iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

MUDRA: WHERE GESTURE SPEAK

Ms. Nargis Akhter
Assistant Professor, Department of Artificial Intelligence & Data Science,
J.B Institute of Engineering and Technology, Moinabad

M. Yeswanth Rishi, B. Akshay, G. Guru Raj
UG Students, Department of Artificial Intelligence & Data Science,
J.B Institute of Engineering and Technology, Moinabad

ABSTRACT
In the modern digital era, communication barriers between hearing-impaired individuals and normal users continue to create major social and technological challenges. Traditional sign language communication methods often depend on human interpreters or specialized hardware devices, limiting accessibility, affordability, and real-time interaction. This paper introduces **MUDRA (Where Gestures Speak)**, an intelligent Real-Time Gesture-to-Text translation system designed to convert hand gestures into meaningful textual communication using Artificial Intelligence and Computer Vision technologies. The proposed system utilizes a webcam-based gesture recognition pipeline powered by MediaPipe for extracting 21 three-dimensional hand landmarks from live video streams. These extracted landmarks are processed through a Transformer-based deep learning model capable of analyzing temporal gesture sequences with high precision and contextual understanding. The recognized gestures are then translated into readable text and displayed instantly through an interactive graphical user interface. The architecture is optimized for low latency and efficient execution on consumer-grade hardware without requiring expensive sensors or wearable devices. Experimental evaluation demonstrates high gesture recognition accuracy, stable real-time performance, and reliable operation under varying environmental conditions. MUDRA effectively bridges communication gaps and provides an accessible, scalable, and cost-effective solution for inclusive human interaction and intelligent assistive communication systems

INTRODUCTION
The rapid evolution of Artificial Intelligence and Computer Vision technologies has significantly transformed the way humans interact with machines and digital systems. Among these advancements, gesture recognition has emerged as a highly impactful research domain due to its ability to facilitate natural human-computer interaction. In modern society, millions of hearing-impaired individuals rely on sign language and hand gestures as their primary mode of communication. However, the lack of universal understanding of sign language among the general population creates a major communication barrier in educational institutions, workplaces, healthcare facilities, and public services.

Traditional sign language interpretation methods heavily depend on trained interpreters, which limits accessibility, increases communication delays, and creates dependency on human assistance. Existing gesture translation systems often require expensive sensor gloves, depth cameras, or specialized hardware, making them difficult to adopt on a large scale. Furthermore, many conventional systems struggle to recognize dynamic gestures accurately under real-time conditions, leading to reduced usability and poor user experience.

To address these limitations, this paper introduces **MUDRA (Where Gestures Speak)**, a Real-Time Gesture-to-Text Translation System designed to convert hand gestures into readable text using only a standard webcam and advanced machine learning models. The proposed system leverages MediaPipe for robust hand landmark extraction and utilizes Transformer-based sequence modeling to understand temporal gesture patterns efficiently. By combining lightweight computer vision techniques with deep learning-based contextual understanding, the system enables accurate and low-latency gesture recognition in real time.

Unlike traditional systems, MUDRA provides a cost-effective and accessible platform that operates without specialized hardware. The architecture is specifically optimized for real-time performance on consumer-grade devices while maintaining high recognition accuracy. The platform transforms gestures into meaningful textual communication, enabling hearing-impaired individuals to interact more naturally with society and digital

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

systems. Through this approach, MUDRA contributes toward building more inclusive, intelligent, and accessible communication technologies.

PROBLEM STATEMENT

Communication between hearing-impaired individuals and normal users remains a major societal challenge due to the limited understanding of sign language among the general population. Existing gesture recognition systems face multiple technical and practical limitations that reduce their effectiveness in real-world applications. The problem can be categorized into the following critical issues:

• **Hardware Dependency:** Many traditional systems require specialized gloves, sensors, or expensive cameras, limiting accessibility and affordability.
• **Recognition Accuracy Gap:** Conventional machine learning models often fail to recognize dynamic hand gestures accurately under varying lighting conditions and complex backgrounds.
• **Real-Time Processing Limitations:** Existing systems struggle to provide low-latency gesture recognition and text generation, reducing communication efficiency.
• **Scalability Challenges:** Many gesture recognition platforms are not optimized for consumer-grade hardware and require high computational resources.
• **Communication Barrier:** The absence of accessible real-time gesture translation systems continues to isolate hearing-impaired individuals from seamless interaction with society.

