RESEARCH PAPER

Real-Time Sign Language Recognition and Speech Synthesis Browser Extension

Dr R Murugesan[1], Mrs Kavitha V[2], Sathiyapriyan B[3], Santhosh S S[4], Bhavesh Kumaran J U[5], Gomathi Nayagam P[6]

[1]Guide, Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur
Email: rmurugesanvsb@gmail.com
[2]Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur
Email: kavithataru2015@gmail.com
[3]Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur
Email: sathiyapriyansathi@gmail.com
[4]Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur
Email: san1282004@gmail.com
[5]Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur
Email: bhaveshkumaran684@gmail.com
[6]Department of Artificial Intelligence and Data Science, V.S.B. Engineering College, Karur
Email: jeevaraja648@gmail.com

Received: 17th Mar, 2026 | Revised: 29th Mar, 2026 | Accepted: 19th Apr, 2026 | Available Online: 5th May, 2026

ABSTRACT

Sign language is the primary mode of communication for deaf and mute individuals. However, communication between sign language users and non-signers remains a major challenge. This project proposes a real-time sign language recognition system that converts hand gestures into text and speech using a browser extension. The system utilises computer vision and deep learning techniques to recognise hand gestures through a live camera feed. MediaPipe is used for hand landmark detection, and a trained classification model converts gestures into textual output. The generated text is displayed as subtitles within the browser extension and further converted into speech using a text-to-speech engine. This solution enables seamless communication in real-time environments such as online meetings, classrooms, and public services. The proposed system is lightweight, accessible, and deployable as a browser extension, making it practical for everyday use.

Index Terms: CNN and LSTM, sign language recognition.

How to cite this article: Murugesan R, Kavitha V, Sathiyapriyan B, Santhosh SS, Bhavesh Kumaran JU, Gomathi Nayagam P., Real-Time Sign Language Recognition and Speech Synthesis Browser Extension. Int J Drug Deliv Technol. 2026;16(44s): 343-352; DOI: 10.25258/ijddt.16.44s.39

I. Introduction

Communication is a fundamental human necessity that enables individuals to express ideas, emotions, and information effectively. However, for people with hearing and speech impairments, communication with the general population remains a significant challenge. According to global disability reports, millions of individuals rely on sign language as their primary mode of interaction. Since sign language is visual-gestural and not widely understood by non-signers, a communication gap exists in everyday environments such as classrooms, workplaces, hospitals, and public service centres.

Sign Language Recognition (SLR) systems aim to bridge this gap by translating hand gestures into textual or spoken language. Traditional SLR systems relied on data gloves, colored markers, or sensor-based devices to capture hand movements. Although these systems provided reasonable accuracy, they were expensive, intrusive, and unsuitable for real-time public use. With the advancement of computer vision and artificial intelligence, vision-based sign language recognition has emerged as a practical and scalable

alternative.

Recent developments in deep learning, particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, have significantly improved gesture recognition performance. These models can learn spatial hand features and temporal motion patterns from image and video sequences. Additionally, pose estimation frameworks such as MediaPipe and OpenPose enable real-time extraction of hand landmarks, reducing computational complexity while maintaining recognition accuracy.

Despite these technological advancements, most existing sign language recognition systems are limited to standalone desktop applications or research prototypes. They often lack real-time subtitle visualisation, speech synthesis, and integration into commonly used digital platforms such as web browsers. This limits their usability in online communication environments like virtual meetings, e-learning platforms, and video conferencing systems.

To address these limitations, this project proposes a Real-Time Sign Language Recognition and Speech Synthesis System implemented as a Browser Extension. The system captures live video input through a webcam and processes frames using a hand-tracking model to extract landmark features. These features are fed into a trained gesture-classification model that converts recognised signs into text.

Furthermore, to enable two-way communication, the system integrates a Text-to-Speech (TTS) engine that converts recognised text into audible speech. This allows sign language users to communicate seamlessly with non-signers in real time, without requiring specialised hardware or software.

The browser-based deployment offers several advantages, including platform independence, ease of access, and real-time usability across web applications. The proposed system is lightweight, cost-effective, and scalable, making it suitable for assistive communication, online education, telehealth services, and inclusive digital interactions.

