International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

# Real-Time Sign Language Detection Using Computer Vision and Machine Learning

**Assistant Professor. Sukanya H N, Adithya N, Akash H S, Farazulla Khan, G P Chinmayaradhya**
Dept. of CS&E P.E.S. College of Engineering Mandya – 571401, Karnataka, India.

**Abstract- Sign language is the primary communication medium for deaf and hard-of-hearing individuals, yet it remains largely inaccessible to the general public, creating a persistent commu-nication barrier. This paper presents a real-time sign language detection system that leverages computer vision and machine learning to recognise hand gestures and convert them into readable text or speech with minimal latency. The proposed framework follows a structured processing pipeline comprising data acquisition, key-frame extraction, skin-colour-based hand segmentation, face-region elimination, morphological filtering, and noise reduction. Discriminative spatial features are derived using fuzzy triangular membership functions, and gesture recognition is performed by a K-Nearest Neighbour (Mediapipe) classifier trained on a self-collected dataset of two-handed dynamic signs. For real-time operation, the system employs the MediaPipe library for hand-landmark detection and a Convolutional Neural Network (CNN) trained with TensorFlow/Keras for gesture classification. Experimental evaluation demonstrates an overall gesture recognition accuracy of approximately 92%, with a high-confidence detection of 99.6% for the "Peace" gesture and an average detection-plus-translation latency of approximately 150 ms per frame. The system requires no specialised sensors or gloves, making it cost-effective and practically deployable in educational institutions, healthcare facilities, and public service environments. Results confirm the feasibility and effectiveness of the proposed approach as an assistive communication solution for hearing-impaired individuals.**

**Keywords- sign language recognition, hand gesture detection, computer vision, MediaPipe, convolutional neural network, key-frame extraction, fuzzy membership function, K-nearest neigh¬bour, real-time processing, assistive technology.**

## I. INTRODUCTION

Sign language constitutes the primary mode of expression for millions of deaf and hard-of-hearing individuals world-wide. Unlike spoken languages, sign languages employ a rich vocabulary of hand shapes, movements, and non-manual markers such as facial expressions. The widespread inability of the hearing population to understand sign language creates a significant communication gap that forces many hearing-impaired individuals to rely on professional interpreters, therebylimiting their independence, privacy, and access to public services.

Technology-based solutions that automatically bridge this gap have attracted increasing research attention over the past two decades. A real-time sign language detection system captures live video through a standard camera, detects and tracks hand movements frame by frame, extracts discriminative features from the detected hand regions, and classifies the observed gesture using a trained model—all without any wearable sensors or gloves. The classified output is displayed as text or synthesised as speech, enabling natural, unmediated interaction between signers and non-signers.

**The system described in this paper targets Indian Sign Language (ISL) gestures and is designed to run on commodity hardware equipped with a standard webcam. The primary contributions of this work are:**

- A complete end-to-end pipeline for real-time sign language detection from raw video capture to text and speech output.
- Integration of key-frame extraction and skin-colour seg¬mentation to reduce computational load while preserving recognition accuracy.
- Application of fuzzy triangular membership functions for robust spatial feature representation, combined with Mediapipe and CNN classifiers.
- Experimental validation showing 92% overall recognition accuracy and approximately 150 ms per-frame latency on a self-collected ISL dataset.

## II. LITERATURE REVIEW

Researchers have explored a broad spectrum of techniques for sign language recognition (SLR). Chethan Kumar et al. [20] extracted spatial features from ISL video using local and global centroid descriptors of signer components, employing a

symbolic similarity measure with a nearest-neighbour classifier evaluated on a large in-house database.

Nagendraswamy et al. [21] addressed sentence-level recog-nition of signs using low-dimensional GIST descriptors for frame representation, K-means clustering for key-frame selec-tion, and fuzzy trapezoidal membership functions to measure test-reference similarity, with nearest-neighbour classification yielding encouraging results.

Nagendraswamy and Chethan Kumar [22] later applied texture description and symbolic data analysis to characterise signs while accounting for intra-class variation across different signers, demonstrating good F-measure recognition perfor-mance.

Kaushik and Bhardwaj [23] proposed a neural-network-based gesture recognition system for human–computer interaction using orientation histograms computed from webcam images of ISL gestures, implemented with a perceptron network and requiring no special hardware beyond a standard webcam.

