Samruddhi Vijay Wakalkar, 2026, 14:1
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Intelligent Sign Language Interpretation System Using Multi- Modal Deep Learning Architectures

Samruddhi Vijay Wakalkar, Sanskruti Vijay Wakalkar, Siddhi Nanasaheb Hon, Shravani Kishor Mahale, Gauri Sanjay Lad

Department of Artificial Intelligence and Data Science, Sanjivani University, India

Abstract- This project presents a real-time American Sign Language (ASL) recognition system using a standard webcam. Communication between deaf or hard-of-hearing individuals and the hearing community is often limited by the high cost and limited availability of professional interpreters. To address this, the proposed system employs an ensemble deep-learning approach that combines a Convolutional Neural Network (CNN) for hand shape recognition, a Graph Neural Network (GNN) to capture finger and joint relationships, and a Vision Transformer to focus on key visual regions while minimizing background noise. By fusing these complementary models, the system achieves enhanced recognition accuracy. The framework was trained and evaluated on a dataset of approximately 87,000 labeled images covering the complete ASL alphabet along with additional gestures such as space and delete. Experimental results demonstrate an accuracy exceeding 95%, outperforming existing methods. The system supports real-time interaction with an average inference time of about 85 milliseconds per gesture. It is deployed through a browser-based interface and requires no specialized hardware beyond a standard webcam. This solution provides an accessible, low-cost alternative to traditional interpretation services and promotes inclusive communication across educational, healthcare, and public environments.

Keywords: "American Sign Language", "ASL recognition, deep learning", "computer vision", "multi- modal models", "CNN", "GNN", "Vision Transformer", "ensemble learning", "real-time detection", "webcam", "accessibility", "sign language interpretation", "machine learning", "assistive technology".

I. INTRODUCTION

Sign language serves as a crucial medium of communication for millions of individuals worldwide, particularly those who are deaf or hard of hearing. Unlike spoken languages, sign language relies on structured hand movements, facial expressions, and body postures to convey meaning, enabling effective and expressive communication without the use of sound. American Sign Language (ASL) is among the most widely recognized sign languages and is the fourth most commonly used language in the United States. With more than 466 million people globally experiencing disabling hearing loss, the importance of sign language continues to grow in everyday communication, education, and social interaction. Sign language is not merely a collection of gestures but a complete and natural language with its own grammar, syntax, and linguistic rules.

This understanding was firmly established in the 1960s through the pioneering work of linguist William Stokoe at Gallaudet University, whose research highlighted the complexity and legitimacy of sign languages. Since then, sign language has gained increasing academic and cultural recognition. Today, more than a hundred colleges and universities in the United States accept ASL to fulfill foreign language requirements, reflecting its growing acceptance as a legitimate academic discipline. Despite its significance, sign language users frequently face communication barriers when interacting with the hearing population.

A major challenge is the limited availability and high cost of professional interpreters, with hourly fees ranging from $50 to $150. This shortage is particularly critical in essential domains such as healthcare, legal proceedings, and emergency situations, where clear communication is vital. Additionally, a large proportion of deaf individuals face difficulties with written language due to limited educational access, making sign language indispensable. Beyond the deaf community, sign language benefits individuals with autism, cerebral palsy, and other conditions affecting verbal

communication. In response to these challenges, this project aims to develop an AI-powered sign language interpretation system that provides accurate, real-time translation, enhancing accessibility, reducing dependency on human interpreters, and promoting social inclusion and communication equity.

RELATED WORK

Sign language recognition has been widely studied using computer vision and machine learning techniques. Early approaches relied on handcrafted features with traditional classifiers such as SVMs and HMMs, which showed limited robustness to lighting variations, background clutter, and signer diversity. With the advancement of deep learning, Convolutional Neural Networks (CNNs) became the primary method for recognizing static sign language gestures, achieving promising accuracy on ASL alphabet datasets. However, CNN-based models mainly focus on visual appearance and often fail to capture structural relationships between hand joints and fingers.

To address temporal dynamics, several studies introduced Recurrent Neural Networks (RNNs) and LSTM-based models for video-based sign recognition. While these approaches improved performance for dynamic gestures, they increased computational complexity and often struggled with real-time deployment. Object detection frameworks such as YOLO were also explored for real-time sign detection but were limited in fine-grained fingerspelling recognition.

Recent research has incorporated attention mechanisms and Vision Transformers to improve spatial focus and robustness against background noise. Additionally, landmark-based methods using MediaPipe combined with deep learning have shown improved accuracy by modeling hand geometry. Despite these advances, most existing systems rely on single-model or limited multimodal architectures. The proposed work differs by employing a multi-modal ensemble of CNN, GNN, and Vision Transformer models, enabling robust visual, structural, and attention-based feature learning for accurate real-time ASL recognition using standard webcam hardware.

II. SYSTEM ARCHITECTURE

The proposed system follows a multi-modal ensemble architecture for real-time American Sign Language (ASL) recognition using a standard webcam. The architecture is designed to capture complementary visual, structural, and contextual features from hand gestures while maintaining real-time performance.

