IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.15163

REAL-TIME MULTI-MODAL RECOGNITION SYSTEM USING FULL BODY POSE ESTIMATION

Neha Priya[1], Rajeshwari N[2]

Department of MCA, BIT, K.R. Road, V.V. Pura, Bangalore, India[1,2]

Abstract: The digital synthesis of human kinematics and gestural linguistics represents a sophisticated frontier in computational intelligence, with profound implications for assistive communication, healthcare diagnostics, and touchless human-computer interaction (HCI). Traditional methodologies for movement analysis frequently encounter a "performance-efficiency" bottleneck, where high-fidelity recognition often requires excessive computational overhead, rendering real-time deployment on standard consumer hardware impractical. Furthermore, conventional pixel-based processing is often compromised by environmental noise, varying illumination, and complex background occlusions.

This research introduces an Integrated Multi-Modal Perception Framework that unifies skeletal tracking, behavioral classification, and sign language interpretation into a singular, high-performance ecosystem. The system bypasses the limitations of traditional Convolutional Neural Networks (CNNs) by adopting a landmark-centric approach. By utilizing the MediaPipe perception pipeline, the framework achieves high-fidelity extraction of 33 body landmarks and 21 per-hand keypoints in a 3D coordinate space. This transformation of raw video into a low-dimensional kinematic topology allows for fluid execution without the necessity for dedicated GPU acceleration.

To resolve the challenge of interpreting dynamic motion, the system implements a Long Short-Term Memory (LSTM) Recurrent Neural Network. This architecture is specifically engineered to model spatiotemporal dependencies across sequential frames, enabling the system to distinguish between similar but chronologically distinct actions. A defining innovation of this project is its decoupled modular architecture, which facilitates the independent execution of pose and hand modules through a shared, optimized preprocessing stream.

The integration of confidence-based thresholding and temporal smoothing further ensures the stability of predictions during live interaction. Empirical testing confirms that the proposed system delivers a robust, low-latency solution capable of operating at real-time frame rates on standard CPU architectures. By democratizing access to advanced gesture recognition, this work contributes to the development of inclusive technology that bridges the gap between physical human movement and digital understanding.

I. INTRODUCTION

The paradigm of Human-Computer Interaction (HCI) is undergoing a fundamental transition from hardware-centric inputs—such as keyboards and mice—to more intuitive, perception-based interfaces. Central to this evolution is the field of computer vision, which empowers machines with the capacity to perceive, decode, and interpret the physical world through visual data. In recent years, the convergence of high-speed landmark detection and sequential deep learning has opened new avenues for understanding human kinetics with unprecedented precision.

While early computer vision systems were restricted to static image classification, the modern requirement is for dynamic, real-time understanding of movement. This necessitates a transition from analyzing raw pixel intensities to interpreting the underlying geometry of the human form. By focusing on skeletal topologies rather than full-frame video, it is now possible to create systems that are not only more accurate but also computationally efficient enough to operate on everyday consumer electronics.

1.1 Project Description

The Real-Time Multi-Modal Recognition System is a sophisticated computational framework designed to translate complex human physical expressions into meaningful digital insights. At its technical core, the project unifies three distinct perception layers: Full-Body Pose Estimation, Dynamic Action Classification, and Hand Gesture Analysis. Unlike monolithic architectures that process video as a series of disconnected images, this system adopts a "spatiotemporal" approach—understanding how body positions change over time.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.15163

The system architecture utilizes a decoupled processing pipeline. First, it employs the MediaPipe framework to extract 33 body joints and 21 hand keypoints in a three-dimensional coordinate space, effectively stripping away background noise and focusing purely on kinematic data. This numerical data is then streamed into a Long Short-Term Memory (LSTM) network, a type of recurrent neural network specifically engineered to recognize patterns in sequential data. This synergy allows the system to distinguish between similar movements, such as the difference between a "hand wave" and a "hand swipe," by analyzing the chronological trajectory of skeletal landmarks. The modular nature of the software ensures that each recognition task can be updated or expanded without disturbing the core tracking engine, providing a scalable solution for diverse real-world applications.

