Journal of Science and Technology – Engineering and Technology for Sustainable Development
Volume 36, Issue 1, March 2026, 097–112

A Comprehensive Review of Vietnamese Sign Language Recognition Techniques

Hoang Quang Huy[1], Nguyen Ngoc Tram[1], Nguyen Huu Trung[1], Nguyen Ngoc Anh[2], Pham Thi Viet Huong[3], Tran Anh Vu[1*]

[1]Hanoi University of Science and Technology, Ha Noi, Vietnam
[2] Viet Duc College of Medicine and Medical Devices, Ha Noi, Vietnam
[3]International University, Vietnam University, Ha Noi, Vietnam
[*] Corresponding author email: vu.trananh@hust.edu.vn

Abstract

The paper presents a systematic quantitative literature review of Vietnamese Sign Language recognition techniques developed between 2015 and 2025. VSL recognition plays a vital role in bridging communication gaps and enhancing accessibility for the deaf and hard-of-hearing community in Vietnam. To identify and synthesize current trends and challenges, we conducted a structured search and screening process across major academic databases. These works were analyzed based on recognition approach (e.g., computer vision, wearable sensors, data-driven methods, and multimodal data fusion), datasets used, feature extraction strategies, classification models, and performance metrics. Descriptive statistics were used to map the evolution of methods over time, while comparative analyses highlighted the strengths and limitations of different techniques across real-time and static recognition tasks. Our findings indicate a growing shift towards deep learning and sensor fusion methods, though limitations persist in dataset availability, model generalizability, and real-world deployment. This review provides critical insights into current research gaps and offers guidance for future work on scalable, culturally adaptive VSL recognition systems.

Keywords: Computer vision, machine learning, sign language recognition, Vietnamese sign language.

1. Introduction

For over one million deaf and hard-of-hearing individuals in Vietnam [1] communication takes a different form – Vietnamese Sign Language (VSL). Sign language provides an alternative linguistic framework in which words are conveyed through hand gestures, facial expressions, and body movements. However, sign language remains relatively unfamiliar in Vietnam. Currently, there are only a few dozen qualified Vietnamese Sign Language interpreters nationwide. As a result, deaf people face significant challenges when participating in activities within families, schools, workplaces, and society [2].

Vietnamese Sign Language is a distinct language with its own vocabulary and grammar, different from the spoken Vietnamese language used by hearing individuals. This is specified in the national standard for sign language for individuals with disabilities, which outlines the system of Vietnamese sign language for use by the hearing and speech-impaired. The Vietnamese sign language system defined in this regulation includes a set of arrow symbols, a table of letter symbols with tone marks, a table of numeral symbols, and a glossary of sign language terms with only 408 words (signs). Vietnamese sign language has six extended vowels: Ă, Â, Ê, Ô, Ơ, Ư which are the combination of four letters A, E, O, U with three markers or accents (˘, ˆ, ᾽) apart from the twenty-three base letters, as seen in Fig. 1, and four loan letters (F, J, W, Z) [3]. All letters are considered static gestures, except for J and W.

Fig. 1. Vietnamese Sign Language Alphabets

In addition, Vietnamese has five special symbols called tones (Fig. 2) which could change the entire meaning of the word[3]. For example, the word "đo" has multiple definitions:

- "đo" – without tone = measure
- "đo" – rising tone "đó" = there
- "đo" – falling tone "đò" = ferry
- "đo" – falling rising tone "đỏ" = red
- "đo" – high rising tone "đõ" = hollow tree trunk to raise bees

p-ISSN 3093-3242
e-ISSN 3093-3579
https://doi.org/10.51316/jst.188.etsd.2026.36.1.12
Received: Jun 11, 2025; Revised: Sep 25, 2025;
Accepted: Sep 25, 2025; Online: Feb 12, 2026.

- "đo" – low constricted tone "đọ" = compete

Understanding the significance of bridging this communication gap and consequently, offering an equal opportunity for deaf students, linguistics and researchers are committed to developing a proficient VSL gesture recognition system through innovations in computer vision, and machine learning. These breakthrough studies have significantly improved the visual interpretation of VSL gestures. This can be seen in the work of Vo *et al.* [1], who employed Hidden Markov Model (HMM) and Dynamic Time Warping (DTW) for dynamic hand gesture recognition. Additionally, the introduction of wearable sensor technology has facilitated highly accurate results. Studies [4-6] that incorporated wearable sensor-embedded devices have yielded exceptional results with a high probability of integrating into real-time applications due to low computational cost.

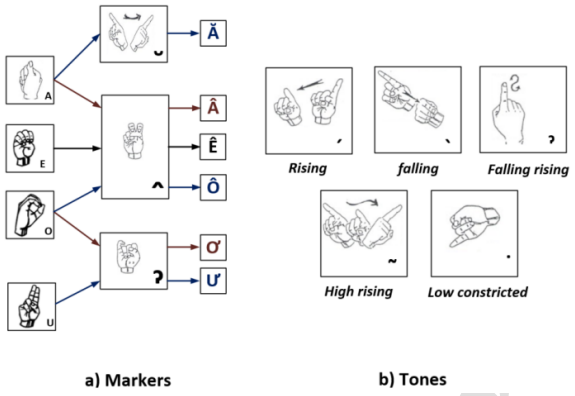

Fig. 2. Vietnamese sign language for six vowels with tones and markers

Progress in VSL recognition has been marked by further milestones, as seen in studies by Vo *et al.* [4]who explored spatial and scene-based features and deep learning techniques for VSL recognition in video sequences. Pham *et al.* [5] further underscore this progress by introducing a combination of Gaussian distribution and correlation coefficient techniques to identify moving objects in video frames. They also used GoogLeNet, a specialized neural network, to extract features from the videos, and Bidirectional Long Short-Term Memory (BiLSTM) to classify these sequences. The authors achieved an accuracy of 98.13%, much higher than that of HMM, SVM, VGG16-BiLSTM, and Alexnet-BiLSTM.

Vietnamese Sign Language Recognition (VSLR) techniques encounter several distinct challenges not as prevalent in the recognition of other sign languages such as American Sign Language (ASL) or British Sign Language (BSL). One of the difficulties is the scarcity of comprehensive VSL datasets necessary for the training and validation of machine learning models, a challenge often not presented for more widely studied sign languages. Moreover, the dynamic nature VSL, including its unique handshapes, movements, and regional variations, requires complex algorithms for accurate interpretation. The development of such algorithms is impeded by the lack of technological infrastructure in Vietnam. The process of annotating VSL data accurately is complicated by the need for annotators who are not only proficient in VSL but also skilled in dataset creation.

Additionally, the lack of standardization across VSL leads to inconsistencies in sign usage and difficulty in developing an effective recognition system. This issue is even aggravated by the scarcity of resources allocated to VSL research, which is often clouded by other projects due to the smaller size of the deaf and hard-of-hearing community in Vietnam. Finally, translating VSL recognition research into user applications involves overcoming obstacles related to user interface design, hardware system, and the incorporation of cultural and contextual nuances.

Most recently, a fusion of computer vision, machine learning, and human-computer interaction has been utilized in an effort to improve the precision in VSL recognition and classification as seen in the study by Khang V. Nguyen *et al.* [7]. The authors proposed an exceptional novel approach that captured data with a Vicon motion capture system, utilized Deep Neural Network (DNN) for training, and a semi-supervised algorithm for gesture recognition and classification in human-robotic applications. However, due to the limited training data, they only achieved moderate recognition rates between 50% and 71.43%. The authors also raised safety concerns over its real-world applications and aimed for safer, more seamless human-robot interactions in the future.