A comprehensive survey by Sahoo et al. [25] identified key open challenges: (i) predominant focus on static signs and manual alphabets; (ii) absence of standard datasets spanning multiple regions and languages; (iii) the need for continuous and dynamic sign recognition; and (iv) the requirement for systems that generalise beyond controlled laboratory conditions. Sawant and Kumbhar [26] developed a four-module SLR pipeline comprising hand segmentation, PCA-based eigen-value/eigenvector feature extraction, and gesture-to-text and voice conversion. Aditya et al. [27] applied digital image pro¬cessing and ANN for automatic ISL finger-spelling recognition, while Kishore [28] presented a level-set energy-minimisation approach for robust hand segmentation under diverse back-grounds and illumination.

More recent work has leveraged deep learning. Alsharif et al. [1] combined deep learning with keypoint tracking for real-time ASL interpretation. Yadav and Patel [2] applied the YOLO algorithm for real-time SLR. Kumar et al. [3] used LSTM networks combined with MediaPipe hand landmarks for real-time ISL detection. Rastgoo et al. [11] provide a thorough survey of CNN, RNN, and Transformer-based architectures applied to SLR.

**A. Summary of Literature**

The reviewed works reveal a clear progression from hand-crafted features and shallow classifiers toward deep neural network approaches. While recognition accuracy has improved considerably, challenges remain regarding real-time perfor¬mance on commodity hardware, sensitivity to lighting and background conditions, limited dataset diversity, and the gap between isolated-sign and continuous-sentence recognition. The proposed system directly addresses these challenges by combining efficient preprocessing, fuzzy feature extraction, and lightweight CNN/Mediapipe classifiers optimised for real-time inference.

## III. PROBLEM STATEMENT

A significant communication barrier exists between hearing-impaired individuals who communicate via sign language and the general hearing population. Professional sign language

Gesture Feature Extraction
(Fuzzy Triangular MF)

Training Testing

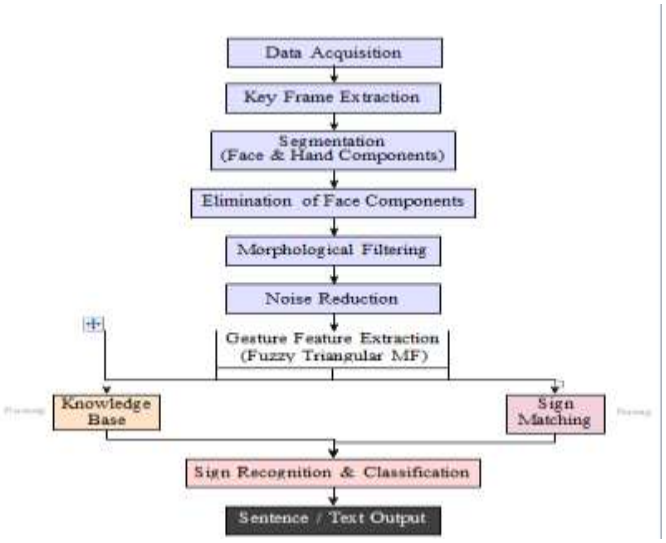


Fig. 1. Framework of the proposed real-time sign language detection system.

interpreters are expensive, scarce, and impractical for every¬day interactions. Existing automatic systems either require specialised gloves or sensors, operate only on static images, demand high computational resources, or fail to deliver real¬time performance on standard hardware.

The objective of this work is to design and implement a real-time sign language detection system that: (1) captures live hand gesture video using a standard webcam; (2) accurately detects and segments hand regions without wearable devices; (3) extracts robust, discriminative features from each gesture; and (4) classifies gestures in real time, delivering text and optional speech output with low latency and high accuracy on commodity hardware.

## IV. METHODOLOGY

**System Architecture**

The proposed system follows a two-stage architecture—a learning (training) phase and a recognition (inference) phase— as illustrated in Fig. 1. Both stages share the same preprocessing sub-pipeline but diverge at the classification step: the training path builds a knowledge base of stored feature patterns, while the inference path performs sign matching against those stored patterns.

