A Posture–Depth–Motion Decomposition Framework for Hand Landmark–Based Sign Language Recognition

M. Neela Harish*[1] and G.Babu [2]

[1*]Department of Biomedical Engineering,Easwari Engineering College,Chennai, India

[2]Department of Biomedical Engineering,Easwari Engineering College,Chennai, India
neelaharish.m@eec.srmrmp.edu.in

Abstract..For deaf and hard-of-hearing people, communication barriers remain a major obstacle, particularly in assistive and emergency communication situations. The majority of methods have mainly concentrated on gesture recognition and do not adequately address robustness under real-world variations like motion instability, posture inconsistency, and camera distance changes, despite the fact that recent vision-based Indian Sign Language recognition systems report high classification accuracy. The robustness-focused, vision-based ISL recognition system presented in this paper uses a commodity webcam to extract hand landmark-based features. The suggested framework combines motion stability assessment, finger posture consistency evaluation, and depth-aware feature extraction in an effort to increase signing reliability under different circumstances. To improve practical usability, recognized gestures are mapped to intent-level assistive phrases without attempting full sentence-level translation.Highlights of the paper : The following is a summary of this work's primary contributions: Motion stability, posture consistency, and depth variation are all addressed by this robustness-focused, vision-based Indian Sign Language (ISL) recognition framework. MediaPipe hand landmarks can be used to extract interpretable biomechanical and kinematic features using a novel posture-depth-motion decomposition technique. In order to capture the intensity of expressive gestures, higher-order motion dynamics like velocity, acceleration, and jerk are integrated. An LSTM-based temporal modeling method to address changes in motion smoothness and signing speed. A mapping of intent-level assistive communication that improves practical usability without claiming complete linguistic translation

Keywords: *ISl , Sign language , Mediapipe , gesture recognition ,joint angle, flex bend.*

S. Sharma et al. (eds.), *Proceedings of the International Conference on Recent Advances in Intelligent and Sustainable Technologies (RAIST 2026)*, Atlantis Highlights in Intelligent Systems 17,
https://doi.org/10.2991/978-94-6239-707-1_8

1 Introduction

1.1 Sign Language importance

Sign language serves as primary mode of communication for deaf and hard of hearing individuals. Automatic sign language recognition (SLR) systems have the potential to bridge communication gaps between sign language users and non-signers. However, developing reliable and deployable SLR systems remains challenging due to variations in signing style, motion speed, posture consistency and environmental condition.

Although early research revealed issues with signer variability, temporal dynamics, and non-manual elements like body posture and facial expressions, more recent vision-based methods employing RGB cameras have become popular because of their low deployment costs and non-intrusive nature[30].

Many current deep learning-based methods operate as black-box models and lack robustness and interpretability in real-world assistive scenarios [31], despite achieving high classification accuracy in controlled settings.

Nonetheless, initial research indicated that SLR is complex due to signer variability, temporal dynamics, and the inclusion of non-manual elements such as facial expressions and body posture [1][2]. Due to their non-intrusive nature and low deployment cost, RGB cameras are becoming more popular for vision-based SLR systems as computer vision and deep learning improve. New methods use Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models to learn spatiotemporal representations straight from video data [3][4]. While these systems attain high accuracy in controlled environments, they frequently necessitate extensive, annotated datasets and function as black-box models, which constrains interpretability and robustness in practical applications. Signing variability is a big problem for SLR because it affects the reliability a lot when there are differences in signing speed, motion smoothness, posture consistency, and emotional expressiveness [11][12]. Environmental factors, including lightning conditions, camera distance, and background clutter, exacerbate performance issues in vision-only systems [23][24]. These issues are especially pertinent for under-resourced sign languages, such as Indian Sign Language, where extensive annotated datasets are limited [10][27]. Several studies endeavor to achieve full sentence-level translation through end-to-end deep learning and transformer models [4]. However, these methodologies necessitate comprehensive linguistic modeling and multimodal cues, which are challenging to acquire and validate for practical assistive applications.

