Bridging communication with machine learning in sign language recognition for Vietnamese

Hoa Tat Thang[1], Pham Van Quoc[2], Doan Van Hoa[3*]

[1]Le Quy Don Technical University, No. 236 Hoang Quoc Viet, Bac Tu Liem, Hanoi, Vietnam ;
[2]VNU University of Science, No. 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam;
[3]Institute of Information Technology, Academy of Military Science and Technology, No. 17 Hoang Sam, Cau Giay, Hanoi, Vietnam.
[*]Corresponding author: doanvanhoa@gmail.com
Received 6 Jul. 2024; Revised 15 Sep. 2024; Accepted 11 Oct. 2024; Published 25 Oct. 2024.
DOI: https://doi.org/10.54939/1859-1043.j.mst.98.2024.139-145

ABSTRACT

Vietnamese Sign Language (VSL) serves as the primary language for deaf and hard-of-hearing individuals in Vietnam. This paper explores the sign language recognition process for VSL, emphasizing the role of machine learning in bridging communication barriers. We delve into the basics of VSL, detailing the one-to-one correspondence between hand signs and Vietnamese alphabet letters and address the formation of words through sequential hand signals and diacritics placement. Furthermore, the paper highlights the importance of pausing between words and the utilization of machine learning algorithms for automated sign recognition. Lastly, we conclude by discussing the potential applications and future directions of VSL recognition technology in Vietnam.

Keywords: Vietnamese Sign Language; Sign language recognition; Deep learning model; KNN.

1. INTRODUCTION

Sign language stands as a paramount medium of communication for individuals with hearing impairments, bridging the gap in verbal interaction and fostering inclusivity in society. Vietnam, like many other nations, has developed its unique sign language system, intricately designed to represent the Vietnamese alphabet through a series of distinct hand gestures. This paper delves into the comprehensive exploration of the sign language recognition process tailored specifically for the deaf community in Vietnam.

Vietnam boasts a rich and complex cultural heritage, where sign language serves as an integral part of communication for the deaf. Unique hand signs correspond to each letter of the Vietnamese alphabet, forming the basis of this visual and expressive language. These signs are not arbitrary; rather, they have evolved organically over time, embodying the linguistic essence of the Vietnamese language while accommodating the visual and spatial nature of sign communication.

The recognition of sign language among the deaf in Vietnam operates on a meticulous and systematic process. Each hand gesture symbolizes a specific letter of the alphabet. To construct words, individuals express the preceding letters sequentially, followed by the application of diacritics to form accurately pronounced words. Crucially, pauses ranging from 0.5 to 1 second delineate the spaces between words, aiding in coherent communication and comprehension.

The integration of machine learning techniques has emerged as a transformative approach in deciphering and interpreting these intricate hand gestures. By leveraging advanced algorithms and data-driven models, the technology facilitates the recognition and translation of these gestures into their corresponding letters, enhancing accuracy, speed, and enabling effective communication for the deaf community.

This paper aims to elucidate the nuances of the sign language recognition process utilized in Vietnam. It examines the intricate relationship between hand gestures and linguistic elements, highlighting the systematic construction of words and spaces within the sign language framework.

Furthermore, the paper explores the pivotal role of machine learning techniques in enhancing the accuracy and efficiency of recognizing and interpreting these gestures, ultimately contributing to fostering a more inclusive and accessible environment for individuals with hearing impairments.

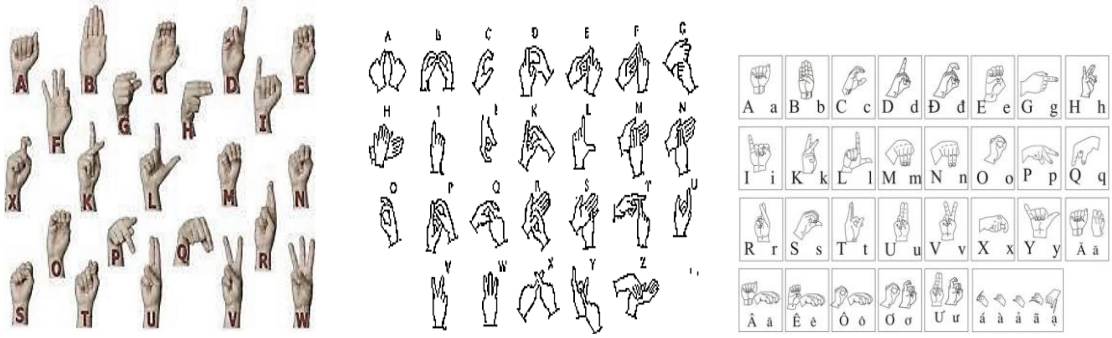
2. RELATED WORKS

Every language encompasses numerous letters, a myriad of words, and an endless array of sentence structures employed by individuals to engage in communication. Likewise, individuals who are deaf convey their emotions and thoughts through the arrangement of hand signs and gestures [1].

