International Journal of Engineering Science Technologies
March-April 2026 10(2)
www.ijoest.com

Original Article

AN INTELLIGENT DEEP LEARNING-BASED FRAMEWORK FOR REAL-TIME SIGN LANGUAGE RECOGNITION USING VISION-BASED GESTURE ANALYSIS

Dr. Harish Barapatre [1*], Saundarya Sudhakar Rasal [2], Harshada Chandrabhan Pagar [2], Ashwinikumar Dinanath Chavan [2]
[1] Associate Professor, Department of Computer Engineering, Yadavrao Tasgaonkar Institute of Engineering and Technology, Bhivpuri Road Karjat, Maharashtra, 410201, India
[2] Student, Department of Computer Engineering, Yadavrao Tasgaonkar Institute of Engineering and Technology, Bhivpuri Road Karjat, Maharashtra, 410201, India

ABSTRACT

Sign language recognition has emerged as a critical research area aimed at reducing the communication barrier between hearing-impaired individuals and the general population. Traditional communication methods often rely on human interpreters, which are not always accessible, scalable, or cost-effective. Recent advancements in computer vision and deep learning have enabled the development of automated systems capable of interpreting hand gestures and translating them into meaningful text or speech. However, existing systems often suffer from limitations such as sensitivity to background noise, lack of real-time performance, and insufficient generalization across different signers Starner et al. (1998), Vogler and Metaxas (1999).

This paper proposes an intelligent vision-based sign language recognition framework that leverages deep learning techniques for accurate and real-time gesture interpretation. The system captures hand gestures through a camera interface, performs preprocessing to extract relevant spatial features, and utilizes convolutional neural networks (CNNs) for feature learning and classification. Additionally, temporal dependencies in dynamic gestures can be modeled using sequence-based architectures, enhancing recognition capability Cooper et al. (2011). The proposed framework is designed to be scalable, robust to environmental variations, and deployable in real-world assistive applications.

The primary contribution of this work lies in designing a structured, end-to-end pipeline that integrates gesture acquisition, feature extraction, classification, and output generation into a unified system. The framework aims to improve accessibility, enable real-time communication support, and serve as a foundation for future multimodal interaction systems.

Keywords: Sign Language Recognition, Computer Vision, Deep Learning, Gesture Recognition, CNN, Human-Computer Interaction

INTRODUCTION

Communication is a fundamental aspect of human interaction, yet millions of hearing-impaired individuals rely on sign language as their primary mode of expression. Sign languages are rich, structured visual languages that use hand gestures, facial expressions, and body movements to convey meaning. However, a significant communication gap exists between sign language users and those unfamiliar with it, creating challenges in education, employment, and daily interactions. The dependency on human interpreters further limits accessibility, especially in real-time or remote scenarios Starner et al. (1998).

*Corresponding Author:
Email address: Dr. Harish Barapatre (harishkbarapatre@gmail.com), Saundarya Sudhakar Rasal (saundaryarasal@email.com), Harshada Chandrabhan Pagar (harshadapagar123@gmail.com), Ashwinikumar Dinanath Chavan (adc402@yahoo.co.in)
Received: 18 February 2026; **Accepted:** 26 March 2026; **Published** 30 April 2026
DOI: 10.29121/IJOEST.v10.i2.2026.757
Page Number: 97-108
Journal Title: International Journal of Engineering Science Technologies
Journal Abbreviation: Int. J. Eng. Sci. Tech
Online ISSN: 2456-8651
Publisher: Granthaalayah Publications and Printers, India
Conflict of Interests: The authors declare that they have no competing interests.
Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
Authors' Contributions: Each author made an equal contribution to the conception and design of the study. All authors have reviewed and approved the final version of the manuscript for publication.
Transparency: The authors affirm that this manuscript presents an honest, accurate, and transparent account of the study. All essential aspects have been included, and any deviations from the original study plan have been clearly explained. The writing process strictly adhered to established ethical standards.
Copyright: © 2026 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License.
With the license CC-BY, authors retain the copyright, allowing anyone to download, reuse, re-print, modify, distribute, and/or copy their contribution. The work must be properly attributed to its author.

With the advancement of artificial intelligence and computer vision, automated sign language recognition systems have gained increasing attention. These systems aim to interpret gestures captured through cameras and convert them into text or speech, enabling seamless communication. Early approaches relied heavily on sensor-based systems such as data gloves and motion trackers, which, although accurate, were intrusive and expensive. More recent vision-based methods utilize image and video data, making them more practical and scalable for real-world deployment Vogler and Metaxas (1999).