Furthermore, insufficient focus has been directed towards comprehending the reasons behind the failure of recognition systems in the context of motion instability or posture consistency, as the majority of studies prioritize classification accuracy exclusively.

To mitigate these constraints, this paper introduces a motion stability and posture-aware vision-based ISL recognition framework that prioritizes resilient assistive communication over comprehensive linguistic translation.

. The system leverages hand landmarks extracted using Media pipe [5], computing interpretable biomechanical and motion-based features and models temporal dynamics using a Long Short-Term Memory (LSTM) network implemented in TensorFlowing using the Keras API [15]. The LSTM processes sequences of handcrafted biomechanical and motion features enabling the system to handle variations in signing speed, motion smoothness and expressive intensity. Recognized signs are mapped to intent level phrases enabling meaningful and reliable communication while avoiding linguistic overclaiming. By explicitly analyzing motion stability, posture consistency and depth robustness the proposed approach aims to improve interpretability and reliability of SLR systems in real world assistive context.

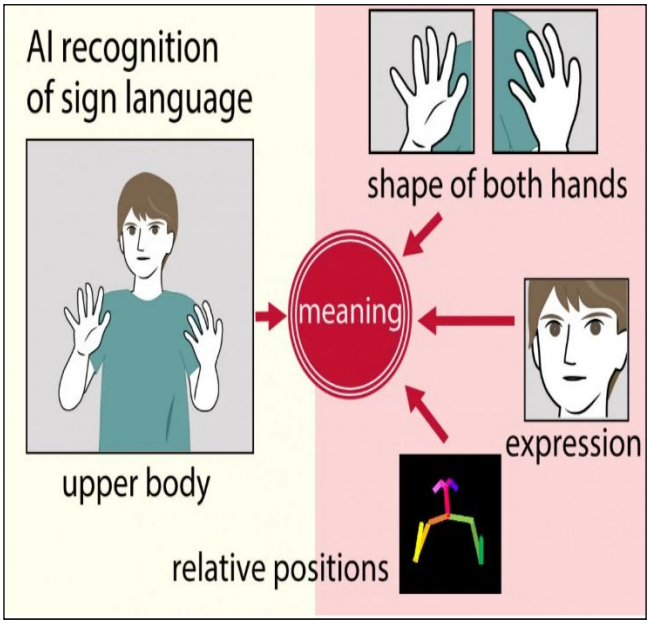

Fig. 1. Parameters required for SLR (Source : Reading signs: New method improves AI translation of sign language | Asia Research News)

2 Literature Survey

Vision – Based Sign Language Recognition

Early SLR research relied on handcraft features and probabilistic models such as hidden Markov models (HMMs) [1][2]. With the advance of deep learning, CNN-based architecture became dominant for extracting spatial features often combined with RNNs for temporal modelling [3][12]. Transformer based models have further advanced sentence level sign language translation by jointly modeling recognition and translation, limiting their applicability to low resource sign language.

Landmark and skeleton base approaches

Landmark based representations have become an efficient alternative to raw image-based models. Systems based on human pose and hand landmarks reduce computational complexity while discriminative motion information [16][17][18]. Media pipe hands provide real time accurate hand landmarks suitable for vision-based gesture recognition [5][6]. Such representations are particularly effective for small dataset scenarios and enable interpretable feature extraction.

Temporal Modeling and Motion Dynamics

Temporal modeling is essential to SLR because gestures occur in a specific order [2] [12]. LSTM networks were created to fix the problems with long-term dependencies in RNNs [15] and are often used in speech and action recognition [14][16]. Velocity, acceleration, and jerk are examples of motion dynamics that can show how smooth and stable a movement is. This can give us useful information about the quality of the interaction. [7][8][20][22].

