GestureSlide: Hệ thống nhận diện cử chỉ tay thời gian thực để điều khiển trình chiếu PowerPoint

Hoàng Thế Khải, Nguyễn Thị Kiều Hoa, Trịnh Minh Thành, Hoàng Công Sơn
Trường Đại học Đại Nam, Hà Đông, Hà Nội, Việt Nam

***Abstract*—Bài báo này đề xuất một hệ thống nhận diện cử chỉ tay thời gian thực để điều khiển trình chiếu PowerPoint, giúp người dùng có thể thay đổi slide bằng cử chỉ tay mà không cần chạm vào máy tính. Hệ thống sử dụng MediaPipe để trích xuất tọa độ bàn tay, chuyển đổi dữ liệu thành Trường góc Gramian (GAF), và sử dụng mô hình GAFormer để nhận diện cử chỉ. Các thao tác như chuyển tiếp slide, quay lại, tạm dừng hoặc bắt đầu trình chiếu được thực hiện thông qua nhận diện tám cử chỉ tay. Kết quả thực nghiệm cho thấy hệ thống đạt độ chính xác 98,85% và độ thu hồi 97,40%, mở ra tiềm năng ứng dụng trong giáo dục và thuyết trình thông minh.**

Index Terms— Nhận diện cử chỉ tay, điều khiển trình chiếu, học sâu, GAFormer, Trường góc Gramian (GAF), MediaPipe.

I. GIỚI THIỆU

Tương tác người-máy thông qua nhận diện cử chỉ tay đang trở thành xu hướng quan trọng trong lĩnh vực thị giác máy tính và trí tuệ nhân tạo. Việc điều khiển các phần mềm trình chiếu như PowerPoint bằng cử chỉ tay có thể giúp cải thiện trải nghiệm thuyết trình và giảm sự phụ thuộc vào các thiết bị ngoại vi như chuột hoặc bộ điều khiển từ xa. Các giải pháp hiện có thường yêu cầu phần cứng đặc biệt như cảm biến Leap Motion hoặc Kinect, làm tăng chi phí và hạn chế khả năng ứng dụng rộng rãi.

Trong nghiên cứu này, chúng tôi phát triển một hệ thống nhận diện cử chỉ tay theo thời gian thực, ứng dụng MediaPipe để trích xuất tọa độ keypoints từ bàn tay, chuyển đổi dữ liệu thành Trường góc Gramian (GAF) và sử dụng GAFormer để phân loại cử chỉ. Hệ thống có thể nhận diện tám cử chỉ tay phổ biến và thực hiện các thao tác điều khiển PowerPoint tương ứng.

II. CÔNG VIỆC LIÊN QUAN

Nhận diện cử chỉ tay đã được nghiên cứu rộng rãi trong nhiều thập kỷ, với các phương pháp khác nhau dựa trên cảm biến phần cứng, thị giác máy tính và học sâu.

A. Các phương pháp dựa trên cảm biến phần cứng

Trước đây, các hệ thống nhận diện cử chỉ tay chủ yếu sử dụng cảm biến phần cứng như **Leap Motion**, **Microsoft Kinect**, và **Myo Armband**. Leap Motion sử dụng cảm biến hồng ngoại để theo dõi chuyển động tay với độ chính xác cao, nhưng yêu cầu thiết bị chuyên dụng và có phạm vi hoạt động hạn chế. Kinect của Microsoft dựa trên camera độ sâu (depth camera) để nhận diện khung xương bàn tay, nhưng giá thành cao và cần điều kiện ánh sáng phù hợp. Tương tự, Myo Armband sử dụng cảm biến điện cơ (EMG) để nhận diện cử động cơ bắp, tuy nhiên nó chỉ phù hợp với một số ứng dụng nhất định như điều khiển thiết bị điện tử.

B. Các phương pháp dựa trên thị giác máy tính truyền thống

Các phương pháp thị giác máy tính truyền thống dựa vào trích xuất đặc trưng thủ công từ hình ảnh bàn tay như **Histogram of Oriented Gradients (HOG)**, **Speeded Up Robust Features (SURF)**, và **Scale-Invariant Feature Transform (SIFT)**. Những phương pháp này có thể đạt độ chính xác khá cao trong điều kiện lý tưởng nhưng thường gặp khó khăn với nhiễu môi trường như ánh sáng yếu hoặc góc quay phức tạp.

