Pose-Based Temporal Convolutional Networks for Isolated Indian Sign Language Word Recognition

Tatigunta Bhavi Teja Reddy , Vidhya Kamakshi
Department of Computer Science & Engineering
National Institute of Technology Calicut
Kozhikode - 673601, Kerala, India
tatigunta_m240602cs@nitc.ac.in, vidhyakamakshi@nitc.ac.in

Abstract

This paper presents a lightweight and efficient baseline for isolated Indian Sign Language (ISL) word recognition developed for the WSLP-AACL-2025 Shared Task.We propose a two-stage framework combining skeletal landmark extraction via MediaPipe Holistic with a Temporal Convolutional Network (TCN) for temporal sequence classification. The system processes pose-based input sequences instead of raw video, significantly reducing computation and memory costs. Trained on the WSLP-AACL-2025 dataset containing 4,398 isolated sign videos across 4,361 word classes, our model achieves 54% top-1 and 78% top-5 accuracy.

1 Introduction

Sign Language serves as a primary means of communication for millions of deaf and hard-of-hearing individuals across the world. However, the lack of mutual intelligibility between signers and non-signers continues to pose substantial barriers in education, healthcare, employment, and daily communication. While human interpreters provide an effective bridge, their limited availability and high cost restrict widespread accessibility. Automated Sign Language Recognition (SLR) systems thus hold significant potential to enhance social inclusion by enabling real-time, scalable translation between sign and spoken languages.

Recent progress in computer vision and deep learning has revitalized research in automatic sign language understanding. Unlike spoken languages, which rely on one-dimensional acoustic signals, sign languages are inherently multimodal—integrating hand configurations, body posture, facial expressions, and spatial-temporal dynamics to convey meaning. This multidimensional structure makes SLR a particularly challenging problem in visual sequence modeling. Conventional frame-based models often struggle to capture the fine-grained temporal dependencies and spatial variations inherent to signing. Consequently, developing models that effectively learn temporal patterns, remain robust to inter-signer variability, and generalize across diverse signing conditions is a key research objective.

Despite these advances, Sign Language Recognition remains a challenging task due to its inherently temporal and highly variable nature. Each sign involves dynamic motion sequences that differ across signers in speed, articulation, and regional style, while transitions between signs often blur semantic boundaries. Moreover, annotated datasets for Indian Sign Language (ISL) are limited in size and diversity, constraining the training of data-intensive deep models. These challenges call for lightweight architectures capable of capturing long-range temporal dependencies using compact representations.

Motivated by these challenges this study, we address the problem of isolated Indian Sign Language (ISL) word recognition as part of the WSLP-AACL-2025 Shared Task. The goal is to design and train an efficient recognition pipeline that performs reliably despite the limited availability of labeled samples per class. Our work explores a lightweight, pose-based approach using MediaPipe Holistic for landmark extraction and Temporal Convolutional Networks (TCNs) for temporal modeling, aiming to balance recognition accuracy, computational efficiency, and real-time deployability on assistive devices.

2 Related Work

Research in Sign Language Recognition (SLR) has evolved from handcrafted visual features to deep neural architectures capable of modeling complex spatio-temporal dynamics. Early vision-based approaches (Tamura and Kawasaki, 1988) relied on geometric and motion descriptors, often

Proceedings of the Workshop on Sign Language Processing (WSLP), pages 51–54
December 20-24, 2025 ©2025 Association for Computational Linguistics

combined with Hidden Markov Models (HMMs) (Starner and Pentland, 1995; Starner et al., 1998) for real-time American Sign Language (ASL) recognition. Subsequent works enhanced temporal modeling through parallel HMMs to mitigate co-articulation effects (Vogler and Metaxas, 1999), and through hybrid CNN–HMM architectures for continuous signing (Koller et al., 2015).

With the rise of deep learning, pose-based representations have gained prominence for their robustness and computational efficiency. MediaPipe Holistic (Lugaresi et al., 2019) enabled real-time extraction of body and hand landmarks, facilitating lightweight recognition pipelines. Leveraging such pose data, transformer-based models have demonstrated strong performance for isolated sign recognition (Alyami et al., 2024), while recurrent GRU-based architectures have been successfully applied to Indian Sign Language (ISL) recognition (Subramanian et al., 2022). More recent studies explore Temporal Convolutional Networks (TCNs) with dilated causal convolutions for efficient temporal reasoning (Xu et al., 2023), and correlation networks enhanced with spatial-temporal attention for continuous SLR (Hu et al., 2023).

Reviewing the literature reflects a paradigm shift toward pose-based and temporally aware architectures that balance recognition accuracy with real-time deployability, forming the foundation for the approach adopted in this work.

3 Methodology

The proposed Indian Sign Language (ISL) word recognition system adopts a two-stage framework integrating pose-based feature extraction with temporal modeling. In the first stage, skeletal landmarks are extracted from each video frame using MediaPipe Holistic (Lugaresi et al., 2019). This pipeline provides 33 pose landmarks and 21 landmarks per hand, resulting in 75 keypoints per frame, each with (x, y, z) coordinates, yielding a 225-dimensional feature vector. This representation retains essential kinematic information while substantially reducing input dimensionality compared to raw RGB frames.

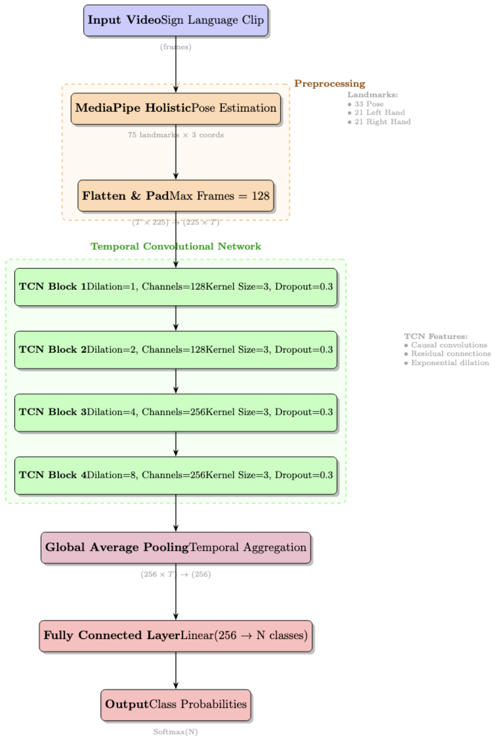
Input VideoSign Language Clip
(frames)
Preprocessing
Landmarks:
• 33 Pose
• 21 Left Hand
• 21 Right Hand
MediaPipe HolisticPose Estimation
75 landmarks × 3 coords
Flatten & PadMax Frames = 128
(T × 225) → (225 × T)
Temporal Convolutional Network
TCN Block 1Dilation=1, Channels=128Kernel Size=3, Dropout=0.3
TCN Block 2Dilation=2, Channels=128Kernel Size=3, Dropout=0.3
TCN Block 3Dilation=4, Channels=256Kernel Size=3, Dropout=0.3
TCN Block 4Dilation=8, Channels=256Kernel Size=3, Dropout=0.3
TCN Features:
• Causal convolutions
• Residual connections
• Exponential dilation
Global Average PoolingTemporal Aggregation
(256 × T) → (256)
Fully Connected LayerLinear(256 → N classes)
OutputClass Probabilities
Softmax(N)

Figure 1: Overview of the proposed Pose-based TCN pipeline for ISL Word Recognition.

In the second stage, the extracted pose sequences are processed by a Temporal Convolutional Network (TCN) designed to capture temporal dependencies across sign sequences. Unlike recurrent networks, TCNs leverage 1D convolutions along the temporal axis, enabling full parallelization during training and inference. The network maps normalized and padded pose sequences directly to sign word labels in an end-to-end manner, achieving a balance between recognition accuracy and computational efficiency suitable for real-time applications.

The overall model design emphasizes three principles: (i) robustness to intra-class and inter-signer variations, (ii) effective temporal modeling through dilated causal convolutions, and (iii) lightweight representation enabling deployment on resource-limited assistive devices. The model architecture is illustrated in Figure 1. Each temporal block consists of two dilated Conv1D layers followed by causal chomp, ReLU activation, and dropout, with residual connections to stabilize gradient flow. The network input is a tensor of shape $(T \times 225)$, where $T = 128$ is the temporal length. Four temporal blocks with exponentially increasing dilation rates are stacked, followed by global average pooling along the temporal dimension and a fully connected layer for classification across N sign classes.

3.1 Temporal Convolutional Network Architecture

Temporal Convolutional Networks (TCNs) serve as a parallelizable alternative to recurrent architectures for sequence modeling. A TCN operates using 1D convolutions over time, where causal convolutions ensure that each timestep prediction depends only on the current and past frames. Dilated convolutions enlarge the receptive field exponentially with minimal parameter overhead, allowing efficient long-range temporal modeling. Residual connections are incorporated to mitigate vanishing gradient problems and facilitate deeper network training. This architecture preserves temporal causality while providing high throughput suitable for real-time recognition.

3.2 Dataset and Preprocessing Pipeline

We employ the WSLP-AACL-2025 Shared Task Word Recognition dataset (Lab, 2025), consisting of 4,398 short video clips of isolated sign language words performed by a single signer in controlled and semi-controlled settings. The dataset spans 4,361 unique word classes, forming an extreme few-shot learning scenario: 80% of classes contain two or fewer samples, the median sample count per class is one, and the maximum is five. Videos range from 2–5 seconds at 30 FPS, with resolutions between 320p and 1080p. Following integrity checks, the dataset is divided into 3,517 training and 879 validation samples using a fixed random seed for reproducibility.

Pose extraction is performed using MediaPipe Holistic configured in non-static mode with detection and tracking confidences set to 0.3. For each frame, 75 landmarks with (x, y, z) coordinates are extracted and normalized relative to the frame dimensions and depth. Missing landmarks are replaced with zeros. Frames are resized to a width of 320 pixels, every third frame is skipped to reduce redundancy, and each sequence is truncated or padded to 128 frames. The resulting pose tensors of shape $(128 \times 75 \times 3)$ are stored in NPZ format for training.

3.3 Implementation Details

The TCN comprises four temporal blocks with hidden channel sizes [128, 128, 256, 256], kernel size 3, and dilation rates [1, 2, 4, 8]. A dropout rate of 0.3 is applied within each block. Training uses the AdamW optimizer with learning rate 10^{-3}, weight decay 0.01, $\beta = (0.9, 0.999)$, and $\epsilon = 10^{-8}$. The learning rate is adaptively reduced using a plateau scheduler (factor 0.5, patience 3, minimum learning rate 10^{-6}). Early stopping based on validation accuracy prevents overfitting.

The objective function is the categorical cross-entropy loss, defined as:

$$\mathcal{L} = -\sum_{k=1}^{N} y_k \cdot \log(\hat{y_k}) \tag{1}$$

where y_k denotes the one-hot encoded ground truth and $\hat{y_k}$ represents the predicted probability corresponding to the k^{th} sign.

This configuration achieves a balance between temporal modeling capacity, generalization on few-shot classes, and computational efficiency suitable for shared-task benchmarking and real-time deployment.

4 Results

The proposed pose-based Temporal Convolutional Network achieved a top-1 classification accuracy of 54% on the validation set. Corresponding precision, recall, and F1-scores were observed to lie consistently within the 52–54% range, indicating balanced performance across most sign classes despite the highly imbalanced few-shot nature of the dataset.

A Top-5 accuracy of approximately 78% further demonstrates that the correct class frequently appeared among the top predicted candidates, highlighting the model's capacity to capture semantically relevant temporal patterns even when the top prediction was incorrect.

An examination of prediction confidence distributions revealed that correctly classified samples exhibited moderate confidence levels, whereas lower confidence was typically associated with visually or temporally ambiguous gestures, signer variation, or partial landmark occlusions. These findings suggest that while the model effectively learns generalizable temporal representations from pose trajectories, performance remains constrained by limited per-class data and subtle intra-class motion variations.

5 Summary

This work presents an efficient baseline for isolated Indian Sign Language (ISL) recognition by integrating MediaPipe-based pose estimation with

Temporal Convolutional Networks (TCNs). The proposed system achieved 54% validation accuracy on a challenging few-shot multi-class dataset, highlighting the effectiveness of skeleton-based representations in capturing essential gesture dynamics. The TCN architecture, leveraging dilated causal convolutions, successfully modeled long-range temporal dependencies while retaining computational efficiency through its fully parallelizable design. Preprocessing strategies such as frame skipping and resolution reduction reduced computational cost by nearly 70% with minimal performance degradation, demonstrating the approach's suitability for real-time deployment. Representing each frame through 75 key landmarks achieved an input size reduction of approximately 4,000× relative to raw video frames, significantly enhancing inference speed without compromising discriminative power. The proposed pipeline establishes a compact and practical foundation for the recognition of ISL in low-resource and assistive environments. Future work can extend this baseline by exploring richer temporal attention mechanisms, synthetic data augmentation, and integration of facial and contextual cues to further enhance recognition accuracy and system robustness.

References

Sultan Alyami, Hamzah Luqman, and Mohammad Hammoudeh. 2024. Isolated arabic sign language recognition using a transformer-based model and landmark keypoints. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–19.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2529–2539.

Oscar Koller, Hermann Ney, and Richard Bowden. 2015. Deep sign: Hybrid cnn-hmm for continuous sign language recognition. *British Machine Vision Conference (BMVC)*.

Exploration Lab. 2025. Wslp-aacl-2025 shared task word recognition dataset. https://huggingface.co/datasets/Exploration-Lab/WSLP-AACL-2025/tree/main/Shared_task_WR.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, and 1 others. 2019. Mediapipe: A framework for building perception pipelines. In *arXiv preprint arXiv:1906.08172*.

Thad Starner and Alex Pentland. 1995. Visual recognition of american sign language using hidden markov models. In *International Workshop on Automatic Face and Gesture Recognition*, pages 189–194. IEEE.

Thad Starner, Joshua Weaver, and Alex Pentland. 1998. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375.

Balamurali Subramanian, Bekhzod Olimov, Sushil M Naik, and 1 others. 2022. An integrated mediapipe-optimized gru model for indian sign language recognition. *Scientific Reports*, 12(1):11964.

Shingo Tamura and Satoru Kawasaki. 1988. Recognition of sign language motion images. *Pattern Recognition*, 21(4):343–353.

Christian Vogler and Dimitris Metaxas. 1999. Parallel hidden markov models for american sign language recognition. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 116–122. IEEE.

Xiujuan Xu, Jian Wang, and Lei Zhang. 2023. Isolated word sign language recognition based on improved skresnet-tcn network. *Journal of Sensors*, 2023:9503961.