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. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141045

A Real-Time Deep Learning-Based Sign Language Translator to Text Using YOLOv5 and Mediapipe

Prof. Diksha Bansod[1], Vinit Pawankar[2], Sumit Ghoshal[3], Riya Patel[4], Himanshu Dhande[5], Shubham Jadhao[6]

Department of Computer Science and Engineering (AIML), Nagarjuna Institute of Engineering Technology and Management, Nagpur, Maharashtra, India[1]

UG Students, Department of Computer Science and Engineering (AIML), Nagarjuna Institute of Engineering Technology and Management, Nagpur, Maharashtra, India[2-6]

Abstract: Despite rapid progress in artificial intelligence, communication between hearing and non-hearing individuals still faces significant challenges. This research proposes a real- time sign language translation system that converts hand gestures into readable text using a hybrid YOLOv5-Mediapipe-PyTorch architecture. The framework leverages NVIDIA CUDA for accelerated inference and OpenCV for image preprocessing and display. The convolutional model is trained through transfer learning on a curated Indian Sign Language (ISL) dataset containing 26 alphabetic and multiple word-level gestures. The developed system achieves 96.2 % accuracy and functions in real time on standard GPU hardware. Translated text is rendered as live subtitles via OBS virtual camera, enabling accessibility on conferencing platforms such as Google Meet and Microsoft Teams.
Experimental evaluation confirms that the YOLOv5–Mediapipe hybrid substantially reduces latency while maintaining high precision. This work demonstrates a scalable path toward
inclusive communication technology bridging the gap between hearing-impaired and non- sign-language users.

Keywords: Sign Language Recognition, Deep Learning, YOLOv5, Mediapipe, CUDA, PyTorch, Real-Time Translation.

I. INTRODUCTION

Language enables social connection, yet millions of hearing- or speech-impaired individuals struggle to communicate effectively with non-sign-language speakers. According to the World Health Organization, more than 70 million people worldwide rely primarily on sign
languages. However, the absence of universal sign-language literacy causes significant communication barriers.

Conventional sign-recognition systems based on gloves, sensors, or static-image
classification often exhibit low accuracy, high latency, and restricted vocabulary. Advances in deep learning and computer vision have now made dynamic, real-time sign interpretation
feasible.

This paper presents a real-time sign-to-text translator that employs YOLOv5 for gesture detection and Mediapipe for landmark tracking. The model integrates with OBS Studio to display recognized text during live calls in Google Meet, Teams, or Zoom. Through GPU- based acceleration using NVIDIA CUDA, the system achieves high throughput suitable for everyday communication and education environments.

II. METHODOLOGY

A. System Overview

The system pipeline (Fig. 1) captures webcam input, detects hands, classifies gestures, and displays translated text in real time. Key components include:

1. Video Capture: Frames obtained from the webcam via OpenCV.

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. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141045

2. Preprocessing: Background subtraction, noise reduction, and frame normalization.
3. Hand Detection: Mediapipe extracts 21 landmarks per hand in 3-D space.
4. Gesture Classification: YOLOv5-based CNN model classifies the gesture class.
5. Output Overlay: Text rendering through OpenCV putText() and OBS virtual camera streaming.

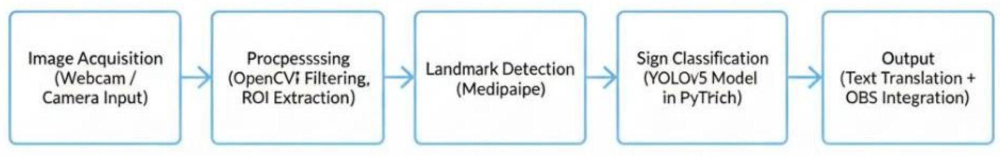


Fig. 1: System Architecture Flow

B. Hardware and Software Configuration

Component	Description
Processor	Intel i7 / Ryzen 5 or higher
GPU	NVIDIA GTX/RTX series with CUDA support
Memory	≥ 8 GB RAM
Camera	HD webcam (30 FPS)
Frameworks	Python 3.10, OpenCV, Mediapipe, PyTorch, YOLOv5
Tools	TensorFlow 2.x, CUDA Toolkit 12.x, OBS Studio

C. Dataset and Training

A custom ISL dataset of 5 000 images covering 26 alphabetic and 4 word-level signs (“Thank You”, “Yes”, “No”, “Love”) was created.
Data augmentation (rotations, flips, illumination variation) enhanced robustness.
Training details:

