International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

H2V: A Real Time Sign to Text and Speech using CNN and Mediapipe

Prachi Agarwal[1],Paridhi Gupta[2], Priyanshu Chauhan[3], Anmol Gaur[4], Priyanshi Sharma[5]

[1] Assistant Professor, Computer Science and Engineering Department, MIT, Moradabad, India

reachtoprachi@gmail.com

[2,3,4,5] Computer Science and Engineering Department, MIT, Moradabad, India

paridhigupta099@gmail.com

priyanshuchauhan065@gmail.com

anmolgaur99565@gmail.com

qapple672@gmail.com

ABSTRACT

Communication between hearing-impaired individuals and the general population is often limited due to the lack of common communication tools. This paper introduces a real-time Hand-to-Voice (H2V) system that translates static sign language gestures into readable text and audible speech. The proposed approach captures live video through a standard webcam and detects hand movements using MediaPipe to obtain 21 landmark points. These landmarks are transformed into a skeletal representation, which is then analyzed by a Convolutional Neural Network (CNN) trained on grouped sign language gestures. To further reduce misclassification among visually similar signs, a rule-based refinement mechanism is applied after CNN prediction. The final output is displayed as text and converted into speech using a non-blocking text-to-speech module. Practical testing indicates that the system operates efficiently in real time for a predefined set of gestures, demonstrating its suitability for assistive communication.

Keywords

Hand Gesture Recognition, Sign Language Translation, Convolutional Neural Network, MediaPipe, Assistive Systems

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

1. Introduction

Sign language is an essential means of communication for individuals with hearing and speech impairments. However, the absence of universal sign language knowledge among the general population often creates communication barriers. With advancements in computer vision and machine learning, automated sign language interpretation systems have gained attention as a potential solution.

The Hand-to-Voice (H2V) system proposed in this work focuses on translating hand gestures into text and speech in real time using vision-based techniques. Unlike hardware-intensive solutions that rely on gloves or sensors, the proposed system requires only a webcam. By integrating hand landmark detection, deep learning-based classification, rule-based correction, and speech synthesis, the system aims to provide an accessible and cost-effective communication aid.

2. Related Work

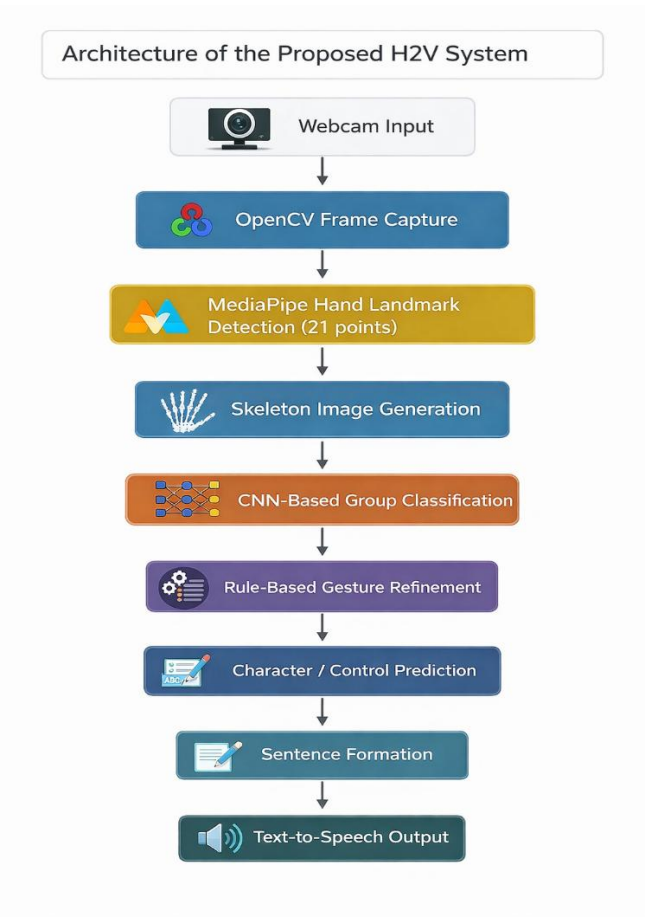
Earlier studies in sign language recognition employed sensor-based gloves, depth cameras, and marker-driven systems. Although these methods offered reasonable accuracy, they were often expensive and inconvenient for everyday use. Recent vision-based approaches utilizing convolutional neural networks and hand landmarks have shown improved flexibility and performance.

Several works have explored CNN-based gesture recognition; however, many lack real-time speech output or rely on raw image inputs that are sensitive to background variations. The proposed system differentiates itself by using skeleton-based representations, hybrid classification strategies, and a real-time graphical interface with speech feedback.

