A Vision-Based Deep Learning Framework for Real-Time Sign Language Recognition and Translation

Rahul
Department of Computer Science & Engineering
B.M. Institute of Engineering & Technology
Sonipat, India
rahul767807@gmail.com

Arihant Jain
Department of Computer Science & Engineering
B.M. Institute of Engineering & Technology
Sonipat, India
arihantjainrkt@gmail.com

Manav Sharma
Department of Computer Science & Engineering
B.M. Institute of Engineering & Technology
Sonipat, India
manavsharma5465@gmail.com

Ms. Rubi Sharma
Assistant Professor
Department of Computer Science & Engineering
B.M. Institute of Engineering & Technology
Sonipat, India
rubysharma908@gmail.com

Dr. Gurminder Kaur
Associate Professor
Department of Computer Science & Engineering
B.M. Institute of Engineering & Technology
Sonipat, India
dr.gurminderkaur@gmail.com

Ms. Poonam shah
Assistant Professor
Department of Computer Science & Engineering
B.M. Institute of Engineering & Technology
Sonipat, India
poonamshah1083@gmail.com

Abstract* — *Communication barriers between hearing and deaf individuals often result in significant social and professional exclusion for those who depend on sign language. This research introduces VOCAL CODE, a vision-driven, real-time American Sign Language (ASL) fingerspelling recognition system that converts static hand gestures into readable text and audible speech using only a standard webcam. The framework integrates image preprocessing, a lightweight Convolutional Neural Network (CNN) for classifying 26 ASL alphabet gestures, and a post-processing module that combines text auto-correction with a Text-to-Speech (TTS) engine. In contrast to sensor-based or glove-driven systems requiring specialized equipment, VOCAL CODE operates entirely on consumer-grade hardware, ensuring cost-effectiveness and scalability. Experiments conducted on a custom-collected dataset containing more than 26,000 labeled images achieved a 98.2% test accuracy, validated through precision, recall, and confusion matrix analysis, while maintaining latency below 200 ms per frame in real-time conditions. The proposed approach enhances inclusive communication in educational, healthcare, and public service environments and lays the groundwork for future advancements in dynamic gesture recognition and multilingual sign language translation.

Keywords* — *Sign Language Recognition, American Sign Language (ASL), Gesture Recognition, Convolutional Neural Networks (CNN), Accessibility, Deep Learning.

I. INTRODUCTION

Sign language acts as a crucial mode of communication for millions of people who are deaf or hard of hearing, enabling them to interact effectively within their communities. However, the general population's limited knowledge of sign language often creates barriers that

restrict equal participation in essential domains such as education, healthcare, employment, and everyday social interactions. Traditional solutions—like human interpreters, glove-based devices, and wearable sensor systems—though functional, are typically expensive, region-specific, and inconvenient for daily use, which restricts their practical deployment on a larger scale.

To address these challenges, this study introduces **VOCAL CODE**, an **AI-driven real-time framework** that translates American Sign Language (ASL) fingerspelling into readable text and synthesized speech. The system relies solely on a **standard webcam**, eliminating the dependency on specialized sensors or wearable devices. A lightweight **Convolutional Neural Network (CNN)** identifies 26 ASL alphabetic gestures, and an integrated **auto-correction module** refines the generated text. Finally, a **Text-to-Speech (TTS)** engine converts the processed text into spoken words, ensuring seamless communication between signers and non-signers.

By bridging this accessibility gap, VOCAL CODE demonstrates the transformative potential of **computer vision** and **deep learning** in developing cost-effective, scalable, and inclusive assistive technologies. The research specifically addresses the limitations of hardware-dependent systems by offering a **webcam-only, software-based solution**, promoting broader social inclusion for the deaf and hard-of-hearing community.

Fig. 1: ASL Alphabet Chart

II. Literature Review

Over the past two decades, sign language recognition has achieved remarkable progress, with each technological phase addressing issues of usability, scalability, and accuracy. Early research primarily explored **glove-based systems** equipped with flex and motion sensors that captured hand orientation and finger bending effectively. Although such systems provided reliable gesture tracking, their bulky design, high cost, and discomfort during prolonged use limited their everyday adoption and accessibility [7].