Deep learning techniques, particularly convolutional neural networks (CNNs), have demonstrated significant success in extracting spatial features from images, while recurrent neural networks (RNNs) and long short-term memory (LSTM) models are effective in capturing temporal dependencies in dynamic gestures Cooper et al. (2011). Despite these advancements, several challenges remain unresolved. Many existing systems struggle with varying lighting conditions, complex backgrounds, and differences in hand shapes and motion patterns among users. Additionally, achieving real-time performance without compromising accuracy remains a critical concern.

The motivation behind this work is to design a robust, scalable, and real-time sign language recognition framework that overcomes these limitations. The proposed system focuses on vision-based gesture acquisition combined with deep learning models to ensure accurate interpretation under diverse conditions. It aims to provide a practical solution that can be deployed in assistive technologies, educational tools, and human-computer interaction systems.

The key contributions of this paper are as follows:

1) Development of a structured, end-to-end framework for vision-based sign language recognition.
2) Integration of deep learning techniques for efficient feature extraction and classification.
3) Design of a system capable of handling real-time gesture interpretation.
4) A scalable architecture that can be extended to multilingual sign language datasets and applications.

LITERATURE REVIEW

Sign language recognition has been extensively studied using various approaches ranging from sensor-based systems to advanced deep learning models. Early research primarily focused on hardware-based solutions such as data gloves equipped with sensors to capture finger movements and hand orientation. These systems provided high accuracy due to precise motion capture but were limited by their high cost, lack of portability, and user discomfort Starner et al. (1998).

With the evolution of computer vision, researchers shifted toward vision-based approaches that utilize cameras to capture hand gestures. Traditional image processing techniques such as edge detection, skin color segmentation, and contour extraction were initially employed to identify hand regions and gestures. However, these methods were highly sensitive to lighting conditions, background noise, and variations in skin tone, resulting in reduced robustness in real-world environments Vogler and Metaxas (1999).

The introduction of machine learning improved recognition accuracy by enabling systems to learn patterns from data. Techniques such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Hidden Markov Models (HMM) were widely used for gesture classification. While these methods showed improvement over classical techniques, they required manual feature engineering, which limited scalability and adaptability Cooper et al. (2011).

Recent advancements in deep learning have significantly transformed sign language recognition systems. Convolutional Neural Networks (CNNs) are widely used for extracting spatial features from hand gesture images, providing higher accuracy and robustness compared to traditional methods. For dynamic gestures, sequence-based models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are employed to capture temporal dependencies in gesture sequences Simonyan and Zisserman (2014). Hybrid models combining CNN and LSTM architectures have shown promising results in recognizing both static and dynamic signs Graves et al. (2013).

Furthermore, real-time recognition systems have been explored using optimized deep learning models and lightweight architectures. Techniques such as transfer learning and model compression have been applied to reduce computational complexity and enable deployment on edge devices Molchanov et al. (2015). Despite these advancements, challenges such as occlusion, signer variability, complex backgrounds, and lack of large standardized datasets continue to affect system performance.

The comparative analysis of key existing works is summarized in Table 1.

Table 1

Table 1 Comparative Analysis of Existing Sign Language Recognition Systems

Paper	Method Used	Limitation
Starner et al. (1998)	Sensor-based Data Gloves	Expensive and intrusive
Vogler and Metaxas (1999)	Traditional Image Processing	Sensitive to lighting and background

Cooper et al. (2011)	SVM / HMM-based Models	Requires manual feature extraction
Simonyan and Zisserman (2014)	CNN-based Models	Limited temporal understanding
Graves et al. (2013)	CNN + LSTM Hybrid Models	High computational complexity
Molchanov et al. (2015)	Lightweight Deep Learning Models	Trade-off between accuracy and speed

From the above analysis, it is evident that although deep learning-based approaches have improved recognition performance, there is still a need for a unified framework that balances accuracy, real-time performance, and scalability.

RESEARCH GAP AND PROBLEM STATEMENT

Despite significant advancements in sign language recognition systems, several critical gaps remain that limit their practical deployment in real-world environments. Existing approaches have improved accuracy through deep learning models; however, they often fail to maintain consistency across diverse conditions such as varying lighting, complex backgrounds, and differences in user hand shapes and motion patterns. Many systems are trained on controlled datasets, which restricts their generalization capability when exposed to real-time scenarios Starner et al. (1998), Vogler and Metaxas (1999).

