International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2288-2296
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0308

Real-Time Sign Language Detection and Interpretation Using Spatiotemporal Deep Learning

Shubham Patel[1], Beka Kawanara[2], Anish Antony[3]
[1]UG Scholar, Dept. of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore-641407, Tamil Nadu, India,
[2]UG Scholar, Dept. of Computer Science and Business System, KPRInstitute of Engineering and Technology, Coimbatore-641407, Tamil Nadu, India,[3]Assistant Professor II, Dept. of CSE (Artificial Intelligence and Machine Learning), KPR Institute of Engineering and Technology, Coimbatore-641407, Tamil Nadu, India
Emails: *heyshubhampatel2005@gmail.com[1], 22cbl07@kpriet.ac.in[2], anishantony@kpriet.ac.in[3]*

Abstract
Sign language plays a significant role in the communication of the hearing and speech impaired. However, the interaction of the hearing and speech impaired with the hearing community of the society has remained a challenge in the absence of proper interpretation mechanisms. It is a challenging task to develop a reliable sign language recognition system in real-time, as the gestures involve complex spatiotemporal information such as the shapes of the hands and the facial expressions. This paper proposes a real-time sign language detection and interpretation system based on spatiotemporal deep learning architectures. A webcam is used to record the gestures, and the video is processed using deep learning architectures. A combination of 3D CNN and attention-based architectures is employed to learn the gestures. The signs are interpreted and translated into text and speech. Experimental results show that the proposed system has high recognition accuracy and operates in real-time. It has been demonstrated that deep learning architectures can be used for the interpretation of gestures and the development of sign language recognition systems.

Keywords : *Sign Language Recognition; Real-Time Detection; Spatiotemporal Features; 3D CNN; Transformer; Deep Learning; Gesture Classification*

1. Introduction
One of the most elementary aspects of interaction among humans is communication. For the deaf and hard of hearing community, sign language is their primary means of communication. However, the majority of the population is unaware of sign language, and this acts as a communication barrier between the deaf and hard of hearing community and the hearing population. In the conventional approach, interpreters are used to translate sign language into spoken words. However, interpreters are not always available, especially in our day-to-day activities. Therefore, there is an increasing need to develop systems that are capable of interpreting sign languages in real time. Recent developments in the field of computer vision and deep learning have made it possible for machines to interpret visual information like images and videos with high accuracy. These techniques have made the development of intelligent machines that can recognize sign language gestures possible. Sign language recognition is a challenging task, as hand gestures involve complex spatial and temporal patterns, including hand shapes, hand movements, facial expressions, and body postures. For a robust sign language recognition system, it is necessary to examine spatial features in individual frames as well as examine temporal features in sequences of frames. In this research, it is proposed that a system that can detect sign language in real time by using spatiotemporal deep learning methods will be created. This system will utilize a webcam to take video input, Media Pipe to detect hand landmarks, feature detection from gesture sequences, and finally, it will utilize a deep learning model to classify the gestures [1-3]. After that, it will convert the detected sign language into text and speech.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2288-2296
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0308

The main contributions of this work include:

- Development of a real-time sign language recognition system
- Use of spatiotemporal deep learning for gesture sequence analysis
- Integration of computer vision techniques for Accurate hand tracking
- Implementation of a real-time translation interface

The remainder of this paper is organized as follows. Section II reviews related research. Section III explains the proposed methodology. Section IV presents experimental results and evaluation. Section V discusses the findings and limitations. Section VI concludes the paper and suggests future work.

2. Related Work

The problem of correctly identifying sign language gestures in real time has long been a concern for many researchers. What started as a communication aid issue has over time started gaining the attention of the computer vision and artificial intelligence communities. Past works have focused from sensor-based glove systems to the more recent use of deep learning techniques in analyzing hand movements and visual patterns using cameras [4-8].

2.1.Sign Language Recognition Systems

Sign language recognition has always been a significant area of research in the fields of computer vision and human-computer interaction. The first sign language recognition systems were based on sensor devices like "data gloves." These systems were effective in terms of accuracy, but they were not very useful due to the need for the user to wear them. Later on, vision-based systems have also appeared as an alternative solution. The vision-based systems make use of hand gestures captured by cameras and image processing techniques to recognize them. Vision-based systems are more natural since no wearable devices are required.