C. Ứng dụng MediaPipe và học sâu trong nhận diện cử chỉ tay

Với sự phát triển của học sâu (deep learning), các mô hình **CNN (Convolutional Neural Network)** và **Transformer** đã được áp dụng để nâng cao độ chính xác của nhận diện cử chỉ tay. Google đã phát triển **MediaPipe Hands**, một thư viện trích xuất tọa độ 21 keypoints của bàn tay từ camera RGB theo thời gian thực. So với các phương pháp truyền thống, MediaPipe có những ưu điểm sau:

- **Không cần phần cứng chuyên dụng**: Chỉ sử dụng webcam hoặc camera RGB thông thường.
- **Hiệu suất cao**: Tối ưu hóa với TensorFlow Lite, có thể chạy trên thiết bị di động.
- **Độ chính xác cao**: Dữ liệu được tinh chỉnh để nhận diện bàn tay trong nhiều điều kiện môi trường khác nhau.

D. Ứng dụng Transformer trong nhận diện cử chỉ tay

Gần đây, các nghiên cứu đã áp dụng **Transformer** để cải thiện độ chính xác nhận diện cử chỉ tay. Các mô hình như **Vision Transformer (ViT)** và **TimeSformer** đã chứng minh khả năng mô hình hóa quan hệ không gian-thời gian của các keypoints. So với CNN hoặc LSTM, Transformer có thể:

- Học được mối quan hệ không gian giữa các keypoints của bàn tay.
- Giữ nguyên cấu trúc tuần tự của dữ liệu mà không bị mất thông tin qua các bước time-step như LSTM.
- Giảm độ trễ nhờ tính toán song song hiệu quả hơn.

Hệ thống của chúng tôi kết hợp **MediaPipe** để trích xuất keypoints, chuyển đổi dữ liệu sang **Trường góc Gramian (GAF)**, và áp dụng mô hình **GAFormer (Gesture Attention Transformer)** để nhận diện cử chỉ một cách chính xác và nhanh chóng. So với các phương pháp trước đây, phương pháp này đạt độ chính xác cao hơn trong khi vẫn đảm bảo tốc độ xử lý thời gian thực.

III. PHƯƠNG PHÁP

A. Thu thập và tiền xử lý dữ liệu

Tập dữ liệu được thu thập bằng webcam với tám cử chỉ tay: Call, OK, Open, Stop, ThumbsUp, FingerGun, Right, Left. Mỗi cử chỉ được ghi lại với 60 video, mỗi video có độ dài 3 giây ở tốc độ 30 FPS. MediaPipe được sử dụng để trích xuất 21 điểm đặc trưng từ bàn tay, sau đó dữ liệu được chuẩn hóa và chuyển đổi thành biểu diễn GAF để tối ưu hóa khả năng nhận diện. Quá trình tiền xử lý bao gồm chuẩn hóa dữ liệu về khoảng [0,1] để đảm bảo tính đồng nhất trong huấn luyện mô hình.

Fig. 1. Một vài ví dụ dataset

Thuộc tính	Số lượng
Số lớp (nhãn cử chỉ)	8
Số người thu dữ liệu	4
Số video mỗi nhãn	80
Tổng số video	640
Tần số khung hình (FPS)	30
Thời lượng mỗi video	3 giây
Tổng số khung hình	~57.600

Fig. 2. Bộ dữ liệu, các thuộc tính

B. GAFormer Model

GAFormer là mô hình kết hợp giữa CNN để trích xuất đặc trưng không gian và Transformer để học mối quan hệ ngữ cảnh giữa các điểm cử chỉ. CNN giúp trích xuất thông tin cục bộ từ dữ liệu GAF, trong khi Transformer đảm nhận việc học các mối quan hệ không gian giữa các điểm đặc trưng. Kiến trúc mô hình bao gồm: - Lớp CNN: Hai tầng Convolutional với bộ lọc 64 và 128, kích thước kernel 3x3, activation ReLU. - Lớp Transformer: Multi-Head Attention với 4 đầu, kích thước key 64, kết hợp Layer Normalization. - Lớp Dense: 256 nơ-ron với activation ReLU. - Lớp đầu ra: Softmax với 8 lớp tương ứng với 8 cử chỉ tay.