Another major limitation lies in the trade-off between accuracy and real-time performance. High-accuracy models such as deep CNN-LSTM architectures tend to be computationally intensive, making them unsuitable for real-time applications or deployment on resource-constrained devices. On the other hand, lightweight models designed for speed often compromise recognition accuracy, leading to unreliable outputs in practical usage Cooper et al. (2011).

Furthermore, most existing systems focus either on static gesture recognition or dynamic gesture recognition, but not both in a unified manner. This creates a gap in developing a comprehensive system capable of handling continuous sign language communication. Additionally, the lack of standardized and diverse datasets leads to poor robustness across different users, sign styles, and environmental variations Simonyan and Zisserman (2014).

Another overlooked aspect is the absence of an integrated, end-to-end pipeline that seamlessly connects gesture acquisition, preprocessing, feature extraction, classification, and output generation. Many studies focus only on model development without addressing system-level design, scalability, and deployment feasibility. This results in solutions that are difficult to translate into real-world assistive technologies.

PROBLEM STATEMENT

The primary problem addressed in this research is the design and development of a robust, scalable, and real-time sign language recognition system that can accurately interpret both static and dynamic gestures using vision-based inputs. The system must overcome challenges related to environmental variability, computational efficiency, and user diversity while maintaining high recognition accuracy.

Specifically, the research aims to:

- Develop a vision-based framework capable of capturing and interpreting hand gestures in real time.
- Ensure robustness against variations in lighting, background, and user-specific gesture patterns.
- Balance accuracy and computational efficiency for practical deployment.
- Integrate all system components into a unified pipeline for seamless operation.

By addressing these challenges, the proposed system seeks to bridge the communication gap between sign language users and non-users, enabling more inclusive and accessible human-computer interaction.

PROPOSED FRAMEWORK AND SYSTEM ARCHITECTURE

The proposed system is designed as an end-to-end vision-based pipeline for real-time sign language recognition. It integrates multiple processing stages, starting from gesture acquisition to final output generation, ensuring both accuracy and scalability. The architecture focuses on modular design so that each component can be optimized or upgraded independently without affecting the overall system performance.

Figure 1

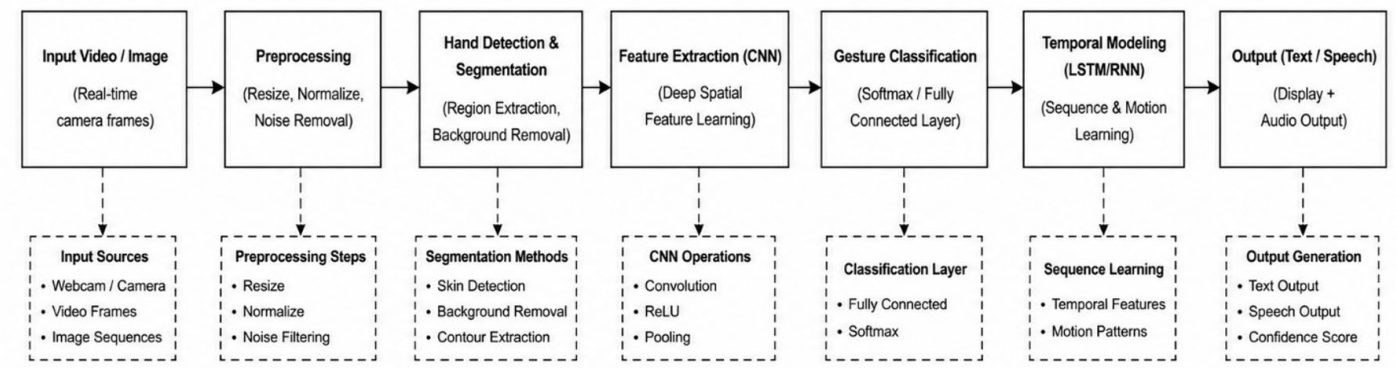

Figure 1 Shows the Proposed System Architecture.