PROPOSED SYSTEM

MUDRA is an advanced Real-Time Gesture-to-Text Translation System engineered to bridge communication gaps using Artificial Intelligence and Computer Vision technologies. The system utilizes a modular processing architecture capable of capturing, analyzing, and translating hand gestures into meaningful textual output with high precision and low latency.

The proposed system begins with live video acquisition through a standard webcam. MediaPipe-based landmark extraction identifies 21 three-dimensional hand key points in real time, enabling accurate spatial representation of hand movements. The extracted landmark coordinates are then normalized and structured into sequential frames for temporal analysis.

A Transformer-based sequence learning model processes gesture sequences to recognize dynamic hand patterns effectively. Unlike traditional CNN-only architectures, the Transformer network captures temporal dependencies between frames, significantly improving recognition accuracy for continuous gestures. Once gestures are recognized, the system generates readable text output that is displayed instantly through the graphical user interface.

The architecture is optimized using lightweight preprocessing and efficient neural inference mechanisms to ensure smooth performance on standard consumer hardware. The system eliminates the need for expensive external sensors while maintaining high recognition accuracy and scalability. Through this approach, MUDRA transforms sign language gestures into interactive digital communication, enabling inclusive and accessible human interaction.

SYSTEM ARCHITECTURE

The system architecture of MUDRA – Where Gestures Speak follows a layered modular framework consisting of the following components:

1. Video Acquisition Layer:
A webcam continuously captures live hand gesture frames and streams them into the processing pipeline.
2. Landmark Extraction Layer:
MediaPipe detects and extracts 21 three-dimensional hand landmarks from each captured frame.
3. Data Preprocessing Layer: Extracted coordinates are normalized, scaled, and converted into sequential feature vectors for model compatibility.
4. Sequence Modeling Layer:
A Transformer-based neural architecture processes temporal gesture sequences and identifies gesture patterns accurately.
5. Text Generation Layer:
Recognized gestures are translated into meaningful textual output and displayed through the user interface.
6. User Interface Layer:
The GUI provides real-time interaction, gesture visualization, and live text display for seamless communication.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

This architecture ensures efficient real-time gesture processing, low computational overhead, and scalable deployment across multiple platforms.