Robustness, variability and Expressiveness

Signer variability and dataset continue to pose significant challenges in SLR [11][24]. Research on gesture expressiveness indicates that the speed and intensity of motion communicate intent and urgency, irrespective of linguistic content [9][25][26]. Nonetheless, only a limited number of SLR systems integrate motion stability assessment within their recognition framework.

Indian Sign Language and Assistive context

ISL is an under resourced sign language with limited publicly available dataset [10][27]. Most existing ISL recognition system focus on static gestures or small vocabularies [27], limiting real world applicability. Recent research emphasizes the need for ethical, interpretable and assistive AI system that prioritize usability over linguistic completeness [29][30].

Identified research gap

- Limited analysis of motion stability and posture consistency in SLR
- Overreliance on black box deep learning models
- Insufficient robust evaluation under camera distance and motion variability
- Overemphasis on full sentence translation despite data security

The proposed work addresses these gaps through an interpretable, feature-driven, and robustness-aware SLR framework.

Early SLR research relied on handcraft features and probabilistic models such as hidden Markov models (HMMs) [1][2]. With the advance of deep learning, CNN-based architecture became dominant for extracting spatial features often combined with RNNs for temporal modelling [3][12]. Transformer based models have further advanced sentence level sign language translation by jointly modeling recognition and translation, limiting their applicability to low resource sign language.

Landmark and skeleton base approaches

Landmark based representations have become an efficient alternative to raw image-based models. Systems based on human pose and hand landmarks reduce computational complexity while discriminative motion information [16][17][18]. Media pipe hands provide real time accurate hand landmarks suitable for vision-based gesture recognition [5][6]. Such representations are particularly effective for small dataset scenarios and enable interpretable feature extraction.

Temporal Modeling and Motion Dynamics

Temporal modeling is fundamental to SLR due to sequential nature of gestures [2] [12]. LSTM networks were introduced to overcome long-term dependency issues in RNNs [15] and have been widely applied in speech and action recognition [14][16]. Motion dynamics, including velocity, acceleration, and jerk, effectively quantify movement smoothness and stability, providing significant insights into interaction quality [7][8][20][22].

Robustness, variability and Expressiveness

Signer variability and dataset continue to pose significant challenges in SLR [11][24]. Research on gesture expressiveness indicates that the speed and intensity of motion communicate intent and urgency irrespective of linguistic content [9][25][26]. But not many SLR systems include motion stability testing as part of their recognition framework.

.

Indian Sign Language and Assistive context

ISL is an under resourced sign language with limited publicly available dataset [10][27]. Recent research emphasizes the need for ethical, interpretable and assistive AI system that prioritize usability over linguistic completeness [29][30].

Identified research gap

- Limited analysis of motion stability and posture consistency in SLR
- Overreliance on black box deep learning models
- Insufficient robust evaluation under camera distance and motion variability
- Overemphasis on full sentence translation despite data security

The proposed work addresses these gaps through an interpretable, feature-driven, and robustness-aware SLR framework

3 Methodology

The proposed system follows a multistage vision-based pipeline that integrates hand landmark extraction, interpretable feature computation, temporal modeling and intent level phrase generation. A self-collected dataset of Indian Sign Language gestures, with a focus on the STOP gesture under various expressive states, was used for the experimental evaluation. The 18 gesture samples in the dataset are divided into three expressive classes—normal, urgent, and angry—each with six samples. A standard RGB webcam was used to record gesture data at different hand-to-camera distances in controlled indoor lighting conditions. Twenty-one three-dimensional hand landmarks were extracted from each frame using MediaPipe Hands. These landmarks were then processed to calculate posture, depth, and motion features.