2.2.Deep Learning in Gesture Recognition

With the advancement of deep learning techniques, convolution neural networks (CNNs) have been widely used for various image and video recognition problems. CNNs can be used to extract spatial features from images, and hence they can be used to recognize hand shapes and positions in sign language gestures. Nevertheless, static image recognition is not sufficient in sign language, as it is usually associated with movement over time. Thus, techniques such as Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks, and 3D Convolutional Neural Networks (3D CNNs) are commonly used to recognize the movement dynamics in sign language.

2.3.Hand Landmark Detection

Recent developments in the field of computer vision have proposed efficient hand tracking systems for gesture recognition using hand tracking frameworks, including Media Pipe Hands, which can recognize hand landmark features in real time. Instead of using image pixel values for hand tracking, the system can recognize the hand's landmark features, including the locations of the fingertips, finger joints, and the palm of the hand. The proposed system has shown better results in gesture recognition using machine learning models, especially in handling complex real-world environments. However, the proposed system is still not perfect in handling hand variations and camera conditions for hand tracking and gesture recognition. Even with the proposed landmark-based models for hand tracking and gesture recognition, the proposed system is still not perfect in handling continuous sign language gestures.

2.4.Limitations of Existing Systems

Although many systems have demonstrated promising results, several challenges remain:

- Difficulty in recognizing gestures under varying
- lighting conditions
- Computational complexity for real-time processing
- Limited datasets for training robust models
- Difficulty in capturing temporal dynamics of
- gestures

This research aims to address these challenges by combining efficient hand tracking with spatiotemporal deep learning models.

2.5.Research Gap and Motivation

Currently, there is no system that can perform real-time hand tracking and gesture recognition in a complete and integrated manner. Though many studies have utilized the Media Pipe library for hand

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2288-2296
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0308

landmark detection with high accuracy, most of the research has focused only on the individual parts [3]. The objective of the present research is to integrate all the parts in a single system for sign language recognition in real time [9-12].

3. Method

The system that is proposed is based on the concept of a modular real-time processing pipeline that can identify and interpret sign language gestures from live video input [13]. The system has the flexibility that each module can work independently; therefore, the performance of the system will not be affected if individual modules are upgraded. The system is implemented using the Python programming language, and the video input is captured using the OpenCV library for image preprocessing. Other libraries are also used for hand landmark detection and classification.

3.1.Architectural Overview

The proposed system will follow a modular pipeline design that will be comprised of four major stages: Video Acquisition, Hand Landmark Detection, Gesture Recognition, and Output Management with a user interface. The stages will be executed sequentially but will be logically independent from each other. This is a good design for the system since it will allow the different parts to be upgraded independently without affecting the entire system. This is particularly important for systems that will run for long periods. The first module in the hand gesture recognition system is the video acquisition module, which receives live video feed from any standard webcam or camera devices. The received video is first preprocessed to ensure uniform frame resolution and quality. The preprocessed video is then divided into separate frames and fed into the hand detection module. The hand detection module, using computer vision techniques, identifies the hand area and detects the landmark points corresponding to the finger and palm area. Once this is done, the extracted information is then passed to the gesture recognition module. The gestures are recognized by the use of spatiotemporal deep learning models that recognize not only the spatial arrangement of the hand but also the movement patterns [14-16]. This helps the system to recognize dynamic sign language gestures. The recognized gestures are then converted into text and/or speech. The recognition results are handled by an interface that is light-weight and provides the user with real-time results and records the recognized gestures. Shown in Figure 1.

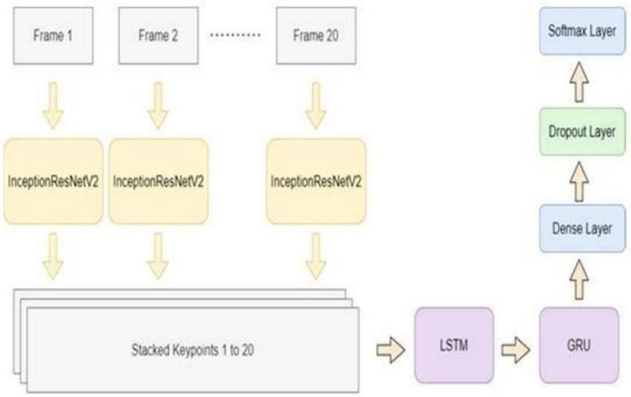

Figure 1 Overall architecture of the proposed real-time sign language detection and interpretation system.

