XÂY DỰNG HỆ THỐNG PHIÊN DỊCH NGÔN NGỮ KÝ HIỆU

SIGNAL LANGUAGE TRANSLATION SYSTEM

SVTH: Nguyễn Văn Mạnh, Nguyễn Văn Hoàng Phúc, Trần Thanh Nguyên

Lớp 20T1, Khoa Công nghệ thông tin, Trường Đại học Bách Khoa - Đại học Đà Nẵng; Email: nguyenvanmanh2001it1@gmail.com, phucknguyen123456789@gmail.com, trannguyen642@gmail.com

GVHD: TS. Ninh Khánh Duy

Khoa Công nghệ thông tin, Trường Đại học Bách Khoa - Đại học Đà Nẵng; Email: nkduy@dut.udn.vn

Tóm tắt - Việc giao tiếp đối với người khiếm thính hay khuyết tật ngôn ngữ là 1 vấn đề khó khăn do số lượng người học phiên dịch ngôn ngữ ký hiệu rất ít. Vì vậy để giúp việc giao tiếp bằng ngôn ngữ ký hiệu trở nên dễ dàng hơn, nghiên cứu này đề xuất phương pháp để nhận diện ngôn ngữ ký hiệu chuyển chúng thành văn bản và thử nghiệm trong thời gian thực**.** Nhóm chúng tôi đã sử dụng MediaPipe kết hợp với các mô hình RNN để giải quyết các vấn đề về nhận dạng ngôn ngữ ký hiệu động. Các mô hình RNN như GRU, LSTM giải quyết vấn đề phụ thuộc khung trong chuyển động của dấu hiệu. Do thiếu bộ dữ liệu dựa trên video cho ngôn ngữ ký hiệu, Bộ dữ liệu DSL25 đã được tạo. DSL25-Bộ dữ liệu chứa 25 từ vựng đã được lặp lại 75 lần bởi 8 người cung cấp (gồm DSL10 và 15 từ vựng mới do chúng tôi tạo) các bước hướng dẫn để tạo một từ như vậy. Các thử nghiệm được thực hiện trên tập dữ liệu của chúng tôi (DSL25-Dataset) sử dụng các mô hình RNN để so sánh độ chính xác của nhận dạng ngôn ngữ ký hiệu động có hay không thực hiện giảm số đặc trưng và tiền xử lý, đồng thời có hay không thực hiện giảm số lớp của mô hình RNN trước đó. Các thí nghiệm cho thấy mô hình của chúng tôi có độ chính xác hơn 98%.

Từ khóa - ngôn ngữ kí hiệu động (DSL); MediaPipe; landmarks; GRU; LSTM;

Abstract - Communicating with the deaf or language-impaired is a challenging issue due to the limited number of sign language interpreters. To make communication through sign language easier, this study proposes a method to recognize sign language and convert it into text in real-time. Our team used MediaPipe combined with RNN models to solve problems related to dynamic sign language recognition. RNN models such as GRU and LSTM address the frame dependency issue in sign movements. Due to the lack of video-based datasets for sign language, the DSL25 dataset was created. The DSL25 dataset contains 25 vocabulary words repeated 75 times by 8 providers (including DSL10 and 15 newly created words by us) with step-by-step instructions to create such a word. Experiments were conducted on our DSL25-Dataset using RNN models to compare the accuracy of dynamic sign language recognition with or without reducing the number of features and pre-processing, as well as with or without reducing the number of layers in the previous RNN model. The experiments showed that our model achieved an accuracy of over 98%.

Key words - dynamic sign language (DSL); MediaPipe; landmarks; GRU; LSTM;

1. Đặt vấn đề

Theo thống kê của Tổ chức Y tế Thế giới (WHO), đến tháng 1 năm 2023 có đến hơn 70 triệu người trên thế giới đang phải đối mặt với khiếm thính. Ở Việt Nam, con số này là khoảng 2.5 triệu người Điếc và Khiếm thính. Mỗi năm, có thêm trung bình 5000 trẻ khiếm thính, và số liệu này được dự đoán sẽ tăng lên nhanh chóng, gấp 1.5 đến 2 lần vào năm 2050.

Một trong những vấn đề chính đối với người khiếm thính đó chính là việc giao tiếp với mọi người xung quanh. Vì họ không có khả năng nghe, nói, nên ngôn ngữ ký hiệu chính là công cụ để người ta có thể biểu đạt suy nghĩ, cảm xúc của mình. Để học được ngôn ngữ đặc biệt này, cần có phương pháp học, và có người chỉ dẫn các cử chỉ tay, ngôn ngữ cơ thể và biểu cảm khuôn mặt đúng cách. Theo một cách thông thường, người ta có thể biểu đạt từ ngữ bằng khẩu hình miệng. Cách này thường không khả quan và rất khó bởi vì có rất nhiều từ có khẩu hình giống nhau. Còn việc thuê hay có chuyên gia phiên dịch viên thì khá là tốn kém để một người bình thường có thể nói chuyện với một người khiếm thính. Ngôn ngữ ký hiệu cho người khiếm thính không chỉ khó, mà còn tốn rất nhiều thời gian để ghi nhớ kí hiệu và biểu đạt lại.

Chính vì vậy, chúng tôi nghĩ, bằng Công nghệ và Trí tuệ nhân tạo, tạo một hệ thống giúp người khiếm thính và bình thường giao tiếp dễ dàng hơn là cần thiết. Cần có một phương pháp nhận dạng cử chỉ hiệu quả có thể phát hiện các cử chỉ trong luồng video ngôn ngữ ký hiệu để triển khai hệ thống này.

Cử chỉ tay có thể được phân thành hai loại: tĩnh và động. Một cử chỉ tĩnh là một hình ảnh duy nhất đại diện cho một hình dạng và tư thế cụ thể của bàn tay. Cử chỉ động là một cử chỉ chuyển động, được thể hiện bằng một loạt hình ảnh. Chiến lược của chúng tôi được xây dựng dựa trên tính năng nhận dạng cử chỉ tay động [31].

