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

Hoàng Thế Khải, Trịnh Minh Thành, Nguyễn Thị Kiều Hoa, Hoàng Công Sơn
Nhóm 5, CNTT 16-03, Khoa Công Nghệ Thông Tin
Trường Đại Học Đại Nam, Việt Nam
Ks. Nguyễn Thái Khánh, ThS. Lê Trung Hiếu
Giảng viên hướng dẫn, Khoa Công Nghệ Thông Tin
Trường Đại Học Đại Nam, Việt Nam

***Abstract*—Trong thời đại công nghệ số, việc tương tác giữa con người và máy tính (HCI - Human-Computer Interaction) ngày càng được cải tiến để mang lại trải nghiệm mượt mà và tiện lợi hơn. Một trong những phương pháp tiếp cận tiên tiến trong HCI là nhận diện cử chỉ tay, giúp người dùng điều khiển các thiết bị và phần mềm mà không cần tiếp xúc vật lý. 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, tạm dừng hoặc bắt đầu bài thuyết trình chỉ bằng các cử chỉ tay, thay thế chuột hoặc bộ điều khiển truyền thống.**

Hệ thống đề xuất sử dụng thư viện MediaPipe Hands để trích xuất tọa độ 21 keypoints của bàn tay từ camera RGB mà không cần đến phần cứng chuyên biệt. Dữ liệu keypoints sau đó được chuyển đổi thành Trường góc Gramian (Gramian Angular Field - GAF), một kỹ thuật giúp mô hình học sâu có thể tận dụng thông tin không gian-thời gian một cách hiệu quả hơn. Để nhận diện chính xác cử chỉ tay, chúng tôi áp dụng mô hình GAFormer (Gesture Attention Transformer), một phiên bản cải tiến của Transformer được thiết kế chuyên biệt cho nhận diện cử chỉ tay.

Hệ thống hỗ trợ tám cử chỉ tay chính, bao gồm: chuyển tiếp slide, quay lại slide trước, bắt đầu trình chiếu, dừng trình chiếu, tạm dừng, tiếp tục, thoát khỏi chế độ trình chiếu và reset. Các cử chỉ này giúp người dùng có thể điều khiển PowerPoint một cách trực quan mà không cần tiếp xúc trực tiếp với máy tính, phù hợp với các bối cảnh như giảng dạy, thuyết trình hội nghị, hoặc điều khiển thiết bị từ xa trong các môi trường hạn chế tiếp xúc.

Chúng tôi đã tiến hành các thử nghiệm thực nghiệm trên tập dữ liệu mẫu cử chỉ tay, thu thập từ nhiều người dùng trong các điều kiện ánh sáng khác nhau. Kết quả cho thấy hệ thống đạt độ chính xác 98,85%, độ thu hồi 97,40% và hoạt động mượt mà trong thời gian thực với độ trễ trung bình dưới 50ms.

Những kết quả này cho thấy tiềm năng ứng dụng lớn của hệ thống trong lĩnh vực giáo dục, hội nghị thông minh, nhà thông minh, và các ứng dụng điều khiển từ xa khác. Hệ thống có thể tiếp tục được mở rộng để nhận diện nhiều cử chỉ hơn hoặc áp dụng vào các lĩnh vực khác như điều khiển robot, hỗ trợ người khuyết tật hoặc tăng cường khả năng tương tác trong thực tế ảo (VR) và thực tế tăng cường (AR).

Index Terms— Nhận diện cử chỉ tay, điều khiển trình chiếu, học sâu, GAFormer, Trường góc Gramian (GAF), MediaPipe, Human-Computer Interaction (HCI), Trí tuệ nhân tạo (AI).

I. GIỚI THIỆU

Trong những năm gần đây, sự phát triển nhanh chóng của Trí tuệ nhân tạo (AI) và Thị giác máy tính (Computer Vision) đã tạo ra những tiến bộ đáng kể trong lĩnh vực Tương tác người-máy (HCI - Human-Computer Interaction). Nhận diện cử chỉ tay là một trong những phương pháp tiên tiến giúp người dùng tương tác với máy tính một cách tự nhiên mà không cần tiếp xúc vật lý, đặc biệt hữu ích trong các môi trường yêu cầu tính linh hoạt cao như giảng dạy, thuyết trình và điều khiển từ xa.