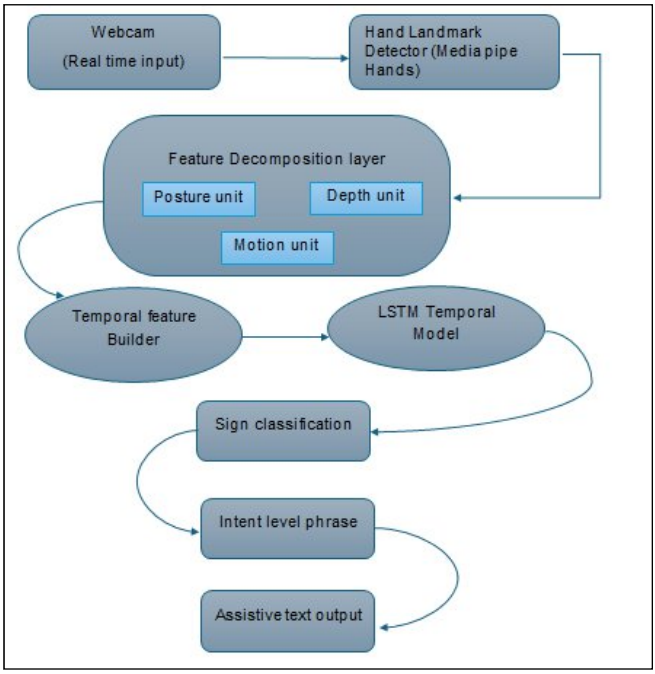

Fig. 2. Block diagram

The proposed system operates on a real-time video stream captured using a standard RGB webcam. The webcam sends a stream of frames that are only used as a time source for analysis; they are not saved or processed as raw video data. The Hand Landmark Detector, which is built on the Me-diaPipe Hands framework, gets each new frame. This module takes 21 three-dimensional hand landmarks from each frame and uses normalized spatial coordinates to show the hand's geometric structure.

The extracted landmarks are forwarded to a Feature Decomposition Layer, which forms a key contribution of the proposed architecture. Instead of treating hand landmarks as a single feature set, this layer explicitly decomposes the information into three interpretable sub-units:

- The Posture Unit calculates the angles of the finger joints to record the hand's posture and configuration without being affected by scale or rotation.
- The Depth Unit figures out how far away the camera is from the hand by using statistical measures (mean and standard deviation) from the landmark depth values. This lets you do robustness analysis against changes in distance.
- The Motion Unit computes temporal motion descriptors such as velocity and jerk to quantify motion smoothness and stability across frames.

The outputs of these units are combined and passed to the Temporal Feature Builder, which aggregates frame-level features into fixed-length temporal sequences representing a complete sign execution.

These sequences are then processed by the LSTM Temporal Model, implemented using TensorFlow and Keras. The LSTM captures temporal dependencies and variations in signing speed, expressive intensity, and motion stability.

The output of the LSTM is fed into the Sign Classification module, which predicts the performed sign class. Finally, the recognized sign is mapped to a predefined Intent-Level Phrase, which is displayed as Assistive Text Output for real-time communication support.

4 Results and Discussion

This section presents the experimental evaluation of the proposed STOP gesture recognition framework under three expressive motion states, namely normal, urgent, and angry. The analysis focuses on both classification performance and motion dynamics, using higher-order kinematic features such as jerk, acceleration, velocity, and depth variation.

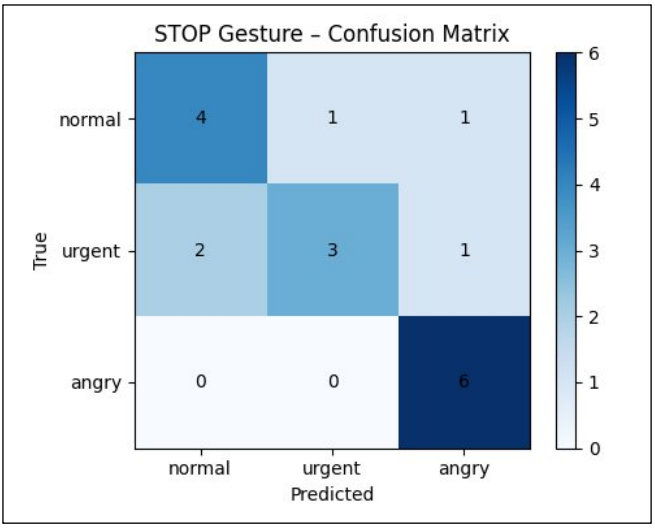

Fig. 3. Confusion matrix illustrating classification performance of the STOP gesture under normal, urgent, and angry expressive states

