International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue5
Sep-Oct- 2026, PP 01-6

AI-Based Indian Sign Language Recognition Using Computer Vision: A Review Of Hand Tracking, Feature Learning And Gesture Classification

Andhale Aniket Pandhari[1] Prof. S. K. Said[2] Prof. M. C. Kolse[3]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India

Abstract—Indian Sign Language (ISL) is a primary means of communication for many Deaf and hard-of-hearing people, yet few hearing people understand it, creating everyday barriers in education, healthcare and public services. This paper reviews AI-based sign language recognition with a focus on isolated ISL gesture recognition from ordinary camera input. It traces the field from early wearable and colour-glove systems to modern vision pipelines that use real-time hand and pose landmark extraction, convolutional and recurrent networks, and three-dimensional video models. Key datasets, including INCLUDE for ISL, WLASL for word-level American Sign Language and RWTH-PHOENIX for continuous recognition, are discussed alongside the difference between isolated recognition, continuous recognition and full translation. A five-layer architecture is proposed that captures video, preprocesses frames, extracts MediaPipe hand landmarks, classifies gestures with a lightweight neural network and outputs text and speech. The design is presented as a proposed system; no implementation results or accuracy claims are made. Research gaps, objectives, a comparison of feature approaches and challenges such as signer variation, lighting, two-hand occlusion and limited ISL data are analysed.

Index Terms—Indian Sign Language, sign language recognition, computer vision, MediaPipe, hand landmarks, CNN, LSTM, accessibility

I. Introduction

Sign languages are complete natural languages with their own grammar, using hand shape, movement, location, facial expression and body posture. Indian Sign Language is used across the country, but interpreters are scarce, so Deaf people often struggle to communicate with doctors, teachers, officials and shopkeepers.

Automatic sign language recognition aims to convert signs into text or speech so that hearing people can understand them without an interpreter. Early systems relied on data gloves or coloured markers; Starner *et al.* [1] demonstrated real-time American Sign Language recognition from desk-mounted and wearable cameras using hidden Markov models. Deep learning and real-time landmark tracking now make camera-only recognition practical on ordinary phones and laptops.

This paper reviews AI-based sign language recognition, focusing on isolated ISL gestures, and proposes a system design. The project is at the research and proposed-design stage, so no accuracy claims are made.

II. Background and Fundamental Concepts

2.1 Levels of the Task

Isolated sign recognition classifies one sign at a time. Continuous recognition segments and recognizes a stream of signs, and sign language translation produces grammatical spoken-language sentences. Fig. 1 organizes the approaches.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue5
Sep-Oct- 2026, PP 01-6

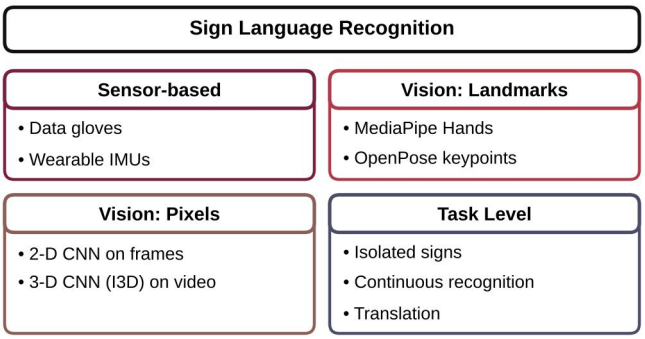


Fig. 1. Approaches to sign language recognition.

2.2 Hand and Body Landmarks

MediaPipe [2] provides real-time perception pipelines, and MediaPipe Hands [3] detects twenty-one three-dimensional landmarks per hand on mobile devices. OpenPose [4] estimates full-body and hand keypoints. Using landmarks rather than raw pixels makes classifiers less sensitive to background and skin tone.

III. Literature Survey

Starner *et al.* [1] showed early that vision-based continuous recognition was feasible with constrained vocabularies. Koller *et al.* [5] introduced the RWTH-PHOENIX-Weather corpus and statistical systems for large-vocabulary continuous recognition with multiple signers, which became a standard benchmark.

