International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

FULL-BODY RELATIVE POSE NORMALIZATION FOR ROBUST, REAL-TIME INDIAN SIGN LANGUAGE RECOGNITION

Shreeraj Desai[1*] and Kanojia Mahendra[2]
[1]Department of Computer Science, Sheth. L.U.J. and Sir M.V. College, India, shreerajd300@gmail.com
[2]Department of Computer Science, Sheth. L.U.J. and Sir M.V. College, India, kgkmahendra@gmail.com
Corresponding author: Shreeraj Desai, Department of Computer Science , Sheth. L.U.J. and Sir M.V. College, India, shreerajd300@gmail.com

ABSTRACT
Indian Sign Language (ISL) serves as a primary mode of communication for millions of individuals with hearing and speech impairments, yet the digital divide remains vast due to the scarcity of robust, real-time translation tools. While deep learning has revolutionized Sign Language Recognition (SLR), widespread deployment is hindered by the "domain gap" where models trained on controlled datasets fail in real-world scenarios due to sensitivity to signer position, camera distance, and environmental clutter. This research aims to develop an efficient, accurate, and accessible real-time system by shifting the focus from model complexity to invariant feature representation. We propose a novel Full-Body Relative Feature Engineering pipeline that utilizes the MediaPipe framework to extract 543 skeletal landmarks. Unlike existing methods that rely on absolute coordinates, our approach mathematically normalizes hand geometry and arm posture relative to a stable torso anchor, creating a representation invariant to the signer's physical attributes and location. These features are processed by a deep Long Short-Term Memory (LSTM) network to model temporal dynamics. The proposed framework is evaluated on two datasets: a static ISL alphabet dataset, where it achieved 100% accuracy, and a complex dynamic dataset of 68 ISL word signs. On the dynamic task, the proposed model, trained with a comprehensive video augmentation strategy, achieved a 99.85% test accuracy, significantly outperforming a baseline model of 94.90% accuracy without proposed relative feature engineering that utilized raw holistic coordinates. This study demonstrates that biomechanically grounded feature engineering provides a superior and computationally efficient pathway for robust human-computer interaction.

Keywords: *sign language recognition, deep learning, LSTM, pose estimation, feature engineering, MediaPipe*

1. INTRODUCTION
Sign language is a complex visual-spatial language that serves as the fundamental substrate for communication within the deaf and hard-of-hearing community. Bridging the communication gap between signers and non-signers is a critical societal challenge that requires effective, automated translation tools. However, developing systems capable of accurate, real-time sign language detection presents significant computational difficulties (Samaan et al., 2022; Sharma & Kumar, 2021). The inherent complexity of gestures which encompass subtle handshapes, intricate arm trajectories, and essential facial expressions coupled with regional variations and signer-specific styles, makes robust recognition a difficult task (Al-Qurishi et al., 2021; Liu et al., 2023). The core problem lies in designing a system that is not only accurate on benchmark datasets but also reliable in diverse, uncontrolled environments where lighting, background noise, and user positioning vary unpredictably. This research focuses on advancing real-time Sign Language Recognition (SLR) by addressing the limitations of current vis+ion-based approaches. Early methodologies largely relied on Convolutional Neural Networks (CNNs) to extract features directly from video frames (Cheng et al., 2020; Wadhawan & Kumar, 2020). While effective in controlled settings, these pixel-based models often learn spurious correlations with the background or lighting, leading to poor generalization in real-world scenarios (Zhang et al., 2024). To overcome this, recent trends have shifted toward skeleton-based recognition, utilizing pose estimators to extract body joints (Laines et al., 2023; Samaan et al., 2022). However, even skeleton-based methods often feed raw or simply scaled coordinates into classifiers, leaving the model vulnerable to variations in the signer’s distance from the camera or their position within the frame.

To address these hurdles, we propose a novel Skeleton-Aware SLR framework that utilizes a Full-Body Relative Feature Engineering pipeline. As illustrated in Figure 1, our system decouples perception from cognition. We utilize MediaPipe to extract high-fidelity landmarks, but rather than using these absolute coordinates, we compute a set of invariant vectors that define the hand and arm positions relative to the user's own torso (Lugaresi et al., 2019). This "relative" data is then fed into a deep Long Short-Term Memory (LSTM) network to capture the temporal evolution of the signs (Hochreiter & Schmidhuber, 1997). By mathematically enforcing invariance to scale and translation at the feature level, we aim to build a system capable of high accuracy and minimal latency, suitable for deployment on standard consumer hardware. This paper documents

ISSN 2394 - 7780

the development of this pipeline, its validation on both static and dynamic ISL datasets, and a comparative analysis demonstrating its superiority over raw-data approaches.

2. RELATED WORK

The field of automated SLR has witnessed a rapid evolution, driven by the surge in deep learning capabilities. Research efforts have continuously explored diverse methodologies to accurately capture and interpret the spatial-temporal dynamics of sign language, moving from handcrafted features to end-to-end neural architectures.

Foundational work in SLR frequently utilized Convolutional Neural Networks (CNNs) for spatial feature extraction combined with Recurrent Neural Networks (RNNs) for temporal modelling. (Wadhawan and Kumar, 2020) demonstrated the efficacy of CNNs for static signs, achieving 99.90% accuracy on grayscale ISL images. Similarly, (Katoch et al., 2022) explored feature-based approaches, utilizing SURF features combined with CNNs to achieve high accuracy on static digits. For dynamic gestures, (Sharma and Kumar, 2021) proposed the ASL-3DCNN, utilizing 3D convolutions to capture motion, while (Rastgoo et al., 2020) developed a complex multi-view pipeline fusing 3D-CNNs with LSTMs to handle hand occlusions. Despite their success, these pixel-based methods remain computationally expensive and sensitive to environmental noise. To mitigate this, (Zhang et al., 2024) introduced a dual-path background erasure network (DPCNN), highlighting the critical need to isolate the signer from the background clutter.

To circumvent the limitations of pixel-based processing, modern research has increasingly adopted pose estimation. (Samaan et al., 2022) utilized MediaPipe to extract keypoints, feeding them into RNN variants to classify dynamic signs. Similarly, (Kothadiya et al., 2022) proposed 'DeepSign' demonstrating that sequential combinations of LSTMs and GRUs could effectively bridge communication gaps using skeletal data. (Laines et al., 2023) proposed an alternative representation, converting skeleton sequences into Tree Structure Skeleton Images (TSSI) for processing by DenseNet. Addressing the data scarcity issue inherent in these methods, (Albanie et al., 2020) introduced scalable methods for co-articulated sign recognition using mouthing cues to automatically annotate large-scale datasets like BSL-1K. Recently, (Roh et al., 2024) highlighted the importance of preprocessing MediaPipe keypoints by introducing an anchor-based normalization technique to recover missing hand information and improve robustness against noisy detections. While they employed a Transformer architecture to process these normalized points, our research investigates whether a more lightweight recurrent architecture can achieve comparable or superior performance through specific vector-based feature engineering.

