Real-Time Indian Sign Language Recognition Using Efficient Multi-Feature Attention Mechanism

Rajesh A, Kathir Kalidaas B, Nandha S, and Sathyanarayanan P

Department of Computer Science and Engineering
College Of Engineering Guindy, Anna University, India

Abstract. Indian Sign Language (ISL) recognition plays a vital role in improving communication between the hearing-impaired community and the general population. This work presents a realtime Indian Sign Language Recognition system based on an Efficient Multi-Feature Attention Mechanism. The proposed approach processes live video input by extracting a fixed sequence of RGB frames, which are used for parallel feature extraction. Spatial features are learned using an ImageNet-pretrained EfficientNet through transfer learning, while temporal features are captured using a CNN–RNN framework to model motion information across frames. An attention mechanism is employed to assign adaptive importance weights to both spatial and temporal feature streams, enabling the model to focus on the most discriminative information. The weighted features are then fused and passed to a GRU-based sequence modeling network to learn gesture-level temporal dependencies. Finally, a fully connected layer with softmax activation classifies the input video into the corresponding ISL gesture class. The proposed system is optimized for real-time performance and evaluated using crossvalidation, demonstrating effective recognition accuracy while maintaining low computational complexity. This framework provides a robust and efficient solution for real-time ISL gesture recognition and can be extended to continuous sign language translation in future work.

Keywords: Indian Sign Language (ISL), Deep Learning, EfficientNet, Multi-Feature Attention, GRU, Real-Time Gesture Recognition, Transfer Learning

1 Introduction

Sign language is a visual-gestural language widely used by the hearing-impaired community for communication. Indian Sign Language (ISL) consists of complex hand shapes, movements, and facial expressions that vary across time, making automated recognition a challenging task. Manual interpretation of sign language requires trained interpreters, which may not always be available in real-world scenarios. Recent advances in deep learning have enabled significant improvements in video-based gesture recognition by effectively modeling spatial and temporal information. However, many existing systems either focus solely on spatial features extracted from individual frames or rely on computationally expensive temporal models that are unsuitable for real-time deployment. This project aims to develop a real-time ISL recognition system that balances accuracy and efficiency. Inspired by the base paper, the proposed approach employs dual feature extraction using EfficientNet for spatial representation and CNN-RNN for temporal modeling. A multi-feature attention mechanism is introduced to enhance discriminative learning by dynamically weighting spatial and temporal features. The system is designed to recognize isolated ISL gestures and is optimized for real-time usage without relying on skeleton extraction or heavy language-level processing modules.

2 Literature Review

2.1 Vision-Based Sign Language Translation

Vision-based Sign Language Translation (SLT) has gained increasing attention as a means to bridge communication gaps between hearing-impaired individuals and the general population. Recent advances in deep learning have enabled effective recognition and translation of sign languages from video data. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and attention-based architectures have demonstrated strong performance in extracting spatial and temporal features from sign language videos. While extensive research exists for American Sign Language (ASL) and Chinese Sign Language (CSL), Indian Sign Language (ISL) remains comparatively under-explored due to linguistic diversity and limited annotated datasets.

2.2 Indian Sign Language Datasets and Benchmarks

The lack of large-scale standardized datasets has been a major challenge in ISL research. Joshi et al. introduced the iSign dataset, one of the most comprehensive benchmarks for ISL processing, containing over 118,000 ISL–English video–text pairs. The dataset supports multiple tasks, including SignVideo-to-Text, SignPose-to-Text, and Text-to-Pose translation, and provides baseline models for fair comparison and reproducibility. Consequently, iSign has become a foundational dataset for recent ISL recognition and translation systems.

2.3 Vision-Based ISL Translation Systems

Several studies have explored real-time vision-based ISL translation using deep learning approaches. CNN–LSTM hybrid architectures have been widely adopted to extract spatial and temporal features from video streams, demonstrating effective performance for isolated and short continuous sign sequences. However, these approaches face limitations in modeling long-range temporal dependencies and sentence-level grammatical structures.