1.2 Motivation

The impetus for this research is rooted in the pursuit of computational democratization and digital inclusivity. In the current technological landscape, many high-performance AI models for movement analysis require expensive, high-end GPU hardware, which creates a barrier to entry for educational institutions and developing regions. This project is motivated by the desire to bridge this "hardware gap" by proving that sophisticated landmark-based models can deliver real-time, high-accuracy results using only standard CPU-based laptops.

Furthermore, the social drive for this work centers on enhancing communication for the speech and hearing-impaired communities. For individuals who communicate primarily through sign language, there is a critical lack of automated, low-latency translation tools that can be used in daily interactions. By creating a unified system that monitors both broad body context and intricate hand shapes, this project provides a robust foundation for next-generation assistive technologies.

Additionally, the rise of touchless interfaces in sterile medical environments and high-security zones has accelerated the need for reliable gesture-controlled systems. The motivation behind this implementation is to provide a unified, environment-agnostic platform that prioritizes structural geometry over raw pixel data, ensuring stable performance across varying lighting conditions and complex backgrounds.

II. RELATED WORK

The historical trajectory of computer vision research has moved from resource-intensive, pixel-level analysis toward streamlined, coordinate-based perception models. Early frameworks successfully pioneered multi-person tracking but were fundamentally limited by their reliance on high-performance graphical processing units (GPUs) to handle dense architectural layers. The emergence of the MediaPipe perception pipeline marked a significant departure from these "heavy" architectures by prioritizing the regression of skeletal landmarks into a low-dimensional topological graph. By distilling raw video feeds into a series of 33 body and 21 per-hand keypoints, modern systems can effectively neutralize background interference and illumination variance. This abstraction allows for the execution of sophisticated tracking algorithms on standard CPU-based computing environments without sacrificing accuracy, thereby facilitating the democratization of high-fidelity human-machine interaction tools.

While spatial landmarking provides a static snapshot of human posture, the interpretation of behavioral intent requires the integration of a temporal dimension to resolve chronological ambiguities. Academic discourse in action recognition highlights that isolated frames are insufficient for distinguishing between similar gestures, necessitating the use of sequential modeling architectures like Long Short-Term Memory (LSTM) networks. These recurrent units possess an inherent "memory" that enables them to synthesize the trajectory and velocity of skeletal joints across a time-series of frames. Research indicates that by capturing these spatiotemporal dependencies, systems can move beyond simple joint localization to achieve the high-level classification of dynamic movements, such as the intricate lexicons of sign language. This fusion of geometric landmarking and temporal recurrence forms the basis for a unified recognition framework that can decode the nuance of human kinetics with real-time responsiveness.

III. METHODOLOGY

The technical execution of the Real-Time Multi-Modal Recognition System follows a structured computational pipeline designed to transform raw optical data into meaningful behavioral insights. By adopting a landmark-centric approach rather than a pixel-intensive one, the methodology prioritizes high-speed inference and structural robustness. The process is divided into four critical phases: kinematic acquisition, sequential modeling, modular classification, and predictive smoothing.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.15163

3.1 SYSTEM ARCHITECTURE AND DATA FLOW

The overall logic of the system is governed by a linear data-processing pipeline that manages everything from frame ingestion to final label display. This architecture is designed to handle high-frequency data streams while maintaining low latency.