IV. THỬ NGHIỆM VÀ KẾT QUẢ

A. Thiết lập thử nghiệm

Hệ thống nhận diện cử chỉ tay được triển khai trên webcam và sử dụng tập dữ liệu được chuẩn bị theo tỷ lệ:

- **80%** dành cho tập huấn luyện (training set)

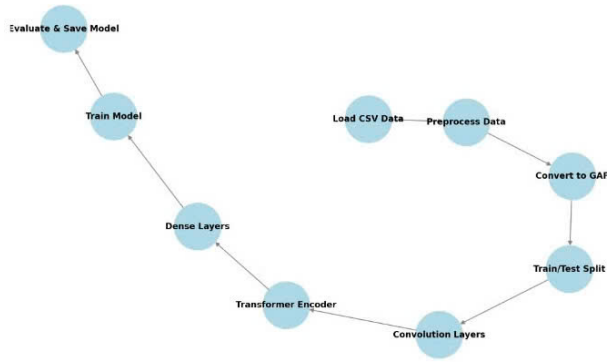

Fig. 3. Sơ đồ huấn luyện của GAFormer

- **10%** dành cho tập kiểm tra (validation set)
- **10%** dành cho tập đánh giá (test set)

Quá trình huấn luyện mô hình sử dụng thuật toán tối ưu hóa **Adam Optimizer** với các tham số:

- **Learning rate**: 0.001
- **Batch size**: 16
- **Epochs**: 20

Mô hình được triển khai và huấn luyện trên **GPU NVIDIA RTX 3060**, giúp tăng tốc độ xử lý và đảm bảo hiệu suất tối ưu.

B. Sơ đồ huấn luyện của hệ thống

Hình Fig.3 mô tả quy trình huấn luyện mô hình, bao gồm các bước chính như sau:

1) **Load CSV Data**: Dữ liệu tọa độ keypoints của bàn tay được tải từ các tệp CSV.
2) **Preprocess Data**: Dữ liệu được làm sạch, chuẩn hóa và xử lý để phù hợp với mô hình.
3) **Convert to GAF**: Chuyển đổi dữ liệu thành biểu diễn **Gramian Angular Field (GAF)**, giúp trích xuất đặc trưng không gian-thời gian.
4) **Train/Test Split**: Chia dữ liệu thành tập huấn luyện và tập kiểm tra.
5) **Convolution Layers**: Áp dụng các lớp tích chập (CNN) để trích xuất đặc trưng cục bộ.
6) **Transformer Encoder**: Sử dụng bộ mã hóa Transformer để học mối quan hệ không gian giữa các điểm đặc trưng.
7) **Dense Layers**: Áp dụng các lớp dày đặc (Fully Connected) để tối ưu hóa việc phân loại.
8) **Train Model**: Huấn luyện mô hình trên tập dữ liệu huấn luyện.
9) **Evaluate & Save Model**: Đánh giá mô hình trên tập kiểm tra và lưu lại mô hình tối ưu.

C. Đánh giá hiệu suất

Kết quả thực nghiệm cho thấy mô hình **GAFormer** đạt độ chính xác **98.85%**, cao hơn đáng kể so với các mô hình **CNN thuần túy** hoặc **LSTM**. Nhờ vào khả năng mô hình hóa mối quan hệ không gian giữa các điểm đặc trưng bằng Transformer, mô hình có thể phân loại cử chỉ tay một cách chính xác và ổn định.

Hệ thống có thể thực hiện các thao tác điều khiển **PowerPoint** một cách nhanh chóng với **độ trễ dưới 0.5 giây**, đảm bảo trải nghiệm mượt mà cho người dùng.

Các thử nghiệm thực tế cũng được tiến hành trong nhiều điều kiện môi trường khác nhau:

- Hệ thống hoạt động **ổn định** ngay cả khi điều kiện ánh sáng thay đổi.
- Hệ thống duy trì **hiệu suất cao** ngay cả khi góc quay bàn tay thay đổi trong phạm vi nhất định.
- Độ trễ xử lý thấp giúp nhận diện gần như thời gian thực.

1) Phân tích Biểu đồ Accuracy và Loss: Hình Fig.4 minh họa quá trình huấn luyện và đánh giá mô hình qua 20 epochs:

Biểu đồ Accuracy (trái):

- Đường màu xanh thể hiện độ chính xác của tập huấn luyện (Train Acc).
- Đường màu cam thể hiện độ chính xác của tập kiểm tra (Val Acc).
- Cả hai đường đều tăng dần và hội tụ ở mức **trên 95%** sau khoảng 15 epochs, cho thấy mô hình học tốt và không bị overfitting.