Ngôn ngữ ký hiệu động (Dynamic Sign Language - DSL) bao gồm một chuỗi các hoạt động bao gồm các chuyển động nhanh với độ tương đồng cao. Kết quả là, nhận dạng ngôn ngữ ký hiệu động, phụ thuộc vào một chuỗi hành động, phải đối mặt với những thách thức để đối phó với sự đa dạng, phức tạp và một loạt các từ vựng trong cử chỉ tay [2]. Hơn nữa, nhận dạng ngôn ngữ ký hiệu động phải đối mặt với những thách thức như xác định vị trí của bàn tay, xác định hình dạng cộng với hướng và phát hiện chuyển động của các dấu hiệu [2,3]. Việc nhận dạng rất phức tạp do sự thay đổi về không gian và thời gian của các cử chỉ được thực hiện bởi những người khác nhau.

Có một số cách tiếp cận để giải quyết các vấn đề trong nhận dạng DSL. Hầu hết các phương pháp có thể được phân loại thành hai loại [2]. Đầu tiên, một thuật toán dựa

trên hình dạng bàn tay và quỹ đạo chuyển động cử chỉ của bàn tay. Thứ hai, một cách tiếp cận dựa trên chuỗi hình ảnh của từng ngôn ngữ ký hiệu.

So sánh với hai phương pháp được giới thiệu ở trên, phương pháp dựa trên trình tự video để nhận dạng ngôn ngữ ký hiệu động đi trước về hiệu suất và tính khả thi của nó. Theo đó, phương pháp trong công việc này dựa trên việc sử dụng khung MediaPipe kết hợp với đổi mới các mô hình mạng thần kinh tái phát (RNN) lấy cảm hứng từ [31]: đơn vị tái phát có cổng (GRU), bộ nhớ dài hạn ngắn hạn (LSTM) [4,5]. MediaPipe nhận dạng các mốc và trích xuất các điểm chính từ các đối tượng như bàn tay, cơ thể và khuôn mặt. Khung này hỗ trợ giải quyết các sự cố DSL phổ biến. Một mặt, MediaPipe có thể xác định hình dạng và vị trí của bàn tay và cơ thể. Nó cũng giải quyết các vấn đề về hướng lòng bàn tay và nét mặt, là các thông số phụ. Ngoài ra, các mô hình RNN được đề cập ở trên có thể giải quyết vấn đề chuyển động của kí hiệu.

Do thiếu bộ dữ liệu DSL dựa trên video, Bộ dữ liệu DSL25 đã được tạo trong bài báo này, cùng với các bước hướng dẫn để tạo bộ dữ liệu đó.

Đóng góp chính của bài báo này là các vấn đề về nhận dạng ngôn ngữ ký hiệu thuộc loại động, tức là phụ thuộc vào chuyển động, đã được giải quyết. Chúng tôi đã so sánh giữa các loại RNN khác nhau và lưu ý cách sử dụng từng loại. Ngoài ra, chúng tôi đã so sánh giữa bao gồm và loại trừ các điểm bên dưới thân, giữa có và không tiền xử lý và chuẩn hoá và nêu rõ những lợi ích và hạn chế của việc sử dụng chúng.

Tóm lại, những đóng góp chính của nghiên cứu này mà chúng tôi hướng tới là:

(1) Đề xuất các ý tưởng tiền xử lý và chuẩn hoá dữ liệu; cải thiện độ chính xác và thời gian thực thi của DSL trên tập dữ liệu lớn hướng đến chạy trên thời gian thực.
(2) Đã tạo bộ dữ liệu dựa trên video mới (DSL25-Dataset) bao gồm 25 từ vựng.
(3) Thử nghiệm để tìm mô hình phù hợp nhất cho việc nhận diện ngôn ngữ ký hiệu động (GRU, LSTM).

Bài nghiên cứu này được trình bày thành 6 phần, trong đó Phần 2 trình bày các nghiên cứu liên quan; phương pháp của nghiên cứu ở Phần 3; Phần 4 dataset và kết quả tương ứng; Phần 5 là thảo luận các vấn đề xoay quanh kết quả; và cuối cùng là phần kết luận của bài báo ở Phần 6.

2. Nghiên cứu liên quan

Trong phần này, công việc liên quan đến nhận dạng ngôn ngữ ký hiệu động sẽ được xem xét từ hai khía cạnh: phương pháp dựa trên hình dạng bàn tay và quỹ đạo chuyển động và phương pháp dựa trên trình tự video.

2.1. Phương pháp quỹ đạo chuyển động và hình dạng bàn tay

Quỹ đạo chuyển động và hình dạng bàn tay là những phương pháp thông thường đối với những thách thức của nhận dạng DSL. Phương pháp này xem xét các thuộc tính và đặc điểm của hình dạng bàn tay và quỹ đạo chuyển động của cử chỉ tay. Một số công trình liên quan chủ yếu dựa trên việc đánh giá các đặc điểm của hình thức bàn tay. Kim et al. [6] đã sử dụng mạng lưới thần kinh sâu để giải quyết thách thức nhận dạng chính tả ngón tay dựa trên các đặc điểm hình dạng bàn tay.

Phương pháp dựa trên hình dạng bàn tay có thể biểu thị ý nghĩa đơn giản của cử chỉ tay chẳng hạn như các ký tự chữ và số. Tuy nhiên, nó vẫn bị giới hạn ở các cử chỉ chuyển động phức tạp do chỉ xem xét dạng bàn tay mà không có chuyển động của bàn tay.

Mặt khác, một số công trình liên quan chỉ tập trung vào việc đánh giá quỹ đạo chuyển động của cử chỉ tay. Hệ thống được phát triển bởi Mohandes et al. [7] đã sử dụng bộ nhớ dài hạn LSTM để phân biệt các cử chỉ tay hoàn toàn dựa trên quỹ đạo chuyển động của tay.

Các công trình [8–12] đã phân loại cử chỉ tay bằng cách sử dụng thông tin về quỹ đạo chuyển động thu được từ các cảm biến bao gồm con quay hồi chuyển, Kinect, gia tốc kế, găng tay điện tử và camera độ sâu. Các kỹ thuật trên được giới hạn chỉ trong một vài cử chỉ tay đơn giản như vẫy và di chuyển bàn tay lên xuống.