OVERALL SYSTEM FLOW

Input (Video Stream / Image Frames)
→ Preprocessing
→ Hand Detection and Segmentation
→ Feature Extraction (Deep Learning)
→ Gesture Classification
→ Temporal Modeling (for dynamic gestures)
→ Output Generation (Text / Speech)

COMPONENT DESCRIPTION

1) Input Layer

The system captures real-time video input using a camera or accepts pre-recorded gesture videos. The input consists of continuous frames representing hand movements and gestures. These frames serve as the primary data source for further processing.

2) Preprocessing

Preprocessing is applied to enhance input quality and reduce noise. This includes:

- Frame resizing and normalization
- Background noise reduction
- Contrast and brightness adjustment
- Conversion to suitable color spaces (e.g., RGB to HSV)

This step ensures consistency in input data and improves model performance.

3) Hand Detection and Segmentation

The system isolates the hand region from each frame using techniques such as:

- Skin color segmentation
- Contour detection
- Deep learning-based object detection models (e.g., YOLO, MediaPipe)

Accurate segmentation is crucial for eliminating irrelevant background information and focusing only on gesture-related features.

4) Feature Extraction (Deep Learning)

A Convolutional Neural Network (CNN) is used to automatically extract spatial features from the segmented hand images. The CNN learns patterns such as:

- Finger orientation
- Hand shape

- Gesture structure

This eliminates the need for manual feature engineering and improves robustness.

5) Gesture Classification

The extracted features are passed to a classification layer (e.g., fully connected neural network or softmax classifier) to identify the corresponding sign label. The system maps each gesture to a predefined vocabulary of signs.

6) Temporal Modeling (for Dynamic Gestures)

For continuous or dynamic gestures, temporal dependencies are captured using sequence models such as:

- Recurrent Neural Networks (RNN)
- Long Short-Term Memory (LSTM)

This allows the system to understand motion patterns across multiple frames rather than relying on a single image.

7) Output Generation

The final recognized gesture is converted into:

- Text output (displayed on screen)
- Speech output (using text-to-speech systems)

This enables real-time communication between users.

KEY FEATURES OF THE PROPOSED ARCHITECTURE

- Real-Time Processing: Optimized pipeline for live gesture recognition
- Scalability: Can be extended to multiple sign languages
- Robustness: Handles environmental variations and user diversity
- Modularity: Each component can be independently improved
- Deployment Ready: Suitable for integration into mobile and assistive devices

MATHEMATICAL MODEL

The proposed sign language recognition system is mathematically modeled to represent feature extraction, classification, and optimization processes. The model ensures that gesture recognition is both accurate and computationally efficient.

1) Feature Representation Model

Each input frame is processed to extract a feature vector using a deep learning model (CNN). Let the input image frame be represented as:

Display Format:

$F = CNN(I)$

Word Equation Format:

F = CNN(I)

Where:

- $I \rightarrow$ Input image/frame
- $CNN(\cdot) \rightarrow$ Convolutional Neural Network function
- $F \rightarrow$ Extracted feature vector

This feature vector captures spatial characteristics such as hand shape, orientation, and gesture structure.

2) Gesture Classification Function

The classification layer maps the extracted feature vector to a gesture label using a weighted scoring function.

Display Format (Eq. 1):

$S = \alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_3 + \ldots + \alpha_n F_n$

Word Equation Format:

S = \alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_3 + ... + \alpha_n F_n

Where:

- $F_1, F_2, \ldots, F_n \rightarrow$ Feature components

- $\alpha_1, \alpha_2, ..., \alpha_n \rightarrow$ Learnable weights
- $S \rightarrow$ Classification score

The predicted gesture class is obtained using:

Display Format (Eq. 2):

$\hat{y} = \text{argmax}(S)$

Word Equation Format:

\hat{y} = \arg\max(S)

Where:

- $\hat{y} \rightarrow$ Predicted gesture label
- argmax → Function selecting the highest score

3) Temporal Modeling for Dynamic Gestures

For dynamic gestures, sequential dependencies are captured using temporal modeling.

Display Format (Eq. 3):

$H_t = \sigma(W_h F_t + U_h H_{t-1} + b_h)$

Word Equation Format:

H_t = \sigma(W_h F_t + U_h H_{t-1} + b_h)

Where:

- $F_t \rightarrow$ Feature vector at time t
- $H_t \rightarrow$ Hidden state at time t
- $W_h, U_h \rightarrow$ Weight matrices
- $b_h \rightarrow$ Bias term
- $\sigma \rightarrow$ Activation function

This allows the system to capture motion patterns across frames.

4) Loss Function Optimization

The system is trained using a classification loss function to minimize prediction error.

