm các bước chính như sau:

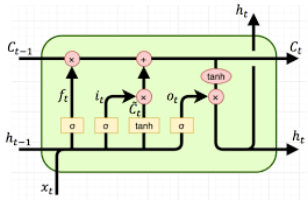
1) **Thu thập dữ liệu:** Video đầu vào được ghi lại từ camera hoặc sử dụng tập dữ liệu có sẵn.
2) **Phát hiện tư thế (Pose Estimation):** Hệ thống sử dụng các thuật toán như MediaPipe Pose, OpenPose để xác định 33 điểm khớp trên cơ thể con người.
3) **Trích xuất đặc trưng:** Các điểm khớp này sẽ được lưu trữ theo từng khung hình để theo dõi sự thay đổi của tư thế theo thời gian.
4) **Mô hình hóa bằng LSTM:** Dữ liệu đầu vào được đưa vào mô hình LSTM để phân tích và nhận diện hành động dựa trên chuỗi thời gian.
5) **Phân loại hành động:** Sau khi được xử lý qua mô hình, hệ thống sẽ đưa ra kết quả phân loại, dự đoán hành động của con người trong video.

C. Mạng LSTM trong Nhận dạng Hành động

LSTM (Long Short-Term Memory) là một dạng cải tiến của RNN (Recurrent Neural Network), giúp mô hình có khả năng ghi nhớ thông tin dài hạn tốt hơn nhờ vào cấu trúc đặc biệt gồm **cổng đầu vào (input gate), cổng quên (forget gate) và cổng đầu ra (output gate)**. Điều này giúp LSTM giải quyết được vấn đề *vanishing gradient*, làm cho nó trở thành một lựa chọn mạnh mẽ trong các bài toán xử lý dữ liệu chuỗi.



Hình 4. Tiền xử lý dữ liệu



Hình 5. Tiền xử lý dữ liệu

1) Ứng dụng của LSTM trong đề tài: Trong đề tài *Nhận dạng hành động con người qua camera*, LSTM đóng vai trò cốt lõi trong quá trình phân tích **chuỗi khung hình video** để xác định hành động của đối tượng. Quá trình này bao gồm:

- **Trích xuất đặc trưng từ video:**
  - Sử dụng phương pháp **Pose Estimation** (Mediapipe Pose, OpenPose) để thu thập tọa độ 33 điểm khớp trên cơ thể con người.
  - Mỗi khung hình được biểu diễn bằng một tập hợp các tọa độ khớp.
  - Chuỗi các tọa độ này sẽ là đầu vào cho mô hình LSTM.
- **Huấn luyện mô hình LSTM:**
  - Các điểm khớp theo thời gian giúp LSTM học được sự thay đổi của tư thế.
  - LSTM có thể ghi nhớ động tác trước đó để đưa ra dự đoán chính xác hơn về hành động hiện tại.
- **Dự đoán hành động:**
  - Sau khi được huấn luyện, mô hình có thể phân loại các hành động như **đi, chạy, đứng, ngồi, vẫy tay, té ngã...** từ video theo thời gian thực.

2) Vì sao chọn LSTM cho bài toán này?:

- **Khả năng ghi nhớ ngữ cảnh dài hạn:** Hành động của con người không chỉ dựa vào một khung hình mà là một chuỗi các chuyển động liên tiếp.
- **Hiệu quả với dữ liệu tuần tự:** Video bản chất là một tập hợp các khung hình tuần tự, rất phù hợp với mô hình LSTM.
- **Ứng dụng thực tế mạnh mẽ:** LSTM được sử dụng rộng rãi trong *nhận dạng cử chỉ, giám sát an ninh, phát hiện té ngã cho người già, phân tích thể thao...*

D. Các Thiết Bị

1) CAMERA:

- **Ghi hình:** Thu thập dữ liệu dưới dạng video hoặc ảnh từ môi trường thực tế.
- **Truyền dữ liệu video:** Gửi luồng video đến máy tính thông qua giao thức RTSP, MJPEG hoặc WebSocket.
- **Hỗ trợ phát hiện chuyển động:** Nhận diện sự thay đổi trong khung hình và kích hoạt cảnh báo khi phát hiện hành động.