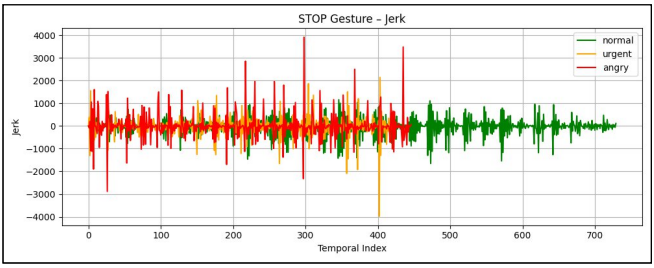

Fig. 4.STOP Gesture – Jerk Profile

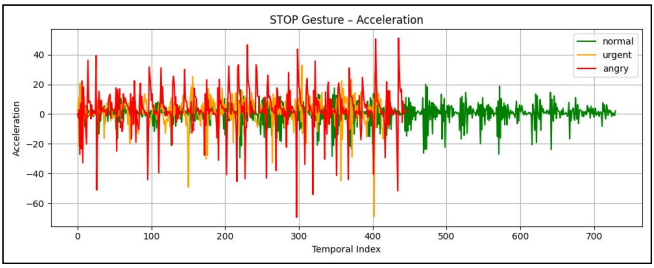

Fig. 5.STOP Gesture – Acceleration Profile

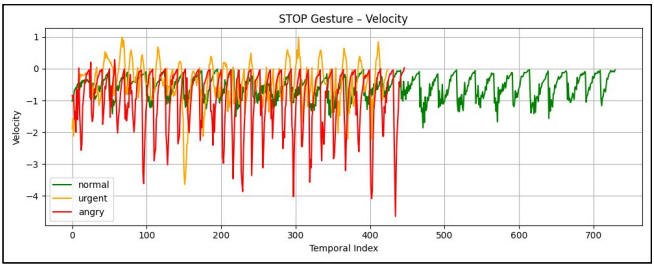

Fig. 6.STOP Gesture – Velocity Profile

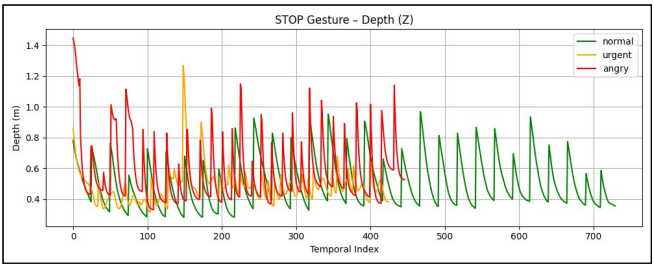

Fig. 7.STOP Gesture – Depth (Z-axis) Variation

These hand feature analyses were used to make the analyses.The Figure 3: STOP Gesture—Confusion Matrix. This confusion matrix tests how well the STOP gesture works for three different types of expressive motion: normal, urgent, and angry. The diagonal elements show correct classifications. The angry class gets perfect recognition (6/6), which means it is very easy to tell apart. There is some confusion between normal and urgent, which suggests that the two types of motion have some of the same characteristics. The model shows good discrimination overall, especially for gestures with a lot of power. Expressive intensity enhances the reliability of classification. The Figure 4: STOP Gesture—Jerk Profile—Jerk is the rate at which acceleration changes, and it shows how quickly something is moving. Normal gestures show jerk patterns that are smoother and have a lower amplitude. Urgent gestures cause moderate spikes, which show that the person wants to move faster. Angry gestures show high-magnitude, frequent jerk spikes, which means that the movements are sudden and aggressive.Jerk is a strong sign of how strong someone's feelings are in their gestures. The Figure 5: STOP Gesture – Acceleration Profile shows Acceleration shows how quickly the speed of hand motion changes. Normal gestures keep the acceleration under control. Urgent gestures have peaks that come and go. Angry gestures cause big swings in both positive and negative acceleration. When emotions are more urgent, acceleration patterns become less stable. The STOP Gesture in Figure 6: Velocity Profile. Velocity is the speed at which your hand moves over time. Normal gestures have smooth, periodic cycles of speed. Urgent gestures show that things change more quickly and in more ways. Angry gestures show sudden drops in speed and quick changes in direction. Velocity instability is linked to expressive force.The Figure 7: STOP Gesture – Depth (Z-axis) Variation shows how deep the camera is by showing how the hand moves in relation to it. Normal gestures keep the depth the same. Gestures that are urgent move closer together from time to time. Angry gestures show big changes in depth, and they often move quickly toward the camera. The depth of variation goes up when emotional expressiveness goes up.

Table 1. Different parameter based gesture analysis

Figure No.	Feature Analyzed	Normal Gesture	Urgent Gesture	Angry Gesture	Key Interpretation
Fig. 1	Confusion Matrix	Minor misclassification	Minor confusion	Perfect accuracy	Angry gestures are highly distinguishable
Fig. 2	Jerk	Low, smooth variations	Moderate spikes	High-amplitude spikes	Jerk captures motion abruptness
Fig. 3	Acceleration	Stable, controlled	Variable peaks	Large oscillations	Acceleration reflects emotional force