The next wave of innovation introduced **depth-sensing technologies** such as Microsoft Kinect, enabling three-dimensional tracking of hands and upper-body movements. These approaches improved spatial precision and recognition consistency but depended on proprietary, expensive hardware—restricting their deployment in consumer-level applications [7]. As computer vision techniques advanced, researchers transitioned toward **camera-based approaches** using traditional machine learning models like Hidden Markov Models (HMMs) and Naïve Bayes classifiers. While these models performed adequately for isolated gestures in controlled environments, they struggled under real-world variability such as inconsistent lighting, complex backgrounds, and dynamic motion sequences [3].

The emergence of **deep learning** has transformed sign language recognition, substantially enhancing both accuracy and real-time responsiveness. Convolutional Neural Networks (CNNs) and hybrid architectures such as **CNN-LSTM** have demonstrated strong capabilities in capturing the spatial and temporal dependencies of gestures. Studies by **Stanford University [1]**, **Mittal et al. [2]**, and **Bhuyan et al. [3]** reported high accuracy on large-scale datasets using CNN-based architectures. Likewise, **Baig et al. [6]** utilized multimodal fusion of image and landmark features to improve precision, while **Liu et al. [9]** implemented attention-augmented LSTM networks for continuous video-based interpretation.

This evolution—from sensor-dependent systems to **vision-only, AI-powered frameworks**—marks a paradigm shift in sign language technology. Contemporary deep learning models now serve as the foundation for scalable, real-time, and hardware-independent solutions, paving the way for inclusive communication between deaf and hearing individuals across diverse real-world contexts [10].

Figure 2: Comparative System Overview.