The primary goal of this project is to design and develop an intelligent assistive system that leverages computer vision, machine learning, and web technologies to minimise the communication barrier between deaf-mute individuals and society.

II. Literature Survey

Several research works have been carried out in the field of sign language recognition. Deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are widely used for gesture classification.

Sequence-to-sequence models like S2VT have been applied to convert sign videos into text. These models process video frames as sequential inputs and generate textual outputs. However, such models often require large computational resources.

Parameter optimisation techniques such as tensor-train decomposition have been proposed to reduce model complexity while maintaining accuracy.

Recent systems also incorporate pose estimation frameworks like MediaPipe and OpenPose to extract hand landmarks, improving recognition performance in real-time applications.

III. Problem Statement

Existing sign language recognition systems are often limited to laboratory environments and require specialised hardware. Most systems do not provide real-time subtitle generation or speech output within commonly used digital platforms such as web browsers.

There is a need for an accessible, real-time, and deployable solution that can recognise sign language gestures using standard webcams and convert them into both text and speech for effective communication.

IV. Objectives

- To develop a real-time sign language recognition model
- To convert hand gestures into textual output
- To display recognised text as subtitles
- To generate speech from detected text
- To implement the system as a browser extension
- To enable accessible communication

V. Proposed System

The proposed system consists of a gesture recognition model integrated into a browser extension. The extension activates the user's webcam to capture live video. Hand landmarks are extracted using MediaPipe, and the trained model classifies gestures into text labels.

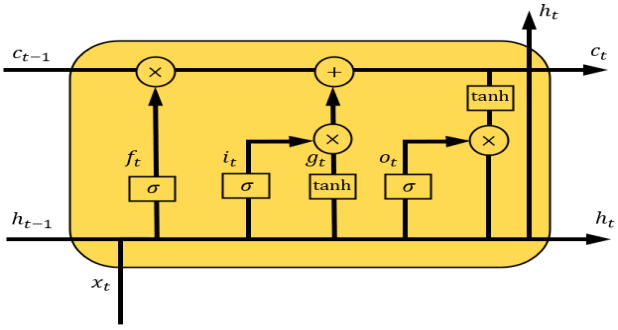
The predicted text is displayed as subtitles within the browser interface. A text-to-speech module converts the text into audio output, enabling two-way communication.

The system operates in real time and does not require external hardware, making it salable and user-friendly.

VI. System Architecture

The proposed system is designed as an end-to-end real-time assistive communication framework that converts sign language gestures into textual subtitles and synthesised speech through a browser extension. The architecture integrates computer vision, deep learning, and web technologies into a unified pipeline capable of operating with standard webcam hardware.

The system consists of six major functional modules: Video Acquisition, Hand Detection, Feature Extraction, Gesture Classification, Subtitle Generation, and Speech Synthesis. These modules operate sequentially to process live video streams and generate multimodal outputs.



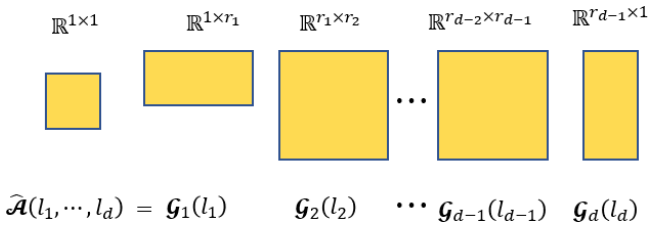
Initially, the video acquisition module activates the user's webcam through browser media APIs. Live frames are continuously captured and forwarded to the processing pipeline. Since raw video contains background noise and irrelevant information, frames are preprocessed and resized to ensure uniform input dimensions.

The hand detection module employs a real-time pose estimation framework such as MediaPipe Hands. This module detects the presence of hands within each frame and extracts 21 three-dimensional landmarks representing finger joints and palm positions. These landmarks form a compact skeletal representation of hand gestures, eliminating the need for heavy image processing.

