TRƯỜNG ĐẠI HỌC SƯ PHẠM THÀNH PHỐ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

NHẬN DẠNG NGÔN NGỮ KÝ HIỆU TIẾNG VIỆT CỦA NGƯỜI KHIẾM THÍNH SỬ DỤNG MEDIAPIPE VÀ LSTM

Trần Tú Quyên

1 TỔNG QUAN ĐỀ TÀI

Ngôn ngữ ký hiệu (NNKH) của người khiếm thính, còn có thể gọi là thủ ngữ, là ngôn ngữ hình ảnh với các quy tắc riêng, cấu trúc và ngữ pháp khác nhiều so với ngôn ngữ nói [1]. Ngôn ngữ này được thể hiện bằng biểu cảm nét mặt, các chuyển động của bàn tay và ngón tay,… Mục tiêu của đề tài là tạo ra một mô hình hỗ trợ người khiếm thính trong giao tiếp hàng ngày, đồng thời đóng góp vào việc phổ cập và nâng cao nhận thức về NNKH trong cộng đồng tiếng Việt, đặc biệt là giáo dục và giao tiếp.

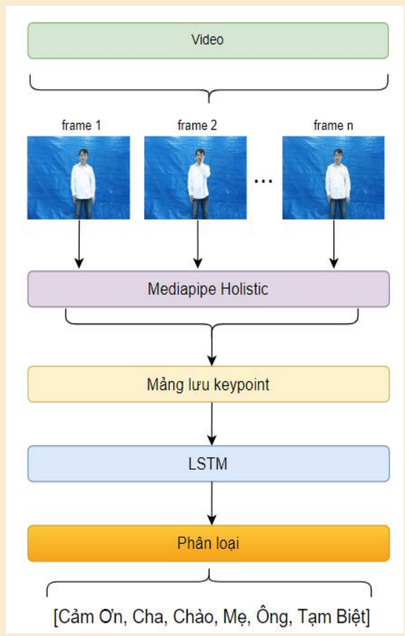
Tính đến nay đã có rất nhiều nhóm tác giả trong và ngoài nước quan tâm đến mô hình nhận diện NNKH. Chẳng hạn như nghiên cứu "Deep Learning for Vietnamese Sign Language Recognition in Video Sequence"[2] của nhóm tác giả Anh H.Vo, Van-Huy.Pham và Bao T.Nguyen đã tiến hành nhận dạng các tư thế từ bộ dữ liệu VSL trên một chuỗi các video. Từ chuỗi video đầu vào các khung chính được trích xuất một cách thủ công, loại bỏ các vùng liên quan đến mặt và chỉ lấy các vùng liên quan đến tay. Sau khi train với 2 mô hình SVM và Deep Learning độ chính xác đạt được lần lượt là 88.5% và 95,83%; "Sign Pose-based Transformer for Word-level Sign Language Recognition"[3] của Matyáš Boháček và Marek Hrúz. Nhóm đã tiến hành ước tính các tư thế từ mỗi khung hình video, sử dụng thuật toán ước tính tư thế tiêu chuẩn từ Vision API để trích xuất 54 điểm mốc cơ thể, kết hợp với mô hình Transformer. Mô hình sau khi được huấn luyện đã đạt được tỷ lệ nhận dạng 63.18% trên tập dữ liệu WLASL100, 43.78% trên bộ WLASL300 và đạt độ chính xác tuyệt đối 100% trên bộ LSA64; "Sign Language recognition based on hand and body skeletal data"[4] của nhóm tác giả Dimitrios Konstantinidis, Kosmas Dimitropoulos và Petros Daras - tiến hành nhận dạng khung xương bằng dựa trên mô hình LSTM kết hợp với khung xương. Khung xương được dựa trên các đặc điểm xương tay, cơ thể được trích xuất từ video RGB giúp phân biệt rõ hơn, 18 điểm cho cơ thể và 21 điểm cho tay. Mô hình sau khi train đạt được 98.09% trên bộ LSA64.

Nội dung của đề tài này là xây dựng một mô hình nhận dạng Ngôn ngữ ký hiệu tiếng Việt, tập trung vào việc nhận diện các từ khóa quan trọng như "cha", "mẹ" , "ông" và các biểu hiện giao tiếp như "xin chào", "tạm biệt", "cảm ơn". Dữ liệu được thu thập từ bộ dữ liệu VSL của nhóm Duc-Hoang Vo trường Đại học Đà Nẵng [5]. Phương pháp của chúng tôi là phát triển một hệ thống thông minh, kết hợp cả các phương pháp học máy và sử dụng dữ liệu lớn để tối ưu hóa độ chính xác và hiệu suất.

