NHẬN DẠNG NGÔN NGỮ KÝ HIỆU TIẾNG VIỆT TRONG VIDEO BẰNG LSTM VÀ I3D ĐA KHỐI

Vũ Hoài Nam*, **Hoàng Mậu Trung***, **Phạm Văn Cường***
*Học Viện Công Nghệ Bưu Chính Viễn Thông

***Tóm tắt*—Ngôn ngữ ký hiệu là một trong những phương tiện không thể thay thế trong giao tiếp hằng ngày của cộng đồng người câm điếc. Ngôn ngữ ký hiệu được biểu diễn bằng cử chỉ phần thân trên của người thể hiện ngôn ngôn ngữ. Với sự phát triển vượt bậc của các công nghệ cao trong lĩnh vực học sâu và thị giác máy tính, hệ thống nhận dạng ngôn ngữ ký hiệu trở thành một cầu nối hiệu quả giữa cộng đồng người câm điếc và thế giới bên ngoài. Nhận dạng ngôn ngữ ký hiệu tiếng Việt (VSLR) là một nhánh của bài toán nhận dạng ngôn ngữ ký hiệu nói chung được sử dụng trong cộng đồng người câm điếc Việt Nam. VSLR hướng đến thông dịch từ cử chỉ của người thực hiện ngôn ngữ ký hiệu sang thành văn bản. Trong bài báo này, chúng tôi đề xuất một phương pháp nhận dạng ngôn ngữ ký hiệu tiếng Việt từ video dựa trên mô hình học sâu. Phương pháp đề xuất bao gồm hai phần chính là mô hình hai luồng mạng nơ ron tích chập (CNN) cho đặc trưng không gian và mạng bộ nhớ dài ngắn (Long-Short Term Memory - LSTM) cho đặc trưng thời gian. Chúng tôi đánh giá mô hình đề xuất với bộ dữ liệu chúng tôi thu thập bao gồm 29 ký tự trong bảng chữ cái tiếng Việt. Thực nghiệm đạt được với độ chính xác 95% chứng minh tính hiệu quả và thực tế của phương pháp đề xuất trong việc nhận dạng ngôn ngữ ký hiệu tiếng Việt.**

***Từ khóa*—Học sâu, nhận dạng, ngôn ngữ ký hiệu.**

I. GIỚI THIỆU

Ngôn ngữ ký hiệu là một ngôn ngữ được phát triển bởi nhu cầu cần thiết trong việc giao tiếp của cộng đồng người khiếm thính. Một quan điểm sai lầm là ngôn ngữ ký hiệu đồng nhất trên toàn thế giới. Trên thực tế tại mỗi quốc gia khác nhau có một bộ ngôn ngữ khác nhau, thậm chí trong cùng một quốc gia tại mỗi khu vực, vùng, miền lại có một bộ ngôn ngữ ký hiệu khác nhau. Chẳng hạn Việt Nam có 3 nhóm ngôn ngữ ký hiệu chính, đó là: ngôn ngữ ký hiệu Hải Phòng, Hà Nội, Thành phố Hồ Chí Minh. Tại Việt Nam cộng đồng người khiếm thính chiếm tổng số 4-5% dân số của cả nước. Bên cạnh đó, hầu hết họ không biết sử dụng ngôn ngữ ký hiệu trong cuộc sống hàng ngày, do đó điều này trở thành rào cản để họ giao tiếp với thế giới bên ngoài. Do đó, việc tất yếu của việc phát triển tập dữ liệu ngôn ngữ kí hiệu tiêu chuẩn và hoàn thiện một hệ thống hỗ trợ giao tiếp cho người khiếm thính tại Việt Nam. Hệ thống nhận dạng ngôn ngữ ký hiệu tự động không chỉ là một cầu nối giữa cộng đồng khiếm thính và thế giới bên ngoài mà chúng còn có vai trò quan trọng trong ứng dụng về rô bốt và hệ thống tương tác người và máy tính. Hơn thế nữa việc hoàn thành nhận dạng ngôn ngữ ký hiệu cũng giúp trẻ em khiếm thính có thể học về nhận thức, xã hội, cảm xúc và ngôn ngữ. Hệ thống nhận dạng ngôn ngữ kí hiệu ghi nhận sự chuyển chộng và phân tích chuyển động của phần trên cơ thể con người. Bởi vậy, có 2 giải pháp chính cho vấn đề trên: tiếp cận theo hướng thị giác máy tính và tiếp cận theo hướng sử dụng cảm biến chuyển động. Phương pháp dựa trên thị giác máy tính sử dụng đầu vào là video, trong khi đó phương pháp còn lại sử dụng tín hiệu thu được từ cảm biến. Trong số hai hướng tiếp cận này, cách tiếp cận dựa trên thị giác máy tính chứng tỏ sự thuận tiện và tự nhiên hơn vì chúng không yêu cầu người khiếm thính phải đeo thiết bị có chứa cảm biến gây khó chịu khi giao tiếp. Cách tiếp cận dựa trên thị giác lấy đầu vào là một loạt các khung hình và phân loại tập các khung hình này thành các từ hoặc ký tự ngôn ngữ ký hiệu tương ứng, tương tự như vấn đề nhận dạng hoạt động video.

Các mô hình học sâu gần đây đã được áp dụng để giải quyết hiệu quả các vấn đề nhận dạng hoạt động

Tác giả liên hệ: Vũ Hoài Nam, email: namvh@ptit.edu.vn
Đến tòa soạn: 20/08/2020, chỉnh sửa: 23/10/2020, chấp nhận đăng: 26/10/2020.

trong video [1], [2], [3]. Đề xuất của chúng tôi tận dụng lợi thế của các cấu trúc mạng học sâu bởi sự kết hợp của I3D [1] và LSTM [4] cho nhận dạng ngôn ngữ ký hiệu tiếng Việt. I3D module được sử dụng để nắm bắt thông tin không gian của chuyển động, còn LSTM module thì lại nắm bắt đặc trưng chuyển động theo theo thời gian. Để xuất của chúng tôi chia tập khung hình đầu vào thành các khối khung hình nhỏ hơn và đưa vào I3D module. Việc chia này dựa trên quan sát hành động mô tả ngôn ngữ ký hiệu trong video được cấu thành bởi nhiều các hành động con rời rạc bao gồm kí tự và dấu thanh. Do đó, việc chia đầu vào thành khối khung hình nhỏ giúp cải thiện độ chính xác của hệ thống.