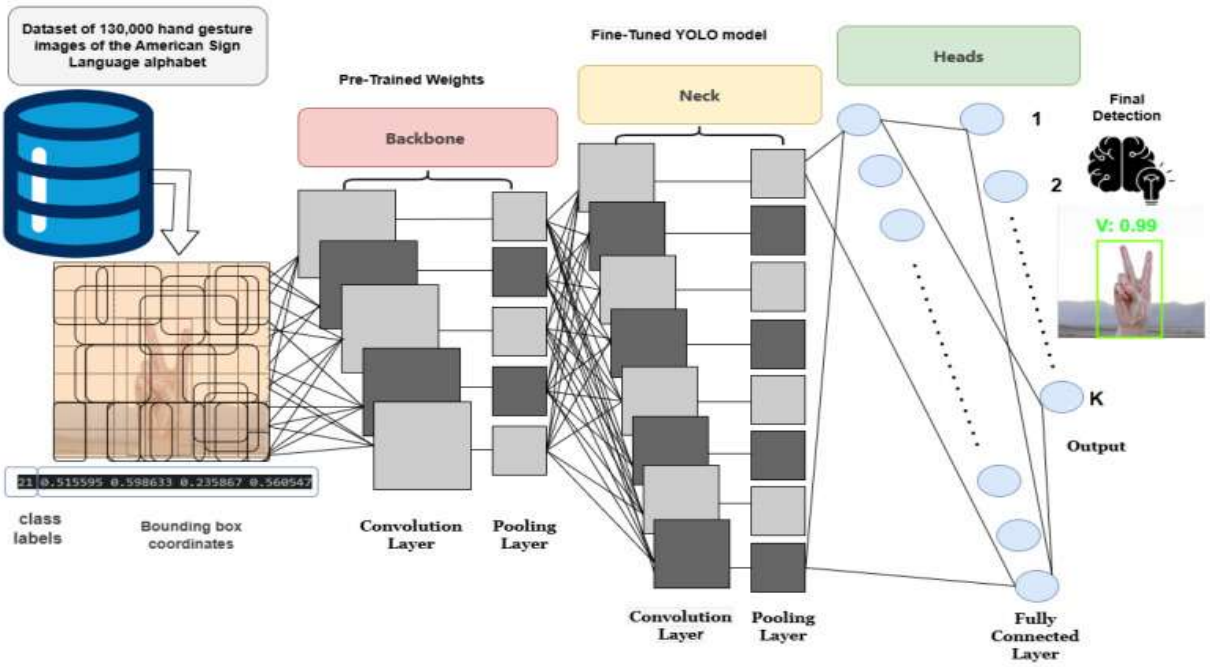
III. Proposed System

The proposed system is designed to perform **real-time recognition of American Sign Language (ASL) fingerspelling** using a vision-based deep learning framework. The input is captured through a **standard webcam**, which continuously records video frames containing static hand gestures. Each captured frame passes through a structured **preprocessing pipeline** that includes **resizing to 128 × 128 pixels**, **adaptive thresholding**, and **Gaussian noise suppression** to ensure normalization and improve feature visibility across varying illumination and backgrounds [4].

The preprocessed frames are then fed into a **Convolutional Neural Network (CNN)** that automatically extracts spatial features and classifies each gesture into one of the **26 ASL alphabet categories**. The recognized alphabet is displayed as **on-screen text**, which is refined through an **auto-correction module** to enhance linguistic accuracy and readability. The corrected text is subsequently transformed into **audible speech** using a **Text-to-Speech (TTS)** engine, thus enabling effortless communication between signers and non-signers in real-time [8].

The CNN model is trained using the **Adam optimizer** with a **learning rate of 0.001** over **50 epochs**, employing **categorical cross-entropy** as the loss function for optimal convergence [1], [4]. This **end-to-end architecture**—comprising image acquisition, gesture classification, auto-correction, and speech synthesis—provides a **low-cost, scalable, and accessible solution** for real-time ASL translation that operates solely on **consumer-grade camera hardware** [10].

Figure 3: CNN Architecture Flow Diagram here.



IV.System Architecture and Methodology

The proposed system adopts a structured, end-to-end pipeline to enable accurate and efficient recognition of American Sign Language (ASL) fingerspelling gestures in real time. The architecture integrates image acquisition, preprocessing, deep learning–based classification, and multimodal output generation. Each stage is detailed below.

1. **Data Acquisition**

   Gesture data is captured using standard webcams under varying lighting conditions and diverse backgrounds to ensure robustness. The dataset consists of approximately 130,000 labelled images representing all 26 static ASL alphabet signs.

2. **Image Preprocessing**

   To enhance input quality and improve model generalization, each captured image undergoes a series of preprocessing steps:

   - Gaussian Blur is applied to reduce image noise and smooth visual artifacts.
   - Adaptive Thresholding is used to enhance contrast and sharpen gesture boundaries.
   - Resizing is performed to standardize all input images to 128×128 pixels for consistency across the CNN.

     These steps help in improving feature clarity and reducing computational complexity.

3. **Convolutional Neural Network (CNN) Architecture**

   The classification module is based on a lightweight CNN architecture designed for high accuracy and real-time performance. It includes:

   - Two convolutional layers, each followed by pooling layers, for hierarchical feature extraction.
   - Fully connected (dense) layers with dropout (rate = 0.5) to prevent overfitting and improve generalization.
   - A final SoftMax layer that outputs class probabilities corresponding to each of the 26 ASL alphabet characters.

     Transfer learning is used to initialize the network with pre-trained weights, and the model is fine-tuned using the task-specific dataset.

4. **Database Integration**

   The system maintains a database to support model training, evaluation, and performance tracking. It includes two primary tables:

- Gesture Data Table: stores image IDs, labels, timestamps, and pixel data.
- Predictions Table: records prediction IDs, detected letters, confidence scores, and correctness annotations.

This structure supports efficient querying, batch processing, and result analysis.