In the context of this article, the author aims to discuss the method of static hand recognition. With this method, the hand's position does not affect the meaning of the gesture, and the gestures are not dependent on time [2]. In American Sign Language (ASL), static one-hand gestures correspond to the English alphabet, as illustrated in figure 1a. Therefore, these symbols are highly significant as users employ them for spelling, such as their names, the names of places they wish to visit, the names of foods they want to eat, or to label anything that lacks a specific symbol [3].

In Indian sign language (ISL), the letters of the alphabet are represented by both 2-static and 1-static hands [1]**,** as shown in figure 1b.

In Vietnamese Sign Language (VSL), static gestures of one hand correspond to the English alphabet, with the additional incorporation of certain Vietnamese letters (Ă, Â, Ê, Ô, Ơ, Ư), as illustrated in figure 1c. Additionally, Vietnamese Sign Language also utilizes diacritics.



(a) *(b)* *(c)*

Figure 1. *Hand gesture alphabet (a) [ASL] American, (b) [ISL] Indian, (c) [VSL] Vietnam.*

In [5], the authors employed an alternative approach using MediaPipe, a machine-learning platform developed by Google. Their method involved applying this technique to the American Sign Language (ASL) alphabet, extracting 3D coordinates for 21 joints from RGB images captured by a webcam to derive distance and angle features. Subsequently, two classification methods, namely the light gradient boosting machine (GBM) and Support Vector Machine (SVM), were implemented. Various multi-type data inputs were utilized, including the Massey dataset, finger spelling A, and ASL alphabets with corresponding images. Remarkably, the first two data types yielded outstanding accuracy of 97.80% and 99.39% with GBM and SVM classifiers, respectively, marking the highest achieved results. However, the system's performance experienced degradation with the third data type, resulting in an accuracy drop to 86.12% and 87.60% with the respective classifiers.

In [6], a similar approach was applied, utilizing the same technique but with a single type of intricate background, and the prediction was carried out using the K-Nearest Neighbors (KNN) algorithm. This configuration exhibited an enhanced accuracy, elevating it from 86% to 91% when compared to the prior study. In the context of a smart home application, the authors in [4] leveraged MediaPipe along with Convolutional Neural Network (CNN) techniques for the prediction phase, achieving an impressive accuracy of 99%. The machine learning model, constructed in Tensor

Flow, comprised three 2D convolutional layers with rectified linear unit activation functions. Video input, captured by the integrated camera and processed frame by frame using OpenCV, resulted in individual frames stored as two-dimensional arrays (480 x 60 pixels x 3 RGB values) during runtime. Subsequently, these video frames were transmitted to a server for further processing.

Some studies on hand gesture recognition have been conducted in Vietnam:

[7, 8] proposed a method based on basic image processing. The data captured by the camera is a hand image. The pre-processing step uses a color filter to eliminate noise. The study applies geometric methods to determine the fingertips and remove the arm. After extracting features represented by vectors, the study uses a multi-layer support vector machine (SVM) learning model for training and recognition. This method has the disadvantage of slow processing speed and low accuracy.

Another study used a depth camera to collect data and extracted features based on the rank order correlation matrix (ROCM). The study was tested on datasets of Vietnamese sign language symbols with single and double symbols. This study requires specialized hardware (depth camera), which is not suitable for popular applications in Vietnam.

In [9], the authors address the problem of Vietnamese Sign Language (VSL) recognition for real-world applications. The study proposes using two types of features: spatial features and scene-based features, to capture important information about language signs. Besides the traditional classification method like SVM commonly used in sign language recognition, this study also uses advanced deep learning techniques to recognize VSL, aiming to find the dependence of each frame in the video sequence. However, this study only applied two VSL datasets of the relative family topic, such as father, mother, uncle, aunt, etc. (VSL-WRF), which were collected. The first dataset includes 12 words in Vietnamese sign language, which only have small changes between frames. While the second dataset was acquired with the corpus of continuous VSL datasets, and consists of 15 words that have the relative position and orientation of motion gestures and the body parts. The study did not apply to Vietnamese characters.

