# RECOGNIZING VIETNAMESE SIGN LANGUAGE USING DEEP NEURAL NETWORKS

**Nguyen Quang Duy, Luong Thai Le***
*University of Transport and Communications*

**ARTICLE INFO**

**Received:** 29/4/2025
**Revised:** 26/6/2025
**Published:**

**KEYWORDS**

Vietnamese sign language
Convolutional neural network
Long short-term memory
Attention mechanism
Computer vision

**ABSTRACT**

Vietnamese sign language plays a pivotal role in enabling effective communication among deaf and hard-of-hearing communities throughout Vietnam. In this study, we propose a deep learning-based recognition system that leverages MediaPipe to accurately extract hand landmarks from video sequences. These landmarks are then processed by an architecture, either a convolutional neural network or a long short-term memory network enhanced with an attention mechanism (such as additive or multi-head attention), to selectively highlight salient temporal patterns in sign gestures. To support robust training and evaluation, we compiled and meticulously annotated a comprehensive dataset of Vietnamese sign language gestures. Experimental results demonstrate that the proposed model attains a remarkable recognition accuracy of 99.51%, outperforming baseline approaches. The system's real-time performance and high precision highlight its potential as the basis for practical assistive communication tools, paving the way for further research in sign language processing and cross-cultural gesture recognition applications within the Vietnamese context.

# NHẬN DIỆN NGÔN NGỮ KÍ HIỆU TIẾNG VIỆT SỬ DỤNG MẠNG HỌC SÂU

**Nguyễn Quang Duy, Lương Thái Lê***
*Trường Đại học Giao thông Vận tải*

**THÔNG TIN BÀI BÁO**

**Ngày nhận bài:**
**Ngày hoàn thiện:**
**Ngày đăng:**

**TỪ KHÓA**

Ngôn ngữ kí hiệu Việt Nam
Mạng tích chập
Mạng bộ nhớ ngắn-dài hạn
Cơ chế chú ý
Thị giác máy tính

**TÓM TẮT**

Ngôn ngữ ký hiệu Việt Nam đóng vai trò thiết yếu trong việc tạo điều kiện giao tiếp hiệu quả cho cộng đồng người điếc và khiếm thính trên khắp lãnh thổ Việt Nam. Trong nghiên cứu này, chúng tôi đề xuất một hệ thống nhận dạng dựa trên học sâu, tận dụng thư viện MediaPipe để trích xuất chính xác các điểm mốc bàn tay từ chuỗi phim. Các điểm mốc này sau đó được đưa vào một kiến trúc mạng nơ-ron, có thể là mạng nơ-ron tích chập hoặc mạng nơ-ron với bộ nhớ ngắn hạn-dài hạn được trang bị cơ chế chú ý (bao gồm chú ý gia tính hoặc chú ý đa đầu) nhằm tập trung chọn lọc các mẫu thời gian nổi bật trong các cử chỉ ký hiệu. Để hỗ trợ quá trình huấn luyện và đánh giá độ chính xác, chúng tôi đã biên soạn và chú thích tỉ mỉ một tập dữ liệu đầy đủ về các động tác ký hiệu Việt Nam. Kết quả thực nghiệm cho thấy mô hình đề xuất đạt độ chính xác lên tới 99,51%, vượt trội so với các phương pháp cơ sở. Khả năng vận hành theo thời gian thực cùng độ chính xác cao của hệ thống nhấn mạnh tiềm năng ứng dụng trong các công cụ trợ giúp giao tiếp, đồng thời mở ra hướng nghiên cứu sâu hơn về xử lý ngôn ngữ ký hiệu và ứng dụng nhận dạng cử chỉ đa văn hóa trong bối cảnh Việt Nam.

**DOI: https://doi.org/10.34238/tnu-jst.12708**

* Corresponding author. *Email: luongthaile80@utc.edu.vn*

## 1. Introduction

The rapid evolution of deep learning has significantly advanced sign language recognition, enabling more effective communication tools for the deaf community. Recent research has demonstrated the effectiveness of integrating computer vision frameworks such as MediaPipe [1] with a convolutional neural network (CNN) to achieve high recognition accuracies for hand sign images, as evidenced by Kumar et al. [2], who reported accuracies exceeding 90% for American sign language (ASL) gestures. This success underscores the potential of deep neural networks in addressing the complexities of hand sign recognition tasks across various sign languages.

In parallel, skeleton-based deep learning approaches have gained traction for their efficiency in modeling human pose using sparse keypoint landmarks rather than dense pixel arrays. Yan et al. [3] introduced spatial–temporal graph convolutional networks (ST-GCN), which treats human joints as graph nodes connected both spatially and temporally, yielding state-of-the-art performance on standard action recognition benchmarks. Shi et al. [4] further proposed the two-stream adaptive graph convolutional network (2s-AGCN), which learns graph topologies end-to-end and incorporates both joint and bone information for enhanced accuracy. In the sign language domain, C. C. De Amorim et al. [5] applied ST-GCN to full-body skeleton sequences for isolated sign recognition, achieving 85% accuracy on the American Sign Language Lexicon Video Dataset.

Despite these advancements, Vietnamese sign language (VSL) presents unique challenges that distinguish it from other sign languages. Unlike systems that map static hand gestures directly to letters or words, VSL incorporates both characters and tone marks. For example, characters such as "ô" and "ê" are formed by combining a base hand gesture representing the vowel with an additional gesture representing the diacritical hat [6]. Consequently, the recognition of VSL cannot rely solely on single-frame image analysis; it requires an understanding of the entire sequence of hand movements to interpret the intended signs accurately.

In this paper, we propose a deep neural network-based approach that integrates MediaPipe for landmark extraction with a bi-directional long short-term memory (Bi-LSTM) network incorporating an attention mechanism for temporal sequence modeling in VSL recognition. Our method leverages MediaPipe to accurately capture detailed spatial landmarks, while the Bi-LSTM, building upon the foundational work on Long Short-Term Memory networks [7] and bidirectional recurrent neural networks [8], processes the dynamic temporal evolution of the signing process. Furthermore, the attention mechanism, inspired by advancements in neural machine translation [9] and the transformer model [10], allows the model to focus on critical transitions and pivotal frames, thereby enhancing its ability to discern isolated gestures and interpret complete signing sequences. This holistic framework addresses the contextual ambiguities inherent in VSL, where the meaning of a gesture can vary significantly depending on its sequence context. Ultimately, our approach not only advances deep learning applications in sign language recognition but also provides a versatile framework that can be adapted to other sign languages.

## 2. Proposed methodology

### *2.1. Proposed architecture of the model*

We propose the following methodology for hand sign recognition, as illustrated in Figure 1:

- Frame extraction: Decompose the input video into individual frames.
- Landmark detection: Utilize MediaPipe to extract 21 landmarks for the left hand and 21 landmarks for the right hand, totaling 42 landmarks in each frame.
- VSL classification: Input the sequence of hand landmarks into a deep learning neural network to classify the gestures into corresponding hand sign language representations.

Gesture classification is performed using four distinct models for comparative analysis: a 1D convolutional neural network (named as CNN-1D), a standard bidirectional long short-term memory (named as Bi-LSTM-Hand) network, a Bi-LSTM enhanced with additive attention

mechanism (referred to as Bi-LSTM-Att-14), and a Bi-LSTM integrated with multi-head attention mechanism (referred to as Bi-LSTM-Att-17).

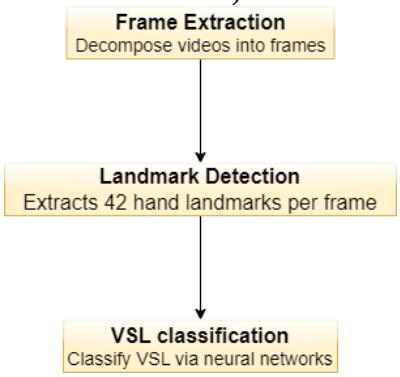


**Figure 1.** *Proposed method for VSL recognition*

Figure 2 illustrates the architecture of the Bi-LSTM model integrated with the attention mechanism, which achieved the highest accuracy among the models evaluated.

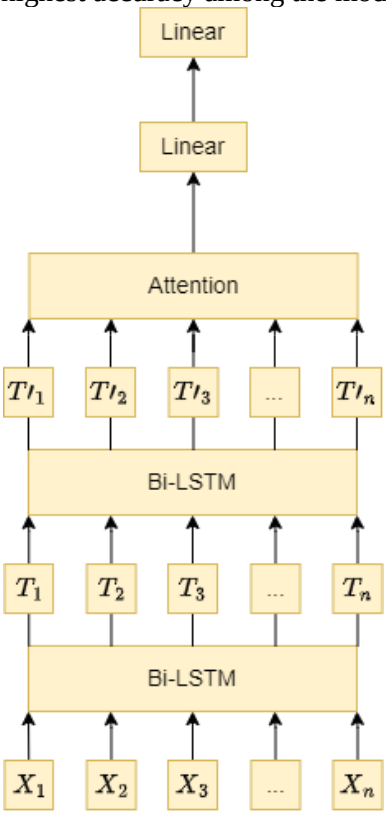


**Figure 2.** *Bi-LSTM with attention architecture*

Figure 2 illustrates the architecture diagram; the batch size dimension is omitted for clarity. The input sequence consists of vectors $X_1, X_2, ..., X_n$, each being a one-dimensional vector of length 126, corresponding to 42 hand landmarks with 3 coordinates each. These vectors are processed through two Bi-LSTM layers. The output from the last Bi-LSTM layer is a sequence of vectors $T'_1, T'_2, ..., T'_n$, representing the time steps. This output is then fed into an attention layer (additive or multi-head). The attention mechanism produces a single vector, which is then passed through two linear layers. Following the attention and fully-connected layers, batch normalization, ReLU activation, and dropout are applied to the output vector.

### *2.2. MediaPipe*

MediaPipe is an open-source, cross-platform framework developed by Google for constructing perception pipelines. It offers pre-trained solutions for real-time video analysis and incorporates state-of-the-art models for various tasks, including face, hand, and pose tracking. In this study, MediaPipe is utilized to extract spatial landmarks from input video data. Specifically, each video is processed frame-by-frame, resulting in the extraction of 42 key landmarks per frame, with each landmark encoded as a three-dimensional coordinate (x, y, z). This automated landmark extraction provides a robust and consistent feature representation for downstream deep learning models. Figure 3 illustrates a sample output generated by MediaPipe.

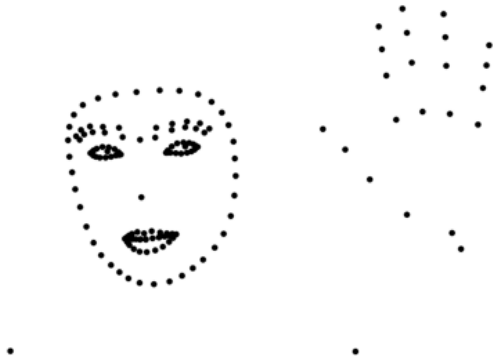

**Figure 3.** *Landmark extraction using MediaPipe, with 21 landmarks for the left hand*

### *2.3. CNN*

To effectively capture the temporal dynamics inherent in Vietnamese sign language gestures, we developed a CNN that processes sequences of hand landmarks extracted using MediaPipe. Each time step in the sequence is represented by a 126-dimensional vector, corresponding to 42 hand landmarks with three coordinates each.

### *2.4. Bi-LSTM*

LSTM networks were originally introduced by Hochreiter and Schmidhuber [7] to address the vanishing gradient problem encountered in traditional recurrent neural networks. They achieve this by maintaining a memory cell that selectively retains information over long sequences. Bi-LSTM extends the conventional LSTM by processing the input sequence in both forward and backward directions, and then concatenating the outputs at each time step. This dual approach, first popularized by Schuster and Paliwal [8], allows the network to capture context from both past and future time steps. Such an arrangement is particularly advantageous for our application, as it helps in understanding the full temporal context of each gesture sequence.

### *2.5. Additive attention*

While Bi-LSTM effectively models temporal dependencies, it treats each time step with equal importance. To allow the model to focus on the most relevant parts of the sequence, we incorporate additive attention introduced by Bahdanau et al. [9]. This attention mechanism computes a weighted sum of the hidden states across time steps. By assigning higher weights to more informative frames, the model can emphasize critical features that are most indicative of a given gesture. Figure 4 illustrates the architecture of the Bidirectional model augmented with additive attention.

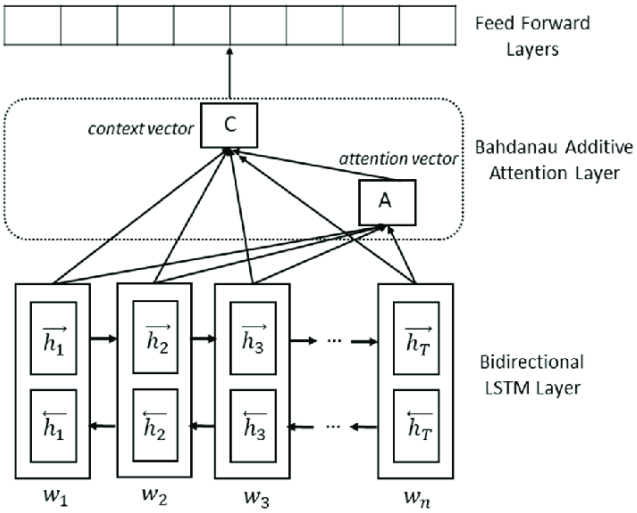


**Figure 4.** *Additive Attention as proposed in [9]*

### *2.6. Multi-head attention*

To further refine the model's focus on different aspects of the sequence, we also experiment with multi-head attention based on Vaswani et al. [10]. Multi-head attention extends the conventional attention mechanism by allowing the model to jointly attend to information from different representation subspaces at various positions. This is achieved by running multiple attention operations in parallel (referred to as "heads"). Figure 5 illustrates this process, in which multiple attention heads, each using Scaled Dot-Product Attention, independently compute weighted representations. These outputs are then concatenated and passed through a linear transformation to yield the final attention output.

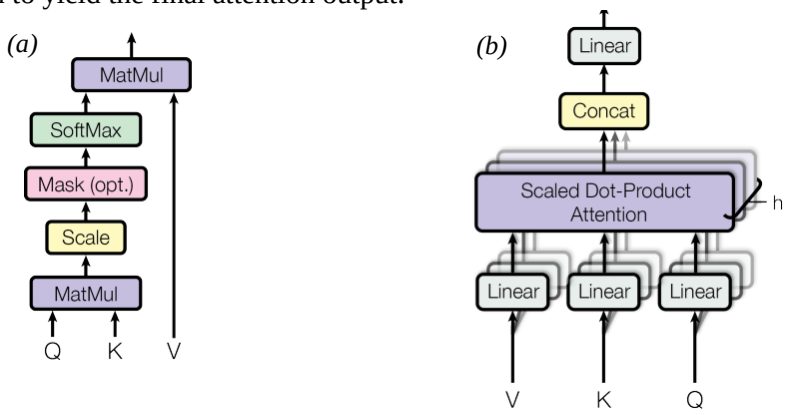


**Figure 5.** *(a) Scaled dot-product attention; (b) Multi-head attention as proposed in [10]*

We obtain a single attention vector from multi-head attention by applying a pooling layer along the first dimension, excluding the batch dimension.

## 3. Experimental data and results

### *3.1. Experimental data*

Our self-constructed dataset includes recordings of 29 Vietnamese characters, supplemented by five tonal diacritics ('sắc', 'huyền', 'hỏi', 'ngã', and 'nặng'), as illustrated in Figure 6. The recordings vary in length from 1 to 5 seconds, with the majority lasting around 1.5 seconds.

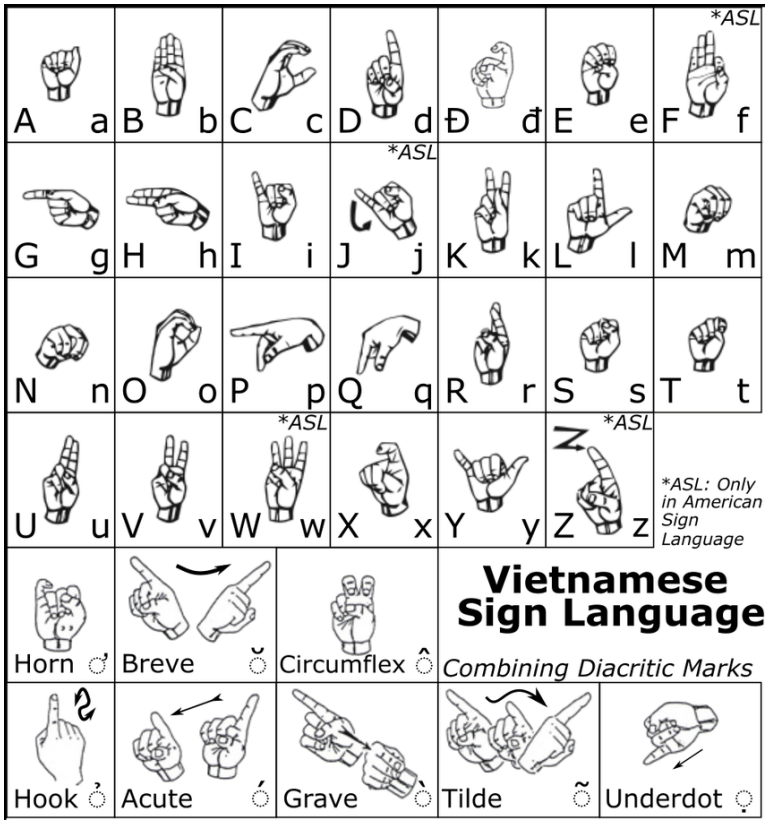


**Figure 6**. *Vietnamese alphabet based on [6]*

Following processing with MediaPipe to extract hand landmarks, the dataset is structured with an overall shape of (1055, 90, 42, 3). Here, 1055 videos consist of 90 selected frames each, where longer videos are trimmed at the end and shorter ones are padded with zero frames, and each frame contains 42 hand landmarks represented in 3 dimensions. The dataset was doubled through horizontal reflection. By selecting six instances per class, the validation and test sets each have shape (408, 90, 42, 3). To enhance the diversity and robustness of the training set, we applied various data augmentation techniques, including spatial rotation, shifting, zooming, random frame removal, and random frame duplication. These augmentations are applied at the start of each training batch and serve as a regularization method, helping to prevent the model from over-fitting to the training data.

### *3.2. Experimental design*

#### *3.2.1. CNN-1D*

Our CNN-1D can be understood as a special case of an ST-GCN. In an ST-GCN, the 126-dimensional joint feature vector for each frame is first processed by a graph convolution that uses the skeleton's adjacency, and then those outputs are convolved across frames. By contrast, our CNN-1D treats the entire 126-dimensional vector as a single channel per time step and omits any spatial graph convolution, applying only a temporal convolution across those vectors. In other words, it is an ST-GCN with the spatial-convolution step removed. The architecture comprises three convolutional blocks. The first block consists of a 1D convolution with 128 output channels and a kernel size of 3, followed by batch normalization and ReLU activation. The second block increases the output to 256 channels, also followed by batch normalization and ReLU activation. The third block further expands the feature representation to 512 channels, maintaining the same configuration. To summarize temporal information, an adaptive max pooling layer is applied

across the time dimension, producing a fixed-length feature vector. For classification, the fully connected head processes this vector through two linear layers, each followed by sequence batch normalization, ReLU activation, and dropout.

*3.2.2. Bi-LSTM-Hand*

In our experimental setup, the Bi-LSTM layers are configured with an output dimensionality of 256 units. Two Bi-LSTM layers with the same architecture are stacked on top of each other. Following these, the first fully connected (Dense) layer comprises 128 neurons, leading up to the final classification layer. Batch normalization, ReLU activation and dropout are applied after each linear layer and attention layer.

*3.2.3. Bi-LSTM-Att-14*

This model is built based on the architecture shown in Figure 1. The Bi-LSTM block has the same architecture as the Bi-LSTM-Hand model described in section 3.2.2. The classification head includes two fully connected layers (the first one has 128 neurons, the second one has 34 neurons for classification).

*3.2.4. Bi-LSTM-Att-17*

This model is constructed using the same architecture as Bi-LSTM-Att-14, with the additive attention mechanism replaced by a multi-head attention mechanism configured with four attention heads.

*3.2.5. Training*

All models were trained using the Adam optimizer with an initial learning rate set to 1e-3 with a batch size of 128. A learning rate scheduler was employed to reduce the learning rate by a factor of 0.5 at epochs 48, 96, 144 and 192. The training process spanned 256 epochs, with the model achieving the lowest cross-entropy loss on the validation set selected for final evaluation.

Given the imbalanced distribution of classes within the dataset, a class-aware sampling strategy was applied. This sampling mechanism decreases the occurrence of samples from overrepresented classes while increasing the sampling frequency of underrepresented classes, thereby promoting more balanced learning and improving the model's generalization across all gesture categories.

### *3.3. Experimental results and discussions*

Table 1 presents the performance outcomes for the evaluated models: CNN-1D, Bi-LSTM-Hand, Bi-LSTM-Att-14, and Bi-LSTM-Att-17.

**Table 1**. *Comparisons of each model's performance*

| Model | Parameters | Test Accuracy |
|---|---|---|
| CNN-1D | 683170 | 98.53% |
| Bi-LSTM-Hand | 695458 | 98.21% |
| Bi-LSTM-Att-14 | 695715 | **99.51%** |
| Bi-LSTM-Att-17 | **959138** | **99.51%** |

All models demonstrated high accuracy, indicating effective data preprocessing and augmentation. Notably, Bi-LSTM models with attention mechanisms outperformed others, with Bi-LSTM-Att-17 and Bi-LSTM-Att-14 achieving the highest accuracy of 99.51%. The attention mechanism enhances the model's focus on relevant parts of the input sequence, improving generalization and learning outcomes. These results highlight the importance of attention-based architectures, especially as datasets grow in size and complexity.

## 4. Conclusion

In this study, we developed a deep neural network system for recognizing Vietnamese sign language (VSL) by integrating MediaPipe for landmark extraction with a Bi-LSTM architecture enhanced by attention mechanisms for sequence modeling. Our approach achieved high accuracy and maintained a reasonable model size while successfully recognizing 29 Vietnamese characters and 5 tone marks. These results demonstrate the effectiveness of combining spatial landmark extraction with advanced sequence modeling techniques for real-time applications. Future work will focus on expanding the dataset to encompass a broader range of gestures and on implementing real-time recognition capabilities to further enhance communication accessibility for the hearing-impaired community in Vietnam.

REFERENCES

[1] V. Bazarevsky *et al.,* "MediaPipe: A Framework for Building Perception Pipelines," *arXiv preprint arXiv:1906.08172,* 2019.

[2] R. Kumar, A. Bajpai, and A. Sinha, "Mediapipe and CNNs for Real-Time ASL Gesture Recognition," *arXiv preprint arXiv:2305.05296*, 2023.

[3] H. P. The, H. C. Chau, V.-P. Bui, and K. Ha, "Automatic feature extraction for Vietnamese sign language recognition using support vector machine," *2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom),* Jan. 2018, pp. 146–151, doi: 10.1109/SIGTELCOM.2018.8325780.

[4] S. Yan *et al.,* "Spatial–Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," in *Proc. 32nd AAAI Conf. on Artificial Intelligence*, 2018, pp. 7444–7452.

[5] L. Shi *et al.,* "Two-Stream Adaptive Graph Convolutional Network for Skeleton-Based Action Recognition," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 5678–5686.

[6] C. C. De Amorim *et al.,* "Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition," in *Proc. International Joint Conference on Neural Networks (IJCNN*), 2019, pp. 1–8.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv preprint arXiv:1409.0473,* 2014.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000 - 6010.