IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, May 2026
Impact Factor: 8.2

Wearable Vision-Based System for Real-Time Sign Language Recognition and Audio Translation

Manpreet Kaur, Vansh Choudhary, Rachit Singh, Adwait, Rahul Diwakar
Raj Kumar Goel Institute of Technology, Ghaziabad
mkaurfio@rkgit.edu.in, 26iojitsh@rkgit.edu.in
26ionavit@rkgit.edu.in, 26ioranit@rkgit.edu.in, 26iobanul@rkgit.edu.in

Abstract: *According to the World Health Organisation (WHO), more than 5% of the global population, representing over 430 million individuals, experience disabling hearing loss. A large portion of this population relies on sign language as a primary mode of communication. Also, less awareness and understanding of sign language among the general public create significant communication barriers. To solve this issue, this paper proposes a wearable vision-based real-time sign language translation system. The proposed system utilises an ESP32-CAM module mounted on wearable to capture hand gestures performed in front of the user. Hand landmark and coordinates are extracted using MediaPipe and processed through a trained machine learning model developed on a custom-built dataset captured from multiple angles. The recognised gestures are mapped to corresponding English text and converted into speech using a text-to-speech (TTS) module. The audio is delivered as output through a wireless Bluetooth speaker integrated into the wearable, enabling instant auditory feedback.*
The experiment's results demonstrate that the system achieves reliable real-time gesture recognition while maintaining low cost, and practical usage. The proposed approach provides an effective assistive communication solution by integrating computer vision, machine learning, and wearable hardware into a well-defined framework.

Keywords: American Sign Language (ASL), Wearable Assistive System, ESP32-CAM, MediaPipe, Hand Landmark Detection, Real-Time Gesture Recognition, Text-to-Speech (TTS)

I. INTRODUCTION

Sign language translation plays an key role in reducing communication hurdles faced by individuals with hearing or speech issues. It enables effective communication in social environments and can also be applied in human-to-machine interaction to support nonverbal communication. It helps in the translation of hand gestures and their conversion into a form that can be easily understood by people, such as text or speech.

Sign language recognition has received noticeable attention in the area of computer vision, aiming at the accurate detection of hand movements and the interpretation of their respective meaning. Sign language recognition applications include assistive communication for hearing-impaired people, interactive systems. Also, there are several challenges involved in the real-time recognition of sign language gestures, including hand shape, speed of gestures, lighting.

Numerous sign languages are used worldwide, including American Sign Language (ASL), Indian Sign Language (ISL), and German Sign Language (GSL). Among these, American Sign Language (ASL) is widely adopted in research due to its one-handed manual alphabet, which simplifies gesture representation and classification. In contrast, sign languages such as ISL and British Sign Language (BSL) predominantly uses two-handed gestures, resulting in increased recognition complexity. ASL is internationally recognised and its strong connection with the English language - one of the most widely used languages globally - increasing its accessibility and practicality . So, ASL is selected in this research as it is well suited for real-time, vision-based wearable systems.

DOI: 10.48175/IJARSCT-34820

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, May 2026
Impact Factor: 8.2

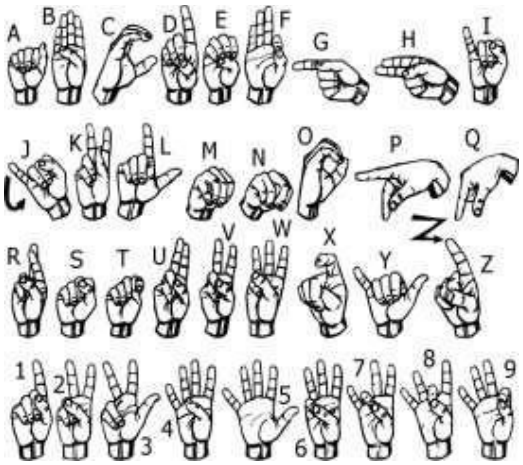

Fig. 1. Basic Gestures in American Sign Language [20]

To enhance real-world usage, word's gesture recognition offers a more efficient communication mechanism by directly translating complete sign gestures into meaningful audio output. Recognising full words instead of single letters significantly reduces processing steps and improves response time for the communication systems.

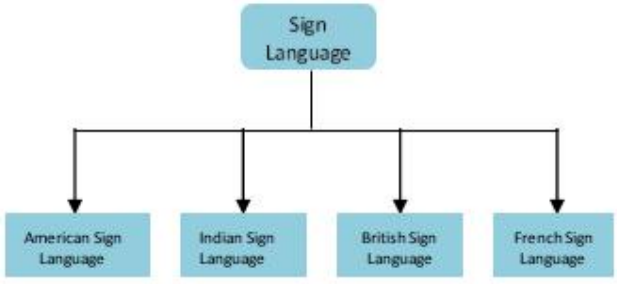

Fig.2 Types of sign language.

