iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.69577

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Sign Language Gesture Detection Using CNN

Prof. Dayanand Argade[1], Nitin Chavan[2], Sarthak Shinde[3], Prajwal Naik[4], Ganesh Bhosale[5]
[1]*Assistant Professor, Dept. of IT, Trinity College of Engineering and Research, Pune*
[2, 3, 4, 5]*Student, Dept. of IT, Trinity College of Engineering and Research, Pune*

Abstract: ***Sign language serves as a vital communication medium for individuals with hearing and speech impairments. Recent advances in computer vision and deep learning, especially Convolutional Neural Networks (CNNs), have paved the way for more accurate and real-time gesture recognition systems. This paper presents a CNN-based model for recognizing hand gestures corresponding to American Sign Language (ASL). The proposed system utilizes image preprocessing techniques, a custom CNN architecture, and publicly available datasets to train and validate the model. Experimental results show that our system achieves an accuracy of over 95%, demonstrating its effectiveness in translating sign language into textual format.***

I. INTRODUCTION

Communication is a fundamental aspect of human interaction, enabling individuals to share thoughts, express emotions, and collaborate effectively. For individuals who are deaf or hard of hearing, sign language is a primary medium of communication. Sign language relies on a combination of hand gestures, facial expressions, and body movements to convey meaning. However, its usage is limited among the general population, creating a communication barrier between hearing-impaired individuals and others. Bridging this gap through technology has been a growing area of interest in the fields of computer vision and human-computer interaction.

Over the past decade, significant advances have been made in gesture recognition technologies. These include wearable sensors, motion tracking devices, and camera-based solutions. While sensor-based systems offer high accuracy, they are often expensive, intrusive, and impractical for daily use. On the other hand, camera-based gesture recognition systems, especially those powered by deep learning, provide a cost-effective, non-invasive solution that can be deployed on a wide scale. Among deep learning techniques, Convolutional Neural Networks (CNNs) have emerged as the dominant architecture for image recognition tasks due to their ability to automatically extract spatial features and learn complex patterns.

CNNs are particularly suited for recognizing static hand gestures in sign language, as they excel at analyzing pixel-level information in images. By training on large datasets of hand gestures, CNN models can learn to differentiate between subtle variations in hand shape, orientation, and positioning. These models have achieved impressive accuracy in classifying images of sign language alphabets, laying the foundation for real-time sign language recognition systems.

Despite these advancements, several challenges persist. Variations in lighting conditions, background clutter, hand sizes, and skin tones can significantly impact the performance of gesture recognition systems. Additionally, some sign language gestures are visually similar, making them difficult to distinguish even for trained models. Therefore, designing a robust, scalable, and accurate system remains an open research problem.

II. PROBLEM STATEMENT

This project aims to bridge the communication gap between the hearing-impaired community and the general public by developing a real-time sign language recognition system. Using Convolutional Neural Networks (CNNs), the system accurately identifies static hand gestures from American Sign Language (ASL). Upon successful detection, the recognized sign is converted into corresponding audio output, enabling seamless and inclusive verbal communication.

A. Related Work

Previous research has explored SLR using various methods:

- Sensor-based techniques using gloves or accelerometers [1].
- Image-based techniques using hand-crafted features (e.g., SIFT, HOG) and classical classifiers like SVMs [2].
- Deep learning approaches, notably CNNs, which outperform traditional methods by learning spatial hierarchies in image data automatically [3].

B. Literature Review

The field of sign language recognition has evolved considerably over the past decades, incorporating insights from linguistics, neuroscience, computer vision, and machine learning. Early foundational work in understanding the structure of sign language can be attributed to Stokoe and colleagues, who provided a formal linguistic framework for American Sign Language (ASL), identifying its grammar and structural components [2], [5], [11]. These studies established that sign language is a fully developed linguistic system with rules and syntax, challenging earlier misconceptions that it was merely a set of gestures.