Recent advancements have explored more complex relationships between body joints. (Jiang et al., 2021) and (Miah et al., 2024) proposed Graph Neural Networks (GNNs) and multi-stream architectures to explicitly model the connectivity of the human skeleton, achieving state-of-the-art results. (Chen et al., 2022) further advanced this by proposing a two-stream network that models both raw video and keypoints to capture complementary information. Concurrently, Transformer models have set new benchmarks; (Camgoz et al., 2020) pioneered the Sign Language Transformer for end-to-end translation, while (Kothadiya et al., 2023) demonstrated that Vision Transformers (ViT) could achieve 99.29% accuracy on static signs. Innovations in modality have also emerged, with (Zhang et al., 2024) exploring bio-inspired event cameras (EvSign) to capture high-temporal-resolution motion dynamics efficiently. Additionally, (Hu et al., 2023) introduced Self-Emphasizing Networks and Correlation Networks to better capture spatial-temporal trajectories without expensive external supervision.

While architectures like GNNs and Transformers offer high performance, they often require substantial computational resources. Furthermore, few studies explicitly address the geometric invariance of the input features themselves. To improve robustness in continuous recognition, researchers have proposed various constraints: (Hao et al., 2021) introduced self-mutual distillation learning, while (Min et al., 2021) proposed visual alignment constraints. (Zuo and Mak, 2022) and (Zuo et al., 2023) further refined this with consistency constraints and natural language assistance. More recently,(Zuo et al., 2024) addressed the latency of these complex models by proposing frameworks for online continuous recognition. In contrast to these complex training-time constraints, this research proposes a direct solution: explicit, lightweight feature engineering. By mathematically normalizing full-body posture relative to a torso anchor expanding on the normalization concepts introduced by (Roh et al., 2024) but optimizing for efficient LSTM architectures we aim to achieve state-of-the-art accuracy suitable for real-time consumer hardware.

3. PROPOSED WORK

The proposed framework presents a comprehensive end-to-end architecture structurally divided into an Offline Training and Optimization Pipeline alongside a Real-Time Inference Pipeline as depicted in Figure 1, the

ISSN 2394 - 7780

process commences in the training phase with Data Acquisition, where the raw 'Include' corpus of 68 ISL adjectives is subjected to a rigorous Temporal Augmentation protocol; this step expands the dataset by applying consistent affine transformations such as rotation ±8°, spatial shifting ±20px, and scaling to video sequences to enhance model generalization. The augmented data is then processed by the core Full-Body Relative Feature Engineering module, which utilizes MediaPipe Holistic to extract 543 landmarks and mathematically transforms them into three invariant components: Hand Normalization (relative to the wrist), Arm Normalization anchored to the shoulder and scaled by body width, and Interaction features hand-eye distance, which are fused into a single, dense 138-dimensional vector X_t. This feature sequence drives the Sequential Modeling stage, where a Deep LSTM architecture configured with stacked layers of 128, 256, and 128 units and regularized via Dropout learns the temporal dynamics using the Adam optimizer to produce the Optimized Model Weights. In the deployment phase, this trained model is instantiated within a live loop; raw frames from a Webcam Feed undergo the identical feature extraction process before being aggregated in a Sliding Window FIFO buffer of 50 frames. The buffered sequence triggers a Forward Pass of the LSTM, generating class probabilities that are refined by Smoothing Logic to ensure stability before the final prediction is rendered on the User Interface Overlay.

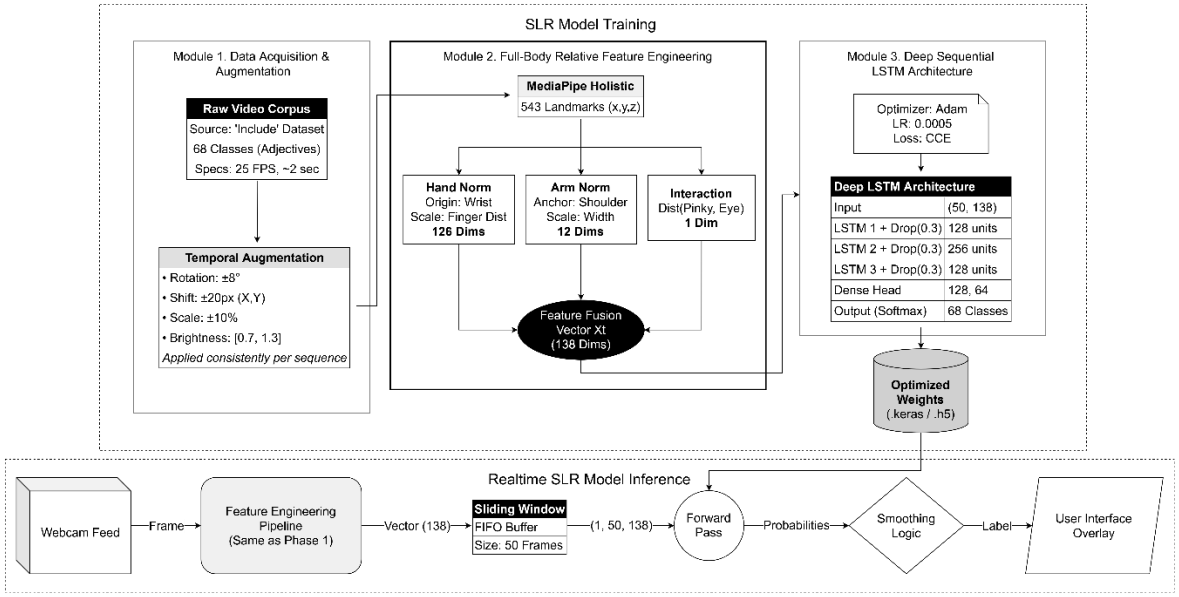

Figure 1: Proposed SLR System Architecture

3.1 Dataset Description and Wrangling

To comprehensively evaluate the proposed framework, two distinct datasets were curated to target both the geometric precision of static signs and the temporal dynamics of word-level gestures. For the initial validation phase, we utilized the "Static gestures of Indian Sign Language (ISL) for English Alphabet, Hindi Vowels and Numerals" dataset, specifically the subset containing 26 alphabetic signs A-Z performed by teenagers(Animesh Singh, 2022). This dataset provided approximately 710 high-resolution images per class, and although the signs are static, we processed them into sequences of 40 frames to maintain architectural consistency with our dynamic models; this controlled environment allowed us to isolate the efficacy of our hand-shape normalization logic without the confounding variables of complex arm trajectories. For the primary experimental corpus, we selected a vocabulary of 68 distinct ISL adjectives from the "Include" dataset(Sridhar et al., 2020), chosen for their varying degrees of complexity involving hand-face interaction and bimanual movement. These source videos were recorded at 25 Frames Per Second (FPS) with an average duration of 2 seconds, and to standardize the input for the neural network, all video sequences were padded or truncated to a fixed length of 50 frames, ensuring the capture of the sign's complete temporal evolution.