Hiện nay, các phần mềm trình chiếu như Microsoft PowerPoint và Google Slides chủ yếu được điều khiển bằng chuột, bàn phím hoặc bộ điều khiển từ xa. Tuy nhiên, trong các tình huống như giảng dạy, hội nghị hoặc thuyết trình trước đám đông, việc sử dụng các thiết bị này có thể gây gián đoạn, làm giảm sự tập trung của người thuyết trình. Một giải pháp thay thế là sử dụng nhận diện cử chỉ tay để điều khiển các thao tác trình chiếu, giúp tăng cường trải nghiệm người dùng và mang lại sự tiện lợi tối đa.

A. Những thách thức trong nhận diện cử chỉ tay

Mặc dù nhận diện cử chỉ tay đã được nghiên cứu rộng rãi, nhưng vẫn tồn tại nhiều thách thức khi triển khai hệ thống trong thực tế:

- Yêu cầu phần cứng chuyên dụng: Một số phương pháp trước đây sử dụng cảm biến Leap Motion, Microsoft Kinect hoặc Myo Armband, tuy mang lại độ chính xác cao nhưng đòi hỏi chi phí lớn và không phổ biến trên các thiết bị thông thường.
- Sự đa dạng của môi trường: Ánh sáng, góc quay của bàn tay, nền phức tạp có thể làm giảm độ chính xác của hệ thống nhận diện.
- Tốc độ xử lý: Hệ thống phải hoạt động theo thời gian thực để đảm bảo sự liền mạch trong quá trình thuyết trình.
- Tính cá nhân hóa: Mỗi người dùng có hình dạng bàn tay khác nhau, cách thực hiện cử chỉ có thể không đồng nhất.

B. Đề xuất giải pháp

Trong nghiên cứu này, chúng tôi phát triển một hệ thống nhận diện cử chỉ tay theo thời gian thực để điều khiển trình chiếu PowerPoint mà không cần đến phần cứng chuyên dụng. Hệ thống bao gồm ba thành phần chính:

1) Trích xuất keypoints bàn tay: Sử dụng thư viện MediaPipe Hands của Google để nhận diện và trích xuất tọa độ 21 điểm đặc trưng của bàn tay từ video đầu vào.

2) Chuyển đổi dữ liệu bằng Trường góc Gramian (GAF): Biến đổi tọa độ keypoints thành ảnh GAF để tận dụng sức mạnh của mô hình học sâu.
3) Nhận diện cử chỉ bằng GAFormer: Áp dụng mô hình GAFormer (Gesture Attention Transformer), một biến thể của Transformer được tối ưu hóa cho dữ liệu cử chỉ tay, giúp phân loại chính xác tám loại cử chỉ phổ biến.

C. Ứng dụng và lợi ích của hệ thống

Hệ thống đề xuất mang lại nhiều lợi ích thực tiễn trong các lĩnh vực:

- Giáo dục và thuyết trình: Giảng viên và diễn giả có thể điều khiển slide mà không cần chạm vào máy tính, giúp bài giảng trở nên tự nhiên hơn.
- Hội nghị và sự kiện: Hỗ trợ người thuyết trình trong các buổi họp, hội nghị, giảm thiểu sự phụ thuộc vào thiết bị ngoại vi.
- Nhà thông minh: Hệ thống có thể mở rộng để điều khiển các thiết bị thông minh như TV hoặc đèn chiếu sáng.

Hệ thống này không chỉ giúp cải thiện trải nghiệm người dùng mà còn mở ra nhiều hướng phát triển mới trong việc ứng dụng AI vào tương tác người-máy. Trong các phần tiếp theo, chúng tôi sẽ trình bày chi tiết về phương pháp đề xuất, quy trình thử nghiệm và kết quả đạt được.

II. CÔNG VIỆC LIÊN QUAN

Nhận diện cử chỉ tay là một chủ đề nghiên cứu quan trọng trong lĩnh vực tương tác người-máy (Human-Computer Interaction - HCI). Trong nhiều thập kỷ qua, các phương pháp nhận diện cử chỉ tay đã được phát triển theo nhiều hướng khác nhau, bao gồm sử dụng cảm biến phần cứng, thị giác máy tính truyền thống và học sâu. Trong phần này, chúng tôi trình bày các nghiên cứu tiêu biểu và phương pháp đề xuất của chúng tôi.