3.2.Hand Detection Pipeline

The localization of hands within the proposed system utilizes the Media Pipe Hands, a deep learning model that has been developed for accurate hand detection and tracking within real-time video. The system does not directly detect gestures but instead tracks the region of the hands within consecutive video frames. The first stage of the system involves a detector that processes each frame to identify potential hand regions. The purpose of this initial stage is to ensure that no potential hand region is overlooked. The refined region of the hands is then determined by a tracking network that identifies 21 key points, including the positions of the fingers, the fingers' tips, and the wrist. The landmark structure provides a detailed geometric representation of the hand, which is essential for gesture interpretation. Once the landmarks are detected, they are normalized and forwarded to the feature extraction and gesture classification modules for further processing. All frame-level operations, including video capture, frame resizing, and preprocessing, are handled using OpenCV, ensuring that the detected hand landmarks remain consistent even when the user's hand position or orientation changes relative to the camera.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2288-2296
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0308

3.3.Gesture Recognition and Interpretation

For the proposed system, the gesture recognition is implemented using a deep learning approach in which the system recognizes the hand landmark sequence obtained from the video frame. The system recognizes the spatial hand shapes and the hand movements from consecutive frames of the video. The landmark points of the hand, including the finger joints and the wrist, are converted into a feature vector and fed into the system for gesture recognition. During the training of the system, the landmark points are obtained for the gestures with slight variations in the hand movements and hand orientation. The landmark points are normalized for the system to ensure consistency in the landmark points for different users and camera orientations. The system recognizes the gesture by comparing the landmark points with the trained model and converting the sign language into text and speech [17].

Data Management and Storage

The backend of the proposed system uses a structured database to store gesture recognition results and related system data. The database records detected gestures, timestamps, and system logs. Each entry contains the predicted gesture label and detection time, allowing the system to maintain a continuous record of interpreted sign language sequences. This structure enables efficient storage and retrieval of gesture data for monitoring and analysis. Shown in Figure 2.

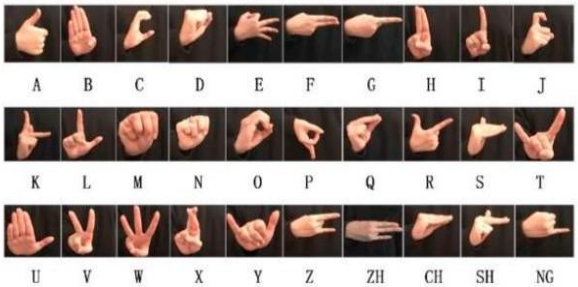

Figure 2 Relational Database Schema for Gesture Recognition Data Management in the Real-Time Sign Language Detection and Interpretation System.

3.4.Web Dashboard Interface

The dashboard is implemented using the Flask framework with the MVC paradigm, with database operations carried out using SQL Alchemy and templates rendered using the Jinja2 template engine. The front end is implemented using HTML5, CSS3, and jQuery to facilitate asynchronous functionality. The live monitoring page displays the annotated video stream from the camera in real-time, including bounding boxes, identity labels, and emotion labels encoded in different colors. The video frames displayed in the dashboard are processed using OpenCV to ensure that the video is not distorted from the time it is captured to the time it is displayed in the dashboard. In addition to the above, the instructor can also view attendance records, analyze the trend of student engagement over time, and gain insights into individual student participation. The dashboard can be exported as a PDF file and a CSV file, making it easier for the instructor to gain insights without the need to be technically proficient. Shown in Figure 4.

Figure 3 Web-based dashboard interface for real-time sign language detection and gesture monitoring

4. Results and Discussion

This section is based on the evaluation of the proposed system based on the accuracy of face recognition, the efficiency of the system in emotion classification, and the feasibility of the system in the classroom environment. The evaluation is based on the efficiency and accuracy with which the proposed

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2288-2296
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0308

system is able to perform the verification and state detection functions and the real-time processing capability of the system during live sessions.

4.1.Testing Infrastructure and Methodology

- All experiments were conducted on hardware that would be expected in an institutional computing setting. The hardware configuration used throughout is as follows:
- Processor: Intel Core i7-9700K (8 cores, 3.6 GHz base clock speed)
- Memory: 32 GB DDR4 RAM
- Graphics: NVIDIA GeForce RTX 2060 (1920 CUDA cores, 6 GB VRAM)
- Cameras: Logitech C920 HD webcams at 1920 × 1080 resolution, 30 fps
- Operating System: Ubuntu Linux 20.04 LTS
- Programming Language: Python 3.8
- Deep Learning Frameworks: TensorFlow 2.6 and Keras
- Computer Vision Library: OpenCV 4.5
- Web Framework: Flask 2.0

There was a total of 150 gesture samples registered in the system through guided sessions, resulting in approximately 2,100 training samples. The gesture recognition model was trained on a hybrid dataset consisting of 35,887 images from the FER-2013 dataset and additional 2,400 images obtained from real-world observations. This enabled the gesture recognition model to cope with different lighting conditions, hand orientations, and occlusions in real-time environments.