Do đó, một số nghiên cứu liên quan được xây dựng dựa trên việc phân tích các khía cạnh hình dạng bàn tay, chuyển động của tay và quỹ đạo chuyển động của cử chỉ tay. Ví dụ, Ding và Martinez [13] đã đề xuất một phương pháp để nhận hình dạng 3D của mọi ngón tay quan trọng và sau đó nhấn chuyển động của bàn tay dưới dạng quỹ đạo 3D. Dogra et al. [14] đã phát triển một mô hình nhận dạng cử chỉ tay đa giác quan. Vị trí ngón tay và lòng bàn tay được ghi lại từ hai góc nhìn khác nhau bằng cảm biến Leap Motion và Kinect. Những hệ thống như vậy dựa trên các thiết bị cảm biến có nhược điểm về sự thoải mái của người dùng.

Trong ngôn ngữ ký hiệu Ba Tư, Zadghorban và Nahvi [15] đã phát triển một phương pháp để phát hiện các ranh giới của từ. Phương pháp này chuyển đổi cử chỉ tay thành lời nói bằng cách sử dụng các tính năng chuyển động và hình dạng của bàn tay.

Nhận dạng DSL dựa trên các đặc điểm hình dạng bàn tay và quỹ đạo chuyển động có một số sai sót rõ ràng. Chiến lược này hoạt động tốt đối với một số khía cạnh của biển báo, chẳng hạn như ký tự chữ và số, nhưng sẽ trở nên khó khăn hơn khi mô hình được nhận dạng chứa nhiều biển báo và từ vựng. Tóm lại, việc sử dụng các tính năng dạng bàn tay và quỹ đạo chuyển động với DSL có những hạn chế.

2.2. Phương pháp trình tự video

Nhận dạng DSL dựa trên chuỗi video không yêu cầu cảm biến tay. Một số công việc liên quan tập trung vào nhận dạng ngôn ngữ ký hiệu dựa trên video xác định chuyển động của tay bằng thuật toán học sâu [16,17].

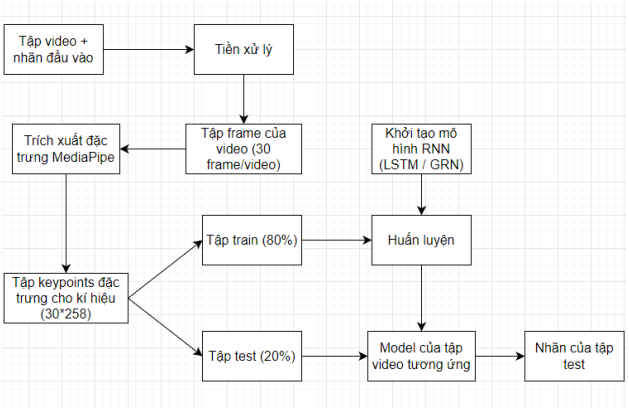
Trương và cộng sự. [18] đã tạo ra một mạng thần kinh tích chập tái phát (RCNN) dựa trên video để nhận dạng ký hiệu tay động. Kishore et al. [19] trình bày việc nhận dạng các cử chỉ của ngôn ngữ ký hiệu Ấn Độ (ISL) bằng cách sử dụng CNN. Phương pháp video ngôn ngữ ký hiệu ở chế độ selfie đã được sử dụng trong nghiên cứu này. Manikanta và cộng sự. [20] ngôn ngữ ký hiệu được nhận dạng bằng cách phân loại mờ các video cử chỉ động bằng cách sử dụng hỗn hợp các hình dạng và tính năng theo dõi.

Để đạt được sự công nhận DSL, nhiều phương pháp có thể được sử dụng như trích xuất các đặc điểm của cử chỉ tay với CNN [21,22], cách tiếp cận để học chuỗi video với RNN [23] và cách tiếp cận để học các đặc điểm chuỗi không gian thời gian bằng cách tích hợp CNN với RNN [24 ,25].

Khi so sánh với các phương pháp dựa trên hình dạng

bàn tay và quỹ đạo chuyển động, các phương pháp như vậy đạt được sự công nhận DSL đáng kể dựa trên chuỗi video.

3. Phương pháp nghiên cứu



Hình 3.1 Pipeline đề xuất đối với hệ thống nhận diện DSL

Phần này lý giải các phương pháp trích xuất đặc trưng các bộ phận trên cơ thể người thể hiện hành động và chi tiết triển khai mô hình Sequential cho nhận diện luồng hành động.

3.1. Tiền xử lí

Chúng tôi đề xuất sử dụng các kĩ thuật tiền xử lí sau:

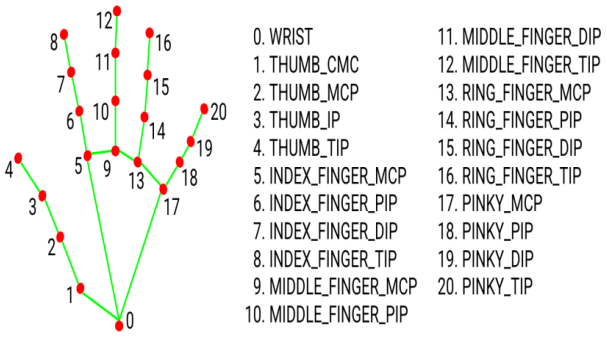
Phương pháp	Mục đích	Thực hiện trong quá trình
Điều chỉnh số frame của video về 30 frame/video	Đồng nhất số frame cho quá trình huấn luyện	Huấn luyện
Cắt bỏ đoạn đầu cuối video không thuộc ngôn ngữ kí hiệu	Tập trung vào các hành động của ngôn ngữ kí hiệu	Huấn luyện
Đưa khuôn mặt vào chính diện	Đưa về hệ toạ độ chung với gốc là khuôn mặt	Nhận dạng trên thời gian thực
Giảm số chiều ảnh hưởng của các keypoints không quan trọng	Tập trung vào các keypoints quan trọng, đặc trưng cho hành động	Huấn luyện
Điều chỉnh tỉ lệ cơ thể về chuẩn chung (Cố định độ rộng vai)	Đưa về hệ toạ độ chung với cùng kích thước	Nhận dạng trên thời gian thực