2 PHƯƠNG PHÁP ĐỀ XUẤT

Mô hình đề xuất

Chúng tôi đề xuất một mô hình gồm hai phần chính là MediaPipe Holistic để trích các đặc trưng về khung xương và mô hình LSTM cho bài toán nhận dạng NNKH.

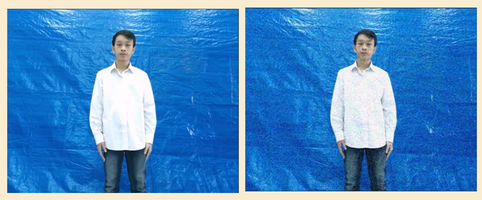


Hình 1: *Kiến trúc mô hình*

Xử lý và tăng cường dữ liệu video

Độ dài video trong tập dữ liệu từ 3 đến 6 giây, để thuận lợi trong quá trình đào tạo mô hình chúng tôi chọn độ dài đồng nhất là 4 giây, lọc các video không đủ chuẩn.

Sau đó, cân bằng dữ liệu giữa các nhãn, chúng tôi tăng dữ liệu bằng cách thêm nhiễu Gaussian vào một số video đã có.



Hình 2: *Hình cắt từ video gốc và video sau khi thêm nhiễu Gaussian*

Sau khi thực hiện các bước xử lý, ta được thông tin dữ liệu như *Bảng 1.*

Bảng 1: *Thông tin bộ dữ liệu*

Định dạng video	AVI
Độ phân giải	512x372
Số khung hình mỗi giây	10
Số video	600
Số người ký hiệu	5
Số label	6
Số video mỗi label	100
Thời lượng	4 giây

Trích xuất khung xương bằng Mediapipe Holistic

Sử dụng giải pháp dùng để ước tính tư thế người từ Mediapipe Holistic là MediaPipe Pose và Mediapipe Hands. Đối với *Pose,* lấy 25 landmarks (các điểm từ đầu đến hông); với Hands, lấy 21 landmarks mỗi tay.

Các landmarks trên của Mediapipe Pose có 4 thông tin là x, y, z, visibility. Mediapipe Hands có 3 thông tin là x, y, z. Nhưng điểm z không được đào tạo đầy đủ để dự đoán độ sâu[6] nên với dữ liệu đầu vào cho mô hình đề xuất, chúng tôi lấy 2 tọa độ điểm x và y của các landmarks.

Vì thế, thu được 67 landmarks và 134 tọa độ điểm trên một khung hình. Các landmarks không xuất hiện trong khung hình thì các điểm toạ độ sẽ được cho bằng 0.

Input cho mô hình dự đoán là (1, 40, 134) tương ứng với 1 video đầu vào có 40 khung hình với mỗi khung hình có 134 điểm.

3 THỰC NGHIỆM VÀ KẾT QUẢ

Môi trường thực nghiệm

Thông tin thiết bị thực nghiệm: ***Laptop Dell Inspiron***

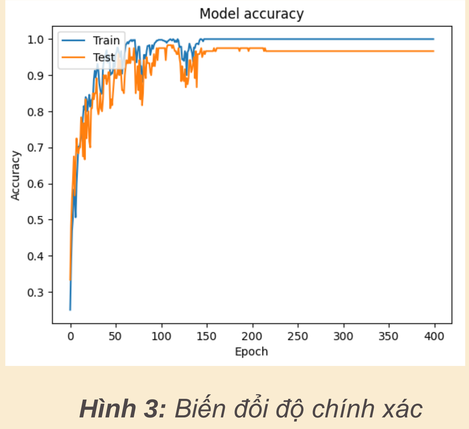
- Hệ điều hành: Windows 11 Home Single Language
- CPU: i7-10750H
- Memory: 16GB RAM
- Ổ cứng: SSD 512GB