The system begins with video acquisition, where live frames are captured from a webcam. Each frame undergoes preprocessing, including resizing, normalization, and background noise reduction. In parallel, hand landmark detection is performed using MediaPipe to extract 21 key hand joints, which represent finger and palm positions.The processed input is then forwarded to three parallel deep-learning streams.

Fig. ASL Dataset

The CNN stream extracts spatial and texture-based visual features, effectively learning hand shapes and appearance patterns. The GNN stream models the geometric relationships between hand landmarks by representing joints as graph nodes and their connections as edges, enabling structural

understanding of finger configurations. The Vision Transformer (ViT) stream applies self-attention mechanisms to focus on the most informative regions of the input image while suppressing background interference.

Each model independently generates class probabilities for ASL gestures. These outputs are combined using a confidence-weighted ensemble fusion strategy, producing a final prediction that is more robust and accurate than any single model. The recognized letters are then passed to a post-processing module, which assembles characters into words and supports control gestures such as space and delete.

Finally, the recognized output is displayed through a web-based user interface, enabling fast, accurate, and accessible real-time communication without requiring specialized hardware.

IV. METHODOLOGY

The system is trained on an ASL dataset containing approximately 87,000 labeled images covering 29 gesture classes. Data augmentation techniques such as rotation and brightness adjustment improve generalization. The CNN and ViT process image inputs, while the GNN operates on hand landmark graphs.

Training is performed using the Adam optimizer with categorical cross-entropy loss. The ensemble approach enhances robustness by compensating for individual model limitations.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed ensemble model achieves an overall recognition accuracy exceeding 95%, demonstrating a clear improvement over the performance of individual deep learning models. This enhanced accuracy is achieved by effectively combining the complementary strengths of the constituent architectures. The Convolutional Neural Network (CNN) excels at extracting fine-grained visual features such as hand shape, edges, and texture patterns, which are essential for distinguishing static ASL gestures. The Graph Neural Network (GNN) contributes by modeling the structural relationships between hand joints and fingers, allowing the system to better differentiate gestures with similar visual appearances but distinct spatial configurations. In addition, the Vision Transformer (ViT) improves recognition by employing self-attention mechanisms that focus on the most informative regions of each frame while minimizing the influence of background noise and illumination variations.
Beyond accuracy, the system is designed to support real-time deployment. Performance evaluation shows an average inference time of approximately 85 milliseconds per gesture, enabling smooth and responsive interaction suitable for live communication scenarios. This low latency ensures that users experience minimal delay between gesture execution and system output, making the solution practical for continuous fingerspelling and interactive use. The combination of high accuracy, robustness, and real-time performance highlights the effectiveness of the proposed ensemble approach for real-world American Sign Language recognition.

A. Individual Model Performance

1. Each deep learning model in the ensemble was first evaluated independently to analyze its strengths and limitations. The Convolutional Neural Network (CNN) achieved high accuracy in recognizing static ASL gestures by effectively capturing visual patterns such as hand contours, finger orientation, and texture details. However, CNN-based models exhibited difficulty in distinguishing visually similar gestures where subtle differences in finger positioning were critical.

2. The Graph Neural Network (GNN) demonstrated its effectiveness in modeling spatial dependencies between hand joints by representing landmarks as nodes in a graph structure. This enabled the system to better understand finger articulation and hand posture geometry. Despite this advantage, the GNN showed comparatively lower standalone accuracy due to its dependence on accurate landmark extraction and limited visual context.

3. The Vision Transformer (ViT) achieved strong performance by leveraging self-attention mechanisms to focus on discriminative regions of the input image. The ViT proved particularly robust to background clutter and lighting variations. However, its higher computational complexity made it less suitable as a standalone solution for lightweight real-time deployment.

B. Ensemble Model Performance

1. The proposed ensemble model integrates the outputs of the CNN, GNN, and Vision Transformer using a confidence-weighted fusion strategy. By combining visual, structural, and attention-based features, the ensemble achieves an overall accuracy exceeding 95%, outperforming all individual models. This improvement highlights the complementary nature of the selected architectures and validates the effectiveness of multi-modal learning for sign language recognition.
2. The ensemble approach significantly reduces misclassification of visually similar signs by compensating for the weaknesses of individual models. For instance, cases where the CNN fails due to structural ambiguity are corrected through the GNN's spatial reasoning, while the ViT enhances robustness under challenging environmental conditions.

C. Real-Time Performance Analysis

1. In addition to accuracy, real-time performance is a critical requirement for practical sign language interpretation systems. The proposed system was evaluated under live webcam input to measure inference latency and responsiveness. The average inference time was recorded at approximately 85 milliseconds per gesture, enabling smooth and continuous interaction without noticeable delay.
2. This low latency allows users to perform fingerspelling naturally, with the system responding almost instantaneously. The model runs efficiently on standard consumer hardware without requiring high-end GPUs or specialized sensors, making it suitable for deployment in real-world environments such as classrooms, hospitals, workplaces, and public service centers.

D. Robustness and Generalization

1. The system was tested under varying lighting conditions, backgrounds, and hand orientations to assess robustness. Data augmentation during training played a key role in improving generalization. The ensemble model consistently maintained high accuracy across different environments, demonstrating its ability to adapt to real-world variability.

