International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-15 Issue-8, July 2026

Vietnamese Sign Language Recognition for People with Hearing Impairment using Digital Library

Tran Thi Thuy Ha, Phan Thi Ha

Abstract: ***This paper proposes a method for building a specialised sign language dataset for deaf-mute sign language recognition in digital libraries, comprising 120 signs and 65 phrases, collected from 25 people with hearing impairments in Hanoi and Ho Chi Minh City. The data were collected under various conditions and yielded approximately 125,000 video samples after processing. Based on experimental results with Random Forest [3], pure LSTM [1], and CNN-LSTM models [4], the CNN-LSTM model yields the best results and is selected for the development of Vietnamese sign language recognition services, demonstrating high performance, stability, and suitability for practical implementation requirements.***

Keyword: Random Forest, Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), CNN-LSTM, Hearing Impairment, Identification, Sign Language, Library, Phrase.

Nomenclature:
LSTM: Long Short-Term Memory
CNN: Convolutional Neural Network

I. INTRODUCTION

Sign language is a natural language system that developed within the deaf community, using hand movement, facial expressions, and body posture to communicate information. Contrary to the popular belief that sign language is an illustration of spoken language, it is an independent language with its own grammatical structure and vocabulary [2]. Sign language recognition is one of the studies of computer vision and artificial intelligence, which is the process of using computer vision techniques as well as machine learning to automatically analyse, process, and convert hand movements, gestures, and expressions in sign language into a form of information that can be understood and processed by a computer, such as text, control commands, or query signals. In the world, this problem has developed diversely, from traditional feature extraction methods to modern deep learning models such as convolutional networks, recurrent networks, and Transformer [6].

In Vietnam, there have been initiatives and studies on the recognition of Vietnamese Sign Language, focusing mainly on building datasets and recognising individual symbols under controlled lighting and setting conditions, with relatively promising results [5][7]. However, many limitations remain unresolved. Most current research focuses only on basic vocabulary, failing to address the needs of recognising phrases used in digital libraries for information retrieval. Furthermore, model evaluation is primarily conducted in laboratory settings, which can reduce effectiveness when applied to real-world conditions with various interfering factors, such as unstable lighting, complex backgrounds, or changing camera angles. To date, few studies have implemented a Vietnamese sign language recognition system integrated into existing information systems, such as digital libraries, public service portals, or online education platforms.

The development of a Vietnamese sign language recognition system not only helps the hearing-impaired people access information more naturally and conveniently but also plays a crucial role in promoting social inclusion, improving quality of life, and creating equal opportunities for learning and development for this community. This has significant practical implications not only for digital library systems but can also be extended to other public services such as healthcare, education, and public administration. In Vietnam, digital library systems primarily focus on digitising documents, managing data repositories, and optimising text search tools. Integrating Vietnamese sign language recognition technology into search systems is being set as an important goal in the development roadmap; however, its implementation still faces many challenges. Firstly, there is no specialised dataset for Vietnamese Sign Language in the library and information retrieval sector. Secondly, existing research is mostly limited to laboratory testing and has not been integrated into real-world systems. Thirdly, the lack of participation by the hearing-impaired community in the design and development process results in solutions that do not meet practical needs. Therefore, a comprehensive, integrated solution is needed that encompasses data collection, recognition model development, system deployment, and user evaluation.

In this paper, we study the application of machine learning models to Vietnamese Sign Language recognition for the hearing-impaired who use digital libraries, helping the community access knowledge more easily.

Unlike other studies that stop at laboratory experiments, this research aims to build a specialised dataset that

Manuscript received on 20 June 2026 | First Revised Manuscript received on 27 June 2026 | Second Manuscript Accepted on 08 July 2026 | Manuscript Accepted on 15 July 2026 | Manuscript published on 30 July 2026.