II. NGHIÊN CỨU LIÊN QUAN

Nhận dạng ngôn ngữ ký hiệu được chia làm hai loại chính: dựa trên dữ liệu cảm biến (sensor-based) và dựa trên thị giác máy tính (vision-based).

A. Phương pháp dựa trên dữ liệu cảm biến

Người khiếm thính phải đeo một hoặc một số thiết bị có gắn các cảm biến khi mô tả các từ ngôn ngữ ký hiệu trong suốt cuộc hội thoại của họ. Bằng cách sử dụng dữ liệu cảm biến này, có thể giúp đơn giản hóa công việc tiền xử lý dữ liệu bởi khả năng lọc nhiễu, và yếu tố phức tạp của môi trường. Bên cạnh đó chuyển động của người khiếm thính không bị giới hạn bởi một ngữ cảnh cụ thể nào như đứng trước một máy thu hình. Trong cách tiếp cận này, tín hiệu từ các cảm biến được truyền không dây đến một thiết bị từ xa để xử lý nhận dạng [5], [6]. Tuy nhiên, với sự phát triển khả năng tính toán của các thiết bị nhúng, một vài hệ thống nhận dạng ngôn ngữ kí hiệu đơn giản có thể chạy trực tiếp trên các thiết bị này chẳng hạn như găng tay điện tử hoặc vòng đeo tay thông minh [7]. Cải tiến này có thể làm cho cách tiếp cận dựa trên cảm biến phù hợp hơn trong các ứng dụng thực tế. Trong một số bài báo, có một số cách tiếp cận được đề xuất để tận dụng nhiều cảm biến để nhận dạng ngôn ngữ ký hiệu. Nhóm tác giả trong [8] đề xuất một phương pháp sử dụng kết hợp các cảm biến gia tốc và cảm biến điện cơ. Các tín hiệu đến từ các cảm biến gia tốc và điện cơ được xử lý trước khi đưa vào bộ phân loại SVM. Theo đề xuất của họ, hệ thống nhận dạng ngôn ngữ kí hiệu có thể đạt được độ chính xác 96,16% trên bộ dữ liệu tự thu thập của họ. Mặc dù các phương pháp tiếp cận dựa trên nhiều cảm biến có thể đạt được độ chính xác tốt hơn nhưng hệ thống trở nên bất tiện hơn cho người thực hiện ngôn ngữ ký hiệu vì họ phải đeo nhiều thiết bị hơn. Hơn thế nữa, cách tiếp cận này không thể nắm bắt được toàn bộ sự thay đổi về hình dạng và chuyển động tương đối của các bộ phận cơ thể.

B. Phương pháp dựa trên thị giác máy tính

Với phương pháp tiếp cận này máy thu hình được sử dụng là công cụ chính giúp ghi lại dữ liệu đầu vào. Lợi thế của sử dụng máy thu hình đó là không cần đeo một thiết bị nào cả và giúp giảm chi phí giá thành của thệ thống. Hơn thế nữa giới hạn góc nhìn của máy thu hình rất lớn giúp cho có thể thu được đồng thời nhiều người trong cuộc hội thoại. Bên cạnh đó ngày này các điện thoại thông minh đều được trang bị máy thu hình với độ phân giải cao đó có thể là một tiềm năng lớn cho dữ liệu đầu vào của hệ thống nhận dạng. Vì thế các tiếp cận dự trên thị giác máy tính cho hệ thống nhận dạng ngôn ngữ kí hiệu khiến cho việc giao tiếp hằng ngày của người khiếm thính tự nhiên hơn và thuật tiện hơn khi sử dụng. Do những lợi ích được đề cập trên, đã có nhiều nhà nghiên cứu tập trung vào đề xuất nhận dạng ngôn ngữ ký hiệu dựa trên thị giác bằng nhiều ngôn ngữ khác nhau như ngôn ngữ ký hiệu của Mỹ [9], [10], [11], ngôn ngữ ký hiệu Trung Quốc [12], ký hiệu Hàn Quốc ngôn ngữ [13] và ngôn ngữ ký hiệu Việt Nam [14], [15]. Trong [11], tác giả đã nghiên cứu hai kỹ thuật trích xuất tính năng mới của Combined Orient Histogram and Statistical and Wavelet feature để nhận dạng ngôn ngữ kí hiệu Mỹ các số từ 0-9. Các đặc trưng được kết hợp lại và được đưa vào một mạng nơ ron để huấn luyện. Tác giả của [12] triển thực hiện nắm bắt thông tin cả 2 chiều không gian và thời gian trong mô hình phân loại ngôn ngữ kí hiệu Trung Quốc. Đầu tiên một mô hình trích đặc trưng của ngôn ngữ kí hiệu được thực hiện, các đặc trưng là đầu vào của bộ phân loại SVM để nhận dạng 30 loại của bảng chữ cái Trung Quốc. Kết quả của họ cho thấy Linear kernel SVM là bộ phân loại phù hợp nhất với nhận dạng ngôn ngữ kí hiệu. Để nhận dạng ngôn ngữ kí hiệu Việt Nam, tác giả của [14] được sử dụng mô tả địa phương. Trong mô đun trích chọn đặc trưng, họ trích xuất đặc trưng không gian và đặc trưng ngữ cảnh để mô tả từ ngữ trong ngôn ngữ ký hiệu. Sau đó một tập các đặc trưng được học bởi bộ phân loại SVM. Đánh giá trên tập dữ liệu của họ cho kết quả đạt được độ chính xác là 86,61%. Từ cách tiếp cận thị giác máy tính, nhận dạng ngôn ngữ ký hiệu được xem là một nhánh của nhận dạng hành