Sign Language	Comparison		
	Similarities	Differences	Region
American Sign Language [1]	Involves both static and dynamic hand gestures	Uses a one-handed manual alphabet; vowels are represented using finger positions	United States and parts of Canada
Indian Sign Language [2]	Involves both static and dynamic hand gestures	Uses a two-handed manual alphabet, increasing gesture complexity	South Asian countries
British Sign Language [3]	Similar to ISL in structure; uses two-handed gestures	Uses thumb and four fingers to represent vowels	England and Northern Ireland

Table.1 Comparison Between Various Sign Languages

In this paper, The gesture capture is performed using an ESP32-CAM module placed on glasses, allowing real-time video streaming to a laptop that works as the main processing unit. The trained classification model recognize the performed gesture, which is then converted into English text and output is given in the form of audio output . The audio is delivered through a wireless Bluetooth speaker placed on wearable, providing immediate feedback.

The respective system focuses on portability, affordability, real-time performance, and practical deployment. By combining gesture recognition of meaningful words, skeleton-based feature extraction, and wearable hardware

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, May 2026
Impact Factor: 8.2

integration, this work moves beyond laboratory-based prototypes toward a more functional and practical communication solution.

II. LITERATURE REVIEW

According to the World Health Organization [1], around 5% of the world's population requires treatment to heal their "disabling" hearing loss (435 million adults and 35 million children). It is roughly assumed that by 2050, over 700 million people, or one in every ten people, will have a hearing impairment. That's why, studies are done to help disabled people to communicate freely.

Early vision-based systems mainly depends on image processing and handcrafted features for gesture classification. Shivashankara and Srinath [2] proposed an effective ASL recognition framework using feature extraction and classification techniques under ideal conditions. Also, Rokade and Jadav [3] developed a recognition -based sign language recognition system focusing on structured gesture representation.

With the rapid advancement of deep learning, Convolutional Neural Networks (CNNs) came to be a dominant method for gesture recognition. Shahriar et al. [4] implemented a real-time ASL recognition system using skin segmentation methods integrated with CNN architecture, resulting in good classification performance. Jain et al. [5] compared Support Vector Machines and CNN models on the Sign Language MNIST dataset [6], reporting comparatively higher accuracy using CNN. Similar CNN-based systems were presented by Kadhim and Khamees [7], showing effective recognition accuracy on real-world datasets.

Even after their promising results, most of these systems focus on alphabet-level classification rather than recognizing complete words. Letter-based recognition requires step by step interpretation to form meaningful text, which may increase time-taken in real-time communication situations. For assistive wearable applications, direct word-level recognition can improve interaction efficiency.

Numerous advancements in hand pose estimation have introduced landmark-based recognition techniques. Instead of processing raw pixel data, these approaches extract structured hand keypoints from RGB images. Zimmermann and Brox [8] introduced a 3D hand pose estimation technique that allows keypoint extraction from single images. These techniques representations reduce sensitivity to background noise and lighting changes and require lower resources compared to heavy CNN models.

For dynamic gestures, temporal modeling techniques such as 3D Convolutional Neural Networks have also been taken into account . Huang et al. [9] utilized 3D CNNs to capture spatio- temporal information in gesture sequences. Sharma and Singh [10] tested multiple deep learning architectures for gesture recognition and resulted in strong performance using deep convolutional architecture. Carneiro et al. [11] proposed a multi-stream CNN-based approach for video gesture datasets, focusing on improved translation through parallel feature extraction.Even though , these studies report high accuracy values, several limitations still remain unattended. Many systems depend heavily on publicly available benchmark datasets and operate under ideal experimental conditions. Also, most implementations remain desktop-based prototypes without integration into portable wearable devices. Real-time translation systems with direct audio feedback are comparatively less explored even after so much time.

In contrast, the proposed work focuses on a practical wearable implementation that integrates real-time gesture capture, skeleton-based feature extraction, classification, and immediate speech output into a well-defined system. Instead of depending solely on benchmark datasets, a self- made dataset containing word-level ASL gestures is developed from different viewing angles to improve generalization. Hand landmark positions are gathered using a pose estimation architecture, and the recognized gestures's audio output is provided via speaker. The integration of an ESP32-CAM module mounted on wearable glasses with wireless audio output provides a complete assistive communication pipeline aimed at real-world usability.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, May 2026
Impact Factor: 8.2

III. PROPOSED METHODOLOGY

In the given flow chart, we have shown our methodology for sign language recognition.We preprocessed the dataset and then split it into training and testing dataset. Then we train our model with training dataset. Thereafter, we test our trained model with testing dataset and optimized it.

