MoViNet-A2 cho bài toán nhận diện ký hiệu Tiếng Việt

MoViNet-A2 for Vietnamese sign language recognition

Trương Duy Việt[1*], Ngô Hữu Gia Huy[1], Phạm Đăng Khôi[1], Nguyễn Trần Thiên Phúc[1]

[1]Trường Cao đẳng Đà Lạt, Thành phố Đà Lạt, Việt Nam

[*]Tác giả liên hệ, Email: truongduyviet@gmail.com

THÔNG TIN

DOI:10.46223/HCMCOUJS.tech.vi.20.2.4201.2025

Ngày nhận: 06/03/2025

Ngày nhận lại: 30/04/2025

Duyệt đăng: 26/05/2025

Từ khóa:

học sâu; nhận dạng hành động; nhận diện ngôn ngữ ký hiệu; MoViNet-A2; tăng cường dữ liệu

TÓM TẮT

Nhận diện ngôn ngữ ký hiệu từ video là một bài toán quan trọng nhằm hỗ trợ giao tiếp cho cộng đồng người khiếm thính. Tuy nhiên, sự đa dạng của cử chỉ, góc quay khác nhau và điều kiện môi trường biến thiên đặt ra nhiều thách thức cho các hệ thống nhận dạng truyền thống. Trong nghiên cứu này, chúng tôi đề xuất một phương pháp nhận diện ngôn ngữ ký hiệu tiếng Việt dựa trên MoViNet-A2, một mô hình tiên tiến được tối ưu hóa cho nhận dạng hành động trong video trên thiết bị di động. Bộ dữ liệu nghiên cứu bao gồm 98 từ hoặc cụm từ, được thực hiện bởi 18 học sinh từ Trường Khuyết tật Lâm Đồng - Đà Lạt, với tổng cộng 4,709 video từ ba góc quay khác nhau, đảm bảo tính đa dạng trong dữ liệu huấn luyện. Kết hợp với MoViNet-A2 là backbone được tiền huấn luyện trên tập Kinetics-600, kết hợp với các kỹ thuật tiền xử lý như cân bằng lớp, chuẩn hóa độ sáng và các phương pháp tăng cường dữ liệu nhằm nâng cao khả năng tổng quát hóa của mô hình. Kết quả thực nghiệm đạt độ chính xác Top-1 88.55% trên tập kiểm tra. Nghiên cứu cho thấy phương pháp đề xuất đạt hiệu suất cao trong việc phân loại và nhận diện các cử chỉ ký hiệu, đồng thời đảm bảo khả năng xử lý thời gian thực trên thiết bị di động. Nghiên cứu này không chỉ góp phần nâng cao độ chính xác của hệ thống nhận diện ngôn ngữ ký hiệu mà còn mở ra tiềm năng ứng dụng thực tế trong hỗ trợ giao tiếp cho cộng đồng người khiếm thính.

ABSTRACT

Sign language recognition from video is an essential task to support communication for the hearing-impaired community. However, the diversity of gestures, different camera angles, and varying environmental conditions pose significant challenges for traditional recognition systems. In this study, we propose a Vietnamese sign language recognition method based on MoViNet-A2, an advanced model optimized for action recognition in videos on mobile devices. The research dataset consists of 98 words or phrases, performed by 18 students from Lam Dong - Da Lat School for the Disabled, with a total of 4,709 videos from three different camera angles, ensuring diversity in training data. MoViNet-A2 serves as the backbone, pre-trained on the Kinetics-600 dataset. It is combined with preprocessing

techniques such as class balancing, brightness normalization, and data augmentation to improve model generalization. Our method achieves a Top-1 Accuracy of 88.55%. Experimental results demonstrate that the proposed method achieves high performance in classifying and recognizing sign language gestures while ensuring real-time processing capabilities on mobile devices. This research not only improves the accuracy of sign language recognition systems but also opens up practical applications in facilitating communication for the hearing-impaired community.

Keywords:

deep learning; action recognition; sign language recognition; MoviNet-A2; data augmentation

1. Giới thiệu

Việc nhận diện ngôn ngữ ký hiệu từ video đang trở thành một vấn đề cấp thiết nhằm hỗ trợ giao tiếp cho cộng đồng người khiếm thính. Trong bối cảnh đó, các hệ thống nhận diện ký hiệu trên thiết bị di động không chỉ cần đảm bảo tính chính xác cao mà còn phải đáp ứng yêu cầu xử lý thời gian thực trên các thiết bị có tài nguyên hạn chế. Tuy nhiên, sự đa dạng của cử chỉ, góc quay khác nhau và điều kiện môi trường biến thiên đã tạo ra không ít thách thức cho các hệ thống nhận dạng truyền thống.

Trong nghiên cứu này, chúng tôi đề xuất một giải pháp nhận diện ngôn ngữ ký hiệu tiếng Việt dựa trên mô hình MoViNet-A2, một kiến trúc tiên tiến được thiết kế tối ưu cho các ứng dụng nhận dạng hành động từ video trên thiết bị di động. Mô hình sử dụng các phép tích chập dạng 2+1D và 2+3D kết hợp với hàm kích hoạt hard_swish, giúp mô hình học được các đặc trưng không gian - thời gian một cách hiệu quả, đồng thời giảm thiểu số lượng tham số và chi phí tính toán. Hơn nữa, việc áp dụng các causal convolutions đảm bảo rằng các dự đoán chỉ dựa trên thông tin từ các khung hình trước đó, từ đó hỗ trợ xử lý liên tục trong môi trường thời gian thực.