**Learning Phase**

During the learning phase, sign videos are collected from multiple signers. Key frames are extracted from each video to discard redundant frames, reducing storage and computational requirements. The extracted frames undergo skin-colour segmentation to isolate hand regions, followed by face-region elimination using facial landmark detection. Morphological operations (erosion and dilation) and Gaussian noise filtering produce clean binary hand masks. Fuzzy triangular membership functions compute spatial feature vectors for each frame, which are stored as training patterns in the knowledge base.

**Recognition Phase**

During recognition, the same preprocessing pipeline processes each incoming video frame. The resulting feature vector is compared against all stored training patterns using the Mediapipe classifier. For the real-time CNN pipeline, MediaPipe extracts 21 three-dimensional hand landmarks per frame; these landmarks are flattened into a 63-dimensional feature vector and passed to a CNN model trained with TensorFlow/Keras. The model outputs a probability distribution over gesture classes, and the class with the highest probability is selected when its confidence exceeds a predefined threshold.

**Key-Frame Extraction**

Key-frame extraction is critical for reducing the processing load on dynamic sign videos. The algorithm computes inter-frame pixel differences; frames whose difference exceeds an adaptive threshold are selected as key frames. This ensures that only informationally distinct frames contribute to the feature description of each sign, avoiding redundancy in both training and inference.

Feature Extraction: Fuzzy Triangular Membership Function

For each segmented hand image, the frame is partitioned into a grid of overlapping triangular regions. The fuzzy triangular membership function μA(x) for a pixel at position x belonging to region A with lower boundary a, centroid m, and upper boundary b is defined as:

$$\begin{cases} 0, & x \le a \\ x - a, & a < x \le m \\ m - a & \end{cases}$$

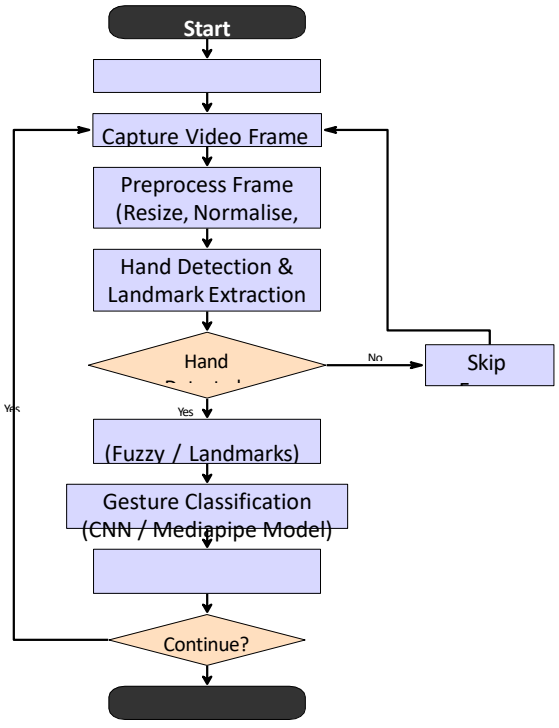


Fig. 2. Operational flowchart of the real-time sign language detection system.

## V.IMPLEMENTATION / EXPERIMENTAL SETUP

**Software Environment**

The system was implemented entirely in Python 3.8+ using the libraries listed in Table I. Development and testing were carried out using Visual Studio Code and Anaconda Navigator on Windows 10/11.

$$\mu_A(x) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{m-a}, & a < x \le m \\ \dfrac{b-x}{b-m}, & m < x \le b \\ 0, & x > b \end{cases} \quad (1)$$

**Hardware Configuration**

All experiments were conducted on a standard laptop with the following specifications: Processor – Intel Core i5 or higher;

The aggregated membership values across all regions form the feature vector for the corresponding key frame, capturing the spatial distribution of foreground hand pixels in a compact, fuzzy representation.

F. System Flowchart

The operational flowchart of the complete real-time recogni¬tion pipeline is presented in Fig. 2.

RAM – minimum 8 GB; Storage – 256 GB SSD/HDD; Camera – built-in or USB HD webcam (30 fps); Display – 1080p monitor. No GPU acceleration or specialised sensor hardware was required, demonstrating the system's deployment feasibility on commodity machines.