Following detection, the feature extraction module converts

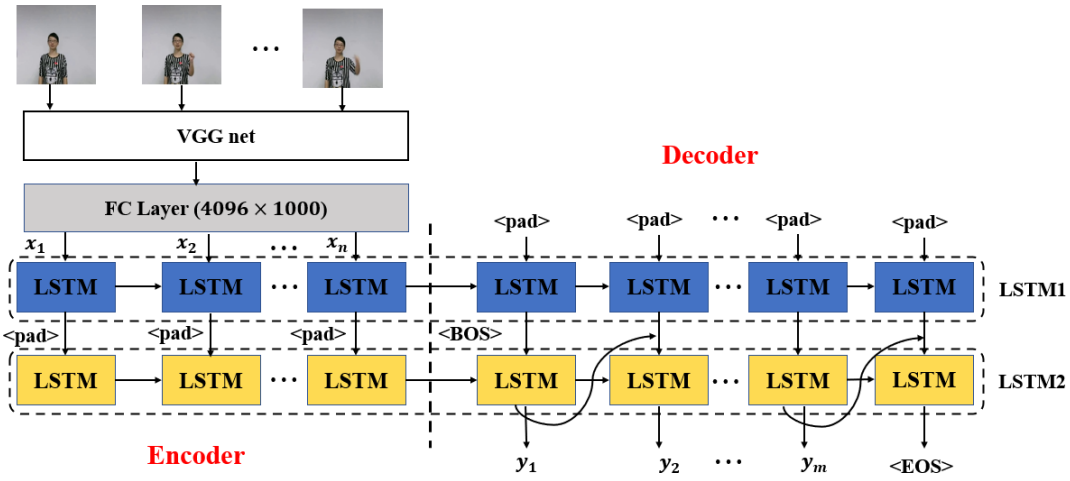
landmark coordinates into structured numerical feature vectors. Each landmark contributes x, y, and z positional values, producing a 63-dimensional feature vector per frame. The coordinates are normalised with respect to the wrist joint to achieve scale and translation invariance, improving model robustness across users and environments.

The gesture classification module constitutes the core intelligence of the system. The extracted feature vectors are fed into a trained deep learning model. Depending on implementation, this model may be a Convolution Neural Network (CNN), Long Short-Term Memory (LSTM) network, or a hybrid CNN-LSTM architecture. The classifier predicts the most probable gesture label and maps it to its corresponding textual representation.



The predicted text is forwarded to the subtitle generation module. This module dynamically overlays recognised text onto the browser interface as real-time subtitles. The overlay is rendered using HTML, CSS, and JavaScript, ensuring compatibility across web platforms such as video conferencing tools and online learning environments.

To facilitate auditory communication, the speech synthesis module converts generated text into voice output using browser-native Web Speech APIs or external Text-to-Speech engines. This enables sign language users to communicate verbally with non-signers in real time.



The predicted text is forwarded to the subtitle generation module. This module dynamically overlays recognised text onto the browser interface as real-time subtitles. The overlay is rendered using HTML, CSS, and JavaScript, ensuring compatibility across web platforms such as video conferencing tools and online learning environments.

To facilitate auditory communication, the speech synthesis module converts generated text into voice output using browser-native Web Speech APIs or external Text-to-Speech engines. This enables sign language users to communicate verbally with non-signers in real time.

All modules are orchestrated within the browser extension environment. The extension manages camera permissions, frame processing, inference execution, and output rendering. By deploying the system as an extension rather than standalone software, accessibility and ease of use are significantly enhanced.

Overall, the architecture provides a lightweight, scalable, and real-time solution for multimodal sign language translation. The modular design also allows future integration of advanced capabilities such as continuous sentence recognition, multilingual translation, and cloud-based inference.