3. System Architecture

The proposed H2V system follows a sequential processing pipeline, beginning with video acquisition and ending with speech generation. Live video frames captured from a webcam are processed to detect hand landmarks. These landmarks are converted into skeleton images, which are classified by a CNN model. The CNN output is refined using geometric rules, after which characters are combined into meaningful text and converted into speech.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)



4. METHODOLOGY

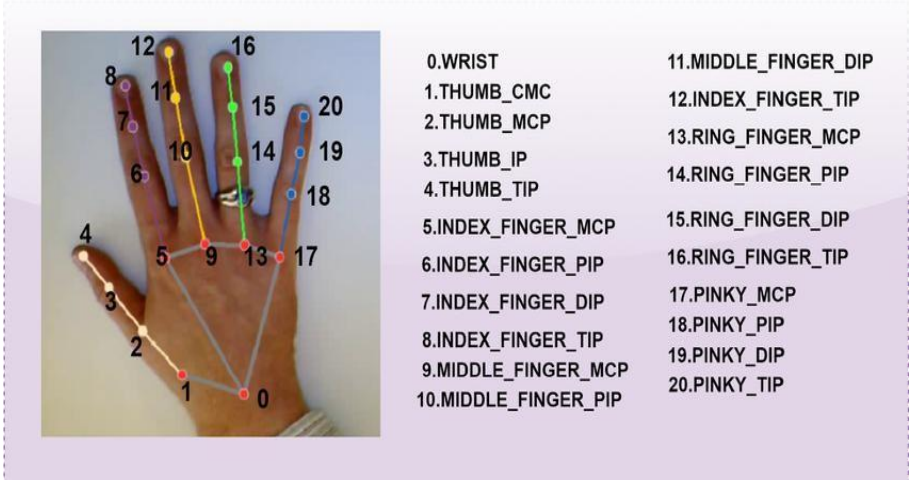
4.1 HAND DETECTION AND LANDMARK EXTRACTION

The system captures continuous video frames using OpenCV. MediaPipe Hands is employed to identify a single hand within each frame and extract 21 landmark points corresponding to finger joints and palm locations.

4.2 SKELETON REPRESENTATION

The extracted landmarks are used to generate a skeletal hand structure on a plain background. This representation emphasizes hand geometry while minimizing the influence of background noise and lighting variations.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)



4.3 DATASET PREPARATION

A custom dataset is developed by capturing skeleton images for individual alphabet gestures. For each class, multiple samples are collected and stored in a structured manner to support supervised training of the CNN model.

4.4 CNN-BASED CLASSIFICATION

A Convolutional Neural Network is designed with multiple convolutional, pooling, normalization, and dense layers. The network is trained using categorical cross-entropy loss and the Adam optimizer to classify gestures into predefined categories.

4.5 GROUPING STRATEGY

To handle visual similarities between certain gestures, alphabets are organized into logical groups. The CNN predicts group-level classes, which simplifies the classification task and reduces confusion between similar signs.

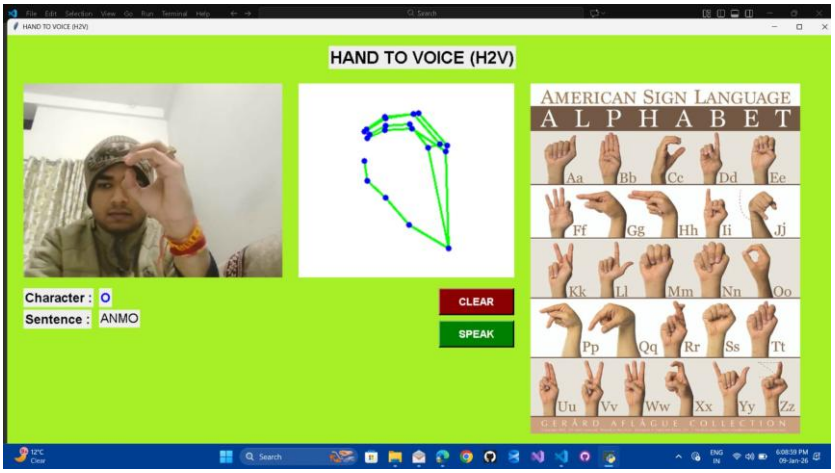
4.6 RULE BASED REFINEMENT

Following CNN prediction, a set of geometric rules based on landmark distances and finger orientations is applied. This step improves recognition accuracy for ambiguous gestures such as C and O, or U and V.