3.2. Trích xuất đặc trưng các bộ phận trên cơ thể người

Ngôn ngữ ký hiệu dựa trên việc sử dụng tay và ước tính tư thế, tuy nhiên, DSL đối mặt với nhiều khó khăn do sự chuyển động liên tục. Những khó khăn này bao gồm việc xác định vị trí của tay, hình dạng và hướng đi. MediaPipe được sử dụng như một giải pháp cho những vấn đề này. Nhóm nghiên cứu đã sử dụng mediapipe để bắt bộ khung trên cơ thể và liên kết chúng thành các keypoints. Các keypoints này là các điểm đặc trưng trên cơ thể, được sử dụng để phân tích và nhận diện các hành động. Hành động được định nghĩa như sự chuyển tiếp giữa các keypoints trong quá trình thực hiện hành động đó.

Đối với mỗi tay, MediaPipe trích xuất 21 điểm chính [26] như được thể hiện trong Hình 3.2. Do đó, số điểm chính được trích xuất của hai tay được tính toán như sau:

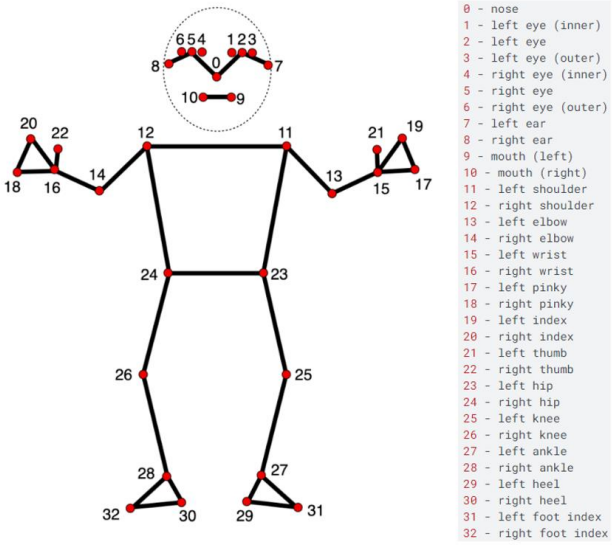
keypoints in hand × Three dimensions × No. of hands = (21 × 3 × 2) = 126 keypoints.



Hình 3.2 Thứ tự và nhãn cho các điểm chính có trong tay của MediaPipe [27].

Để ước tính tư thế, MediaPipe trích xuất 33 điểm chính [26] như được thể hiện trong Hình 3.3. Do đó, số điểm chính được trích xuất từ ước tính tư thế được tính toán như sau:

keypoints in pose × (Three dimensions + Visibility) = (33 × (3 + 1)) = 132 keypoints.



Hình 3.3 Thứ nhãn cho các điểm chính tự và có trong pose [28].

Các điểm dưới thân sẽ nằm trong khoảng 25-32. Do đó khi thực hiện giảm chiều của keypoints các điểm này sẽ bị loại bỏ:

keypoints in pose without legs × (Three dimensions + Visibility) = (25 × (3 + 1)) = 100 keypoints.

Bao gồm các điểm chân, tổng số điểm chính cho mỗi khung được tính như sau:

keypoints in hands + keypoints in pose = (126 + 132) =

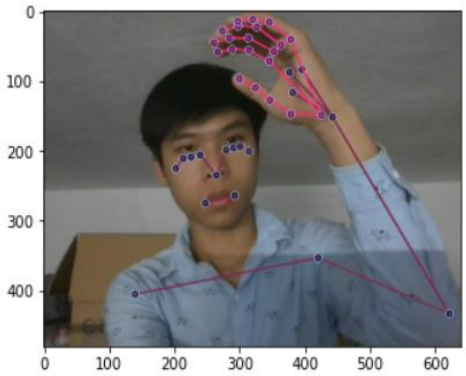
258 keypoints.

Nếu không bao gồm các điểm chân, tổng số điểm chính cho mỗi khung được tính như sau:

keypoints in hands + keypoints in pose without legs = (126 + 100) = 226 keypoints.

Thao tác này được lặp lại trong toàn bộ video để trích xuất các điểm chính cho mỗi khung trong tất cả các video của DSL25-Dataset.

Như vậy sau khi thực hiện trích xuất đặc trưng của cơ thể bằng mediapipe sẽ thu được tọa độ các điểm trên không gian 3 chiều. Các dữ liệu thu thập được mà nhóm nghiên cứu chọn để thực hiện đặc trưng cho hành động của cơ thể người là tay trái, tay phải, khung xương. Xét riêng về ngôn ngữ ký hiệu thì chỉ cần có đủ 3 thành phần trên thì có thể nhận diện được toàn bộ các từ vựng của ngôn ngữ này.



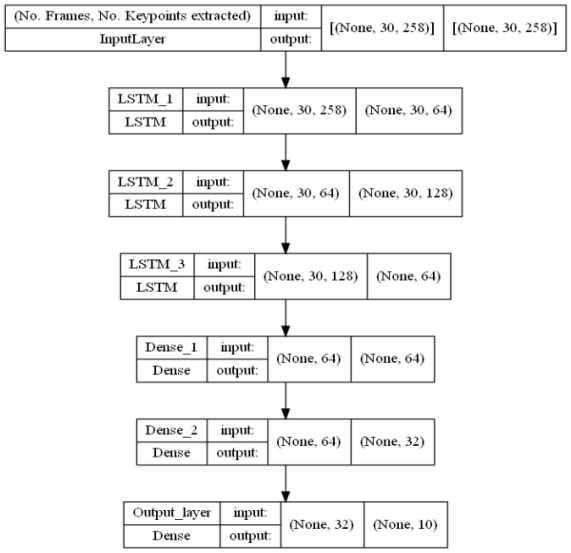
Hình 3.4 Trích xuất đặc trưng các bộ phận trên cơ thể (tay, khung xương)