Bộ dữ liệu của dự án được xây dựng nhằm hỗ trợ nghiên cứu nhận diện ngôn ngữ ký hiệu tiếng Việt trên thiết bị di động. Bộ dữ liệu bao gồm tổng cộng 98 từ hoặc cụm từ, được thực hiện bởi 18 học sinh từ Trường Khuyết tật Lâm Đồng - Đà Lạt. Các video được quay đồng thời từ ba góc khác nhau (chính diện, bên trái, bên phải) với định dạng 25fps và độ phân giải 1080×1080 pixel, đảm bảo chất lượng hình ảnh đủ cao cho việc nhận diện cử chỉ chính xác. Tổng số video thu thập được là 4,709 video, với số khung hình trung bình mỗi video đạt 62.96, thời lượng trung bình khoảng 2.52 giây, và tổng số khung hình vượt qua 296,500. Những thông số này tạo nên một nguồn dữ liệu phong phú, góp phần làm nền tảng vững chắc cho việc huấn luyện và đánh giá hệ thống.

Trong quy trình huấn luyện, chúng tôi sử dụng backbone của MoViNet-A2 đã được tiền huấn luyện trên tập dữ liệu Kinetics-600 để tận dụng các đặc trưng chung của hành động được học từ một tập dữ liệu lớn và đa dạng. Sau đó, một classifier head với 98 lớp đầu ra, tương ứng với 98 ký hiệu trong bộ dữ liệu được gắn thêm nhằm chuyển đổi các đặc trưng trích xuất từ video thành các nhãn phân loại cụ thể. Quá trình tiền xử lý dữ liệu bao gồm chuẩn hóa độ sáng, cân bằng lớp và áp dụng các kỹ thuật augmentation như random cropping, frame jittering và các biến đổi không gian, giúp mô hình có khả năng tổng quát hóa cao trong các điều kiện quay khác nhau.

Nhờ vào sự kết hợp giữa một kiến trúc hiện đại, quy trình tiền huấn luyện và tinh chỉnh kỹ lưỡng trên bộ dữ liệu custom, hệ thống nhận diện ngôn ngữ ký hiệu tiếng Việt mà chúng tôi đề xuất hứa hẹn sẽ đạt hiệu năng cao trong việc phân loại và nhận diện các cử chỉ phức tạp. Hệ thống này không chỉ mang lại những kết quả ấn tượng về độ chính xác mà còn

đáp ứng được yêu cầu xử lý thời gian thực trên các thiết bị di động, từ đó hỗ trợ hiệu quả cho giao tiếp của người khiếm thính. Trong bài báo này, chúng tôi sẽ trình bày chi tiết về kiến trúc mô hình, quy trình tiền xử lý, chiến lược huấn luyện cũng như các kết quả thực nghiệm minh họa hiệu quả của hệ thống trên bộ dữ liệu được xây dựng.

2. Cơ sở lý thuyết

2.1. Các phương pháp truyền thống dựa vào trích xuất đặc trưng thủ công

Trong những năm đầu tiên, khi công nghệ xử lý hình ảnh còn hạn chế, các nghiên cứu về nhận diện ngôn ngữ ký hiệu chủ yếu dựa vào việc trích xuất các đặc trưng thủ công từ dữ liệu video. Công trình của Starner và cộng sự (1998) đã tập trung vào việc nhận dạng ngôn ngữ ký hiệu Mỹ trong thời gian thực. Các tác giả đã thiết kế hệ thống bằng cách xác định các đặc trưng thị giác cơ bản như hình dạng bàn tay, chuyển động và hướng di chuyển của các bộ phận cơ thể. Sau đó, họ sử dụng mô hình ẩn Markov (HMM) để mô hình hóa chuỗi các cử chỉ này. Kết quả cho thấy, dù các đặc trưng được trích xuất có tính chất đơn giản và thủ công, sự kết hợp với HMM vẫn cho phép hệ thống hoạt động ổn định và đáp ứng yêu cầu thời gian thực, từ đó mở ra hướng nghiên cứu mới về xử lý tín hiệu động.

Tương tự, Vogler và Metaxas (2001) đã đề xuất một khuôn khổ toàn diện cho việc nhận diện ngôn ngữ ký hiệu thời gian thực. Công trình này không chỉ tập trung vào việc trích xuất đặc trưng từ từng khung hình mà còn chú trọng đến quá trình tiền xử lý, nhằm khắc phục các vấn đề về nhiễu, biến đổi ánh sáng và góc quay. Việc kết hợp các bước tiền xử lý kỹ lưỡng với mô hình thống kê như HMM đã giúp cải thiện đáng kể hiệu năng của hệ thống trong môi trường thực tế. Nhờ đó, các giải pháp truyền thống đã tạo ra nền tảng để các nghiên cứu sau này có thể chuyển sang khai thác các kỹ thuật học sâu hiện đại.

2.2. Ứng dụng học sâu trong nhận diện ngôn ngữ ký hiệu

Với sự ra đời của các thuật toán học sâu và khả năng tính toán ngày càng mạnh mẽ của máy tính, giới nghiên cứu nhanh chóng chuyển hướng từ các phương pháp truyền thống sang khai thác khả năng học tự động các đặc trưng từ dữ liệu. Các công trình của Koller và cộng sự (2015), Koller và cộng sự (2016) đã mở ra một hướng tiếp cận mới bằng cách kết hợp giữa mạng nơ-ron tích chập (CNN) với HMM. Trong Koller và cộng sự (2015), tác giả tập trung xây dựng một hệ thống có khả năng xử lý một từ vựng lớn với nhiều người ký hiệu khác nhau, giúp giải quyết vấn đề đa dạng trong cách thể hiện của người ký hiệu. Tiếp theo, công trình "Deep Sign" được trình bày ở Koller và cộng sự (2016) đã phát triển kiến trúc CNN-HMM, trong đó CNN được dùng để tự động học các đặc trưng phức tạp từ video, còn HMM đảm nhận việc mô hình hóa tính liên tục của chuỗi ký hiệu. Qua đó, hệ thống đạt được độ chính xác cao hơn và khả năng khái quát hóa vượt trội so với các phương pháp dựa vào đặc trưng thủ công.