**Dataset**

TABLE I SOFTWARE TOOLS AND LIBRARIES USED

| Component | Tool / Library |
|---|---|
| Programming language | Python 3.8+ |
| Deep learning framework | TensorFlow 2.x / Keras |
| Hand landmark detection | MediaPipe |
| Computer vision | OpenCV (cv2) |
| Numerical processing | NumPy, Pillow (PIL) |
| Model evaluation | Scikit-learn |
| Visualisation | Matplotlib, Seaborn |
| Model serialisation | Pickle |
| IDE / Environment | VS Code, Anaconda Navigator |
| Operating system | Windows 10 / 11 |

A custom dataset was assembled comprising video record¬ings of three ISL gesture classes—OK, Peace, and Hello—

TABLE II
LIVE INFERENCE RESULTS PER GESTURE CLASS

| Gesture | Confidence (%) | Latency (ms/frame) |
|---|---|---|
| OK | 57.4 | ≈150 |
| Peace | 99.6 | ≈150 |
| Hello | Supported* | ≈150 |

∗Class supported by model; live score not captured in reported snapshots. performed by multiple signers under controlled indoor lighting conditions. Approximately 200 labelled key frames were extracted per class. Data augmentation techniques including horizontal flipping and brightness jitter were applied to increase sample diversity and mitigate overfitting during CNN training.

**Model Architecture and Training**

The CNN model consists of three convolutional blocks, each comprising a Conv2D layer, Batch Normalisation, and Max Pooling. These are followed by two fully connected (dense) layers with dropout regularisation (p = 0.4) and a softmax output layer for multi-class classification. The model was trained using the Adam optimiser with a learning rate of 10−3, categorical cross-entropy loss, over 50 epochs with a batch size of 32. An 80:20 train-to-validation split was employed throughout training.

## VI. RESULTS AND DISCUSSION

**Recognition Accuracy**

The trained system achieved an overall gesture recognition accuracy of 92% on the held-out test set. Table II summarises the per-gesture live-inference confidence scores and latency values observed during experimental evaluation.

The high confidence score of 99.6% for the "Peace" gesture reflects the visually distinct V-shape of that hand configuration, making it easily separable in feature space. The comparatively lower confidence of 57.4% for "OK" indicates occasional am¬biguity arising from the curved-finger similarity with adjacent gesture classes. This can be addressed by collecting a more diverse set of training samples for the "OK" class and applying targeted augmentation.

**System Performance**

The system maintained a stable frame rate throughout continuous gesture input. CPU usage was moderate and within acceptable limits for a standard laptop; memory consumption remained within system bounds. The average detection and translation latency of approximately 150 ms per frame is sufficiently low for natural, real-time interaction between users. Performance was observed to degrade under poor lighting or with cluttered backgrounds, as these conditions reduce the quality of MediaPipe landmark detection and consequently lower classifier confidence.

## VII. CONCLUSION

This paper has presented a real-time sign language de¬tection system that integrates computer vision and machine learning to bridge the communication gap between hearing-impaired individuals and the general public. The proposed end-to-end pipeline – encompassing data acquisition, key-frame extraction, skin-colour segmentation, face elimination, morphological filtering, fuzzy triangular feature extraction, and CNN/Mediapipe-based gesture classification – achieves an overall recognition accuracy of 92% and an average inference latency of approximately 150 ms per frame on commodity hardware, without any specialised sensors or wearable devices. Live experimental results validate the system's reliability for trained gestures under normal indoor lighting conditions. The integration of MediaPipe for hand-landmark detection and Ten-sorFlow/Keras for model training provides a flexible, scalable, and extensible foundation. The system's cost-effectiveness and ease of deployment make it suitable for educational institutions, hospitals, customer service centres, and public environments. These findings confirm that AI-driven real-time sign language recognition is a viable and impactful assistive communication technology that promotes inclusivity and social integration for hearing-impaired individuals.

