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 4, November 2026
Impact Factor: 8.2

Toward Real-Time Recognition of Continuous Indian Sign Language: A Multi-Modal Approach Using RGB and Pose

Kakade Yashpal Jagannath, Prof. Y. G. Suryawanshi, Dr. H. B. Jadhav
Department of Computer Engineering
Adsul Technical Campus, Chas, Ahilyanagar, India

Abstract: *Indian Sign Language (ISL) is an important means of communication for many Deaf and hard-of-hearing people in India. Although sign-language recognition has advanced through computer vision and machine learning, recognising isolated signs is different from understanding continuous signing, where gestures flow naturally from one another and meaning depends on movement, timing, hand shape, facial expression, and body posture. This paper presents a review-oriented multi-modal framework for real-time recognition of continuous ISL using complementary Red-Green-Blue (RGB) video and human-pose information. RGB frames retain visual details such as hand appearance, clothing-independent context, and facial cues, while pose representations describe the movement and spatial relationships of the hands, arms, shoulders, and upper body. The proposed approach discusses a pipeline involving video capture, frame preparation, pose estimation, temporal feature learning, sequence decoding, and readable text output. It also considers practical challenges including signer variation, co-articulation, occlusion, lighting, camera quality, dataset diversity, annotation effort, and real-time computational limits. The central idea is that combining visual appearance with structured pose information may provide a more informative representation than relying on either modality alone. The paper outlines an evaluation strategy based on recognition quality, sequence-level error, latency, and robustness across signers. It emphasises that an accessible ISL recognition system should be developed with Deaf-community participation and should support communication rather than claim to replace interpreters or human interaction*

Keywords: Indian Sign Language, Continuous Sign Language Recognition, RGB Video, Human Pose Estimation, Multi-Modal Learning, Temporal Modelling, Real-Time Recognition

I. INTRODUCTION

Sign languages are complete natural languages with their own linguistic structure, grammar, and regional variation. Indian Sign Language is used across diverse communities, and its expression involves more than a sequence of hand gestures. Hand configuration, movement direction, location, facial expression, and body posture can all contribute to meaning. For this reason, a system that recognises only a single hand pose may miss important information needed to interpret a natural signed sentence.

Recent developments in computer vision and deep learning have created opportunities to build systems that interpret sign-language video. However, many recognition tasks focus on isolated signs, in which each clip contains one clearly separated gesture. Continuous signing is more demanding because there may be no pause between signs. The system must identify meaningful temporal boundaries and interpret connected movements without relying on artificial breaks.

This paper explores a multi-modal direction for continuous ISL recognition by combining RGB video with pose-based features. RGB input preserves rich visual information, while pose data offers a compact description of body and hand

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 4, November 2026
Impact Factor: 8.2

movement. Their combination may help a model learn both what a signer looks like in motion and how the relevant body parts move over time.

II. PROBLEM STATEMENT

Continuous ISL recognition remains challenging because signs are produced in a connected stream rather than as isolated, neatly separated actions. Natural signing includes transitions between signs, variations in speed and style, differences in camera position, and possible occlusion of the hands or face. RGB-only systems may be affected by background and appearance variation, while pose-only systems can lose useful visual details or become unreliable when keypoints are missed. A practical recognition framework therefore needs to combine complementary information while maintaining accurate sequence interpretation and low processing delay.

III. OBJECTIVES

- To examine the requirements and challenges of recognising continuous Indian Sign Language from video.
- To describe a multi-modal framework that combines RGB appearance features with pose-based movement information.
- To outline suitable stages for video preprocessing, pose extraction, temporal modelling, and sequence decoding.
- To discuss how signer variation, co-articulation, occlusion, and environmental conditions affect recognition.
- To identify evaluation measures for recognition quality, real-time performance, and robustness.

IV. BACKGROUND AND RELATED CONCEPTS

Sign-language recognition commonly involves visual feature extraction followed by a model that learns the relationship between observed movements and linguistic units. In isolated-sign recognition, a clip is often labelled with one sign. In continuous sign-language recognition, the input is a longer sequence and the system must infer a sequence of labels from uninterrupted signing. This creates a temporal alignment problem: the model must learn where one sign ends and another begins even when the signer does not pause.

RGB video provides colour and texture information and can preserve details that are difficult to express through keypoints alone. Pose estimation converts selected body landmarks into coordinates or structured features. Depending on the pose estimator and camera view, these landmarks may include shoulders, elbows, wrists, hands, and facial points. Pose sequences can reduce some background variation and make movement patterns easier to model, but inaccurate landmark detection can introduce noise.

V. PROPOSED MULTI-MODAL FRAMEWORK

A. Video acquisition and preprocessing

The system begins with a camera stream or recorded video of a signer. Frames are sampled at a consistent rate, resized to the model's input dimensions, and normalised. Preprocessing should preserve the signer's hands and upper body while avoiding unnecessary loss of spatial detail. The capture setup should be tested under different lighting conditions, camera distances, and backgrounds.

B. RGB feature extraction

An appearance encoder processes RGB frames to learn visual patterns associated with hand shape, hand-object relationships where relevant, facial expression, and overall signing context. A convolutional or vision-transformer-based encoder may be considered, depending on the available dataset and computational budget. The feature extractor should be selected and validated for the target hardware rather than assumed to be real-time by design.

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 4, November 2026
Impact Factor: 8.2

C. Pose estimation and motion representation

A pose-estimation stage detects body and hand landmarks for each frame. Landmark coordinates can be normalised relative to a body reference point or frame dimensions, and temporal differences can describe movement direction and speed. Missing or low-confidence landmarks should be handled explicitly, because dropped keypoints can affect later sequence interpretation.

D. Multi-modal feature fusion

RGB and pose features can be combined at an early, intermediate, or late stage. Early fusion joins the input representations before deeper learning, while intermediate fusion combines learned features and late fusion merges predictions from separate branches. Intermediate fusion is a practical option to investigate because each modality can first learn its own useful representation. The fusion design should be tested against RGB-only and pose-only baselines to determine whether combining the modalities provides a measurable benefit.

E. Temporal modelling and sequence decoding

Continuous signing requires a model that captures information across time. Recurrent networks, temporal convolutional networks, and attention-based architectures are possible choices for learning sequence dependencies. The model output may be decoded using a sequence-learning objective such as Connectionist Temporal Classification (CTC), or through another alignment strategy suited to the available annotations. A language model or constrained decoder may help produce readable output, but it should not silently change the recognised signs or conceal uncertainty.

F. Real-time output

For interactive use, the pipeline should process incoming frames in a streaming manner and display partial or final predictions with an indication of uncertainty where appropriate. Buffer size, frame sampling, model complexity, and hardware acceleration influence end-to-end delay. A useful system should balance recognition quality with responsiveness rather than optimise accuracy alone.

VI. KEY CHALLENGES

- Co-articulation: adjacent signs influence one another, making sign boundaries difficult to identify.
- Signer diversity: hand size, signing speed, preferred signing space, and regional or personal variation affect the observed data.
- Occlusion: one hand may hide the other, or hands may be partially outside the camera frame.
- Non-manual features: facial expression and head or body movement can contribute to meaning and should not be ignored.
- Dataset limitations: continuous, accurately annotated ISL video is difficult and costly to collect at sufficient scale.
- Environmental variation: lighting, background, camera angle, and image quality can reduce feature reliability.
- Real-time constraints: pose estimation and temporal modelling must operate within practical latency and hardware limits.

VII. EVALUATION STRATEGY

A reliable evaluation should compare the proposed multi-modal model with RGB-only and pose-only baselines using the same data splits. Data should be separated by signer where possible so that performance on unseen signers can be assessed. Word Error Rate (WER) is a common sequence-level measure for continuous recognition and counts substitutions, deletions, and insertions relative to a reference sequence. Sign-level accuracy or other task-appropriate measures may also be reported. Real-time evaluation should include end-to-end latency, throughput in frames per second, memory use, and performance under different lighting and occlusion conditions. Results should be reported on a clearly described dataset and should not be inferred from model design alone.

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 4, November 2026
Impact Factor: 8.2

VIII. POTENTIAL APPLICATIONS

A continuous ISL recognition system could support communication interfaces, educational demonstrations, sign-language learning tools, public information kiosks, and accessibility research. It may also help explore ways to make selected digital services easier to navigate for sign-language users. These applications require careful user testing: recognition errors can change meaning, and automated text output should not be treated as a guaranteed substitute for a qualified interpreter in important medical, legal, or emergency situations.

IX. ETHICAL AND ACCESSIBILITY CONSIDERATIONS

ISL recognition should be developed with Deaf signers and community organisations as active partners, not merely as sources of training data. Participants should understand how video and pose data will be collected, stored, shared, and used. Consent, privacy, fair representation, and the ability to withdraw data are important considerations. System designers should avoid describing sign language as a universal gesture code; language variation and community preferences must be respected. The intended role of the technology should be clearly communicated, including its limitations and the possibility of incorrect predictions.

X. FUTURE SCOPE

Future work can investigate larger and more diverse continuous-ISL datasets, improved hand and facial landmark tracking, efficient temporal architectures, and adaptive fusion strategies that respond to missing or unreliable modalities. Research may also explore signer-independent learning, regional variation, uncertainty estimation, and deployment on affordable edge devices. Evaluation with Deaf users should examine not only model scores but also usability, trust, communication success, and the situations in which the tool is or is not appropriate.

XI. CONCLUSION

Real-time recognition of continuous Indian Sign Language is a complex sequence-understanding task that requires attention to hand movement, body posture, facial information, and the natural flow between signs. RGB video and pose features provide complementary views of signing: one preserves rich visual context, while the other describes structured movement. A multi-modal framework that combines these signals offers a promising direction to investigate, but its effectiveness must be demonstrated through careful comparison, signer-independent testing, and real-time measurement. Above all, the development process should be guided by the needs and expertise of the Deaf community. The goal is to create a dependable accessibility aid that supports communication while respecting the linguistic richness of Indian Sign Language and the continuing importance of human interaction.

ACKNOWLEDGMENT

The authors express their sincere gratitude to their academic guides and institution for their encouragement and support. They also acknowledge the researchers and members of the sign-language community whose work and experience contribute to progress in accessible communication technologies.

REFERENCES

[1] S. C. W. Ong and S. Ranganath, “Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 873–891, 2005.
[2] O. Koller, “Quantitative Survey of the State of the Art in Sign Language Recognition,” arXiv:2008.09918, 2020.
[3] 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 Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2020.
[4] A. M. Sincan and H. Y. Keles, “AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods,” IEEE Access, vol. 8, pp. 181340–181355, 2020.

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 4, November 2026
Impact Factor: 8.2

[5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 369–376.
[6] World Federation of the Deaf, "Sign Languages," World Federation of the Deaf. Available: https://wfdeaf.org/.
[7] Indian Sign Language Research and Training Centre (ISLRTC), Department of Empowerment of Persons with Disabilities, Government of India. Available: https://islrtc.nic.in/.