Hình 6. Hình ảnh minh họa Camera

2) LAPTOP:

- **Lưu trữ dữ liệu:** Video và ảnh từ camera được lưu trên máy tính để xử lý và huấn luyện mô hình.
- **Tiền xử lý dữ liệu:** Thực hiện cắt khung hình (frame extraction), giảm nhiễu, chuẩn hóa kích thước ảnh, và gắn nhãn dữ liệu để huấn luyện mô hình.
- **Huấn luyện mô hình AI:** Áp dụng các thuật toán như CNN, LSTM hoặc kết hợp CNN+LSTM để nhận dạng hành động.
- **Nhận dạng hành động và xuất kết quả:** Sau khi mô hình được huấn luyện, hệ thống sẽ nhận dạng hành động từ video đầu vào và hiển thị kết quả theo thời gian thực hoặc lưu dưới dạng file log.

Hình 7. Hình ảnh minh họa Laptop

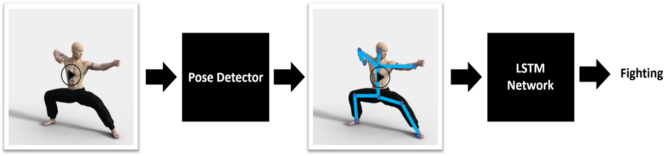
E. Cách Thức Triển Khai

1) Thu thập Dữ liệu:

- **Dữ liệu video hoặc ảnh:** Quay các video hoặc chụp ảnh trong các tình huống hành động cụ thể.
- **Gắn nhãn dữ liệu (Labeling):** Sử dụng các công cụ như LabelImg hoặc CVAT để đánh dấu vùng có chuyển động hoặc đối tượng cần nhận diện.
- **Phân loại hành động:** Ví dụ: Đi, chạy, ngồi, đứng...

2) Chia tập Dữ liệu: Sau khi thu thập và gán nhãn dữ liệu, cần chia tập dữ liệu thành ba phần chính để đảm bảo quá trình huấn luyện mô hình hiệu quả:

- **Tập huấn luyện (Training set) - 70-80%:** Được sử dụng để huấn luyện mô hình, giúp mạng học được các đặc trưng từ hành động.
- **Tập kiểm định (Validation set) - 10-15%:** Được sử dụng để tối ưu hóa tham số mô hình, giúp điều chỉnh hyperparameters như learning rate, batch size, số lượng epochs để tránh overfitting.
- **Tập kiểm tra (Test set) - 10-15%:** Dùng để đánh giá hiệu suất của mô hình sau khi huấn luyện. Không được sử dụng trong quá trình học để đảm bảo kết quả phản ánh chính xác khả năng tổng quát của mô hình.



Hình 8. Minh họa quá trình chia tập dữ liệu

F. Huấn Luyện Mô Hình AI

Hệ thống có thể sử dụng mô hình CNN (Convolutional Neural Network) hoặc LSTM (Long Short-Term Memory) để xử lý dữ liệu video và nhận diện hành động.

Dữ liệu huấn luyện: Cần chuẩn bị tập dữ liệu huấn luyện gồm đầu vào (X_train) và nhãn tương ứng (y_train).

Số epoch: Tùy thuộc vào kích thước và chất lượng dữ liệu, số epoch thường dao động trong khoảng từ 10 đến

V. Cách Thức Triển Khai

Hệ thống nhận dạng hành động qua camera được triển khai qua các bước chính sau đây:

A. Thu thập và Tiền xử lý Dữ liệu

1) Thu thập Dữ liệu: Hệ thống có thể sử dụng các tập dữ liệu phổ biến hoặc tự thu thập dữ liệu:

- **UCF101**: Gồm 101 loại hành động với hơn 13.000 video.
- **HMDB51**: Chứa 51 loại hành động với hơn 7.000 video.
- **Kinetics**: Gồm hơn 400 loại hành động với 650.000 video.