Tham số mô hình

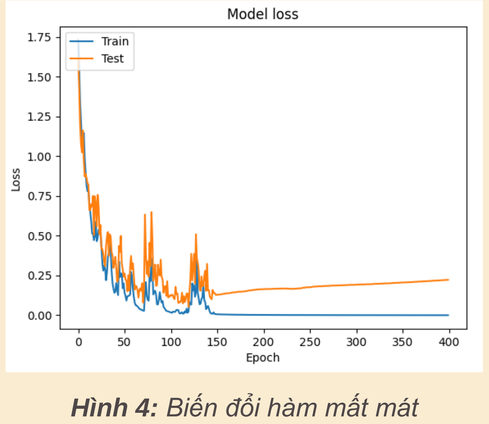
Bảng 2: *Số lượng thông số huấn luyện của mô hình đề xuất*

Layer	Output Shape	Param
lstm (LSTM)	(None, 40, 256)	400384
lstm_1 (LSTM)	(None, 40, 128)	197120
lstm_2 (LSTM)	(None, 40, 64)	49408
lstm_3 (LSTM)	(None, 40, 32)	12416
lstm_4 (LSTM)	(None, 16)	3136
dense (Dense)	(None, 6)	102
Total params: 662566		

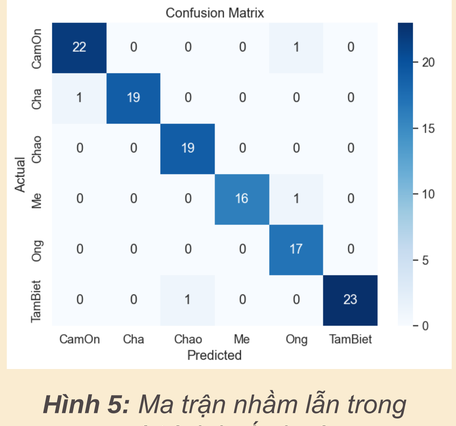
Kết quả thực nghiệm: với mô hình đề xuất và các tham số kể trên, chúng tôi đạt kết quả tốt nhất ở 400 epoch.



Hình 3: *Biến đổi độ chính xác*



Hình 4: *Biến đổi hàm mất mát*



Hình 5: *Ma trận nhầm lẫn trong quá trình huấn luyện*

Bảng 3: *Kết quả mô hình đề xuất*

Accuracy	Precision	Recall	F1 Score
0.9666	0.97	0.9667	0.9667

4 KẾT LUẬN

Kết luận

Nhìn chung đề tài đã có thể nhận diện và phân loại 6 ký hiệu mà chúng tôi đã đề cập bằng cách kết hợp Mediapipe và LSTM. Đạt kết quả khả quan với độ chính xác là 96,66%.
Bên cạnh những kết quả đã đạt được, mô hình còn nhiều hạn chế khi còn nhận sai giữa các ký hiệu.

Hướng phát triển

- Cải thiện, thu thập và tăng cường thêm dữ liệu.
- Tìm hiểu và thử thêm các mô hình mới để nâng cao độ chính xác hơn.
- Tìm hiểu thêm về các cách tiền xử lý với các đặc trưng khung xương.
- Tìm hiểu về Xử lý ngôn ngữ tự nhiên để có thể sinh các câu mẫu từ các từ được biểu diễn.

5 TÀI LIỆU THAM KHẢO

[1] X. M. Cao Thị, "Quá trình hình thành và phát triển ngôn ngữ kí hiệu", *Tạp Chí KHOA HỌC ĐHSP TPHCM*, vol. 46, p. 181, 2013
[2] A. Vo, V.-H. Pham, and B. Thien, "Deep Learning for Vietnamese Sign Language Recognition in Video Sequence," *Int. J. Mach. Learn. Comput.*, vol. 9, Jul. 2019, doi: 10.18178/ijmlc.2019.9.4.823.
[3] M. Bohacek and M. Hruz, "Sign Pose-based Transformer for Word-level Sign Language Recognition," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA: IEEE, Jan. 2022, pp. 182–191. doi: 10.1109/WACVW54805.2022.00024.
[4] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "SIGN LANGUAGE RECOGNITION BASED ON HAND AND BODY SKELETAL DATA," in *2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Helsinki: IEEE, Jun. 2018, pp. 1–4. doi: 10.1109/3DTV.2018.8478467.
[5] D. H. Vo, "Data for Dynamic Vietnamese Sign Language." [Online]. Available: http://test101.udn.vn/d-VSL/
[6] MediaPipe Holistic, https://github.com/google/mediapipe/blob/master/docs/solutions/holistic.md