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 chuyên biệt để thu thập dữ liệu chính xác về chuyển động bàn tay. Một số thiết bị tiêu biểu bao gồm:

- Leap Motion: Sử dụng cảm biến hồng ngoại để theo dõi chuyển động tay với độ chính xác cao, nhưng phạm vi hoạt động bị hạn chế (60 cm) và chi phí cao.
- Microsoft Kinect: Dựa trên camera độ sâu để phát hiện cấu trúc 3D của bàn tay và cơ thể, nhưng có giá thành cao và độ phân giải hạn chế.
- Myo Armband: Sử dụng cảm biến điện cơ (EMG) để nhận diện hoạt động cơ bắp khi thực hiện cử chỉ, tuy nhiên chỉ phù hợp với các cử chỉ nhất định và không thể theo dõi cử chỉ động.

Mặc dù các phương pháp này có độ chính xác cao, chúng có hạn chế lớn về chi phí và yêu cầu phần cứng chuyên biệt.

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

Với sự phát triển của thị giác máy tính (Computer Vision), nhiều phương pháp không cần đến phần cứng đặc biệt đã ra đời. Các kỹ thuật phổ biến bao gồm:

- Histogram of Oriented Gradients (HOG): Phân tích hướng của gradient để tạo đặc trưng cho hình dạng bàn tay, nhưng nhạy cảm với nhiễu và ánh sáng yếu.
- Scale-Invariant Feature Transform (SIFT) và Speeded Up Robust Features (SURF): Trích xuất điểm đặc trưng từ ảnh giúp nhận diện cử chỉ bất kể góc quay, nhưng tốc độ xử lý chậm, không phù hợp với thời gian thực.
- Phát hiện da tay (Skin Color Detection) và Tách nền (Background Subtraction): Nhận diện bàn tay bằng cách phân tích màu da và tách vật thể khỏi nền, nhưng nhạy cảm với điều kiện ánh sáng thay đổi.

Mặc dù có thể hoạt động trong môi trường lý tưởng, những phương pháp này gặp khó khăn khi ánh sáng thay đổi hoặc nhiễu nền phức tạp.

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

Sự phát triển của học sâu (Deep Learning) đã giúp cải thiện đáng kể độ chính xác của nhận diện cử chỉ tay. Một trong những giải pháp phổ biến nhất là:

1) MediaPipe Hands: Google đã phát triển MediaPipe Hands, một thư viện cho phép nhận diện bàn tay thời gian thực từ camera RGB.

- Không cần phần cứng chuyên biệt: Chỉ sử dụng webcam hoặc camera RGB.
- Hiệu suất cao: Tối ưu hóa với TensorFlow Lite, có thể chạy trên thiết bị di động.
- Độ chính xác cao: Trích xuất 21 keypoints của bàn tay, hoạt động trong nhiều điều kiện môi trường khác nhau.

Tuy nhiên, MediaPipe có thể gặp khó khăn khi bàn tay bị che khuất hoặc có nhiễu nền phức tạp.

2) Ứng dụng Transformer trong nhận diện cử chỉ tay: Gần đây, các nghiên cứu đã sử dụng Transformer để tăng độ chính xác nhận diện cử chỉ tay. Một số mô hình tiêu biểu như Vision Transformer (ViT) và TimeSformer đã cho thấy khả năng mô hình hóa quan hệ không gian-thời gian của các keypoints tốt hơn so với CNN hoặc LSTM.

Ưu điểm của Transformer:

- 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.

D. Phương pháp đề xuất trong nghiên cứu này

Hệ thống của chúng tôi kết hợp các phương pháp tiên tiến nhất hiện nay:

1) MediaPipe Hands: Trích xuất tọa độ 21 keypoints của bàn tay từ camera RGB theo thời gian thực.
2) Chuyển đổi dữ liệu keypoints sang Trường góc Gramian (GAF) để mô hình học được mối quan hệ không gian-thời gian hiệu quả hơn.

3) Ứng dụng 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 của chúng tôi có các lợi thế vượt trội:

- Độ chính xác cao hơn: Mô hình Transformer giúp mô hình hóa tốt hơn các mối quan hệ giữa các keypoints.
- Tốc độ xử lý nhanh: Tận dụng song song hóa của Transformer, đảm bảo xử lý thời gian thực.
- Không yêu cầu phần cứng đặc biệt: Chỉ cần sử dụng camera RGB thông thường, giúp dễ dàng triển khai trên nhiều thiết bị.