While several reviews on sign language recognition exist, this paper presents a unique contribution to the literature through three key distinctions. Firstly, it offers a specialized scope as the first comprehensive review focused exclusively on Vietnamese Sign Language, addressing its unique linguistic characteristics. Secondly, its updated timeframe covers the critical decade of development from 2015 to 2025, capturing the rapid evolution from traditional machine learning to the most recent state-of-the-art architectures. Finally, this paper employs a systematic, evidence-based approach to analyze a curated set of publications, providing a clear and structured overview of the trends, challenges, and future directions specific to VSL recognition.

We will examine various methodologies, including machine learning, deep learning, data-driven approaches, wearable devices, sensor fusion, gesture segmentation, and real-time applications. The remaining paper is organized as follows: Section 2 presents the methodology for the systematic review and bibliometric analysis of the selected articles for studies. Section 3 summarizes different approaches to VSL recognition.

Section 4 describes the methodology. Finally, Section 5 concludes the paper.

2. Bibliometric Analysis

This paper provides a comprehensive review and analysis of different approaches for recognizing VSL gestures. In conducting this review, we selected research papers that significantly contribute to advancements in VSL recognition methodologies.

We analyzed a total of more than 60 referenced papers published between 2015 and 2025. Keywords such as “Vietnamese Sign Language”, “Vietnamese Sign Language Recognition”, “Machine Learning”, “Deep Learning” were applied to filter through academic databases and journals including Scopus, Web of Science, Science Direct, IEEE Xplore, Springer and Google Scholar as in Fig 3. Specifically, we focused on papers demonstrating Vietnamese Sign Language recognition and classification. Major VSL-related papers and their focus were listed in Section 3, Table 4.

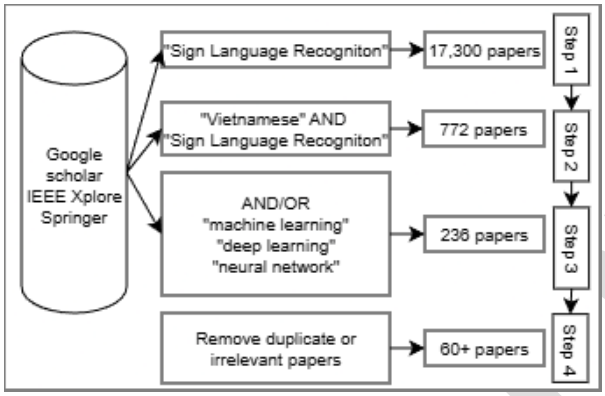

Fig. 3. Flowchart of gathering and analyzing publication data from Google scholar, IEEE Xplore, and Springer database

3. Vietnamese Sign Language Recognition Approach

3.1. Wearable-Sensor-Based Approach

Wearable devices, especially sensor gloves, have significantly advanced the precision VSL recognition capturing detailed hand movements and gestures,. These gloves incorporate flex sensors, accelerometers, and gyroscopes, each catering to different motion tracking [6, 11]. Flex sensors are mounted on fingers to determine finger bending with their resistance varying with the degree of bend and on the palm [9]. Accelerometers measure both static and dynamic acceleration forces on the hand, while gyroscopes record orientation and rotational movement [11]. Values from these sensors are then processed by the main processor of the gloves to turn hand movements into digital signals before displaying the corresponding letter according to rule sets. The application of glove-based approach could be seen in the work by Lam *et al.* [3]. The authors developed a sensor glove with ten flex sensors and one accelerometer to recognize VSL. The results were astounding with precision rate reaching 100% for twelve letters and only three letters, Ă, Ơ, and Ư, were on the lower side due to its complex Z-axis hand rotation.

A different application of wearable sensor devices could be seen in the work by Hong Quan Nguyen *et al.* [6]. They proposed a new set of twelve dynamic hand gestures (G1 to G12) to develop a system for controlling home appliances through hand gestures captured by wrist-worn devices and achieved impressive accuracy of 98.48% and 96.23%. One drawback is its inconvenience since participants are required to wear gloves throughout the signing process[4].

3.2. Vision-Based Approach

Computer vision has emerged as an innovative and highly efficient VSL recognition approach. It is currently one of the most applied techniques. In vision-based methods, imaging and processing techniques are applied to extract features from pictures of hand movements. As such, it is often preferred over contact-based approach since signers are not required to wear sensor devices at all [8, 14]. One significant improvement of computer based VSL recognition technique is the ability to recognize rapid and complex gestures compared to traditional video analysis. Visual data is, respectively, processed and interpreted more accurately. It employs a combination of filters and edge detectors to detect hand movements.

Multiple studies [9-12] that incorporated the use of Microsoft Kinect devices and depth cameras have demonstrated exemplary accuracy. In [8], the authors achieved an accuracy of 91% utilizing Support Vector Machine (SVM) for recognition on a sequence of depth images captured by a Microsoft Kinect sensor. A similar study also utilized skeletal joint maps collected from Kinect device yielded exceptional accuracy of 74% – 100% for static letters, 97% – 100% for dynamic diacritics, and 90% for real-time application [12].

It is worth noting that the accuracy of vision-based method is tremendously affected by noises, brightness, viewpoint variation, and complex backgrounds [13]. As such, achieving such outstanding results requires complete and precise background segmentation of the visual data. However, simple segmentation algorithms are often unable to remove crowded backgrounds with many unwanted objects.

3.3. Data-Driven Approach

Given the dynamic nature of Vietnamese sign languages, it is essential to have a vast dataset that covers a wide range of gestures, hand orientations, and facial expressions. However, collecting such comprehensive datasets poses significant challenges [14, 15]. Our research reveals a critical gap in dataset resources. As of now, there are only a handful of publicly available VSL datasets. Only 4 referenced papers in our review utilized publicly VSL available datasets, which are D-VSL dataset [10], VSL-ADT

dataset [16], VNTC dataset [17], and Multi-VSL [18] for VSL recognition and classification.

One major issue with generating sign language dataset is annotating data signals coming from sensors annotation is often expensive and time-consuming. Additionally, the lack of signs representing regional dialects causes the incompleteness of the dataset [12].

To overcome some of these challenges, recent studies have explored the use of digitally generated data. This refers to the process of converting a common Vietnamese sentence into image, video, or 3D model of its corresponding signs [19, 20]. There exist various sign language translation systems worldwide but machine translation of VSL remains mostly undiscovered [21]. It could not only provide data augmentation but also help in representing hard-to-capture signs. The author proposed a simple method of data augmentation based on Wordnet with augmented data scoring much higher Bilingual Evaluation Understudy (BLEU) score than the original one. The limitation of this, however, lies in its unreal gestures compared to human signers [19].

In [4, 16] the authors used a different technique of data augmentation to increase the quality of their dataset. Due to the discrepancy among signers, they rotated the original image ±5 and ±10 degrees and added salt/pepper noise with 0.05 probability to gain more information of the hand positions. The newly introduced Multi-VSL dataset [18] provides the first large-scale, multi-view resource dedicated to VSL. This corpus features over 84,000 multi-view videos, 1,000 glosses (signs), and 30 signers, positioning it as the largest multi-view dataset of its kind across all sign languages. The utilization of this multi-view data demonstrated substantial model accuracy improvement, achieving up to 19.75% greater performance compared to single-view data by effectively differentiating between visually ambiguous signs that appear similar from a frontal perspective.

3.4. Data Fusion Approach

Another solution to improve the completeness of the dataset is to combine inputs from various approaches. When these data are blended using advanced algorithms, such as neural networks, it results in minimized sensor discrepancies and improved gesture recognition results. Previous studies have utilized this approach and achieved promising results for gesture recognition, however, their limited data size led to overfitting during the training process.