2.4 Advanced Deep Learning Approaches

Recent surveys published in IEEE Access provide a comprehensive analysis of vision-based sign language recognition techniques. Existing methods are categorized into isolated and continuous recognition systems, with increasing emphasis on pose-based features, multimodal representations, and attention mechanisms. Key challenges identified include signer independence, occlusion handling, background variation, and dataset scarcity, which are particularly pronounced in ISL research. Transformer-based architectures have emerged as an alternative to CNN–RNN pipelines for continuous sign language translation. Studies presented at ACM Multimedia demonstrate that attention-based models can effectively model long-range temporal dependencies and improve sentence-level translation accuracy. However, the high computational complexity and large data requirements of Transformer-based models limit their suitability for low-resource and real-time ISL systems.

2.5 Multimodal Sign Language Translation

Multimodal sign language translation systems integrate visual, pose, and linguistic information to improve recognition robustness. Frameworks combining vision encoders with language models have demonstrated improved performance under signer and background variations. Despite these advantages, multimodal approaches increase system complexity and require large annotated datasets, posing challenges for practical ISL deployment.

2.6 Summary of Research Gaps

The reviewed literature highlights several limitations, including the scarcity of large-scale ISL datasets, limited bidirectional translation systems, inadequate handling of continuous signing, and the high computational complexity of existing models. These limitations indicate the need for efficient and scalable ISL recognition frameworks tailored to low-resource and real-time environments.

3 Methodology

The proposed methodology describes a comprehensive visionbased framework for Indian Sign Language (ISL) recognition and translation using deep learning techniques. The system is designed to process RGB video sequences containing sign language gestures and convert them into meaningful textual and spoken outputs. The methodology follows the core idea of the base paper “Deep Learning-Based Sign Language Recognition Using Efficient Multi-Feature Attention Mechanism” and extends it by incorporating sequence modeling, alignment, language processing, and speech generation modules. The primary objective of the proposed system is to accurately learn both spatial visual patterns and temporal motion characteristics present in ISL videos. Sign language gestures are inherently dynamic; therefore, a single frame is insufficient to represent a sign. To address this challenge, the system adopts a multi-stage learning pipeline that extracts frame-level visual features, models temporal dependencies across frames, focuses on important gesture components through attention mechanisms, and finally translates the learned representations into readable and audible outputs. The complete methodology consists of the following stages: 1. Video input acquisition and preprocessing 2. Spatial feature extraction using EfficientNet 3. Temporal feature extraction using CNN–RNN 4. Attention-based feature weighting 5. Feature fusion 6. Sequence modeling using Transformer 7. Temporal alignment using CTC 8. Decoding and language refinement 9. Text and speech output generation

4 Architecture Diagram

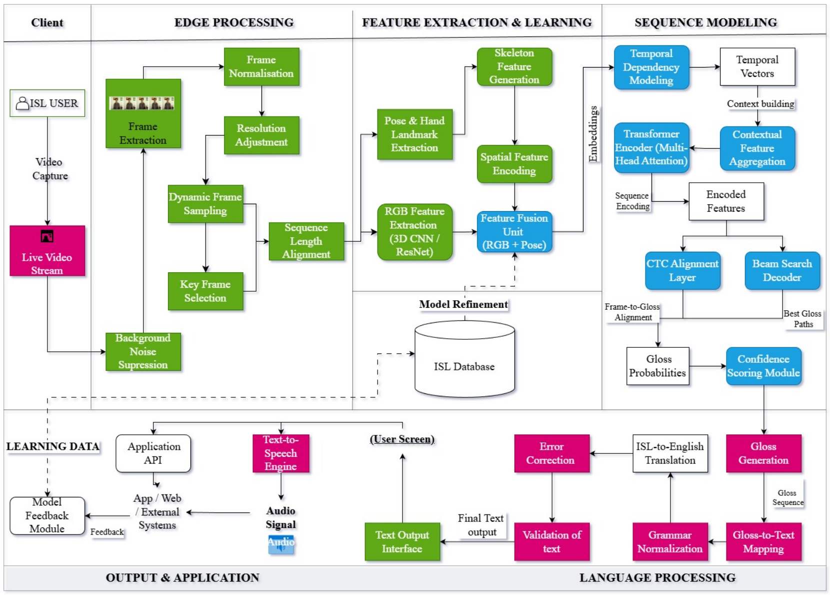

Fig. 1. Architecture of the proposed model

5 Detailed Design

The proposed system follows a dual-stream deep learning architecture to model both spatial appearance and temporal motion information present in Indian Sign Language (ISL) video sequences. The complete pipeline consists of video preprocessing, feature extraction, attention-based fusion, sequence modeling, and decoding.