3.3. Nhận diện hành động với mô hình RNN:

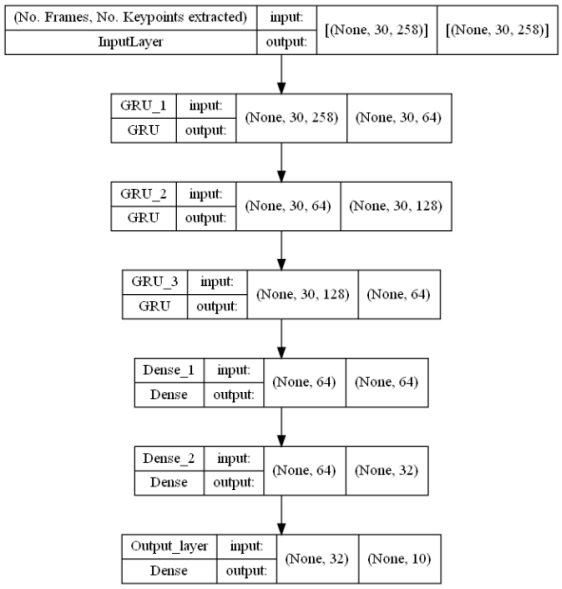
Mạng nơ-ron tuần hoàn (RNNs) là một loại mạng nơ-ron nhân tạo (ANN) sử dụng dữ liệu chuỗi thời gian và tuần tự. RNNs được gọi là tuần hoàn vì chúng thực hiện cùng một chức năng cho mỗi phần tử trong chuỗi, với sự phụ thuộc tính toán của các trạng thái trước đó và có đặc điểm chính là mạng có các kết nối phản hồi [29]. Công việc của chúng tôi thay đổi độ phức tạp hai mô hình liên quan đến RNN: GRU và LSTM [0].

GRU tương tự như LSTM với một cổng quên; tuy nhiên, nó có độ phức tạp thấp hơn và ít tham số hơn. LSTM được tạo ra để giải quyết vấn đề gradient biến mất có thể xảy ra khi cố gắng huấn luyện RNN truyền thống.

Đầu ra của các mô hình này dựa trên chuỗi đầu vào, cải thiện khả năng phát hiện chuyển động của DSL. Cấu trúc của các mô hình được thể hiện trong Hình 3.5-3.6. Ba lớp đầu tiên thuộc về mô hình RNN trong khi ba lớp cuối cùng là các fully-connected. Sau đó, các lớp được biên dịch bằng cách lựa chọn giá trị tối ưu nhất của tham số tối ưu hóa [5], giá trị của các tham số của lớp có thể được điều chỉnh bằng cách chọn bất kỳ giá trị nào từ Bảng 1 để chuẩn bị cho giai đoạn huấn luyện.



Hình 3.5 Cấu trúc mô hình LSTM



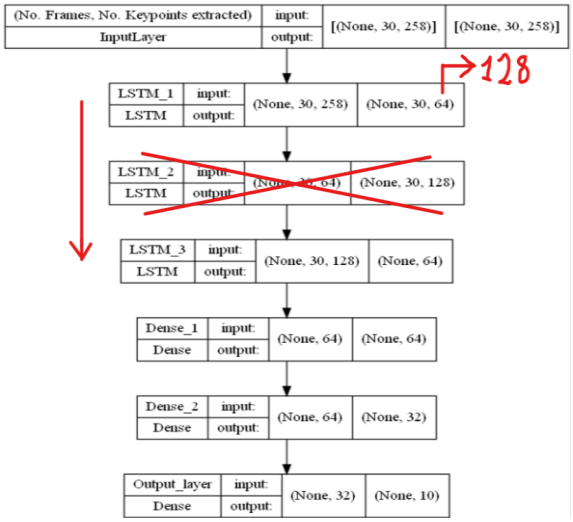
Hình 3.6 Cấu trúc mô hình GRU

Các đầu vào cho các mô hình là độ dài chuỗi và tổng số điểm chính. Độ dài chuỗi là số khung hình có trong mỗi đoạn. Tổng số điểm chính là 258 nếu bao gồm phần chân hoặc 226 nếu không bao gồm phần chân.

Bảng 1: Tham số huấn luyện

Parameters	**Value**
RNN Model	GRU, LSTM
Number of Nodes	Between (64, 256)
Activation	‘Relu’ or ‘Softmax’
Optimizer	‘Adam’

Tham khảo kết quả trên bài báo gốc của các mô hình được đề xuất trên có thể thấy được kết quả trên tập huấn luyện và kiểm thử của mô hình GRU là khả quan hơn so với mô hình LSTM. Do đó chúng tôi dự đoán độ phức tạp của model nhận diện ngôn ngữ ký hiệu có thể được giảm lại. Theo đó, chúng tôi thực hiện giảm số lớp RNN từ 3 xuống còn 2 lớp và thực hiện các thử nghiệm. Hình sau minh họa các thay đổi khi giảm số lớp RNN.



Hình 3.7 Minh họa các thay đổi số lớp RNN

4. Thực hiện phương pháp nghiên cứu

4.1. Dữ liệu

Các thử nghiệm để nhận dạng ngôn ngữ ký hiệu động được thực hiện bằng cách sử dụng tập dữ liệu lớn từ 2 tập khác nhau được nhóm đặt tên là DSL25-Dataset:

+ DSL10-Dataset: Tập dữ liệu nhỏ thu thập từ bài báo gốc, gồm ký 10 ký hiệu từ 5 người quay khác nhau (chứa 750 video)

+ Tập dữ liệu tự thu thập: gồm 15 ký hiệu từ 3 người quay khác nhau (chứa 1125 video)

Mỗi ký hiệu được quay 75 video, bộ dữ liệu DSL25-Dataset có tất cả 1875 video, được chia ngẫu nhiên để huấn luyện và thử nghiệm. Tất cả video trong DSL25-Dataset được ghi lại trong môi trường trong nhà với ánh sáng thông thường và máy ảnh di động trung bình (720p). Mỗi video được ghi ở 30 khung hình mỗi giây (FPS) và có cùng số khung hình và thời lượng (1s).