*Correspondence Author(s)
Dr. Tran Thi Thuy Ha, Lecturer, Faculty of Information Technology, Posts and Telecommunications Institute of Technology (PTIT) in Vietnam, Ha Dong, (Ha Noi), Vietnam. Email ID: hatt@ptit.edu.vn, thuyhadt@gmail.com, ORCID ID: 0009-0005-6734-1509
Phan Thi Ha*, Lecturer, Faculty of Information Technology, Posts and Telecommunications Institute of Technology (PTIT) in Vietnam, Ha Dong, (Ha Noi), Vietnam. Email ID: hapt@ptit.edu.vn, hapt37@fe.edu.vn, phantha.72@gmail.com, ORCID ID: 0009-0003-2521-0717

closely reflects the real-world context of digital library use, directly serving the information retrieval. needs of the hearing-impaired. Therefore, the process of defining the scope, collecting, processing, and organizing the data is carried out systematically, rigorously, and under control.

II. METHODOLOGY

A. Scope and Content of Dataset

Unlike general sign language recognition studies, this study focuses on a specific context of use: assisting the hearing-impaired in searching for and accessing resources in digital libraries. Therefore, the dataset does not aim to cover the entire Vietnamese sign language, but is selectively built, focusing on symbols and phrases with high use frequency in library search activities. This approach reduces the complexity of the problem while enhancing the system's practical applicability.

- List of symbols and phrases for search: The dataset is designed to include two main content groups:
  - **120 Single Symbols:** Represent basic keywords that frequently appear in digital library search queries.
  - **65 Phrases:** Reflecting symbol combinations commonly used by users when expressing their complete information search needs.
- The symbols and phrases belong to the categories below:
  - **Document:** Book; Textbook; Thesis; Dissertation; Article; Digital Document
  - **Document Attribute:** Author; Publication Year; Publisher; Topic; Field of Study; Language
  - **Search Action:** Search; Look Up; View; Download; Filter; Sort
  - **Query Phrase:** "Information Technology Textbook"; "Master's Thesis"; "Books from 2020"; "Vietnamese Document"; "Author Nguyen Van A"

Building phrases rather than single symbols helps the model more closely match real-world search behaviour, where users often express intent using a series of consecutive symbols.

B. Context of Digital Library

A key differentiating factor of this dataset is its contextuality. Each symbol and phrase is selected based on its direct relationship to the functionality and data of the digital library system. Specifically:

- Symbols are not interpreted independently but are linked to the search action.
- The recognition results do not stop at "identifying symbol" but are directly mapped to the query keyword in the library system.

This allows the system to seamlessly transition from sign language to search operation, minimizing intermediate steps for hearing-impaired users.

C. Data Collection

The process for collecting Vietnamese sign language data is systematically structured to ensure the dataset accurately reflects the characteristics of sign language in real-world use while also meeting the requirements for training and evaluating machine learning models. Because sign language data is highly visual and heavily dependent on participants and recording conditions, the collection process is tightly organised and controlled, adhering to consistent principles.

First, the team clearly identified the target participants for data collection as hearing-impaired individuals proficient in Vietnamese Sign Language for daily communication. These participants included both men and women of various ages, reflecting the diversity in sign language expression. Participant selection was based not only on the ability to perform sign language correctly but also on diversity in expressive styles, execution speed, and hand shapes, helping the model learn the natural variations found in real-world data. The data collection process involved 25 hearing-impaired individuals, each of whom repeatedly performed the identified symbols and phrases in the dataset. The data were collected repeatedly to increase the sample size and help the machine learning model recognise subtle variations in symbol expression within the same individual and between individuals. This is a crucial factor in improving the model's generalizability in real-world operating environments.

Regarding data collection conditions, data were collected across various settings to ensure the model performed well in a range of environments. Specifically, videos were filmed under different lighting conditions, including natural, classroom, and artificial indoor lighting. The backgrounds ranged from monochrome to detailed to simulate real-world usage scenarios when users access digital libraries in different spaces. Camera angle was primarily set to frontal, chest-level, or face-level, ensuring full capture of hand and finger movements and some facial expressions – crucial elements in sign language.

Regarding the equipment and tools used, data were primarily collected using the built-in cameras on laptops and smartphones at full HD resolution, reflecting the devices that normal users might use when accessing the system. The absence of specialized recording equipment ensures the feasibility and widespread deployment of the system in the future. Videos are recorded at a stable frame rate and in a common format, making them convenient for processing and feature extraction in subsequent steps.