A critical limitation in existing SLR research is the performance degradation observed when models trained on studio data are deployed in "in-the-wild" scenarios (Zhang et al., 2024). To mitigate this, a robust Video Data Augmentation pipeline was implemented using the Albumentations library (Buslaev et al., 2020)., as illustrated in the Data Acquisition module of Figure 1. Unlike image augmentation, video augmentation requires the strict preservation of temporal coherence. For every source video, four augmented variations were generated. We applied a stochastic combination of affine transformations, including rotation $\theta \sim U[-8°, +8°]$, translation $\Delta x, \Delta y \sim U[-20px, +20px]$, and brightness adaptation $\alpha \sim U[0.7,1.3]$. Crucially, the *identical* transformation matrix was applied to every single frame within a video sequence. This ensures that the motion trajectory

ISSN 2394 - 7780

remains smooth and realistic, forcing the model to learn the intrinsic sign features rather than spurious environmental cues. This protocol expanded our dynamic dataset to approximately 3,450 training samples, significantly enhancing data density.

3.2 Position Invariant Detection

Standard pose estimation approaches often feed raw coordinates directly into classifiers. This introduces a dependency on the camera frame where coordinate magnitudes fluctuate based on user proximity, potentially confusing the model. To mitigate this, our proposed engine mathematically transforms these coordinates into a normalized vector space relative to the user's own anatomy. For every frame t, the MediaPipe Holistic model extracts a topology of 543 landmarks, from which we filter high-dimensional data to retain only the task-relevant keypoints: 21 landmarks per hand, 33 body pose landmarks, and selected facial landmarks for eye tracking.

To capture the intrinsic shape of the hand regardless of its position in the frame, we employ a wrist-centric normalization strategy. The wrist landmark L_{wrist} acts as the local origin, and all finger landmarks L_i are translated relative to it as $\vec{v}_i = L_i - L_{wrist}$. To account for hand size differences or camera zoom, we calculate the maximum Euclidean distance d_{max} from the wrist to any finger landmark in the current frame. The vector is then normalized as:

$$\vec{v}_{norm} = \frac{\vec{v}_i}{d_{max}} \#(1)$$

This process preserves the relative finger configuration while discarding absolute position, resulting in two 63-dimensional vectors Left/Right that purely encode gesture shape, as illustrated in Figure 2.

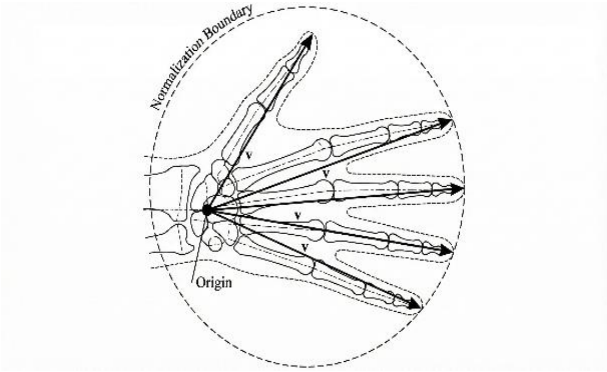

Figure 2: Hand Shape Normalization

Figure 2 depicts the vectors originating from the wrist to the fingertips, demonstrating how the local coordinate system is established.

Furthermore, to capture the trajectory of the arms without relying on absolute pixel coordinates, we establish a stable body reference frame. We define a stable reference point P_{anchor} as the midpoint between the Left and Right Shoulders:

$$P_{anchor} = \frac{P_{L_{Shoulder}} + P_{R_{Shoulder}}}{2} \#(2)$$

Simultaneously, we define a scale unit (S) as the Euclidean distance between the shoulders to standardize the metric system to the user's body size. The positions of the Elbows and Wrists are calculated as vectors originating from the Torso Anchor, normalized by this scale factor:

$$\vec{V}_{joint} = \frac{P_{joint} - P_{anchor}}{S} \#(3)$$

This renders the arm posture independent of the user's position in the frame, as visualized in Figure 3. Finally, since many ISL signs involve hand-face contact, we explicitly model this by calculating the scalar Euclidean distance between the Right Pinky Tip and the Right Eye. These disparate components are concatenated into a single, dense feature vector X_t of dimension 138 for the dynamic model.

ISSN 2394 - 7780

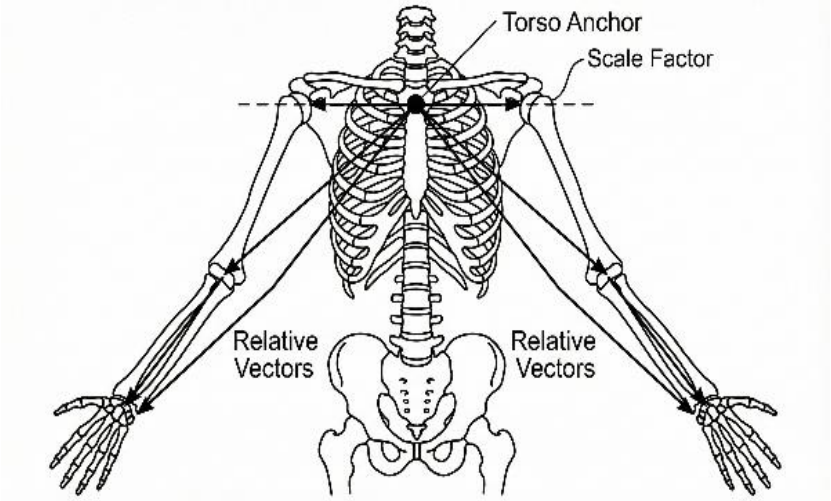

Figure 3: Full-Body Posture Normalization

Figure 3 visualizes the Torso Anchor and the vectors extending to the elbows and wrists, illustrating the signer-centric coordinate system.