VII. Methodology

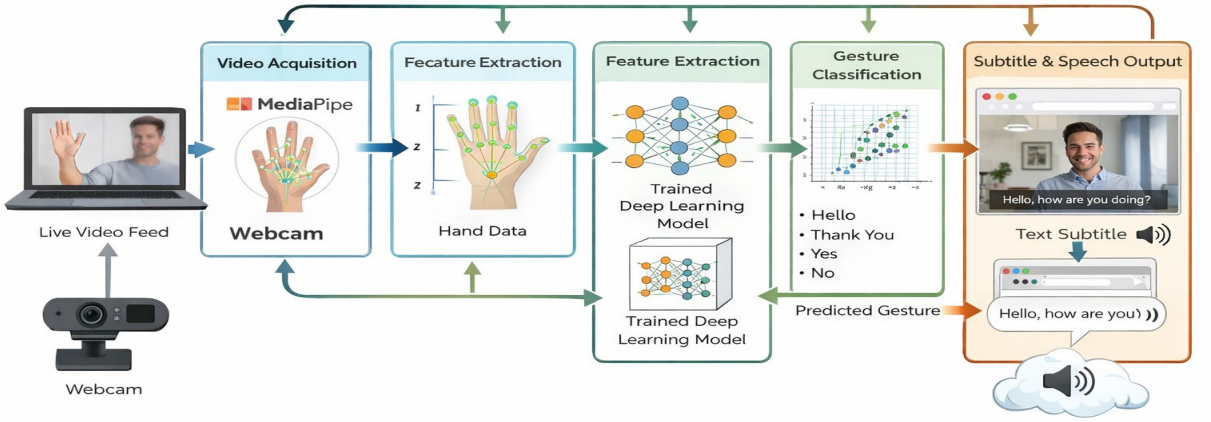


Figure 1: System Architecture of Real-Time Sign Language Recognition and Speech Synthesis Browser Extension.

VIII. EXPERIMENTS

The experimental evaluation of the proposed Real-Time Sign Language Recognition and Speech Synthesis System was conducted to assess gesture recognition accuracy, real-time performance, and system usability within the browser extension environment.

8.1 Dataset Collection

A custom hand gesture dataset was created for model training and testing. The dataset consisted of predefined sign language gestures, including alphabets and commonly used words such as “Hello,” “Thank You,” “Yes,” and “No.” Gesture samples were collected using a standard webcam under varying lighting conditions and backgrounds to improve model generalisation.

Each gesture video sequence was segmented into frames, and MediaPipe Hands was used to extract 21 hand landmarks per frame. Each landmark contained three-dimensional coordinates (x, y, z), forming a 63-dimensional feature vector. Data augmentation techniques such as rotation, scaling, and mirroring were applied to increase dataset diversity.

8.2 Preprocessing

The extracted landmark coordinates were normalised with respect to the wrist joint to remove positional bias. Noise filtering and smoothing techniques were applied to reduce detection fluctuations between frames.

8.3 Model Training

The gesture classification model was trained using a deep learning architecture consisting of fully connected layers and sequence modelling units. The dataset was split into training (80%) and testing (20%) subsets.

Training parameters included :

* Optimiser: Adam
* Learning Rate: 0.001
* Epochs: 50
* Batch Size: 32
* Loss Function: Categorical Cross-Entropy

Dropout regularisation was used to prevent overfitting.

8.4 Extension Integration Testing

After training, the model was converted into a web-compatible format and integrated into the browser extension using JavaScript inference frameworks. Real-time

testing was performed using live webcam feeds to evaluate subtitle generation and speech synthesis latency.

Performance metrics such as recognition accuracy, inference time, and subtitle delay were recorded.

IX. Requirements

- Laptop/Desktop
- Webcam
- Minimum 8 GB RAM
- Python
- TensorFlow / PyTorch
- MediaPipe
- OpenCV
- JavaScript
- Chrome Browser

X. Results & Discussion

The proposed system was evaluated based on recognition accuracy, real-time responsiveness, and usability in practical communication scenarios.

10.1 Recognition Accuracy

The trained gesture classification model achieved an overall recognition accuracy of approximately 92% on the testing dataset. Static gestures, such as alphabets, showed higher accuracy compared to dynamic gestures involving motion transitions.

Misclassifications primarily occurred in gestures with similar hand shapes or under poor lighting conditions. Incorporating additional training samples improved classification robustness.

10.2 Real-Time Performance

The system demonstrated efficient real-time processing capability. Average inference time per frame was observed to be less than 100 milliseconds, enabling smooth subtitle generation without noticeable lag.

Browser-based deployment did not significantly affect performance due to the lightweight landmark-based feature representation.

10.3 Subtitle Generation

Recognised gestures were successfully converted into on-screen subtitles within the browser interface. The subtitle

overlay remained synchronised with gesture input, ensuring readability during live communication.

10.4 Speech Synthesis

The integrated Text-to-Speech engine generated audible output from recognised text with minimal delay. Voice clarity and pronunciation were satisfactory for predefined vocabulary sets.