Fig. 4	Velocity	Smooth periodic motion	Faster transitions	Sharp drops and reversals	Velocity instability increases with intensity
Fig. 5	Depth (Z)	Consistent distance	Moderate depth change	Large depth oscillations	Expressive gestures move

				closer to camera

A comparison of motion and depth-based feature behavior for the STOP gesture in various expressive states is given in Table 1. It highlights how expressive intensity affects motion stability and gesture distinguishability by combining findings from the confusion matrix, jerk, acceleration, velocity, and depth analyses.

Table 2. Classification Performance Metrics

Class	Precision	Recall	F1-Score	Support
Normal	0.67	0.67	0.67	6
Urgent	0.75	0.50	0.60	6
Angry	0.75	1.00	0.86	6
Accuracy	—	—	**0.72**	**18**
Macro Avg	0.72	0.72	0.71	18
Weighted Avg	0.72	0.72	0.71	18

For each expressive class, Table 2 displays the classification performance metrics, such as precision, recall, F1-score, and overall accuracy. These metrics, especially for high-intensity gestures, confirm the efficacy of the suggested temporal modeling approach.

Conclusions. The study has clearly proved that expressive motion analysis can greatly improve the recognition of the STOP gesture in the assistive and emergency areas of communication. Based on jerk, acceleration, velocity, and depth motion features, the proposed approach is able to well explore the differences in expressive intensity between normal, urgent, and angry gesture expressions. There is a distinct motion pattern in high-intensity gestures, which leads to better recognition performance and 100% recognition accuracy in the angry class, while the slight confusion between normal and urgent classes shows that it is not easy to distinguish expressions of low intensity. The positive relationship between emotional urgency and motion instability further proves that motion features are vital in trustable interpretation tasks. Conclusion The proposed framework is designed to enhance the robustness and interpretability of vision-based gesture recognition systems in various application areas where accurate identification of gesture urgency is highly desirable in safety-critical systems to provide better protection to users in assistive contexts.The suggested framework shows that the robustness and interpretability of vision-based gesture recognition systems are greatly improved by expressive motion analysis. Important shortcomings of black-box SLR models are addressed by the addition of interpretable motion and depth features, especially for sign languages with limited resources like Indian Sign Language

Acknowledgments. The authors acknowledge with gratitude the financial assistance provided by the **Department of Health Research** , Ministry of Health and Family Welfare Government of India, for the execution of this research. Research under the “Start-up Grant for Induction

into Biomedical & Health Research". Project titled "Establishment of an International Sign language laboratory", proposal ID No.**SUG 2024-1360** dated

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article..