động với hạn chế chuyển động của một số bộ phận trên cơ thể. Có một xu hướng trong cộng đồng nhận dạng ngôn ngữ ký hiệu trong đó các nhà nghiên cứu đang cố gắng thay thế các đặc trưng thủ công bằng mô hình học sâu để cải thiện độ chính xác và độ tin cậy. [15] đã sử dụng CNN-LSTM cho nhận dạng ngôn ngữ kí hiệu Việt Nam. Kết quả của họ đã cho thấy rằng phương pháp học sâu có kết quả vượt trội so với phương pháp truyền thống. Tác giả [13] đã phát triển một hệ thống nhận dạng ngôn ngữ kí hiệu Hàn Quốc dựa trên mạng nơ-ron tích chập CNN từ đầu vào là các video. Tập dữ liệu của họ bao gồm 10 từ được chọn trong ngôn ngữ kí hiệu Hàn Quốc. Phương pháp của họ đạt độ chính xác 84,5% Tắc giả của [16] đã xuất một phướng pháp kết hợp hai kĩ thuật mạnh nhất của học sâu là CNN trích đặc trưng không gian và LSTM trích đặc trưng thời gian. Kết quả hệ thống của họ được đánh giá trên tập dữ liệu gồm 40 từ vựng thông dụng hằng ngày. Đánh giá của họ chỉ ra rằng mô hình dựa trên CNN-LSTM có thể được thực thi trong thời gian thực cho các ứng dụng thực tế. Trong [17], việc nhúng CNN từ đầu đến cuối vào mô hình Markov ẩn (HMM) đã được giới thiệu. CNN-HMM lai tận dụng khả năng phân biệt đối xử mạnh mẽ của CNN và khả năng mô hình hóa trình tự của HMM. Phương pháp được đề xuất của họ có thể nhận ra ngôn ngữ ký hiệu liên tục đạt tỷ lệ lỗi lần lượt là 30% và 32,5% trên bộ dữ liệu Phoenix 2012 [18] và bộ dữ liệu Phoenix 2014 [19].

III. PHƯƠNG PHÁP ĐỀ XUẤT

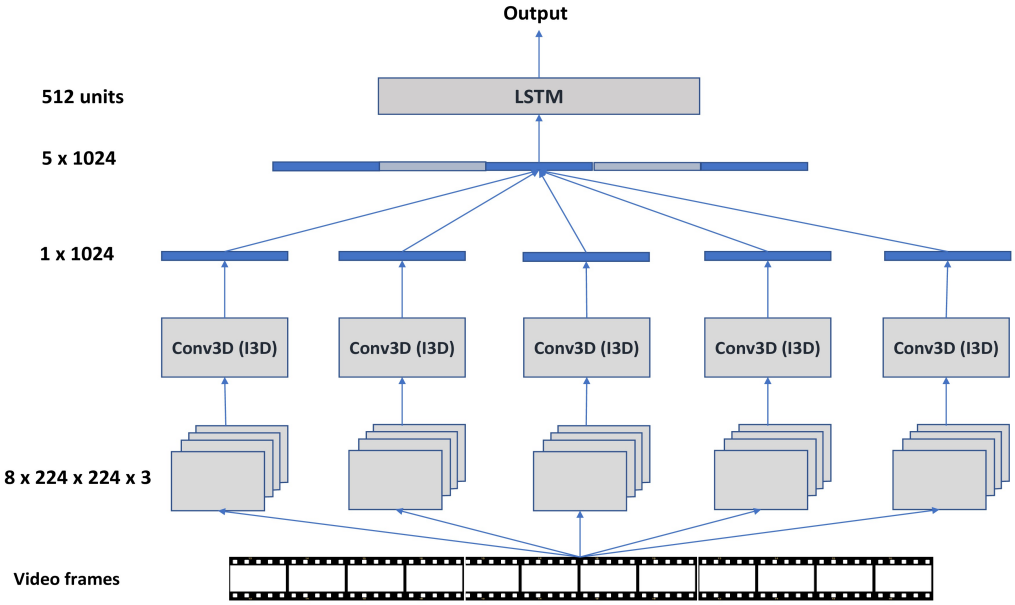
Đề xuất của chúng tôi được mô tả trong Hình 1 bao gồm 2 phần chính: mô hình I3D để trích rút đặc trưng về mặt không gian và mô hình LSTM để trích rút đặc trưng về mặt thời gian. Đầu vào là từng khung hình được lấy ra từ video, chúng tôi chia tập khung hình thành các khối con. Sau đó với mỗi khối sẽ trở thành đầu vào của một mô đun I3D, số lượng mô đun I3D bằng số lượng khối khung hình con. Trong bài báo này chúng tôi tối ưu số lượng các khối con đầu vào dựa trên kết quả thực nghiệm trên các bộ cơ sở dữ liệu. Độ dài của mỗi khối video con sẽ ảnh hưởng đến số lượng của các khối sau khi được cắt nhỏ. Trong thực tế, nếu mô hình này được đưa ra để nhận dạng hành động trong video nói chung thì sẽ cho độ hiệu quả không cao. Tuy nhiên với bài toán nhận dạng ngôn ngữ ký hiệu, các hành động của người thực hiện ngôn ngữ ký hiệu là tập hợp của rất nhiều hành động nhỏ của tay và cảm xúc trên khuôn mặt, những hành động nhỏ này sẽ xuất hiện trong các video của những nhóm khác khi người đó thực hiện từ khác. Do vậy, lấy động lực từ phương pháp túi từ (Bag of word), nhóm nghiên cứu đề xuất có thể chia video của từng từ trong ngôn ngữ ký hiệu thành các video nhỏ hơn để có thể áp dụng hiệu quả trong bài toán nhận dạng ngôn ngữ ký hiệu này. Mỗi hành động Đầu ra của mô đun I3D là vector đặc trưng 1024 chiều, sau đó được đưa qua các lớp LSTM để phân loại thành các nhóm ngôn ngữ kí hiệu.