A. Dataset Creation

Since the system focuses on word-level recognition instead of alphabet-level classification, a self-made dataset was developed specifically for predefined American Sign Language (ASL) word gestures. These gesture samples were collected using a ESP-32 camera and processed through a hand landmark detection framework. To improve model robustness and generalisation, data was captured from multiple viewing angles and under different hand orientations. Each gesture class contains multiple samples to make sure balanced training data.Unlike benchmark datasets that depends on static lab oriented images, this dataset shows practical real-world usage conditions, making it suitable for wearable deployment.

B. Hand Landmark Extraction

In place of using raw RGB images as model input, a skeleton- based feature representation was selected. Hand landmarks were extracted using a real-time hand pose estimation framework.For every detected hand, 21 key landmark points were collected , representing the spatial coordinates (x, y, z) of important joints such as fingertips, knuckles, and wrist. These structured coordinate values creates the feature vector for classification.

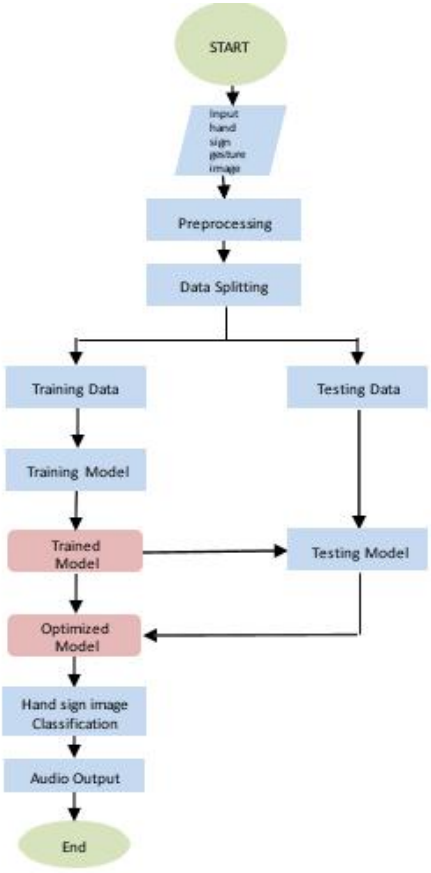

Fig.3 Flowchart of Proposed Methodology

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, May 2026
Impact Factor: 8.2

The advantages of using landmark-based features include:
• Less sensitive to lighting conditions
• Lower computational complexity
• No Background dependency
• Comparatively faster real-time processing
This approach makes the system suitable for wearable real- time implementation.

C. Data Preprocessing
After landmark extraction, the following preprocessing steps were applied:
1. Firstly, Normalization of landmark coordinates to reduce scale variation takes place.
2. Then, Flattening of 3D coordinate values into a structured feature vector
3. Later on, Label encoding of gesture classes
4. Splitting the dataset into training and testing sets to ensure data accuracy.
These preprocessing steps results in consistency and stability during model training.

D. Model Training and Classification
The processed landmark feature vectors were used to train a supervised machine learning classification model. The model learns the patterns of finger positions respective to each word- level ASL gesture.
As the system recognizes complete words instead of letters, the classification output directly maps the respective audio clip, reducing translation delay.
The trained model was tested against all odds using standard performance metrics such as:
• Training accuracy
• Testingaccuracy
• Confusion matrix
The evaluation confirmed that the model can effectively
differentiate between the predefined word gestures.

E. Real-Time Gesture Recognition Pipeline
For real-time operation, an ESP32-CAM module is mounted on wearable glasses to capture live video streams. The video stream is transmitted wirelessly to a laptop, which acts as the main processing unit.
The real-time pipeline operates as follows:
1. Video frame capture from ESP32-CAM
2. Hand landmark detection
3. Feature vector generation
4. Model-based gesture prediction as per training
5. Text mapping of predicted gesture
This pipeline make sure of low-latency recognition best for real-time communication.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, May 2026
Impact Factor: 8.2

F. Speech Output Generation

After classification, the predicted gesture is converted into English language text and its respective audio output will played. The generated audio output is played through the Bluetooth speaker placed on the wearable glasses.

This allows immediate feedback, allowing the user to hear the translated word instantly. The integration of gesture recognition and audio feedback completes the communication loop.

VI. RESULTS & DISCUSSION

The performance of the wearable word-level sign language translation system was tested through both offline dataset testing and real-time deployment testing. The self-made dataset consisting of predefined ASL word gestures was divided into training and testing sets to ensure unbiased testing.

