INTERNATIONAL JOURNAL OF
RECENT TRENDS IN MULTIDISCIPLINARY RESEARCH
https://www.doi.org/10.59256/ijrtmr.20250506025
November-December 2025, Vol 5(06), 195-199.
ISSN No: 2583-0368

Sign Language Detection in Video and Real-Time Video Calls

VM Saravana Perumal[1], Manjunath Suresh Patil[2], Nikhil S H[3], Mallinath[4], Rajath K N[5]

[1]*Professor, Department of Computer Science and Engineering, Raja Rajeswari College of Engineering, Bengaluru, Karnataka, India.*

[2, 3, 4,5]*Department of Computer Science and Engineering Raja Rajeswari College of Engineering, Bengaluru, Karnataka, India.*

OPEN ACCESS
Article Citation:
VM Saravana Perumal[1], Manjunath Suresh Patil[2], Nikhil S H[3], Mallinath[4], Rajath K N[5], "Sign Language Detection in Video and Real-Time Video Calls", International Journal of Recent Trends in Multidisciplinary Research, November-December 2025, Vol 5(06), 195-199.

Abstract: Real-time communication between signers and non-signers remains one of the biggest technological challenges. Most of the current systems use raw video frames, which are prone to a number of factors such as variation in lighting conditions, camera stability, and background clutter and image quality. This paper proposes a novel landmark-based sign language detection system that is designed for both videos and real-time video calls and provides a robust and efficient option against conventional image-based methods. The system integrates Media Pipe land- mark extraction, CNN-LSTM temporal modeling and WebRTC communication, effortlessly decoding both static and dynamic gestures with high accuracy. Representing gestures as numer- ical coordinate sequences rather than raw images, the model shows tremendous adaptability in heterogeneous environments and demonstrates smooth performance during live interactions. Experimental evaluations validate that the proposed architecture ensures reliable detection accuracy, low latency and smooth frame rates, which makes it suitable for realistic communication scenarios.

Keywords: Sign language recognition, real-time processing, gesture detection, MediaPipe landmarks, CNN-LSTM, WebRTC communication.

1. Introduction

The broad utilization of digital communication technologies has greatly influenced the ways in which people connect, collaborate, and acquire information. However, persons relying on sign languages still suffer from many barriers to participate in conversations online through Zoom, Google Meet, or What- sApp video calls. The barriers are partly based on the fact that sign language demands interpretation of complex hand shapes, finger arrangements, and patterns of fluent motion. Human translators are very effective but rarely available during spon- taneous or personal conversations; therefore, this accessibility gap continues to persist.

Traditional research in sign language recognition has fo- cused on pixel-based image processing that was highly susceptible to environmental inconsistencies. Changes in skin tone, illumination, occlusions and camera resolution easily affect recognition accuracies. Moreover, the static image-based models poorly capture the dynamism of sign language, where meaning often emerges from the motion trajectory rather than just one frame.

The recent advancements in human-pose estimation tech- nologies have enabled new frontiers in gesture understanding. MediaPipe, in particular, is capable of tracking skeletal joints and hand landmarks with high precision in real time. Instead of dealing with large image datasets, MediaPipe provides nu- merical coordinates that are structured for every frame, hence making representations of gestures consistent and independent from visual noise.

In this work, the authors present a real-time sign language detection framework that effectively integrates landmark extraction with deep sequential learning. The system is designed to be effective for both offline video processing and live video call scenarios. By combining the strengths of Medi- aPipe, CNN–LSTM modeling, and WebRTC for low-latency streaming, the system attempts to bridge the communication gap between signers and non-signers in a practical and scalable manner.

2. Literature Survey

Research in automated sign language recognition has reached a high degree of sophistication with rapid advances in computer vision and deep learning. Initial work focused on the development of large-scale recognition pipelines for regional sign languages. Sharma et al. [1] presented the first deep learning-based Indian Sign Language (ISL) recognition, demonstrating the ability of CNN architectures to extract discriminative spatial features from hand gestures effectively. Other related works regarding static hand gesture classifica- tions further established the appropriateness of convolutional neural networks for image-based recognition tasks, pointing out that their accuracies were well improved compared to traditional handcrafted feature methods [2].

Fundamental deep learning developments further shaped the sign language recognition architectures. Chollet presented the availability of accessible and efficient deep learning frame- works, such as Keras, allowing rapid prototyping of neural models for gesture recognition [3]. Introducing Long Short- Term Memory (LSTM) networks by Graves presented an effective mechanism for modeling temporal dependencies in sequential visual gestures that could recognize dynamic signs over several frames [4]. Hinton's work on deep neural net- works emphasized hierarchical feature learning and established the basis for end-to-end recognition pipelines widely used in gesture understanding [5].