Danh sách các kí hiệu được sử dụng trong tập DSL25-Dataset:

Bảng 2 Danh sách các kí hiệu trong DSL25-Dataset

hello	howare	love	mask	no
please	sorry	thanks	wear	you
ảnh hưởng	cảm cúm	căm thù	công nhận	đau
đau bụng	đau lưng	hòa bình	khám	lắng nghe
lạnh	máu	mệt	mệt mỏi	mỏi cổ

Nguồn: Nhóm đã học các kí hiệu ngôn ngữ trên nguồn thông tin chính thống và phổ biến là website https://tudienngonngukyhieu.com/. Sau khi đã có được nguồn kí hiệu chính xác thì nhóm thực hiện tự quay video trên chính các thành viên của nhóm.

Hình 4.1 Hình ảnh 15 video của một ngôn ngữ kí hiệu

Hình 4.2 Các góc quay khác nhau của 3 người khác nhau

Chia tập dữ liệu: Nhóm thực hiện chia tập train, test theo tỉ lệ 80:20 và thực hiện huấn luyện trên tập train đồng thời thống kê các kết quả trên tập test.

4.2. Xử lý dữ liệu/ đặc trưng

a. Đồng nhất số khung frame của mỗi kí hiệu ngôn ngữ

Trong quá trình nhận diện và phân loại kí hiệu ngôn ngữ, việc đồng nhất số khung frame của mỗi kí hiệu là một yếu tố quan trọng. Điều này đảm bảo rằng mỗi kí hiệu được xử lý và phân tích theo cùng một đơn vị thời gian, tạo ra tính nhất quán và tin cậy cho quá trình phân loại.

Để đạt được sự đồng nhất này, chúng ta sử dụng một số khung frame cố định cho mỗi kí hiệu ngôn ngữ. Trong trường hợp này, mô hình của chúng ta sử dụng một số khung frame là 30. Điều này có nghĩa là mỗi kí hiệu ngôn ngữ được chia thành 30 khung frame tương ứng với một đoạn video nhất định.

b. Trích xuất ra keypoints tương ứng mỗi frame

Mỗi frame được xử lý để tìm ra 258 điểm keypoints tương ứng. Đối với xử lý giảm chiều dữ liệu, thực hiện loại bỏ các điểm phần chân số lượng keypoints là 226. Các điểm keypoints này bao gồm các điểm tay và pose, cung cấp thông tin quan trọng về tư thế và cử chỉ của người trong video.

Việc trích xuất keypoints từ các frame giúp chúng ta có một biểu diễn số học của hành động, tạo ra một tập dữ liệu số liệu để huấn luyện và phân loại các hành động khác nhau. Các điểm keypoints chính xác và chi tiết cung cấp thông tin quan trọng về vị trí và động tác của tay và cơ thể, giúp chúng ta hiểu và phân tích hành động một cách toàn diện.

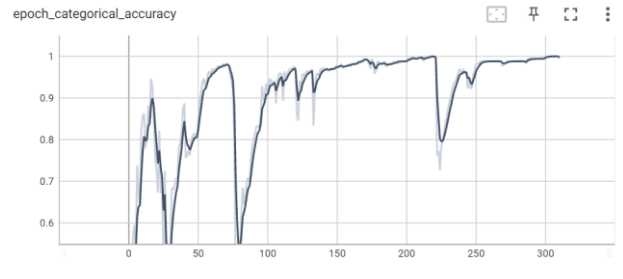
c. Đánh nhãn:

Mỗi đoạn video sau các bước trên sẽ thu được 30 frame tương ứng với 30 danh sách keypoints. Các keypoints này thay đổi tuần tự theo thời gian của video và đặc trưng cho video đó. Sau đó mỗi video sẽ được gán nhãn tương ứng với kí hiệu ngôn ngữ được thực hiện trong video.

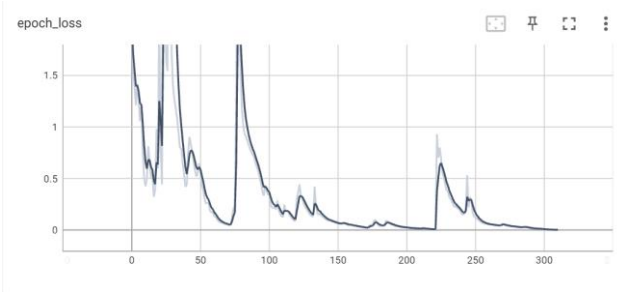
4.3. Kết quả thực nghiệm

4.3.1. Kết quả huấn luyện

Hầu hết các thử nghiệm trên tập train của các mô hình đều cho thấy hình dạng loss và accuracy là như nhau, các đồ thị sau minh hoạ quá trình huấn luyện.



Hình 4.3 Đồ thị accuracy khi huấn luyện trên tập train



Hình 4.4 Đồ thị loss function khi huấn luyện trên tập train

4.3.2. *Kết quả kiểm thử:*

a) Trên tập kiểm thử:

Bảng 3: Bảng thống kê kết quả chạy trên các mô hình

		Không tiền xử lý	Chuẩn hóa	Giảm chiều
GRU	3 lớp	98.67%	95.20%	98.13%
	2 lớp	98.13%	97.07%	97.67%
LSTM	3 lớp	96.53%	93.07%	96.53%
	2 lớp	98.40%	96.27%	97.87%

Bảng 4: Bảng thống kê thời gian dự đoán trên tập kiểm thử (375 video)

GRU		LSTM	
3 lớp	2 lớp	3 lớp	2 lớp
0.2124	0.1838	0.2914	0.2712

b) Kết quả thực nghiệm trên thời gian thực:

Hình 4.5 Kết quả thực nghiệm trên thời gian thực

Hiệu quả khi nhận diện trên thời gian thực là tương đối tốt, tuy nhiên vì xử lý song song việc trích xuất đặc trưng và nhận diện hành động nên hệ thống sẽ dễ bị lag và gây nhầm lẫn.