The classification model trained on normalized hand landmark coordinates shows stable convergence during training, resulting that skeleton-based spatial features effectively capture the structural differences among different gestures. The testing phase confirmed that the model was able to consistently differentiate between visually similar finger configurations, suggesting that multi-angle data collection obviously improved generalization capability. Unlike conventional image-based CNN systems that depend on raw pixel values, the use of 21 key landmark coordinates per frame significantly reduced feature dimensionality, resulting in faster inference time and lower computational complexity. During real-time evaluation, live video frames streamed from the ESP32-CAM placed on wearable glasses were processed sequentially with least possible lag. The complete recognition pipeline—from frame collection to audio output generation— operated smoothly, demonstrating low-latency performance suitable for effective communication. The average response gap between gesture presentation and speech output was minimum, allowing natural communication flow. The Bluetooth-based audio delivery system provided immediate feedback without requiring visual confirmation, hence enhancing accessibility for users in dynamic environments. Also, the system showcased stable prediction performance under moderate lighting variations and different indoor backgrounds, ensuring the accuracy of landmark-based feature extraction. However, slight reductions in accuracy were observed when hand detection confidence dropped due to extreme lighting conditions. Overall, the experiment's results shows that connecting skeleton-based gesture recognition with wearable hardware components creates a efficient and practically working communication framework. The architecture is scalable and can be extended to support a larger vocabulary set, dynamic gestures, or edge-based processing in future implementations.

VII. CONCLUSION

This work presented a wearable real-time sign language translation system designed to improve communication for individuals depending on American Sign Language. Unlike many existing systems that mainly focuses on alphabet recognition under laboratory ideal conditions, the given system focuses on word-level gesture recognition connected into a practical wearable framework. By using hand landmark extraction instead of raw image-based processing, the system maintains computational efficiency while maintaining reliable gesture difference. The use of a self-made multi-angle dataset further enhances model generalization in real-world scenarios.The practical connection of an ESP32-CAM module for live gesture capture and a wireless Bluetooth speaker for audio output enables a complete end-to-end communication system. The system successfully showcases real-time performance with low latency, making it effective for communicating purposes . Its lightweight hardware configuration and low-cost components ensures and improves portability and accessibility, which are key factors in assistive technology applications.Even if the current implementation supports a limited vocabulary of predefined gestures, the framework is scalable and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, May 2026
Impact Factor: 8.2

adaptable for future expansion. Overall, the proposed approach highlights the effectiveness of combining skeleton-based gesture recognition with wearable hardware integration to create a practical, efficient, and user- friendly assistive communication solution.

REFERENCES

[1]. World Health Organization, "Deafness and hearing loss," 2021.
[2]. S. Shivashankara and S. Srinath, "American Sign Language Recognition System: An Optimal Approach," Int. J. Image, Graphics and Signal Processing, 2018.
[3]. Y. Rokade and P. Jadav, "Indian Sign Language Recognition System," Int. J. Engineering and Technology, 2017.
[4]. S. Shahriar et al., "Real-Time American Sign Language Recognition Using Skin Segmentation and CNN," IEEE TENCON, 2018.
[5]. V. Jain et al., "American Sign Language Recognition using SVM and CNN," Int. J. Information Technology, 2021.
[6]. Kaggle, "Sign Language MNIST Dataset," 2017.
[7]. R. A. Kadhim and M. Khamees, "A Real-Time ASL Recognition System using CNN for Real Datasets," 2020.
[8]. C. Zimmermann and T. Brox, "Learning to Estimate 3D Hand Pose from Single RGB Images," arXiv, 2017.
[9]. J. Huang et al., "Sign Language Recognition using 3D Convolutional Neural Networks," IEEE ICME, 2015.
[10]. S. Sharma and S. Singh, "Vision-based Hand Gesture Recognition using Deep Learning," Expert Systems with Applications, 2021.
[11]. A. L. C. Carneiro et al., "Efficient Sign Language Recognition System using Deep Learning," ICDIP, 2021.
[12]. O. Koller, H. Ney, and R. Bowden, "Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016.
[13]. N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition," Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017.
[14]. A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," Proc. Int. Conf. Machine Learning (ICML), 2006.
[15]. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Proc. Int. Conf. Learning Representations (ICLR), 2015.
[16]. A. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv:1704.04861, 2017.
[17]. F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017.
[18]. T. Starner and A. Pentland, "Real-Time American Sign Language Recognition from Video Using Hidden Markov Models," MIT Media Laboratory Technical Report, 1995.
[19]. S. Escalera et al., "ChaLearn Gesture Challenge 2014: Dataset and Results," Proc. European Conf. Computer Vision (ECCV) Workshops, 2014.
[20]. Rao, G. Anantha, K. Syamala, P. V. V. Kishore, and A. S. C. S. Sastry. "Deep convolutional neural networks for sign language recognition." In 2018 Conference on Signal Processing And Communication Engineering Systems (SPACES), pp. 194-197. IEEE, 2018.