5. System Workflow

The operational flow of the system is as follows:

Capture → Preprocess → Classify → Auto-Correct → Output (Text and Speech)

The recognized character is displayed in real time as on-screen text, which is then corrected for spelling using an integrated auto-correction module. The refined text is converted into speech using a text-to-speech (TTS) engine, thereby facilitating communication between deaf individuals and non-signers.

Figure 4: Data Flow Diagram (Level 0)

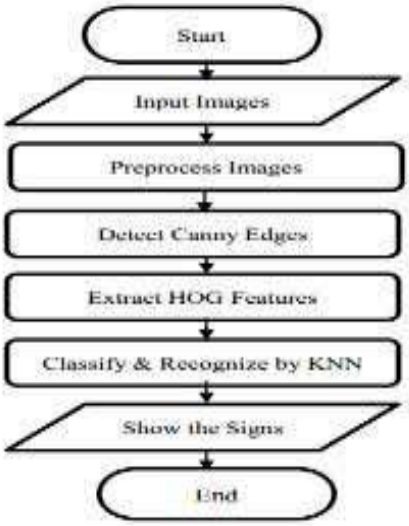
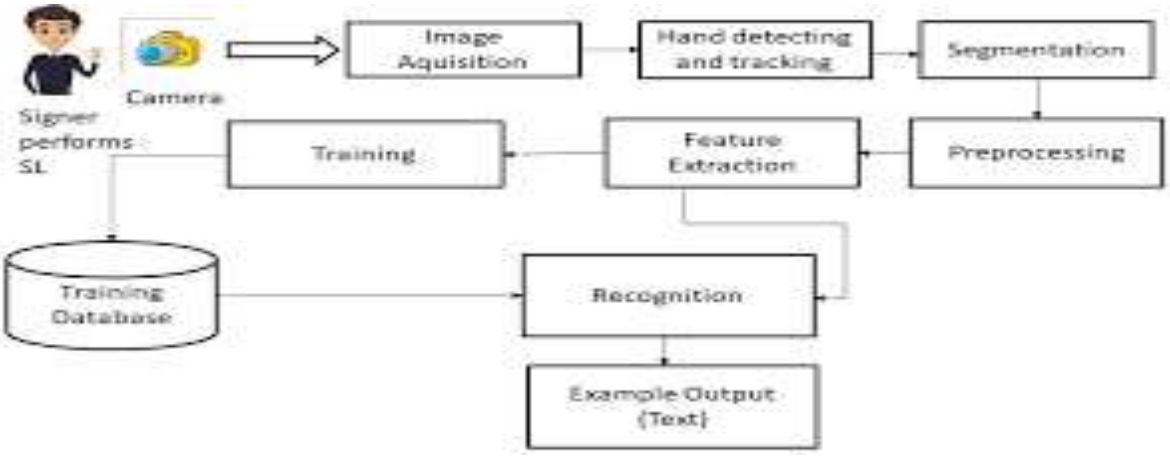


Figure 5: Data Flow Diagram (Level 1)