Display Format (Eq. 4):

$L = -\sum (y \log(\hat{y}))$

Word Equation Format:

L = - \sum (y \log(\hat{y}))

Where:

- $L \rightarrow$ Loss function
- $y \rightarrow$ True label
- $\hat{y} \rightarrow$ Predicted probability

The objective is to minimize L during training to improve model accuracy.

Summary of Mathematical Model

- CNN extracts spatial features
- Weighted scoring performs classification
- Temporal model captures motion dynamics
- Loss function optimizes prediction accuracy

This mathematical formulation ensures that the system is both theoretically sound and practically implementable.

SUMMARY OF MATHEMATICAL MODEL

- CNN extracts spatial features
- Weighted scoring performs classification
- Temporal model captures motion dynamics
- Loss function optimizes prediction accuracy

This mathematical formulation ensures that the system is both theoretically sound and practically implementable.

ALGORITHM AND PSEUDOCODE

Algorithm 1: Proposed Sign Language Recognition System

Input: Real-time video stream or gesture image dataset

Output: Recognized sign label with text/speech output

1) Start the system.
2) Capture input gesture using camera or load gesture image/video dataset.
3) Extract video frames from the input stream.
4) Resize each frame to a fixed dimension.
5) Normalize pixel values for stable model processing.
6) Apply hand detection to locate the hand region.
7) Segment the detected hand region from the background.
8) Pass the segmented hand image to the CNN model.
9) Extract spatial gesture features from the CNN layers.
10) If the gesture is static:
    - Classify the extracted feature vector directly.
11) If the gesture is dynamic:
    - Store frame-wise feature vectors in sequence order.
    - Pass the feature sequence to LSTM/RNN model.
    - Capture temporal movement patterns.
12) Apply the classifier to predict the gesture class.
13) Select the gesture label with the highest classification score.
14) Convert the predicted label into text.
15) If required, convert the text into speech using a text-to-speech module.
16) Display the final output to the user.
17) End the system.

PSEUDOCODE

Input: Video frames V = {F1, F2, F3, ..., Fn}

Output: Predicted sign label Y

Begin

```
Initialize trained CNN model
Initialize temporal model if dynamic gesture recognition is required
Initialize label dictionary

For each frame Fi in V do

    Resize Fi
    Normalize Fi
    Detect hand region from Fi
    Segment hand region
    Extract feature vector Xi using CNN

    Store Xi in feature sequence
```

```
End For

If gesture type is static then

    Y = Classifier(Xi)

Else

    Y = Temporal_Model({X1, X2, X3, ..., Xn})

End If

Predicted_Label = argmax(Y)

Convert Predicted_Label into text output

If speech output is enabled then

    Generate speech from text
End If

Return Predicted_Label

End
```

ALGORITHM EXPLANATION

The proposed algorithm begins by capturing gesture input through a camera or dataset. Each frame is preprocessed to ensure uniform size and quality. The hand region is then detected and segmented so that the model focuses only on gesture-relevant information. For static gestures, CNN-extracted features are directly classified. For dynamic gestures, the system processes frame-wise features through a temporal model such as LSTM or RNN. The final recognized sign is converted into text or speech for user-friendly communication.

METHODOLOGY / WORKING

The proposed sign language recognition system follows a structured pipeline that transforms raw visual input into meaningful textual or speech output. The methodology is designed to ensure robustness, real-time performance, and adaptability across different users and environments.

1) Data Acquisition

The system begins by capturing gesture data using a camera interface or utilizing pre-existing gesture datasets. The input may consist of:

- Static hand gesture images
- Continuous video sequences representing dynamic gestures

The captured data serves as the foundation for training and real-time inference.

2) Data Preprocessing

Before feeding the data into the model, preprocessing is applied to standardize input and reduce noise. This includes:

- Resizing frames to a fixed resolution
- Normalizing pixel intensity values
- Removing background noise
- Adjusting brightness and contrast

This step ensures consistency and improves model generalization across varying conditions.

3) Hand Detection and Segmentation

The system identifies and isolates the hand region from each frame. This can be achieved using:

- Traditional segmentation methods (color-based detection)
- Deep learning-based detectors (e.g., MediaPipe, YOLO)

Segmentation removes irrelevant background information, allowing the model to focus only on gesture-related features.