4.7 TEXT CONSTRUCTION AND STABILITY CONTROL

To ensure stable output, predictions are confirmed only after consistent detection across multiple frames. Special gestures such as space and delete are processed separately to support sentence construction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

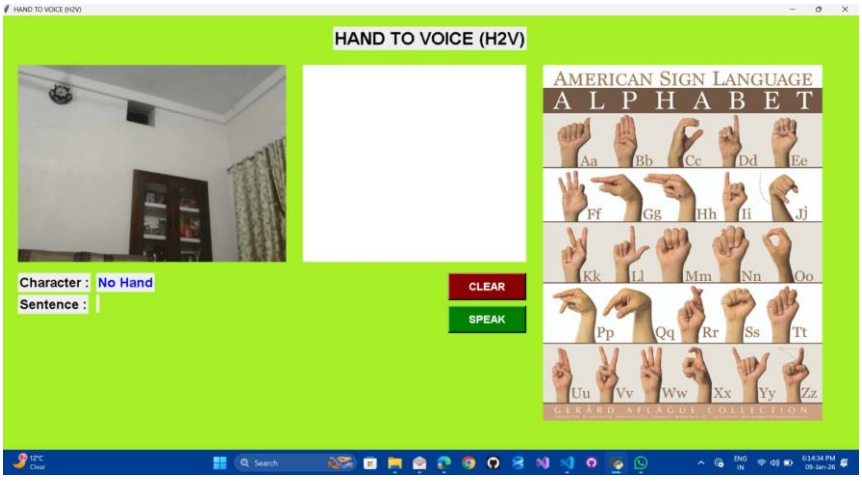


4.8 TEXT-TO- SPEECH

The recognized text is converted into speech using a queue-based text-to-speech engine. This design prevents blocking of the graphical interface and ensures smooth audio output.

5. GRAPHICAL USER INTERFACE

A user-friendly graphical interface is developed using Tkinter. The interface displays the live video feed, skeletal visualization of the detected hand, recognized characters, and the generated sentence. Additional controls allow users to clear text or trigger speech output.



6. EXPERIMENTAL OBSERVATIONS

The system was evaluated under controlled indoor lighting conditions using a standard webcam. It demonstrated real-time responsiveness and reliable recognition for static hand gestures. Performance was observed to vary with hand orientation, distance from the camera, and lighting consistency.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

7. ADVANTAGES

1. Requires only a standard webcam
2. Operates in real time
3. Combines deep learning with rule-based correction
4. Provides both text and speech output
5. Easy to use graphical interface

8. LIMITATIONS

1. Supports only static gestures
2. Sensitive to poor lighting conditions
3. Limited gesture vocabulary

9. FUTURE ENHANCEMENT

Future work may extend the system to dynamic gesture recognition, incorporate two-hand interactions, enable continuous word-level recognition, and support deployment on mobile platforms with multilingual speech output.

10. CONCLUSION

This paper presented a real-time Hand-to-Voice (H2V) system for translating sign language gestures into text and speech. By combining MediaPipe-based hand landmark detection, CNN-driven skeleton classification, and rule-based refinement, the system offers an effective and accessible solution for assistive communication. The proposed approach demonstrates practical feasibility and provides a strong foundation for further enhancements.

11. REFERENCES

[1] K. Yaseen, O.-J. Kwon, J. Kim, S. Jamil, J. Lee, and F. Ullah, “Next-Gen Dynamic Hand Gesture Recognition: MediaPipe, Inception-v3 and LSTM-Based Enhanced Deep Learning Model,” Electronics, vol. 13, no. 16, p. 3233, 2024, doi:10.3390/electronics13163233.

[2] A. Khatak and S. Naaz, “Real-Time Multi-Mode Hand Gesture Recognition Using MediaPipe and Deep Learning for Human-Computer Interaction,” Journal of Computational Analysis and Applications, vol. 33, no. 08, pp. 6610–6621, 2024.

[3] W. Utomo, Y. Suhanda, H. Ar-Rasyid, and A. Dharmalau, “Indonesian Language Sign Detection using Mediapipe with Long Short-Term Memory (LSTM) Algorithm,” J. of Informatics and Web Engineering, 2025.

[4] F. S. Takouchouang and H. T. Vinh, “Reconnaissance Automatique des Langues des Signes: Hybrid CNN-LSTM Approach Based on Mediapipe,” arXiv:2510.22011, 2025.

[5] K. Madhurima and M. Maneesha, “Sign Language Recognition Using CNN and Hand Gestures Tracking,” Int. J. of Eng. Research and Science & Technology, vol. 21, no. 4, pp. 341–345, 2025.

[6] S. Kamble, “SLRNet: A Real-Time LSTM-Based Sign Language Recognition System,” arXiv:2506.11154, 2025.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

[7] S. Huse, R. Makode, T. Wankhade, and T. Nachane, “Real-Time ISL Recognition using CNN and MediaPipe,” Int. J. for Multidisciplinary Research, 2025.

[8] “Hand Gesture Detection for Sign Language using CNN and MediaPipe,” IJCRT, vol. 13, no. 4, 2025.