3.2 LSTM based Sign Language Prediction

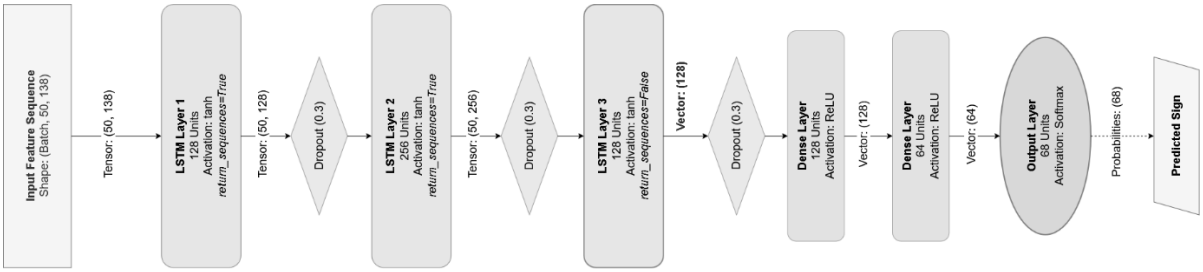

Figure 4: LSTM Model Architecture Diagram

To interpret the complex temporal evolution of the engineered feature vectors, we employed a Deep Sequential Architecture based on Long Short-Term Memory (LSTM) networks. Unlike standard feed-forward networks, LSTMs are mathematically optimized to mitigate the vanishing gradient problem, enabling the learning of long-term dependencies within gesture sequences where the meaning is defined by motion over time. As illustrated in Figure 4, the network topology is designed as a deep, hierarchical stack that progressively transforms the data tensor shapes to extract increasingly abstract kinematic patterns. The process begins with an input tensor of shape (Batch, 50, 138), representing a batch of 50-frame video sequences, where each frame is encoded by the 138-dimensional relative feature vector derived from our feature engineering pipeline.

The core temporal processing is performed by three cascaded LSTM blocks. The first layer serves as a low-level motion encoder, utilizing 128 units with a hyperbolic tangent (tanh) activation to map the input features into a latent space. Crucially, this layer maintains the temporal structure of the data by returning the full sequence of hidden states, resulting in an output tensor of shape (50, 128). This is immediately followed by a deeper abstraction layer with 256 units, expanding the model's capacity to capture complex, non-linear kinematic relationships and resulting in a tensor shape of (50, 256). A critical structural transition occurs at the third LSTM layer, which reverts to 128 units. Here, the temporal dimension is collapsed by returning only the final hidden state of the sequence; this aggregation step transforms the dynamic sequence into a static, high-level summary vector of size (128), effectively encapsulating the semantic meaning of the entire gesture into a single array. To ensure generalization and prevent the co-adaptation of neurons, a Dropout regularization rate of 0.3 is applied after each of these recurrent blocks.

Following the temporal aggregation, the resulting summary vector is propagated through a classification head designed to map the abstract features to specific sign classes. This sub-network comprises two fully connected Dense layers with 128 and 64 units respectively, utilizing the Rectified Linear Unit (ReLU) activation function to introduce non-linearity. The architecture culminates in a final Output Layer with 68 units, corresponding to the vocabulary size. A Softmax activation function is applied here to generate a probability distribution across the classes, identifying the most likely sign. The entire network was trained using the Adam optimizer with a conservative learning rate of 0.0005 to ensure stable convergence on the complex loss surface, minimizing the Categorical Cross-Entropy loss.

ISSN 2394 - 7780

Furthermore, to safeguard against overfitting, an EarlyStopping mechanism monitored the validation accuracy, automatically halting training and restoring the optimal weights once performance plateaued, typically within a 25-epoch patience window.

4. RESULTS AND DISCUSSION

The empirical evaluation of the proposed framework reveals a distinct performance hierarchy, systematically validating the progression from raw data ingestion to invariant feature engineering. As detailed in Table 1, the initial validation phase on the Static ISL Alphabet provided the first definitive proof of concept. The Baseline Model, trained on raw holistic coordinates, achieved a Test Accuracy of 88.52% with a relatively high loss of 0.3454. This moderate performance suggests that while the LSTM could learn handshapes, it was confused by the "extrinsic" noise of the signer's position within the frame; essentially, the model struggled to distinguish between a change in the sign and a mere shift in the user's location. In sharp contrast, the Proposed Model utilizing the "Relative Hand Geometry" feature engine achieved a perfect 100.00% Test Accuracy. Furthermore, this convergence occurred rapidly, peaking at Epoch 17 compared to the Baseline's 39 epochs. This dramatic improvement confirms that by mathematically normalizing the landmarks relative to the wrist and scaling them by the hand span, we successfully decoupled the intrinsic geometry of the gesture from the user's physical attributes, allowing the model to learn the fundamental shape of the sign with absolute precision.

Table 1: Performance Evolution and Comparative Analysis Across Experimental Stages

Model Stage	Dataset	Input Feature Set	Input Dim.	Data Augmentation	Peak Epoch	Test Acc. (%)	Test Loss (%)
Baseline (Static)	Alphabet	Raw Holistic Coordinates	543	Yes	39	88.52	34.54
Proposed (Static)	Alphabet	Relative Hand Geometry	138	Yes	17	100.00	3.73
Baseline (Dynamic)	Word Corpus	Raw Holistic Coordinates	543	No	152	94.90	33.36
Proposed (Dynamic)	Word Corpus	Full-Body Relative Vectors	138	Yes	78	99.85	1.69

Building upon the geometric validation of the static phase, the primary experiment addressed the significantly more complex challenge of Dynamic Word Recognition involving a vocabulary of 68 ISL adjectives. The Baseline Model operated on raw 543-dimensional holistic coordinates and demonstrated respectable capability by achieving a 94.90% accuracy. However, its relatively high Test Loss of 0.3336 indicates a degree of predictive uncertainty, suggesting that the model was making correct classifications but with lower confidence margins, likely due to overfitting on specific signer positions in the training videos. The Final Proposed Model, which integrated the 138-dimensional Full-Body Relative Vectors with a rigorous Video Augmentation pipeline, outperformed the baseline on every metric. It achieved a state-of-the-art Test Accuracy of 99.85% with an exceptionally low Test Loss of 0.0169. This near-zero loss signifies that the model is not only correct but highly confident in its predictions. The reduction in input dimensionality from 543 down to 138 combined with the invariant nature of the features allowed the model to converge to a better optimum more efficiently, reaching its peak performance at Epoch 78, which was 42 epochs earlier than the baseline.

