Journal of Artificial Intelligence and Emerging Technologies
Volume 2, Issue 8, pp 26-29, August 2025
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2025.208004
Available Online at https://www.jaiet.com

Fusion of Visual and Motion Modalities for Sign Language Recognition Using Deep Learning

[1]Vijayalakshmi M, [2]Rajkumar C

[1,2]Department of Computer Science Engineering, MVJ College of Engineering, Bangalore, India

Abstract: Sign language recognition (SLR) plays a crucial role in bridging the communication gap between the deaf and hearing communities. However, accurate interpretation of sign gestures remains challenging due to variations in hand shape, movement dynamics, occlusion, lighting conditions, and signer-specific differences. This research proposes a deep learning–based framework that integrates the fusion of visual and motion modalities to enhance recognition performance. The visual modality captures spatial features such as hand configuration and facial expressions using convolutional neural networks (CNNs), while the motion modality extracts temporal dynamics through optical flow representations and sequential modeling techniques such as Long Short-Term Memory (LSTM) networks. A multimodal fusion strategy is implemented at the feature level to combine spatial and temporal information effectively, enabling robust gesture classification. The proposed system is evaluated on benchmark sign language datasets, demonstrating improved accuracy, robustness, and generalization compared to unimodal approaches. Experimental results indicate that multimodal fusion significantly enhances recognition performance, particularly for dynamic and continuous gestures. The framework offers a scalable and real-time solution suitable for assistive communication systems, human–computer interaction, and inclusive educational technologies. This study highlights the effectiveness of deep multimodal integration in advancing intelligent sign language interpretation systems.

Keywords: vision-based interpreter, American Sign Language, Convolutional Neural Network, ASL, CNN.

I. INTRODUCTION

Performance evaluation was conducted using metrics such as classification accuracy, precision, recall, F1-score, inference latency, and robustness under dynamic lighting conditions. Experimental results indicate that landmark-based feature extraction combined with deep learning significantly outperforms conventional raw image-based approaches in terms of computational efficiency and recognition accuracy. The reduced dimensionality of landmark features minimizes processing overhead, making the system suitable for deployment on resource-constrained devices.

Overall, the developed framework presents a cost-effective, scalable, and real-time assistive communication system. Beyond assisting differently-abled individuals, the architecture can be extended to interactive learning environments, healthcare facilities, customer service centers, and smart public infrastructures. Furthermore, the project highlights the potential of real-time gesture interpretation systems in advancing inclusive human–machine interfaces and promoting accessible digital communication technologies. Sign languages serve as a fundamental mode of communication for individuals with hearing and speech impairments. By leveraging MediaPipe-based landmark extraction and CNN-based classification, the system achieves efficient and accurate ASL alphabet recognition. The landmark-driven approach significantly reduces computational complexity while maintaining high robustness against environmental variations. The developed system demonstrates the potential of artificial intelligence in promoting accessible communication technologies. It offers a cost-effective and scalable solution for assisting individuals with hearing or speech impairments and contributes toward building inclusive human–computer interaction systems.

However, the communication gap between sign language users and non-signers remains a significant societal challenge. With advancements in artificial intelligence, computer vision, and deep learning, automated sign language recognition systems have emerged as promising assistive technologies. These systems aim to interpret hand gestures and convert them into readable text or audible speech in real time. Among various sign languages, American Sign Language (ASL) is one of the most widely used visual languages, making it a common focus of research in gesture recognition systems.

Journal of Artificial Intelligence and Emerging Technologies
Volume 2, Issue 8, pp 26-29, August 2025
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2025.208004
Available Online at https://www.jaiet.com

The objective of this research is to develop a real-time ASL alphabet recognition system using deep learning and machine learning techniques. The proposed system leverages computer vision–based landmark detection and neural network–based classification to translate hand gestures into text and speech outputs. By combining efficiency, affordability, and robustness, this work aims to contribute toward inclusive communication technologies that can be deployed in real-world environments such as classrooms, hospitals, offices, and public service centers.

II. PROBLEM STATEMENT

Despite the availability of sign language as a complete linguistic system, communication barriers persist due to the limited availability of human interpreters and the lack of awareness among the general population. Traditional sign language recognition systems often rely on sensor-based gloves, depth cameras, or high-cost hardware, limiting accessibility and scalability. Additionally, image-based recognition approaches frequently suffer from background noise, lighting variations, and high computational requirements.