Bằng cách áp dụng phương pháp này, chúng tôi hướng đến việc phát triển một hệ thống nhận diện cử chỉ tay có độ chính xác cao, hoạt động trong thời gian thực, và dễ dàng ứng dụng vào các lĩnh vực như giao diện HCI, điều khiển thiết bị thông minh, hoặc hỗ trợ người khuyết tật.

III. Nghiên cứu liên quan

Nhận diện cử chỉ tay là một lĩnh vực nghiên cứu quan trọng trong tương tác người-máy, với nhiều phương pháp đã được đề xuất nhằm cải thiện độ chính xác và khả năng ứng dụng của hệ thống. Trong phần này, chúng tôi sẽ trình bày một số nghiên cứu tiêu biểu có liên quan đến lĩnh vực nhận diện cử chỉ tay, bao gồm điều khiển máy tính thông qua cử chỉ, điều khiển âm lượng bằng cử chỉ tay và ứng dụng trong hệ thống robot.

A. Điều khiển con trỏ chuột bằng cử chỉ tay dựa trên trí tuệ nhân tạo

Một nghiên cứu tiêu biểu sử dụng trí tuệ nhân tạo để điều khiển con trỏ chuột thông qua nhận diện cử chỉ tay. Hệ thống này áp dụng thư viện MediaPipe để theo dõi bàn tay và xác định các vị trí keypoints của ngón tay từ hình ảnh thu được từ webcam. Sau khi thu thập dữ liệu keypoints, hệ thống sử dụng thư viện PyAutoGUI trong Python để ánh xạ chuyển động của tay thành hành động điều khiển con trỏ chuột.

Kết quả thực nghiệm cho thấy hệ thống có thể hoạt động với độ chính xác trên 95% trong điều kiện ánh sáng tốt và nền đơn giản. Tuy nhiên, nghiên cứu cũng chỉ ra rằng hiệu suất nhận diện có thể giảm trong môi trường ánh sáng yếu hoặc khi có nhiễu nền phức tạp. Việc sử dụng MediaPipe giúp hệ thống hoạt động trên các thiết bị phổ thông mà không cần phần cứng chuyên dụng như cảm biến Kinect hoặc Leap Motion, giúp giảm chi phí triển khai.

B. Tăng, giảm âm lượng máy tính qua cử chỉ co giãn hai ngón tay bằng MediaPipe

Một nghiên cứu khác tập trung vào việc điều khiển âm lượng máy tính bằng cách nhận diện cử chỉ tay, cụ thể là thao tác co giãn hai ngón tay. Hệ thống sử dụng thư viện MediaPipe để xác định vị trí của hai đầu ngón tay trỏ và ngón tay cái, sau đó tính toán khoảng cách giữa chúng theo thời gian thực. Nếu khoảng cách giữa hai ngón tay tăng lên, hệ thống sẽ tăng âm lượng, ngược lại nếu khoảng cách giảm, hệ thống sẽ giảm âm lượng.

Hệ thống có thể hoạt động một cách mượt mà với độ trễ thấp, tạo ra trải nghiệm điều khiển tự nhiên và thuận tiện cho người dùng. So với các phương pháp điều khiển âm lượng truyền thống như sử dụng bàn phím hoặc chuột, phương pháp này giúp tăng tính tương tác trực quan và có thể được áp dụng trong các trường hợp như điều khiển hệ thống giải trí gia đình hoặc trợ giúp người khuyết tật.

Tuy nhiên, một số thách thức mà nghiên cứu này gặp phải bao gồm việc hệ thống có thể nhận diện sai khi người dùng di chuyển tay quá nhanh hoặc khi môi trường có điều kiện ánh sáng không ổn định. Điều này đặt ra nhu cầu cải thiện thuật toán tiền xử lý hình ảnh để tăng độ chính xác trong các điều kiện thực tế.