Overall, hand gesture recognition research in Vietnam is still in its early stages. More research is needed to develop more robust and accurate methods that can be applied to a wider range of applications.

3. PROPOSED METHODS

To recognize Vietnamese sign language, features of the hand are used to distinguish between signs. The rule classifier is an effective and efficient algorithm that is used to detect hand signs.

Figure 2. *Block diagram of hand sign recognition.*

Figure 2 implements the block diagram of the hand sign recognition system. This system is used to identify the alphabets and the characters that are provided using the sign. The basic steps that is involved in the conversion of image to identification is shown in the above block diagram.

3.1. Input image

Images are sourced from either the laptop camera or the provided webcam, forming the input for image processing. These images, submitted by the user, encompass various representations of the Vietnamese Sign Language alphabet. Captured with a notebook webcam, these visuals serve as the raw material for subsequent image processing stages.

To effectively collect data, hand images were automatically gathered using the Mediapipe tool, combined with manual selection. The hand is placed in front of a computer camera, and the computer continuously captures images of the hand. Each gesture involves collecting around 400

hand images under different lighting conditions and backgrounds, ensuring the images can be effectively trained and recognized later on. During the data collection process, some images may not correctly represent the intended gestures, so the next step involves manual selection. Inappropriate gestures are discarded, leaving about 300 images that accurately represent the correct gestures (such gestures are usually very few). The result of this step is directories containing images of gestures representing the characters.

3.2. Hand identification using MediaPipe

MediaPipe is a library similar to familiar frameworks that enable developers to build cross-platform, multimodal machine learning pipelines for video, audio, and time series data. The MediaPipe library boasts a large collection of pre-trained models for body and hand detection and tracking, developed using Google's extensive datasets. These models are built by Google developers using TensorFlow Lite to ensure smooth information flow within easily adaptable and modifiable graphs.

The hand detection model is a state-of-the-art Google model that accurately identifies and crops the palm image before forwarding it to the next model. This process also minimizes the need for data augmentation techniques like rotations, flipping, and scaling.

Figure 3 is a reference to the images that are included in the research paper. The images show a person's hands with the palm facing the camera. The hand is being detected by MediaPipe, which is a library that enables developers to build cross-platform, multimodal machine-learning pipelines.

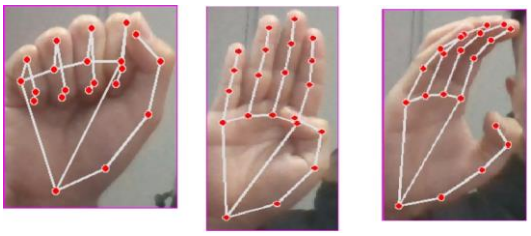


Figure 3. *Hand identification using MediaPipe.*

3.3. Image pre-processing

Images of hands recognized with the help of the MediaPipe library can vary greatly in size depending on the gesture to be represented and the distance of the hand from the camera figure 3. However, in order to train and recognize hand images, the images must be of the same size. In our research, we resize the images to 300 x 300 pixels. The hand image is resized to a maximum of 300 pixels in length or width and placed in the center of the image. The missing width or length on both sides is filled with white pixels.

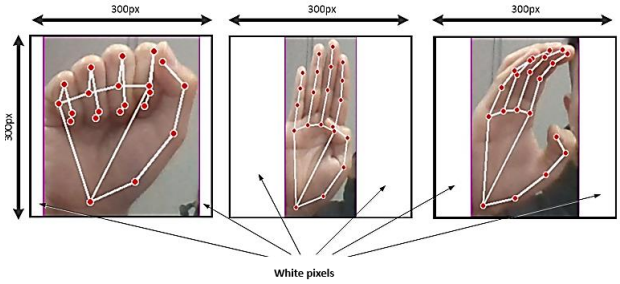


Figure 4. *Hand images pre-processing.*

Figure 4 is a reference to the images that are included in the research paper. The images show a number of hand images that have been preprocessed. The preprocessing steps include resizing the images to 300 x 300 pixels and filling the missing width or length on both sides with white pixels.

3.4. VLS database

Vietnamese Sign Language (VSL) uses 29 distinct signs to represent the Vietnamese alphabet. It is one of the most common communication methods used by deaf people in Vietnam. Of these, 22 letters A, B, C, D, E, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y use the same signing method as in English. The letter Đ uses its own unique sign, while the 6 letters Ă, Â, Ê, Ô, Ơ, Ư are represented by two hand signs, as shown in figure 1c.