Vision-based gesture analysis has also evolved into appli- cations in real time. Nagi et al. reported early contributions toward real-time hand gesture recognition, addressing the challenges of illumination variation, complex background, and motion blur [6]. With increased computational powers, Kim and Byun adopted 3D CNNs that jointly learn spatiotemporal features, thereby attaining better performance in real-time hu- mancomputer interaction systems [7]. Zhang et al. showed that combining CNN features with LSTM architectures enhances dynamic gesture classification and thus allows for continuous recognition in video sequences [8].

Recent progress in real-time sign language detection has been accelerated by the development of pose tracking and lightweight deep learning frameworks. MediaPipe, newly in- troduced by Google Research, provides cross-platform hand pose estimation for rapid landmark extraction suitable for live video call environments [9]. Optimization techniques like the Adam optimizer have continued to support stable and efficient model training, proposed by Kingma and Ba, improving convergence in sign recognition networks [10]. Sridhar and Lu further extended deep learning-based gesture recognition to real-time vision systems, proving scalability for real-world deployment [11].

Historically, sign language recognition systems evolved from early studies of video-based ASL. Starner presented one of the first real-time ASL recognition systems using wearable computing, emphasizing temporal modeling of ges- tures in continuous video streams [12]. Kelly et al. later introduced person-independent gesture recognition, highlight- ing the importance of generalization across different signers and environments [13]. Li et al. expanded spatiotemporal action recognition using 3D point representations, influencing modern skeleton and pose-based gesture detection approaches [14]. The existing literature shows promising advances in deep learning architectures, temporal modeling, real-time pose extraction, and system scalability. However, most studies deal with a controlled environment or static background, or use pre-recorded datasets. On the other hand, the real-time sign language detection among live video calls introduces extra challenges such as fluctuating network quality, interaction between multiple participants, different camera angles, and limited frame resolution. Such constraints need lightweight, efficient, and robust models that can detect and interpret efficiently within dynamic communication platforms, an area which still continues as an active research opportunity.

3. Methodology

The approach taken in this study is based on a well- structured pipeline: dataset creation, landmark extraction, pre-processing, model development, evaluation, and real-time de- ployment within a WebRTC-enabled communication frame-work. Every stage of the pipeline was deliberately designed in such a manner to ensure system robustness, scalability, and re-sponsiveness, especially in real time during human interaction scenarios. The broader aim behind this kind of methodological construction is to develop a model that generalizes well across multiple users with different execution speeds, environmental backgrounds, camera positions, and lighting conditions-all common factors during regular video communications.

The developed dataset in this work consists of video record- ings of a predefined vocabulary that includes alphabets, com-monly used conversational words, and dynamic sentence-level gestures. All videos were captured using the same consumer-grade webcam to mimic real-world human-computer interac- tion environments like online meetings or video calls. Subjects were asked to perform the gestures without constriction in any way concerning background and clothing to make sure that the dataset consists of variations. Videos have been recorded with a specified frame rate and resolution to minimize domain shift by ensuring continuity. This realistic yet controlled acquisition setup forms the proper basis for the development of sign language systems focused on being deployed in the real world. After data collection, landmark extraction was conducted for each frame using the hand-tracking and full-body pose esti- mation modules of MediaPipe. Twenty-one three-dimensional hand keypoints and thirty-three body pose keypoints were extracted for every frame, totaling fifty-four landmark co- ordinates per frame. The skeletal representation was pre- ferred because of computational efficiency and its robustness to visual variations, such as illumination changes, cluttered backgrounds, and moderate camera movement. Landmarks abstract away raw pixel information; thus, this representation is compact and invariant for human gestures, allowing faster model inferences for real-time systems.

First, after landmark extraction, multiple preprocessing op- erations were performed in order to improve the consistency of data and model learnability. In particular, all landmark coordinates were normalized with respect to reference joints, mainly shoulders and torso, to compensate for scale dif- ferences due to the distance between the subject and the camera. Centering the coordinates was followed by scaling along the dimensions to guarantee uniform spatial ranges across samples. Real-time pose tracking may introduce jitter or missing points; thus, smoothing filters across consecutive frames were applied to reduce noise

from tracking. Gesture sequences naturally have a different number of frames; hence, standardization of sequence length was conducted using either zero-padding or truncation in order to make fixed-length input tensors possible for batch-based training. Outlier removal and sequence validation ensured that only high-quality samples were retained for model learning.

