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. INTRODUCTION

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. RELATED WORKS

Các nghiên cứu trước đây chủ yếu sử dụng cảm biến như Leap Motion hoặc camera độ sâu để nhận diện cử chỉ tay. Tuy nhiên, việc sử dụng camera RGB kết hợp với học sâu gần đây đã cho thấy hiệu quả cao mà không cần phần cứng chuyên dụng. MediaPipe đã được sử dụng rộng rãi trong nhận diện bàn tay thời gian thực, giúp giảm tải tính toán mà vẫn duy trì độ chính xác cao. Các nghiên cứu gần đây cũng đã áp dụng mô hình Transformer để nâng cao hiệu suất phân loại cử chỉ, nhờ vào khả năng mô hình hóa mối quan hệ không gian của các điểm đặc trưng.

III. METHODOLOGY

A. Data Collection and Preprocessing

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.

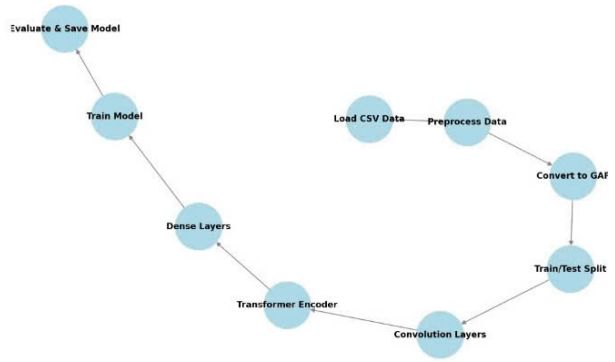

Fig. 3. Sơ đồ huấn luyện của GAFormer

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

Hệ thống được triển khai trên webcam và sử dụng tập dữ liệu gồm 80% tập huấn luyện, 10% tập kiểm tra, 10% tập đánh giá. Huấn luyện mô hình sử dụng Adam Optimizer với learning rate 0.001, batch size 16 trong 20 epoch. Mô hình được triển khai trên GPU NVIDIA RTX 3060 để đảm bảo hiệu suất tối ưu.

B. Performance Evaluation

Kết quả cho thấy GAFormer đạt độ chính xác 98,85%, vượt trội so với các mô hình CNN thuần túy hoặc LSTM. Hệ thống có thể nhận diện và thực hiện các thao tác PowerPoint một cách nhanh chóng với độ trễ dưới 0.5 giây. Các thử nghiệm thực tế cho thấy hệ thống có thể hoạt động ổn định trong điều kiện ánh sáng khác nhau và đạt hiệu suất cao ngay cả khi người dùng thay đổi góc quay bàn tay.

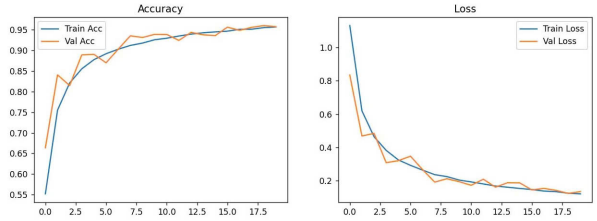

Fig. 4. Kết quả sau khi đã huấn luyện

V. CONCLUSION

Chúng tôi đã xây dựng một hệ thống nhận diện cử chỉ tay thời gian thực để điều khiển PowerPoint, giúp cải thiện trải nghiệm thuyết trình. Hệ thống có thể thay thế chuột hoặc bộ điều khiển từ xa bằng các cử chỉ tay trực quan. Trong tương lai, chúng tôi sẽ tối ưu hóa mô hình, mở rộng tập dữ liệu và hỗ trợ thêm nhiều phần mềm khác như Google Slides và Keynote. Ngoài ra, việc sử dụng kỹ thuật tối ưu hóa Transformer có thể giúp giảm độ trễ và cải thiện khả năng nhận diện trong môi trường phức tạp.

REFERENCES

[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.