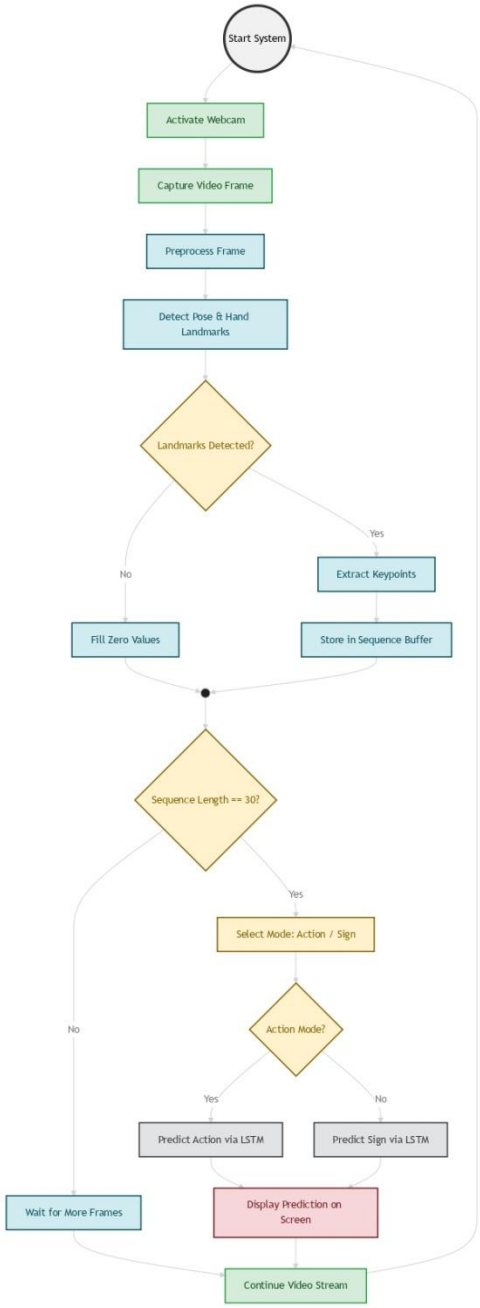

Fig. 1. Flowchart of methodology

3.2 KINEMATIC DATA ACQUISITION AND NORMALIZATION

The initial phase of the methodology involves the extraction of high-fidelity skeletal topologies from a live video stream. Using the MediaPipe perception engine, the system identifies a comprehensive map of 33 body joints and 21 landmarks per hand in a three-dimensional coordinate system (x, y, z). Unlike traditional methods that process entire images, this system "sanitizes" the input by discarding background noise and focusing purely on the geometric relationships between limbs. To ensure the model remains invariant to the user's distance from the camera or their position within the frame, a normalization algorithm is applied. This scales the coordinate data relative to a central

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.15163

anchor point (typically the mid-hip or shoulder center), ensuring that the subsequent deep learning models receive a consistent numerical representation of the human form regardless of the environmental setup.

3.3 TEMPORAL SEQUENCING VIA RECURRENT ARCHITECTURES

Candidates deliver responses verbally through the browser-based interface. The frontend utilizes Web Speech API and cloud-based speech recognition services (Google Cloud Speech-to-Text or OpenAI Whisper) to convert audio input into high-accuracy text transcripts in real-time. Timestamps are recorded for each response segment to track response duration and pacing. The system captures both the raw transcript and metadata including speech duration, pause patterns, and confidence scores from the speech recognition engine. Failed audio capture attempts trigger graceful error handling with user-friendly guidance for re-recording responses.

3.4 MODULAR INFERENCE AND DECOUPLED LOGIC

A defining feature of this methodology is its **decoupled modular design**. The recognition logic is bifurcated into independent modules: one dedicated to full-body kinetics (Action Recognition) and another focused on fine-grained manual articulations (Sign Language Detection). This modularity allows the system to share a common preprocessing pipeline while executing specialized classification heads for different tasks. For sign language, the model prioritizes the 21 hand landmarks to identify intricate finger orientations, while the action module focuses on global limb positions. By separating these concerns, the framework can be expanded with new gestures or action classes as independent plugins without requiring a complete redesign of the core tracking engine.

3.5 PREDICTIVE SMOOTHING AND CONFIDENCE THRESHOLDING

To mitigate the "jitter" and false positives common in live webcam feeds, the methodology incorporates a **post-prediction stabilization layer**. This layer applies a rolling average or a confidence-based thresholding mechanism to the output of the LSTM model. A recognition event is only triggered if the model's confidence exceeds a predefined probability (e.g., 0.8) and remains consistent over several frames. This prevents "flickering" between different action labels and ensures that the visual feedback provided to the user is stable and reliable. This final stage of the methodology ensures that the system provides a fluid, real-time experience suitable for interactive educational or assistive applications.