5. Bàn luận

Kết quả đạt được tốt nhất là 98.67% đối với mô hình GRU (kiến trúc gồm 3 lớp GRU) và không thực hiện tiền xử lý dữ liệu.

So sánh việc thực hiện giảm số lượng lớp của kiến trúc LSTM và kiến trúc GRU:

+ Đối với GRU thì hiệu suất giảm nhẹ từ 98.67% xuống còn 98.13%.

+ Đối với LSTM thì hiệu suất lại tăng mạnh từ 96.53% lên 98.40%

Việc giảm số lớp đi chính là việc giảm độ phức tạp của kiến trúc được sử dụng. Với kết quả thu được, LSTM có kiến trúc phức tạp hơn GRU khi giảm đi số lớp thì hiệu suất cũng tăng mạnh. Có thể trong bài báo gốc dự đoán sai về độ phức tạp của ngôn ngữ ký hiệu. Nhiều kết quả sử dụng SVM để nhận diện ngôn ngữ ký hiệu đạt hiệu suất rất tốt [30] là mô hình có độ phức tạp thấp hơn so với nhận diện bằng luồng hành động. Vì vậy, trong các nghiên cứu tiếp theo có thể thực hiện đánh giá về độ phức tạp cần thiết để nhận diện một ngôn ngữ kí hiệu.

So sánh việc chuẩn hóa dữ liệu và không chuẩn hóa:

+ Kết quả khá bất ngờ khi việc chuẩn hoá không đem lại hiệu quả trên tập dữ liệu.

+ Tuy nhiên, có thể giải thích vấn đề này do tập dữ liệu được quay theo chuẩn chung về kích thước, về vị trí người thực hiện trên khung hình nên các bước tiền xử lý là thừa đối với trường hợp này.

+ Mặt tích cực, mô hình dự đoán đã dự đoán được khá chính xác những thay đổi về vị trí và kích thước của người thực hiện.

Về giảm chiều keypoints: Kết quả giảm chiều giảm ít so với khi không thực hiện giảm chiều. Tuy nhiên lợi ích nó mang lại là giảm đi bộ nhớ, giảm thời gian huấn luyện.

Về thời gian đưa ra dự đoán: Nhìn chung các mô hình đều đưa ra dự đoán khá nhanh, chậm nhất là LSTM với kiến trúc 3 lớp có thời gian dự đoán 375 video kiểm thử là 0.2914 giây và nhanh nhất là trên GRU với kiến trúc 2 lớp có thời gian là 0.1838 giây. Kết quả nảy có thể xem xét chọn lựa mô hình áp dụng trong thời gian thực đảm bảo tốc độ nhận diện và độ chính xác.

6. Kết luận

Trong bài báo này, chúng tôi đã trình bày một mô hình nhận diện hành động sử dụng Deep Learning và phương pháp nhận diện keypoints trên cơ thể người. Mô hình được xây dựng dựa trên kiến trúc LSTM hoặc GRU (3 lớp và 2 lớp) và đã được huấn luyện trên một tập dữ liệu lớn với số khung frame đồng nhất cho mỗi kí hiệu ngôn ngữ.

Kết quả đạt được của mô hình là rất đáng khích lệ. Qua quá trình huấn luyện, chúng tôi đã theo dõi sự tiến triển của mô hình thông qua TensorBoard và đã đạt được độ chính xác cao trên tập dữ liệu kiểm tra. Kết quả này cho thấy mô hình đã học được cách nhận diện và phân loại các hành động ngôn ngữ với hiệu suất tốt.

Bên cạnh đó, việc trích xuất các keypoints tương ứng với mỗi frame từ dữ liệu đầu vào đã cung cấp thông tin quan trọng về vị trí và động tác của tay và pose trong quá trình nhận diện hành động. Số lượng keypoints là 258, bao gồm các điểm tay và pose, giúp tăng độ phức tạp và độ chính xác của mô hình. Đồng thời thực hiện tiền xử lý loại bỏ các điểm chân không quan trọng đối với ngôn ngữ ký hiệu. Kết quả nghiên cứu là phép đánh giá để áp dụng hệ thống vào thời gian thực.

Đóng góp:

Kết quả của nghiên cứu này có thể ứng dụng trong

nhiều lĩnh vực như giao tiếp ngôn ngữ cử chỉ, nhận dạng hành động trong video, hệ thống giám sát an ninh, và nhiều ứng dụng khác. Tiềm năng của mô hình này là rất lớn và cung cấp một cơ sở vững chắc cho các nghiên cứu và ứng dụng trong tương lai.

Tổng kết lại, bài báo này đã trình bày một mô hình nhận diện hành động hiệu quả sử dụng Deep Learning và phương pháp trích xuất keypoints. Kết quả và tiềm năng của mô hình đã được chứng minh thông qua quá trình huấn luyện và đánh giá. Hy vọng rằng công trình này sẽ góp phần vào sự phát triển và ứng dụng của nhận diện hành động trong các lĩnh vực thực tế.

Hướng phát triển:

Điều chỉnh mô hình để đạt hiệu suất tốt hơn trên các tập dữ liệu lớn được công bố.

Phát triển hệ thống trên thời gian thực, kết hợp với dự đoán ngôn ngữ, liên kết các đơn vị ngôn ngữ kí hiệu để thành lập một câu giao tiếp hoàn chỉnh.

Xây dựng ứng dụng dựa trên hệ thống nhận diện ngôn ngữ kí hiệu được đề xuất để góp phần giúp đỡ về mặt giao tiếp của người khiếm thính với những người xung quanh.

Tài liệu tham khảo

[1] Abdalla, M.S.; Hemayed, E.E. Dynamic hand gesture recognition of arabic sign language using hand motion trajectory features. Glob. J. Comput. Sci. Technol. 2013, 13, 27–33.

[2] Liao, Y.; Xiong, P.; Min, W. Weiqiong Min, and Jiahao Lu. Dynamic sign language recognition based on video sequence with blstm-==3d residual networks. IEEE Access 2019, 7, 38044–38054.