Bên cạnh đó, Pigou và cộng sự (2015) đã chứng minh tiềm năng của các mạng CNN khi được áp dụng trực tiếp lên các khung hình video mà không cần các bước tiền xử lý phức tạp. Phương pháp này cho thấy các đặc trưng hình ảnh quan trọng có thể được tự động phát hiện và trích xuất, giúp đơn giản hóa quy trình thiết kế hệ thống nhận diện ngôn ngữ ký hiệu. Điều này đánh dấu một bước tiến quan trọng, khi giảm thiểu sự phụ thuộc vào các kỹ thuật trích xuất thủ công vốn tốn thời gian và nhạy cảm với các yếu tố ngoại cảnh.

Mới đây, Kumari và Anand (2024) đã đưa cơ chế attention vào khung CNN-LSTM dành cho bài toán WLASL-100, một tập dữ liệu quy mô trung bình gồm 100 ký hiệu và hơn 40 người ký hiệu. Giai đoạn CNN (ResNet-18) chịu trách nhiệm mã hóa không gian, trong khi

hai lớp Bi-LSTM ghi nhận động học thời gian và khối attention học trọng số khung hình theo ngữ cảnh. Mô hình đạt 84.6% độ chính xác top-1 - cao hơn 6% so với mô hình không attention - đồng thời giảm ≈ 12% thời gian suy luận nhờ chiến lược frame skipping hướng dẫn bởi attention. Qua đó, nghiên cứu khẳng định rằng cơ chế tập trung không chỉ cải thiện độ chính xác trên câu dài mà còn hỗ trợ giảm chi phí tính toán, phù hợp triển khai thời gian thực.

2.3. Các mô hình hiện đại: 3D CNN, Transformer và sự kết hợp giữa chúng

Để khai thác tối đa thông tin không gian - thời gian từ dữ liệu video, các nghiên cứu đã chuyển sang sử dụng mô hình 3D CNN. Molchanov và cộng sự (2015) đã đi tiên phong trong việc ứng dụng 3D CNN cho bài toán nhận diện cử chỉ, bằng cách khai thác mối quan hệ giữa các khung hình liên tiếp. Việc sử dụng các lớp 3D cho phép mô hình hiểu được sự biến đổi liên tục của cử chỉ, từ đó phân loại chính xác hơn các hành động.

Khắc phục nhược điểm "hòa trộn" bàn tay và thân người trong khối 3D CNN, Camgoz và cộng sự (2017) giới thiệu SubUNets - kiến trúc hai nhánh song song: một nhánh chuyên biệt cho ROI-bàn tay (crop từ heat-map) và nhánh còn lại cho toàn bộ khung hình.

Song song với đó, để xử lý các chuỗi video liên tục, Camgoz và cộng sự (2018), Camgoz và cộng sự (2020) đã đề xuất hệ thống Neural Sign Language Translation. Ở công trình Camgoz và cộng sự (2018), kiến trúc kết hợp giữa CNN và các mô hình tuần tự như LSTM đã được sử dụng để chuyển đổi chuỗi video thành văn bản hoặc nhãn ký hiệu, giúp hệ thống xử lý được tính liên tục và ngữ cảnh trong giao tiếp. Công trình Camgoz và cộng sự (2020) tiếp tục phát triển mô hình này với cấu trúc SubUNets, nhằm cải thiện khả năng phân tích các chi tiết nhỏ trong cử chỉ, từ đó tăng cường hiệu quả tổng hợp thông tin từ các khung hình. Những nghiên cứu này mở rộng phạm vi ứng dụng của nhận diện ngôn ngữ ký hiệu từ việc chỉ phân loại từng cử chỉ riêng lẻ sang nhận dạng các câu hoặc đoạn giao tiếp liên tục.

Trong bối cảnh suy luận di động, Kondratyuk và cộng sự (2021) phát triển MoViNets - họ mạng 3D CNN siêu nhẹ dùng kiến trúc Factorized Temporal Depthwise Convolution. Cơ chế "stream buffer" duy trì trạng thái ẩn giữa các clip 08 khung kế tiếp, cho phép suy luận online không chồng lấn tính toán. Trên GPU T4, MoViNet-A2 (224×224) chỉ tiêu thụ 9.1 ms/khung mà vẫn đạt 78.1 % top-1 trên Kinetics-600; trong các thử nghiệm nội bộ với dữ liệu ký hiệu, biến thể fine-tuned đạt 25fps trên điện thoại Pixel 5, chứng minh tiềm năng ứng dụng thời gian thực.

2.4. Nghiên cứu về ngôn ngữ ký hiệu tiếng Việt

Trong bối cảnh ngôn ngữ ký hiệu tiếng Việt, mặc dù số lượng công trình nghiên cứu còn hạn chế so với các ngôn ngữ ký hiệu khác, nhưng các nghiên cứu trong nước đã có những bước tiến đáng kể nhằm giải quyết các thách thức đặc thù. Pham và cộng sự (2018) là nhóm đầu tiên khai thác cảm biến Kinect v1/v2 để xây dựng hệ thống nhận dạng bảng chữ cái VSL theo thời gian thực. Khung xương 3D (25 khớp) được trích xuất chính xác tới 30fps, sau đó chuyển thành vectơ tọa độ, góc khớp và tốc độ khớp; thuật toán Relief-F tự động chọn 128/493 đặc trưng có tính phân biệt cao nhất. Bộ SVM phi tuyến với hạt RBF được huấn luyện trên 3,960 mẫu (29 chữ cái + 02 dấu thanh), đạt 92.3 % top-1 và giữ độ trễ dưới 45 ms/khung, đặt viên gạch đầu tiên cho hướng tận dụng độ sâu nhằm giảm nhiễu nền và biến thiên ánh sáng.