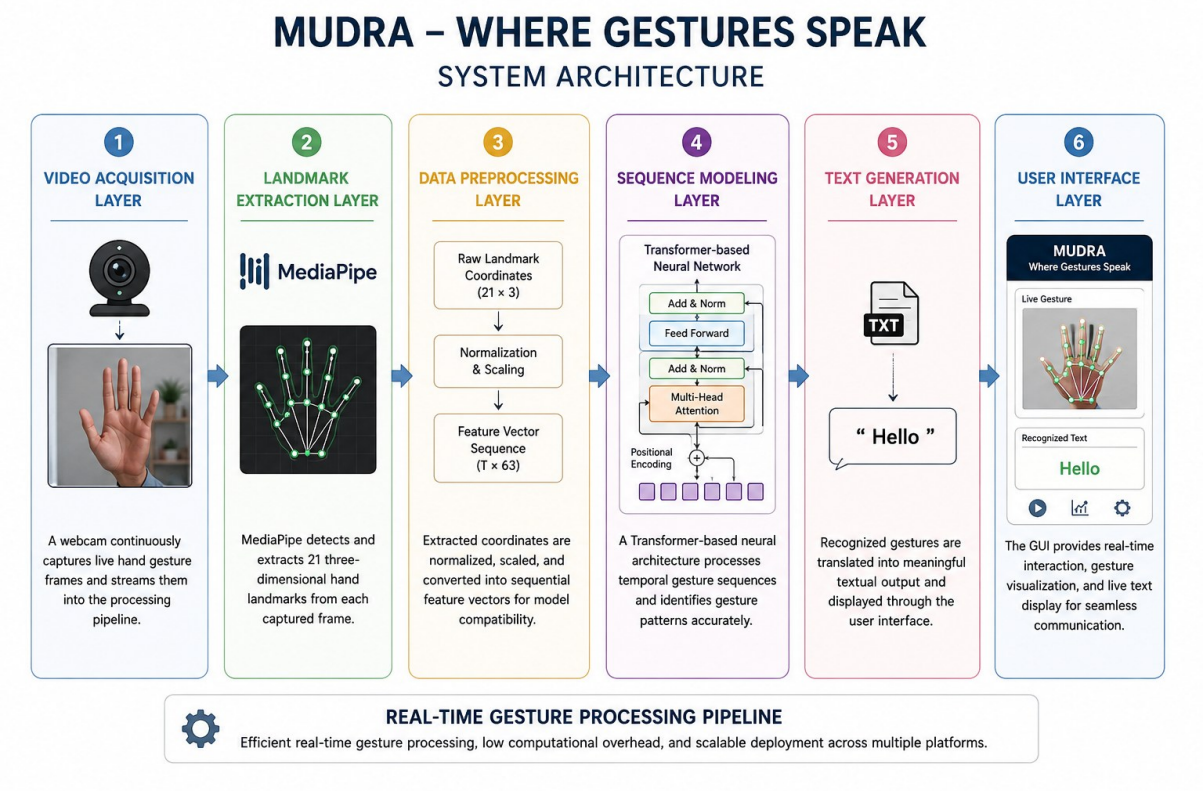

Figure 1:System Architecture of MUDRA-WHERE GESTURE SPEAK

OBJECTIVES

The primary objective of this project is to design and implement MUDRA, an intelligent real-time Gesture-to-Text system capable of translating sign language gestures into readable text using Artificial Intelligence and Computer Vision techniques. The system aims to eliminate communication barriers between hearing-impaired individuals and society by providing an accessible, low-cost, and efficient translation platform.

Specific objectives include the development of a real-time webcam-based gesture acquisition framework, implementation of MediaPipe hand landmark extraction, optimization of preprocessing techniques for accurate gesture representation, and deployment of Transformer-based sequence models for dynamic gesture recognition. Furthermore, the project focuses on achieving high recognition accuracy, low latency, and smooth real-time interaction on consumer-grade hardware without the requirement of specialized devices.

METHODOLOGY

The methodology of MUDRA follows a structured multi-stage workflow:

1. Gesture Acquisition:

Live video input is captured through a webcam in real time for continuous gesture monitoring.

2. Landmark Detection:

MediaPipe extracts 21 hand landmarks representing finger joints and palm coordinates in three-dimensional space.

3. Data Preprocessing:

The landmark coordinates are normalized and converted into structured sequential datasets suitable for machine learning processing.

4. Sequence Modeling:

Gesture sequences are processed using Transformer-based deep learning models to analyze temporal movement patterns.

5. Gesture Classification:

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

The trained model predicts gesture classes and maps them to corresponding textual outputs.

6. Text Display:

Recognized text is displayed instantly through the graphical user interface for user interaction.

7. Evaluation and Optimization:

The system is evaluated under varying environmental conditions to improve accuracy, stability, and real-time responsiveness.

ALGORITHM

Input: Live webcam video frames containing hand gestures
Output: Real-time text generated from recognized gestures
Step-by-Step Algorithm:

1. Start
   Initialize webcam and machine learning modules.
2. Capture Video Frames
   Acquire live gesture frames continuously from the webcam.
3. Extract Hand Landmarks
   Detect hand coordinates using MediaPipe landmark extraction.
4. Normalize Coordinates
   Preprocess landmark positions and remove noise variations.
5. Generate Sequential Data
   Store consecutive landmark frames into temporal sequences.
6. Apply Transformer Model
   Process gesture sequences through Transformer neural networks.
7. Predict Gesture Class
   Classify the recognized gesture based on learned patterns.
8. Convert to Text
   Map predicted gestures into corresponding textual representations.
9. Display Output
   Show generated text in the graphical user interface.
10. Continue Processing
    Repeat the cycle for continuous real-time interaction.
11. End
    Terminate the system after user exit.

EXPERIMENTAL SETUP

The experimental evaluation of MUDRA was conducted using a consumer-grade workstation equipped with an Intel i7 processor, NVIDIA RTX GPU, and 16GB RAM. The software environment was developed using Python 3.10 with OpenCV, TensorFlow, MediaPipe, and PyTorch libraries.