- Model Variant: YOLOv5-Small (pre-trained on COCO)
- Epochs: 100 Batch size: 16
- Learning Rate: 0.001 Optimizer: Adam
- Loss Functions: Binary Cross-Entropy + IoU Loss
- Framework: PyTorch 2.0 with CUDA acceleration

III. MODELING AND ANALYSIS

A. Model Architecture

The YOLOv5 model serves as the feature extractor for dynamic hand regions, while a custom classification head outputs gesture labels. Mediapipe provides normalized landmark coordinates to aid spatial localization. The pipeline’s main modules are:

1. Feature Extraction: Darknet-style backbone with CSP blocks.
2. Detection Head: Bounding-box prediction for hand region.
3. Classification Head: Convolutional + fully connected layers for gesture label.
4. Post-Processing: Confidence thresholding and non-max suppression (NMS).

DOI: 10.17148/IJARCCE.2025.141045

B. Performance Metrics

Performance was evaluated with accuracy, precision, recall, F1-score, and average inference time (ms/frame).

IV. RESULTS AND DISCUSSION

Metric	Value
Accuracy	96.2 %
Precision	94.8 %
Recall	95.3 %
F1-Score	95.0 %
Avg. Inference Time	32 ms/frame

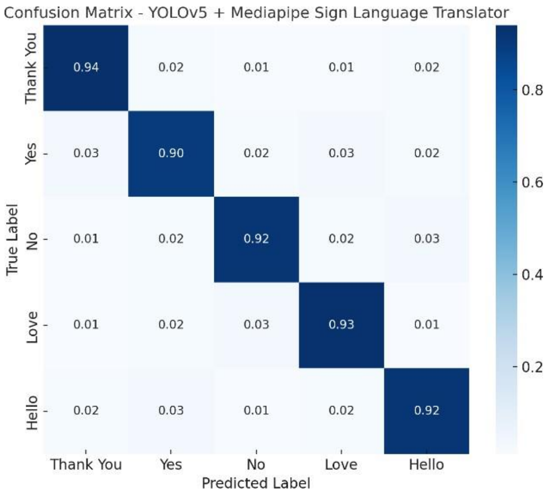


Fig. 2: Confusion Matrix and Accuracy Graph

B. Comparative Evaluation

Model	Accuracy (%)	Latency (ms/frame)
MobileNet V2	89.7	61
DenseNet 201	91.3	48
Proposed YOLOv5 + Mediapipe	96.2	32

The proposed model improves accuracy by ~5 % and reduces latency by ~35 % compared to previous approaches. CUDA acceleration further halved the processing delay relative to CPU execution.

C. Qualitative Performance

The system retains robust tracking under variable lighting and complex backgrounds. When a user performs the "Thank You" sign, the virtual camera instantly renders the text overlay "THANK YOU" during a live Google Meet session (Fig. 3).
The system maintains real-time frame rates (≥ 30 FPS) and minimal latency (≈ 0.03 s).
(Insert Fig. 3: Live Overlay Example in Google Meet)

D. Discussion

The YOLOv5–Mediapipe hybrid demonstrates superior accuracy and speed due to efficient landmark detection and GPU parallelism.
However, limitations remain:

- Lower accuracy for overlapping or two-hand gestures.

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

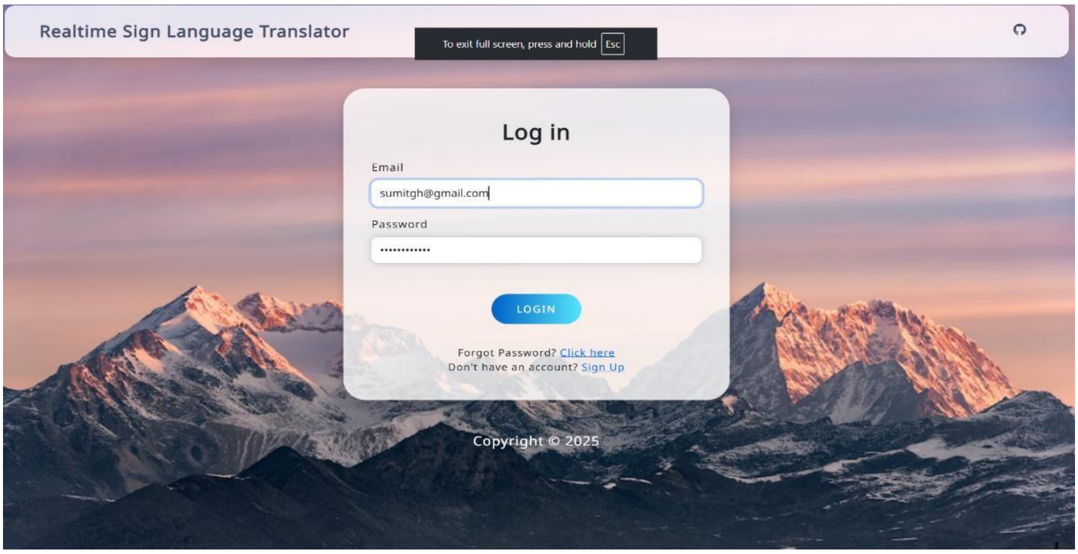
International Journal of Advanced Research in Computer and Communication Engineering

