Nhận Dạng Hành Động Con Người Qua CAMERA

[*]Quàng Minh Anh*, Nguyễn Anh Tuấn*, Phạm Đình Tuấn*, Nguyễn Tuấn Dũng*, Nguyễn Thị Mai Lan*

Giảng viên hướng dẫn: ThS Lê Trung Hiếu*, Nguyễn Văn Nhân*

*Khoa Công Nghệ Thông Tin, Trường Đại học Đại Nam, Hà Nội, 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. Công trình liên quan

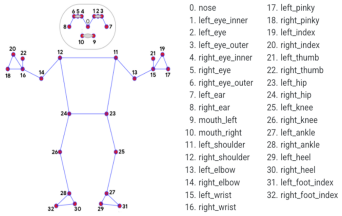
Duy trì tính toàn vẹn của các thông số kỹ thuật Trong bài báo này, chúng em trình bày một phương pháp tiếp cận để nhận dạng hành động của con người thông qua camera, tập trung vào việc nhận dạng các hành động như chạy, đấm bốc, vẫy tay và chào hỏi. Phương pháp của chúng tôi bao gồm xử lý dữ liệu video, trích xuất các tính năng chính và áp dụng các kỹ thuật học máy để phân loại các hành động khác nhau của con người theo thời gian thực.

Duy trì tính toàn vẹn của các thông số kỹ thuật đảm bảo rằng mô hình đề xuất của chúng em có thể được tích hợp hiệu quả vào các ứng dụng tương tác giữa người và máy tính và thị giác máy tính rộng hơn. Bằng cách tuân thủ các hướng dẫn này, chúng tôi cung cấp một phương pháp tiếp cận có cấu trúc phù hợp với các tiêu chuẩn của ngành đồng thời chứng minh tính khả thi của việc nhận dạng hành động theo thời gian thực thông qua các hệ thống dựa trên camera.

III. 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.
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.



Hình 1. Tiền xử lý dữ liệu

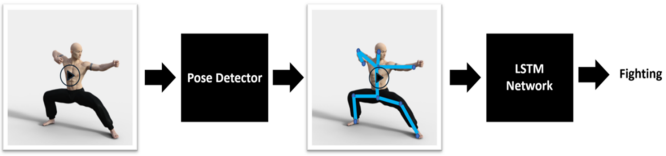
Hình 2. mục tiêu

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.

B. 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...



Hình 3. Minh họa quá trình chia tập dữ liệu

2) Chia tập dữ liệu:

C. 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

D. PYCHARM VÀ ỨNG DỤNG TRONG BÀI TOÁN

- **1. Thu thập và Tiền xử lý Dữ liệu**
  Đây là bước quan trọng giúp mô hình học được các đặc trưng cần thiết từ video đầu vào.

LAPTOP

Hình 4. 2

- **1.1 Thu thập dữ liệu**
  Sử dụng tập dữ liệu có sẵn như:
  - UCF101 (101 loại hành động, hơn 13.000 video).
  - HMDB51 (51 loại hành động, hơn 7.000 video).
  - Kinetics (hơn 400 loại hành động, 650.000 video).

  Nếu dữ liệu chưa đủ, bạn có thể tự thu thập video bằng cách quay video và gán nhãn thủ công.
- **1.2 Tiền xử lý dữ liệu**
  Cắt và chuẩn hóa video:
  - Chuyển đổi tất cả video về cùng độ dài, độ phân giải, tốc độ khung hình (fps).
  - Định dạng chuẩn là RGB hoặc grayscale.

  Trích xuất khung hình (frames):
  - Ví dụ: Trích xuất mỗi 5 khung hình/giây để giảm tải dữ liệu.

  Gán nhãn dữ liệu:
  - Mỗi video phải có nhãn ứng với hành động như "chạy", "boxing", "vẫy tay".