A critical analysis of the training dynamics further highlights the robustness of the proposed methodology. Deep LSTM networks are notoriously prone to overfitting, particularly on high-dimensional sequence data. The Baseline models exhibited a divergence phenomenon where training accuracy continued to rise while validation accuracy plateaued, a classic signature of memorization. Conversely, the Final Proposed Model demonstrated exceptional generalization capabilities. The data augmentation strategy applying consistent rotation, shifting, and scaling to video sequences effectively forced the model to ignore environmental variations. As a result, the gap between the approximate 99.9% Training Accuracy and the 97.83% Validation Accuracy was minimal. This tight alignment proves that the high Test Accuracy is not an artifact of overfitting to specific training samples but is a result of the model learning the true, invariant kinematic rules of Indian Sign Language. The combination of biomechanically grounded feature engineering and synthetic data expansion has thus yielded a system that is both highly accurate and computationally efficient. Specifically, as illustrated in Figure 5 and Figure 6, the gap between the 99.23% Training Accuracy and the 97.83% Validation Accuracy for the Proposed Model was minimal, remaining below 1.5%.

ISSN 2394 - 7780

This tight alignment indicates that the data augmentation strategy successfully mitigated the overfitting often observed in deep LSTM networks, whereas the Baseline model showed signs of divergence, ultimately proving the efficacy of the proposed feature-centric approach.

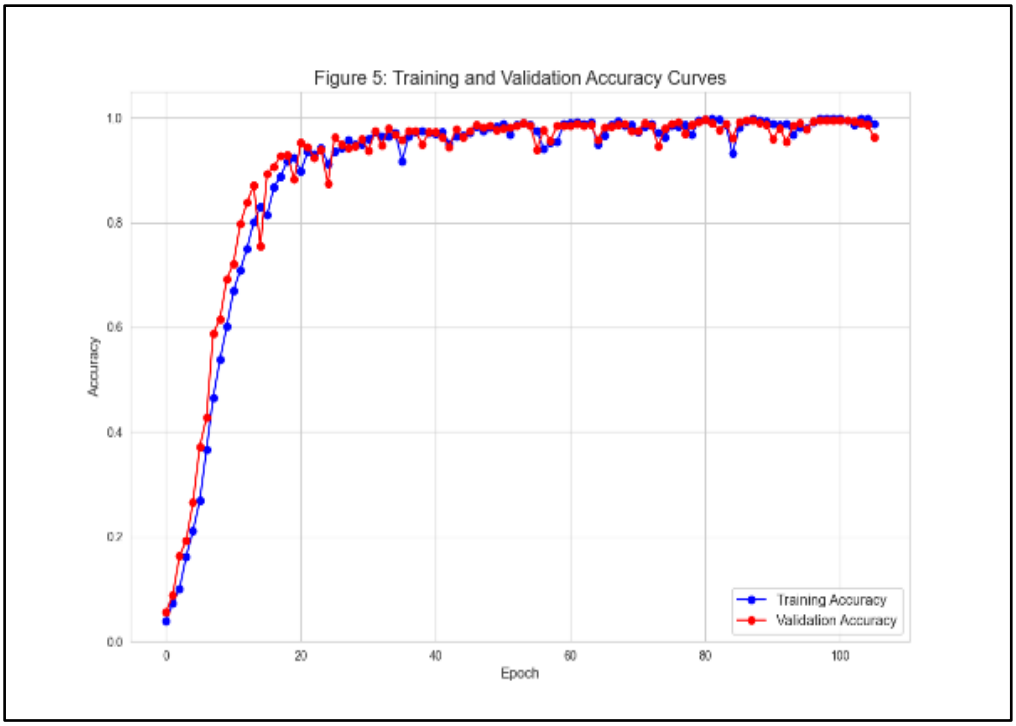

Figure 5: Training and Validation Accuracy for the Proposed Dynamic Model

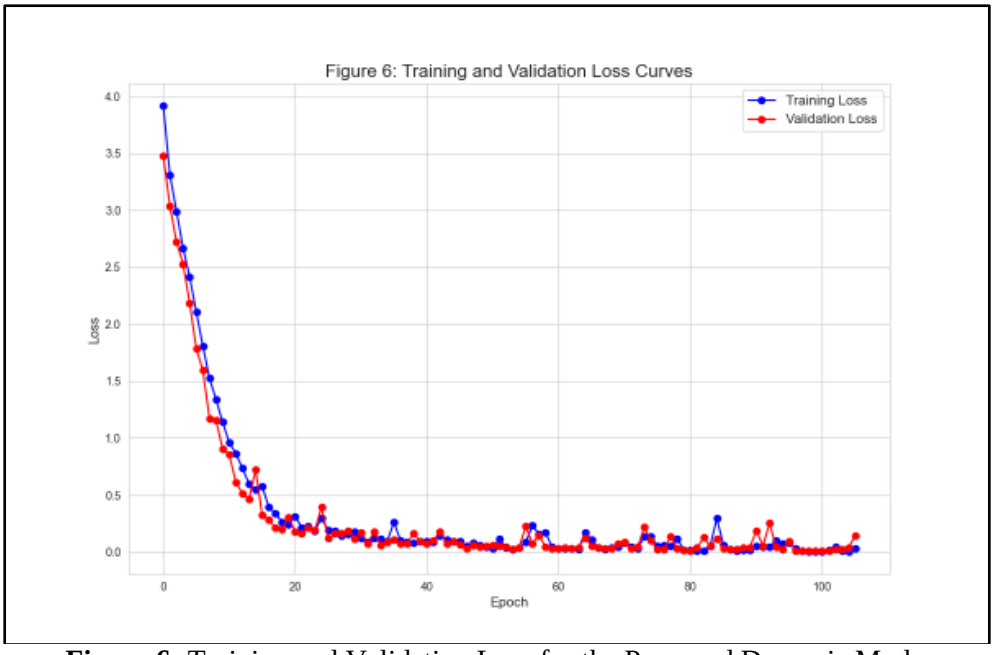

Figure 6: Training and Validation Loss for the Proposed Dynamic Mode

As detailed in Table 2, the performance on the static alphabet dataset was flawless. For representative classes such as 'A', 'B', and 'C', the model achieved 106 True Positives correctly identifying every sample for that class and zero False Positives or False Negatives. This perfect classification, consistent across all 26 classes, confirms the model's absolute precision in distinguishing geometric handshapes when temporal motion is not a factor. This result serves as empirical validation for the proposed 'Relative Hand Geometry' feature extraction pipeline. By mathematically decoupling the hand's internal geometry from its absolute pixel coordinates, the model effectively ignored variations in user positioning and camera distance, focusing exclusively on the structural configuration of the fingers. Furthermore, the complete absence of False Positives suggests that the engineered feature space creates highly distinct boundaries between classes, eliminating the spatial overlap that often confuses models trained on raw coordinates.