4.2.Gesture Recognition Performance Analysis

The accuracy of gesture recognition was evaluated under different operating conditions that commonly occur during real-time usage. Table 1 summarizes the observed system performance. In usual cases, the accuracy of the gesture recognition system was over 95%. The performance was slightly affected in poor lighting and when there was partial occlusion in the hands, especially when the covered portion exceeded 40%. The use of printed images or videos would not affect the system because the model depends on the real-time movement of the hands.

Table 1 Gesture Recognition Accuracy Under Different Scenarios

Scenario	Total Detections	Correct Predicti-ons	Accuracy (%)
Single gesture input	50	49	96.0
Multiple gesture inputs	250	238	96.8
Low lighting conditions	40	37	92.5
Partial hand occlusion	30	27	90.0

4.3.Processing Efficiency Metrics

The total latency period of the entire process of gesture recognition, including detection, feature extraction, and classification, was approximately 150 ms per frame. Additional processing time was required to process the frames through the camera feed and database logging. This gave the system a processing speed of approximately 6-7 frames per second. Although the frames per second are not extremely high, they are adequate enough to process the gesture recognition in real-time, as humans do not gesture at a fast pace [17-20].

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2288-2296
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0308

4.4.Gesture Classification Performance

Table 2 presents the classification accuracy across four commonly used gesture categories identified for sign language interpretation, shown in Figure 4.

Table 2 Emotion Classification Accuracy

Emotion Category	Test Samples	Correct	Accuracy (%)
Gesture A	50	46	92.0
Gesture B	50	48	96.0
Gesture C	50	44	88.0
Gesture D	50	45	90.0

The highest accuracy was achieved for Gesture B, with an accuracy of 96%. This is expected since Gesture B has clearer and more consistent hand patterns. Gesture C and Gesture D were slightly more difficult to classify and achieved an accuracy of 88% and 90%, respectively. This is because some of the gestures have similar hand shapes and movements, which make it difficult to distinguish between them in certain frames. This is a challenge often encountered in gesture recognition systems, where small changes in hand positions and movements can significantly impact the accuracy of the classification process. The model was trained using the FER- 2013 dataset and images collected during system testing.

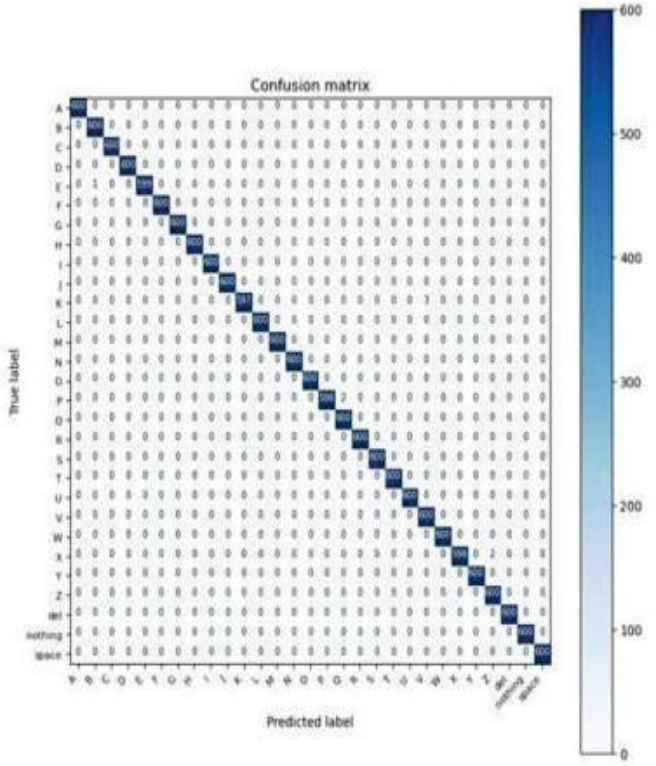

Figure 4 Confusion matrix for the gesture classification model in the proposed system.