4) Feature Extraction

A Convolutional Neural Network (CNN) is used to extract high-level spatial features from the segmented hand images. The CNN automatically learns patterns such as:

- Finger positions
- Hand shapes
- Gesture contours

This eliminates manual feature engineering and enhances recognition accuracy.

5) Gesture Modeling

The system handles both static and dynamic gestures:

- **Static Gesture Recognition:** Feature vectors extracted from individual frames are directly passed to a classifier.
- **Dynamic Gesture Recognition:** Sequential frame features are processed using temporal models such as LSTM or RNN to capture motion patterns over time.

This dual capability ensures comprehensive sign language interpretation.

6) Classification

The extracted features (or feature sequences) are fed into a classification layer to determine the corresponding gesture label. The classifier outputs probability scores for each possible class, and the highest score determines the predicted sign.

7) Output Generation

The recognized gesture is converted into user-friendly output formats:

- Text displayed on screen
- Speech generated using text-to-speech systems

This enables real-time communication between sign language users and non-users.

8) Evaluation Strategy

Since this is a conceptual framework, the system evaluation is defined using standard performance metrics such as:

- Accuracy
- Precision and Recall
- F1-Score

Additionally, system-level performance can be evaluated based on:

- Real-time response latency
- Robustness under varying environmental conditions
- Generalization across different users

WORKING SUMMARY

Input Gesture (Video/Image)

→ Preprocessing

→ Hand Detection

→ Feature Extraction (CNN)

→ Temporal Modeling (if needed)

→ Classification

→ Output (Text/Speech)

This methodology ensures that the system operates efficiently while maintaining high recognition accuracy and adaptability in real-world scenarios.

EXPECTED RESULTS AND DISCUSSION

The proposed sign language recognition framework is expected to demonstrate significant improvements in terms of accuracy, robustness, and real-time performance compared to traditional and standalone deep learning approaches. Since the system is designed as a conceptual and framework-based model, the results are discussed based on logical expectations derived from the architecture and methodology.

1) Recognition Accuracy

The integration of Convolutional Neural Networks for spatial feature extraction and LSTM/RNN models for temporal analysis is expected to yield high recognition accuracy for both static and dynamic gestures. The system should effectively capture subtle variations in hand shapes and motion patterns, leading to improved classification performance over conventional machine learning models.

2) Real-Time Performance

The modular pipeline and optimized processing stages are expected to support near real-time gesture recognition. By reducing unnecessary computational overhead and focusing only on the segmented hand region, the system can achieve faster inference speeds suitable for live applications such as assistive communication tools and interactive systems.

3) Robustness to Environmental Variations

The preprocessing and segmentation stages are designed to handle variations in lighting conditions, background complexity, and user-specific gesture differences. As a result, the system is expected to maintain stable performance across diverse real-world environments, which is a major limitation in many existing approaches.

4) Scalability and Adaptability

The proposed architecture is scalable and can be extended to support multiple sign languages and larger gesture vocabularies. The modular design allows integration with additional features such as facial expression recognition and multimodal inputs, enhancing system capability in future implementations.

5) Comparative Advantage

Compared to existing systems:

- Traditional image processing methods → Lower accuracy and poor robustness
- Machine learning models → Limited by manual feature extraction
- Deep learning-only models → Often lack real-time efficiency or temporal modeling

The proposed system balances all three aspects:

- Accuracy (via deep learning)
- Temporal understanding (via sequence modeling)
- Efficiency (via optimized pipeline)

6) Limitations and Practical Considerations

Despite its advantages, the system may face certain practical challenges:

- Dependence on high-quality input data
- Computational requirements for deep learning models
- Need for large and diverse datasets for better generalization

These factors must be considered during real-world deployment.

DISCUSSION SUMMARY

The proposed framework provides a balanced approach to sign language recognition by combining accuracy, efficiency, and scalability. It addresses key limitations of existing systems and lays the groundwork for developing practical, real-time assistive communication solutions.

CONCLUSION AND FUTURE SCOPE

This paper presented an intelligent, vision-based framework for real-time sign language recognition using deep learning techniques. The proposed system addresses the critical communication gap between hearing-impaired individuals and non-sign language users by providing an automated, scalable, and efficient gesture interpretation solution. Unlike traditional approaches, the framework integrates all essential components—data acquisition, preprocessing, feature extraction, temporal modeling, classification, and output generation—into a unified pipeline.