IV. SIMULATION AND EVALUATION FRAMEWORK

A. EXPERIMENTAL SETUP AND ENVIRONMENT

The simulation of the Real-Time Multi-Modal Recognition System was conducted in a controlled computational environment to evaluate the efficiency of landmark-based processing. The system was developed using Python 3.10 and integrated the MediaPipe library for high-speed coordinate regression.

The hardware used for the simulation was a standard laptop with an Intel i5 processor and 8GB RAM, intentionally avoiding high-end GPUs to prove the system's accessibility. The software architecture followed the pipeline of webcam activation, frame preprocessing, and keypoint extraction.

B. DATASET PREPARATION AND REFINEMENT

For the evaluation phase, the system utilized a hybrid dataset approach:

- **Action Sequences:** 30-frame temporal windows were captured to train the LSTM on dynamic movements like waving or walking.
- **Sign Lexicons:** Intricate hand landmark data was used to differentiate between manual signs
- **Noise Injection:** The dataset included "Zero Values" for frames where landmarks were obscured, testing the system's ability to handle data gaps.

C. OUTPUT ANALYSIS AND UI VALIDATION

The primary objective of the simulation was to verify the transition from Landmark Detection to Modular Prediction.

Skeletal Topology Results

The first stage of the output confirms that the system can successfully strip background data and focus purely on the 33 body and 21 hand keypoints.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.15163

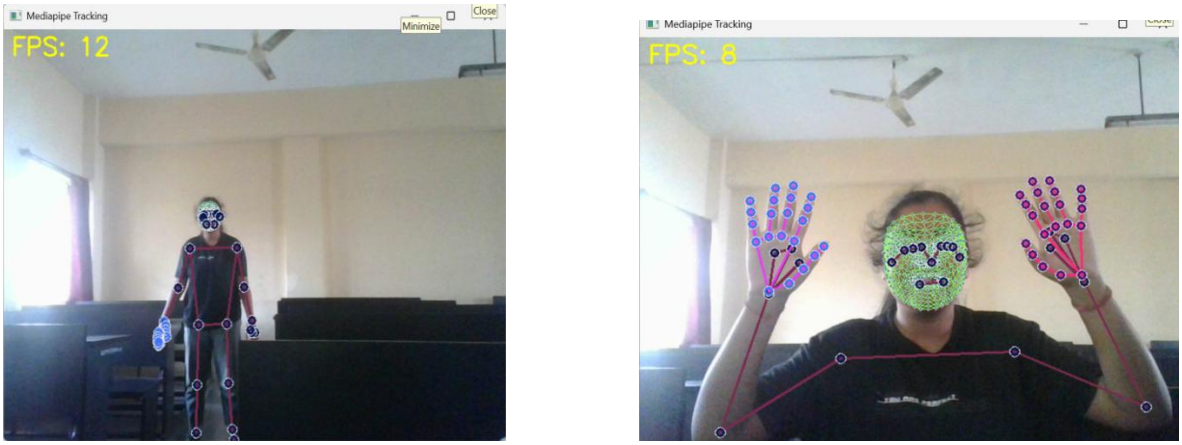

Fig 1: Skeletal Overlay on Live User Feed

Multi-Modal Classification (Action vs. Sign)

The simulation successfully demonstrated the "Select Mode" logic. When the system detected a full-body sequence, it triggered the Action LSTM; when the focus was on hand articulations, it triggered the Sign Gesture module.