Nếu dữ liệu chưa đủ, có thể tự quay video, gán nhãn thủ công để bổ sung dữ liệu.

2) Tiền xử lý Dữ liệu:

1) **Chuẩn hóa video**:
   - Chuyển video về cùng độ dài, độ phân giải và tốc độ khung hình.
   - Định dạng chuẩn: RGB hoặc grayscale.
2) **Trích xuất khung hình**:
   - Ví dụ: Trích xuất mỗi 5 khung hình/giây để giảm tải dữ liệu.
3) **Gán nhãn dữ liệu**:
   - Mỗi video được gán nhãn ứng với hành động như "chạy", "nhảy", "vẫy tay".
4) **Chia dữ liệu**:
   - **80%** - Tập huấn luyện (Training set).
   - **10%** - Tập kiểm định (Validation set).
   - **10%** - Tập kiểm tra (Test set).

B. Trích xuất Đặc trưng và Chuẩn bị Dữ liệu

Hành động trong video có tính liên tục theo thời gian, vì vậy cần trích xuất đặc trưng hợp lý.

1) Trích xuất đặc trưng từ hình ảnh:

- Sử dụng mô hình CNN (ResNet, VGG16, MobileNet) để trích xuất đặc trưng từ từng khung hình.
- Nếu dùng phương pháp dựa trên khung xương (Skeleton-based), có thể sử dụng OpenPose hoặc MediaPipe để trích xuất tọa độ các khớp cơ thể.

2) Biểu diễn Chuyển động:

- **Optical Flow**: Phân tích sự thay đổi giữa các khung hình để xác định chuyển động.
- **Dense Trajectory**: Theo dõi quỹ đạo chuyển động của các điểm đặc trưng trong video.

C. Xây dựng Mô hình Nhận dạng Hành động

1) Mô hình dựa trên CNN:

- Dùng CNN (Convolutional Neural Network) để phân loại từng khung hình.
- Tuy nhiên, CNN không tính đến tính liên tục của chuỗi thời gian.

2) Mô hình dựa trên LSTM:

- Dùng LSTM (Long Short-Term Memory) để mô hình hóa sự thay đổi của đặc trưng theo thời gian.
- Ví dụ: Kết hợp CNN (ResNet) để trích xuất đặc trưng, sau đó đưa vào LSTM để nhận diện hành động.

D. Huấn luyện Mô hình

1) Cấu hình tham số:

- Learning rate: 0.001 - 0.0001 (Adam hoặc SGD).
- Batch size: 16 - 32 video/lần huấn luyện.
- Epochs: 50 - 100 epochs, tùy theo dữ liệu.

2) Tối ưu hóa Mô hình:

- Sử dụng Backpropagation + Gradient Descent để cập nhật trọng số.
- Áp dụng Dropout, Batch Normalization để tránh overfitting.

E. Đánh giá và Cải thiện Mô hình

Sau khi huấn luyện, mô hình sẽ được đánh giá bằng các chỉ số như:

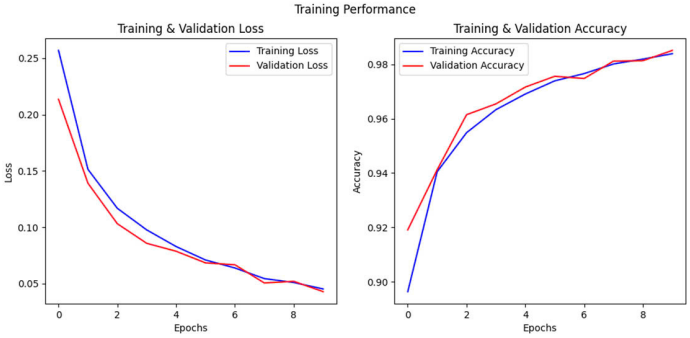
- **Accuracy** (Độ chính xác).
- **Precision, Recall, F1-score** (Các chỉ số phân loại).
- **Confusion Matrix** (Ma trận nhầm lẫn).