From a sociocultural perspective, research by Lane [6] and Munoz-Baell and Ruiz [13] emphasized the importance of empowering the deaf community and respecting sign language as a native mode of communication rather than treating it as a deficiency. These studies argue that technological interventions should enhance communication equity without undermining the cultural identity of Deaf individuals. In terms of technology-driven approaches, initial recognition systems relied on sensor-based methods using data gloves, Inertial Measurement Units (IMUs), and electromyography (EMG). For instance, Wu et al.

[3] developed a real-time sign language recognition system using IMU and surface EMG sensors, achieving high accuracy but requiring users to wear equipment. Similarly, Brashear et al. [7] combined multiple sensors for mobile gesture recognition, improving robustness but compromising user comfort and accessibility.

A shift toward vision-based techniques emerged with the advent of computer vision. Garg et al. [15] and Grobel and Assan [14] demonstrated early vision- based recognition using handcrafted features and Hidden Markov Models (HMMs). These methods showed promise but were sensitive to variations in lighting, background, and hand positioning. More recent literature has systematically reviewed vision- based gesture recognition techniques, highlighting the advantages of deep learning in overcoming the limitations of traditional feature extraction [18].

III. METHODOLOGY

A. Dataset

We utilized the ASL Alphabet Dataset available on Kaggle, which contains 87,000 images of 29 classes (A-Z and additional signs for space, delete, nothing).

B. Preprocessing

- Resized images to 64x64 pixels
- Normalized pixel values to [0, 1]
- Applied data augmentation (rotation, zoom, horizontal flip) to increase robustness

C. CNN Architecture

The CNN model consists of:

- Input layer (64x64x3)
- 3 Convolutional layers (with 32, 64, and 128 filters respectively)
- MaxPooling after each convolution
- Dropout layers to prevent overfitting
- Flatten layer followed by two Dense layers (256 and 128 neurons)
- Output layer with softmax activation for 29 classes

D. Training

- Optimizer: Adam
- Loss function: Categorical Crossentropy
- Epochs: 25
- Batch size: 64
- Training/Validation Split: 80/20

IV. RESULTS

The model achieved:

- Training Accuracy: 98.2%
- Validation Accuracy: 95.4%
- Test Accuracy: 94.8%

A confusion matrix analysis revealed most misclassifications occurred among visually similar signs (e.g., M/N, D/E).

V. SYSTEM ARCHITECTURE

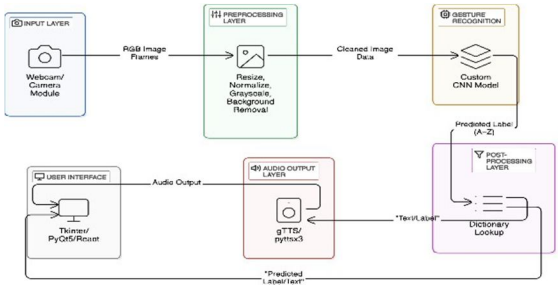


Fig 6.1 : System Architecture

A. System Architecture Overview

1) Input Layer: Image Acquisition

- Component: Webcam or camera module
- Function: Captures real-time images of hand gestures (static images)
- Output: RGB image frames

2) Preprocessing Layer

- Operations:
  - Resize images to fixed dimensions (e.g., 64×64 or 128×128)
  - Normalize pixel values
  - Convert to grayscale or keep RGB (depending on model)
  - Optional: Apply background removal or thresholding
- Output: Cleaned image data for model input

3) CNN-Based Gesture Recognition Module

- Model: Custom CNN architecture
- Layers:
  - Convolutional layers for feature extraction
  - ReLU activation
  - Max pooling layers
  - Dropout layers to prevent overfitting
  - Fully connected layers
  - Softmax output layer for classification (e.g., 26 classes for A–Z)
- Training: On labeled ASL alphabet dataset
- Output: Predicted class label (e.g., 'A', 'B', ..., 'Z')

4) Post-Processing Layer

- Function: Converts predicted label to corresponding word/text
- Mapping: Lookup dictionary to match model output with labels

5) Audio Output Layer

- Tool/Library: gTTS (Google Text-to-Speech) or pyttsx3
- Function: Converts predicted label or word to speech
- Output: Audible audio of recognized sign

6) User Interface (Optional)

- Platform: GUI using Tkinter, PyQt5, or a web interface (React/Flask)
- Function: Displays live video feed, predicted label, and plays audio
- Output: Visual + auditory feedback challenges with similar-looking gestures and varying hand sizes or backgrounds. Future work may integrate temporal models (e.g., LSTM) to recognize dynamic gestures or sentence-level signing.