A. I3D

I3D được đề xuất để giải quyết vấn đề cho nhận dạng hành động con người (Human Activity Recognition - HAR). I3D sử dụng Inception V1 được đào tạo trước để thực hiện học tập chuyển đổi từ bộ dữ liệu ImageNet sang bộ dữ liệu video hoạt động của con người. Các hạt nhân của mạng Inception V1 [20] truyền thống được mở rộng thành các hình dạng 3 chiều (3D) để phù hợp với dữ liệu đầu vào của chuỗi khung. Thành công của mô hình I3D dựa trên quan sát rằng không có bộ dữ liệu HAR nào có sẵn lớn như ImageNet. Trong tài liệu, các mô hình mạng nơ ron nhân chập 3 chiều (3DCNN) được sử dụng cho các vấn đề phân loại video là các mô hình nông vì thiếu dữ liệu. Mô hình của chúng tôi sử dụng mô hình I3D được đào tạo trước để tinh chỉnh với tập dữ liệu của chúng tôi. Mô hình I3D được đào tạo trước phù hợp với các vấn đề phân loại video HAR ngắn vì nó không chỉ nắm bắt thông tin không gian một cách hoàn hảo mà còn tìm hiểu các đặc điểm tạm thời của các hoạt động cục bộ. Tuy nhiên, áp dụng mô hình I3D trực tiếp vào bộ dữ liệu ngôn ngữ ký hiệu là không hiệu quả vì video ngôn ngữ ký hiệu chứa một số hành động phụ trong video thời lượng dài. Do đó, thay vì áp dụng I3D trực tiếp để nhận dạng ngôn ngữ ký hiệu, chúng tôi chia khung đầu vào thành các khối phụ để lấy đầu vào cho lớp I3D. Sau đó, đầu ra của lớp I3D được chuyển cho các lớp LSTM để khai phá các đặc trưng toàn cục.

B. LSTM

LSTM là một trong những biến thể nổi tiếng nhất của mô hình mạng thần kinh hồi quy (Recurrent Neural Network - RNN) để giải quyết vấn đề của mô hình dữ liệu biến đổi theo thời gian. Ý tưởng chính chính của RNN là sử dụng trực tiếp thông tin tuần tự. Mô hình RNN thực hiện cùng một nhiệm vụ cho mọi phần tử của chuỗi, với đầu ra phụ thuộc vào các tính toán trước đó. Ngoài ra, mô hình RNN có



Hình 1. Sơ đồ khối phương pháp đề xuất.

thể nắm bắt thứ tự dữ liệu chuỗi thời gian để dự đoán chính xác đầu ra. Tuy nhiên RNN gặp phải hai vấn đề đó là vanishing gradient và exploding gradient. Vanishing gradient xảy ta khi sự đóng góp không đáng kể thông tin cho gradient của các bước thời gian xảy ra trước đó. Do đó mô hình càng sâu thì càng khó đào tạo. Exploding gradient xảy ra khi bùng nổ thông tin của các bước thời gian trước đó dẫn đến sự tích lũy gradient, dẫn đến cập nhật rất lớn cho trọng số của mô hình trong quá trình huấn luyện. LSTM là một trong những đề xuất được đưa ra để giải quyết các nhược điểm của RNN. Một tế bào LSTM được mổ tả trong Hình 2 bao gồm cổng đầu vào i_t cổng đầu ra o_t, và cổng quên f_t. Với thiết kế gồm 3 cổng như vậy LSTM có khả năng giải quyết vấn đề phụ thuộc dài hạn mà mô hình RNN không thể học được. Trong một bài viết, LSTM vượt trội hơn RNN trong vấn đề liên qua đến dữ liệu thay đổi theo chuỗi thời gian. Đạo hàm công thức cụ thể của LSTM được minh họa trong Công thức (1) - (11). Trong phương pháp đề xuất của chúng tôi, lớp LSTM được xếp chồng lên nhau sau các mô-đun I3D để tìm hiểu mối quan hệ giữa hành động phụ trong các video ngôn ngữ ký hiệu. Đầu ra của các tế bào LSTM là trạng thái của tế bào đó (c_t) và trạng thái ẩn (h_t). Đầu vào của các tế bào LSTM là trạng thái tế bào trước đó (c_{t-1}), trạng thái ẩn trước đó (h_{t-1}) và đầu vào của trạng thái thứ i (x_t).

$$sigmoid(x) = \frac{1}{1+e^{-x}} \quad (1)$$

$$tanh(x) = \frac{e^{2x}-1}{e^{2x}+1} \quad (2)$$

$$f_t = sigmoid(U_f * x_t + W_f * h_{t-1} + b_f) \quad (3)$$

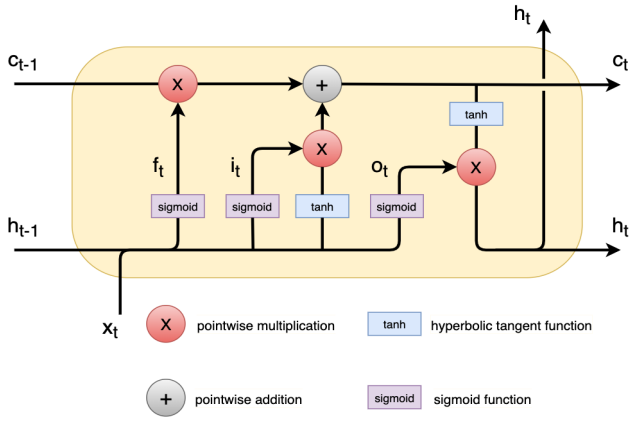
$$i_t = sigmoid(U_i * x_t + W_i * h_{t-1} + b_i) \quad (4)$$

$$o_t = sigmoid(U_o * x_t + W_o * h_{t-1} + b_o) \quad (5)$$

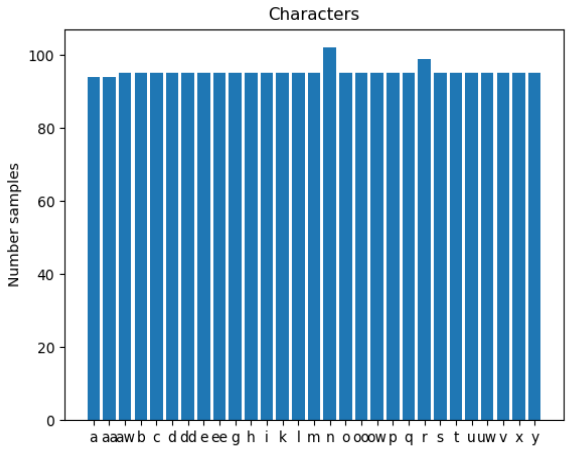
$$c_t = f_t * c_{t-1} + i_t * tanh(U_c * x_t + W_c * h_{t-1} + b_c) \quad (6)$$