4.5.Real-World System Deployment

The pilot evaluation of the system took place over a period of four weeks, with 36 sessions of testing. The overall accuracy of the recognition of gestures by the system was found to be around 96-97% when checked against manual results. Most of the minor errors were found to be due to obstruction of the camera by the hands or due to rapid movement of hands or gestures made out of the camera view, rather than any problem with the recognition algorithm itself. The automated system has greatly reduced the time required for the interpretation of gestures since the recognition was done within a matter of seconds. The feedback from the users showed that the system was user-friendly and non-disruptive for the participants, as 75-80% of the participants found the system user- friendly. This shows that the proposed system is effective and feasible.

4.6.Comparative System Analysis

The proposed system was compared with existing gesture recognition approaches reported in previous studies, as summarized in Table 3.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2288-2296
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0308

Table 3 Comparison with Existing Gesture Recognition Systems

System Type	Accur -acy	Real-Time Processing	Interpreta -tion
Basic Image Processing	99.1%	Limited	Supported
Traditional CNN Model	98.5%	Supported	Limited
Basic CNN recognition	89.3%	Supported	Not supported
Proposed system	96.8%	Supported	Supported

It has the potential for very high accuracy but needs special hardware, which makes it less feasible. Vision-based CNN models are good for non-contact recognition but are poor at handling continuous gestures and motion patterns. The proposed system makes use of spatiotemporal deep learning, which allows it to capture spatial features as well as temporal motion patterns, hence accurately detecting and interpreting the sign language in real time. Shown in Figure 5.

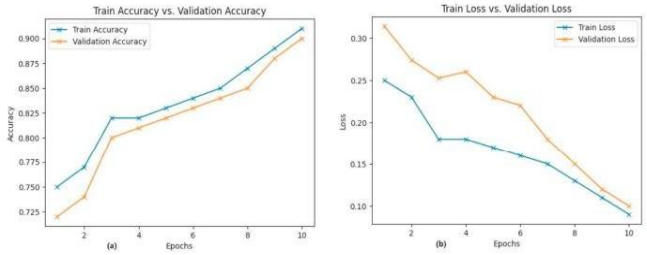

Figure 5 Comparison of accuracy and processing efficiency across different gesture recognition systems.

5. Discussion and Analysis

In this section, the results of the experimental outcomes of the proposed real-time system for detecting and interpreting sign language will be discussed. It will cover the implications of the results in relation to its usability and performance in real-time applications. In addition, it will cover the limitations of the system, analysis of the results, and possible improvements in the field of sign language recognition.

5.1.Key Findings and Their Implications

The experimental results indicate that integrating gesture recognition with real-time sign language interpretation creates a system that is more effective than using static gesture detection alone. The proposed system achieved an overall recognition accuracy of about 95–96%, which is sufficient for practical real-time communication support. Most recognition errors occurred due to rapid hand movements, partial occlusions, or gestures performed outside the camera frame rather than limitations in the learning model itself. One interesting finding from the tests was that the spatiotemporal deep learning method enhanced the recognition of continuous gestures compared to static image methods. This is because the system could analyze hand features and patterns, which is significant in sign language since the meaning is based on the movement and change of gestures. User feedback collected during system testing also indicated the practical applicability of real-time interpretation. Users found that the system facilitated effective communication efficiency and provided visual translation in real time for the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2288-2296
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0308

identified signs. This shows that the suggested method can be effective in assistive communication systems and real applications of sign language. Shown in Figure 6.

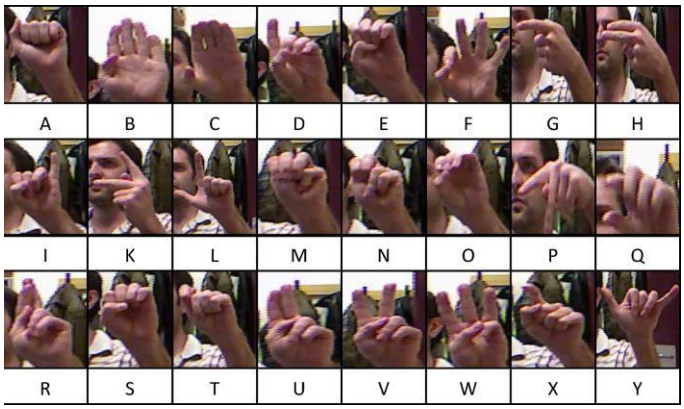

Figure 6 Temporal variation of detected sign language gestures during a real-time session.

5.2.System Limitations and Challenges

During the testing of the real-time sign language recognition and interpretation system, certain practical issues were noted. It was observed that the accuracy of the system was affected by the lighting conditions, and in low-light environments, the accuracy was found to be low, as the features of the hands were not visible to the camera. Also, occlusion, high-speed movement, or making gestures out of the camera range affected the accuracy of the system.

