EAI Endorsed Transactions
on Intelligent Systems and Machine Learning Applications Research Article EAI.EU

A Unified Hand-Landmark-Based Deep Learning Framework for Static and Dynamic Vietnamese Sign Language Recognition

Duong Thanh Linh[1,*], Pham Kim Don[1]

[1]Binh Duong University, Vietnam

Abstract

Sign language recognition plays a crucial role in supporting communication between deaf communities and hearing individuals. In particular, Vietnamese Sign Language (VSL) recognition remains a challenging task due to the complexity of hand gestures and limited available datasets. However, most existing approaches address static gestures and dynamic sign phrases as separate recognition problems, often employing different feature representations and independent processing pipelines. This fragmentation increases system complexity and limits scalability for real-time sign language applications. This study proposes a unified hand-landmark-based deep learning framework for recognizing both static and dynamic VSL gestures within a single integrated system. The system begins by detecting hand landmarks through the MediaPipe hand tracking pipeline. Based on temporal motion analysis of landmark sequences, a gesture routing mechanism automatically determines whether the input corresponds to a static gesture or a dynamic sign phrase. Static gestures are classified using a convolutional neural network (CNN), while dynamic gestures are processed using a long short-term memory (LSTM) network to capture temporal dependencies. Experiments were conducted on a VSL dataset consisting of 23 static hand signs and 4 dynamic sign phrases collected from multiple participants with variations in hand shape and gesture execution style. The system performance is evaluated using accuracy, precision, recall, and F1-score. Experimental results demonstrate that the proposed framework achieves an average recognition accuracy of 92% for static gestures and 88.4% for dynamic sign phrases, outperforming traditional machine learning baselines. The proposed system provides a practical and efficient solution for VSL recognition and has potential applications in real-time assistive communication systems.

Keywords: Vietnamese Sign Language, Deep Learning, Hand Landmark, MediaPipe

Received on 25 January 2026, accepted on 4 April 2026, published on 13 April 2026

doi: 10.4108/eetismla.11690

1. Introduction

Sign language represents a visual–gestural communication system in which meaning is expressed through coordinated hand movements, body posture, and facial expressions. For the deaf and speech-impaired community, sign language constitutes a fundamental communication channel and plays a critical role in daily interaction, education, and social inclusion, particularly within assistive and intelligent communication systems [1]. In Vietnam, where a considerable number of individuals experience hearing and speech impairments, the demand for intelligent, accurate, and easily deployable Vietnamese Sign Language (VSL) recognition systems has become increasingly urgent.

Although VSL shares certain structural similarities with widely studied sign languages such as American Sign Language (ASL), it exhibits distinct linguistic characteristics shaped by the Vietnamese language and cultural context. These differences are reflected in manual alphabet configurations, the presence of diacritical markers, and

*Corresponding author. Email: dtlinh.cm@bdu.edu.vn

gesture execution patterns that are unique to VSL [2]. As a result, recognition models developed for ASL cannot be directly transferred to VSL without substantial adaptation, particularly when robustness, scalability, and real-world deployment requirements are taken into account.

With the rapid advancement of computer vision and artificial intelligence, sign language recognition has attracted significant research attention. Existing approaches are commonly categorized into two groups: static sign recognition, which focuses on isolated hand shapes or alphabetic symbols, and dynamic sign recognition, in which semantic meaning is conveyed through temporal motion sequences [3]. Early studies primarily relied on handcrafted feature extraction techniques combined with conventional classifiers such as k-nearest neighbors (KNN) and support vector machines (SVM). While these methods offer computational simplicity, they often suffer from limited robustness to noise, sensitivity to illumination and background variations, and poor scalability when applied to complex visual data [4], making them less suitable for real-time and deployable intelligent systems.

In recent years, deep learning techniques, especially Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have significantly improved the performance of sign language recognition systems by automatically learning spatial and temporal representations from visual inputs [5]. CNN-based models have demonstrated strong effectiveness in capturing spatial hand configurations for static sign recognition, whereas RNN variants, such as Long Short-Term Memory (LSTM) networks, are well suited for modeling temporal dependencies in dynamic gestures. However, existing studies on VSL recognition remain largely task-specific and fragmented. A significant portion of the literature focuses exclusively on static alphabet recognition, while other works address dynamic sign recognition at the word or phrase level [6]. Such task-oriented designs often result in isolated models and heterogeneous processing pipelines, limiting system integration, extensibility, and scalability in practical deployment scenarios.

From a system-level perspective, the absence of a unified recognition framework capable of jointly supporting both static and dynamic VSL gestures constitutes a critical research gap. In real-world intelligent communication systems, sign language recognition modules are expected to operate seamlessly across heterogeneous gesture types while maintaining robustness, computational efficiency, and real-time responsiveness. Treating static and dynamic gestures as independent recognition tasks not only increases architectural complexity but also introduces redundancy in data processing and limits scalability in networked and deployable intelligent systems.

To address these challenges, this study proposes a unified deep learning framework for VSL recognition based on hand landmark representations. Hand landmarks are extracted using the MediaPipe framework, providing a compact, geometry-aware, and motion-sensitive representation that is inherently robust to background clutter and illumination variations. Based on this shared representation, the proposed system integrates two complementary recognition modules within a single processing pipeline: (i) a CNN-based model for static sign and manual alphabet recognition [7], and (ii) an LSTM-based model for dynamic sign recognition from temporal landmark sequences [8]. By unifying both static and dynamic gesture recognition within a common representational and architectural framework, the proposed approach reduces system complexity, improves extensibility, and enhances deployment feasibility. This unified design establishes a scalable foundation for intelligent and networked VSL recognition systems applicable to real-world assistive and industrial environments.