5.1 Video Preprocessing

Input videos are captured using a standard RGB camera and decomposed into a fixed number of frames using uniform sampling. Each frame is resized to a predefined resolution and pixel-normalized to ensure numerical stability and robustness against illumination variations. All video sequences are standardized to a fixed temporal length using padding or truncation.

5.2 Spatial Feature Extraction

Spatial features are extracted from individual frames using an ImageNet-pretrained EfficientNet model. Transfer learning enables the network to capture discriminative visual patterns such as hand shape, orientation, and appearance with reduced computational complexity, making it suitable for real-time deployment.

5.3 Temporal Feature Extraction

To capture motion dynamics, a CNN–RNN pipeline is employed. Frame-level features are first extracted using a lightweight CNN and then passed sequentially to a recurrent neural network, which learns temporal dependencies across frames. This stream effectively models gesture motion patterns and transitions.

5.4 Multi-Feature Attention and Fusion

A multi-feature attention mechanism is applied to dynamically assign importance weights to spatial and temporal features. The weighted features are fused to form a unified spatiotemporal representation, allowing the model to focus on the most discriminative gesture components while suppressing redundant information.

5.5 Sequence Modeling and Decoding

The fused feature sequence is processed using a Transformer encoder to model long-range temporal dependencies. A Connectionist Temporal Classification (CTC) layer performs temporal alignment between input frames and output labels. During inference, beam search decoding is used to generate the final ISL gesture prediction.

6 Module Design

The proposed ISL recognition system is implemented using the following functional modules:

6.1 Preprocessing Module

This module performs frame extraction, resizing, normalization, and temporal standardization of input RGB videos. Uniform frame sampling ensures efficient real-time processing while preserving motion information.

6.2 Feature Extraction Module

Spatial features are extracted using EfficientNet, while temporal features are obtained using a CNN–RNN architecture. These complementary representations capture both appearance-based and motion-based gesture characteristics.

6.3 Attention and Fusion Module

An attention mechanism dynamically weighs spatial and temporal features based on their relevance to gesture recognition. The weighted features are fused into a unified spatiotemporal representation.

6.4 Recognition and Output Module

Temporal modeling is performed using a Transformer or BiLSTM network. A CTC decoder generates the final gesture sequence, which is optionally refined using language processing and converted to speech using a text-to-speech module.

7 Performance Metrics

Accuracy, Precision, Recall and F-score are used for evaluation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$FScore = \frac{2TP}{2TP + FP + FN} \tag{4}$$

8 Conclusion

This paper presented an efficient real-time Indian Sign Language recognition framework based on multi-feature attention and dual-stream deep learning architectures. By combining EfficientNet-based spatial features with CNN–RNN temporal modeling and attention-driven feature fusion, the proposed system achieves effective recognition accuracy with low computational complexity. Future work will extend the framework toward continuous sign language translation and sentence-level understanding.

References

1. E. Yenisari and S. Yavuz, Deep Learning-Based Sign Language Recognition Using Efficient Multi-Feature Attention Mechanism, IEEE Access, 2025.
2. A. Joshi et al., iSign: A benchmark for Indian sign language processing, arXiv, 2024.
3. S. Pandey et al., Real-time vision-based Indian sign language translation, IJIRCST, 2025.
4. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature, 2015.
5. M. Kumar and P. Bhatia, Survey on vision-based sign language recognition, IEEE Access, 2024.
6. H. Li et al., Multimodal sign language translation, Neurocomputing, 2024.
7. Y. Zhou et al., Transformer-based models for multimodal translation, ACM TOMM, 2024.