Camgoz *et al.* [6] formulated sign language translation as a neural machine translation problem from video to spoken-language text, highlighting that recognition and translation are different tasks. Li *et al.* [7] released WLASL, a large word-level dataset, and compared pose-based and appearance-based models, finding that three-dimensional convolutional models such as I3D [8] performed strongly while pose-based recurrent models were lighter.

For Indian Sign Language specifically, Sridhar *et al.* [9] released INCLUDE, a large ISL dataset of isolated words recorded by multiple signers, and showed that pose-based features with recurrent or transformer models give good accuracy at low computational cost. Rastgoo *et al.* [10] surveyed deep sign language recognition and identified data scarcity and signer independence as central challenges. Table I summarizes these works.

TABLE I Summary of Literature Survey

Ref.	Method	Key Findings	Limitations
[1]	HMM on camera-based hand tracking	Real-time ASL sentence recognition	Small vocabulary
[5]	Statistical continuous recognition	RWTH-PHOENIX benchmark	German Sign Language only
[6]	Neural sign language translation	Video-to-text translation framing	Needs large parallel corpora
[7]	WLASL dataset; pose vs. I3D	Large word-level benchmark	ASL, not ISL
[9]	INCLUDE ISL dataset; pose features	Accurate ISL word recognition	Isolated words only
[3]	MediaPipe Hands	Real-time 3-D hand landmarks on mobile	Occlusion between hands
[4]	OpenPose part affinity fields	Multi-person body and hand keypoints	Heavier than MediaPipe
[10]	Survey of deep SLR	Mapped methods and challenges	Survey only

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue5
Sep-Oct- 2026, PP 01-6

IV. Research Gap

- 1. Publicly available ISL datasets are small compared with those for other sign languages.
- 2. Many systems ignore facial expression and body posture, which carry grammatical meaning.
- 3. Models trained on few signers generalize poorly to new signers, lighting and backgrounds.
- 4. Real-time recognition on low-cost phones is rarely evaluated with Deaf users.
- 5. Most work stops at isolated words; continuous ISL translation remains largely unsolved.

4.1 Objectives

- **O1:** To review vision-based sign language recognition methods and datasets.
- **O2:** To compare landmark-based and pixel-based feature approaches.
- **O3:** To propose an isolated ISL gesture recognition architecture using MediaPipe and a lightweight classifier.
- **O4:** To define evaluation criteria suitable for a future implementation.
- **O5:** To identify challenges in moving toward continuous ISL translation.

4.2 Proposed System Architecture

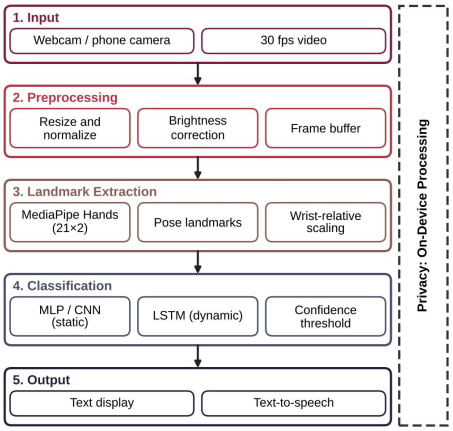


Fig. 2 shows the proposed five-layer design.

Fig. 2. Proposed architecture for isolated ISL gesture recognition.

A. Input Layer

A webcam or phone camera records the signer at around thirty frames per second.

4.4 Preprocessing Layer

Frames are resized, colour-normalized and, where needed, brightness-corrected. A short buffer stores the sequence for one sign.

4.5 Landmark Extraction Layer

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue5
Sep-Oct- 2026, PP 01-6

MediaPipe Hands extracts landmarks for both hands; optionally, pose landmarks capture arm position. Coordinates are normalized relative to the wrist and scaled to remove the effect of distance from the camera.