Vu và cộng sự (2025) nâng cấp pipeline sang học sâu, kết hợp ResNet-18 với MediaPipe Hands để tập trung vào ROI bàn tay. Tập dữ liệu tự thu thập 48,000 ảnh (25 ký hiệu + dấu) được tăng cường bằng phép quay ± 15°, nhiễu Gaussian và điều chỉnh gamma,

giúp mô hình học bền vững với điều kiện quay đa dạng. Khi chạy offline, hệ thống đạt 95.0% chính xác và 98.7% AUC; khi port sang điện thoại Realme 7 (Snapdragon 720G) với TensorFlow-Lite-GPU, tốc độ suy luận đạt 23fps, chứng minh tính khả thi của triển khai di động cho người dùng khiếm thính

Nguyen và cộng sự (2025) công bố bộ dữ liệu đa góc nhìn đầu tiên cho VSL: 84,000 video độ dài 04 - 06s, quay đồng thời từ bốn camera HD đặt cách 90° quanh người ký hiệu (30 signer, 1,000 gloss phổ biến). Pipeline gắn nhãn bán tự động, chuẩn hóa ánh sáng và đánh dấu thời gian bắt-kết thúc chuyển động, đồng thời phát hành tập chia train/val/test chuẩn. Thí nghiệm nội bộ cho thấy 3D CNN-LSTM (I3D + Bi-LSTM 512) fine-tune trên tập dữ liệu này giảm WER thêm 19.7% so với mô hình huấn luyện đơn-góc, khẳng định lợi ích của dữ liệu nhiều quan sát không gian cho VSL.

2.5. Hướng nghiên cứu đa phương thức và xây dựng bộ dữ liệu chuẩn

Các nghiên cứu gần đây đã tập trung vào việc xây dựng các bộ dữ liệu chuẩn và tối ưu hóa các siêu tham số của mô hình học sâu, nhằm tăng cường hiệu quả của hệ thống nhận dạng. Al-Qurishi và cộng sự (2021) tổng hợp 167 công trình học sâu về NNKH và cho thấy các hệ RGB-chỉ bị suy giảm tới 14% mAP khi điều kiện ánh sáng đổi, trong khi phương pháp kết hợp RGB + Depth + Pose chỉ suy giảm ≤ 5%. Tác giả minh họa mô hình Deep-Depth-Pose (DDP) - ghép CNN phân giải cặp điểm ứng cử viên, rồi ước lượng pose 3D bằng depth - tăng mAP trung bình 6% và giảm lỗi vị trí tay-không gian 6mm so với đối thủ tốt nhất. Phân tích siêu tham số cho thấy việc kéo dài clip từ 08 khung lên 16 khung và tăng batch-size từ 16 lên 32 đồng thời nâng mAP thêm 6%, cung cấp hướng dẫn thực nghiệm quan trọng cho cộng đồng.

Tiếp thu các khuyến nghị đó, Nguyen và cộng sự (2025) cài sẵn luồng RealSense D435 thu RGB + Depth 60fps, đồng bộ với MediaPipe Pose/Hands, rồi chia sẻ cùng bộ dữ liệu đa góc nhìn kể trên như một chuẩn benchmark "3-Modality". Thử nghiệm so sánh cho thấy I3D-RGB đạt Top-1 79.4%, I3D-RGBD tăng lên 85.1%, còn khi thêm pose-landmark (Early-Fusion) thì đạt mAP 88.2% và F1-score 0.873. Bộ benchmark này thúc đẩy cộng đồng VSL chuyển từ báo cáo độ chính xác đơn lẻ sang hệ số đánh giá toàn diện (WER, BLEU, F1) cũng như kiểm thử fairness giữa signer nam-nữ, qua đó tạo động lực chuẩn hóa dữ liệu và phương pháp cho các nghiên cứu tiếp theo.

Tổng hợp lại, hành trình phát triển của lĩnh vực nhận diện ngôn ngữ ký hiệu đã trải qua nhiều giai đoạn, từ các phương pháp truyền thống dựa trên trích xuất đặc trưng thủ công và HMM đến các mô hình học sâu tiên tiến như CNN, 3D CNN và Transformer. Sự ra đời của các mô hình tối ưu cho xử lý video như MoViNet-A2 càng khẳng định tiềm năng khai thác hiệu quả thông tin không gian và thời gian. Những công trình nghiên cứu này không chỉ thể hiện sự phát triển về mặt lý thuyết mà còn hướng tới các ứng dụng thực tiễn, góp phần tạo ra những giải pháp hỗ trợ giao tiếp hiệu quả cho cộng đồng người khiếm thính, đặc biệt trong bối cảnh nhận diện ngôn ngữ ký hiệu tiếng Việt. Qua đó, các hướng nghiên cứu đa phương thức và xây dựng bộ dữ liệu chuẩn đã mở ra những triển vọng phát triển mới, hứa hẹn sẽ đưa lĩnh vực này tiến xa hơn nữa trong tương lai.

3. Bộ dữ liệu

Bộ dữ liệu được thu thập nhằm hỗ trợ nghiên cứu về nhận diện video ngôn ngữ ký hiệu tiếng Việt trên thiết bị di động. Bộ dữ liệu bao gồm tổng cộng 98 từ/cụm từ trong ngôn ngữ ký hiệu tiếng Việt, được thực hiện bởi 18 người tham gia, là các học sinh của Trường Khuyết tật Lâm Đồng - Đà Lạt. Video được quay đồng thời bằng ba điện thoại từ ba góc quay khác nhau: bên trái, chính diện, bên phải.