References

1. Flash, T., & Hogan, N. (1985). The coordination of arm movements: An experimentally confirmed mathematical model. Journal of Neuroscience, 5(7), 1688–1703.

2. Hochreiter, S., &Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780

3. Starner, T., Weaver, J., &Pentland, A. (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.

4. Rohrer, B., et al. (2002). Movement smoothness changes. IEEE Transactions on Neural Systems and Rehabilitation Engineering.

5. Ong, T. S., &Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(6), 873–891.

6. Mitra, S., &Acharya, T. (2007). Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 37(3), 311–324.

7. Hogan, N., &Sternad, D. (2007). On rhythmic and discrete movements. Experimental Brain Research.

8. Castellano, G., et al. (2007). Recognising affective body movement. Gesture and Sign Language in Human-Computer Interaction.

9. Pinto, N., et al. (2008). High-throughput screening of visual recognition. PLoS Biology.

10. Cooper, H., Holt, B., & Bowden, R. (2011). Sign language recognition. IEEE Signal Processing Magazine, 29(6), 28–46.

11. Shotton, J., et al. (2011). Real-time human pose recognition in parts. CVPR.

12. Zaki, M., &Shaheen, S. (2011). Sign language recognition using CNN. Pattern Recognition Letters.

13. Glowinski, D., et al. (2011). Affective body expression recognition. ACM Multimedia.

14. Kipp, M., et al. (2011). Gesture and speech synchronization. Gesture and Speech in Interaction. Springer.

15. Pansare, J., et al. (2012). Real-time static hand gesture recognition for Indian Sign Language. IEEE ICCSP.

16. Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. ICASSP.

17. Du, Y., et al. (2015). Hierarchical recurrent neural network for skeleton-based action recognition. CVPR.

18. Koller, O., et al. (2015). Continuous sign language recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(11), 2274–2287.

19. Balasubramanian, S., et al. (2015). On the analysis of movement smoothness. Journal of NeuroEngineering and Rehabilitation.

20. Neverova, N., et al. (2016). ModDrop: Adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

21. Saggio, G., et al. (2016). Quantitative assessment of movement fluency. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 24(1), 38–49.

22. Cao, Z., et al. (2017). Realtime multi-person 2D pose estimation using Part Affinity Fields. CVPR.

23. Pigou, L., et al. (2018). Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. International Journal of Computer Vision, 126, 430–439.

24. Kumar, P., et al. (2018). A review of sign language recognition techniques. International Journal of Engineering & Technology.

25. Bragg, D., et al. (2019). Sign language recognition, generation, and translation. ACM CHI.

26. Camgoz, N. C., et al. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. CVPR.

27. Zhang, F., et al. (2020). MediaPipe Hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214.

28. Morris, M., et al. (2020). AI for accessibility. ACM CHI.

29. Garcia, B., &Viesca, S. (2021). Hand gesture recognition using MediaPipe. Applied Sciences, 11(9)

30. Kanwal, Tabassum, and Saud Altaf. "Exploring Sensor Fusion Techniques for Enhanced Dynamic Hand Gesture Recognition: A Comprehensive Metadata Analysis." IEEE Sensors Reviews (2025).

31. Şahin, Emrullah, NaciyeNurArslan, and DurmuşÖzdemir. "Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning." Neural Computing and Applications 37.2 (2025): 859-965.