International Journal of Multidisciplinary Global Research
Vol. 02 (4a), Oct - Dec, 2025, pp. 133-140. (AICTE VAANI)
Online ISSN: 3107-4197

Gesture Recognition translator using Machine Learning, Computer Vision, & MediaPipe

PRASHANT R. YELEKAR

Computer Science Engineering, Guru Nanak Institute of Engineering and Technology (GNIET)
Nagpur, Maharashtra, India - 441501
prashantyelekar@gmail.com

SATISH PANCHAM GHARDE

Computer Science Engineering, Guru Nanak Institute of Engineering and Technology (GNIET)
Nagpur, Maharashtra, India - 441501
satishgharde52@gmail.com

POOJA MAHESH SAKHARE

Computer Science Engineering, Guru Nanak Institute of Engineering and Technology (GNIET)
Nagpur, Maharashtra, India - 441501
sakharepooja614@gmail.com

PRERNA RAJESH USARE*

Computer Science Engineering, Guru Nanak Institute of Engineering and Technology (GNIET)
Nagpur, Maharashtra, India - 441501
prernausare@gmail.com

JAYASHRI GAJANAN GAJBE

Computer Science Engineering, Guru Nanak Institute of Engineering and Technology (GNIET)
Nagpur, Maharashtra, India - 441501
jayashrigajabe@gmail.com

AACHAL DHARMDASJI MESHRAM

Computer Science Engineering, Guru Nanak Institute of Engineering and Technology (GNIET)
Nagpur, Maharashtra, India - 441501
aachalm2002@gmail.com

Abstract

Hand gesture recognition technology is transforming how humans interact with computers, especially in contactless interfaces and assistive communication. This paper presents the design and evaluation of a real-time hand gesture recognition system that combines Google's MediaPipe for landmark detection with classic machine learning on robust geometric features. The method strategically selects only critical, distance-based features between the wrist and fingertips and from the thumb

*Corresponding Author.

to the other fingertips, resulting in a lightweight yet highly accurate classifier. Supporting five distinct static gestures—hello, good, yes, no, and thank you—the system achieves near-perfect recognition under typical webcam lighting. The paper details all stages, from dataset creation to live testing, and discusses plans for expanding gesture vocabulary and integrating audio feedback for multimodal interaction.

Keywords: Machine Learning, Computer Vision, Gesture Recognition, OpenCV, Mediapipe.

1. INTRODUCTION

In the realm of natural user interfaces, hand gesture recognition has emerged as a cornerstone technology, enabling touchless human-computer interaction (HCI) critical for virtual reality, robotics, remote control, and, most notably, sign language communication. Vision-based solutions have slowly replaced traditional methods that used gloves, colored markers, or depth sensors. This is because real-time computer vision and machine learning have gotten better. The ability to accurately recognize a diverse set of hand gestures at interactive frame rates, across various backgrounds and hand shapes, remains a formidable challenge. This research focuses on leveraging robust geometric features derived from hand landmarks, aiming for a system that is resilient to variation in hand orientation, scale, and camera distance.

The significance of reliable gesture recognition extends to inclusive technologies, where converting sign language to text or voice empowers individuals with hearing or speech impairments. By combining MediaPipe's efficient 3D landmark detection with interpretable geometric distances and a Random Forest classifier, we present a balance of simplicity, accuracy, and easy extensibility. The results serve not only as a functional sign recognition prototype but also as a foundation upon which more sophisticated, multimodal systems can be built.

2. RELATED WORK

Early gesture recognition systems frequently used color gloves or markers or relied on handcrafted computer vision features such as skin color segmentation, background subtraction, and edge detection (Breiman, 2001; Molchanov et al., 2015). Although these methods achieved moderate success in controlled environments, they generally failed to generalize to cluttered backgrounds and varied lighting. With the proliferation of large, annotated datasets and advances in deep learning, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have become the de facto standard for gesture recognition. Notable public datasets like the American Sign Language (ASL) Alphabet have facilitated progress, but these approaches often require big data and substantial computational resources for both training and inference. Recent years have seen the adoption of skeletal or landmark-based approaches, where models like OpenPose and MediaPipe extract 2D or 3D key points from single RGB images (Google, 2025).