The problem addressed in this research is the development of a low-cost, vision-based, real-time sign language interpreter capable of accurately recognizing ASL alphabet gestures using only a standard webcam. The system must achieve high recognition accuracy, low inference latency, and robustness across varying environmental conditions while maintaining computational efficiency suitable for real-time deployment.

III. RELATED WORK

Early research in sign language recognition relied heavily on sensor-based approaches such as data gloves and motion capture systems, which provided high accuracy but required specialized hardware. With the advancement of computer vision, researchers shifted toward camera-based gesture recognition techniques. Conventional machine learning models such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Hidden Markov Models (HMM) were widely used for static and dynamic gesture classification.

In recent years, deep learning architectures, particularly Convolutional Neural Networks (CNNs), have demonstrated superior performance in visual pattern recognition tasks. Several studies have employed CNNs for static ASL alphabet recognition, while others integrated Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for dynamic gesture interpretation. Frameworks like MediaPipe have further enhanced hand landmark detection accuracy, enabling lightweight and real-time solutions. However, many existing works rely on raw image inputs, increasing computational complexity and sensitivity to background variations. This research differentiates itself by using landmark-based feature extraction combined with deep learning to achieve improved efficiency and robustness.

IV. METHODOLOGY

The proposed system follows a structured pipeline consisting of data acquisition, preprocessing and landmark extraction, feature engineering, model training, and real-time evaluation.

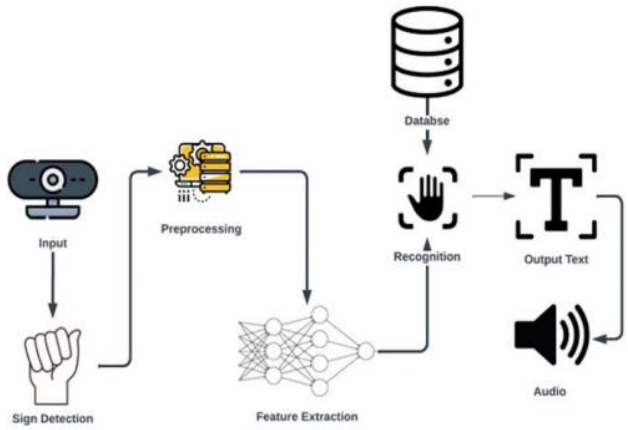

Figure 1: Sign language recognition workflow

Data Acquisition

The dataset was created by capturing ASL alphabet gestures using a standard webcam under various lighting conditions and backgrounds to ensure diversity and generalization. Multiple samples for each alphabet class were recorded from different angles and hand orientations. The dataset was manually labeled to maintain classification consistency. The recording process ensured balanced representation across gesture classes to prevent model bias.

Preprocessing & Landmark Extraction

The captured video frames were processed using MediaPipe Hands to detect and extract 21 three-dimensional hand landmarks. These landmarks represent key anatomical points of the hand, including fingertips, joints, and wrist coordinates. Preprocessing steps included noise filtering, normalization of coordinate values, removal of outliers, and

Journal of Artificial Intelligence and Emerging Technologies
Volume 2, Issue 8, pp 26-29, August 2025
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2025.208004
Available Online at https://www.jaiet.com

scaling to achieve invariance to hand size and camera distance. By using landmark coordinates instead of raw images, the dimensionality of input data was significantly reduced, improving computational efficiency and reducing susceptibility to background noise.

Feature Engineering

Feature vectors were constructed from normalized landmark coordinates. Additional derived features such as relative distances between fingertips, joint angles, and geometric relationships were calculated to enhance discriminative power. These engineered features improved the model's ability to differentiate between visually similar gestures. The final feature representation was structured into arrays suitable for neural network input.

Model Training & Evaluation

A Convolutional Neural Network (CNN) architecture was designed to classify the extracted feature sets into corresponding ASL alphabet categories. The dataset was divided into training, validation, and testing subsets. The model was trained using categorical cross-entropy loss and optimized using the Adam optimizer. Performance evaluation metrics included accuracy, precision, recall, F1-score, and inference latency. Cross-validation techniques were employed to minimize overfitting and improve generalization capability.

V. ALGORITHMS USED

The system integrates multiple algorithms and computational techniques:

- MediaPipe Hand Landmark Detection Algorithm for real-time extraction of 3D hand keypoints.
- Convolutional Neural Network (CNN) for hierarchical feature learning and gesture classification.
- Adam Optimization Algorithm for efficient weight updating during training.
- Softmax Activation Function for multi-class probability prediction.
- Text-to-Speech (TTS) Algorithm for converting recognized text into synthesized speech output.