Các video được quay ở định dạng 25fps, với độ phân giải 1080 x 1080 pixel nhằm đảm bảo chất lượng hình ảnh đủ cao cho việc nhận diện cử chỉ chính xác. Tổng số video thu thập được là 4,709 video.

Danh sách đầy đủ 98 lớp ký hiệu trong bộ dữ liệu bao gồm các nhóm từ về động vật, màu sắc, trái cây, quan hệ gia đình, hành động, cảm xúc, trạng thái, phẩm chất cá nhân, ... Một số ví dụ trong danh sách lớp ký hiệu:

• **Động vật**: Con chó, con mèo, con gà, con vịt, con rùa, con thỏ, con trâu, con bò, con dê, con heo.

• **Màu sắc**: Màu đen, màu trắng, màu đỏ, màu cam, màu vàng, màu hồng, màu tím, màu nâu.

• **Trái cây:** Quả dâu, quả mận, quả dứa, quả đào, quả đu đủ, quả cam, quả bơ, quả chuối, quả xoài, quả dừa.

• **Quan hệ gia đình**: Bố, mẹ, con trai, con gái, vợ, chồng, ông nội, bà nội, ông ngoại, bà ngoại.

• **Hành động**: Ăn, uống, xem, thèm, mách, khóc, cười, học, dỗi, chết, đi, chạy, bận, hát, múa, nấu, nướng, quan sát, cắm trại, cung cấp, bắt chước, bắt buộc, báo cáo, mua bán.

• **Cảm xúc**: Không quen, không nên, không cần, không cho, không nghe lời, hài hước, thú vị, dũng cảm.

• **Hương vị**: Mặn, đắng, cay, ngọt, đậm, nhạt, ngon miệng.

• **Tính từ/trạng thái**: Xấu, đẹp, chật, hẹp, rộng, dài, cao, lùn, ốm, mập.

• **Phẩm chất cá nhân**: Ngoan, hư, khỏe, mệt, đau, giỏi, chăm chỉ, lười biếng, tốt bụng, sáng tạo.

Bộ dữ liệu này cung cấp nguồn tài nguyên phong phú cho việc nghiên cứu nhận diện ngôn ngữ ký hiệu tiếng Việt trên thiết bị di động. Các thống kê chi tiết sẽ được cập nhật sau quá trình tiền xử lý và kiểm tra chất lượng dữ liệu.

Bảng 1

Phân Tích Bộ Dữ Liệu

Thông số	**Số liệu**
Tổng video	4709
Số khung hình trung bình mỗi video	62.96
Số khung hình tối thiểu mỗi video	33
Số khung hình tối đa mỗi video	114
Độ dài trung bình mỗi video (s)	2.52
Độ dài tối thiểu mỗi video (s)	1.32
Độ dài tối đa mỗi video (s)	4.56
Tổng khung hình	296500
Tổng thời lượng (s)	11860.00

Ghi chú: Thống kê này được thực hiện trong bổ dữ liệu của nhóm nghiên cứu

4. Phương pháp nghiên cứu

4.1. Giới thiệu về mô hình MoViNet-A2 và kiến trúc sử dụng trong nhận diện video ngôn ngữ ký hiệu

Mô hình MoViNet-A2 là một trong những mô hình tiên tiến được phát triển để nhận diện các đặc trưng trong dữ liệu video, với mục tiêu chủ yếu là nhận diện các hoạt động hoặc hành động trong video. Đây là mô hình thuộc dòng "motion enhanced video classification" của Google, và đặc biệt thích hợp cho các bài toán nhận diện hành động trong video, bao gồm cả nhận dạng ngôn ngữ ký hiệu. Trong nghiên cứu này, chúng tôi sử dụng mô hình mạnh mẽ này để dễ dàng triển khai ứng dụng trên các thiết bị di động.

Kiến trúc của MoViNet-A2:

• Backbone: MoViNet-A2 sử dụng một bộ backbone được xây dựng dựa trên kiến trúc mạng thần kinh có cấu trúc "2+1D" và "2+3D". Đây là hai kiến trúc tối ưu để xử lý dữ liệu video, giúp giảm thiểu số lượng tham số mà vẫn giữ được hiệu năng cao.

– 2+1D: Đây là một cấu trúc convolutional với kernel 2D cho không gian (chiều cao và chiều rộng), nhưng chỉ sử dụng kernel 1D cho chiều thời gian. Điều này giúp giảm thiểu độ phức tạp tính toán trong khi vẫn bảo tồn khả năng phân tích không gian và thời gian.

– 2+3D: Là sự kết hợp của một số lớp convolutional 2D trong không gian và một số lớp 3D trong không gian-thời gian. Kiến trúc này giúp mạng học được những mối quan hệ không chỉ trong không gian mà còn trong thời gian, rất phù hợp với dữ liệu video.

Bảng 2

Kiến Trúc MoViNet-A2 dùng TuNAS, chạy 50 Khung Hình trên Kinetics 600

Model Stage	Operation	Output Size
data	Stride 5, RGB	50×224^2
conv1	1×3^2, 16	50×112^2
block2	1×5^2, 16, 40 3×3^2, 16, 40 3×3^2, 16, 64	50×56^2
block3	3×3^2, 40, 96 3×3^2, 40, 120 3×3^2, 40, 96 3×3^2, 40, 96 3×3^2, 40, 120	50×28^2
block4	5×3^2, 72, 240 3×3^2, 72, 160 3×3^2, 72, 240 3×3^2, 72, 192 3×3^2, 72, 240	50×14^2
block5	5×3^2, 72, 240 3×3^2, 72, 240 3×3^2, 72, 240 3×3^2, 72, 240 1×5^2, 72, 144 3×3^2, 72, 240	50×14^2