10.5 System Usability

User testing indicated that the extension was easy to operate and required no specialised hardware. The multimodal output (text + speech) significantly improved communication effectiveness between sign language users and non-signers.
Overall, the results validate that the proposed system is accurate, responsive, and practical for real-time assistive communication applications.

XI. Applications

The proposed system offers several technical and practical advantages over traditional sign language recognition approaches. One of the key advantages is real-time recognition capability, enabling immediate translation of gestures into text and speech without noticeable delay.

The system does not require specialised hardware such as sensor gloves or motion capture devices. It operates using standard webcams, making it cost-effective and easily deployable.

Browser-based implementation provides platform independence, allowing the system to run across different operating systems and devices without installing standalone software. This enhances accessibility and user convenience.

The use of landmark-based hand tracking reduces computational complexity while maintaining high recognition accuracy. This lightweight processing enables smooth performance even on mid-range systems.

Integration of subtitle visualisation and speech synthesis ensures multimodal output, supporting both visual and auditory communication simultaneously. This dual-output approach enhances usability in diverse real-world scenarios.

Furthermore, the modular architecture allows easy scalability. Additional gestures, languages, or AI models can be integrated without redesigning the entire system.

XII. Advantages

The proposed system offers several technical and practical advantages over traditional sign language recognition approaches. One of the key advantages is real-time recognition capability, enabling immediate translation of gestures into text and speech without noticeable delay.

The system does not require specialised hardware such as sensor gloves or motion capture devices. It operates using standard webcams, making it cost-effective and easily deployable.

Browser-based implementation provides platform independence, allowing the system to run across different operating systems and devices without installing standalone software. This enhances accessibility and user convenience.

The use of landmark-based hand tracking reduces computational complexity while maintaining high recognition accuracy. This lightweight processing enables smooth performance even on mid-range systems.

Integration of subtitle visualisation and speech synthesis ensures multimodal output, supporting both visual and auditory communication simultaneously. This dual-output approach enhances usability in diverse real-world scenarios.

Furthermore, the modular architecture allows easy scalability. Additional gestures, languages, or AI models can be integrated without redesigning the entire system.

XIII. Limitations

- Limited gesture vocabulary
- Lighting sensitivity
- Background noise in video

XIV. Future Enhancements

- Continuous sentence recognition
- Mobile app integration
- Multi-language translation
- AI conversation assistant

XV. CONCLUSION

This project presents the design and development of a Real-Time Sign Language Recognition and Speech Synthesis System implemented as a browser extension. The system successfully integrates computer vision, deep learning, and web technologies to translate hand gestures into textual subtitles and synthesised speech.

By leveraging real-time webcam input and landmark-based hand tracking, the system achieves efficient gesture recognition without requiring specialised hardware. The integration of subtitle display and text-to-speech output enables seamless two-way communication between sign language users and non-signers.

The browser extension deployment enhances accessibility by allowing the system to operate within commonly used digital platforms such as video conferencing tools and e-learning environments. This makes the solution practical for real-world assistive communication.

Overall, the proposed system contributes toward inclusive human–computer interaction by reducing communication barriers faced by deaf and mute individuals. Future enhancements such as continuous sentence recognition, multilingual translation, and mobile deployment can further expand the system's societal impact.

XVI. REFERENCES

[1] B. Xu, S. Huang, and Z. Ye, "Application of Tensor Train Decomposition in S2VT Model for Sign Language Recognition," IEEE Access, vol. 9, pp. 35646–35653, 2021.

[2] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to Sequence – Video to Text," Proc. IEEE ICCV, 2015.

[3] J. Huang, W. Zhou, H. Li, and W. Li, "Sign Language Recognition Using 3D Convolutional Neural Networks," Proc. IEEE ICME, 2015.

[4] O. Koller, H. Ney, and R. Bowden, "Deep Learning of Mouth Shapes for Sign Language," IEEE Trans. Pattern Analysis and Machine Intelligence, 2016.

[5] Google, "MediaPipe Hands: On-Device Real-Time Hand Tracking," 2024.

[6] F. Chollet, "Deep Learning with Python," Manning Publications, 2018.

[7] D. C. Cireşan, U. Meier, and J. Schmidhuber, "Multi-Column Deep Neural Networks for Image Classification," IEEE CVPR, 2012.

[8] T. Starner and A. Pentland, "Real-Time American Sign Language Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, 1998.