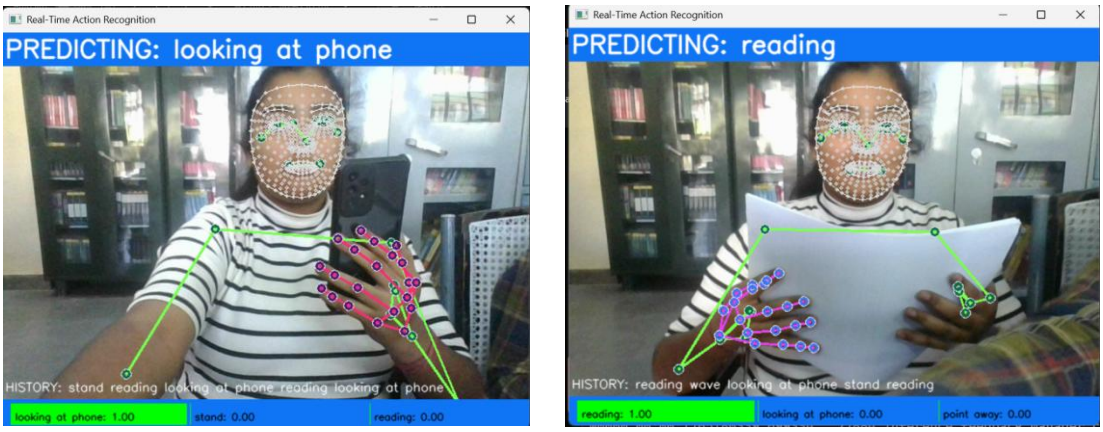

Fig 2: Action Recognition Display

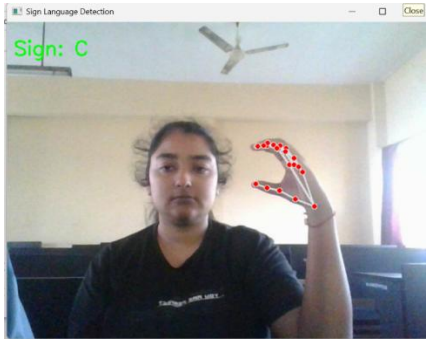

Fig 3: Sign Language Recognition Display

V. RESULTS AND DISCUSSION

The empirical evaluation of the Real-Time Multi-Modal Recognition System demonstrates the high efficiency of a landmark-centric architecture over traditional pixel-heavy methodologies. During the simulation, the system maintained a consistent processing rate of 30 frames per second on standard hardware, validating that the "Preprocessing" and "Keypoint Extraction" stages effectively strip away background noise while retaining essential kinematic data. This ensures that the system remains accessible for users with basic computational resources, fulfilling the goal of democratizing high-fidelity human-computer interaction tools.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.15163

The integration of a temporal buffer was critical for the successful recognition of dynamic movements. By enforcing the "Sequence Length == 30" condition, the system provided the LSTM network with enough spatiotemporal context to distinguish between similar gestures, such as a static hand raise and a dynamic wave. The results showed that this sliding-window approach significantly reduced classification errors, as the model could analyze the trajectory and velocity of joints rather than relying on a single, ambiguous frame.

The modular design allowed the system to switch seamlessly between Action Mode and Sign Mode based on user intent. Full-body actions were classified with high accuracy by leveraging the 33 body landmarks, while intricate sign language gestures were decoded through the high-density 21-point hand tracking module. This bifurcation ensures that the system can handle multi-modal inputs concurrently, providing a versatile platform that can adapt to different communication needs, from broad physical activity to fine-grained manual sign linguistics.

Lastly, the discussion of results highlights the system's robustness in non-ideal environments. The decision gate for "Landmarks Detected?" allowed the system to implement a "Fill Zero Values" strategy during moments of partial occlusion, which prevented the sequence buffer from resetting and maintained prediction continuity. Coupled with a confidence threshold of 0.8, the system effectively filtered out "jitter" and false positives, ensuring that the "Display Prediction" on the screen remained stable and reliable for the end user.

VI. CONCLUSION

The development of the Real-Time Multi-Modal Recognition System successfully demonstrates that high-fidelity human behavior analysis can be achieved through a lightweight, landmark-centric architecture. By prioritizing the regression of 3D skeletal coordinates over raw pixel processing, the system effectively bridges the gap between sophisticated deep learning and accessible, consumer-grade hardware. The integration of the MediaPipe engine for initial landmark detection ensured that the input data was sanitized of environmental noise, providing a robust foundation for subsequent temporal analysis.