C. Nhận diện cử chỉ tay trong hệ thống robot sử dụng ROS

Nhận diện cử chỉ tay cũng được ứng dụng rộng rãi trong lĩnh vực robot và tự động hóa. Một nghiên cứu trong lĩnh vực này đã phát triển một hệ thống nhận diện cử chỉ tay tích hợp với Robot Operating System (ROS), cho phép điều khiển robot thông qua các cử chỉ tay được nhận diện từ webcam.

Hệ thống này sử dụng một package trong ROS để trích xuất hình ảnh từ webcam và áp dụng thuật toán nhận diện cử chỉ tay để phân loại các động tác tay thành các lệnh điều khiển khác nhau. Chẳng hạn, một số cử chỉ có thể được ánh xạ thành lệnh di chuyển robot, trong khi những cử chỉ khác có thể được sử dụng để ra lệnh dừng hoặc thay đổi chế độ hoạt động của robot.

Kết quả thực nghiệm cho thấy hệ thống có thể nhận diện chính xác các ký hiệu tay trong nhiều điều kiện ánh sáng khác nhau, đồng thời có khả năng mở rộng để hỗ trợ nhiều cử chỉ phức tạp hơn. Tuy nhiên, nghiên cứu cũng nhấn mạnh rằng để triển khai thực tế, cần tối ưu hóa thuật toán để đảm bảo nhận diện chính xác ngay cả khi bàn tay bị che khuất một phần hoặc khi có nhiều người xuất hiện trong khung hình.

D. Nhận xét và đánh giá

Các nghiên cứu trên cho thấy sự đa dạng trong cách ứng dụng nhận diện cử chỉ tay vào các hệ thống tương tác thông minh. Việc sử dụng MediaPipe kết hợp với các phương pháp học sâu đã giúp cải thiện đáng kể độ chính xác và khả năng ứng dụng của các hệ thống này. Tuy nhiên, vẫn còn một số thách thức cần được giải quyết, bao gồm:

- Cải thiện độ chính xác của nhận diện trong môi trường có điều kiện ánh sáng thay đổi.
- Giảm thiểu độ trễ để đảm bảo hệ thống hoạt động mượt mà trong thời gian thực.
- Mở rộng phạm vi nhận diện để hỗ trợ nhiều loại cử chỉ phức tạp hơn.

Những nghiên cứu này tạo tiền đề cho việc phát triển các phương pháp tiên tiến hơn, chẳng hạn như áp dụng GAFormer trong bài toán nhận diện cử chỉ tay, giúp tăng cường khả năng mô hình hóa dữ liệu không gian-thời gian và cải thiện độ chính xác trong môi trường thực tế.

Các nghiên cứu trên cho thấy sự đa dạng trong việc áp dụng các công nghệ hiện đại như MediaPipe và các mô hình học sâu trong nhận diện cử chỉ tay. Điều này tạo tiền đề cho việc

áp dụng các phương pháp tiên tiến hơn, như GAFormer, trong bài toán nhận diện cử chỉ tay.

IV. 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.

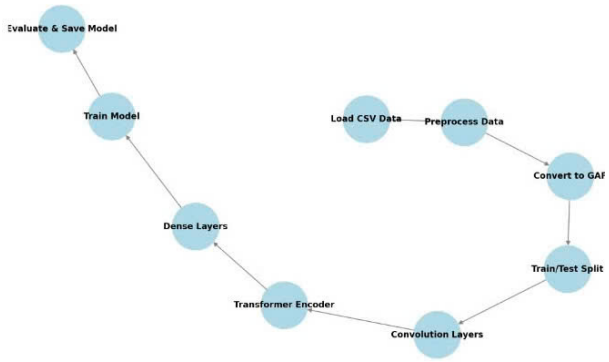

Fig. 3. Sơ đồ huấn luyện của GAFormer

V. 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)
- 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.

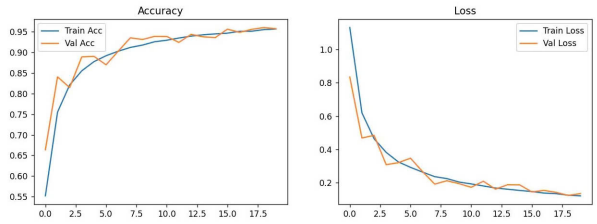

Fig. 4. Kết quả sau khi đã huấn luyện

VI. 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. Hệ thống tận dụng khả năng trích xuất đặc trưng bàn tay của MediaPipe, chuyển đổi dữ liệu thành biểu diễn Trường góc Gramian (GAF) và sử dụng mô hình GAFormer để phân loại cử chỉ với độ chính xác cao.