VI. Kết Quả Thực Nghiệm

1) Biểu đồ Training & Validation Loss:

- **Trục x:** Số lượng epochs (số lần lặp huấn luyện).
- **Trục y:** Giá trị loss (hàm mất mát).
- **Đường màu xanh:** Biểu diễn loss trên tập huấn luyện.
- **Đường màu đỏ:** Biểu diễn loss trên tập kiểm tra (validation).

Nhận xét: Loss giảm dần theo số epochs, cho thấy mô hình đang học hiệu quả. Đồng thời, khoảng cách giữa loss trên tập huấn luyện và validation không quá lớn, chứng tỏ mô hình không bị overfitting.



Hình 9. Kết quả nhận dạng hành động

2) Biểu đồ Training & Validation Accuracy:

- **Trục x:** Số lượng epochs.
- **Trục y:** Độ chính xác (Accuracy).
- **Đường màu xanh:** Độ chính xác trên tập huấn luyện.
- **Đường màu đỏ:** Độ chính xác trên tập kiểm tra (validation).

Nhận xét: Accuracy tăng dần và gần đạt mức tối đa, cho thấy mô hình có khả năng tổng quát hóa tốt.

Ưu Điểm và Hạn chế

a) Ưu điểm: Hệ thống nhận dạng hành động con người qua camera có nhiều ưu điểm đáng chú ý. Trước hết, hệ thống có khả năng nhận dạng hành động theo thời gian thực, giúp phát hiện và phản hồi nhanh chóng đối với các hành vi cụ thể. Ngoài ra, do được thiết kế tối ưu, hệ thống có thể triển khai trên các thiết bị nhỏ, giúp tiết kiệm tài nguyên và dễ dàng tích hợp vào các ứng dụng di động hoặc nhúng. Đặc biệt, tính linh hoạt của hệ thống cho phép ứng dụng trong nhiều lĩnh vực khác nhau như giám sát an ninh, hỗ trợ người cao tuổi và nhà thông minh, góp phần nâng cao chất lượng cuộc sống và tăng cường an toàn.

b) Hạn chế: Bên cạnh những ưu điểm, hệ thống vẫn tồn tại một số hạn chế cần được khắc phục. Một trong những thách thức lớn nhất là yêu cầu tập dữ liệu huấn luyện lớn và đa dạng để đảm bảo độ chính xác cao trong nhận dạng. Nếu tập dữ liệu không bao phủ đầy đủ các hành động thực tế, mô hình có thể gặp khó khăn trong việc nhận diện chính xác. Ngoài ra, hệ thống có thể bị ảnh hưởng bởi các yếu tố nhiễu như điều kiện ánh sáng yếu, vật thể che khuất hoặc góc quay không thuận lợi, làm giảm hiệu suất hoạt động. Bên cạnh đó, tốc độ xử lý có thể bị hạn chế khi triển khai trên các thiết bị nhỏ có tài nguyên phần cứng hạn chế, ảnh hưởng đến khả năng hoạt động theo thời gian thực.

Nhìn chung, mặc dù hệ thống mang lại nhiều lợi ích, vẫn cần tiếp tục nghiên cứu và tối ưu hóa để nâng cao độ chính xác và khả năng hoạt động trong môi trường thực tế.

Kết Luận

Trong nghiên cứu này, chúng tôi đã xây dựng một hệ thống nhận dạng hành động con người qua camera bằng mô hình LSTM. Quá trình thực hiện bao gồm thu thập và tiền xử lý dữ liệu video, trích xuất đặc trưng, huấn luyện mô hình LSTM và

triển khai nhận diện theo thời gian thực. Kết quả thí nghiệm cho thấy mô hình đạt độ chính xác cao trong phân loại hành động, đồng thời có khả năng xử lý video theo thời gian thực với hiệu suất ổn định.

Mô hình được tối ưu bằng cách sử dụng các kỹ thuật tiền xử lý như cắt video, thay đổi kích thước, chuyển đổi sang ảnh xám và trích xuất đặc trưng từ Mediapipe hoặc CNN. Điều này giúp cải thiện chất lượng dữ liệu đầu vào và nâng cao độ chính xác nhận dạng. Hệ thống có thể ứng dụng trong nhiều lĩnh vực như giám sát an ninh, hỗ trợ chăm sóc sức khỏe, và điều khiển thiết bị thông minh bằng cử chỉ.