3.5. Classifier

Keras is an open-source deep-learning framework written in Python. It can run on top of other deep learning frameworks such as TensorFlow, Theano, and CNTK. In this research, we use it with TensorFlow to build a deep learning model for hand sign classification. For training Vietnamese hand signs, we collect images of 29 signs**,** as shown in figure 1c. Specifically, the letters Ă, Â, Ê, Ô, Ơ, Ư are each represented by 2 signs, resulting 29 base signs + 3 signs  ,   (for hats, hook). In total**,** we have 32 image folders for training.

3.6. Identified alphabet

During classification, the letters A, B, C, D, Đ, E, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y are classified using 1 sign, while the letters Ă, Â, Ê, Ô, Ơ, Ư are classified using 2 signs. For example, the letter Â will first be detected sign  as the letter A, but if the next sign is the "hat sign  ", it will be converted to the letter Â.

Vietnamese deaf people use gestures to represent utilizes diacritics ("\", "/", "?", "~", "."). This poses a challenge when combined with signs. In this study, we only use machine learning to classify Vietnamese letters without classifying utilizing diacritics.

3.7. Vietnamese word formation

Vietnamese sentences are constructed from words. Vietnamese words are constructed from characters. The characters are detected by the machine learning model introduced above. Since the images received by the computer are continuous, if consecutive images have the same character, they are ignored, and only the first character image is kept to avoid repeated characters. Special Vietnamese characters (Ă, Â, Ê, Ô, Ơ, Ư) will be detected through two hand signs.

If no hand is detected in an image within 0.5 to 1 second, a space will be added after the word. These words and spaces will create Vietnamese sentences.

4. EXPERIMENTS AND RESULTS

The experiment was conducted on a regular laptop with a Core i7 CPU, 8GB RAM, a Windows 11 operating system, and a full HD webcam. The Python programming environment was used in combination with the MediaPipe library for hand detection.

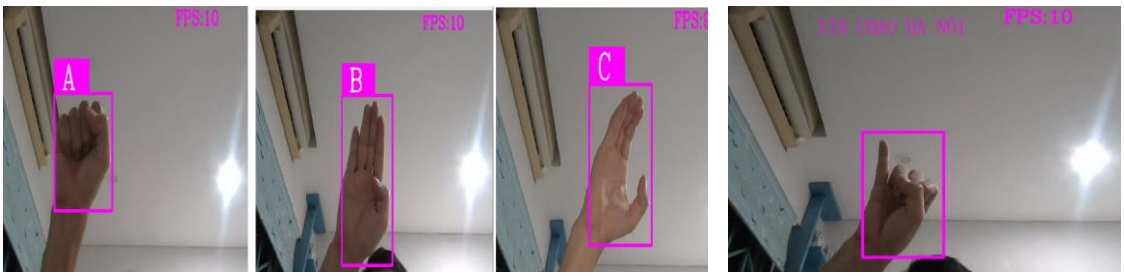
In our experiment, we tested and accurately evaluated the recognition for the first 10 characters in the alphabet and 6 unique characters of the Vietnamese language. Each character was tested with gestures using 500 consecutive images. The results are shown in table 1 below.

The accuracy can be explained as follows: The accuracy for single-character representations is higher. Hand gestures are more stable, leading to more precise recognition. For characters requiring two-character representations, the need to switch between gestures results in less stable representations, leading to lower recognition accuracy. The author himself has not been able to demonstrate precise and quick transitions between the gestures.

Table 1. *Results of Vietnamese sign language recognition.*

Letters	Accuracy
A	95,20%
B	94,40%
C	96,00%
D	94,60%
Đ	95,40%
E	95,40%
G	94,60%
H	95,80%
I	95,40%
K	95,00%
Ă	93,20%
Â	92,00%
Ê	93,40%
Ô	92,20%
Ơ	92,00%
Ư	93,40%

The recognition of individual characters (A, B, C) is shown in figure 5a and the entire Vietnamese sentence "XIN CHAO HA NOI" as shown in figure 5b. It is important to note that we have not yet been able to process Vietnamese utilizing diacritics because this requires gesture recognition, not just sign recognition like characters. The accuracy of the method depends on the person performing the hand signs. The accuracy was about 95%, depending on the characters and how they were performed. The recognition of short sentences (about 4 - 5 words) had an accuracy of about 85%. The recognition speed ranges from 9 to 11 frames per second, suitable for real-time applications.



(a) *(b)*

Figure 5. *Hand signs recognition.*

(a) Individual characters recognition; (b) Entire sentence recognition.