In an attempt to address this drawback, authors in [15] contributed a large dataset, a fusion of video data and sensor data from two smartwatches and proposed the first self-training fusion model for SLR (termed STSLR). Experimental results reflect that STSLR performs significantly better when being trained with VSL and other datasets compared to competitive models. Additionally, merging collected data from vision-based and sensor-based methods increases depth layer to hand gestures. Table 1 shows a comparison of VSL recognition methods in terms of cost, input device, efficiency, and implementation. In this study, we focused mostly on vision-based approach.

The preceding sections have categorized VSL recognition research into four primary approaches. While each has its unique characteristics, they often share a common underlying workflow. The following section provides a detailed breakdown of the typical five-stage methodology–from initial data acquisition to final sign translation–that forms the backbone of most modern VSL recognition systems.

Table 1. Comparison table of VSL recognition approach

Method	Cost	Input Device	Efficiency	Implementation
Sensor Devices	Avg-High	Sensor-embedded devices	High	Real-time processing Integration with computer vision
Vision-based	Avg	Depth camera, Microsoft Kinects	High	Complex High computationa resources
Data-driven	High	Cameras, digitally generated	High (depends on data quality)	Data augmentation Pre-processing & post-processing
Data fusion	High	Multiple sensor	Very High	Integration of multiple data sources

4. Methodology

4.1. Workflow

There are different methods of sign language recognition, but they usually consist of five steps that start with capturing data through cameras, webcam etc. This is followed by pre-processing, where the data is refined through segmentation and normalization. Before collected data could be classified into distinct signs, relevant features are captured during feature extraction step. In the final step, classified sign gestures are translated into text or speech.

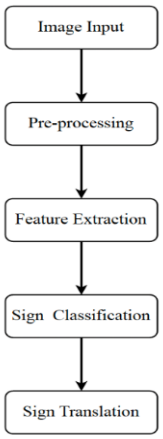

Fig. 4. Methodology flowchart

4.1.1. Image acquisition and preprocessing

Image acquisition refers to the capturing of visual data, typically using Microsoft Kinect or depth camera. The data collected are then subjected to preprocessing.

Preprocessing steps generally include segmentation, normalization, and thresholding. The main objective of segmentation is to eliminate redundancies and take only the Region of Interest (ROI) [14]. During normalization, images are scaled to reduce computational cost and increase processing speed as well as remove illumination effects [3]. Smoothing and blurring are also applied to increase the quality of the data [4].

Vo *et al.* [20] introduced depth information for segmentation to address the issue of environmental background when using skin color filers. They also applied some morphological/spatial filters for noise removal and boundary smoothing. Nguyen *et al.* [27] used histogram equalizer and low pass filter to remove the noise.

However, segmenting continuous sign language into individual gestures remains a challenge, especially given the fluidity of human signing without clear pauses. Researchers have highlighted the importance of identifying the start and end points of signs [15, 22]. Recently, deep learning has been incorporated into data segmentation process to increase accuracy [26].

4.1.2. Feature extraction

Feature extraction is the next critical step after pre-processing. Its purpose is to reduce dimensionality of the image into individual features and extract relevant features for classification. Vo *et al.* [13] proposed the application of both spatial feature and scene-based local descriptors to extract features from each VSL word in a video sequence. Recently, machine learning has been taking a dominant position in the field of feature extraction, especially Convolutional Neural Networks (CNNs).

4.1.3. Sign classification

Sign classification refers to the categorization of the extracted features into different groups of sign language gestures. This stage utilizes various machine learning and deep learning methods. SVM is a supervised learning algorithm most recently used in classification.

4.1.4. Sign translation

The final stage involves translating the classified sign gestures into text or speech, making the recognition process complete. In this stage, a built-in function or a predefined library is applied to convert signs into readable text or audible speech.

4.2. Feature Extraction Methods

4.2.1. Traditional feature extraction techniques

Despite the rise of machine learning in feature extraction, these manual features are still applicable when working with dataset that lacks labeled data and computational resources are limited. HOG descriptor is one of the appearance descriptors that captures the edge boundary or gradient structure of the image. It is geometric and transformations (translation, rotation) invariant. It calculates how local intensity gradient orientation of a detection window [13]. GIST [16] is a scene-based descriptor, aims to form a low-dimensional representation of the data. It is computed by binding the image with 32 Gabor filters at 4 scales, 8 orientations, then obtaining 32 same-sized feature maps of the same image. Histograms of Oriented Optical Flow [21] is a motion-based descriptor. Each optical flow vector $v = [x, y]^T$, with direction $\theta = \tan^{-1}(\frac{x}{y})$, in the range $\frac{-\pi}{2} + \pi\frac{b-1}{B} \leq \theta < \frac{-\pi}{2} + \pi\frac{b}{B}$, will add $\sqrt[2]{x^2 + y^2}$ to the sum in bin b, with $1 \leq b \leq B$, out of B bins. The sum is then normalized to 1. Some researchers apply a single descriptor, while others combine two or three descriptors, as demonstrated in [12, 16].

4.2.2. Convolutional Neural Networks

CNN is known for its universal application in feature extraction and classification due to its exceptional ability to process and analyze image data. They are particularly effective in extracting spatial features such as hand shapes and finger movements from images or video sequences. Its application in Natural Language

Processing (NLP) model is illustrated in Fig. 5.

CNN is a neural network with a multilayer of convolution and pooling. It consists of nonlinear activation functions such as *RELU* or *tanh* to create information for the next layers. It is worth noting that CNN is location and composition sensitive. If an object is put under different angles, the accuracy of the model will vary significantly [6].

One efficient CNN architecture is GoogLeNet, as seen in [5]. GoogLeNet is a 22-layers deep convolutional network with 7 million parameters. It has been trained on over a million images and can classify into 1000 object categories. GoogLeNet works by converting video inputs into feature vectors as the output of the activations function on the last pooling layer of the network. For each input, there is a default vector with 1024 features.

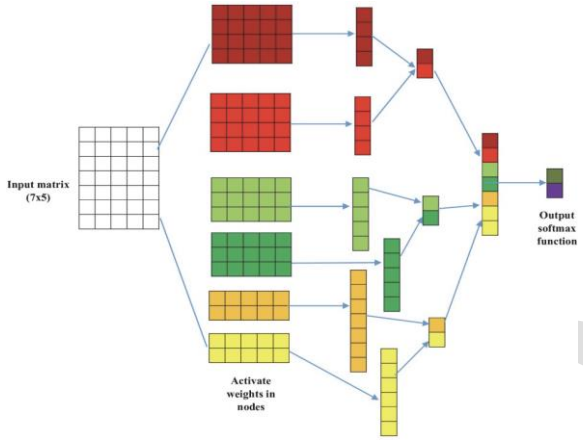

Fig. 5. Process of image recognition using Convolutional Neural Networks

Other CNN models are MobileNet v1 and VGG16, which were used in [22] to extract features. ResNet-101 is another CNN model that yielded exceptional results. It has similar architecture as the VGG model but with more layers. However, it is not often preferred due to higher computational cost, time, and memory.

4.2.3. Long short-term memory networks

Long short-term memory (LSTM) is another network frequently applied in dynamic feature extraction. It is built to learn from other parameters and prevent long-term dependency by remembering information over an extended period. Its critical feature, the cell state (Fig. 6), allows for easy information transmission through linear interaction. Nguyen *et al.* [23] proposed the integration of LSTM to increase performance since there was no coherence among different gestures.

Firstly, the model needs to determine which information should be removed from the cell state. A sigmoid function or "forget gate layer" carries out this decision which takes inputs h_{t-1} and x_t and returns C_{t-1}. If output equals 1, information is kept, otherwise, all information will be discarded. Secondly, the model must determine which new information to save into the cell state. The input gate layer assesses which values are to be updated, and the tanh layer will create a new vector. These two values are then combined to form a new state. Finally, the system runs through the sigmoid layer to decide which cell state part to export. Cell state (Fig. 6) is processed by tanh function and then multiplied with cell state output (Fig. 7).