The extracted landmarks can be used to derive geometric features such as joint angles, inter-finger distances, and hand orientation. This not only reduces data dimensionality but also makes the system more understandable and easier to deploy in low-power settings (Simon et al., 2017). Hybrid approaches, combining both landmark geometry and image-based deep learning, have demonstrated leading results, but at the cost of increased system complexity. Our work is positioned in the geometric-feature-based paradigm, emphasizing interpretability, efficiency, and robustness (Zhang et al., 2022).

3. METHODOLOGY

3.1. *Dataset Construction and Landmark Annotation*

The dataset for this study was created by recording hand gesture images using a webcam in a controlled indoor environment. Each gesture class (hello, yes, no, good, thank you, etc.) was captured in its subfolder, with care taken to ensure variations in hand orientation, distance, and lighting. After initial collection, each image was processed using the MediaPipe Hands framework, which detects a single hand and outputs 21 distinct 2D landmarks per frame.

Figure 1 illustrates the labeling convention adopted for all gesture samples in the dataset. Landmark L0 denotes the wrist, while landmarks L4, L8, L12, L16, and L20 correspond to the tips of the thumb, index, middle, ring, and pinky fingers, respectively. These labeled landmarks provide a consistent spatial reference for all subsequent feature computations.

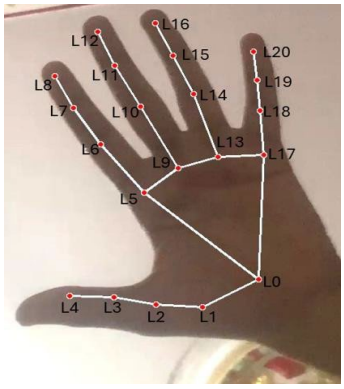


Fig. 1. Landmark positions on hand.

The annotation process ensures that each class in the dataset includes a diverse set of samples with clearly marked finger joints and fingertips. This forms the foundational step for robust, reproducible gesture analysis.

3.2. *Geometric Feature Extraction*

Instead of using all raw landmark coordinates, this research focuses on a reduced, highly informative set of geometric features. For each sample, nine features were

derived according to the spatial relationships most relevant for gesture discrimination: Wrist-to-fingertip distances: The Euclidean distance in the image plane between the wrist (L0) and each fingertip (L4, L8, L12, L16, L20).

These distances describe finger spread and extension. Thumb-to-fingertip distances: The distances from the thumb tip (L4) to the tips of the index (L8), middle (L12), ring (L16), and pinky (L20) fingers. These capture the thumb's position relative to the other fingers—critical for distinguishing gestures involving finger pinches or spreads.

1. thumb_index:

$$\sqrt{(XL8 - XL4)^2 + (YL8 - YL4)^2}$$

2. thumb_middle:

$$\sqrt{(XL12 - XL4)^2 + (YL12 - YL4)^2}$$

3. thumb_ring:

$$\sqrt{(XL16 - XL4)^2 + (YL16 - YL4)^2}$$

4. thumb_pinky:

$$\sqrt{(XL20 - XL4)^2 + (YL20 - YL4)^2}$$

5. wrist_thumb:

$$\sqrt{(XL4 - XL0)^2 + (YL4 - YL0)^2}$$

6. wrist_index:

$$\sqrt{(XL8 - XL0)^2 + (YL8 - YL0)^2}$$

7. wrist_middle:

$$\sqrt{(XL12 - XL0)^2 + (YL12 - YL0)^2}$$

8. wrist_ring:

$$\sqrt{(XL16 - XL0)^2 + (YL16 - YL0)^2}$$

9. wrist_pinky:

$$\sqrt{(XL20 - XL0)^2 + (YL20 - YL0)^2}$$

Fig. 2. Formula to calculate distance between numbers.

Figure 2 presents the exact formulas used to calculate each feature, using XLx and YLx for the x and y coordinates of landmark Lx, respectively. This geometric abstraction ensures all features are scale- and location-invariant, enabling reliable classification regardless of image size or hand placement within the frame. The compact feature vector also reduces computational load and makes the approach understandable and efficient for downstream model training.