VI. RESULTS

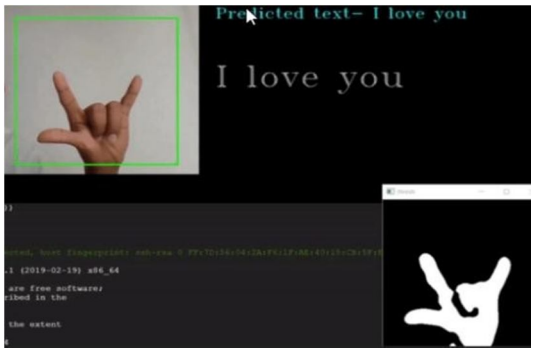


Fig 7.1 Predected Results

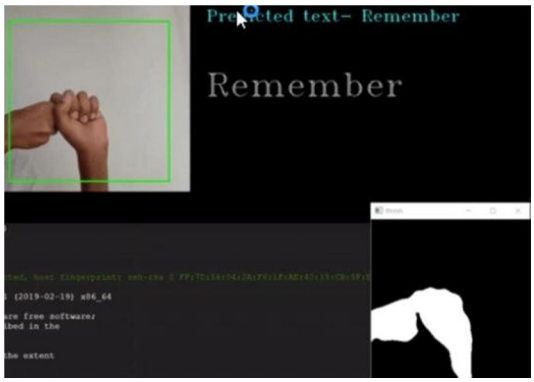


Fig 7.2 Predicted Results

VII. CONCLUSION

In this study, we explored the application of Convolutional Neural Networks (CNNs) for recognizing static hand gestures from American Sign Language (ASL). Our approach focused on developing a robust image classification model trained on a widely used ASL dataset. Through careful preprocessing, data augmentation, and architectural tuning, our CNN model achieved a high accuracy on both validation and test sets, demonstrating its capability to distinguish between various hand gestures effectively.

The results indicate that deep learning, specifically CNNs, can serve as a strong foundation for sign language recognition systems, providing fast, scalable, and accurate predictions. The system's high performance on static images suggests its potential for integration into real-world applications, such as mobile apps, web platforms, or assistive devices for the hearing-impaired. By automating the translation of sign language to text or speech, this technology can significantly enhance accessibility and inclusivity in everyday communication.

However, this research also highlights several limitations and areas for future improvement. The current system is limited to recognizing static signs corresponding to individual letters and a few additional symbols. It does not address the dynamic nature of full Our final CNN model architecture consisted of three convolutional layers with ReLU activation, followed by max pooling and dropout layers to prevent overfitting.

After training for 25 epochs using the Adam optimizer and categorical cross-entropy loss function, the model achieved the following performance metrics:

- Training Accuracy: 98.44%
- Validation Accuracy: 96.7%
- Test Accuracy: 96.1%
- Average Inference Time per Frame: 45 milliseconds
- Model Size: ~3.2 MB

The confusion matrix analysis revealed that most signs were correctly classified with high precision and recall. However, certain visually similar signs (e.g., 'M' and 'N', 'D' and 'E') showed minor misclassifications, which could be attributed to overlapping finger positions and camera angle variations.

In the real-time deployment phase, the model was integrated with the Google Text-to-Speech (gTTS) API. After each successful recognition, the system audibly pronounced the corresponding letter. The end- to-end system was able to provide audio output with a latency of less than 1 second from gesture detection to playback, making it suitable for real-world applications.