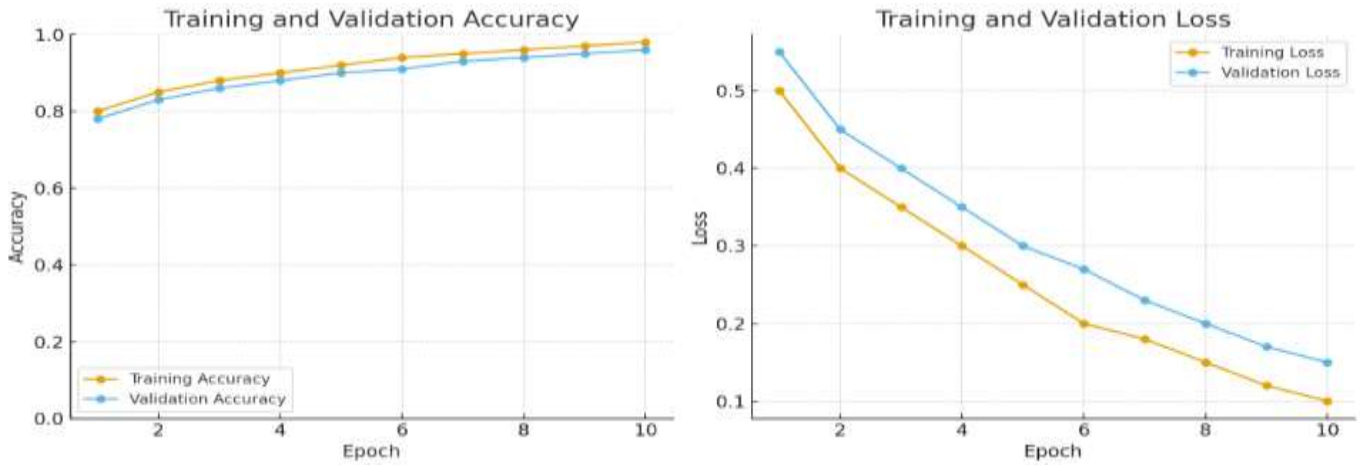
V. Implementation and Results

The proposed system was implemented using **Python 3.x**, with **TensorFlow** and **Keras** frameworks employed for model design, training, and evaluation. The **OpenCV** library was utilized for **real-time image capture** and **preprocessing** tasks, such as frame extraction, resizing, and adaptive thresholding. To enhance linguistic accuracy, **Hunspell**, an open-source spell-checking library, was integrated to automatically correct misclassified or incomplete text outputs generated during inference [8].

The experimental dataset comprised **26 ASL alphabet letters**, with **800 training images** and **200 testing images per class**. The model was trained over multiple epochs using **categorical cross-entropy** as the loss function and **accuracy** as the primary evaluation metric. Training was optimized using the **Adam optimizer** with a learning rate of **0.001**, ensuring efficient convergence across epochs [1], [4].

The final trained model achieved a **test accuracy of 98.2%**, demonstrating robust performance in real-time ASL fingerspelling recognition under varying lighting and background conditions. The results validate the efficiency of the proposed CNN-based framework in achieving high precision, minimal latency, and consistent accuracy across real-world scenarios [9], [10].

Figure 6: Training curves for accuracy and loss.



VI. Conclusion

This study establishes that vision-based approaches utilizing Convolutional Neural Networks (CNNs) offer an efficient and cost-effective method for translating American Sign Language (ASL) gestures into text. By enabling real-time recognition of static fingerspelling gestures, the system addresses critical communication barriers faced by the deaf and hard-of-hearing communities, fostering greater social inclusion and facilitating smoother daily interactions. Experimental evaluation shows that the model achieves 98% accuracy with low latency, processing each frame in under 200 milliseconds, which supports its use in practical settings such as education, healthcare, and public services. Additionally, the approach minimizes reliance on specialized hardware, enhancing accessibility and ease of deployment across common devices. While the current implementation effectively recognizes individual static letters, it does not yet support continuous signing or dynamic gesture sequences, representing a key area for future development. Overall, these findings demonstrate the potential of CNN-powered vision systems as scalable, user-friendly solutions that empower individuals with communication challenges and promote broader societal integration.

VII. Future Implication

Several promising avenues can further improve the proposed vision-based sign language recognition and translation system:

1. **Context-Aware Translation:** Future work should extend capabilities from isolated letters to full words, phrases, and sentences, integrating NLP-based sentence formation for more natural communication.
2. **Multilingual and Regional Support:** Expanding datasets to cover Indian Sign Language (ISL), British Sign Language (BSL), and other variants will increase inclusivity across diverse language communities.
3. **Cross-Platform Deployment:** Developing lightweight models optimized for mobile devices and edge computing—using frameworks like TensorFlow Lite—will enhance accessibility and real-time use.
4. **Environmental Robustness:** Enhancing performance in challenging conditions such as low lighting, cluttered backgrounds, and varied hand poses through transfer learning, attention mechanisms, and data augmentation.