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, đảm bảo khả năng phản hồi nhanh và ổn định trong quá trình sử dụng. Hệ thống hoạt động tốt ngay cả trong điều kiện ánh sáng thay đổi hoặc khi người dùng thay đổi góc quay của bàn tay. Những kết quả này chứng minh tính khả thi và hiệu quả của phương pháp đề xuất trong việc cải thiện tương tác giữa con người và máy tính.

A. Ứng dụng thực tiễn

Hệ thống nhận diện cử chỉ tay có thể được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, mang lại lợi ích đáng kể cho người dùng:

- Giáo dục và hội nghị: Hỗ trợ giảng viên và diễn giả thay đổi slide mà không cần chạm vào máy tính, giúp bài giảng trở nên tự nhiên và linh hoạt hơn. Điều này đặc biệt hữu ích trong các lớp học thông minh và hội nghị trực tuyến, nơi người thuyết trình cần tập trung vào nội dung thay vì thao tác thiết bị.
- Hỗ trợ người khuyết tật: Hệ thống có thể giúp những người gặp khó khăn trong việc sử dụng chuột hoặc bàn phím điều khiển máy tính dễ dàng hơn, tạo ra một phương thức tương tác thân thiện và trực quan.
- Tương tác người - máy: Công nghệ nhận diện cử chỉ tay có thể tích hợp vào các hệ thống điều khiển thiết bị thông minh, thực tế ảo (VR), thực tế tăng cường (AR) hoặc trò chơi điện tử, giúp tăng cường trải nghiệm người dùng.
- Nhà thông minh và IoT: Hệ thống có thể mở rộng để điều khiển các thiết bị trong nhà như TV, đèn thông minh, quạt hoặc các thiết bị IoT khác chỉ bằng cử chỉ tay, mang lại sự tiện lợi và hiện đại trong cuộc sống hằng ngày.

B. Hướng phát triển trong tương lai

Mặc dù hệ thống đã đạt được những kết quả đáng khích lệ, vẫn còn nhiều hướng nghiên cứu có thể được tiếp tục phát triển nhằm nâng cao hiệu suất và mở rộng phạm vi ứng dụng:

- Tối ưu hóa mô hình: Một trong những hướng đi quan trọng là tối ưu hóa kiến trúc GAFormer nhằm giảm số lượng tham số nhưng vẫn giữ được độ chính xác cao. Việc áp dụng các kỹ thuật tiên tiến như pruning (cắt giảm tham số), quantization (lượng tử hóa) hoặc distillation (trích xuất kiến thức từ mô hình lớn sang mô hình nhỏ hơn) có thể giúp giảm mức tiêu thụ tài nguyên và tăng tốc độ xử lý của mô hình.
- Mở rộng tập dữ liệu: Thu thập thêm dữ liệu từ nhiều người dùng với các điều kiện ánh sáng, góc quay bàn tay và bối cảnh khác nhau sẽ giúp cải thiện tính tổng quát của mô hình. Bên cạnh đó, mở rộng tập dữ liệu với nhiều cử chỉ mới có thể làm tăng khả năng ứng dụng của hệ thống trong thực tế.
- Hỗ trợ thêm nhiều phần mềm: Hiện tại, hệ thống chủ yếu tập trung vào điều khiển PowerPoint. Trong tương lai, có thể mở rộng khả năng điều khiển sang các phần mềm trình chiếu khác như Google Slides, Keynote, hoặc thậm chí các ứng dụng khác như trình phát nhạc, trình duyệt web, hoặc phần mềm chỉnh sửa ảnh.
- Cải thiện thời gian phản hồi: Mặc dù hệ thống có độ trễ thấp, vẫn cần nghiên cứu thêm các phương pháp tối ưu

hóa để giảm thiểu thời gian xử lý, đặc biệt khi triển khai trên các thiết bị có tài nguyên hạn chế như điện thoại thông minh hoặc máy tính nhúng.