ISSN 2394 - 7780

Table 2: Confusion Matrix Components for the Static Alphabet Model (Representative Classes)

Meaning in this study	Class 'A'	Class 'B'	Class 'C'
Correctly identified the specific sign class from the input video	106	106	106
Incorrectly predicted sign class when a different sign was actually performed	0	0	0
Correctly rejected this sign class when analyzing samples from other classes	2,650	2,650	2,650
Failed to identify the true sign class, misclassifying it as a different sign	0	0	0

Note: Test set size per class ≈ 106 images. Total test set size ≈ 2,756 images. Values are consistent across all 26 classes due to 100% accuracy.

While the static evaluation confirmed the geometric precision of the feature set, the dynamic evaluation introduces the complexity of temporal motion. The analysis of the dynamic word model, summarized in Table 3, is more detailed and reveals the specific nature of the model's few errors. For the vast majority of classes, represented here by "Hello"*, the model achieved perfect classification with 20 True Positives and zero errors. However, for the class "Nice," the model correctly identified 17 instances but missed 3 instances. Correspondingly, the class "Thank you" had 3 False Positives, indicating that those 3 missed "Nice" samples were incorrectly classified as "Thank you." A similar relationship is observed between "Good evening" and "Good night". This quantitative breakdown confirms that the errors were not random but were highly specific and reciprocal swaps between signs that are kinematically and semantically very similar. This suggests the model has learned the correct fundamental kinematics to a very high degree of precision and is now grappling only with the most ambiguous sign pairs in the dataset.

Table 3: Confusion Matrix Components for the Dynamic Word Model (Selected Classes)

Meaning in this study	Nice	Thank you	Good evening	Good night	Hello*
Correctly identified the specific sign class from the input video	17	20	18	20	20
Incorrectly predicted sign class when a different sign was actually performed	0	3	0	2	0
Correctly rejected this sign class when analyzing samples from other classes	1,360	1357	1,360	1358	1,360
Failed to identify the true sign class, misclassifying it as a different sign	3	0	2	0	0

Note: "Hello" represents the 65 other classes that achieved perfect classification. The test set contained 20 videos per class (Total N=1,380).

The experimental results provide compelling validation for the hypothesis that invariant feature representation, particularly when combined with data augmentation, constitutes a superior strategy for SLR compared to raw data ingestion. While the Baseline Model achieved a respectable 94.90% accuracy, demonstrating the inherent capability of LSTMs to learn from raw coordinates, qualitative real-world testing exposed its fragility; the model frequently failed when the signer shifted their position or adjusted their distance from the camera. In contrast, the Proposed Model's significant performance leap to 99.85% can be directly attributed to the Full-Body Relative Feature Engine, which effectively solved the "Position" and "Scale" problems by mathematically anchoring limb positions to the torso and wrist, thereby forcing the network to learn the intrinsic gesture rather than absolute pixel coordinates. This robustness was empirically confirmed through the integration of the trained model into a real-time application using a standard webcam. As illustrated in Figure 7, the system successfully tracks the user's relative posture and identifies signs with high confidence exceeding 99% even in cluttered environments. Furthermore, the system meets critical requirements for assistive technology regarding latency, as the combined pipeline executes in under 15 milliseconds per frame on the test hardware, enabling a seamless, real-time feedback loop at 30+ FPS without perceptible lag.

ISSN 2394 - 7780

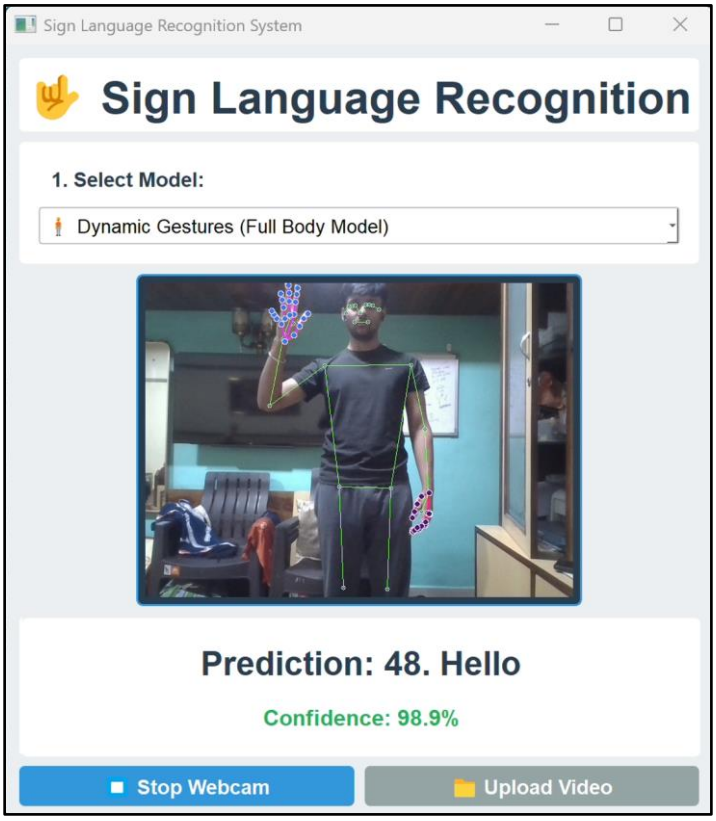

Figure 7: Real-Time System Interface showing a correct prediction using the Dynamic Word Model

Complementing the dynamic gesture analysis, we also validated the system's precision on the fundamental building blocks of sign language. Figure 8 illustrates the deployment of the static alphabet model, explicitly highlighting the real-time visualization of the skeletal landmarks. The overlay demonstrates the system's capability to maintain robust tracking of high-density hand meshes even in the presence of background clutter. The high confidence scores associated with these static letters confirm that the relative feature engineering pipeline effectively isolates the intrinsic geometry of the handshape, operating with the same high fidelity as it does for complex arm trajectories.