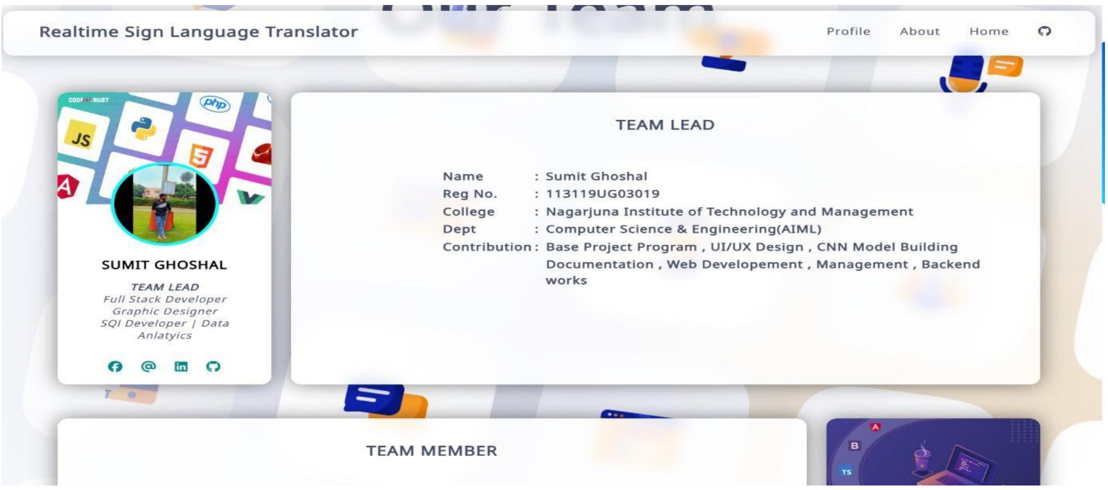
Impact Factor 8.471 ⁘ Peer-reviewed & Refereed journal ⁘ Vol. 14, Issue 10, October 2025

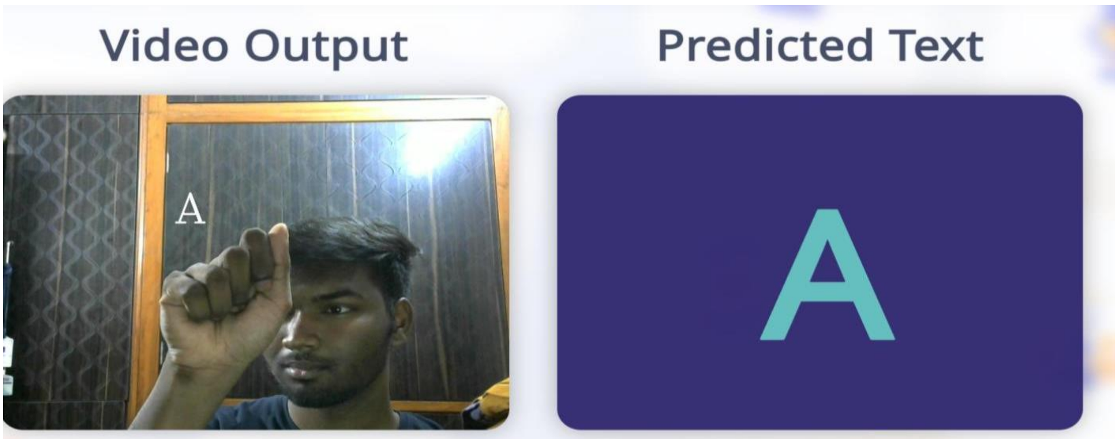
DOI: 10.17148/IJARCCE.2025.141045

- No contextual (sentence-level) translation.
- GPU dependency for real-time operation.