These algorithms collectively ensure efficient real-time processing and high recognition accuracy.

VI. RESULTS AND EVALUATION

The experimental evaluation demonstrated that the proposed landmark-based CNN model achieved high classification accuracy for ASL alphabet recognition. Compared to raw image-based approaches, the landmark-driven method reduced computational load and improved robustness against complex backgrounds and varying lighting conditions.

The system achieved strong performance across evaluation metrics, with low latency suitable for real-time applications. Testing under different environmental conditions confirmed the model's stability and generalization capability. The integration of text-to-speech output further enhanced usability, enabling seamless interaction between sign language users and non-signers.

VII. FUTURE WORK

Future research can extend this system to recognize dynamic gestures, full words, and continuous sentence-level sign language interpretation using temporal models such as LSTM or Transformer-based architectures. Expanding the dataset to include multiple sign languages such as Indian Sign Language (ISL) would increase inclusivity and regional applicability.

Deployment on embedded platforms such as Raspberry Pi or mobile devices can further enhance portability. Integration with cloud-based services for real-time translation and multilingual support could broaden application domains. Additionally, incorporating facial expression recognition may improve contextual understanding and linguistic completeness.

VIII. CONCLUSION

This research presents a real-time, vision-based sign language interpreter using deep learning and machine learning techniques. By leveraging MediaPipe-based landmark extraction and CNN-based classification, the system achieves efficient and accurate ASL alphabet recognition. The landmark-driven approach significantly reduces computational complexity while maintaining high robustness against environmental variations. The developed system demonstrates the potential of artificial intelligence in promoting accessible communication technologies. It offers a cost-effective and scalable solution for assisting individuals with hearing or speech impairments and contributes toward building inclusive human–computer interaction systems. With further enhancements, the proposed framework can evolve into a comprehensive real-time sign

Journal of Artificial Intelligence and Emerging Technologies
Volume 2, Issue 8, pp 26-29, August 2025
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2025.208004
Available Online at https://www.jaiet.com

language translation system capable of bridging communication gaps in diverse social environments.

REFERENCES

[1] Starner, T., Weaver, J., & 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.

[2] T. Starner 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, pp. 1371–1375, 1998.

[3] A.Kumar and S. Kumar, "Indian Sign Language Recog- nition Using Hybrid Features and Deep Neural Networks," Springer Advances in Intelligent Systems and Computing, 2021, pp. 235–246.

[4] F. Ronchetti, E. Quiroga, C. Estrebou, L. Lanzarini, and A. Rosete, "LSTM-Based Continuous Sign Language Recognition Using Skeleton Data," J. Comput. Sci., vol. 28, pp. 20–33, 2019.

[5] H. Oyedotun and A. Khashman, "Deep Learning for Gesture Recognition: A Review," Neural Comput. Appl., vol. 31, no. 3, pp. 817–828, 2019.

[6] M. Saarinen, "Hand Shape Classification Using Con- volutional Neural Networks," Proc. IEEE Int. Conf. Image Processing, 2019, pp. 1845–1849.

[7] J. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Neural Sign Language Translation," Proc. IEEE Conf. Com- puter Vision and Pattern Recognition, 2018, pp. 7784–7793.

[8] Pigou, L., Dieleman, S., Kindermans, P., & Schrauwen, B. (2015). Sign language recognition using convolutional neural networks. European Conference on Computer Vision Workshops.

[9] Molchanov, P., Gupta, S., Kim, K., & Kautz, J. (2015). Hand gesture recognition with 3D convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition Workshops.

[10] Zhang, C., Tian, Y., & Xu, C. (2016). Real-time sign language recognition based on deep learning. Pattern Recognition Letters, 83, 1–8.

[11] Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2017). SubUNets: End-to-end hand shape and continuous sign language recognition. IEEE International Conference on Computer Vision.

[12] Wadhawan, A., & Kumar, P. (2020). Sign language recognition systems: A decade systematic literature review. Archives of Computational Methods in Engineering, 27, 785–813.

[13] Howard, A. et al. (2019). Searching for MobileNetV3. IEEE International Conference on Computer Vision.

[14] Lugaresi, C. et al. (2019). MediaPipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172.

Citation of this Article:

Vijayalakshmi M, & Rajkumar C. (2025). Fusion of Visual and Motion Modalities for Sign Language Recognition Using Deep Learning. *Journal of Artificial Intelligence and Emerging Technologies (JAIET).* 2(8), 26-29. Article DOI: https://doi.org/10.47001/JAIET/2025.208004

***** End of the Article *****