The proposed deep learning architecture integrates spatial and temporal feature modeling. Landmark matrices are first processed by convolutional layers to capture the spatial de- pendencies of hand joints with respect to body keypoints, thereby learning structural patterns that define one particular gesture from another. These spatial embeddings are fed to bidirectional Long Short-Term Memory (BiLSTM) layers that model the temporal evolution of a gesture, thereby allowing the network to capture motion dynamics from both forward and backward temporal contexts. An attention mechanism has been incorporated to better recognize gestures whose recognition relies on subtle but critical transitional frames. This module adaptively weights temporally informative frames, ensuring that the network focuses on key transitions rather than treat- ing all frames as equal. The model was trained by using standardized learning schedules, an appropriate loss function, early stopping, dropout-based regularization, and the Adam optimizer to prevent overfitting while ensuring convergence. The model was then integrated into a real-time WebRTC- enabled application capable of processing live video streams from webcams or video conferencing interfaces. The incoming frames are continuously translated into landmarks, prepro- cessed, and fed into the model with minimal latency, which outputs gesture predictions. The classification head makes a prediction of the most probable gesture from the predefined vocabulary, thus enabling real-time interpretation suitable for communication accessibility, automated captioning, or inter- active user interfaces.

4. Purpose and System Interface

The main objective of the proposed system is to provide a practical, inclusive communication tool that would allow users who know sign language and those who do not to interact with each other with ease. Different from other systems that have relied on high-resolution images or controlled backgrounds, the proposed framework relies solely on the numerical land- mark data; thus, consistent accuracy in diverse environments can be guaranteed.

The system is designed to have a simple user interface: the user logs into the platform, and live video goes through a landmark extraction pipeline, displaying the recognized ges- ture in textual form. In addition, when necessary, recognized text is converted to synthesized speech, which is used to enable signers and non-signers to communicate with each other in real-time. Due to its integration with WebRTC, the system will have low latency during video calls, thus providing natural, uninterrupted communication.

5. Optimization Modes

To guarantee real-time performance, several optimization strategies are applied to the system. Efficient processing of hand and body detection on the CPU using MediaPipe de- creases the major computational load of the system. Unneces- sary parameters in the CNN–LSTM model were reduced to op- timize the model, while batch normalization was added to sta- bilize model training. At the time of deployment, lightweight quantization techniques further reduced the model size and enhanced the inference speed. The WebRTC communication layer adaptively changes bitrate and frame quality based on network conditions to make sure that gesture recognition is always smooth, even with suboptimal internet connectivity.

6. Access And Authentication

It was designed with secure access mechanisms for respon- sible usage of the recognition system. All users are required to log in before initiating or receiving a recognition session. Once logged in, the user interface then loads the real-time recognition panel where gestures are interpreted and displayed. The system, in fact, provides a controlled environment in which the processing of video data is handled in a secure fashion, considering user privacy and responsible use of the platform.

7. Results and Discussion

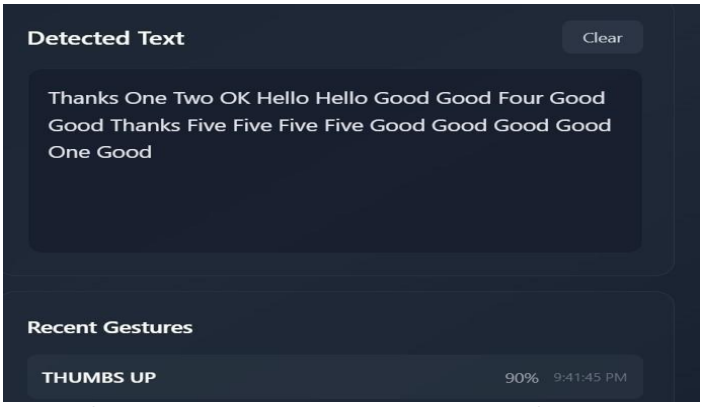

Fig. 1. System output interface displaying the accumulated sequence of recognized gestures in real time.

Extensive experiments were carried out to investigate the system's accuracy, efficiency, and real-time performance. The model showed strong performance, revealing high recogni- tion accuracy for static and dynamic gestures. For static, the recognition accuracy reached nearly ninety-eight percent, while dynamic achieved accuracy in the range from ninety- two to ninety-five percent. The system ran under CPU-only conditions at smooth inference speeds of about twenty-eight to thirty-two frames per second. Misclassifications primarily occurred where gestures either had similar intermediate transi- tions or small differences in motion. However, the introduction of body and hand landmarks enhanced clarity for complex gestures and thereby increased overall accuracy by almost ten percent. The bidirectional LSTM layers substantially enhanced the recognition of dynamic gestures through the capturing of temporal dependencies in both forward and backward directions. The attention module further enhanced such iden- tification and emphasis on the most meaningful frames during gesture execution. In summary, the results confirm that the coordinate-based representation used in this study is highly robust and much more efficient than traditional, image-based approaches.