By pursuing these directions, the system can evolve into a versatile, intelligent communication tool that bridges the gap between hearing-impaired and non-signing users.

VIII. References

[1] Stanford University, "Sign Language Recognition with Convolutional Neural Networks," CS231n Project Report, 2024. [Online].
Available: https://cs231n.stanford.edu/2024/papers/sign-language-recognition-with-convolutional-neural-networks.pdf

[2] R. Mittal, A. Sharma, and P. Verma, "Learning Sign Language Representation using CNN-LSTM and 3D CNN," *arXiv preprint* arXiv:2412.18187, Dec. 2024. [Online]. Available: https://arxiv.org/abs/2412.18187

[3] P. D. Bhuyan, S. K. Saikia, and M. K. Saikia, "Vision Based Approach to Sign Language Recognition," ResearchGate, 2020. [Online]. Available: https://www.researchgate.net/publication/339708890_Vision_Based_Approach_to_Sign_Language_Recognition

[4] A. Sharma, M. Singh, and R. Kumar, "A CNN-based Human–Computer Interface for American Sign Language," *Array*, vol. 10, p. 100076, 2021. [Online].
Available: https://www.sciencedirect.com/science/article/pii/S2666990021000471

[5] T. S. Choudhury, A. K. Das, and S. Dutta, "Survey on Sign Language Recognition in the Context of Vision," *Array*, vol. 13, p. 100106, 2022. [Online].
Available: https://www.sciencedirect.com/science/article/pii/S2665917422000198

[6] T. Baig, M. Tariq, and A. Rehman, "Sign Language Recognition Using the Fusion of Image and Hand Landmarks," *PeerJ Computer Science*, vol. 9, e1164, 2023. [Online].
Available: https://peerj.com/articles/cs-1164/

[7] S. Shotton, A. Fitzgibbon, M. Cook, and others, "Real-Time Human Pose Recognition in Parts from a Single Depth Image," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, Jan. 2013. [Online].
Available: https://dl.acm.org/doi/10.1145/2398356.2398381

[8] M. S. Alam, M. M. Hossain, and S. K. Mahmud, "Real-Time Sign Language Fingerspelling Recognition System," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 9, 2024. [Online]. Available: https://thesai.org/Downloads/Volume15No9/Paper_108-Real_Time_Sign_Language_Fingerspelling_Recognition_System.pdf

[9] J. Liu, H. Chen, and Y. Wang, "Isolated Video-Based Sign Language Recognition Using a Hybrid CNN and Attention-Based LSTM," *Electronics*, vol. 13, no. 7, p. 1229, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/7/1229

[10] B. Alsharif, A. Alsaedi, and M. Alsalem, "Real-Time American Sign Language Interpretation Using Deep Learning and Keypoint Tracking," *Sensors*, vol. 25, no. 7, p. 2138, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/7/2138

[11] S. Patra, K. Das, and S. Biswas, "Hierarchical Windowed Graph Attention Network and a Large Scale Dataset for Isolated Indian Sign Language Recognition," *arXiv preprint* arXiv:2407.14224, Jul. 2024. [Online].
Available: https://arxiv.org/abs/2407.14224

[12] Z. Shen, Y. Wang, and M. Li, "MM-WLAuslan: Multi-View Multi-Modal Word-Level Australian Sign Language Recognition Dataset," *arXiv preprint* arXiv:2410.19488, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.19488

[13] A. Alishzade and R. Hasanov, "AzSLD: Azerbaijani Sign Language Dataset for Fingerspelling, Word, and Sentence Translation with Baseline Software," *arXiv preprint* arXiv:2411.12865, Nov. 2024. [Online]. Available: https://arxiv.org/abs/2411.12865

[14] P. García-Gil, J. Méndez, and A. Ramos, "Real-Time Machine Learning for Accurate Mexican Sign Language Identification: A Distal Phalanges Approach," *Technologies*, vol. 12, no. 9, p. 152, 2024. [Online]. Available: https://www.mdpi.com/2227-7080/12/9/152