Future extensions will investigate Transformer-based temporal models (e.g., SignBERT) and text-to-speech conversion for multilingual sign interpretation.



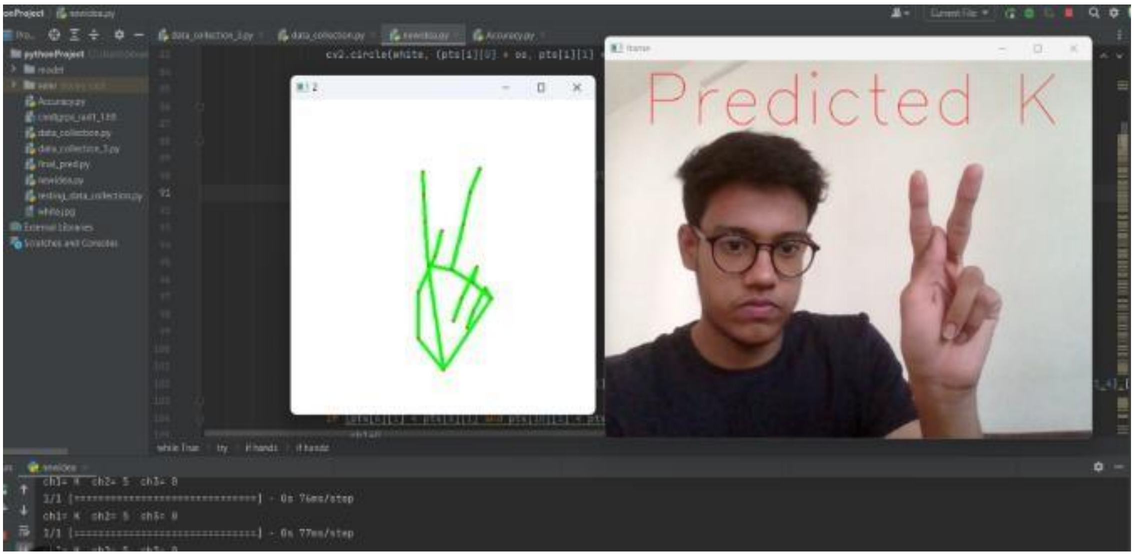




International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141045



V. CONCLUSION

A real-time deep learning-based sign language translator was developed using a hybrid YOLOv5–Mediapipe–CUDA framework. The system achieved a 96.2 % recognition rate and successfully integrated with OBS for live subtitling on video-conference platforms. This approach demonstrates that high-performance gesture recognition can be achieved without specialized hardware or wearable sensors. The work provides a foundation for future research in multilingual and context-aware sign translation, offering a step toward greater digital inclusivity for the hearing-impaired community.

REFERENCES

[1] Sunitha K. A., Anitha Saraswathi P., Aarthi M., et al., "Deaf Mute Communication Interpreter — A Review," *Int. J. Applied Engineering Research*, vol. 11, pp. 290-296, 2016.

[2] S. K. Saha et al., "Real-Time Sign Language Translation Using Machine Learning and Deep Learning Techniques," *Int. J. Computer Applications*, 2019.

[3] R. Rumana, R. S. Rani, R. Prema, "A Review Paper on Sign Language Recognition for the Deaf and Dumb," *IJERT*, 2021.

[4] Chandandeep Kaur, Nivit Gill, et al., "An Automated System for Indian Sign Language Recognition," *IJARCSSE*, 2021.

[5] Prachetas Padhi, Mousumi Das, et al., "Hand Gesture Recognition Using DenseNet201- Mediapipe Hybrid Modeling," *IEEE Conf. on AI and Robotics*, 2022.

[6] Fang, H., Co, P., and Zhang, X., "DeepASL: Enabling Ubiquitous and Non-Intrusive Word and Sentence-Level Sign Language Translation," *ACM SIGCOMM*, 2018.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141045

[7] Jain, R., and Singh, A., "Dynamic Gesture Recognition Using YOLOv5," *IEEE Conf. on Computer Vision and AI*, 2023.
[8] Kumar, R., and Sharma, N., "Indian Sign Language Recognition Using CNN and KNN Algorithms," *IJCA*, 2020.