$$h_t = o_t * tanh(c_t) \quad (7)$$

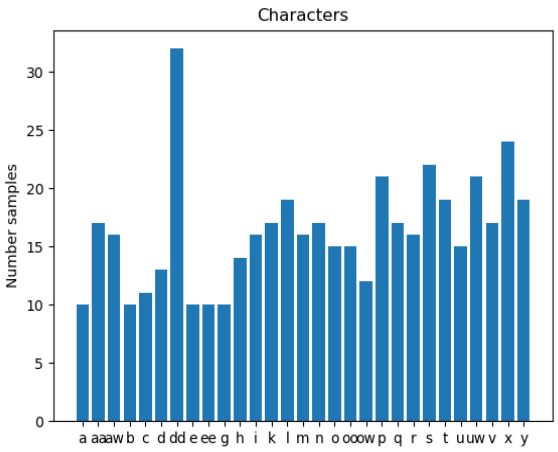
Trong đó U_f, U_i, U_o, U_c lần lượt là các tham số đầu vào ; W_f, W_i, W_o, W_c lần lượt là các tham số hồi quy; b_f, b_i, b_o, b_c lần lượt là các tham số độ lệch;



Hình 2. Kiến trúc của LSTM.



Hình 3. Phân bố mẫu huấn luyện.



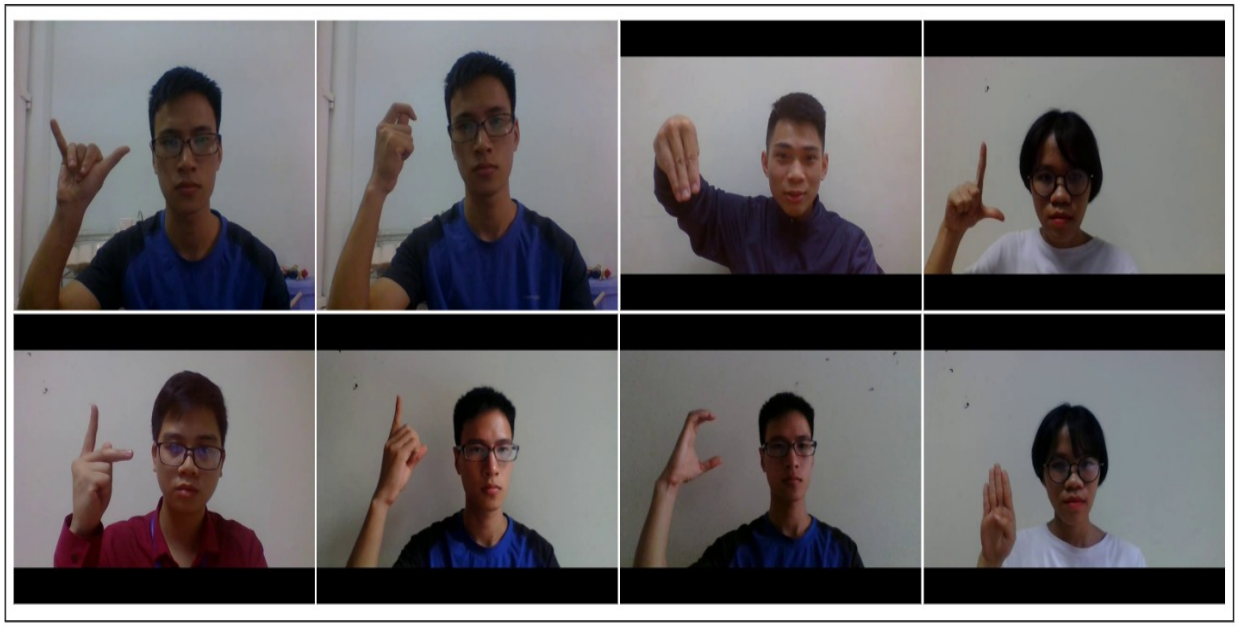
Hình 4. Phân bố mẫu kiểm tra.

C. Chiến lược chia khối con.

Đặc điểm khác biệt của phương pháp chúng tôi đề xuất là phương pháp phân chia khối con. Phương pháp này được bắt nguồn từ việc quan sát rằng mỗi ký tự trong ngôn ngữ ký hiệu được biểu diễn đã kết hợp một loạt các hành động con. Do đó việc phân đoạn video thành các đoạn nhỏ cho kết quả tốt hơn, khi mà, mô hình có khả năng tìm hiểu và mô hình hóa mối quan hệ giữa các hành động phụ với nhau. Do đó, chúng tôi chia đầu vào video thành các khối con kích thước bằng nhau. Sau đó, các khối con này là đầu vào của I3D và LSTM như trong Hình 1. Độ dài của khối con là một tham số quan trọng cần được chọn cẩn thận. Sự lựa chọn sai của tham số này có thể làm giảm đáng kể độ chính xác của phương pháp được đề xuất. Tuy nhiên, kích thước các khối con được cố định để áp dụng vào trong các trường hợp thức tế. Trong phần kết quả thử nghiệm, chúng tôi đã triển khai hệ thống với các độ dài khác nhau để có được độ dài tối ưu.