The data collection process followed a standardised scenario. Participants were provided with a list of symbols and phrases to perform, and were guided to use the symbols naturally, without being constrained by rigidity. This approach makes the collected data better reflect real-life sign language usage, especially in the context of information retrieval in digital libraries.

The result of the collection process is a rich video dataset that varies in environmental conditions and symbol expressions. This is an important foundation for pre-processing, labelling, and training machine learning models in subsequent steps, while also ensuring practicality.

Table I: Data Used to Build Sample Dataset

Requirements	Number
Number of participants	10
Number of symbols per participant	30
Number of samples per symbol	20
Total samples per person	600
Total collected samples	6,000

D. Data Pre-Processing

i. Normalize Symbol Sequence Length.

Each symbol or phrase is represented by a sequence of hand movements over time; therefore, the number of frames in each sample can vary with each individual's execution speed. To ensure uniformity in the model's input data, the project normalised the sequence length to a fixed number of frames.

Specifically, the sequences that are longer than the defined threshold are sampled to reduce redundance frames, while shorter sequences are supplemented with frames using iterative or interpolation methods. This normalisation helps the machine learning model process data in a fixed-size time-series format while limiting discrepancies caused by varying symbol execution speeds among participants.

ii. Remove Noise and Invalid Data.

During data collection, some samples may be affected by factors such as frame loss, hand movement out of the field of view, or recognition errors in the extraction tool. These samples are checked and removed to avoid negative impacts on the training process.

In addition, anomalies in the coordinate sequence, such as illogical abrupt changes between consecutive frames, are also smoothed out by simple filtering techniques. Removing noise helps the data more accurately reflect the actual hand movement during the gesture.

iii. Normalize Hand Coordination

The model's input data is represented as 21-point hand coordinates, each point comprising 3 coordinates (x, y, z). However, absolute coordinates are highly dependent on the camera's position and the distance between the participant and the recording device. Therefore, we normalized the coordinates to reduce the influence of external factors.

The coordinates are normalized relative to the frame size or using a fixed point on the hand (e.g., wrist) as a reference point. This process helps the model focus on learning the relative hand shape and movement rather than relying on absolute spatial position.

iv. Balance Data Between Classes

Because the frequency of the symbols and phrases is not entirely uniform, some classes have significantly more samples than others. To avoid model bias during training, we applied data balancing measures. Classes with fewer samples are enhanced by exploiting additional valid sequences or applying small transformations in the coordinate sequences. In contrast, classes with excessively large numbers of samples have fewer samples available for training. As a result, the pre-processed dataset achieves a relative balance between classes, helping the model improve its recognition capabilities uniformly across the entire symbol set.

E. Data Label and Organization

After completing the preprocessing steps, Vietnamese Sign Language data needs to be labelled and organised for training, evaluation, and deployment of machine learning models. This is a crucial step to ensure that the input data accurately reflects the linguistic meaning of each sign, while also facilitating future expansion and reuse of the dataset.

i. Data Labelling Method

Labelling is based on the semantic content of the signs and phrases that have been pre-selected in the list for digital library search. Each data sample corresponds to a single sign or a complete phrase and is assigned a unique label that represents the keyword or meaning to be searched within the library system.

The labelling process is performed semi-manually, combining:

- A pre-built list of standard labels (120 symbols and 65 phrases)
- Visually inspecting the original videos or motion sequences to ensure the symbol correctly represents the content
- Cross-check between samples to minimize errors due to confusion between symbols with similar visual appearance.
- For phrases, labels are assigned based on the complete semantic unit rather than to individual symbols, aligning with the objective of document searching in digital libraries. This approach allows the model to learn the temporal relationships between consecutive signs within a phrase.

ii. Data Organization and Storage

The labelled data is organized into a well-structured folder hierarchy, facilitate data management and model training. Each label is stored in a separate folder, containing files corresponding to samples collected from different participants.

Each data sample is stored as a file containing a sequence of hand coordinates over time in a consistent format (e.g., JSON or NumPy array). This organization allows for quick retrieval of each sample based on label and also supports batch processing during training.