[3] Escobedo, E.; Ramirez, L.; Camara, G. Dynamic sign language recognition based on convolutional neural networks and texture maps. In Proceedings of the 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Rio de Janeiro, Brazil, 28–30 October 2019; pp. 265–272.

[4] Chaikaew, A.; Somkuan, K.; Yuyen, T. Thai sign language recognition: An application of deep neural network. In Proceedings of the 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, Cha-am, Thailand, 3–6 March 2021; pp. 128–131

[5] Hoang, M.T.; Yuen, B.; Dong, X.; Lu, T.; Westendorp, R.; Reddy, K. Recurrent Neural Networks for Accurate RSSI Indoor Localization. IEEE Internet Things J. 2019, 6, 10639–10651.

[6] Kim, T.; Keane, J.; Wang, W.; Tang, H.; Riggle, J.; Shakhnarovich, G.; Brentari, D.; Livescu, K. Lexicon-free fingerspelling recognition from video: Data, models, and signer adaptation. Comput. Speech Lang. 2017, 46, 209–232.

[7] Mohandes, M.; Deriche, M.; Liu, J. Image-based and sensor-based approaches to Arabic sign language recognition. IEEE Trans. Hum.-Mach. Syst. 2014, 44, 551–557.

[8] Sonawane, T.; Lavhate, R.; Pandav, P.; Rathod, D. Sign language recognition using leap motion controller. Int. J. Adv. Res. Innov. Ideas Edu. 2017, 3, 1878–1883.

[9] Li, K.; Zhou, Z.; Lee, C.H. Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications. ACM Trans. Access. Comput. (TACCESS) 2016, 8, 1–23. [CrossRef]

[10] Yang, X.; Chen, X.; Cao, X.; Wei, S.; Zhang, X. Chinese sign language recognition based on an optimized tree-structure framework. IEEE J. Biomed. Health Inform. 2016, 21, 994–1004. [CrossRef] [PubMed]

[11] Liu, T.; Zhou, W.; Li, H. Sign language recognition with long short-term memory. In Proceedings of the 2016 IEEE international conference on image processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 2871–2875.

[12] Ma, Z.; Lai, Y.; Kleijn, W.B.; Song, Y.Z.; Wang, L.; Guo, J. Variational Bayesian learning for Dirichlet process mixture of inverted Dirichlet distributions in non-Gaussian image feature modeling. IEEE Trans. Neural Netw. Learn. Syst. 2018, 30, 449–463. [CrossRef] [PubMed]

[13] Ding, L.; Martinez, A. Three-Dimensional Shape and Motion Reconstruction for the Analysis of American Sign Language. In Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), New York, NY, USA, 17–22 June 2006; pp. 146–146. [CrossRef]

[14] Kumar, P.; Gauba, H.; Roy, P.P.; Dogra, D.P. A multimodal framework for sensor based sign language recognition. Neurocomputing 2017, 259, 21–38. [CrossRef]

[15] Zadghorban, M.; Nahvi, M. An algorithm on sign words extraction and recognition of continuous Persian sign language based on motion and shape features of hands. Pattern Anal. Appl. 2018, 21, 323–335.

[16] Moussa, M.M.; Shoitan, R.; Abdallah, M.S. Efficient common objects localization based on deep hybrid Siamese network. J. Intell. Fuzzy Syst. 2021, 41, 3499–3508. [CrossRef]

[17] Abdallah, M.S.; Kim, H.; Ragab, M.E.; Hemayed, E.E. Zero-shot deep learning for media mining: Person spotting and face clustering in video big data. Electronics 2019, 8, 1394. [CrossRef]

[18] Cui, R.; Liu, H.; Zhang, C. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7361–7369.

[19] Rao, G.A.; Syamala, K.; Kishore, P.; Sastry, A. Deep convolutional neural networks for sign language recognition. In Proceedings of the 2018 Conference on Signal Processing And Communication Engineering Systems (SPACES), Vijayawada, India, 4–5 January 2018; pp. 194–197.

[20] Kishore, P.; Kumar, D.A.; Goutham, E.; Manikanta, M. Continuous sign language recognition from tracking and shape features using fuzzy inference engine. In Proceedings of the 2016 International Conference on Wireless Communications, Signal

Processing and Networking (WiSPNET), Chennai, India, 23–25 March 2016; pp. 2165–2170.
[21] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Commun. ACM 2017, 60 , 84–90. [CrossRef]
[22] Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
[23] Srivastava, N.; Mansimov, E.; Salakhudinov, R. Unsupervised learning of video representations using lstms. In Proceedings of the International Conference on Machine Learning. PMLR, Lille, France, 7–9 July 2015; pp. 843–852.
[24] Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential deep learning for human action recognition. In Proceedings of the International Workshop on Human Behavior Understanding; Springer: Berlin/Heidelberg, Germany, 2011; pp. 29–39.
[25] Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
[26] Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.G.; Lee, J.; et al. Mediapipe: A framework for building perception pipelines. arXiv 2019, arXiv:1906.08172.
[27] Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.L.; Grundmann, M. Mediapipe hands: On-device real-time hand tracking. arXiv 2020, arXiv:2006.10214.
[28] Bazarevsky, V.; Grishchenko, I.; Raveendran, K.; Zhu, T.; Zhang, F.; Grundmann, M. Blazepose: On-device real-time body pose tracking. arXiv 2020, arXiv:2006.10204.
[29] Siami-Namini, S.; Tavakoli, N.; Namin, A.S. The performance of LSTM and BiLSTM in forecasting time series. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 3285–3292.
[30] Halder, Arpita, and Akshit Tayade. "Real-time vernacular sign language recognition using mediapipe and machine learning." Journal homepage: www. ijrpr. com ISSN 2582 (2021): 7421.
[31] Samaan, G.H.; Wadie, A.R.; Attia, A.K.; Asaad, A.M.; Kamel, A.E.; Slim, S.O.; Abdallah, M.S.; Cho, Y.-I. MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition. Electronics 2022, 11, 3228. https://doi.org/10.3390/electronics11193228