Model Stage	Operation	Output Size
block6	5×3^2, 144, 480 1×5^2, 144, 360 1×5^2, 144, 360 1×5^2, 144, 480 1×5^2, 144, 480 3×3^2, 144, 480 1×3^2, 144, 576	50×7^2
conv7	1×1^2, 640	50×7^2
pool8	50×7^2	1×1^2
dense9	1×1^2, 2084	1×1^2
dense10	1×1^2, class num	1×1^2

Ghi chú: Dữ liệu từ "MoViNets: Mobile video networks for efficient video recognition" bởi D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, và B. Gong, 2021 (https://doi.org/10.1109/CVPR46437.2021.01576)

• Activation functions: sử dụng hard_swish làm hàm kích hoạt, một biến thể của hàm swish có thể giúp cải thiện sự hội tụ trong quá trình huấn luyện, đồng thời duy trì được khả năng phi tuyến mạnh mẽ.

• Causal Conv: sử dụng causal convolutions, có nghĩa là mô hình chỉ tính toán các tính năng từ các khung hình trước đó (dữ liệu trước). Điều này rất quan trọng trong các bài toán video, đặc biệt khi việc dự đoán phải đảm bảo không làm rò rỉ thông tin tương lai.

4.2. Phần mô hình: Sử dụng backbone đã được pretrained và classifier head

Backbone: được tải trọng số đã được huấn luyện từ trước (pretrained), điều này giúp mạng có thể tận dụng các đặc trưng đã học từ một tập dữ liệu lớn và cải thiện khả năng nhận diện trên các tập dữ liệu nhỏ hơn. Mô hình backbone này chỉ chịu trách nhiệm trích xuất các đặc trưng từ video và không thực hiện phân loại trực tiếp. Đây là một kỹ thuật transfer learning, nơi mà mô hình đã được huấn luyện trên một nhiệm vụ liên quan (ví dụ: nhận diện hành động trong video) và chỉ cần fine-tuning lại một chút cho nhiệm vụ mới (như nhận diện ngôn ngữ ký hiệu).

Classifier Head: Để chuyển đổi các đặc trưng đầu ra từ backbone thành các nhãn (classes) trong bài toán nhận diện ngôn ngữ ký hiệu, một classifier head được thêm vào mô hình. Đây là phần thêm vào cuối mô hình, có nhiệm vụ phân loại các đặc trưng mà mô hình học được từ video đầu vào thành các lớp mong muốn. Trong trường hợp này, mô hình được cấu hình để phân loại thành 98 lớp (tương ứng với 98 ký hiệu trong bộ dữ liệu).

Phần forward pass và input shapes: Khi dữ liệu video được đưa vào mô hình, nó có thể có kích thước [batch_size, num_frames, resolution, resolution, 3], trong đó: batch_size là số lượng video trong một batch, num_frames là số lượng khung hình trong mỗi video (ở đây là 13 khung hình), resolution là kích thước của mỗi khung hình (ở đây là 224 x 224 pixel), 3 là số kênh màu (RGB).

4.3. Các đặc điểm nổi bật của MoViNet-A2

Mô hình là một lựa chọn mạnh mẽ cho các bài toán nhận diện hành động trong video, đặc biệt là với những cải tiến về xử lý không gian-thời gian và khả năng hoạt động hiệu quả trên các thiết bị di động. Các đặc điểm nổi bật bao gồm:

• Kiến trúc nhẹ và hiệu quả: được thiết kế để có khả năng chạy trên các thiết bị có phần cứng hạn chế, như điện thoại di động, nhưng vẫn duy trì được hiệu năng cao.

• Khả năng xử lý video mạnh mẽ: nhờ vào việc kết hợp các lớp convolutional 2D và 3D, mô hình có thể khai thác thông tin không gian và thời gian từ video một cách hiệu quả.

• Cấu trúc causal convolutions: Điều này giúp mô hình xử lý video trong thời gian thực mà không gặp phải vấn đề rò rỉ thông tin từ tương lai, giúp nó phù hợp cho các ứng dụng thực tế, chẳng hạn như nhận diện ngôn ngữ ký hiệu trong video thời gian thực.

5. Thử nghiệm

5.1. Quá trình huấn luyện mô hình

Mô hình MoViNet-A2 được huấn luyện trên bộ dữ liệu ngôn ngữ ký hiệu tiếng Việt với 98 từ/cụm, nhằm mục tiêu phân loại các lớp ký hiệu. Quá trình huấn luyện sử dụng phương pháp tối ưu hóa Adam optimizer và hàm mất mát Sparse Categorical Cross Entropy. Việc huấn luyện được thực hiện trên GPU để đảm bảo hiệu suất và tốc độ hội tụ nhanh chóng.

5.2. Phương pháp tối ưu hóa và hàm mất mát

Phương pháp tối ưu hóa sử dụng trong quá trình huấn luyện là Adam optimizer. Đây là một phương pháp tối ưu hóa phổ biến nhờ khả năng điều chỉnh tốc độ học theo từng tham số và tính toán gradient dựa trên các thông số của các lần cập nhật trước đó. Với Adam, tốc độ học được đặt là 0.001, giúp mô hình hội tụ ổn định mà không gặp phải hiện tượng quá khớp (overfitting).

Hàm mất mát được sử dụng là Sparse Categorical Cross Entropy. Hàm mất mát này phù hợp cho bài toán phân loại đa lớp với các nhãn số nguyên, giúp đo lường sự khác biệt giữa phân phối xác suất dự đoán của mô hình và phân phối xác suất thực tế.