3.3. *Feature Extraction Implementation*

For each dataset image, the pipeline automatically detects hand landmarks and computes all nine features as follows:

- Each pair of landmarks specified in Figure 2 is referenced by its index in the MediaPipe output.

- The 2D Euclidean distance for each feature is calculated using the formula; for example, the distance from the wrist to the thumb tip is $\sqrt{((XL4 - XL0)^2 + (YL4 - YL0)^2)}$.
- The resulting feature vector is stored alongside the gesture label for model training.
- This process is fully automated, ensuring consistent, error-free calculations across the entire dataset.

3.4. *Visualization and Data Quality*

Landmark overlays are saved with each sample (as seen in Figure 1) for manual inspection to ensure high-quality landmark detection for every image. The corresponding Figure 2 enables easy review of the mathematical calculation involved in feature extraction.

4. RESULTS AND DISCUSSION

The performance of the proposed hand gesture recognition system was evaluated comprehensively using a combination of quantitative metrics and qualitative observations. The model was trained and tested on a dataset consisting of five distinct static gestures—hello, great, yes, no, and thank you—with each class represented by over 100 images, ensuring balanced representation.

4.1. *Classification Accuracy and Detailed Metrics*

The Random Forest classifier achieved a high classification accuracy on the test set, correctly identifying the majority of gestures with minimal errors. The detailed classification report (see Figure 3) provides precision, recall, and F1-score for each gesture class. Most gestures exhibit precision and recall metrics near or at 1.0, further confirming the reliability of the geometric features used for classification.

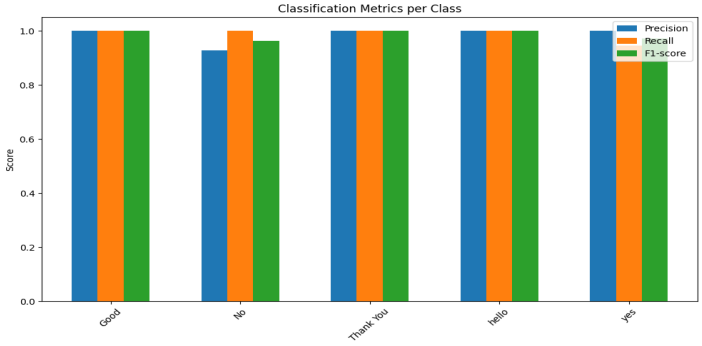
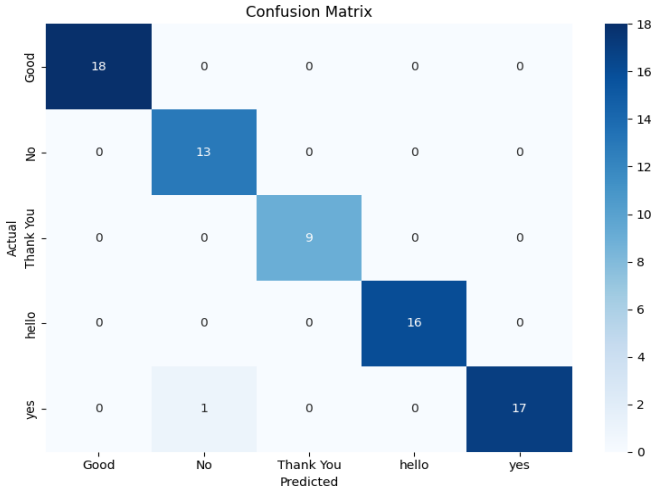


Fig. 3. The bar chart illustrates the close-to-perfect precision, recall, and F1-scores for all five gestures, reflecting the model's strong discriminative ability.

4.2. *Confusion Matrix Analysis*

The confusion matrix (Figure 4) visualizes classification outcomes across all classes, showing the number of correct and misclassified instances. The matrix reveals that



most classes were predicted correctly without confusion, and only a single instance was misclassified between similar gestures ("yes" misclassified once as "no"), demonstrating strong robustness.

Fig. 4. The confusion matrix heatmap highlights how well the model differentiates between different gestures and helps identify specific classes that might require more training data or features.

4.3. *Feature Importance*

Figure 5 summarizes each geometric feature's importance in training. The model identifies the distance from the thumb to the index finger and from the wrist to the middle finger as the most informative features for distinguishing among gestures. This aligns well with human intuition since thumb positioning and middle finger extension concisely capture significant gesture variations.