5. CONCLUSIONS AND FUTURE WORKS

Our proposed method has shown its effectiveness in representing both individual characters and entire sentences in Vietnamese in real time. The hand recognition method, which utilizes deep learning techniques and Google's MediaPipe library, exhibits superior performance compared to traditional image processing methods. The system is suitable for supporting communication with hearing-impaired people through hand sign recognition. Currently, we have successfully recognized individual characters (including Vietnamese-specific characters such as Đ, Ă, Â, Ê, Ô, Ơ, Ư), but we have not yet been able to recognize Vietnamese utilizes diacritics (“\”, “/”, “?”, “~”, “.”). In the future, we will continue to research to improve the accuracy of the recognition results and add recognition for Vietnamese utilizing diacritics.

REFERENCES

[1]. A. Ghotkar, *"Study of Vision Based Hand Gesture Recognition Using,"* vol. 7, no. 1, pp.96–115, (2014).

[2]. M. A. Almasre and H. Al-Nuaim, *"A comparison of Arabic sign language dynamic gesture recognition models,"* Heliyon, vol. 6, no. 3, p. e03554, (2020), doi: 10.1016/j.heliyon.2020.e03554.

[3]. V. Bheda and D. Radpour, *"Using Deep Convolutional Networks for Gesture Recognition in American Sign Language",* (2017), [Online]. Available: http://arxiv.org/abs/1710.06836.

[4]. B. Bagby, D. Gray, R. Hughes, Z. Langford, and R. Stonner, *"Simplifying Sign Language Detection for Smart Home Devices using Google MediaPipe,"* (2021), [Online]. Available: https://bradenbagby.com/Portfolio/Resources/PDFs/ResearchPaper.pdf

[5]. J. Shin, A. Matsuoka, M. A. M. Hasan, and A. Y. Srizon, *"American sign language alphabet recognition by extracting feature from hand pose estimation,"* Sensors, vol. 21, no. 17, pp. 1–19, (2021), doi: 10.3390/s21175856.

[6]. K. Gomase, A. Dhanawade, P. Gurav, and S. Lokare, *"Sign Language Recognition using Mediapipe," Int. Res. J. Eng. Technol.*, vol. 9, no. 1, pp. 744–746, (2022), [Online]. Available: https://www.irjet.net/archives/V9/i1/IRJET-V9I1133.pdf.

[7]. Duc-Hoang Vo, Huu-Hung Huynh, Thanh-Nghia Nguyen, and Jean Meunier. *"Automatic hand gesture segmentation for recognition of Vietnamese sign language."* In Proceedings of the Seventh Symposium on Information and Communication Technology, pp. 368-373. ACM, (2016).

[8]. Duc-Hoang Vo, Huu-Hung Huynh, Phuoc-Mien Doan and Jean Meunier, *"Dynamic Gesture Classification for Vietnamese Sign Language Recognition"*, International Journal of Advanced Computer Science and Applications(IJACSA), 8.3, pp. 415-420, (2017).

[9]. Anh H. Vo, Van-Huy. Pham, and Bao T. Nguyen *"Deep Learning for Vietnamese Sign Language Recognition in Video Sequence",* International Journal of Machine Learning and Computing, Vol. 9, No. 4, (2019).

TÓM TẮT

Xây dựng cầu nối giao tiếp với máy học trong nhận dạng ngôn ngữ ký hiệu cho tiếng Việt

Ngôn ngữ ký hiệu tiếng Việt (VSL) đóng vai trò là ngôn ngữ chính cho những người khiếm thính và khó nghe ở Việt Nam. Bài báo này khám phá quá trình nhận dạng ngôn ngữ ký hiệu cho VSL, nhấn mạnh vai trò của máy học trong việc thu hẹp rào cản giao tiếp. Chúng tôi đi sâu vào những điều cơ bản của VSL, trình bày chi tiết sự tương ứng một-một giữa các ký hiệu bằng tay và các chữ cái trong bảng chữ cái tiếng Việt và giải quyết sự hình thành các từ thông qua các ký hiệu bằng tay tuần tự và vị trí dấu phụ. Hơn nữa, bài báo nhấn mạnh tầm quan trọng của việc tạm dừng giữa các từ và việc sử dụng các thuật toán máy học để nhận dạng ký hiệu tự động. Cuối cùng, chúng tôi kết luận bằng cách thảo luận về các ứng dụng tiềm năng và hướng đi trong tương lai của công nghệ nhận dạng VSL tại Việt Nam.

Từ khoá: Ngôn ngữ ký hiệu Việt Nam; Nhận dạng ngôn ngữ ký hiệu; Mô hình học sâu; KNN.