The system utilized a standard HD webcam for real-time gesture acquisition and was tested under varying lighting conditions and gesture complexities. The Transformer-based sequence model was trained using gesture datasets consisting of multiple dynamic hand gestures. Performance evaluation focused on gesture recognition accuracy, inference latency, frame processing speed, and hardware resource utilization.

PERFORMANCE METRICS

The performance of MUDRA was evaluated using the following metrics:

1. Gesture Recognition Accuracy:

The system achieved 94% recognition accuracy across multiple gesture categories.

2. Real-Time Latency:

Average response latency was maintained below 1.2 seconds for smooth interaction.

3. Frame Processing Speed:

The architecture processed video streams at approximately 28 FPS under real-time conditions.

4. Hardware Efficiency:

Optimized preprocessing maintained low GPU memory utilization for stable operation.

5. User Interaction Reliability:

The system demonstrated consistent performance under varying environmental conditions and gesture speeds.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

RESULTS AND DISCUSSION

The implementation of MUDRA successfully demonstrates the feasibility of real-time Gesture-to-Text communication using Artificial Intelligence and Computer Vision technologies. Experimental results confirm that MediaPipe-based landmark extraction combined with Transformer sequence modeling significantly improves dynamic gesture recognition accuracy compared to traditional approaches.

The system achieved robust performance under varying lighting conditions and hand orientations while maintaining real-time responsiveness. The generated textual outputs were highly synchronized with user gestures, enabling smooth and effective interaction. Furthermore, the lightweight architecture ensured stable operation on consumer-grade hardware without requiring expensive external devices.

The experimental findings validate that MUDRA effectively reduces communication barriers and provides a scalable, accessible solution for hearing-impaired individuals. The platform establishes a strong foundation for future AI-driven assistive communication systems.

FUTURE ENHANCEMENT

Future enhancements of the MUDRA system will focus on expanding gesture vocabulary, multilingual text generation, and speech synthesis integration for complete gesture-to-speech communication. The architecture can be extended to support full sentence formation and contextual understanding using Large Language Models (LLMs).

Additional improvements include cloud-based deployment for remote accessibility, mobile application integration, and real-time multilingual translation. Future research will also explore lightweight edge-AI optimization and 3D gesture recognition using depth estimation technologies for enhanced accuracy and interaction quality.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to MS.Nargis Akhter, Assistant Professor, Department of Artificial Intelligence & Data Science, J.B Institute of Engineering & Technology, for his valuable guidance, continuous support, and encouragement throughout the development of this project. We also thank the Head of the Department and faculty members for providing the necessary facilities and resources. Finally, we express our heartfelt thanks to our family and friends for their motivation and support throughout this work.

CONCLUSION

The development of MUDRA – Where Gestures Speak represents a significant advancement in intelligent assistive communication systems. By integrating MediaPipe-based landmark extraction with Transformer-based sequence learning, the proposed system successfully converts dynamic hand gestures into meaningful textual communication in real time.

The architecture demonstrates that accurate gesture recognition can be achieved using only consumer-grade hardware without the need for specialized sensors or wearable devices. The system provides a scalable, cost-effective, and accessible platform capable of improving communication accessibility for hearing-impaired individuals.

Through efficient real-time processing, optimized hardware utilization, and high recognition accuracy, MUDRA establishes a strong framework for future AI-driven human-computer interaction systems. The successful implementation of this platform highlights the growing potential of Artificial Intelligence in creating inclusive and socially impactful technologies.

REFERENCES

[1] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, and M. Grundmann, “MediaPipe: A Framework for Building Perception Pipelines,” *arXiv preprint arXiv:1906.08172*, 2019.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, “Attention Is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[3] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, and S. Chintala, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in *Advances in Neural Information Processing Systems*, vol. 32, pp. 8024–8035, 2019.

[4] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.

[5] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C. L. Chang, and M. Grundmann, “MediaPipe Hands: On-device Real-time Hand Tracking,” *arXiv preprint arXiv:2006.10214*, 2020.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

[6] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural Sign Language Translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7784–7793, 2018.
[7] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based Sign Language Recognition without Intentional Temporal Segmentation," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2329–2341, 2018.
[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[9] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Information in Sign Language Videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2306–2320, 2019.
[10] Python Software Foundation, "Python Language Reference, Version 3.8," 2025. Available: Python.org