Fig. 2. Media Pipe landmark detection showing the "Five" hand gesture

Fig. 3. Detection of the "Good" (thumbs-up) gesture with MediaPipe hand landmarks

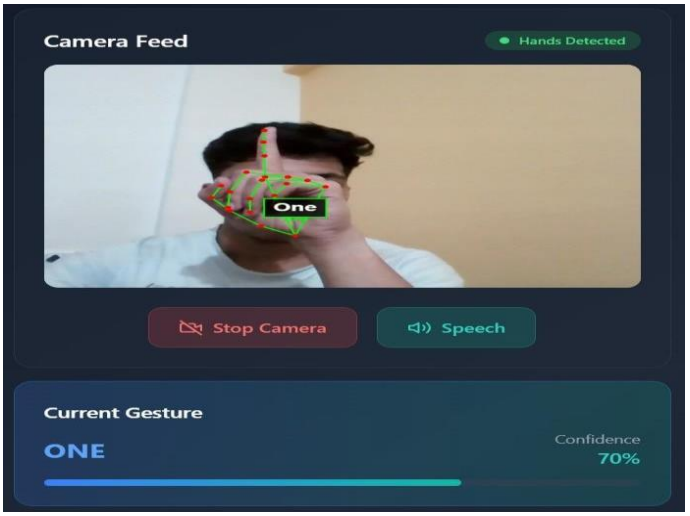

Fig. 4. Live camera feed illustrating detection of the "One" gesture, including confidence score visualization within the interface.

8. Applications

Applications of the proposed system are wide-ranging across a number of fields. An immediate application would be in real-time platforms, which enable conversation between signers and non-signers. The system may be integrated with educational tools to facilitate the learning of sign languages. In customer service situations, it can help persons who are hearing or speech-impaired to get services without relying on humans as intermediaries. In health and rehabilitation centers, it will be useful for tracking the improvement of patients during therapy with the help of gesture recognition. The system may also be integrated into smart home environments and IoT devices.

9. Future Work

Though the model proposed here demonstrates superior performance, many paths have yet to be pursued for further enhancements. The model currently focuses on a limited vo- cabulary; expanding the dataset to include other sign languages such as American Sign Language, British Sign Language, and Indian Sign Language would make this model more practical. Generative models, such as GANs, may be implemented to generate synthetic variations in gesture for better diversity in the model. Including facial expression recognition could be one more way of refining details of complex grammatical structures of the language. Finally, deploying the model using TensorFlow Lite on mobile devices can enable widespread portable use.

10. Conclusion

This research presented a robust landmark-driven sign lan- guage detection system that could function in both video and real-time video call environments. Shifting from image- based processing to coordinate-based modeling significantly improved robustness, efficiency, and scalability of the system. The integration of MediaPipe, CNN-LSTM architectures, and WebRTC enabled the system to attain accurate and low-latency gesture recognition suitable for real-world communication scenarios. This is a useful step toward bridging the gap in communication between signers and non-signers and holds promise for future extension into multilingual and multi- gesture domains.

Acknowledgment

Authors would like to thank Dr. V. M. Saravana Perumal and Dr. Richard William A of Department of CSE, RRCE, Bengaluru, for their continuous guidance, encouragement, and support throughout this research.

References

1. J. Sharma, R. Singh and P. Kumar, “Indian Sign Language Recogni- tion using Deep Learning Approaches,” *IEEE Access*, vol. 7, pp. 12312–12320, 2019.
2. S. Kumar and A. Roy, “Static Hand Gesture Recognition Using Convo- lutional Neural Networks,” *Procedia Computer Science*, vol. 167, pp. 1234–1245, 2021.
3. F. Chollet, “Deep Learning with Python,” Manning Publications, 2018.
4. A. Graves, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
5. G. Hinton et al., “Deep Neural Networks for Acoustic Modeling,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
6. Z. Nagi et al., “Real-Time Hand Gesture Recognition Using Vision- Based Techniques,” *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2790–2799, 2017.
7. M. Kim and H. Byun, “Hand Gesture Recognition Using 3D CNNs for Real-Time Human–Computer Interaction,” *IEEE Access*, vol. 8, pp. 14532–14541, 2020.
8. Y. Zhang, C. Zhang and M. Wang, “Dynamic Hand Gesture Recog- nition Based on LSTM Networks,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 188–197, Jan. 2020.
9. Google Research, “MediaPipe: Cross-platform Machine Learning Pipelines,” https://mediapipe.dev, 2020.
10. D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *International Conference on Learning Representations (ICLR)*, 2015.
11. S. Sridhar and A. Lu, “Real-Time Vision-Based Gesture Recognition Using Deep Learning,” *IEEE Transactions on Image Processing*, vol. 29, pp. 255–268, 2020.
12. T. Starner, “Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video,” *IEEE Transactions on Pat- tern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.
13. D. Kelly et al., “A Person-Independent System for Recognizing Hand Gestures,” *Pattern Recognition Letters*, vol. 31, pp. 1624–1632, 2010.
14. W. Li, Z. Zhang and Z. Liu, “Action Recognition Based on a Bag of 3D Points,” *IEEE CVPR Workshops*, pp. 9–14, 2010.