Real-Time 3D Hand Pose-Based ASL Recognition: A Computer Vision System Using Enhanced Landmark Tracking and Attention Models

Raj Lucka, Northeastern University

# Abstract

We present a real-time American Sign Language (ASL) gesture recognition system that combines MediaPipe-based hand tracking with a deep learning architecture integrating convolutional, recurrent, and attention-based components. Our model accurately classifies eight distinct ASL gestures from hand landmark sequences. Through rigorous preprocessing and augmentation techniques, we achieve a validation accuracy of 96.8% and real-world performance of 94.2%. The system demonstrates low latency and high stability, making it well-suited for accessibility applications. This work highlights how computer vision techniques can support interactive communication systems in real time.

# 1. Introduction

Sign language recognition plays a pivotal role in enhancing communication accessibility for the Deaf and hard-of-hearing communities. This paper presents a real-time ASL detection system that leverages computer vision, deep learning, and landmark-based gesture modeling. Unlike image-based classification approaches, we use a skeletal representation of hand landmarks to model gestures as temporal sequences. This enables robust and efficient recognition across users and conditions. Our contribution focuses on real-time gesture classification using three-dimensional hand pose estimation, which remains a central challenge in computer vision.

# 2. Related Work

[1] Zhang et al. (2020), "HandGestureNet: A Real-Time Hand Gesture Recognition Network" (CVPR) proposed an end-to-end model for real-time hand gesture classification from RGB data. We extend this by using 3D landmarks instead of RGB frames.

[2] Molchanov et al. (2016), "Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks" (CVPR) introduced temporal modeling using 3D CNNs and LSTM. Our system similarly incorporates LSTM for temporal learning but operates on lightweight landmark data.

[3] Yang et al. (2020), "Deephand: Robust Hand Gesture Recognition Using 3D Pose Normalized Skeleton" (ECCV) explored 3D skeleton-based hand recognition. Our system also uses normalized skeletons and enhances discriminability with an attention mechanism.

# 3. Methods

**3.1 Data Collection and Preprocessing**

Our data collection experiments were carefully structured to capture the nuanced movements of ASL signs across different users and environments. We developed a specialized capture interface that guided participants through a systematic protocol of performing each gesture 30 times in succession, with each recording lasting exactly 30 frames (approximately one second). This design ensured temporal consistency. The interface displayed real-time feedback showing the MediaPipe hand tracking overlay alongside frame counts and sequence numbers, helping participants maintain consistent hand positions and timing.

To evaluate the robustness of the MediaPipe hand tracking system, which is a key component of our computer vision pipeline, we experimented with various camera angles, distances (0.5 to 1.0 meters), and lighting conditions. Diffuse, moderate lighting placed at a 45-degree angle consistently produced optimal tracking with minimal occlusion. We also experimented with different backdrop colors, discovering that neutral, matte surfaces significantly improved the reliability of depth estimation.

To further improve robustness under varied lighting conditions, we integrated an advanced image enhancement pipeline. This pipeline begins by converting input frames to the YUV color space to separate luminance from chrominance. We then apply histogram equalization to enhance global contrast and use Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve local contrast. Additionally, adjustable brightness and contrast controls are provided through an interactive trackbar system, allowing users to tune parameters in real time. We also implemented an auto-adaptation module that dynamically adjusts enhancement settings based on live brightness and contrast metrics.

To promote dataset diversity, we recorded gestures with varying hand orientations (within ±15 degrees of the frontal view), speeds (deliberate slow and fast motions), and environmental conditions. Sequences were only retained if they met strict quality criteria: at least 70 percent of frames containing valid landmarks and an average landmark confidence score above 0.7. This rigorous filtering ensured that the data provided a strong foundation for downstream computer vision and machine learning tasks.

**3.2 Data Augmentation**

To improve generalization, we applied rotation, scaling, translation, and Gaussian noise with controlled probabilities. This expanded the dataset to 720 sequences. Ablation studies confirmed a 12.3 percent improvement in validation performance due to augmentation.

**3.3 Model Architecture**

Our model consisted of a Conv1D layer (64 filters), LSTM layer (128 units), and a self-attention mechanism. The output was flattened and passed to a dense layer and a softmax classifier. Residual connections helped retain sequence context. We trained the model using the Adam optimizer and categorical cross-entropy over 80 epochs.

**3.4 Real-Time Inference**

Real-time performance was achieved using asynchronous threading and prediction smoothing over an 8-frame window. A state machine with cooldown logic stabilized output transitions and improved prediction consistency.

# 4. Experiments and Results

Validation accuracy reached 96.8 percent, with the gesture “hello” achieving 99.2 percent precision. The lowest performance was observed in the “neutral” class, which achieved 92.4 percent precision. Real-world testing with five users yielded an average accuracy of 94.2 percent. The system exhibited an average latency of 0.47 seconds. Ablation studies demonstrated that the attention mechanism contributed a 6.2 percent increase in accuracy, and that data augmentation significantly improved the model's robustness.

# 5. Discussion and Summary

This project demonstrates that landmark-based, attention-enhanced temporal modeling is an effective approach for real-time ASL recognition. Key contributions include robust preprocessing, a lightweight inference pipeline, and high recognition accuracy in diverse, practical settings. The experiments reinforce the importance of high-quality computer vision techniques, particularly those focused on landmark detection, tracking, and normalization, as foundational elements for successful deep learning applications. Future work may explore vocabulary expansion, integration of non-manual cues such as facial expressions, and optimization for mobile deployment.

# 6. References

[1] Zhang, M., Xu, H., & Wang, L. (2020). HandGestureNet: Real-Time Hand Gesture Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[2] Molchanov, P., Gupta, S., Kim, K., & Kautz, J. (2016). Online Detection and Classification of Dynamic Hand Gestures. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[3] Yang, J., Liang, H., Li, Y., & Chen, Y. (2020). Deephand: Robust Hand Gesture Recognition Using 3D Pose Normalized Skeleton. European Conference on Computer Vision (ECCV).