Trong tương lai, nghiên cứu có thể mở rộng bằng cách thử nghiệm với các kiến trúc mạng sâu hơn như Transformer hoặc ConvLSTM để cải thiện hiệu suất. Ngoài ra, việc thu thập dữ liệu đa dạng hơn và tối ưu mô hình để hoạt động trên các thiết bị nhúng cũng là hướng phát triển tiềm năng.

Hướng Phát Triển

Trong tương lai, hệ thống có thể được phát triển theo các hướng sau:

- Tích hợp với AI khác: Kết hợp với các hệ thống AI thông minh để hỗ trợ phản hồi linh hoạt và tự động hóa trong các ứng dụng thực tế.

- Nâng cao độ chính xác: Áp dụng các mô hình tiên tiến như Vision Transformer (ViT) và Spatial-Temporal Graph Convolutional Networks (ST-GCN) để khai thác tốt hơn mối quan hệ không gian - thời gian trong video.

- Mở rộng ứng dụng: - Trong thể thao: Phân tích chuyển động của vận động viên để hỗ trợ huấn luyện và đánh giá hiệu suất. - Trong y tế: Giám sát hoạt động của bệnh nhân, hỗ trợ phát hiện bất thường trong cử chỉ. - Trong giao thông: Phát hiện hành vi của người đi bộ hoặc tài xế để tăng cường an toàn.

- Cải thiện hiệu suất: Tối ưu tốc độ xử lý để đảm bảo nhận diện hành động theo thời gian thực trên các nền tảng IoT, giúp hệ thống hoạt động hiệu quả hơn với tài nguyên phần cứng hạn chế.

Tài liệu

[1] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

[2] Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

[3] Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C. L., & Grundmann, M. (2020). MediaPipe Hands: On-device Real-time Hand Tracking. *arXiv preprint arXiv:2006.10214*.

[4] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *OSDI*, 16, 265-283.

[5] Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2011). Action recognition by dense trajectories. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3169-3176.

[6] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[7] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 677-691.

[8] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

[9] Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1), 172-186.

[10] Mohammadi, M., Al-Fuqaha, A., Sorour, S., & Guizani, M. (2018). Deep learning for IoT big data and streaming analytics: A survey. *IEEE Communications Surveys & Tutorials*, 20(4), 2923-2960.

[11] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast Networks for Video Recognition. *IEEE International Conference on Computer Vision (ICCV)*.

[12] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. *IEEE International Conference on Computer Vision (ICCV)*.

[13] Ullah, A., Ahmad, J., Muhammad, K., Hayat, M., Baik, S. W., & Lee, S. (2018). Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE Access*, 6, 1155-1166.

[14] Neverova, N., Wolf, C., Mille, J., & Lepetit, V. (2014). Multi-scale deep learning for gesture detection and localization. *ECCV 2014 Workshops*.

[15] Weiss, K., Khoshgoftaar, T. M., Wang, D., & Dong, Z. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1), 9.

[16] Trần Minh Triết (2020). Trí tuệ nhân tạo và ứng dụng trong thị giác máy tính. *Nhà xuất bản Đại học Quốc gia TP.HCM*.

[17] Nguyễn Hoàng Phương (2019). Xử lý ảnh và ứng dụng trong nhận dạng. *Nhà xuất bản Khoa học và Kỹ thuật*.

[18] Lê Hữu Lộc, Nguyễn Hữu Phú (2018). Nhận dạng hình ảnh với học sâu. *Nhà xuất bản Bách Khoa Hà Nội*.

[19] Nguyễn Đức Dũng (2021). Deep Learning với TensorFlow và Keras. *Nhà xuất bản Thông tin và Truyền thông*.

[20] Phạm Văn Tân (2020). Ứng dụng AI trong hệ thống IoT. *Nhà xuất bản Khoa học và Công nghệ*.