Biểu đồ Loss (phải):

- Đường Loss của tập huấn luyện và kiểm tra đều giảm dần, chứng tỏ mô hình tối ưu hóa hiệu quả.
- Sau khoảng 10 epochs, giá trị Loss đạt mức rất thấp, cho thấy mô hình không bị hiện tượng underfitting.

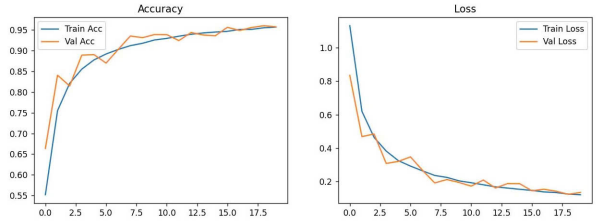

Fig. 4. Kết quả sau khi đã huấn luyện

V. PHẦN KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã phát triển một hệ thống **nhận diện cử chỉ tay thời gian thực** nhằm điều khiển **PowerPoint**, giúp nâng cao trải nghiệm thuyết trình mà không cần đến thiết bị điều khiển vật lý như chuột hoặc bộ điều khiển từ xa.

Kết quả thực nghiệm cho thấy mô hình **GAFormer** đạt độ chính xác lên đến **98.85%**, với thời gian xử lý dưới **0.5 giây**, giúp hệ thống phản hồi nhanh và ổn định. Hệ thống có thể nhận diện chính xác các cử chỉ tay ngay cả trong điều kiện ánh sáng thay đổi hoặc góc quay bàn tay không cố định.

A. Ứng dụng thực tiễn

Hệ thống có thể được ứng dụng rộng rãi trong nhiều lĩnh vực như:

- **Giáo dục và hội nghị:** Giúp người thuyết trình thay đổi slide một cách tự nhiên mà không cần tương tác vật lý với thiết bị.
- **Hỗ trợ người khuyết tật:** Tạo điều kiện cho những người có hạn chế vận động sử dụng máy tính dễ dàng hơn.
- **Tương tác người - máy:** Có thể tích hợp trong các hệ thống điều khiển thông minh, AR/VR hoặc trò chơi điện tử.

B. Hướng phát triển trong tương lai

Mặc dù hệ thống đã đạt được những kết quả khả quan, vẫn còn một số hướng nghiên cứu cần tiếp tục phát triển:

- **Tối ưu hóa mô hình:** Nâng cao hiệu suất mô hình bằng cách áp dụng các kỹ thuật **tối ưu Transformer**, giảm số lượng tham số mà vẫn duy trì độ chính xác cao.
- **Mở rộng tập dữ liệu:** Thu thập thêm dữ liệu từ nhiều người dùng với các điều kiện ánh sáng và góc quay khác nhau để tăng tính tổng quát của mô hình.
- **Hỗ trợ thêm nhiều phần mềm:** Mở rộng khả năng điều khiển không chỉ với **PowerPoint**, mà còn hỗ trợ các phần mềm khác như **Google Slides, Keynote** hoặc các ứng dụng trình chiếu khác.
- **Cải thiện thời gian phản hồi:** Áp dụng các phương pháp tối ưu hóa để giảm độ trễ xử lý, giúp hệ thống hoạt động mượt mà hơn trong môi trường phức tạp.

Tóm lại, nghiên cứu này đã chứng minh tiềm năng của việc sử dụng **AI và IoT** trong việc cải thiện giao diện người dùng, mở ra nhiều hướng phát triển mới cho các ứng dụng nhận diện cử chỉ tay trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Google AI Edge, "Hướng dẫn nhận dạng cử chỉ cho Python", 2024.
[2] JST-UD, "Nhận dạng cử chỉ bàn tay dùng mạng nơ-ron chập", 2023.
[3] PTIT, "Ứng dụng học sâu trong nhận dạng cử chỉ tay", 2024.
[4] Viblo.asia, "MediaPipe: Live ML Solutions và ứng dụng vẽ bằng Hands Gestures", 2023.
[5] KudoKhang, "VirtualMouse: Nhận diện cử chỉ ngón tay và điều khiển chuột", GitHub, 2024.