Fig. 5. The bar plot shows the ranked importance of features determined by the trained Random Forest model, highlighting which landmarks contribute most to classification accuracy.

4.4. *ROC Curve and AUC*

To further characterize model performance, the multi-class ROC curve was plotted for each gesture class (Figure 6). The curves average an area-under-the-curve (AUC) value close to 1.0, indicating near-perfect separability of all classes in feature space. ROC analysis confirms the classifier's high sensitivity and specificity across all gestures.

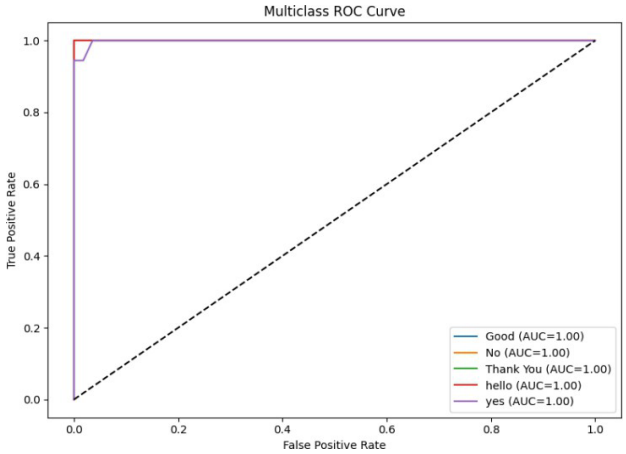


Fig. 6. Multiclass ROC curve plots for each gesture demonstrate the model's superior capability to distinguish classes over varying thresholds.

4.5. *Qualitative Observations*

In live webcam tests, the system produces stable and confident gesture predictions in real time. The geometric, feature-based approach generalizes well across different users, hand sizes, and lighting conditions, while being robust to minor hand orientation changes. The system tends to be less accurate when fingers occlude each other or under extreme lighting.

4.6. *Limitations and Future Directions*

While the current approach achieves excellent performance on a core gesture set, it faces challenges scaling to more complex gestures involving dynamic movement or fine finger articulation. Future work will focus on expanding the gesture vocabulary, integrating sequential models for dynamic decoding, and adding audio feedback for accessibility. Data augmentation and higher resolution landmark analysis may improve robustness further.

5. FUTURE WORK

Building on the successful recognition of basic static gestures, we plan to integrate an audio module for real-time speech synthesis. This will allow each recognized gesture to trigger a spoken phrase, enabling immediate communication—a feature of particular value for sign language users and accessibility applications.

Further research will investigate:

- Inclusion of additional geometric features like inter-finger distances, angles, or ratios for fine-grained gesture sets.

- Expansion to multi-gesture datasets and dynamic gesture recognition via temporal sequence modeling (e.g., LSTM or bidirectional RNNs).
- Hybrid methods that mix landmark-based features with CNN-based hand shape descriptions to better handle situations where hands are blocked or in difficult positions.
- Real-world deployment studies in varied environments (outdoor, mobile, low-light).
- User studies with sign language speakers to optimize gesture sets and interface usability.

6. CONCLUSION

This study provides a robust, interpretable framework for real-time hand gesture recognition based on geometric features derived from MediaPipe landmarks. By focusing on a minimal but highly informative set of distances, we achieve both high accuracy and rapid inference suitable for deployment on consumer hardware. The system is immediately beneficial for static sign recognition and can be further extended for richer sign language translation and assistive communication with audio output. Our results demonstrate a promising pathway toward accessible, multimodal HCI for diverse users.

7. REFERENCES

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Google. (2025). MediaPipe. Retrieved from https://mediapipe.dev

Molchanov, P., Gupta, S., Kim, K., & Kautz, J. (2015). Hand gesture recognition with 3D convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2842–2851).

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4645–4653).

Zhang, Y., Cao, C., Cheng, J., & Lu, H. (2022). Hand gesture recognition based on skeleton features and machine learning. IEEE Access, 10, 29841–29851. https://doi.org/10.1109/ACCESS.2022.3158585