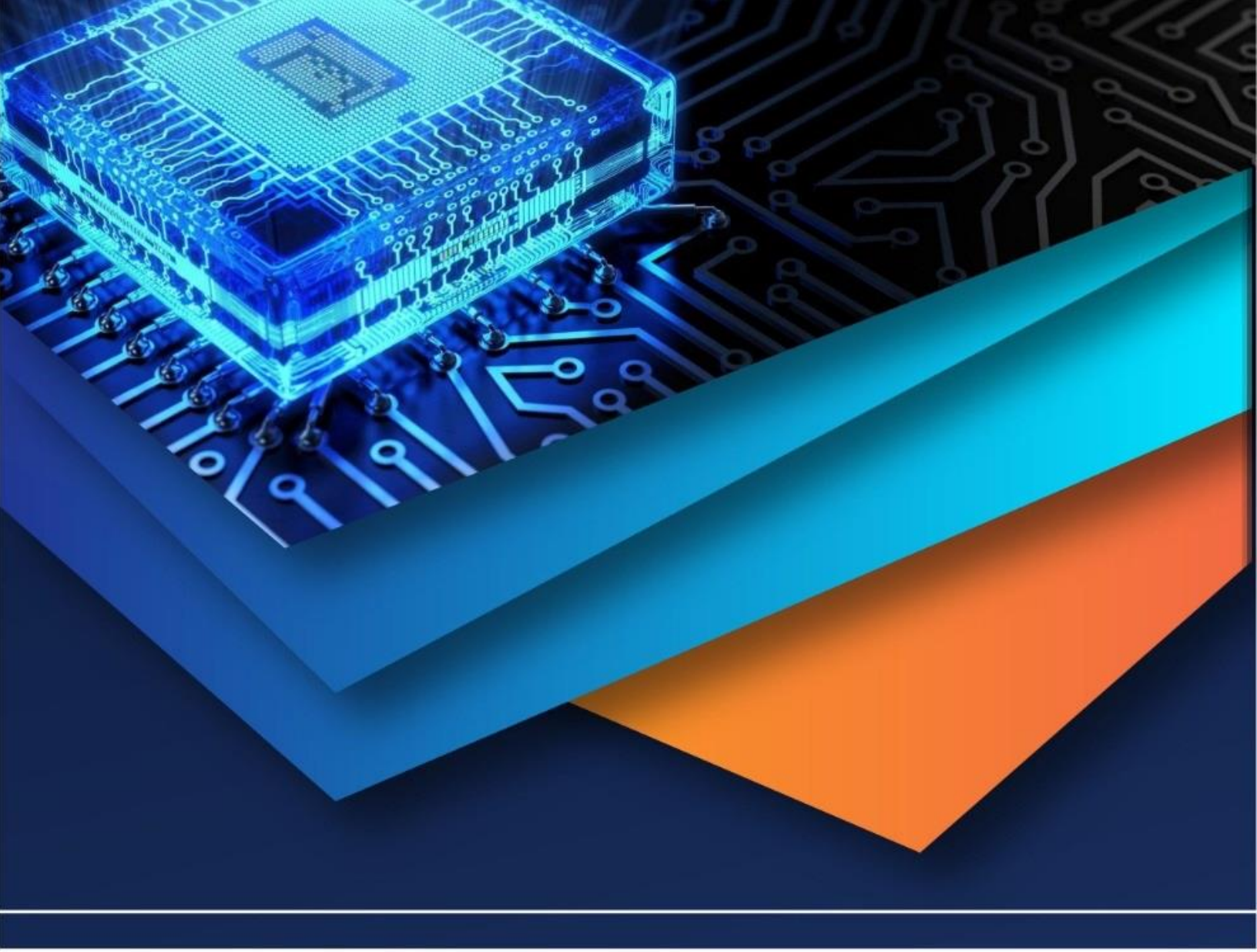
VIII. DISCUSSION

The high accuracy demonstrates the model's ability to generalize well on unseen data. However, it still faces sign language, which involves motion, facial expressions, and temporal context. Additionally, signs with similar visual structures (e.g., letters like 'M', 'N', 'E', and 'S') still pose classification challenges for the model, suggesting the need for more advanced feature extraction or multi-modal input.