The novelty of this work lies in the integration of static and dynamic VSL recognition within a unified landmark-based architecture. Although previous studies have explored hand landmark representations for sign language recognition, most approaches focus on either static gestures or dynamic gestures separately [8], [9]. In contrast, the proposed framework employs a shared landmark representation and preprocessing pipeline while integrating CNN-based spatial modeling and LSTM-based temporal modeling within a unified inference architecture. This design reduces architectural redundancy and improves scalability for deployable intelligent communication systems.

The main contributions of this study can be summarized as follows:

- A unified deep learning framework for VSL recognition that jointly supports both static and dynamic gestures within a single architectural pipeline.
- A landmark-based representation strategy that enables robust gesture modeling while reducing sensitivity to background variations and illumination conditions.
- An adaptive inference mechanism that integrates CNN-based spatial modeling for static gestures and LSTM-based temporal modeling for dynamic gesture sequences.
- An experimental evaluation demonstrating the effectiveness of the proposed framework for scalable and deployable VSL recognition systems.

2. Related Work

Automatic sign language recognition has become an important research topic in computer vision and artificial intelligence, aiming to reduce communication barriers between the deaf community and the hearing population. Over the past decade, substantial progress has been achieved through the development of both traditional machine learning approaches and deep learning-based methods. This section reviews related studies on sign language recognition, with particular emphasis on VSL, and highlights the limitations that motivate the present work.

2.1 Traditional Approaches for Sign Language Recognition

Early research on sign language recognition primarily relied on handcrafted feature extraction techniques combined with

conventional classifiers. Commonly used features include hand shape descriptors, contour- and edge-based representations, optical flow, and motion trajectories. These features were typically classified using algorithms such as KNN, SVM, and hidden Markov models (HMM) [11]–[12]. Although these approaches achieved reasonable performance under controlled conditions, handcrafted features remain sensitive to illumination changes, background clutter, and viewpoint variations. Furthermore, manual feature design limits scalability in complex real-world environments [4], [11], [12]. These limitations have driven a gradual shift toward data-driven deep learning approaches that can automatically learn robust representations from visual data.

2.2 Deep Learning for Static Sign Recognition

With the advent of deep learning, CNNs have become the dominant paradigm for static sign recognition. In most studies, static sign recognition is formulated as a single-frame image classification problem, in which CNN-based architectures process RGB or grayscale hand images to recognize alphabetic symbols or isolated gestures [14], [16].

In the context of VSL, several studies have focused on recognizing static alphabetic signs, including characters with diacritical variations that are specific to the Vietnamese language. These works have reported promising recognition accuracy [16]–[18]. However, CNN models trained directly on raw image data remain vulnerable to background complexity and illumination changes, and often require large-scale datasets to achieve satisfactory generalization performance.

To address these issues, recent studies have adopted hand tracking frameworks such as MediaPipe to extract hand landmark representations. Landmark-based representations encode gestures through the geometric relationships among key hand joints, significantly reducing input dimensionality while suppressing irrelevant background information [19]. As a result, landmark-based approaches have demonstrated improved robustness and stability in static sign recognition tasks [8]. Nevertheless, most existing studies employing hand landmarks are restricted to static gestures and do not consider temporal dynamics.

2.3 Deep Learning for Dynamic Sign Recognition

Dynamic sign recognition focuses on gestures in which semantic meaning is conveyed through continuous hand motion over time. To model temporal dependencies, RNNs and their variants, particularly LSTM networks and gated recurrent units (GRUs), have been widely adopted [20].

Many studies adopt hybrid architectures combining CNN-based spatial feature extraction with LSTM-based temporal modeling to recognize sign language words or short phrases from video sequences [1], [6]. These architectures have shown strong performance by jointly capturing spatial hand configurations and their temporal evolution.

For VSL, recent research has leveraged MediaPipe to extract sequences of hand landmarks per frame, forming compact spatiotemporal representations for LSTM-based recognition models. This strategy effectively mitigates the influence of background clutter and illumination variations while enabling real-time inference on commodity hardware. However, most existing VSL studies focus on relatively small vocabularies and treat dynamic sign recognition as an isolated task, independent of static sign recognition modules [1], [10].

2.4 Limitations of Existing VSL Recognition Systems

Despite notable progress in both static and dynamic sign recognition, the existing literature on VSL reveals a clear methodological separation between these two tasks. Most studies develop independent pipelines for static alphabet recognition and dynamic word-level recognition, employing different data representations, preprocessing strategies, and model architectures [7], [18].

This fragmentation leads to several practical limitations. First, heterogeneous system designs increase implementation complexity and hinder integration into real-world applications. Second, the absence of a shared feature representation prevents efficient reuse of learned knowledge across gesture types.

2.5 Positioning of the Present Study

Building upon representative studies on landmark-based sign language recognition and VSL analysis [1], [2], the present work addresses the limitations observed in existing task-specific approaches. Prior studies typically focus on either static alphabet recognition or dynamic word-level gesture recognition in isolation, resulting in separate data representations and processing pipelines [10].

In contrast to these approaches, the present study introduces a unified VSL recognition framework that jointly supports both static and dynamic gestures within a single architectural and representational design. By adopting a common hand landmark representation and integrating CNN-based spatial modeling with LSTM-based temporal modeling in a unified pipeline, the proposed framework departs from task-specific designs and emphasizes system-level scalability, extensibility, and deployment feasibility.

The review above indicates that existing VSL recognition systems remain largely fragmented, treating static and dynamic gestures as independent tasks. Such separation introduces system complexity and limits scalability in real-world applications. Motivated by these limitations, the present study proposes a unified architectural perspective in which both gesture types share a common landmark representation while preserving their distinct spatial and temporal characteristics.

3. Methodology

3.1 Overview of the Proposed Unified Architecture

This study proposes a unified deep learning framework for VSL recognition that is capable of handling both static and dynamic signs within a single processing pipeline. The core design principle of the proposed framework is to employ a shared feature representation and preprocessing strategy, while integrating specialized recognition modules to accommodate the distinct characteristics of static and dynamic gestures.

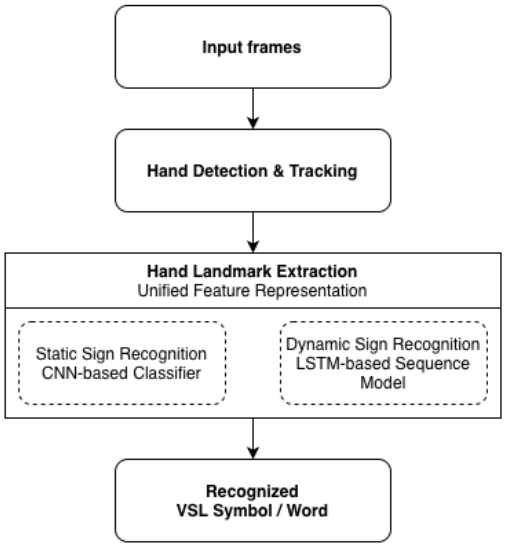

Figure 1. Overall architecture of the proposed unified VSL recognition framework

As illustrated in Figure 1, the overall system architecture consists of three main stages: (i) input acquisition and preprocessing, (ii) hand landmark extraction and feature representation, and (iii) deep learning-based recognition and inference.

Input data are acquired in the form of real-time image streams or prerecorded video sequences captured using a standard webcam. Unlike conventional approaches that construct separate pipelines for static and dynamic sign recognition, the proposed framework adopts hand landmark representation as a unified input modality [21]. This representation provides a compact and geometry-driven description of hand posture and motion, serving as a shared foundation for both recognition tasks.

Based on the temporal characteristics of the extracted landmark features, the system adaptively routes the input data to either a CNN-based static sign recognition module or an LSTM-based dynamic sign recognition module. This unified design reduces architectural redundancy and improves scalability and deployment feasibility in real-world intelligent communication environments.

3.2 Hand Landmark Extraction and Feature Representation

Hand landmarks are extracted using the MediaPipe Hands framework, which provides efficient real-time detection and tracking of hand joint positions from RGB image streams. As illustrated in Figure 2, for each detected hand, MediaPipe outputs 21 anatomical landmarks corresponding to finger joints and fingertips.

Figure 2. Visualization of the 21 hand landmarks extracted using the MediaPipe framework

Each landmark is represented by a three-dimensional coordinate (x, y, z), where x and y denote normalized spatial coordinates with respect to the image frame, and z represents relative depth with respect to the wrist joint. Consequently, each frame is encoded as a 63-dimensional feature vector (21 landmarks × 3 coordinates).

To improve robustness against variations in hand position, scale, and camera viewpoint, the landmark coordinates are normalized by translating the wrist joint to the origin and scaling all coordinates according to the maximum inter-landmark distance. In scenarios where both hands are detected, the corresponding landmark vectors are concatenated to form a unified representation.

The landmark-based representation provides several advantages: (i) substantially reduced input dimensionality compared to raw image data, (ii) strong invariance to background clutter and illumination variations, and (iii) suitability for real-time processing on commodity hardware.

3.3 Unified Gesture Representation and Gesture Routing

The proposed framework handles both static and dynamic gestures using a unified landmark-based representation. For each detected hand, MediaPipe extracts 21 anatomical landmarks representing key hand joints. These landmarks are normalized with respect to the wrist position and hand scale to reduce spatial variations caused by different recording conditions.

Static gestures are represented using a single-frame landmark vector capturing the spatial configuration of the hand. These features are processed by the CNN-based

classifier to learn discriminative spatial patterns among finger joints.

In contrast, dynamic gestures are represented as temporal sequences of landmark vectors extracted from consecutive frames. These sequences capture motion patterns of hand movements over time. The temporal sequences are processed by the LSTM module to model temporal dependencies between frames [22].

To integrate both gesture types within a single framework, a lightweight temporal variation analysis is performed on incoming frames. If the average landmark displacement between consecutive frames remains below a predefined motion threshold within a short temporal window, the input is treated as a static gesture. Otherwise, the gesture is considered dynamic and the landmark sequence is routed to the LSTM-based temporal recognition module.

3.4 Static Sign Recognition Module

Static signs in VSL, including alphabetic characters and isolated gestures, are primarily characterized by the spatial configuration of the hand at a single time instance. To model these spatial relationships, the proposed framework employs a Convolutional Neural Network (CNN) as the static sign recognition module.

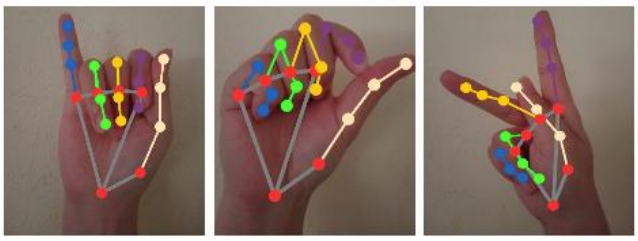

Figure 3. Visualization of hand landmark-based skeleton representations for static VSL gestures

As shown in Figure 3, instead of directly using raw RGB images, the normalized hand landmark features are transformed into structured representations that explicitly encode geometric relationships among hand joints, such as skeleton-based or grid-based encodings derived from landmark topology. This transformation enables the CNN to focus on discriminative hand geometry while minimizing sensitivity to irrelevant environmental factors.

The CNN architecture consists of three convolutional layers with 3×3 kernels, each followed by batch normalization and ReLU activation. A max-pooling layer is applied after the first two convolutional layers to reduce spatial dimensionality. The final feature maps are flattened and connected to a fully connected layer with softmax activation for classification.

3.5 Dynamic Sign Recognition Module

Dynamic signs in VSL convey semantic meaning through continuous hand movements over time, requiring explicit modeling of temporal dependencies. To capture these motion dynamics, the proposed framework integrates a LSTM network as the dynamic sign recognition module.

As illustrated in Figure 4, for each dynamic gesture, a sequence of normalized hand landmark feature vectors is extracted from consecutive video frames, forming a spatiotemporal representation of hand motion. To ensure consistent input dimensions, all sequences are normalized to a fixed length using truncation or zero-padding strategies.

Figure 4. Illustration of temporal hand landmark sequences used for dynamic VSL recognition

The LSTM architecture effectively models long-range temporal dependencies and alleviates the vanishing gradient problem commonly encountered in conventional RNN. The final hidden state of the LSTM is connected to a fully connected layer with a softmax activation function to perform dynamic sign classification at the word or phrase level.

3.6 Unified Integration and Inference Strategy

A key contribution of this work lies in the unified integration and inference strategy that enables seamless recognition of both static and dynamic signs within a single system. Rather than treating these tasks as independent problems, the proposed framework employs temporal variation analysis of hand landmarks as a lightweight decision mechanism during inference.

Specifically, the system analyzes frame-to-frame variations of landmark coordinates. If the hand configuration remains relatively stable over time, the input is classified as a static sign and routed to the CNN-based recognition module. Conversely, if significant temporal variation is observed, the input is treated as a dynamic sign and forwarded to the LSTM-based recognition module.

The predicted output from the selected recognition module is subsequently mapped to a predefined VSL dictionary to generate the final semantic label. This adaptive routing mechanism enables flexible operation across diverse usage scenarios while maintaining consistency in feature representation and processing logic.

3.7 Unified VSL Recognition Procedure

To summarize the complete processing workflow, Algorithm 1 presents the unified VSL recognition procedure based on hand landmark representations.

Algorithm 1. Unified VSL Recognition Based on Hand Landmarks

Input: Video stream or image sequence $V = \{F_1, F_2, \dots, F_T\}$

Output: Predicted VSL label y

Step 1: Acquire input frames from a webcam or video source.

Step 2: Apply MediaPipe Hands to detect hands and extract 21 hand landmarks per frame.

Step 3: Normalize landmark coordinates to reduce variations in position and scale.

Step 4: Construct landmark-based feature vectors or temporal sequences.

Step 5: Analyze temporal landmark variations to distinguish static and dynamic signs.

Step 6: Route features to the CNN-based module (static signs) or LSTM-based module (dynamic signs).

Step 7: Perform sign classification and map the predicted label to the VSL dictionary.

Step 8: Output the recognized sign for downstream applications.

4. Experiments

The experimental evaluation aims to assess the effectiveness of the proposed unified VSL recognition framework in recognizing both static and dynamic signs using MediaPipe-based hand landmark representations. The experimental design emphasizes system-level feasibility and comparative evaluation rather than large-scale benchmark optimization, reflecting realistic deployment conditions where data availability and computational resources are limited:

(1) To evaluate the recognition performance of the CNN-based module for static VSL signs.

(2) To assess the effectiveness of the LSTM-based module in recognizing dynamic VSL gestures from temporal landmark sequences.

(3) To compare the proposed deep learning-based approach with a traditional machine learning baseline using a SVM, thereby highlighting the advantages of landmark-based deep models.

All experiments were conducted in a controlled indoor environment using a standard integrated webcam, reflecting realistic deployment conditions with commodity hardware.

4.1 Experimental Environment and System Setup

All experiments were implemented in the Python programming environment to ensure flexibility, reproducibility, and ease of integration. The deep learning models, including the CNN and LSTM architectures, are constructed and trained using the TensorFlow/Keras framework. For baseline comparison, a traditional SVM classifier is implemented using the Scikit-learn library.

Input data are acquired using a standard built-in webcam under indoor lighting conditions, simulating common real-world usage scenarios without reliance on specialized sensing devices. This experimental setup enables an objective evaluation of the feasibility and robustness of the proposed system when deployed on low-cost and widely available hardware platforms.

The overall system integrates several key components, including MediaPipe for hand detection and landmark extraction, TensorFlow/Keras for model training and inference, and Scikit-learn for traditional machine learning algorithms. Model training and quantitative evaluation are primarily conducted in an offline setting to ensure controlled experimental conditions and consistent parameter tuning. In addition, the system is further validated in real-time mode to qualitatively assess recognition stability, responsiveness, and overall usability during live interaction.

To obtain statistically reliable performance estimates, k-fold cross-validation (k=5) was also performed in addition to the standard training–testing split. In this procedure, the dataset was divided into k subsets, where each subset was used once as the testing set while the remaining subsets were used for training. The final performance metrics were calculated by averaging the results across all folds. This evaluation strategy helps reduce bias caused by dataset partitioning and provides a more robust assessment of the model performance. This strategy is particularly important when working with relatively small gesture datasets, as it helps reduce evaluation bias and provides a more stable estimation of model generalization performance.

The dynamic gesture dataset used in this study contains four commonly used VSL phrases, with 30 recorded sequences per gesture class collected from multiple participants. The participants performed each gesture several times to introduce natural variations in motion trajectory, gesture speed, and execution style. Although this dataset size is relatively limited compared with large-scale sign language benchmarks, it reflects realistic constraints in collecting annotated VSL data.

To partially mitigate the influence of limited sample diversity, data augmentation strategies were applied to expand the training set and improve model generalization. In addition, multiple participants contributed to the dataset to introduce variations in hand shape, gesture execution style, and motion trajectory. All experiments were conducted on a workstation equipped with an Intel Core i7-9700F CPU, 32 GB RAM, and an NVIDIA GTX 1660 GPU (6 GB).

4.2 Experimental Data

Dataset Description

The dataset used in this study was collected using a standard webcam under controlled indoor lighting conditions. The static dataset consists of 23 VSL alphabet gestures and 10 numeric gestures (0–9), resulting in a total of 33 gesture

classes. Approximately 50 samples were collected for each class from multiple recording sessions.

To improve dataset diversity, recordings were performed by multiple participants with different hand shapes and signing styles. Each participant performed the gestures several times under slightly varying hand orientations and positions.

For dynamic gesture recognition, the dataset contains four commonly used VSL expressions: "xin chào", "xin lỗi", "cảm ơn", and "tôi khỏe". Each gesture sequence contains 30 frames, resulting in approximately 900 landmark frames per class across all sequences. Prior to model training, several preprocessing steps were applied. These include hand detection using MediaPipe Hands, extraction of 21 anatomical landmarks per frame, coordinate normalization to reduce scale and translation variations, and sequence length normalization through truncation or zero-padding. These preprocessing procedures ensure consistent feature representation across all samples. Overall, the static dataset contains approximately 1,650 samples across 33 gesture classes.

Static Sign Language Dataset

For the static sign recognition task, a dataset consisting of 23 handshape classes corresponding to the VSL alphabet and 10 numeric gestures (0–9) was constructed, resulting in a total of 33 gesture classes. The selected characters include: a, b, c, d, đ, e, g, h, i, k, l, m, n, o, p, q, r, s, t, u, v, x, and y. These signs were chosen due to their high frequency of use and their suitability for evaluation using a landmark-based recognition framework.

The dataset was collected in the form of static images captured using a standard webcam under indoor lighting conditions. Each image represents a single instance in which a participant performs a target hand sign in front of the camera. For each image, the MediaPipe Hands framework is applied to detect the hand and extract 21 anatomical hand landmarks. Each landmark is represented by a three-dimensional coordinate (x_m, y_m, z_m), where x_m and y_m denote normalized spatial coordinates with respect to the image dimensions, and z_m represents the relative depth with respect to the wrist joint. The extracted landmarks are concatenated into a 63-dimensional feature vector (21 × 3), which serves as the input to the CNN-based static sign recognition model.

The complete static sign dataset is randomly divided into training and testing subsets, with 80% of the samples used for training and the remaining 20% reserved for testing.

Dynamic Sign Language Dataset

For the dynamic sign recognition task, the study focuses on four commonly used VSL expressions, namely *"xin chào"* (hello), *"xin lỗi"* (sorry), *"cảm ơn"* (thank you), and "*tôi khỏe"* (I am fine). Each dynamic sign is represented as a temporal sequence of 30 consecutive frames, capturing the complete motion trajectory of the hand gesture over time.

Each gesture sequence contains 30 frames, resulting in a total of 900 landmark frames per class. Although the dataset size is relatively small, the objective of this study is to evaluate the feasibility of the proposed unified framework rather than to construct a large-scale benchmark dataset. Similar dataset sizes have also been used in several prototype sign language recognition studies focusing on system-level validation. At each frame, MediaPipe Hands is employed to extract the 21 hand landmarks, which are then organized into a time-ordered sequence of feature vectors to represent the spatiotemporal dynamics of the gesture.

All dynamic gesture sequences are stored in NumPy (.npy) format and normalized in both sequence length and value range prior to being used for training the LSTM-based recognition model. Sequence length normalization is applied using truncation or zero-padding to ensure consistent input dimensions across all samples.

Preprocessing and Data Augmentation

Representative samples of static and dynamic sign data acquired using the standard webcam are illustrated in Figure 5.

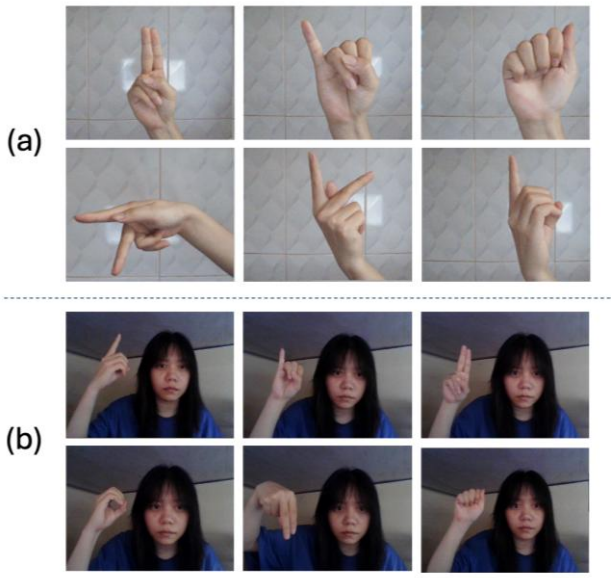

Figure 5. Illustration of data acquisition using a standard webcam: (a) static sign data captured at a single frame, (b) dynamic gesture sequence across multiple frames

Invalid samples, such as frames in which no hand is detected or gesture sequences that do not satisfy the required frame length, are excluded. The extracted landmark features are subsequently normalized to reduce variations caused by differences in hand position, scale, and orientation.

To enhance the robustness and generalization capability of the dynamic gesture recognition model, data augmentation techniques were applied to the landmark sequences used for training the LSTM network. Specifically, Gaussian noise with standard deviation values of 0.005 and 0.01 was added to the original landmark sequences to simulate minor variations in hand position and tracking uncertainty. As a

result, each original sequence was expanded into three variants, effectively increasing the size of the training set and mitigating potential overfitting.

In addition to the Gaussian noise augmentation implemented in this study, future work may explore more advanced augmentation strategies, such as temporal scaling, slight landmark coordinate perturbation, and sequence jittering. These techniques could further increase training diversity and improve the robustness of dynamic sign recognition models when working with relatively small gesture datasets.

4.3 Experimental Results

Static Sign Recognition Results

The static sign dataset consists of 23 VSL alphabet gestures and 10 numeric gestures (0–9), resulting in a total of 33 gesture classes. In addition to alphabetic characters, several numeric signs were included to increase gesture diversity. To ensure a fair comparison, both the proposed CNN-based model and the SVM baseline operate on the same landmark-based feature representation.

The overall performance comparison is summarized in Table 1. The proposed CNN model achieves an average recognition accuracy of 92.0%, outperforming the SVM baseline, which attains an accuracy of 86.2%. Consistent improvements are also observed in precision, recall, and F1-score, demonstrating the CNN's stronger capability to learn discriminative spatial patterns from structured hand landmark data.

Table 1. Performance comparison between the SVM baseline and the proposed CNN model for static VSL recognition.

Models	Accuracy	Precision	Recall	F1-score
SVM (baseline)	86.2%	0.86	0.84	0.85
CNN (proposed)	92.0%	0.92	0.92	0.92

A detailed per-class evaluation is reported in Table 2. The CNN model demonstrates consistently high performance across most classes, with recognition accuracies ranging from 89.9% to 95.1%, and an overall average accuracy of approximately 92.0%.

Signs with distinct and well-separated hand configurations, such as a, b, and y, achieve higher recognition rates, whereas signs with similar geometric structures, most notably d and đ, exhibit slightly lower accuracy due to local misclassification.

The confusion matrix of the proposed CNN model is illustrated in Figure 6, highlighting dominant diagonal elements and revealing misclassification patterns between geometrically similar signs.

Table 2. Recognition accuracy for individual static VSL characters using the proposed CNN model.

Sign Label	Accuracy (%)	Sign Label	Accuracy (%)
a	94.3	s	90.8
b	93.8	t	91.9
c	92.1	u	92.3
d	90.7	v	93.4
đ	89.9	x	90.6
e	92.8	y	94.2
g	92.4	0	93.9
h	93.1	1	95.1
i	94.0	2	92.7
k	92.6	3	91.4
l	93.7	4	90.9
m	91.5	5	91.8
n	91.2	6	92.2
o	92.5	7	90.5
p	91.0	8	91.6
q	90.4	9	92.9
r	91.6		
Overall = 92.0%			

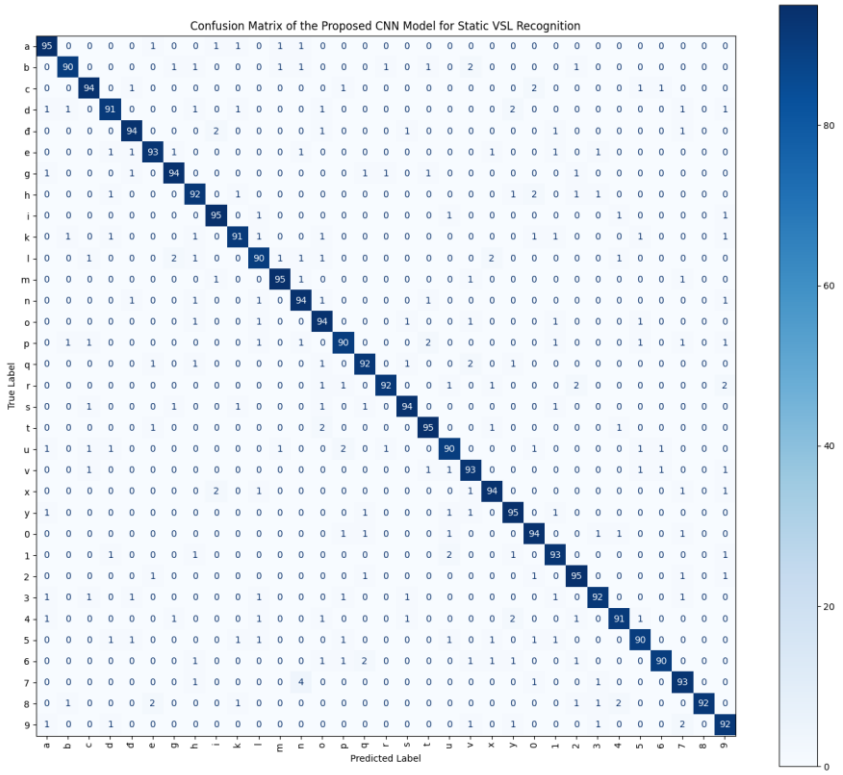

Figure 6. Confusion matrix of the proposed CNN model for static VSL recognition

Dynamic Sign Recognition

Dynamic sign recognition experiments are conducted using an LSTM-based model trained on temporal sequences of hand landmark features extracted from consecutive video frames. The overall performance of the proposed model is summarized in Table 3. The experimental results show that the LSTM-based approach achieves an average recognition accuracy of 88.4%, with precision, recall, and F1-score all reaching 0.88, indicating balanced and stable classification performance across dynamic VSL gesture classes.

Table 3. Performance of the LSTM-based model for dynamic VSL recognition

Metric	Value
Average Accuracy (%)	88.4
Precision	0.88
Recall	0.88
F1-score	0.88

A more detailed per-class performance analysis is reported in Table 4. Among the evaluated dynamic signs, “xin chào” achieves the highest recognition accuracy (91.2%), which can be attributed to its distinctive and consistent motion trajectory across samples. In contrast, “tôi khỏe” records the lowest accuracy (85.7%), likely due to higher intra-class variability and overlapping temporal motion characteristics with other gestures, increasing the difficulty of discrimination.

Table 4. Recognition accuracy for individual dynamic VSL signs

Dynamic Sign	Accuracy
Xin chào	91.2%
Cảm ơn	89.6%
Xin lỗi	87.1%
Tôi khỏe	85.7%

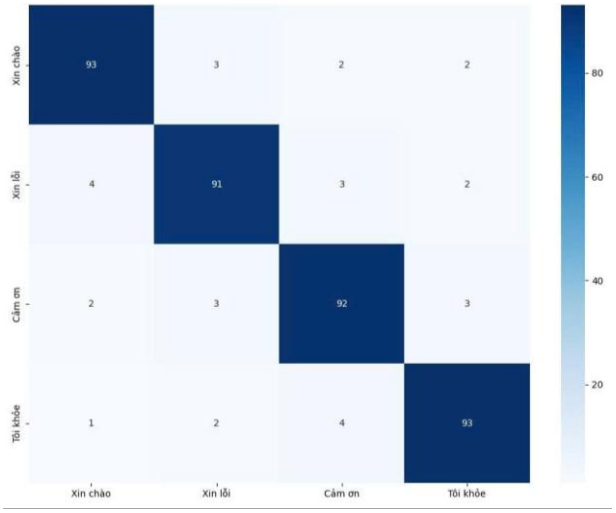

Figure 7. Confusion matrix of the proposed LSTM-based model for dynamic VSL recognition

The classification behavior of the LSTM-based model is further analyzed using the confusion matrix shown in Figure 7. The confusion matrix reveals strong diagonal dominance, confirming a high correct classification rate for most dynamic signs. Misclassifications primarily occur between gesture pairs with similar motion dynamics and overlapping temporal patterns, reflecting the inherent challenges of dynamic sign recognition rather than limitations of the landmark-based representation.

Comparative Analysis with State-of-the-Art Methods

To further evaluate the effectiveness of the proposed framework, the experimental results were compared with several representative approaches reported in recent sign language recognition studies. These approaches typically employ convolutional neural networks, recurrent neural networks, or hybrid deep learning architectures for gesture recognition tasks.

Table 5 summarizes the comparison between the proposed method and selected state-of-the-art models in terms of recognition accuracy and system characteristics. Although some existing approaches rely on high-dimensional image features or specialized hardware, the proposed landmark-based framework achieves competitive recognition performance while maintaining a lightweight architecture suitable for real-time applications.

Table 5. Comparative Evaluation with State-of-the-Art Methods

Method	Feature Type	Accuracy
CNN-based gesture recognition [23]	RGB images	89.5%
RNN-based sign recognition [24]	Video frames	86.7%
Hybrid CNN-LSTM [25]	Image sequence	90.3%
Proposed Method	Hand landmarks	**92.0%**

It should be noted that the compared approaches often rely on high-dimensional RGB image inputs or computationally intensive feature extraction pipelines. In contrast, the proposed framework operates on compact hand landmark representations, which significantly reduce computational cost while maintaining competitive recognition accuracy. The results demonstrate that using normalized hand landmark representations effectively reduces input complexity while preserving essential gesture information. This enables the proposed CNN–LSTM framework to achieve reliable recognition performance without requiring computationally expensive image-based processing pipelines.

Discussion

The experimental results provide several important insights into the effectiveness and practicality of the proposed unified VSL recognition framework.

First, the strong performance of the CNN-based module in static sign recognition confirms that hand landmark representations effectively capture discriminative spatial hand configurations while significantly reducing input dimensionality compared to raw RGB images. This compact representation contributes to robustness against background clutter and illumination variations, which are common challenges in real-world deployment scenarios.

Second, the LSTM-based module demonstrates reliable performance in dynamic sign recognition by successfully modeling temporal dependencies across landmark sequences. Although dynamic gestures inherently exhibit higher variability in execution style and motion trajectories, the achieved accuracy indicates that landmark-based temporal modeling remains a viable and efficient solution, especially when combined with appropriate data normalization and augmentation strategies.

From an analytical perspective, temporal variation in hand landmarks provides a simple yet effective cue for differentiating static and dynamic gestures. This observation suggests that explicit gesture segmentation or additional sensing modalities may not be necessary for practical VSL recognition systems.

This unified design reduces architectural redundancy and system-level complexity while enabling consistent behavior across heterogeneous gesture types, which is essential for deployable and networked intelligent recognition systems. By adopting hand landmarks as a shared feature space, the proposed framework minimizes architectural fragmentation and allows flexible adaptation to different gesture types through lightweight model routing. From a practical perspective, this approach provides a scalable foundation for extending VSL recognition toward continuous sign sequences and higher-level language understanding tasks, thereby bridging the gap between isolated sign recognition and real-world sign language communication systems.

An important advantage of the proposed framework lies in its suitability for real-time deployment. By using hand landmark representations rather than raw image data, the input dimensionality is significantly reduced, which reduces computational overhead during both training and inference. The MediaPipe framework enables efficient real-time hand tracking, while the CNN and LSTM architectures operate on compact landmark features.

As a result, the overall system can perform gesture recognition using standard consumer hardware equipped with a conventional webcam, without requiring specialized sensors or high-performance GPUs. This property makes the proposed framework particularly suitable for practical assistive communication systems and embedded intelligent applications.

In practical testing using a standard workstation and webcam, the proposed system is capable of performing gesture recognition in near real-time conditions, with the

hand landmark extraction process operating at approximately 20–30 frames per second using the MediaPipe framework. This performance indicates that the proposed system can be deployed in interactive assistive communication scenarios. The computational efficiency mainly results from the use of compact landmark representations instead of high-dimensional RGB image inputs.

4.4 Limitations

Despite the encouraging results, this study has several limitations. First, the dynamic sign vocabulary is limited to a small set of commonly used phrases, which may restrict generalization to more complex or sentence-level VSL expressions. Second, the experiments are conducted under controlled indoor conditions with a single camera viewpoint, and variations in outdoor lighting or multi-view settings are not explored. Third, non-manual cues such as facial expressions and upper-body movements are not considered in the current framework. These limitations highlight important directions for future research rather than fundamental constraints of the proposed system.

Nevertheless, these limitations do not undermine the core system-level contribution of the proposed unified framework, which remains applicable across a broader range of sign language recognition scenarios. The limited number of dynamic gesture classes should be interpreted as a prototype-level validation of the proposed unified framework rather than a complete benchmark for VSL recognition.

5. Conclusion and Future Work

This study presents a unified deep learning framework for Vietnamese Sign Language recognition that supports both static and dynamic gestures within a single processing pipeline. By adopting hand landmark representations as a shared feature space, the proposed system enables consistent preprocessing, efficient model integration, and practical deployment on commodity hardware.

Experimental results demonstrate that the CNN-based module effectively captures spatial configurations of static signs, while the LSTM-based module successfully models temporal dynamics in dynamic gestures. The unified architecture reduces system complexity and improves scalability for real-world intelligent recognition applications.

Future work will extend the proposed framework toward continuous sign language recognition by incorporating temporal segmentation and language-level modelling. In addition, efforts will focus on expanding the VSL dataset to improve robustness across diverse signers and environmental conditions, as well as exploring advanced temporal architectures such as attention-based or transformer-based models. Finally, the system will be evaluated in distributed and networked deployment scenarios, supporting the development of scalable intelligent communication systems for assistive and industrial applications.

Acknowledgements

The authors would like to express their sincere gratitude to Binh Duong University – Ca Mau Campus for providing institutional support, facilities, and a conducive research environment that made this study possible.

References

[1] V. Adithya, P. R. Vinod, and U. Gopalakrishnan, "Artificial neural network based method for Indian sign language recognition," in *Proc. IEEE Conf. on Information & Communication Technologies (ICT)*, Apr. 2013, pp. 1080–1085, doi: 10.1109/CICT.2013.6558259.

[2] R. J. Ruben, "Sign language: Its history and contribution to the understanding of the biological nature of language," *Acta Otolaryngol. (Stockh.)*, vol. 125, no. 5, pp. 464–467, May 2005, doi: 10.1080/00016480510026287.

[3] G. Plouffe and A.-M. Cretu, "Static and dynamic hand gesture recognition in depth data using dynamic time warping," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 2, pp. 305–316, Feb. 2016, doi: 10.1109/TIM.2015.2498560.

[4] Z. Zhou *et al.*, "Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays," *Nat. Electron.*, vol. 3, no. 9, pp. 571–578, Sep. 2020, doi: 10.1038/s41928-020-0428-6.

[5] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1610–1618, doi: 10.1109/CVPR.2017.175.

[6] S. Masood, A. Srivastava, H. C. Thuwal, and M. Ahmad, "Real-time sign language gesture (word) recognition from video sequences using CNN and RNN," in *Intelligent Engineering Informatics*, V. Bhateja *et al.*, Eds. Singapore: Springer, 2018, pp. 623–632, doi: 10.1007/978-981-10-7566-7_63.

[7] A. H. Vo, V.-H. Pham, and B. T. Nguyen, "Deep learning for Vietnamese sign language recognition in video sequence," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 4, pp. 440–445, Aug. 2019, doi: 10.18178/ijmlc.2019.9.4.823.

[8] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4207–4215.

[9] O. Koller, "Quantitative survey of the state of the art in sign language recognition", *arXiv:2008.09918*, 2020.

[10] G. H. Samaan *et al.*, "MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition," *Electronics*, vol. 11, no. 19, Oct. 2022, doi: 10.3390/electronics11193228.

[11] G. R. S. Murthy and R. S. Jadon, "A review of vision-based hand gestures recognition," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 405–410, 2009.

[12] P. Garg, N. Aggarwal, and S. Sofat, "Vision-based hand gesture recognition," *World Acad. Sci. Eng. Technol.*, vol. 49, pp. 972–977, 2009.

[13] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural*

Comput. Appl., vol. 28, no. 12, pp. 3941–3951, 2017, doi: 10.1007/s00521-016-2294-8.

[14] F. R. Cordeiro, T. L. M. Barreto, J. P. Teixeira, and A. G. Guimarães, "A convolutional neural network with feature fusion for real-time hand posture recognition," *Appl. Soft Comput.*, vol. 73, pp. 748–766, 2018, doi: 10.1016/j.asoc.2018.09.010.

[15] P. Rathi, N. S. Chauhan, and R. Bhardwaj, "Sign language recognition using ResNet50 deep neural network architecture," in *Proc. Int. Conf. on Next Generation Computing Technologies (NGCT)*, 2020.

[16] P. Bhatia and A. Wadhawan, "Deep learning-based sign language recognition system for static signs," *Neural Comput. Appl.*, vol. 32, pp. 7957–7968, 2020.

[17] W. Wang, C. Wang, and J. Wu, "American sign language recognition using multidimensional hidden Markov models," *J. Inf. Sci. Eng.*, vol. 22, pp. 1109–1123, 2006.

[18] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Sign language recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2015.

[19] C. Lugaresi *et al.*, "MediaPipe: A framework for building perception pipelines," *arXiv preprint* arXiv:1906.08172, 2019.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015, doi: 10.1038/nature14539.

[21] S. Duan, L. Wu, A. Liu, and X. Chen, "Alignment-enhanced interactive fusion model for complete and incomplete multimodal hand gesture recognition", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4661–4671, 2023.

[22] C. Li, D. Zhao, B. Zhang, W. Chu, and Y. Luan, "Research on a fast signal recognition method for a laser screen measurement system based on LSTM", *Measurement*, vol. 254, p. 117905, 2025.

[23] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: adaptive multi-modal gesture recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2015.

[24] R. Bowden, "Deep sign: Hybrid CNN-HMM for continuous sign language recognition", in *Procedings of the British Machine Vision Conference 2016*, 2016.

[25] G. Devineau, F. Moutarde, W. Xi, and J. Yang, "Deep learning for hand gesture recognition on skeletal data", in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 106–113.