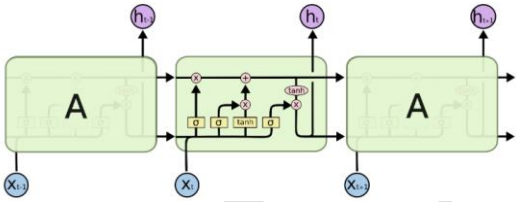

Fig. 6. LSTM cell state diagram

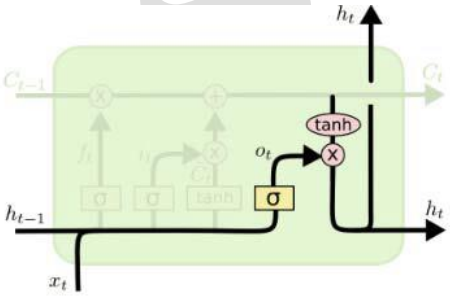

Fig. 7. LSTM output diagram

4.2.4. Hybrid models

The combination of CNNs for spatial feature extraction and LSTM networks for capturing temporal dependencies is an effective and increasingly common approach. These hybrid models are particularly powerful due to their ability to simultaneously analyze both the shape and movement of the hands, thereby yielding higher recognition accuracy [15]. A prime example of this approach in [2], the authors introduced 1D-CNN-LSTM and 1D-CNN-biLSTM, two fusions of deep convolutional network and LSTM models, which have demonstrated superior performance across various datasets. By using a 1D-CNN instead of a 2D-CNN, the model can learn feature patterns directly from the raw data sequence, rather than processing complex spatial features. This approach also helps reduce the risk of overfitting when training on limited datasets. Furthermore, the Bi-LSTM architecture enhances the model's reliability by processing the data sequence in both forward and backward directions. This allows the model to gather context from both past and future elements in the sequence, leading to a deeper understanding of complex temporal patterns.

4.3. Classification Methods

In VSL recognition, classification methods have been evolving rapidly, especially with the birth of more

advanced technology. Trending classification methods are characterized by their enhanced accuracy, efficiency, and ability to handle complex features.

4.3.1. Support vector machine

SVM, first introduced by Chervonenkis and Vapnik, is a powerful, supervised machine learning model for pattern recognition, data analysis, classification, and regression.

4.3.2. Hidden markov models

HMM is a statistical model, particularly effective for dealing with time-series data. It is assumed that the system is a Markov process with hidden states. It has an exceptional ability to deal with the complex nature and temporal of sign language. They are often used in conjunction with other feature extraction techniques to improve overall classification results.

Modern deep learning methods often perform classification in an end-to-end fashion, combining feature extraction and classification within a single neural network. Several representative models for VSL include the following.

4.3.3. Self-training fusion learning model

Self-training fusion learning model for SLR (STSLR) was first introduced by [15] to address the issue of time-consuming, costly manual data labeling, and time synchronization of multiple datasets. This model is applied to convert video input into sensor data that simulates data collected from wearable sensors. The model architecture comprises two main stages: (1) generating simulated sensor features and (2) fusing these with video data features.

In stage one, a pre-trained 1D inception network to extract a 128D feature vector from the sensor signal and a 3D inception network to create a 128D feature vector of the video signal are deployed. In stage two, input video is processed through two separate Inception 3D networks. This process extracts the 128D vector of the simulated sensor signal, and the 1024D feature vector of the video input. These vectors are then combined with Mixer MLP (Multi-Layer Perceptron) to reach the final result. So far, the authors in [15] have been the first to apply STSLR in classifying VSL.

However, since the model requires time-synchronized video and real sensor data for training, its accuracy is heavily dependent on the quality of the data. The use of multiple neural networks also leads to complex computational issues.

4.3.4. Diffusion-guided graph convolutional network

The paper [24] proposes a new architecture combining Graph Convolutional Network (GCN), attention mechanisms, and diffusion models. For classification, an Image-to-Graph Transfer Model is employed, which converts video frames into skeletal graphs, representing the structure of the sign language gestures. A Text Embedding Model is used to generate embeddings from textual descriptions, which assist in guiding the recognition of gestures. Diffusion-based Attention Module enhances the model's ability to capture temporal and spatial relationships within the gestures, improving overall recognition accuracy.

4.3.5. Transformer-based model

The authors in [18] used MViT v2, Video Swin Transformer and VTNHCPF, transformer-based models that leverage self-attention mechanisms. MViT uses a Vision Transformer (ViT) backbone that processes video as a sequence of patches (temporal and spatial tokens). It incorporates multiple views or perspectives of the data, which is particularly useful in scenarios like multi-view video or action recognition. VTNHCPF is an extended version of Video Transformer Network (VTN), where additional components like hand crop and pose flow are integrated. It enhances the ability of VTN to focus on critical areas (hands, body poses) and track temporal changes in video.

Transformer-based models are particularly well-suited for action recognition and gesture tracking. They can outperform traditional CNNs in tasks like video classification, making them a strong candidate for dynamic VSL recognition.

4.3.6. Uniformer-based model

The paper [25] proposes the UF3V (UniFormer 3-View) architecture, a model based on the UniFormer backbone. UniFormer is a hybrid model that combines a Convolutional Neural Network (CNN) and a Transformer. Its architecture includes four stages; the first two stages utilize 3D CNN layers for local feature extraction and strong inductive bias, while the latter two employ a Multi-Head Self Attention (MHSA) mechanism to capture global dependencies. This hybrid design is particularly effective for limited-size datasets, which is a common characteristic of Isolated Sign Language Recognition (ISLR) datasets. The UF3V model specifically uses three UniFormer-S backbones to process three data streams from different views simultaneously. It integrates a Dual Co-Attention (DCA) mechanism to intelligently fuse information from adjacent views. This combination helps optimize information exchange and leverages the discrepancies between views to distinguish difficult signs, such as Visually Indistinguishable Signs (VISigns).

4.4. Datasets

Based on the analysis results above, while existing studies in the field of Vietnamese sign language recognition have made significant advancements, they still face notable limitations. These include the small scale of datasets (as in Table 2), which restricts the generalizability of the models, and the difficulty in verifying the quality of the data due to the lack of public access except paper [24]. Additionally, the

reproducibility of results remains challenging because the code and implementation details are not publicly available, preventing independent verification and comparison. A comparison with other sign languages can be seen in Table 3. Some languages can have more than one dataset.

To further contextualize these findings, a comparison with datasets from more extensively studied sign languages, particularly ASL, is instructive. The data ecosystem for ASL is notably mature, featuring numerous large-scale and publicly accessible datasets. For instance, corpora such as ASL Citizen (over 83,000 videos) and PopSign ASL (over 214,000 videos) provide a wealth of resources for training complex deep learning models. Similarly, other languages, such as German Sign Language, are supported by substantial sentence-level datasets like RWTH-Phoenix-Weather-2014 (over 45,000 videos).

This comparison starkly highlights the disparity with VSL. Although the field has made a significant advancement with the introduction of the Multi-VSL dataset (over 84,000 videos), many other VSL datasets listed in the table are considerably smaller in scale and, critically, have an "Unknown" availability status. The lack of large-scale, publicly accessible, and meticulously annotated datasets remains one of the most significant barriers, limiting the developmental pace and generalizability of VSL recognition models compared to their counterparts in more widely-resourced sign languages.