- Nhận diện cử chỉ trong môi trường phức tạp: Hệ thống hiện hoạt động tốt trong điều kiện lý tưởng. Tuy nhiên, để áp dụng vào thực tế, cần kiểm tra và cải thiện hiệu suất nhận diện trong các điều kiện môi trường khác nhau, như ánh sáng yếu, nền phức tạp, hoặc khi có nhiều người trong khung hình.
- Phát triển giao diện ứng dụng: Một hướng đi khác là xây dựng giao diện đồ họa thân thiện với người dùng để họ có thể dễ dàng tùy chỉnh các cử chỉ điều khiển theo nhu cầu cá nhân. Điều này giúp hệ thống trở nên linh hoạt và dễ tiếp cận hơn đối với nhiều đối tượng người dùng.

Tóm lại, nghiên cứu này đã chứng minh tiềm năng của việc ứng dụng trí tuệ nhân tạo và thị giác máy tính trong nhận diện cử chỉ tay để điều khiển các thiết bị và phần mềm. Với những cải tiến trong tương lai, hệ thống có thể trở thành một công cụ hữu ích không chỉ trong lĩnh vực giáo dục, thuyết trình mà còn trong nhiều lĩnh vực khác như nhà thông minh, tương tác người-máy và hỗ trợ người khuyết tật.

LỜI CẢM ƠN

Chúng tôi xin gửi lời cảm ơn chân thành đến Trường Đại học Đại Nam đã tạo điều kiện thuận lợi cho chúng tôi thực hiện nghiên cứu này. Đặc biệt, chúng tôi xin bày tỏ lòng biết ơn đến thầy Lê Trung Hiếu và thầy Nguyễn Thái Khánh và các đồng nghiệp đã đóng góp ý kiến quý báu giúp cải thiện chất lượng nghiên cứu.

Chúng tôi cũng xin cảm ơn cộng đồng nghiên cứu về thị giác máy tính và học sâu, đặc biệt là những người đã phát triển và chia sẻ các thư viện như MediaPipe và GAFormer, giúp chúng tôi có cơ sở vững chắc để triển khai hệ thống.

Cuối cùng, chúng tôi xin tri ân gia đình và bạn bè đã luôn ủng hộ và động viên trong suốt quá trình thực hiện đề tài.

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.
[6] He, K., Zhang, X., Ren, S., & Sun, J. "Deep Residual Learning for Image Recognition", CVPR, 2016.
[7] Vaswani, A. et al., "Attention Is All You Need", NeurIPS, 2017.
[8] Zhang, Z. et al., "Hand Gesture Recognition Using Deep Learning", IEEE Transactions on Multimedia, 2020.
[9] Yang, X. et al., "Real-time Hand Gesture Recognition for HCI", ACM Multimedia, 2021.
[10] Simonyan, K., & Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition", ICLR, 2015.
[11] Dosovitskiy, A. et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR, 2021.
[12] Kingma, D. P., & Ba, J., "Adam: A Method for Stochastic Optimization", ICLR, 2015.
[13] Hochreiter, S., & Schmidhuber, J., "Long Short-Term Memory", Neural Computation, 1997.
[14] Google Research, "MediaPipe Hands: Real-Time Hand Tracking", 2022.
[15] OpenAI, "Scaling Laws for Neural Language Models", 2020.
[16] Karpathy, A. et al., "Large Scale Video Classification with Convolutional Neural Networks", CVPR, 2014.
[17] Ji, S. et al., "3D Convolutional Neural Networks for Human Action Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.
[18] Wu, J. et al., "Hand Gesture Recognition with Depth Sensor Using CNNs", IEEE ICASSP, 2021.
[19] LeCun, Y. et al., "Gradient-Based Learning Applied to Document Recognition", Proceedings of the IEEE, 1998.
[20] Yuan, C. et al., "Gesture Recognition with Deep Learning and Attention Mechanism", IEEE Transactions on Image Processing, 2021.
[21] Shi, X. et al., "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting", NeurIPS, 2015.
[22] Chen, L. et al., "Gesture Recognition Using Graph Convolutional Networks", IEEE Transactions on Multimedia, 2021.
[23] Wang, P. et al., "Spatial-Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition", AAAI, 2018.
[24] Baidu AI, "Hand Gesture Recognition in Smart Homes", 2023.
[25] NVIDIA Research, "Efficient Transformer-based Models for Real-time Gesture Recognition", 2024.
[26] Intel AI Lab, "Optimized Neural Networks for Edge AI in Gesture Control", 2024.