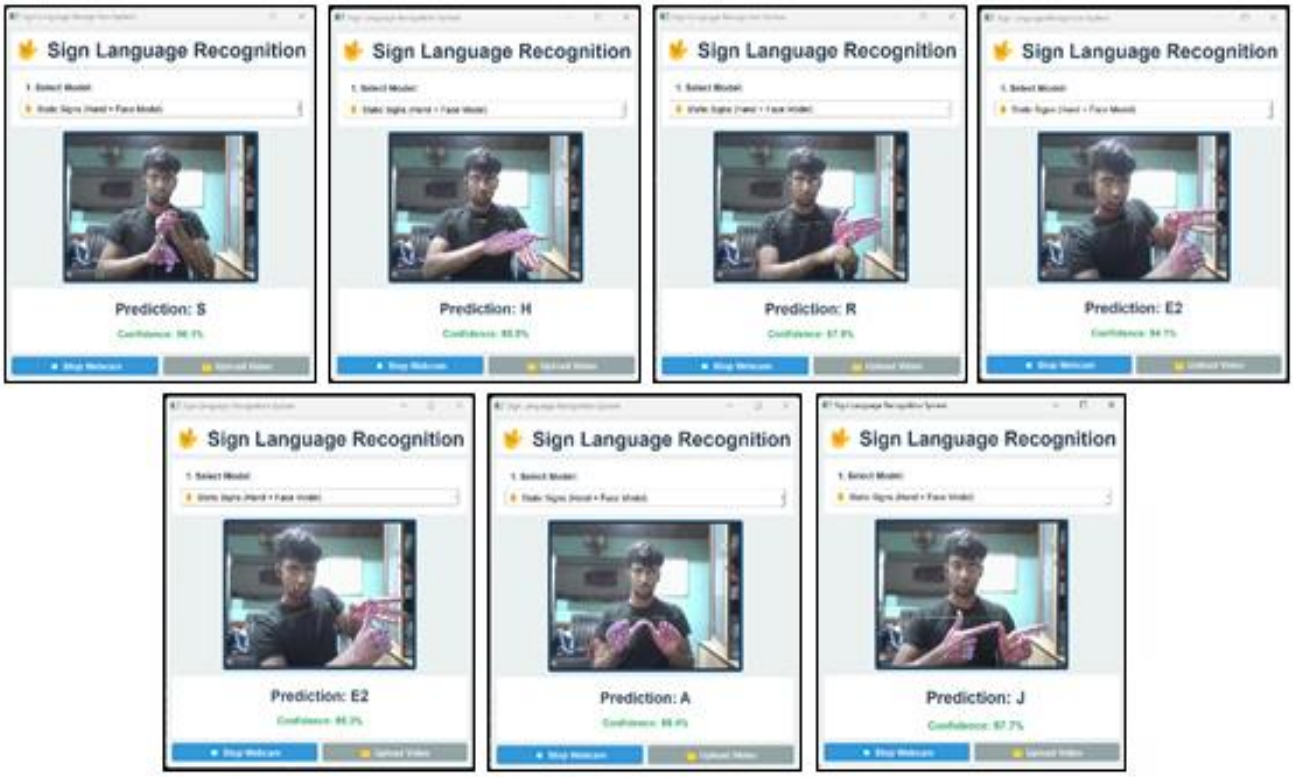

Figure 8: Real-Time Landmark Visualization and classification results using the Static Alphabet Model

ISSN 2394 - 7780

5. CONCLUSION AND FUTURE SCOPE

This research successfully developed and validated a robust, real-time framework for Indian Sign Language recognition. By shifting the focus from model complexity to intelligent feature engineering, we demonstrated that a signer-invariant representation normalizing skeletal data against the user's own anatomy yields state-of-the-art results. The final system, powered by a deep LSTM and trained on augmented data, achieved a 99.85% accuracy, effectively bridging the gap between academic benchmarks and real-world utility. This work provides a scalable blueprint for developing accessible communication technologies for the deaf and hard-of-hearing community. Furthermore, this study empirically challenges the prevailing trend in computer vision that relies on increasingly complex architectures, such as Graph Neural Networks or Transformers, to solve gesture recognition tasks. By achieving near-perfect accuracy with a computationally efficient LSTM, we have proven that the quality of data representation is often more critical than the depth of the network. The dual-phase validation achieving 100% accuracy on static primitives and 99.85% on dynamic gestures confirms that the proposed "Full-Body Relative" feature set is a generalizable solution capable of capturing both fine-grained handshapes and gross motor arm trajectories. With an end-to-end inference latency of under 15 milliseconds, the system stands not just as a theoretical success, but as a viable, low-latency solution ready for deployment on standard consumer hardware.

To evolve this system from a research prototype into a comprehensive translation tool, future work must first address the challenge of scaling the vocabulary to over 1,000 words, which will inevitably increase class overlap. We propose implementing a hierarchical "Router-Expert" Mixture of Experts (MoE) architecture, where a top-level model classifies the semantic context and delegates inference to specialized sub-models to mitigate confusion between similar signs. Furthermore, research will shift toward Continuous Sign Language Recognition (CSLR) to enable natural conversation; this entails adopting advanced techniques such as Connectionist Temporal Classification (CTC) loss or Transformer-based architectures capable of recognizing continuous sentences without requiring artificial pauses. Finally, to democratize access, we aim to focus on edge deployment by quantizing the model to TensorFlow Lite, enabling efficient on-device processing for smartphones to maximize user privacy and accessibility for individuals without high-end hardware.

Recommendations

Based on the empirical evidence presented in this study, we offer several strategic recommendations for the advancement of automated Sign Language Recognition systems. Foremost, researchers and developers should deprioritize the use of raw, absolute skeletal coordinates in favor of geometrically invariant feature representations. Our results conclusively demonstrate that normalizing landmarks relative to the signer’s anatomy is a far more effective method for achieving robustness against positional and scale variations than relying on increasing model depth alone. Furthermore, the critical role of temporally consistent data augmentation cannot be overstated; it is recommended that future training pipelines rigorously integrate geometric transformations to proactively bridge the domain gap between controlled datasets and the unpredictability of real-world deployment. Finally, for applications demanding low-latency interaction on consumer hardware, we recommend retaining efficient recurrent architectures like LSTMs, which provide an optimal balance of computational lightness and predictive accuracy, reserving more computationally intensive Transformer models for tasks involving long-form continuous translation where global context is paramount.

Acknowledgement

The authors express their sincere gratitude to the Department of Computer Science at Sheth L.U.J College of Arts & Sir M.V. College Of Science & Commerce, Mumbai, for providing the necessary academic environment and infrastructural support to conduct this research. We would also like to acknowledge the creators of the 'Include' and 'Static ISL' datasets for making their resources publicly available, which were instrumental in the development and validation of the proposed framework. Finally, we thank the deaf and hard-of-hearing community for their continued advocacy, which serves as the primary inspiration for this work.

Funding Support

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The project was entirely self-funded by the authors, who contributed equally to the procurement of hardware components and development resources.

Ethical Statement

This study did not involve the collection of new data from human or animal participants. All analyses were conducted using publicly available datasets and synthetically generated data. Therefore, ethical approval was not required for this research.

Conflicts of Interest
The authors declare that they have no conflicts of interest to this work.

Data Availability Statement
The datasets used in this study are publicly available or synthetically generated for experimental purposes. The data and supporting materials are available from the corresponding author upon reasonable request.

REFERENCES

Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J. S., Fox, N., & Zisserman, A. (2020). BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 35–53). Springer. https://doi.org/10.1007/978-3-030-58621-8_3

Al-Qurishi, M., Khalid, T., & Souissi, R. (2021). Deep learning for sign language recognition: Current techniques, benchmarks, and open issues. *IEEE Access*, *9*, 126917–126951. https://doi.org/10.1109/ACCESS.2021.3110912

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 125. https://doi.org/10.3390/info11020125

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10023–10033. https://doi.org/10.1109/CVPR42600.2020.01004

Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., & Mak, B. (2022). Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, *35*, 17043–17056.

Cheng, K. L., Yang, Z., Chen, Q., & Tai, Y.-W. (2020). Fully convolutional networks for continuous sign language recognition. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 697–714). Springer. https://doi.org/10.1007/978-3-030-58586-0_41

Hao, A., Min, Y., & Chen, X. (2021). Self-mutual distillation learning for continuous sign language recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11303–11312. https://doi.org/10.1109/ICCV48922.2021.01111

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, L., Gao, L., Liu, Z., & Feng, W. (2023). Continuous sign language recognition with correlation network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2529–2539. https://doi.org/10.1109/CVPR52729.2023.00249

Hu, L., Gao, L., Liu, Z., & Feng, W. (2023). Self-emphasizing network for continuous sign language recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(1), 854–862. https://doi.org/10.1609/aaai.v37i1.25164

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). Skeleton aware multi-modal sign language recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3408–3418. https://doi.org/10.1109/CVPRW53098.2021.00380

Katoch, S., Singh, V., & Tiwary, U. S. (2022). Indian Sign Language recognition system using SURF with SVM and CNN. *Array*, *14*, 100141. https://doi.org/10.1016/j.array.2022.100141

Kothadiya, D., Bhatt, C., Sapariya, K., Patel, K., Gil-González, A.-B., & Corchado, J. M. (2022). Deepsign: Sign language detection and recognition using deep learning. *Electronics*, *11*(11), 1780. https://doi.org/10.3390/electronics11111780

Kothadiya, D. R., Bhatt, C. M., Saba, T., Rehman, A., & Bahaj, S. A. (2023). SIGNFORMER: DeepVision transformer for sign language recognition. *IEEE Access*, *11*, 4730–4739. https://doi.org/10.1109/ACCESS.2022.3231130

Laines, D., Gonzalez-Mendoza, M., Ochoa-Ruiz, G., & Bejarano, G. (2023). Isolated sign language recognition based on tree structure skeleton images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 276–284. https://doi.org/10.1109/CVPRW59228.2023.00034

Liu, T., Tao, T., Zhao, Y., Li, M., & Zhu, J. (2024). A signer-independent sign language recognition method for single-frequency datasets. *Neurocomputing*, *582*, 127479. https://doi.org/10.1016/j.neucom.2024.127479

Lugaresi, C., et al. (2019). MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*. https://arxiv.org/abs/1906.08172

Miah, A. S. M., Hasan, M. A. M., Nishimura, S., & Shin, J. (2024). Sign language recognition using graph and general deep neural network based on large scale dataset. *IEEE Access*, *12*, 34553–34569. https://doi.org/10.1109/ACCESS.2024.3372425

Min, Y., Hao, A., Chai, X., & Chen, X. (2021). Visual alignment constraint for continuous sign language recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11542–11551. https://doi.org/10.1109/ICCV48922.2021.01134

Rastgoo, R., Kiani, K., & Escalera, S. (2020). Hand sign language recognition using multi-view hand skeleton. *Expert Systems with Applications*, *150*, 113336. https://doi.org/10.1016/j.eswa.2020.113336

Roh, K., Lee, H., Hwang, E. J., Cho, S., & Park, J. C. (2024). Preprocessing MediaPipe keypoints with keypoint reconstruction and anchors for isolated sign language recognition. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Mesch, & M. Schulder (Eds.), *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages* (pp. 323–334). ELRA and ICCL. https://aclanthology.org/2024.signlang-1.36

Samaan, G. H., Wadie, A. R., Attia, A. K., Asaad, A. M., Kamel, A. E., Slim, S. O., Abdallah, M. S., & Cho, Y.-I. (2022). MediaPipe's landmarks with RNN for dynamic sign language recognition. *Electronics*, *11*(19), 3228. https://doi.org/10.3390/electronics11193228

Sharma, S., & Kumar, K. (2021). ASL-3DCNN: American Sign Language recognition technique using 3-D convolutional neural networks. *Multimedia Tools and Applications*, *80*(17), 26319–26331. https://doi.org/10.1007/s11042-021-10768-5

Singh, A. (2022). *Static gestures of Indian Sign Language (ISL) for English alphabet, Hindi vowels and numerals* [Data set]. Mendeley Data. https://doi.org/10.17632/7TSW22Y96W.1

Sridhar, A., Ganesan, R. G., Kumar, P., & Khapra, M. (2020). INCLUDE: A large scale dataset for Indian Sign Language recognition. *Proceedings of the 28th ACM International Conference on Multimedia*, 1366–1375. https://doi.org/10.1145/3394171.3413528

Wadhawan, A., & Kumar, P. (2020). Deep learning-based sign language recognition system for static signs. *Neural Computing and Applications*, *32*(12), 7957–7968. https://doi.org/10.1007/s00521-019-04691-y

Zhang, J., Bu, X., Wang, Y., Dong, H., Zhang, Y., & Wu, H. (2024). Sign language recognition based on dual-path background erasure convolutional neural network. *Scientific Reports*, *14*, 11360. https://doi.org/10.1038/s41598-024-62008-z

Zhang, P., Yin, H., Wang, Z., Chen, W., Li, S., Wang, D., Lu, H., & Jia, X. (2024). EvSign: Sign language recognition and translation with streaming events. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2024* (pp. 338–355). Springer. https://doi.org/10.1007/978-3-031-72652-1_20

Zuo, R., & Mak, B. (2022). C2SLR: Consistency-enhanced continuous sign language recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5131–5140. https://doi.org/10.1109/CVPR52688.2022.00506

Zuo, R., Wei, F., & Mak, B. (2023). Natural language-assisted sign language recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14890–14900. https://doi.org/10.1109/CVPR52729.2023.01430

Zuo, R., Wei, F., & Mak, B. (2024). Towards online continuous sign language recognition and translation. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11050–11067. https://aclanthology.org/2024.emnlp-main.619