Table 2. Brief comparison between sign language dataset

Dataset	Language	Year	Videos	Number of signs	Level	Availability
ASLLVD	American	2008	~9800	3300	word	Public
MS-ASL	American	2019	>25000	1000	word	Public
WLASL	American	2020	21083	2000	word	Public
ASL Citizen	American	2023	83399	2731	word	Public
PopSign ASL	American	2023	214326	250	word	Public
RWTH-Phoenix-Weather-2014	German	2014	45760	1200	sentence	Public
KETI	Korean	2019	14672	105	sentence	Public
INCLUDE	Indian	2020	4287	263	word	Public
CSL	Chinese	2018	5000	178	words	Public
SIGNUM	German	2010	33210	450	sentence	Contact author
DGS Kinect	German	2012	3000	40	word	Public
DEVISIGN-D	Chinese	2016	24000	500	word	Contact author
LSA-T	Argentina	2022	14880	N/A	sentence	Public
LSFB-CONT	Belgium	2021	85000	6883	sentence	Public
LSFB-ISOL	Belgium	2021	50000	400	word	Public
Slovo	Russian	2023	20000	1000	word	Public
AUTSL	Turkish	2020	38336	226	word	Public
MMAuslan	Australia	2024	282000	3215	word	Public
VSL-WRF-01	Vietnamese	2019	480	12	word	Contact author
VSL-WRF-02	Vietnamese	2019	600	15	word	Contact author
ViSL oneshot	Vietnamese	2024	436,400	4364	word	Public
ViSL	Vietnamese	2024	50M	3234	sentence	Contact author
P-SL	Vietnamese	2025	5068	50	word	Contact author
QIPEDC-based	Vietnamese	2025	500	500	word	Contact author
Multi-VSL	Vietnamese	2025	84764	1000	word	Public

Table 3 A comparative analysis of leading global sign language lechnology providers and applications

Application/ Platform	Primary Function	Supported Sign Language(s)
Hand Talk	One-way Translation (Text -> Sign Video), Language Learning	ASL, Brazilian Sign Language (Libras)
Sign-Speak	Two-way Translation (Sign <-> Speech)	ASL
Signapse (SignStudio, SignStream)	One-way Translation (Text -> Sign Video)	British Sign Language (BSL), ASL
SignAll	Sign Recognition (Sign -> Text), Language Learning	ASL
KinTrans	Sign Recognition (Sign -> Text/Speech)	ASL, Unified Arabic Sign Language
It's a Sign	Language Learning, Dataset Creation	ASL
SignGemma	AI model designed to perform sign-language translation	ASL
Lingvano	Language Learning	ASL, BSL, Austrian Sign Language (ÖGS)
The ASL App	Language Learning	ASL, with development support for Icelandic (ÍTM) and Quebec (LSQ)
SignSchool	Language Learning	ASL
WeSign	Sign Recognition (Sign -> Text), Language Learning	VSL

A critical distinction emerges when analyzing the scope of these datasets. A significant number of corpora for other sign languages provide data for sentence-level recognition, which is essential for developing continuous sign language translation systems. For example, established datasets such as RWTH-Phoenix-Weather-2014 (German), LSA-T (Argentinian), and LSFB-CONT (Belgian) are all explicitly designed for sentence-level analysis. In stark contrast, VSL research and dataset collection have been overwhelmingly focused on word-level recognition. With the exception of one sentence-level dataset whose availability is unknown, virtually all existing VSL corpora consist of isolated signs. This fundamental gap significantly impedes progress toward developing systems capable of true linguistic comprehension and continuous VSL translation, representing a major area for future research and data development.

Having detailed the technical pipeline and the evolution of specific methods for feature extraction and classification, the discussion now shifts from theoretical frameworks to practical implementation. The following section explores the real-world applications of sign language recognition technology, examining how these research advancements are being translated into commercial products and assistive solutions that create a tangible impact on the deaf and hard-of-hearing community globally.

5. Applications of Sign Language Recognition: From Research to Real-World Impact

This technological maturation from research concepts to functional systems has enabled a diverse ecosystem of applications. These solutions range from consumer-focused language learning tools and dictionaries to enterprise-level, real-time translation services that promote accessibility in various sectors.

Table 3 provides a comparative analysis of several leading global platforms and research projects, highlighting their primary functions and the specific sign languages they support, which will be discussed in this section.

5.1. Technological Foundations: From Sensor Gloves to Computer Vision

The evolution of sign language recognition technology has transitioned from intrusive, device-based methods to accessible, computer vision-centric solutions. Initially, systems relied on sensor gloves to capture precise motion data; however, these were cumbersome and impractical for daily use.

Today, the predominant approach is vision-based, utilizing standard 2D cameras (e.g., webcams, smartphones) and advanced AI models. CNN and Tranformers are fundamental for interpreting both static and dynamic signs. The introduction of frameworks like Google's MediaPipe has revolutionized the field by enabling real-time, accurate tracking of hand, face, and body landmarks. This allows developers to focus on the core challenge of language recognition rather than the complexities of motion tracking.

This technology has advanced in two primary directions.

- Recognition (Sign-to-Text/Speech): It is a complex AI task focused on interpreting video of a signer and translating it into written text or synthesized speech.

- Synthesis (Text/Speech-to-Sign): It is the process of generating sign language videos from text input. This is typically achieved using 3D avatars (e.g., Hand Talk) or advanced Generative AI to create photorealistic videos of human signers (e.g., Signapse).

5.2. Global Implementations: From Consumer Applications to Enterprise Solutions

On the global market, sign language recognition technology has been successfully commercialized, primarily through a business-to-business (B2B) model.

- Business-to-Consumer (B2C) Applications: Applications like Brazil's Hand Talk have gained popularity as "pocket dictionaries" and learning tools, employing 3D avatars to translate text into signs.

- Enterprise Solutions (B2B SaaS): This has emerged as the most lucrative and scalable business model. Companies provide "Accessibility-as-a-Service" to other businesses rather than selling standalone applications.

Sign-Speak: Offers the first real-time, two-way translation system, which can be integrated into virtual meeting platforms or accessed via QR codes at customer service points. **Signapse**: Specializes in high-quality text-to-sign synthesis using Generative AI to produce photorealistic videos. Their technology is widely deployed in the transportation sector for station and airport announcements. SignAll, KinTrans: Provide solutions that require more specialized hardware (e.g., multiple cameras, depth sensors) to achieve high accuracy in controlled environments such as classrooms or banking kiosks.

The exploration of global applications demonstrates the significant commercial and social potential of sign language recognition technology as in Table 3. Building on this broader context, the concluding section of this paper synthesizes the key findings of our systematic review, focusing specifically on the trajectory of VSL recognition research between 2015 and 2025, summarizing the methodological trends, persistent challenges, and critical directions for future work.

6. Conclusion

An analysis of the research landscape from 2015 to 2025 indicates a significant transformation in VSL recognition. This evolution, detailed in Table 4, is primarily defined by progressive methodologies, dynamic dataset strategies, and substantial gains in reported model accuracies. The chronological evolution of methodologies began with traditional machine learning algorithms like HMM and Dynamic Time Warping (DTW) in the early period (2015-2017), often combined with Support Vector Machines (SVM) for tasks such as dynamic hand gesture recognition, achieving up to 87.9%. This era also saw the introduction of wearable sensor technology, like sensor gloves, capable of high precision for individual letters. From 2018 to 2020, there was a distinct move towards deep learning, particularly CNNs and LSTM, which captured both spatial and temporal features in video sequences, leading to accuracies such as 95.83%. The most recent period (2021-2025) highlights advanced deep learning architectures, including GoogLeNet combined with BiLSTM (achieving 98.13% test accuracy), and the increasing prominence of multimodal data fusion which integrates video and sensor data. Cutting-edge approaches now include Transformer-based models (MViT v2, Video Swin Transformer) and hybrid models like Uniformer-based (UF3V) and Diffusion-guided Graph Convolutional Networks (TeDG), designed to handle complex spatiotemporal dynamics and leverage multi-view data for enhanced recognition.

An analysis of trends in dataset sizes and sources highlights the persistent scarcity of comprehensive VSL corpora, a challenge that is notably more pronounced for VSL than for other sign languages. The initial research period (2015-2020) was predominantly reliant on small-scale, self-captured datasets, often collected with devices like the Microsoft Kinect sensor or sourced from limited university databases. To mitigate these data limitations, augmentation techniques such as image rotation and noise injection were introduced. More recently (2021-2025), there has been a significant shift toward the creation of larger and more diverse datasets, including the exploration of synthetic data generation. The introduction of the Multi-VSL dataset [38], a large-scale, multi-view video corpus with 84,764 videos representing 1,000 words marks a crucial advancement toward enabling more robust model training.

Table 4. Summary of reviewed research and methods for Vietnamese Sign Language Recognition

Paper	Year	Pre-processing	Classification	Data source	Dataset size	Data level	Application /Scope	Accuracy
[1]	2015	HSV model color filter, hand extraction, feature extraction	DTW, HMM	SKIG - University of Sheffield	1080 video sequences from 6 people	Dynamic hand gestures	Dynamic hand gesture recognition	DTW (85.0%), HMM (87.9%)
[9]	2015	Gabor Filtering, Fisher's Discriminant Analysis, and Cosine Metric Distance, HMM, Gaussian Process Dynamical Model, Neural Networks	$1 Rcognizer	OpenCV, CLAPACK	420 training images (static), 42 VSL gestures (static), 7 VSL gestures (dynamic)	Static and dynamic gestures	VSL Reader	93.89% (static), 97.14% (dynamic)
[10]	2015	Rank Matrix, Alphabetic Rules	k-NN, Decision Tree, Naïve Bayes, SVM	D-VSL Database	Accent dataset: 3 gestures (613 images). Character dataset: 23 gestures (4637 images)	3D Depth Images	VSL Recognition	42.44% - 100%
[8]	2016	Key frame detection, Local Contour Sequence (LCS) feature extraction	Support Vector Machine (SVM)	Microsoft Kinect sensor	11,500 images - 23 gestures - 20 volunteers (6325 training, 4175 testing)	3D depth images	VSL Recognition	91%
[11]	2017	Otsu thresholding technique, feature extraction, HMM	One-against-one SVM	Microsoft Kinect sensor	3000 samples (1800 training, 1200 testing)	Sequences of depth image	VSL Recognition	95%
[12]	2018	Image Map Generation, Static Gesture Extraction, Dynamic Gesture Extraction	One-against-all Multi-class SVM	Microsoft Kinect v1	Static dataset: 25 classes, 23 Vietnamese letters (exclude F, J, W, Z) + 2 static diacritics Dynamic dataset: 6 classes, 6 dynamic diacritics	Depth image, RGB images, skeletal joint maps	VSL Extraction	74% - 100% (Static letters) 97% - 100% (Dynamic diacritics)
[4]	2019	Skin Detection, Data Augmentation, Spatial and scene-based features, CNN	Long Short-Term Memory (LSTM)	Camera 2D	12 words dataset, 15 words dataset	Video sequences, depth image	VSL Recognition	95.83%
[26]	2019	Histogram Equalizer, Low Pass Filter, HOG Descriptor	NIL	20bn-jester + Self-captured	1000 images containing initiation gestures (20bn-jester dataset)	2D hand gestures	Hand Gestures Recognition	> 90%

[23]	2019	InceptionV3 CNN	Long Short-Term Memory (LSTM)	UCF101 dataset, HMDB51 dataset	Three main videos (walking, baby waking up, and falling)	Video sequences	Gestures Recognition	~90%
[16]	2019	Local Binary Pattern (LBP), Local Phase Quantization (LPQ), (HOG), GIST descriptor, HOG Optical Flow	Support Vector Machine (SVM)	VSL-ADT dataset	24 alphabets, 3 diacritic marks, 5 tones)	Video sequences	Video-based VSL Recognition	86.61%
[27]	2020	Histogram balance, Low Pass Filter, HOGDescriptor	Artifical Neural Networks (ANN) - based algorithm	MSRGesture 3D, Cambridge-Gesture, 20bn-jester	2000 images each for 5 gestures, 1500 static images	Image and video sequences	Gestures Recognition	92.6% (images), > 91% (videos)
[28]	2020	MobileNet-V2, Two-stream ConvNet	Long Short-Term Memory (LSTM)	20BN-jester Dataset	1440 images (6 classes)	RGB video frames, optical flow images	Hand Gestures Recognition	91.25%
[17]	2020	vnTokenizer, Word2vec	Long Short Term Memory (LSTM)	VNTC dataset	33756 articles for training, 50373 articles for testing	Texts	Vietnamese Text Classification	93.8%
[2]	2021	Not available	1D convolutional neural networks + 1D-CNN-biLSTM	GesHome dataset	18,000 gestures with 18 labels (20 subjects)	Not available	Gestures Recognition	89% (F1 score)
[29]	2021	Recurrent Neural Network (RNN) + Mediapipe	Long Short Term Memory (LSTM)	Self-captured	42 landmarks pairs	Video sequences	VSL Detection	63.5% - 77.5%
[5]	2021	Sobel Edge Detection, Bounding box observed object using Gauss distribution, GoogLeNet	BiLSTM	Self-captured	2700 (videos), 5 (signer), 27 (VSL words), 100 videos for each word	AVI videos	VSL Recognition	99.38% (validation) 98.13% (test)
[30]	2021	TensorFlow	MobilenetV2 feature extraction + Single-Shot Detector (SSD) networks.	Self-captured + UCF101 + BU203	4400 training images, 1100 testing images (8 actions via internet, 3 action self-designed)	Single-frame and multi-frame videos	Gestures Recognition System	> 90%
[21]	2023	Normalization and Tokenization, Parsing	Rule-based Translation	VSL Dictionary	10,000 bilingual sentence pairs	Textual data	VSL Translation	89.73% (BLEU score)
[15]	2023	1D inception network, 3D	Self-learning fusion model	A camera & two	~8000 videos	Video and sensor data	SLR	NIL

		inception network, data fusion	for SLR (STSLR)	smartwatches				
[6]	2023	Resampling Data, Gunnar-Farneback	Combined RGB + optical flow using a MoviNet-based architecture	Wrist-worn RGB camera	5,408 gesture samples	Multimodal data (visual frames and motion data)	Hand Gestures Recognition	>99%
[7]	2023	Structuring Time Series Data	XGBoost + DNN	Vicon Motion Capture System (self-captured)	6000-time frames & various gestures from four categories	Motion data	Gestures Recognition	~71.43%, ~58.85%, ~50%
[31]	2024	Pose vectorization using MediaPipe, random skeletal deviations, synthetic data generation	GRU, LSTM	Generated dataset (synthetic variations of real VSL videos)	436,400 samples from 4,364 original videos	Word-level (Isolated SLR)	Data augmentation for VSLR	95.26%
[32]	2024	Text segmentation, dictionary encoding, Markov model, n-gram probability calculation	Statistical Markov model with breadth-first search	Vietnamese text corpus & VSL dataset	50M Vietnamese sentences, 3,234 sign language words	Sentence-level	Automatic sentence generation in VSL	88%
[33]	2024	Heatmap volume generation, skeleton extraction, depth normalization, augmentation	Enhanced 3D-CNN (Ad2C)	RGB-D dataset with heatmap-depth	18,948 videos (train), 2,369 (val), 2,369 (test)	Word-level (Isolated SLR)	VSL word recognition	Max 80.15% (Ad2C with Heatmap-Depth)
[34]	2025	Hand keypoints extraction	CNN-based	Mobile phone captured images & videos	875 videos (~358,000 images)	Alphabet-level	VSL alphabet recognition	Max 96.37% (bone data)
[24]	2025	Skeleton Extraction, Text Embedding	TeDG (Diffusion-guided Graph Convolutional Network)	P-SL Dataset	5,068 videos, 50 words, 12 signers	Word-level	VSL word recognition	90.5% 96.8% on P-SL dataset.
[18]	2025	Multi-view; video standardization	I3D, Swin Transformer, MViTV2, VTNPF	Multi-view video dataset	84,764 videos, 30 signers, 1,000 words	Word-level (Isolated SLR)	General-purpose SLR for VSL	Max 87.99% (VTNPF multi-view)
[25]	2025	Multi-view, Data Augmentation	M3-SLR, Ù3V	Multi-view video dataset	Multi-VSL200, MM-WLAuslan	Word-level (Isolated SLR)	General-purpose SLR for VSL	92.11%

Despite these advancements, significant gaps remain. These include the lack of standardized, large-scale VSL datasets; limited integration of multimodal cues such as facial expression and body posture; and challenges in real-time, resource-efficient deployment. Furthermore, existing systems often overlook the linguistic and regional diversity of VSL, constraining their cultural adaptability and real-world applicability.

Future Research Directions

To address these limitations, future work should focus on the following key areas.

Development of Benchmark VSL Databases: A primary obstacle hindering progress in VSL recognition is the scarcity and small scale of comprehensive, publicly available datasets. This critical gap makes it difficult to train and validate robust models, limits their generalizability, and prevents standardized, objective comparisons between different techniques. Future research must prioritize the development of large-scale, open-access, and meticulously annotated benchmark databases. To be effective, these datasets must encompass the linguistic diversity of VSL, including a broad spectrum of dialectal variations, different signing rates, and varied signer demographics. Addressing this challenge will mitigate issues of cultural adaptability and regional variation, creating a standardized foundation for developing and evaluating more scalable and effective VSL recognition systems.

Integration of Advanced Deep Neural Architectures: The field has seen a clear trend toward deep learning, with models like CNNs, LSTMs, and their hybrid architectures demonstrating significant success. However, to handle the complexities of continuous signing and subtle gesture variations, more advanced architectures are needed. State-of-the-art neural networks, such as Transformers and Vision Transformers, have shown considerable potential for robust action and gesture recognition and are particularly well-suited for modeling complex video data. Future investigations should therefore explore the application of these models to both isolated and continuous VSL recognition, with a specific focus on their capacity to capture long-range temporal dynamics and ensure spatiotemporal consistency. While these models can be computationally intensive, their ability to learn intricate patterns in sequential data is essential for pushing the boundaries of recognition accuracy.

Fusion of Facial Expressions and Advanced Feature Recognition: VSL is a multimodal language where meaning is conveyed not only through hand gestures but also through facial expressions and body movements.

A significant limitation of many current systems is their almost exclusive focus on hands, thereby neglecting the critical role of non-manual cues that carry grammatical and emotional information. To advance from word-level classification to true sentence-level comprehension, future models must integrate facial expression recognition with hand and body movement analysis. This fusion is essential for correctly interpreting grammatical nuances and the full meaning of an utterance. While this presents technical challenges in data synchronization and model complexity, it is a necessary step toward creating systems that can understand VSL with the depth and accuracy required for fluid communication.

Real-Time, Low-Resource Implementation: For VSL recognition technology to have a meaningful impact, it must be accessible and functional in real-world scenarios. This requires deploying models on low-power, readily available devices such as smartphones and augmented reality (AR) glasses. However, many of the most accurate deep learning models have high computational costs, creating a barrier to practical implementation. Therefore, future research must address this challenge by exploring lightweight model architectures, such as MobileNet and efficient Transformer variants, and by optimizing existing models for resource-constrained environments. Achieving a balance between computational efficiency and recognition accuracy is essential to bridge the gap between research prototypes and user-centric, accessible assistive technologies.

Toward End-to-End VSL Translation Systems: While most current research focuses on classifying individual signs or gestures, the ultimate goal is to facilitate seamless communication through full translation. Moving beyond gesture classification, future research should investigate the development of end-to-end systems capable of translating continuous VSL directly into text or speech. This ambitious endeavor will require leveraging advanced frameworks like encoder-decoder models, pre-trained language models, and multimodal learning pipelines that can handle complex sequence-to-sequence tasks. Such systems must also be designed to understand the unique vocabulary and grammar of VSL, which differs from spoken Vietnamese. Building these end-to-end translation systems represents a paradigm shift from simple recognition to comprehensive linguistic interpretation.

Addressing these research directions will contribute to the development of scalable, culturally adaptive, and real-time VSL recognition systems. Such advancements will not only improve communication accessibility for the deaf and hard-of-hearing community in Vietnam but also establish a foundation for inclusive assistive technologies across diverse linguistic contexts.

Acknowledgments

This research is funded by the Ministry of Education and Training (MOET) under grant number B2026.VKG.09: "Research on the application of technology to support teaching Vietnamese sign language to primary school students with students with

deaf or hard of hearing".

References

[1] D. H. Vo, H. H. Huynh, and J. Meunier, Geometry-based dynamic hand gesture recognition, Journal of Science and Technology: Issue on Information and Communications Technology, vol. 1, pp. 1–10, 2015. https://doi.org/10.31130/jst.2015.6

[2] K. Nguyen-Trong, H. N. Vu, N. N. Trung, and C. Pham, Gesture recognition using wearable sensors with Bi-Long short-term memory convolutional neural networks, IEEE Sensors Journal, vol. 21, no. 13, pp. 15065–15079, 2021. https://doi.org/10.1109/JSEN.2021.3074642

[3] L. T. Phi, H. D. Nguyen, T. Q. Bui, and T. T. Vu, A glove-based gesture recognition system for Vietnamese sign language, in 15th International Conference on Control Automation and Systems, Busan, 2015. https://doi.org/10.1109/ICCAS.2015.7364604

[4] A. H. Vo, V.-H. Pham, and B. T. Nguyen, Deep learning for Vietnamese sign language recognition in video sequence, International Journal of Machine Learning and Computing, vol. 9, no. 4, 2019. https://doi.org/10.18178/ijmlc.2019.9.4.823

[5] Q. P. Van and B. N. Thanh, Vietnamese sign language recognition using dynamic object extraction and deep learning, in IEEE Eighth International Conference on Communications and Electronics (ICCE), Phu Quoc Island, 2020.

[6] H.-Q. Nguyen, T.-H. Le, T.-K. Tran, H.-N. Tran, T.-H. Tran, T.-L. Le, H. Vu, C. Pham, T. P. Nguyen, and H. T. Nguyen, Hand gesture recognition from wrist-worn camera for human–machine interaction, IEEE Access, vol. 11, pp. 53262–53274, 2023. https://doi.org/10.1109/ACCESS.2023.3279845

[7] K. H. V. Nguyen, A.-D. Phan, T. B. Minh, T. T. T. Phan, and X. P. Do, Gesture recognition model with multi-tracking capture system for human-robot interaction, in International Conference on System Science and Engineering (ICSSE), 2023. https://doi.org/10.1109/ICSSE58758.2023.10227183

[8] D. H. Vo, H. H. Huynh, T. N. Nguyen, and J. Meunier, Automatic hand gesture segmentation for recognition of Vietnamese sign language, in ACM International Conference Proceeding Series, Association for Computing Machinery, Dec. 2016, pp. 368–373. https://doi.org/10.1145/3011077.3011135

[9] V. D. Nguyen, M. T. Chew, and S. Demidenko, Vietnamese sign language reader using intel creative Senz3D, in 2015 6th International Conference on Automation, Robotics and Applications, IEEE, Queenstown, New Zealand, Apr. 2015, pp. 77–82 https://doi.org/10.1109/ICARA.2015.7081128

[10] D.-H. Vo, T.-N. Nguyen, H.-H. Huynh, and J. Meunier, Recognizing Vietnamese sign language based on rank matrix and alphabetic rules, in International Conference on Advanced Technologies for Communications (ATC), Ho Chi Minh City, 2015.

[11] D. H. Vo, H. H. Huynh, P. M. Doan, and J. Meunier, Dynamic gesture classification for Vietnamese sign language recognition, (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 8, no. 3, 2017. https://doi.org/10.14569/IJACSA.2017.080357

[12] P. T. Hai, H. C. Thinh, B. Van Phuc, and H. H. Kha, Automatic feature extraction for Vietnamese sign language recognition using support vector machine, in Proceedings - 2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications and Computing, SIGTELCOM 2018, Institute of Electrical and Electronics Engineers Inc., Mar. 2018, pp. 146–151. https://doi.org/10.1109/SIGTELCOM.2018.8325780

[13] C. M. Jin, Z. Omar, and M. H. Jaward, A mobile application of American sign language translation via image processing algorithms, in IEEE Region 10 Symposium (TENSYMP), Bali, 2016. https://doi.org/10.1109/TENCONSpring.2016.7519386

[14] L. D. Quach and C.-N. Nguyen, Conversion of the Vietnamese grammar into sign language structure using the example-based machine translation algorithm conversion of the Vietnamese grammar into sign language structure using the example-based machine translation algorithm in International Conference on Advanced Technologies for Communications (ATC), 2018. https://doi.org/10.1109/ATC.2018.8587584

[15] H.-N. Vu, T. Hoang, C. Tran and C. Pham, Sign language recognition with self-learning fusion model, IEEE Sensors Journal, vol. 23, no. 22, pp. 27828–27840, 2023. https://doi.org/10.1109/JSEN.2023.3314728

[16] A. H. Vo, N. T. Q. Nguyen, N. T. B. Nguyen, H. V. Pham, and B. T. Nguyen, Video-based Vietnamese sign language recognition using local descriptors, intelligent information and database systems, vol. 11432, 2019.

[17] N. H. Phat and N. T. M. Anh, Vietnamese text classification algorithm using long short term memory and WORD2VEC, *Informatics and Automation*, vol. 19, no. 6, pp. 1255–1279, Dec. 2020. [In Russia]: Алгоритм классификации вьетнамского текста с использованием долгой краткосрочной памяти и Word2Vec, Искусственный интеллект, инженерия данных и знаний. https://doi.org/10.15625/1813-9663/18025

[18] Dinh, S. N., Tran, T. D., Pham, H. N., Tran, H. T., Tong, A. N., Hoang, H. Q., and Nguyen, L. P., Sign language recognition: a large-scale multi-view dataset and comprehensive evaluation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 7876–7886, 2025. https://doi.org/10.1109/WACV61041.2025.00766

[19] T. B. D. Nguyen, T. N. Phung, and T. T. Vu, A study of data augmentation and accuracy improvement in machine translation for Vietnamese sign language, Journal of Computer Science and Cybernetics, vol. 39, no. 2, 2023. https://doi.org/10.15625/1813-9663/18025

[20] T. D. Ngo, D. H. L. Nguyen, and H. L. Luong, Sign language representation using virtual characters with 3D animation, VNU Journal of Science: Computer Science and Communication Engineering, vol. 41, no. 1, pp. 54–68, 2025. https://doi.org/10.25073/2588-1086/vnucsce.3768

[21] T.-B.-D. Nguyen and T.-T. Nguyen, Rule-based machine translation for the automatic translation of Vietnamese sign language, International Journal of Language and Linguistics, vol. 11, iss. 6, Dec. 2023. https://doi.org/10.11648/j.ijll.20231106.12

[22] P. N. Huu, T. L. Ngoc, and Q. T. Minh, Proposing gesture recognition algorithm using two-stream convolutional network and LSTM, in International Conference on Communications and Electronics (ICCE), Phu Quoc Island, 2021. https://doi.org/10.1109/ICCE48956.2021.9352147

[23] P. N. Huu and H. N. T. Thu, Proposal gesture recognition algorithm combining CNN for health monitoring, in Proceedings - 2019 6th NAFOSTED Conference on Information and Computer Science, NICS 2019, 2019. https://doi.org/10.1109/NICS48868.2019.9023804

[24] Hoai, N. V., and Anh, D. T., Diffusion-guided graph convolutional networks for sign language recognition, Signal, Image and Video Processing, 2025. https://doi.org/10.1007/s11760-025-04007-9

[25] T. T. D. Nguyen, T. T. N. Do, Q. H. Hoang, P. Le Nguyen, and A. V. Tran, M^3-SLR: self-supervised pretraining with maxflow maskfeat for improved multi-view sign language representation, IEEE Access, vol. 13, pp. 148170–148191, 2025. https://doi.org/10.1109/ACCESS.2025.3601235

[26] P. N. Huu and H. L. The, Proposing recognition algorithms for hand gestures based on machine learning model, in Proceedings - 2019 19th International Symposium on Communications and Information Technologies, ISCIT 2019, Ho Chi Minh City, Vietnam, Sep. 2019. https://doi.org/10.1109/ISCIT.2019.8905194

[27] P. N. Huu, Q. T. Minh, and H. L. The, An ANN-based gesture recognition algorithm for smart-home applications, KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS, vol. 14, no. 5, pp. 1967–1983, 2020. https://doi.org/10.3837/tiis.2020.05.006

[28] D.-T. Pham, A two-Stream graph convolutional network for dynamic hand gesture recognition, in Advances in Data Science and Optimization of Complex Systems, H. M. and N. Q. T. Le Thi Hoai An and Le, Ed., Cham: Springer Nature Switzerland, 2025, pp. 288–297. https://doi.org/10.1007/978-3-032-00267-9_26

[29] K. D. Bach, P. T. Duong, P. T. T. Ha, B. N. Anh, and N. T. son, Vietnamese sign language detection using mediapipe, in Proceedings of the 2021 10th International Conference on Software and Computer Applications (ICSCA), Kuala Lumpur, 2021.

[30] P. N. Huu, H. N. T. Thu, and Q. T. Minh, Proposing a recognition system of gestures using mobilenetV2 combining single shot detector network for smart-home applications, Journal of Electrical and Computer Engineering, vol. 2021, 2021. https://doi.org/10.1155/2021/6610461

[31] Dang Khanh, Bessmertny I. A. ViSL one-shot: generating Vietnamese sign language data set. Scientific and Technical Journal of Information Technologies, Mechanics and Optics, vol. 24, no. 2, pp. 241–248, 2024. https://doi.org/10.17586/2226-1494-2024-24-2-241-248

[32] Dang Kh., Bessmertny I. A. ViSL model: the model automatically generates sentences of Vietnamese sign language, Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2024, vol. 24, no. 5, pp. 779–787. https://doi.org/10.17586/2226-1494-2024-24-5-779-787

[33] Xuan-Phuoc Nguyen, Thi-Huong Nguyen, Duc-Tan Tran, Tien-Son Bui, and Van-Toi Nguyen, An isolated Vietnamese sign language recognition method using a fusion of heatmap and depth information based on convolutional neural networks, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2024. https://doi.org/10.1109/APSIPAASC63619.2025.10848961

[34] V. Tran, V. K. Phung, Q. H. Hoang, and T. V. H. Pham, Vietnamese Sign Language Alphabet Recognition Using Deep Learning and Mediapipe Methods, Smart Systems and Devices, vol. 35, no. 1, pp. 10–19, Jan. 2025. https://doi.org/10.51316/jst.179.ssad.2025.35.1.2