5.3. Tiền xử lý và augmentation dữ liệu

Dữ liệu đầu vào là các video, mỗi video sẽ được trích xuất thành một chuỗi các khung hình (frames). Các bước tiền xử lý và augmentation dữ liệu được thực hiện như sau:

• Chuẩn hóa độ sáng: Các giá trị pixel của các khung hình video được chuẩn hóa về dạng float32 trong phạm vi [0, 1] để giảm thiểu ảnh hưởng của độ sáng và độ tương phản trong quá trình huấn luyện. Quá trình chuẩn hóa này giúp tăng độ ổn định của mô hình khi xử lý các video có độ sáng khác nhau.

• Cân bằng lớp: Để tránh mô hình thiên lệch về các lớp có tần suất xuất hiện cao, dữ liệu huấn luyện được cân bằng tỷ lệ giữa các lớp ký hiệu. Việc này giúp mô hình học được cách phân loại chính xác các lớp ít xuất hiện trong dữ liệu.

• Augmentation dữ liệu: Các kỹ thuật augmentation được áp dụng nhằm tăng cường khả năng tổng quát hóa của mô hình, bao gồm:

- Random cropping: Cắt ngẫu nhiên các phần của khung hình, giúp mô hình học được các đặc trưng không bị ảnh hưởng bởi vị trí cố định của đối tượng trong ảnh.

- Frame jittering: Thực hiện thay đổi ngẫu nhiên các giá trị pixel trong mỗi khung hình, giúp mô hình học được sự biến thiên của cảnh vật trong video.

- Spatial transformations: Các phép biến đổi không gian như quay, phóng to, thu nhỏ ảnh được áp dụng để mô phỏng sự thay đổi về góc độ và tỷ lệ, từ đó cải thiện khả năng tổng quát hóa của mô hình.

5.4. Xử lý ảnh

Các khung hình video được xử lý trước khi đưa vào mô hình. Mỗi khung hình sẽ được thay đổi kích thước về một kích thước chuẩn (224 x 224 pixel) với tỷ lệ khung hình được giữ nguyên. Trong trường hợp khung hình có kích thước không phù hợp, chúng được padding để đạt được kích thước đầu ra yêu cầu mà không làm mất đi các thông tin quan trọng của đối tượng trong ảnh. Sau khi xử lý, khung hình sẽ được chuyển đổi thành định dạng chuẩn và chuẩn hóa độ sáng để mô hình có thể tiếp nhận và phân tích.

5.5. Quá trình huấn luyện

Mô hình được huấn luyện với các batch dữ liệu, mỗi batch bao gồm một chuỗi các khung hình video được trích xuất từ các video huấn luyện. Tập huấn luyện được chia thành các phân đoạn nhỏ (mini-batches) nhằm tối ưu hóa quá trình tính toán và tăng tốc độ huấn luyện. Mô hình sẽ được đánh giá qua các chỉ số như độ chính xác (accuracy) và giá trị mất mát (loss) trên tập kiểm tra sau mỗi epoch huấn luyện. Đào tạo trên NVIDIA GeForce RTX 3090 Ti (12 GB VRAM) trong 10 epochs với batch size = 8 và learning rate = 1e-3 (Adam). Thời gian huấn luyện khoảng 03 giờ cho toàn bộ dataset gồm 4,709 video.

6. Kết quả nghiên cứu

Sau khi hoàn tất quá trình huấn luyện, mô hình được đánh giá bằng cách sử dụng tập kiểm tra tách ra từ bộ dữ liệu ban đầu. Các chỉ số đánh giá được báo cáo bao gồm:

• Độ chính xác tổng thể (Top-1 Accuracy): là chỉ số đo lường tỷ lệ dự đoán đúng của mô hình, phản ánh phần trăm số lượng mẫu được phân loại chính xác vào đúng lớp. Độ chính xác tổng thể là một chỉ số cơ bản, quan trọng để đánh giá hiệu quả tổng quát của mô hình trong các bài toán phân loại.

• Hàm mất mát (Sparse Categorical Cross Entropy loss): là chỉ số đo lường sự khác biệt giữa phân phối xác suất dự đoán của mô hình và nhãn thực tế. Hàm mất mát thấp cho thấy mô hình dự đoán sát với giá trị thực tế, trong khi giá trị cao có thể chỉ ra mô hình chưa học được đặc trưng của dữ liệu.

Kết quả đánh giá: sau quá trình huấn luyện, mô hình đạt được độ chính xác cao trên tập huấn luyện và kiểm tra, trong khi hàm mất mát giảm dần theo số epoch. Điều này cho thấy mô hình có khả năng học tốt từ dữ liệu và tổng quát hóa được trên tập kiểm tra.

Bảng 2

Kết Quả Tốt Nhất của Mô Hình sau 10 Epoch khi Sử Dụng Backbone đã được Pretrained

	Top-1 Accuracy	**Sparse Categorical Cross Entropy Loss**
Epoch = 10	0.8855	0.4181

Ghi chú: Dữ liệu do nhóm nghiên cứu chạy thực nghiệm và ghi nhận

Chúng tôi đã tiến hành phân tích mẫu dữ liệu bị phân loại sai và rút ra một số kết luận chính. Đầu tiên, những cử chỉ phức tạp như "bắt chước" và "cung cấp" thường bị nhầm lẫn do biểu diễn của chúng có nhiều nét tương đồng. Thứ hai, ở những video quay trong điều kiện ánh sáng yếu, chất lượng khung hình giảm đáng kể, khiến độ nhạy của mô hình giảm khoảng 12%. Cuối cùng, góc quay side-view cho kết quả kém hơn so với front-view, với tỷ lệ lỗi cao hơn khoảng 8%.