VI. VISUALIZATION AND PERFORMANCE EVALUATION

To further analyze model behavior, performance metrics such as accuracy, precision, recall, F1-score, and confusion matrices were examined. Confusion matrix analysis revealed that most classification errors occurred between gestures with extremely subtle differences, such as 'M' and 'N'. However, these errors were significantly reduced in the ensemble model compared to individual architectures.

Loss and accuracy curves indicated stable convergence during training, with minimal overfitting due to early stopping and regularization techniques. The ensemble approach demonstrated balanced performance across all gesture classes, confirming its suitability for comprehensive ASL recognition.

VII. APPLICATIONS AND SOCIAL IMPACT

The proposed system has wide-ranging applications across multiple domains. In educational settings, it can assist students and instructors by enabling interactive ASL learning and assessment. In healthcare environments, it can facilitate communication between medical professionals and deaf patients, reducing misunderstandings and improving patient care. Public service institutions and workplaces can integrate the system to promote inclusivity and equal access to services.

Beyond technical contributions, this work addresses an important social challenge by reducing

dependence on costly human interpreters and enabling independent communication for deaf and hard-of-hearing individuals.

VIII. CONCLUSION

This paper presented a real-time American Sign Language recognition system based on a multi-modal ensemble of CNN, GNN, and Vision Transformer architectures. The proposed approach achieves high accuracy exceeding 95% while maintaining real-time performance with minimal latency. By leveraging complementary feature representations, the system overcomes limitations of single-model approaches and demonstrates strong robustness in real-world conditions.

REFERENCES

1. Kothadiya, D., Bhatt, C., Sapariya, K., Patel, K., Gil-González, A. B., & Corchado, J. M. (2022). Deepsign: Sign Language Detection and Recognition Using Deep Learning. Electronics, 11(11), 1780.
2. Biravi, K. N., & Krishnaveni, N. (2023). Sign Language Recognition Using Deep Learning. International Journal of Engineering Research & Technology, 12(03), 1456-1462.
3. Al-Qurishi, M., Khalid, T., & Souissi, R. (2021). Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues. IEEE Access, 9, 126917-126951.
4. Baihan, A., Alutaibi, A. I., Alshehri, M., & Sharma, S. K. (2024). Sign language recognition using modified deep learning network and hybrid optimization. Scientific Reports, 14, 8642.
5. Kodandaram, S. R., Kumar, N. P., & Sunil, G. L. (2021). Sign Language Recognition using Deep Learning Techniques. International Journal of Advanced Research in Computer Science, 12(4), 89-94.
6. Saiful, M. N., Isam, A. A., Moon, H. A., Jaman, R. T., Das, M., Alam, M. R., & Rahman, A. (2022). Real-Time Sign Language Detection Using CNN. International Conference on Computer Communication and Informatics, 1-6.
7. Imran, A., Shashishekhara, H. M., & Gardi, H. A. A. (2024). Real Time American Sign Language Detection Using YOLO-v9. arXiv preprint arXiv:2407.10294.
8. Vyavahare, P., Dhawale, S., Takale, P., Koli, V., Kanawade, B., & Khonde, S. (2023). Detection and Interpretation of Indian Sign Language Using LSTM Networks. International Journal of Research in Engineering, Science and Management, 6(7), 234-238.
9. Alsolai, H., Alsolai, L., Al-Wesabi, F. N., Othman, M., Rizwanullah, M., & Abdelmageed, A. A. (2023). Automated Sign Language Detection and Classification using Reptile Search Algorithm with Hybrid Deep Learning. Computer Systems Science and Engineering, 47(3), 3523-3538.
10. Kumar, S., Rani, R., & Chaudhari, U. (2024). Real-time sign language detection: Empowering the disabled community. Journal of Assistive Technologies, 18(3), 145-157.
11. Zuo, R., Wei, F., & Mak, B. (2023). Natural Language-Assisted Sign Language Recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision, 14327-14337.
12. Verma, A. R., Singh, G., Meghwal, K., Ramji, B., & Dadheech, P. K. (2024). Enhancing Sign Language Detection through Mediapipe and Convolutional Neural Networks. International Journal of Computer Applications, 186(24), 12-18.
13. Hossain, M. T. B., Ahad, M. A. R., Das, A., Sugiura, N., & Kise, K. (2019). Sign Language Recognition Analysis using Multimodal Data. Pattern Recognition Letters, 126, 116-125.
14. Ningsih, M. R., Nurriski, Y. J., Sanjani, F. A. Z., Hakim, M. F. A., Unjung, J., & Muslim, M. A. (2024). Sign Language Detection System Using YOLOv5 Algorithm. Indonesian Journal of Artificial Intelligence and Data Mining, 7(1), 45-52.
15. Pathan, R. K., Biswas, M., Yasmin, S., Khandaker, M. U., Salman, M., & Youssef, A. A. F. (2023). Sign language recognition using fusion of image and hand landmarks through multi-headed convolutional neural network. Scientific Reports, 13, 6699.