Additionally, metadata such as participant ID, collection conditions, frame count, and symbol execution time are also stored for further analysis when needed.

iii. Splitting the Dataset into the Training, Testing, and Evaluation Datasets

To ensure objectivity in the model evaluation process, the dataset is split into three separate datasets:

- **Training Set (70%):** largest portion of the dataset, used to train the machine learning model
- **Validation Set (15%):** used to adjust hyperparameters and monitor overfitting during training
- **Test Set (15%):** used for final evaluation of the model's performance on previously unseen data

The dataset was partitioned so that every label was represented in all three subsets. Additionally, samples from the same participant were prevented from appearing in both the training and evaluation set to improve the generality of the results.

The resulting dataset is complete, with a clear structure and correct labels. It is ready for the training and optimization phase of the machine learning models in the next part of the project.

F. Model Selection and Evaluation

Characteristic of input data: the data used in the project is time-series data, extracted from video via the MediaPipe library. Each sample is a sequence of consecutive frames, where each frame is represented by

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

a 21 x 3 matrix, corresponding to 21 points on the hand and 3 spatial coordinates (x, y, z). Thus, the entire data sample can be considered a three-dimensional tensor with shape (number of frames x number of marker points x number of coordinates).

This characteristic indicates that the data contains both spatial information (hand configuration at each time point) and temporal information (changes in hand movement across the sequence of frames). Therefore, the selected model should be capable of effectively capturing both aspects.

Besides, data were collected from many different participants under inconsistent lighting conditions, camera angles, and symbol execution speeds. This requires a model with good generalisation capability to avoid overreliance on the characteristics of each individual.

- Comparison of feasible methods: Based on data characteristics and project objectives, three main methods were considered and tested:
- Traditional machine learning, represented by the Random Forest model, uses features manually extracted from landmark data
- Deep learning combined with a convolutional neural network (CNN) and LSTM, where the CNN handles learning spatial features, while LSTM learns the temporal relationships between frames
- Pure LSTM model on landmark data, aiming to evaluate the effectiveness of using only temporal elements with a simpler architecture

Implementing and comparing multiple methods helps comprehensively evaluate each model's performance while highlighting the strengths and limitations of each approach in the context of Vietnamese sign language recognition.

The selection of the aforementioned methods aims not only to achieve the highest accuracy, but also to serve the project's objectives, including:

- Evaluate the applicability of traditional machine learning models in the symbol recognition problem
- Utilize deep learning to learn feature representations from time-series data automatically
- Compare complex and simple architectures to find a balanced solution between accuracy, processing time, and practical implementation

Based on the objectives, the research team conducted training, optimization, and detailed evaluation for each model, forming the basis for selecting the final model to integrate into the digital library system. Table II shows that the CNN-LSTM model achieves the highest accuracy, significantly surpassing Random Forest and pure LSTM. This confirms the effectiveness of combining spatial feature extraction (CNN) and temporal dependency learning (LSTM) for hand sign recognition.

Table II: Comparing Accuracy Between Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Random Forest	78.6	77.9	76.8	77.3
Pure LSTM	86.4	85.9	85.1	85.5
CNN-LSTM	92.3	92.0	91.7	91.8

Based on experimental results and the requirements of the digital library system, the CNN-LSTM model was selected as the official recognition model for the next phase of recognition service development research. This model not only achieved the highest accuracy (92.3%), but also demonstrated the ability to handle dynamic symbols, long phrases, and data collected in real-world scenarios.

The selection of CNN-LSTM strikes a balance among accuracy, stability, and integration capabilities, fully meeting the objective of building a Vietnamese sign language recognition system that can operate effectively within an existing digital library environment.

G. Deploy Sign Language Recognition Services

The sign language recognition service is built with FastAPI in Python and serves as the backend component responsible for all tasks related to sign recognition. The service operates independently of the digital library system and provides key APIs such as:

- Receive points-on-hand data over time.
- Pre-process input data.
- Perform inference using the trained CNN-LSTM model.
- Return recognition result with confidence level.

A machine learning model is preloaded into memory when the service starts, reducing inference response time. The video is not stored on the server; it is processed only in temporary memory and deleted immediately after recognition is complete, ensuring user data security.

System response time assessment: System response time was measured during real-world testing when users performed searches using sign language. Results were recorded across multiple test sessions.

Table III: Average Response Time of the System

Processing Stage	Average Time (ms)
Receive and pre-process video	180
Recognize symbol	220
Query and return results	140
Total	~540

With the average response time of under 1 second, the system meets real-time interaction requirements, ensuring a seamless user experience, especially for the hearing impaired. This sign language recognition service can be integrated into an existing digital library system, providing an additional accessibility tool that enables people with hearing impairment to access and utilise digital resources effectively.

III. CONCLUSION

This paper presents a methodology for constructing a specialised sign language dataset for the deaf-mute sign language recognition problem applied in digital libraries. The dataset includes 120 symbols and 65 phrases, collected from 25 hearing-impaired people in Hanoi and Ho Chi Minh City. Data were collected under various conditions and, after all processing, yielded approximately 125,000 video samples. Experimental results show that the CNN-LSTM model was chosen for the recognition service due to its high performance (92.3% accuracy, average response time: 1 second), stability, and suitability for practical implementation.

The proposed solution enables integration of the service across various digital library platforms, expanding access

to information and enhancing the user experience for hearing-impaired individuals.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted objectively and without external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally by all participating individuals.

REFERENCES

1. Al-Selwi, S. M., Hassan, M. F., Abdulkadir, S. J., et al. (2024). RNN-LSTM: From applications to modelling techniques and beyond—a systematic review. Journal of King Saud University – Computer and Information Sciences, 36(5), 102068. DOI: https://doi.org/10.1016/j.jksuci.2024.102068
2. Ministry of Education and Training (MOET), Vietnam. (2024). Circular No. 27/2024/TT-BGDĐT promulgating the Regulations on the Organization and Operation of Schools and Classes for Persons with Disabilities.t. URL: https://vbpl.vn/TW/Pages/vbpq-toanvan.aspx?ItemID=174444
3. Chen, Y., Cheung, K. C., Sun, R. Z., & Yam, S. C. P. (2024). A user guide of CART and random forests with applications in FinTech and InsurTech. Japanese Journal of Statistics and Data Science, 7(2), 999–1038. DOI: https://doi.org/10.1007/s42081-024-00258-x
4. Luqman, H., & ElAlfy, E. (2022). Utilising motion and spatial features for sign language gesture recognition with cascaded CNN-LSTM models. Turkish Journal of Electrical Engineering and Computer Sciences, 30(7), 2508–2525. DOI: https://doi.org/10.55730/1300-0632.3952
5. Nguyen Quang Duy, & Luong Thai Le. (2025). Recognizing Vietnamese Sign Language Using Deep Neural Networks. TNU Journal of Science and Technology, 230(7), 208–215. DOI: https://doi.org/10.34238/tnu-jst.12708
6. Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign Language Recognition: A Deep Survey. Expert Systems with Applications, 164, 113794. DOI: https://doi.org/10.1016/j.eswa.2020.113794
7. Tran, A. V., Phung, V. K., Hoang, Q. H., & Pham, T. V. H. (2025). Vietnamese Sign Language Alphabet Recognition Using Deep Learning and Mediapipe Methods. Smart Systems and Devices, 35(1), 10–19. DOI: https://doi.org/10.51316/jst.179.ssad.2025.35.1.2

AUTHOR'S PROFILE

Dr. Tran Thi Thuy Ha is currently a lecturer and researcher at the Faculty of Electronics Engineering at the Posts and Telecommunications Institute of Technology (PTIT) in Vietnam. She received a B.Sc. in Radio-Electronics Engineering in 1998 (Hanoi University of Science, Vietnam National University), an M.Sc. in Radio-Electronics Engineering in 2002 (Faculty of Technology, Vietnam National University) and a PhD. in Electronic Engineering in 2021. Her research interests include Electronics Engineering and Micro Electro Mechanical Systems (MEMS).

Dr. Phan Thi Ha is currently a lecturer at the Faculty of Information Technology at Posts and Telecommunications Institute of Technology (PTIT) in Vietnam, and at the Computing Fundamental Department, FPT University, Hanoi 10000, Vietnam. She received a B.Sc. in Math & Informatics, an M.Sc. in Mathematical Guarantee for Computer Systems, and a PhD. in Information Systems in 1994, 2000, and 2013, respectively. Her research interests include machine learning, natural language processing and mathematical applications.