  Tách dữ liệu thành 3 tập:
  - Training set (80%) – Dùng để huấn luyện mô hình.
  - Validation set (10%) – Dùng để tối ưu tham số.
  - Test set (10%) – Đánh giá hiệu suất mô hình.
- **2. Trích xuất Đặc trưng và Chuẩn bị Dữ liệu**
  Hành động trong video có tính thời gian, vì vậy cần trích xuất đặc trưng hợp lý.
- **2.1 Trích xuất đặc trưng từ hình ảnh**
  - Dùng Pretrained CNN (ResNet, VGG16, MobileNet) để trích xuất đặc trưng hình ảnh từ từng khung hình.
  - Nếu dùng Skeleton-based approach, có thể dùng OpenPose hoặc MediaPipe để lấy tọa độ khớp cơ thể.
- **2.2 Biểu diễn chuyển động**
  - Optical Flow: Phân tích sự thay đổi giữa các khung hình để mô hình hiểu được chuyển động.
  - Dense Trajectory: Theo dõi quỹ đạo chuyển động của các điểm đặc trưng.
- **3. Xây dựng Mô hình Nhận Dạng Hành Động**
  Mô hình có thể được xây dựng theo nhiều phương pháp khác nhau.

Hình 5. NHẬN DẠNG CHUYỂN ĐỘNG QUA CAMERA

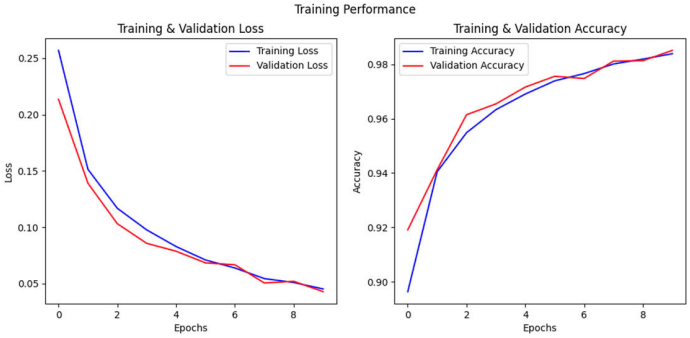
- **3.1 Mô hình dựa trên CNN**
  - Dùng CNN (Convolutional Neural Network) để phân loại từng khung hình.
  - Tuy nhiên, phương pháp này không tính đến sự liên tục của các khung hình theo thời gian.
- **3.2 Mô hình dựa trên RNN/LSTM**
  - Sử dụng LSTM (Long Short-Term Memory) để mô hình hóa sự thay đổi của đặc trưng theo thời gian.
  - Ví dụ: Dùng ResNet để trích xuất đặc trưng ảnh rồi đưa vào LSTM để nhận diện hành động.
- **4. Huấn luyện Mô Hình**
  Quá trình huấn luyện giúp mô hình học được cách nhận dạng hành động.
- **4.1 Cấu hình tham số**
  - Learning rate: 0.001 - 0.0001 (dùng Adam hoặc SGD).
  - Batch size: 16 - 32 video/lần huấn luyện.
  - Epochs: 50 - 100 epochs tùy theo dữ liệu.
- **4.2 Tối ưu hóa mô hình**
  - Dùng Backpropagation + Gradient Descent để cập nhật trọng số.
  - Áp dụng Dropout, Batch Normalization để tránh overfitting.
- **5. Đánh giá và Cải thiện Mô Hình**
  Sau khi huấn luyện, cần đánh giá mô hình để kiểm tra độ chính xác.

E. 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 6. 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.

F. 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] J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[2] A. Graves, “Long Short-Term Memory,” in *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A Closer Look at Spatiotemporal Convolutions for Action Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6450–6459.

[4] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and others, “TensorFlow: A System for Large-Scale Machine Learning,” in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.

[5] F. Chollet, “Deep Learning with Python,” Manning Publications, 2017.

[6] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, “Searching for MobileNetV3,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324.

[7] P. Huang, W. Zhang, and H. Xu, “Real-Time Human Action Recognition Using LSTM and OpenPose,” in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4508–4520, 2020.