D. VSLB-C: Bộ dữ liệu ngôn ngữ ký hiệu tiếng Việt ở mức độ ký tự

Hệ thống bảng chữ cái tiếng Việt giống với hệ thống bảng chữ cái tiếng Anh hơn là bảng chữ cái như hệ thống ký hiệu của Trung Quốc, Nhật Bản và Hàn Quốc. Tuy nhiên Tiếng Việt thậm chí còn phức tạp hơn tiếng Anh vì đặc điểm âm sắc của chúng bao gồm sáu âm khác nhau và ba dấu phụ. Do đó, việc xây dựng bộ dữ liệu mới là cần thiết để nghiên cứu việc nhận dạng ngôn ngữ ký hiệu tiếng Việt trong video. Trong bài báo này, chúng tôi đã thu thập một bộ dữ liệu bao gồm tất cả chữ cái tiếng Việt trong từ vựng ngôn ngữ ký hiệu tiếng Việt. Trong quy trình thu thập dữ liệu này, người tham gia được yêu cầu thực hiện các cử chỉ ngôn ngữ ký hiệu trước máy thu hình. Bên cạnh đó, người tham gia được tự do mặc các loại quần áo khác nhau như trong Hình 5 . Mỗi người tham gia được yêu cầu thực hiện đầy đủ 29 ký tự trong bảng chữ cái ngôn ngữ ký hiệu tiếng Việt. Mỗi người thực hiện được ghi lại nhiều lần với các góc và khoảng cách khác nhau từ người tham gia và máy thu hình. Kết quả là bộ dữ liệu này bao gồm tổng cộng 3248 video. Chúng tôi chia dữ liệu thành phần huấn luyện và phần thử nghiệm. Tổng số video cho mỗi phần được chi tiết trong Hình 3 và Hình 4. Tổng số video cho mỗi người tham gia trong phần huấn luyện gần như bằng nhau. Trong khi tổng số video cho mỗi người tham gia trong phần thử nghiệm là



Hình 5. Ảnh mẫu từ tập dữ liệu video.

khác nhau đáng kể. Chiến lược chia tách này làm cho quá trình huấn luyện hiệu quả hơn nhưng đảm bảo tính khách quan của hệ thống. Các tham số huấn luyện của phương pháp đề xuất của chúng tôi được thể hiện trong Bảng I và Bảng II. Tổng số tham số có thể huấn luyện là khoảng 17 triệu. Để với quá trình huấn luyện hiệu quả, tỷ lệ học của chúng tôi được điều chỉnh ở số lượng epoch khác nhau. Trình tối ưu hóa của chúng tôi sử dụng là Stochastic Gradient Descent, trong khi hàm mất mát là cross entropy.

Bảng I
CÁC THAM SỐ CỦA MÔ HÌNH ĐỀ XUẤT

Parameters	Value	Notes
Input shape	5 blocks x 8 frames x 224 x 224 x 3	RGB image
Output I3D	1024 dimensions	
Output model	29 classes	
Epoch	40	
Batch size	16	
Learning rate	1e-2	Epoch <= 10
Learning rate	1e-3	10 < Epoch < 20
Learning rate	5*1e-4	Epoch >= 20
Optimizer	SGD	Decay = 1e-6
Loss function	Cross entropy	

Kết quả của quá trình huấn luyện được thể hiện trong Hình 6 và Hình 7. Giá trị mất mát và độ chính xác của quá trình huấn luyện có xu hướng dao động mạnh trong những epoch đầu tiên, sau đó ổn định dần dần trong những epoch sau này. Nếu độ mất mát và độ chính xác không ổn định trong quá trình huấn luyện, điều này cho thấy không có dấu hiệu hội tụ, thì mô hình đề xuất không phù hợp với tập dữ liệu. Mô hình đề xuất của chúng tôi có xu hướng hội tụ đến giá trị tối ưu sau 20 epoch. Kết quả này cũng cho thấy mô hình hoạt động hiệu quả trên bộ dữ liệu kiểm tra và xác nhận hợp lệ. Quá trình huấn luyện của chúng tôi dừng lại sau 40 epoch.

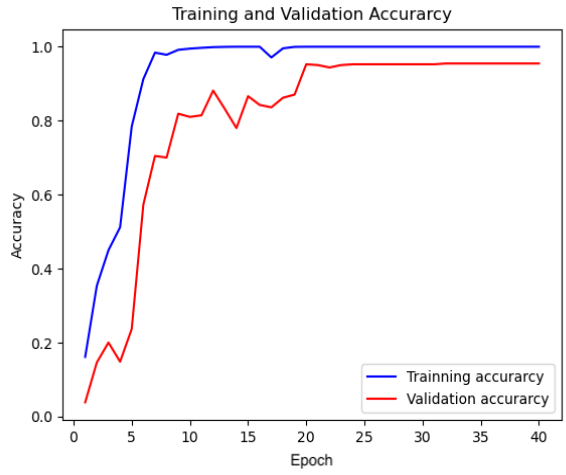
Bảng II
SỐ LƯỢNG TRỌNG SỐ HUẤN LUYỆN ĐƯỢC CỦA MÔ HÌNH ĐỀ XUẤT

Layer	Output shape	No. of param
Time distributed	(None, 5, 1024)	13,344,144
LSTM	(None, 5, 512)	3,147,776
LSTM	(None, 128)	328,192
Dropout	(None, 128)	
Dense	(None, 29)	3,741
Total Params: 16,823,853		

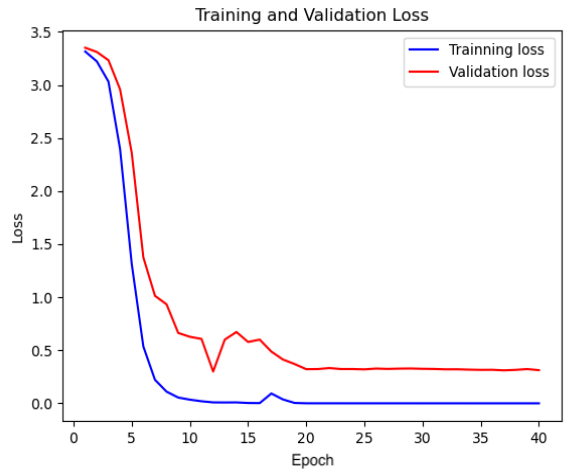
IV. KẾT QUẢ THỰC NGHIỆM

A. Đánh giá độ chính xác

Đối với 29 ký tự, cách tiếp cận của chúng tôi phải phân loại các video đầu vào thành 29 nhãn khác nhau. Chúng tôi đánh giá các mô hình bằng cách sử dụng độ đo F1, xem xét phân loại chính xác của từng lớp quan trọng như nhau. Chiến lược chia tách theo khối được mô tả trong phần trước. Từ kết quả