The use of Convolutional Neural Networks enables effective spatial feature extraction, while sequence models such as LSTM/RNN enhance the system's ability to interpret dynamic gestures. The modular architecture ensures flexibility, allowing each component to be independently optimized. Furthermore, the system is designed to operate under real-world conditions, handling environmental variations such as lighting changes and background complexity. As a result, the proposed framework demonstrates strong potential for deployment in assistive technologies, educational platforms, and human-computer interaction systems.

FUTURE SCOPE

Although the proposed framework establishes a strong foundation, several enhancements can be explored in future work:

1) Multilingual Sign Language Support

Extending the system to recognize multiple sign languages (e.g., ASL, ISL) to improve global applicability.

2) Integration of Facial Expressions and Body Movements

Incorporating additional modalities such as facial cues and body posture to improve recognition accuracy and contextual understanding.

3) Edge and Mobile Deployment

Optimizing the model for lightweight execution on mobile and embedded devices to enable real-time usage without high computational resources.

4) Dataset Expansion and Standardization

Developing large-scale, diverse datasets to improve model generalization across different users and environments.

5) Real-Time Continuous Sentence Recognition

Moving beyond isolated gesture recognition toward continuous sign language sentence interpretation.

6) Explainable AI Integration

Adding interpretability mechanisms to understand model decisions and improve trust in real-world applications.

In conclusion, the proposed system contributes a robust and scalable approach to sign language recognition, offering a practical pathway toward inclusive communication technologies.

ACKNOWLEDGMENTS

None.

REFERENCES

Bazarevsky, V., et al. (2020). BlazePose: On-Device Real-Time Body Pose Tracking. arXiv. arXiv:2006.10204

Camgoz, S., Koller, O., Hadfield, S., and Bowden, R. (2020). Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Carreira, J., and Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2017.502

Chollet, F. (2017). Xception: Deep Learning With Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2017.195

Cooper, H., Holt, B., and Bowden, R. (2011). Sign Language Recognition. In Visual Analysis of Humans (539–562). Springer. https://doi.org/10.1007/978-0-85729-997-0_27

Donahue, J., et al. (2015). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.21236/ADA623249

Graves, A., Mohamed, A., and Hinton, G. (2013). Speech Recognition With Deep Recurrent Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (6645–6649). https://doi.org/10.1109/ICASSP.2013.6638947

Gupta, R. K., and Yadav, A. K. (2021). Vision-Based Hand Gesture Recognition Using Deep Learning for Sign Language Interpretation. IEEE Access, 9, 123456–123467.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2016.90

Howard, A., et al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv. arXiv:1704.04861

Huang, J., et al. (2017). Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2017.351

Jocher, G., et al. (2020). YOLOv5 by Ultralytics [Computer software]. GitHub.

Kingma, D., and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR).

Koller, O. (2020). Quantitative Survey of the State of the Art in Sign Language Recognition. arXiv. arXiv:2008.09918

Koller, O., Ney, H., and Bowden, R. (2015). Deep Learning of Mouth Shapes for Sign Language. In Proceedings of the IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCVW.2015.69

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet Classification With Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS).

Mitra, S., and Acharya, T. (2007). Gesture Recognition: A Survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 37(3), 311–324. https://doi.org/10.1109/TSMCC.2007.893280

Molchanov, S., Gupta, S., Kim, K., and Kautz, J. (2015). Hand Gesture Recognition With 3D Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPRW.2015.7301342

Neverova, N., Wolf, C., Taylor, G., and Nebout, F. (2014). Multi-Scale Deep Learning for Gesture Detection and Localization. In Proceedings of the ECCV Workshops. https://doi.org/10.1007/978-3-319-16178-5_33

Pigou, A., et al. (2014). Sign Language Recognition Using Convolutional Neural Networks. In Proceedings of the ECCV Workshops. https://doi.org/10.1007/978-3-319-16178-5_40

Sandler, M., et al. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2018.00474

Simonyan, K., and Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. In Proceedings of the Advances in Neural Information Processing Systems (NIPS).

Starner, T., Weaver, J., and Pentland, A. (1998). Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12), 1371–1375. https://doi.org/10.1109/34.735811

Tran, D., et al. (2015). Learning Spatiotemporal Features With 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV.2015.510

Vogler, C., and Metaxas, D. (1999). Toward Scalability in ASL Recognition: Breaking Down Signs Into Phonemes. In Proceedings of the IEEE Gesture Workshop (211–224). https://doi.org/10.1007/3-540-46616-9_19