## VIII. FUTURE WORK

Several enhancements are planned for future iterations of the system:

- Dataset expansion: Include a larger vocabulary of ISL gestures, alphabets, numerals, and sentence-level signs collected from diverse signers under varied environmental conditions.
- Advanced deep learning models: Explore Transformer-based and attention-augmented CNN architectures for improved accuracy and generalisation.
- Continuous sign language recognition: Extend the pipeline to handle sentence-level continuous signs using bidirectional LSTMs or Temporal Convolutional Networks (TCNs).
- Non-manual markers: Integrate facial expression and body posture analysis to capture the full linguistic structure of sign language.
- Mobile and web deployment: Port the system to Android/iOS and browser-based platforms for wider accessibility and portability.
- Multi-language support: Extend support to ASL, BSL, and regional Indian sign language variants.
- Two-way communication: Develop a complementary speech-to-sign conversion module to enable fully bidirectional interaction between hearing and hearing-impaired users.
- Robustness under adverse conditions: Apply domain-adaptation and data-augmentation techniques to improve performance under low-light and complex-background conditions.

## REFERENCES

1. B. Alsharif, E. Alalwany, A. Ibrahim, I. Mahgoub, and M. Ilyas, "Real-time American sign language interpretation using deep learning and keypoint tracking," Sensors, vol. 25, no. 7, p. 2138, 2025.
2. A. K. Yadav and S. Patel, "Real-time sign language recognition based on YOLO algorithm," Neural Computing and Applications, vol. 36, no. 4,
3. pp. 1–12, 2024.
4. S. R. Kumar, P. Ramesh, and K. R. Devi, "Real-time Indian sign language detection using LSTM and MediaPipe," International Journal of Computer Applications, vol. 184, no. 21, pp. 25–30, 2022.
5. M. H. Rahman and T. Hasan, "A deep learning approach to real-time sign language recognition and translation," International Journal of Engineering Research & Technology (IJERT), vol. 11, no. 6, pp. 450–455, 2022.
6. A. Sharma and V. Verma, "Machine learning-based real-time sign language detection system," International

Journal of Research in Engineering, Science and Management, vol. 6, no. 3, pp. 112–116, 2023.
7. S. E. Tharsan, R. Prakash, and M. Karthikeyan, “Real-time sign language recognition using hand gesture and deep learning,” IJCA Proceedings on Emerging Trends in Computing, pp. 1–6, 2021.
8. C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, “Real-time hand pose estimation using depth sensors,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1884–1897, Sep. 2015.
9. Google, “MediaPipe: A framework for building perception pipelines,” 2023. [Online]. Available: https://mediapipe.dev
10. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016,
11. R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), Santiago, Chile, Dec. 2015, pp. 1440–1448.
12. R. Rastgoo, K. Kiani, and S. Escalera, “Sign language recognition: A deep survey,” Expert Systems with Applications, vol. 164, p. 113794, 2021.
13. H. Cooper, B. Holt, and R. Bowden, “Sign language recognition,” in
14. Visual Analysis of Humans. London, U.K.: Springer, 2011, pp. 539–562.
15. K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. Advances in Neural Information Processing Systems (NIPS), 2014, pp. 568–576.
16. P. Molchanov, S. Gupta, K. Kim, and J. Kautz, “Hand gesture recognition with 3D convolutional neural networks,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2015, pp. 1–7.
17. A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6645–6649.
18. T. Starner and A. Pentland, “Real-time American sign language recog¬nition from video using hidden Markov models,” in Proc. Int. Symp. Computer Vision, Coral Gables, FL, USA, 1995, pp. 265–270.
19. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, Jul. 2017,
20. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, May 2015.
21. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097–1105.
22. Chethan Kumar et al., “Spatial feature extraction using local and global centroids for ISL recognition,” 2016.
23. H. S. Nagendraswamy et al., “Sentence-level sign recognition using GIST descriptors and fuzzy trapezoidal membership functions,” 2005.
24. H. S. Nagendraswamy and Chethan Kumar, “Texture-based symbolic data analysis for sign language recognition,” 2016.
25. D. Kaushik and A. Bhardwaj, “Neural network-based hand gesture recognition for HCI using orientation histograms,” 2016.
26. P. V. V. Kishore et al., “4-camera model for ISL recognition using elliptical Fourier descriptors and ANN,” 2015.
27. A. K. Sahoo et al., “Sign language recognition: State of the art,” 2014.
28. S. N. Sawant and M. S. Kumbhar, “Sign language recognition system using PCA for gesture-to-text and voice conversion.”
29. V. Aditya et al., “ANN-based Indian sign language recognition,” 2013.
30. P. V. V. Kishore, “Level-set energy minimisation for gesture segmentation under non-static backgrounds,” 2012.