Hình 6. Biến đổi độ chính xác trong quá trình huấn luyện



Hình 7. Biến đổi hàm mất mát trong quá trình huấn luyện

trong Bảng III, chúng ta có thể thấy rằng phương pháp được đề xuất của chúng tôi đạt được chỉ số F1 cao hơn so với phương pháp cơ sở sử dụng mạng I3D tiêu chuẩn là phương pháp tốt nhất hiện tại và các phương pháp khác như CNN1D kết hợp LSTM và 3DCNN. Kết quả này có thể được giải thích bởi thực tế là mỗi hoạt động từ video đầu vào bao gồm một vài hoạt động phụ. Do đó, mô hình của chúng tôi tìm ra được cơ chế phân chia theo khối hiệu quả cho thấy hiệu suất tốt hơn. Do đó, điểm F1 cho việc sử dụng mạng I3D chỉ là 89,2% trong khi con số này cho phương pháp được đề xuất của chúng tôi đạt 92,3%. Ma trận sai số chi tiết của mô hình phân loại được đề xuất được đưa ra trong Hình 8. Như được hiển thị trong ma trận sai số, hầu hết các ký tự cụ thể đều có thể được phân loại chính xác, ngoại trừ một vài ký tự rất giống nhau trong biểu diễn ngôn ngữ ký hiệu như u và ô, m và n, l và đ.

Bảng III
KẾT QUẢ SO SÁNH

Method	F1 score
Standard I3D	89.2
CNN1D+LSTM	87.6
3DCNN	86.2
Our proposed method	**92.3**

B. Thử nghiệm thực tế

Trong thực nghiệm này, chúng tôi cũng tích hợp mô hình vào ứng dụng trong thế giới thực khi một cá nhân muốn giao tiếp với người câm điếc. Họ thực hiện các hoạt động ngôn ngữ ký hiệu trước một máy thu hình. Trong tiếng Việt, giống như các ngôn ngữ Latinh khác, một từ là sự kết hợp một tập hợp các ký tự. Từ quan điểm này, chúng tôi xây dựng một ứng dụng dựa trên web để người dùng nhập một loạt ký tự ngôn ngữ ký hiệu. Nếu người dùng muốn nói "tôi". họ sẽ nhập t, oo, i bằng tiếng Việt theo thứ tự (tôi). Các thí nghiệm cũng cho thấy hệ thống có thể hoạt động trong miền thời gian thực. Thời gian xử lý để xác định một ký tự riêng lẻ là khoảng 200 mili giây với cạc màn hình GTX 1070 TI .

V. KẾT LUẬN

Bằng cách so sánh độ chính xác của mô hình được đề xuất với I3D tiêu chuẩn, mô hình của chúng tôi cho kết quả cao hơn, nhưng độ phức tạp tính toán tương tự như I3D tiêu chuẩn. Để mô hình được triển khai trong thực tế, bộ sưu tập cơ sở dữ liệu cần thêm một số ký tự n Unicode để mã hóa sáu âm và ba dấu phụ trong ngôn ngữ ký hiệu tiếng Việt. Nếu một ký tự được đặt thành chuyển đổi câu là cần thiết, ký tự "khoảng trắng" cũng phải được thêm vào cơ sở dữ liệu. Vào thời điểm đó, nhóm nghiên cứu của chúng tôi sẽ tham khảo ý kiến các chuyên gia ngôn ngữ ký hiệu của Việt Nam để liên kết hoạt động ngôn ngữ ký hiệu liên quan đến kí hiệu "khoảng trắng". Mô hình đề xuất có thể được sử dụng để xây dựng một từ điển cho cả cộng đồng người câm điếc và những người khác. Một thử nghiệm thực nghiệm được tiến hành để xác minh phương pháp được đề xuất của chúng tôi, dựa trên cơ sở dữ liệu VSLB-C. Kết quả đánh giá đã chứng minh tính khả thi của việc nhận biết ngôn ngữ ký hiệu tiếng Việt. Công việc trong

Ground-truth/ Predict	a	aa	aw	b	c	d	dd	e	ee	g	h	i	k	l	m	n	o	oo	ow	p	q	r	s	t	u	uw	v	x	y
a	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
aa	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
aw	1	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
dd	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ee	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
i	0	0	0	0	1	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
l	0	0	0	0	0	0	3	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	1	0	0	0	0	0	0
o	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0
oo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	3	0	0	0	0	0
ow	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0
p	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0
q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	1	0	0	0
s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0
u	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0
uw	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	18	0	0	0
v	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0
x	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0
y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19

Hình 8. Ma trận sai số của phương pháp đề xuất.

tương lai nên điều tra các mô hình phân cấp sâu để học tập hiệu quả hơn và xây dựng cơ sở dữ liệu ngôn ngữ ký hiệu dựa trên tiếng Việt để giao tiếp thuận tiện hơn giữa người câm điếc và người khác.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi chương trình học bổng trong nước của Quỹ đổi mới của tập đoàn VinGroup mã số: VINIF.2019.TS.41.

TÀI LIỆU THAM KHẢO

[1] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299-6308. 2017.

[2] Hong, Jongkwang, Bora Cho, Yong Won Hong, and Hyeran Byun. "Contextual Action Cues from Camera Sensor for Multi-Stream Action Recognition." Sensors 19, no. 6 (2019): 1382.

[3] Wang, Xianyuan, Zhenjiang Miao, Ruyi Zhang, and Shanshan Hao. "I3D-LSTM: A New Model for Human Action Recognition." In IOP Conference Series: Materials Science and Engineering, vol. 569, no. 3, p. 032035. IOP Publishing, 2019.

[4] Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM." (1999): 850-855.

[5] Das, Abhinandan, Lavish Yadav, Mayank Singhal, Raman Sachan, Hemang Goyal, Keshav Taparia, Raghav Gulati, Ankit Singh, and Gaurav Trivedi. "Smart glove for Sign Language communications." In 2016 International Conference on Accessibility to Digital World (ICADW), pp. 27-31. IEEE, 2016.