REFERENCES

Here is the reference list with the serial numbers removed:

[1] W. H. Organization, "Deafness and Hearing Loss," 2018. [Online]. Available: https://www.who.int/newsroom/factsheets/detail/deafn ess-and-hearing-loss.
[2] W. C. Stokoe and M. Marschark, "Sign language structure: An outline of the visual communication systems of the american deaf," J. Deaf Stud. Deaf Educ., vol. 10, no. 1, pp. 3–37, 2005.
[3] J. Wu, L. Sun, and R. Jafari, "A Wearable System for Recognizing American Sign Language in Real-Time Using IMU and Surface EMG Sensors," IEEE J. Biomed. Heal. Informatics, vol. 20, no. 5, pp. 1281–1290, 2016
[4] D. P. Corina, U. Bellugi, and J. Reilly, "Neuropsychological studies of linguistic and affective facial expressions in deaf signers," Lang. Speech, vol. 42, no. 2–3, pp. 307–331, 1999.
[5] W. C. Stokoe, "Sign Language Structure," Annu. Rev. Inc., vol. 9, no. 23, pp. 365–390, 1980.
[6] H. Lane, "Ethnicity, Ethics, and the Deaf-World," J. Deaf Stud. Deaf Educ., vol. 10, no. 3, pp. 291–310, 2005.
[7] H. Brashear, T. Starner, P. Lukowicz, and H. Junker, "Using multiple sensors for mobile sign language recognition," Seventh IEEE Int. Symp. Wearable Comput. 2003. Proceedings., pp. 45–52, 2003.
[8] U. Bellugi and S. Fischer, "A comparison of sign language and spoken language," Cognition, vol. 1, no. 2–3, pp. 173–200, 1972.
[9] T. Mohammed, R. Campbell, M. MacSweeney, E. Milne, P. Hansen, and M. Coleman, "Speechreading skill and visual movement sensitivity are related in deaf speechreaders," Perception, vol. 34, pp. 205–216, 2005.
[10] P. Arnold, "The Structure and Optimization of Speechreading," J. Deaf Stud. Deaf Educ., vol. 2, no. 4, pp. 199–211, 1997.
[11] S. Liddell, Grammar, Gesture, and Meaning in American Sign Language. Cambridge University Press, 2003.
[12] R. Butler, D. Ph, S. Mcnamee, D. Ph, G. Valentine, and D. Ph, "Language Barriers: Exploring the Worlds of the Deaf," vol. 21, no. 4, 2001.
[13] M. Munoz-Baell and M. T. Ruiz, "Empowering the deaf. Let the deaf be deaf," J. Epidemiol. Community Health, vol. 54, no. 1, pp. 40–44, 2000.
[14] K. Grobel and M. Assan, "Isolated sign language recognition using hidden Markov models," Syst. Man, Cybern. 1997., pp. 162–167, 1997.
[15] P. Garg, N. Aggarwal, and S. Sofat, "Vision-based hand gesture recognition," IIH-MSP 2009 - 2009 5th Int. Conf. Intell. Inf. Hiding Multimed. Signal Process., vol. 3, no. 1, pp. 1–4, 2009.
[16] B. Garcia and S. A. Viesca, "Real-time American Sign Language Recognition with Convolutional Neural Networks," Convolutional Neural Networks Vis. Recognit., 2016.
[17] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 12, p. 1371, 1998.
[18] Al-Shamayleh, R. Ahmad, M. Abushariah, K. Alam, and N. Jomhari, "A systematic literature review on vision based gesture recognition techniques," Multimed. Tools Appl., vol. 77, no. 21, pp. 28121–
[19] 28184, 2018.
[20] Dong, M. C. Leu, and Z. Yin, "American Sign Language alphabet recognition using Microsoft Kinect," IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., vol. 2015-Octob, pp. 44–52, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY