Vision-Based Sign Language Recognition and Multilingual Translation for Facilitating Deaf and Mute Communication

S. V. Vasantha[1], A. Ashwini[1], M. Avinash[1](✉), M. Yuvaraj[1], R. Manisha[1], and Shirina Samreen[2]

[1] Department of CSE, Vardhaman College of Engineering, Hyderabad, India
machikaavinash@gmail.com

[2] College of Computer and Information Sciences, Majmaah University, Al Majma'ah, KSA, Saudi Arabia

Abstract. Sign language serves as the primary means of communication for individuals who are both deaf and mute. Nevertheless, communicating with those who do not understand sign language poses a significant challenge. Due to the structural differences between sign language and written/spoken languages, a communication barrier exists. Consequently, the interaction between deaf and mute individuals heavily relies on visual-based communication. To address this issue, a vision-based interface system has been developed to facilitate communication between deaf and mute individuals and the broader public. This system offers an interface capable of translating sign language gestures into text, enabling those unfamiliar with sign language to readily comprehend the message. The proposed system involves real-time video analysis for sign language identification and recognition, followed by the conversion of these visual inputs into English and other native languages. In this paper, French and Japanese languages are supported through Google API translator services, utilizing the SSD MobileNetV2 model for sign language recognition. The system achieved an outstanding overall accuracy of 0.98, with individual sign tokens demonstrating average accuracy scores of 0.99 for "Hello," 0.99 for "I love you," 0.98 for "Thank you," 0.97 for "Yes," 0.96 for "No," and 0.96 for "Help."

Keywords: Sign Language · Object Detection · Language Translation · Vision-based Interface · SSD MobileNetV2

1 Introduction

Deaf and mute individuals predominantly rely on sign language as their primary means of communication. Sign language serves not only as a means for effective communication but also as a gateway to conveying intricate concepts, engaging in nuanced conversations, and nurturing profound social connections within society. Sign language is a multifaceted and expressive mode of communication, incorporating a blend of hand gestures, facial

X. Cheng (Ed): BROADNETS 2024, LNICST 602, pp. 116–123, 2025.
https://doi.org/10.1007/978-3-031-81171-5_12

expressions, and body movements to convey a wide spectrum of messages. Deaf and mute individuals have developed their distinct sign languages, each shaped by regional and cultural influences. These sign languages exhibit remarkable versatility, enabling users to deliberate on various subjects, express their thoughts and emotions, and partake in in-depth dialogues, much akin to those who use spoken language.

In its essence, sign language empowers deaf and mute individuals to bridge the communication divide and participate fully in society. This, in turn, contributes to their personal growth and enriches their overall quality of life, emphasizing the vital role that embracing and accommodating diverse forms of communication plays in promoting inclusivity and understanding within our communities [1]. However, establishing successful communication with these people is a huge problem for those who do not understand sign language. The syntactic or structural differences between sign language and traditional written and spoken languages present this problem, creating a linguistic barrier that obstructs effective communication [2, 3]. The vision-based interface systems are designed to recognize and interpret sign language gestures and movements, allowing for communication between sign language users and individuals who may not be fluent in sign language closing the communication gap [4]. Once the system recognizes a sign or gesture, it can convert the detected signs into text. It uses computer vision and machine learning algorithms to recognize and analyze the gestures, handshapes, and movements made by signers. Developing accurate and reliable sign language detection systems presents several challenges, including variations in sign language gestures, lighting conditions, and background noise.

2 Related Work

This paper introduces a GUI-based technique that employs Visual Words set to recognize the sign alphabet(A-Z) and sign digits (0–9) in real-time video streams, providing both textual and spoken predictions for the identified signs. Segmentation is achieved through a combination of skin color and background subtraction techniques. Image analysis involves the extraction of SURF features, and histograms are used to associate signs with their respective labels. Classification is carried out using SVM and CNN [5]. The authors [6] have created a CNN-based system for recognizing 30 Arabic Sign Language gestures, designed to assist impaired individuals. The proposed approach involves segmenting the sign into a set of positional embedding stripes, which are subsequently fed into a transformer unit comprising 4 layers of self-attention and an MPN [7]. A hybrid network developed combines convolution and transformers for recognition of sign tokens, extracting initial features with a fine-grained approach and achieving impressive accuracy: 89% on a 77-label KSL dataset and 98% on a lab dataset, all with reduced computational requirements [8]. Authors in [9], introduced GlossFreeSLT, an innovative approach based on GFSLTVLP, enhancing SLT by leveraging language-oriented prior knowledge from pre-trained systems, without requiring gloss annotations. In paper [10], sign greeting communication system for Indonesia was developed. The researchers in [11] created a mobile app that translates Turkish Signs into text, with a specific focus on accuracy, achieved a 96.3% for three signs. In [12], an ML-based real-time model for instant translation of signs to English text was created. [13] Detect signs using a

pre-trained PoseNet ML model and classify them with ml5.neuralnetwork. System in [14], Recognized Pakistani Signs based harnessing DL Models with restricted data. [15] Introduced FingerSpeller, a camera-free smart ring solution for text entry using TapStrap's accelerometers. It employs Hidden Markov Models for precise fingerspelling recognition, proven in real-world testing.

3 Proposed System

In the initial phase of our research, we initiated the process of creating a dataset of sign language images through webcam-based capture. It involves capturing images for five specific sign language words: "thank you", "no", "I love you", "hello", "help," and "yes." These signs in the dataset are meticulously annotated, categorizing each image according to its corresponding sign language word. To streamline the annotation process, we employed a user-friendly graphical interface for adding bounding boxes to the images. Subsequently, we thoughtfully divided this annotated dataset into distinct training and testing subsets to assess our model's performance. Additionally, we expanded the proposed system's horizons by incorporating a translation service. The system seamlessly integrates a sign language gesture translator, enabling the translation of sign language gestures into different languages, including French and Japanese. This innovative feature serves as a bridge for cross-cultural communication on a global scale. This comprehensive system depicted in Fig. 1, spanning from image capture to translation, underscores the commitment to leveraging technology for universal accessibility and represents a substantial advancement in improving communication for individuals with disabilities. For the core recognition task, we turned to the TensorFlow deep learning framework, utilizing the SSDMobileNetV2 model. The SSDMobileNetV2 capitalizes on MobileNet architecture for feature extraction, characterized by depthwise separable convolutions that reduce parameters and computational complexity while preserving feature extraction capabilities, making it ideal for real-time applications. Additional convolutional layers were incorporated within the SSD architecture to create a feature pyramid, facilitating object detection across various sizes and scales. The function ReLU, introduced non-linearity to enhance the model's capacity to capture complex patterns. With an input size of 320 $\times$ 320 pixels, the SSDMobileNetV2 model showcased robustness in recognizing the chosen sign language words within our images. The implementation process of the proposed system is detailed in Algorithm 1.

Algorithm 1: Sign Language Recognition and Multilingual Translation
Input:
Webcam for image capture
Sign language dataset
Output:
Trained sign language recognition model

Step 1: Data Collection
1.1 Initialize an empty dataset D.
1.2 Activate the webcam for capturing sign language images.
1.3 Continuously capture images using the webcam.
1.4 Add each captured image to the dataset D.
1.5 Repeat steps 1.3 and 1.4 until a sufficient and diverse dataset is collected.

Step 2: Data Annotation
2.1 Initialize an empty set A for annotations.
2.2 After data collection, label each image in dataset D with its corresponding sign token.
2.3 Add the labelled images to the set of annotations A.

Step 3: Data Splitting
3.1 Divide the annotated dataset A into two distinct subsets: a training set A_train and a testing set A_test.
3.2 Ensure that the split maintains a balanced distribution of sign language words.
3.3 This division facilitates the assessment of the model's performance on new, unseen data.

Step 4: Fine-Tuning Pre-trained Model
4.1 Load a pre-trained SSD MobileNetV2 model.
4.2 Fine-tune the model using the annotated training dataset A_train.
4.3 The fine-tuning process adapts the model to recognize specific sign language words.
4.4 Observe that the use of a pre-trained model serves as a valuable starting point, saving time and computational resources.

Step 5: Feature Extraction
5.1 Utilize the MobileNetV2 architecture to extract meaningful features from sign language images.

Step 6: Object Detection
6.1 Perform object detection using the SSD framework for recognizing sign language gestures in real-time.

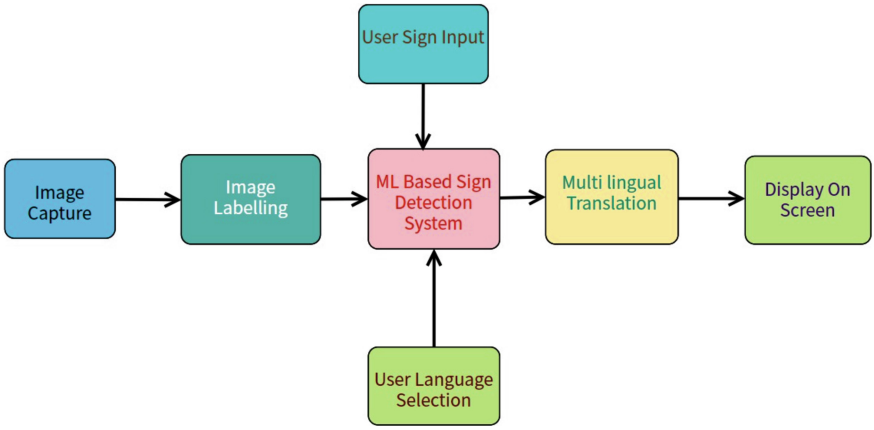

Fig. 1. The Proposed Vision-based Sign Language Recognition and Multilingual Translation System

4 Results and Comparative Analysis

The images are carefully annotated using labelImage.py script and the following libraries are utilized in the system implementation:

OpenCV: For webcam access and image processing
TensorFlow: For model inference and object detection
Matplotlib: For visualization of detection results.
Google API Translator: For multilingual translation

The sign recognized is shown with a bounding box and a sign word as a label, in this paper sign words are translated into English, Japanese and French languages as depicted in Fig. 2. Table 1 and Fig. 3 include the average accuracy levels of the considered six signs. Which range from 0.96 to 0.99 and mean model outstanding performance of 0.98 accuracy level. The proposed SSDMobileNetV2 exhibits notable improvement in accuracy of 0.98 when compared to contemporary models as detailed in Table 2.

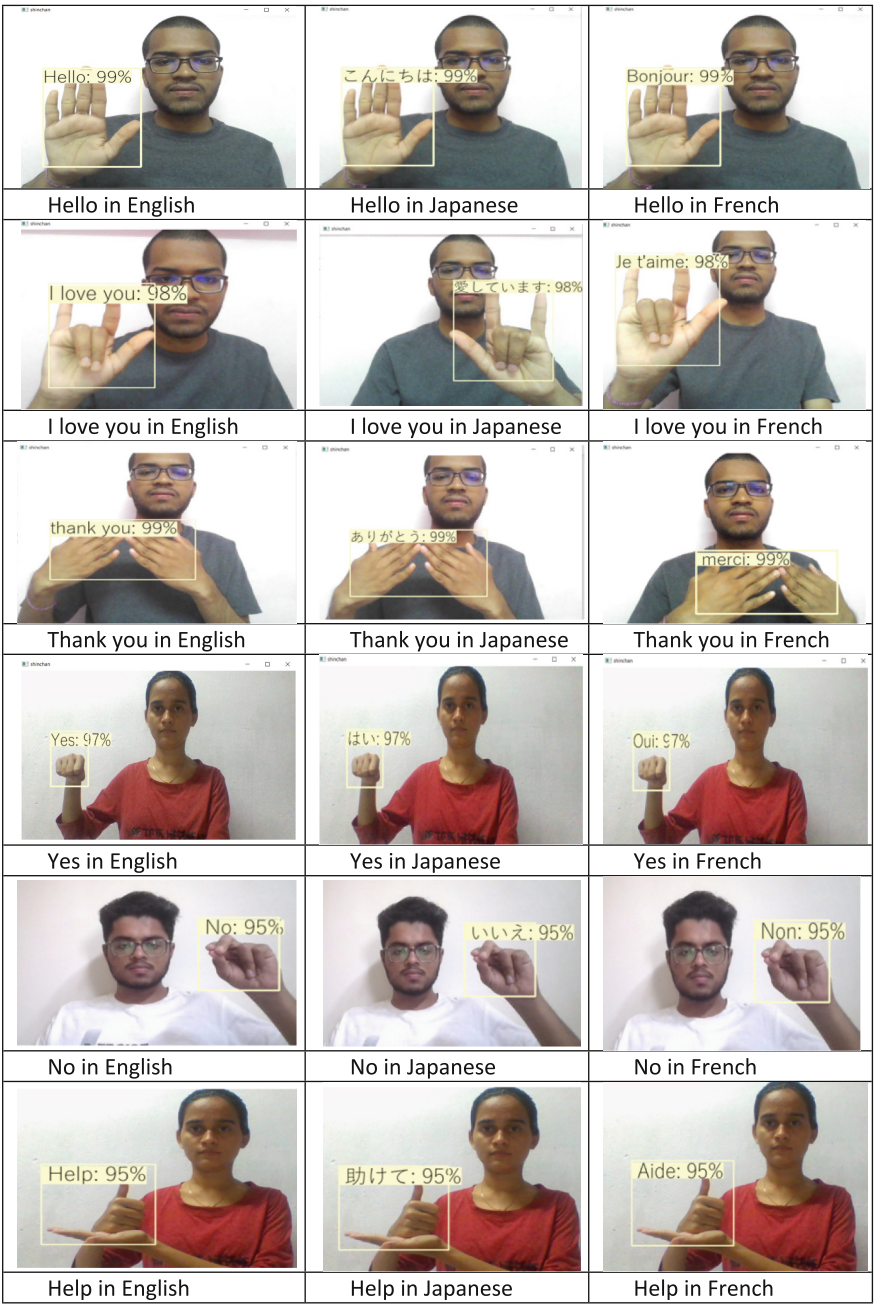

Fig. 2. Recognition and Multilingual Translation Results of the Sign Tokens

Table 1. Average Accuracy of Sign Tokens.

Sign Token	Accuracy
Hello	0.99
I love you	0.99
Thank you	0.98
Yes	0.97
No	0.96
Help	0.96
Model Overall Accuracy	0.98

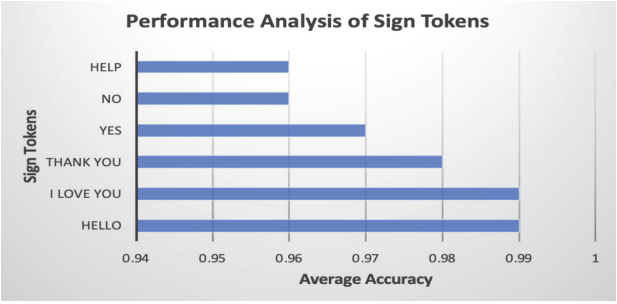

Fig. 3. Average accuracy levels of six sign tokens

Table 2. Comparison with Contemporary Models.

Model	Accuracy
Erdem et al.(2023) [11]	0.96
Abhirami et al.(2023) [13]	0.93
Hafiz et al.(2023) [14]	0.93
David et al.(2023) [15]	0.90
Proposed SSDMobileNetV2Model	**0.98**

5 Conclusion

The development of a vision-based interface system marks a significant advancement in enabling communication between individuals who are deaf and mute and the general public. This innovative system serves as a bridge, translating sign language into easily understandable text, and it extends its utility to support multiple languages, with a particular focus on French and Japanese. At the heart of this system lies the SSD MobileNetV2 model, which has undergone extensive training using an annotated dataset carefully

curated for the purpose of sign language recognition. The results are indeed impressive, with the model achieving an outstanding overall accuracy score of 0.98. Coming to the specific sign language tokens considered, it is observed that the system excels in recognizing these different signs. The "Hello" and "I love you" tokens exhibit exceptional accuracy, standing at an impressive 0.99. For "Thank you" it is at 0.98, while for "Yes" it's score is 0.97. Even for the slightly more challenging sign tokens, "No" and "Help," the system maintains a substantial accuracy level of 0.96.

References

1. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10023–10033 (2020)
2. Bragg, D., et al.: Sign language recognition, generation, and translation: an interdisciplinary perspective. In: Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, pp. 16–31 (2019)
3. Kahlon, N.K., Singh, W.: Machine translation from text to sign language: a systematic review. Universal Access in the Information Society **22**(1), 1–35 (2023)
4. De Castro, G.Z., Guerra, R.R., Guimarães, F.G.: Automatic translation of sign language with multi-stream 3D CNN and generation of artificial depth maps. Expert Systems with Applications **215**, 119394 (2023)
5. Katoch, S., Singh, V., Tiwary, U.S.: Indian Sign Language recognition system using SURF with SVM and CNN. Array. **1**(14), 100141 (2022)
6. Rwelli, R.E., Shahin, O.R., Taloba, A.I.: Gesture based Arabic Sign Language Recognition for Impaired People based on Convolution Neural Network (2022). arXiv preprint arXiv: 2203.05602
7. Kothadiya, D.R., Bhatt, C.M., Saba, T., Rehman, A., Bahaj, S.A.: SIGNFORMER: deepvision transformer for sign language recognition. IEEE Access **11**, 4730–9 (2023)
8. Shin, J., et al.: Korean sign language recognition using transformer-based deep neural network. Applied Sciences **13**(5), 3029 (2023)
9. Zhou, B., et al.: Gloss-free sign language translation: improving from visual-language pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20871–20881 (2023)
10. Andriana, A., et al.: Converter of Indonesian sign language into text and voice, text and voice to sign language to build between inclusion vocasional school student and teacher. In: AIP Conference Proceedings, Vol. 2510, No. 1. AIP Publishing (2023)
11. Demiroglu, E., Ayakdas, F., Tanribuyurdu, A., Kaya, G.A.: Sign language recognition mobile application for Turkish language. In: 9th International IFS and Contemporary Mathematics and Engineering Conference, p. 91 (2023)
12. Deepika, S., Rastogi, K., Jeevika, M.Y., Kumar, M.: Developing a machine learning model to translate sign language to English text in real time. J. Advanc. Softw. Eng. Testing. **6**(2), 39–47 (2023)
13. Abhirami, A., Anisha, G.S.: Indian sign language phrase estimation using PoseNet. In: 2023 3rd International Conference on Intelligent Technologies (CONIT), pp. 1–6. IEEE (2023)
14. Hamza, H.M., Wali, A.: Pakistan sign language recognition: leveraging deep learning models with limited dataset. Mach. Vis. Appl. **34**(5), 71 (2023)
15. Martin, D., et al.: FingerSpeller: camera-free text entry using smart rings for american sign language fingerspelling recognition. In: Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 1–5 (2023)