[6] Praveen, Nikhita, Naveen Karanth, and M. S. Megha. "Sign language interpreter using a smart glove." In 2014 International Conference on Advances in Electronics Computers and Communications, pp. 1-5. IEEE, 2014.

[7] Dai, Qian, Jiahui Hou, Panlong Yang, Xiangyang Li, Fei Wang, and Xumiao Zhang. "The sound of silence: end-to-end sign language recognition using smartwatch." In Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking, pp. 462-464. 2017.

[8] Wu, Jian, Lu Sun, and Roozbeh Jafari. "A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors." IEEE journal of biomedical and health informatics 20, no. 5 (2016): 1281-1290.

[9] Starner, Thad, Joshua Weaver, and Alex Pentland. "Real-time american sign language recognition using desk and wearable computer based video." IEEE Transactions on pattern analysis and machine intelligence 20, no. 12 (1998): 1371-1375.

[10] Zafrulla, Zahoor, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti. "American sign language recognition with the kinect." In Proceedings of the 13th international conference on multimodal interfaces, pp. 279-286. 2011.

[11] Thalange, Asha, and S. K. Dixit. "COHST and wavelet features based Static ASL numbers recognition." Procedia Computer Science 92 (2016): 455-460.

[12] Yang, Quan. "Chinese sign language recognition based on video sequence appearance modeling." In 2010 5th IEEE Conference on Industrial Electronics and Applications, pp. 1537-1542. IEEE, 2010.

[13] Shin, Hyojoo, Woo Je Kim, and Kyoung-ae Jang. "Korean sign language recognition based on image and convolution neural network." In Proceedings of the 2nd International Conference on Image and Graphics Processing, pp. 52-55. 2019.

[14] Vo, Anh H., Nhu TQ Nguyen, Ngan TB Nguyen, Van-Huy Pham, Ta Van Giap, and Bao T. Nguyen. "Video-Based Vietnamese Sign Language Recognition Using Local Descriptors." In Asian Conference on Intelligent Information and Database Systems, pp. 680-693. Springer, Cham, 2019.
[15] Vo, Anh H., Van-Huy Pham, and Bao T. Nguyen. "Deep Learning for Vietnamese Sign Language Recognition in Video Sequence."International Journal of Machine Learning and Computing 9, no. 4 (2019).
[16] Yang, Su, and Qing Zhu. "Continuous Chinese sign language recognition with CNN-LSTM." In Ninth International Conference on Digital Image Processing (ICDIP 2017), vol. 10420, p. 104200F. International Society for Optics and Photonics, 2017.
[17] Koller, Oscar, Sepehr Zargaran, Hermann Ney, and Richard Bowden. "Deep sign: enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs." International Journal of Computer Vision 126, no. 12 (2018): 1311-1325.
[18] Forster, Jens, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H. Piater, and Hermann Ney. "RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus." In LREC, vol. 9, pp. 3785-3789. 2012.
[19] Cihan Camgoz, Necati, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. "Neural sign language translation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7784-7793. 2018.
[20] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. 2015.

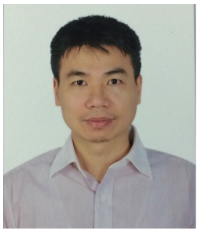
VIETNAMESE SIGN LANGUAGE RECOGNITION IN VIDEO BY MULTI-BLOCK I3D AND LSTM

Abstract: Sign language is an irreplaceable means in the daily communication of the deaf-mute community. Sign language is represented by the gesture of the upper body part. With the development of advanced technology, the Sign language recognition system has become an effective bridge between the deaf-mute community with the outside world. Vietnamese sign language recognition (VSLR) is a branch of sign language recognition used by the community of Vietnamese deaf-mute people. VSLR aims to correctly interpret the gestures in sign language into their corresponding text. In this paper, we propose a method for identifying sign language from videos based on deep learning framework. The proposed method includes two main parts which are two streams convolutional neural network (CNN) for the spatial features and long-short term memory (LSTM) network for the temporal features. We evaluated the framework with our acquired dataset including 29 Vietnamese alphabets, 5 tone marks, and a space symbol. The experiments achieved satisfactory results of 95% F1 score which proves the feasibility and applicability of the proposed approach.

Keywords: Vietnamese sign language, video recognition, deep learning

Vu Hoai Nam nhận bằng kỹ sư Điện tử Viễn thông tại Đại học Bách Khoa Hà Nội năm 2013 và bằng thạc sỹ Khoa học Máy tính tại Đại học Quốc gia Chonnam, Hàn Quốc năm 2015. Hiện tại, Thạc sỹ Nam đang là nghiên cứu sinh nghành Khoa học Máy tính tại Học viện Công nghệ Bưu chính Viễn thông. Từ năm 2016, thạc sỹ Nam là giảng viên bộ môn Khoa học máy tính, Học viện Công nghệ Bưu chính Viễn thông. Hướng nghiên cưu của thạc sỹ Nam bao gồm xử lý ảnh UAV, học máy, và học sâu.

Hoang Mau Trung là sinh viên đại học nghành Khoa học máy tính, Học viện Công nghệ Bưu chính Viễn thông. Hướng nghiên cứu chính của Trung là xử lý ảnh và học sâu.



Pham Van Cuong là Phó giáo sư nghành Khoa học máy tính tại Học viện Công nghệ Bưu chính Viễn thông (PTIT). Trước khi tham gia giảng dạy tại Học viện, Phó giáo sư Cường là nghiên cứu viên chính tại trung tâm nghiên cứu phát triển của Philips tại Hà Lan. Phó giáo sư Cường nhận bằng cử nhân Khoa học máy tính tại Đại học Quốc gia Hà Nội năm 1998, và nhận bằng Thạc sỹ nghành Khoa học máy tính tại Đại học New Mexico, Mỹ năm 2005. Phó giáo sư Cường nhận bằng Tiến sỹ tại Đại học Newcastle, Anh năm 2012. Hướng nghiên cứu chính của Phó giáo sư Cường là tính toán khắp nơi, tính toán trên các thiết bị đeo dán, nhận dạng hoạt động người và học sâu.