4.6 Classification Layer

For static signs, a small multilayer perceptron or CNN classifies one frame of landmarks. For dynamic signs, an LSTM [11] processes the sequence of landmark vectors.

E. Output Layer

The recognized word is displayed as text and spoken through text-to-speech. A confidence threshold avoids displaying uncertain predictions.

V. Methodology and Working

Fig. 3 shows the processing steps. Each landmark $p_i = (x_i, y_i, z_i)$ is normalized as

$$\hat{p_i} = (p_i - p_wrist) / max\|p_j - p_wrist\| \quad (1)$$

so that features are invariant to hand position and scale. The classifier outputs a probability for each of C signs, and the prediction is accepted when the maximum probability exceeds a threshold.

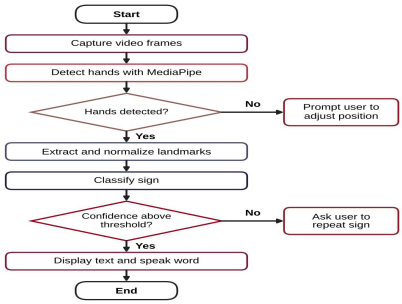


Fig. 3. Processing workflow of the proposed ISL recognizer.

A planned evaluation would use signer-independent splits, in which test signers do not appear in training, and report accuracy and per-class F1-score.

VI. Comparative Analysis

Table II compares feature approaches. Landmark-based methods are fast and robust to background; pixel-based CNNs capture hand shape detail but need more data; 3-D video models are most accurate on large datasets but heavy.

TABLE II Comparison of Feature Approaches

Approach	Speed	Data Need	Robustness
Hand landmarks + MLP	Very fast	Low	Good to background
Landmarks + LSTM	Fast	Medium	Handles motion
2-D CNN on frames	Medium	High	Lighting sensitive
3-D CNN (I3D)	Slow	Very high	High on large data

6.1 Challenges and Future Scope

Table III lists the main challenges. Future work includes larger ISL datasets collected with the Deaf community, incorporating facial expression, continuous recognition with sequence models and transformers, and two-way systems that also render text as signing avatars.

Challenge	Mitigation
Small ISL datasets	Community data collection; augmentation
Signer variation	Signer-independent training and testing
Two-hand occlusion	Pose context; temporal models
Facial grammar ignored	Add face landmarks
Continuous signing	Sequence models with segmentation

TABLE III Challenges and Mitigation

VII. Conclusion

AI-based sign language recognition can reduce communication barriers for Deaf people. The reviewed literature shows that real-time landmark tracking combined with lightweight neural classifiers makes isolated sign recognition feasible on ordinary devices, while larger video models and translation architectures point toward continuous understanding. For ISL, progress depends most on larger, diverse datasets and on evaluation with real users.

7.1 Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

References

1. T. Starner, J. Weaver, and A. Pentland, “Real-time American sign language recognition using desk and wearable computer based video,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 12, pp. 1371–1375, 1998.
2. C. Lugaresi et al., “MediaPipe: A framework for building perception pipelines,” arXiv:1906.08172, 2019.
3. F. Zhang et al., “MediaPipe Hands: On-device real-time hand tracking,” arXiv:2006.10214, 2020.
4. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime multi-person 2D pose estimation using part affinity fields,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, 2021.
5. O. Koller, J. Forster, and H. Ney, “Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers,” Comput. Vis. Image Underst., vol. 141, pp. 108–125, 2015.
6. N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, “Neural sign language translation,” in Proc. IEEE/CVF CVPR, 2018, pp. 7784–7793.
7. D. Li, C. Rodriguez, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2020, pp. 1459–1469.

8. J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the Kinetics dataset,” in Proc. IEEE CVPR, 2017, pp. 6299–6308.
9. [9] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, “INCLUDE: A large scale dataset for Indian sign language recognition,” in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 1366–1375.
10. [10] R. Rastgoo, K. Kiani, and S. Escalera, “Sign language recognition: A deep survey,” Expert Syst. Appl., vol. 164, Art. no. 113794, 2021.
11. [11] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.