Để khắc phục các hạn chế này, chúng tôi đề xuất ba hướng cải tiến chính. Thứ nhất, tích hợp các lớp Spatial-Temporal Transformer để nâng cao khả năng phân biệt giữa các cử chỉ tương tự. Thứ hai, kết hợp dữ liệu từ cảm biến chiều sâu (depth-sensor) hoặc kỹ thuật ước lượng tư thế (pose-estimation) nhằm bổ sung thông tin 3D, giảm thiểu nhầm lẫn đối với những cử chỉ có biểu diễn gần giống. Thứ ba, mở rộng bộ dữ liệu huấn luyện bằng cách bổ sung các video với đa dạng điều kiện ánh sáng và nhiều góc quay hơn, nhằm cải thiện khả năng tổng quát hóa của mô hình.

7. Kết luận

Trong nghiên cứu này, chúng tôi đã huấn luyện một mô hình phân loại dựa trên kiến trúc MoViNet-A2 với tập dữ liệu huấn luyện và kiểm tra đã được chuẩn bị trước. Quá trình huấn luyện kéo dài 10 epoch và cho thấy tại epoch cuối cùng, mô hình đạt được độ chính xác Top-1 trên tập kiểm tra là 88.55%. Đồng thời, giá trị hàm mất mát trên tập kiểm tra đạt mức thấp nhất là 0.4181 tại epoch thứ 8. Khoảng cách nhỏ giữa độ chính xác huấn luyện và kiểm tra cho thấy mô hình có khả năng tổng quát hóa tốt.

Tuy nhiên, vẫn còn một số khía cạnh cần được xem xét trong các nghiên cứu tiếp theo. Thứ nhất, việc tối ưu hóa siêu tham số, chẳng hạn như điều chỉnh tốc độ học và kích thước batch, có thể giúp cải thiện hiệu suất của mô hình. Thứ hai, so sánh trực tiếp MoViNet-A2 với các kiến trúc khác sẽ cung cấp cái nhìn đầy đủ hơn về tính ưu việt và hạn chế của phương pháp này trong bài toán nhận diện ký hiệu tiếng Việt. Tóm lại, kết quả thực nghiệm khẳng định hiệu quả cao của MoViNet-A2 cho nhiệm vụ nhận diện ký hiệu tiếng Việt, và các cải tiến tương lai hứa hẹn sẽ nâng cao hơn nữa độ chính xác cũng như khả năng khái quát hóa trên các tập dữ liệu lớn hơn.

TUYÊN BỐ KHÔNG CÓ XUNG ĐỘT LỢI ÍCH

Các tác giả cam kết, tuyên bố không có bất kỳ xung đột lợi ích nào liên quan đến việc công bố bài báo này.

Tài liệu tham khảo

Al-Qurishi, M., Khalid, T., & Souissi, R. (2021). Deep learning for sign language recognition: Current techniques, benchmarks, and open issues. *IEEE Access, 9,* 126917-126951. https://doi.org/10.1109/ACCESS.2021.3110912

Camgoz, N. C., Hadfield, S., Koller, O., & Bowden, R. (2017). *SubUNets: End-to-end hand shape and continuous sign language recognition*. https://doi.org/10.1109/ICCV.2017 .332

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). *Neural sign language translation*. https://doi.org/10.1109/CVPR.2018.00812

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). *Sign language transformer: Joint end-to-end sign language recognition and translation*. https://arxiv.org/abs/2003. 13830

Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large-vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding, 141,* 108-125. https://doi.org/10.1016/j.cviu.2015. 09.013

Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2016). *Deep sign: Hybrid CNN-HMM for continuous sign language recognition.* https://doi.org/10.5244/C.30.136

Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., & Gong, B. (2021). *MoViNets: Mobile video networks for efficient video recognition*. https://doi.org/10.1109/CVPR46437. 2021.01576

Kumari, D., & Anand, R. S. (2024). Isolated video-based sign language recognition using a hybrid CNN-LSTM framework based on attention mechanism. *Electronics, 13*(7), Article 1229. https://doi.org/10.3390/electronics13071229

Molchanov, P., Gupta, S., Kim, K., & Kautz, J. (2015). *Hand gesture recognition with 3D convolutional neural networks*. https://doi.org/10.1109/CVPRW.2015.7301342

Nguyen, D. S., Nguyen, D. T., Tran, T. D., Pham, H. N. D., Tran, H. T., Tong, A. N., Hoang, H. Q., & Nguyen, L. P. (2025). *Sign language recognition: A large-scale multi-view dataset and comprehensive evaluation*. https://openaccess.thecvf.com/content/WACV2025/papers/Dinh_Sign_Language_Recognition_A_Large-Scale_Multi-View_Dataset_and_Comprehensive_Evaluation_WACV_2025_paper.pdf

Pham, H. T., Huynh, T. C., Bui, P. V., & Ha, K. H. (2018). *Automatic feature extraction for Vietnamese sign language recognition using support vector machine*. https://doi.org/10.1109/SIGTELCOM.2018.8325780

Pigou, L., Dieleman, S., Kindermans, P.-J., & Schrauwen, B. (2015). Sign language recognition using convolutional neural networks. In *ECCV 2014 workshops LNCS 8925* (pp. 572-578). https://doi.org/10.1007/978-3-319-16178-5_40

Starner, T. E., Weaver, J., & Pentland, A. (1998). Real-time American sign language recognition using desk and wearable computer-based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(12), 1371-1375. https://doi.org/10.1109/ 34.735811

Vogler, C., & Metaxas, D. N. (2001). A framework for recognizing the simultaneous aspects of American Sign Language. *Computer Vision and Image Understanding, 81*(3), 358-384. https://doi.org/10.1006/cviu.2000.0895

Vu, A. T., Phung, K. V., Hoang, H. Q., & Pham, H. T. V. (2025). Vietnamese sign language alphabet recognition using deep learning and Mediapipe methods. *Journal of Smart Systems and Devices, 35*(1), 10-19. https://doi.org/10.51316/jst.179.ssad.2025.35.1.2