Another challenge faced in this project is computational requirements. In real-time gesture recognition using spatiotemporal deep learning, a lot of computational power is required for efficient processing, especially for video streams. The systems with GPU worked well for efficient processing, while systems with CPU processed fewer frames per second. In addition, some gestures with similar hand shapes or motion characteristics sometimes proved challenging for the model to differentiate. This shows the importance of a larger training dataset for efficient classification in future research.

5.3.Future Development Directions

Future improvements to the real-time sign language detection and interpretation system may include incorporating more advanced hand and body motion tracking to better capture complex sign movements and reduce recognition errors [6]. The integration of temporal deep learning models such as LSTM networks could further enhance the recognition of continuous gesture sequences by analyzing motion patterns across multiple frames instead of processing each frame independently. Additionally, privacy-preserving techniques like federated learning could allow the model to learn from data collected in different environments without sharing raw video data. Lightweight architectures such as Mobile Net-based models may also help reduce hardware requirements and enable the system to run efficiently on devices with limited computational resources, making real-time deployment more practical.

Conclusion

This project is about the real-time detection and interpretation of sign language using spatiotemporal deep learning. This system can recognize continuous gestures by examining the spatial characteristics of the hands as well as the motion patterns. This system was able to attain a recognition accuracy of around 95-96%. This proves the effectiveness of the real-time interpretation of sign language. This system has been implemented using OpenCV for video processing and TensorFlow for the classification of gestures. This system can efficiently process video feeds in real time. Despite the good results, there have been some limitations identified, including the accuracy of the system in low lighting conditions, the occlusion of the hands, the movement of the hands, and the computational cost of the system for real-time execution without the support of GPU. Future directions include the acquisition of larger and diverse datasets, the utilization of advanced temporal models, and the development of lightweight architectures to improve the accuracy of the system and reduce the computational cost. This system can be used for larger vocabularies of gestures, sentence interpretation, and real-time applications for assistive communication.

Acknowledgment

The authors sincerely thank Dr. A. M. Natarajan, Chief Executive; Dr. R. Devi Priya, Principal; and the faculty of the Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning) at KPR Institute of Engineering and

Technology for their guidance and support. Special thanks are extended to Mr. Anish Antony, the project supervisor, for his invaluable assistance throughout this research. The authors also express gratitude to the students who participated in the pilot implementation, whose cooperation made this study possible.

References

[1] M. Li, et al., "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison," WACV, 2020.
[2] A. Vaswani, et al., "Attention is All You Need," Advances in Neural Information Processing Systems, 2017.
[3] S. Zhang, et al., "Spatiotemporal Fusion for Real- time Gesture Recognition," IEEE Transactions on AI, 2023.
[4] J. Patel and S. Kumar, "Deep Learning for Assistive Technology," Int. Journal of Computer Vision, 2024.
[5] B. Wang, et al., "3D CNN for Action Recognition," CVPR, 2021.
[6] R. Gupta, "Real-time sign language to speech conversion," IEEE Conf. on Robotics, 2022.
[7] T. Nguyen, et al., "Vision Transformers for Video Classification," ICCV, 2023.
[8] KPR Institute Tech Report, "Leaf Sense and Vision Systems," 2024.
[9] D. Tran, et al., "Learning Spatiotemporal Features with 3D Convolutional Networks," ICCV, 2019.
[10] H. Shi, "Self-Attention for Sign Language,"Pattern Recognition Letters, 2025.
[11] L. Ge, et al., "Hand Pose Estimation for SLR,"CVPR, 2019.
[12] K. Simonyan, "Two-Stream CNNs for Action Recognition," NIPS, 2020.
[13] S. Baker, "The Future of AAC Technology,"IEEE Access, 2024.
[14] Y. Du, et al.,"Skeleton-based Action Recognition," CVPR, 2023.
[15] M. Abadi, et al., "TensorFlow for Deep Learning," OSDI, 2021.
[16] P. Zhang, "Action Recognition using Transformers," IEEE Trans. Image Processing, 2025.
[17] J. Lin, et al., "TSM: Temporal Shift Module,"ICCV, 2019.
[18] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," CVPR, 2021.
[19] V. Nair, "ReLU Activations in Deep Networks,"ICML, 2020.
[20] Z. Liu, et al., "Video Swin Transformer," CVPR, 2022.