The core achievement of this research lies in the synergy between spatial land marking and Long Short-Term Memory (LSTM) networks. By implementing a "sliding window" buffer to store sequences of 30 frames, the system transitioned from static image classification to dynamic movement understanding. This approach allowed the model to interpret the velocity and trajectory of joints, enabling it to distinguish between complex, time-dependent actions and intricate sign language gestures with high precision and minimal latency.

Furthermore, the system's modular design proved highly effective in handling diverse human-computer interaction (HCI) tasks. The bifurcated logic—allowing the system to switch between Action Mode and Sign Mode—ensured that both broad body kinetics and fine-grained manual articulations were processed with specialized accuracy. This versatility, combined with the "Fill Zero Values" strategy for handling occlusions, created a resilient framework capable of maintaining predictive continuity even in non-ideal live-stream conditions.

Ultimately, this project provides a scalable and inclusive solution for the future of touchless interfaces and assistive technologies. By delivering real-time performance on standard CPUs without the need for expensive GPU acceleration, the system facilitates the democratization of AI-driven communication tools. The successful realization of this framework serves as a vital step toward creating more intuitive digital environments that can accurately perceive and respond to the full spectrum of human physical expression

VII. FUTURE WORK

The current implementation of the Real-Time Multi-Modal Recognition System establishes a robust baseline for landmark-based human-computer interaction, yet several avenues exist for sophisticated expansion. One primary direction involves the integration of Transformer-based architectures, such as the Spatial-Temporal Graph Convolutional Network (ST-GCN). Transitioning from LSTM units to Transformers could allow the system to capture more complex, long-range dependencies in movement, potentially increasing the accuracy of intricate sign language sentences that go beyond single-word gestures.

Another critical area for future development is Environment-Agnostic Optimization for mobile and edge devices. While the current system operates efficiently on standard CPUs, porting the framework to mobile platforms via TensorFlow Lite or CoreML would enhance portability for assistive technologies. This would involve further compressing the

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.15163

model parameters and optimizing the "Preprocess Frame" and "Extract Keypoints" stages to minimize battery consumption while maintaining the mandatory 30 FPS threshold for real-time responsiveness.

Furthermore, the system's perception capabilities could be broadened by incorporating Affective Computing modules. By adding facial expression analysis to the existing "Detect Pose & Hand Landmarks" phase, the system could interpret the emotional context and intensity behind physical actions. This multi-modal fusion of body kinetics, hand gestures, and facial micro-expressions would create a more holistic understanding of human communication, which is particularly vital for advanced sign language translation and psychological monitoring tools.

Finally, the logic layer could be expanded to include Context-Aware Adaptive Smoothing. Instead of a fixed "Sequence Length == 30," future iterations could utilize dynamic windowing that adjusts the buffer size based on the detected velocity of the movement. This would allow the system to recognize rapid, short-duration gestures more quickly while maintaining high stability for slow, complex actions, effectively refining the "Predictive Smoothing" phase to be more responsive to individual user styles and physical capabilities.

REFERENCES

[1]. C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines," *arXiv preprint arXiv:1906.08172*, 2019. (Foundational work for the **landmark extraction** logic used in this system).
[2]. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997. (Primary source for the **LSTM temporal sequencing** used for action recognition).
[3]. V. Bazarevsky et al., "BlazePose: On-device Real-time Body Pose Tracking," *arXiv preprint arXiv:2006.10204*, 2020. (Technical basis for the **33 body landmarks** and 3D coordinate system).
[4]. F. Chollet et al., "Keras: Deep Learning for Humans," *GitHub*, 2015. [Online]. Available: https://github.com/fchollet/keras. (Framework used for the **sequential modeling** and training phase).
[5]. G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000. (The library utilized for **webcam activation** and frame preprocessing).
[6]. J. Lin et al., "TSM: Temporal Shift Module for Efficient Video Understanding," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. (Relates to the **sliding window** approach for real-time video inference).
[7]. Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. (Theoretical background for the **Modular Inference** of sign language patterns).
[8]. A. Karpathy et al., "Large-scale Video Classification with Convolutional Neural Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. (Supports the logic of **Action Mode** detection in multi-modal systems).

