SignSei: A Real-Time Indian Sign Language Interpreter Using Deep Learning

Mrs. Ilakkia A
Assistant Professor, Dept. of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Puducherry, India

Yuvanthikhaa A
Dept. of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Puducherry, India

Hariynie C
Dept. of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Puducherry, India

Abstract - Sign language is the primary mode of communication for the hearing impaired, and bridging the communication gap between the hearing and hearing-impaired communities is crucial for inclusivity. This paper presents SignSei, a real-time Indian Sign Language (ISL) recognition system designed to enhance communication for deaf and hard-of-hearing individuals through accurate gesture interpretation via live video feeds. The proposed system utilizes Python as the core programming language, leveraging TensorFlow and Keras for developing and training deep learning models, while OpenCV is employed for efficient video processing tasks that enable real-time capture and analysis of gestures. SignSei combines convolutional and recurrent neural architectures with an attention-based module to achieve high accuracy in gesture recognition with low latency, making it suitable for real-time applications. By training on a diverse dataset of ISL signs, the system is designed to generalize effectively across different users, improving reliability and reducing variability in gesture recognition. Ultimately, SignSei aims to empower the hearing-impaired community by fostering independence and a sense of belonging.

Index Terms - Accessibility, computer vision, deep learning, Indian Sign Language, Keras, OpenCV, real-time communication, sign language recognition, TensorFlow.

I. INTRODUCTION

Communication is one of the most fundamental parts of human interaction, yet millions of people are excluded from it because they are deaf or hard of hearing. For the hearing-impaired, sign language is the primary means of communication. Indian Sign Language (ISL), mainly used across South Asia, enables users to communicate effectively through hand gestures, facial expressions, and body movements. The gap this creates between signing and non-signing communities affects everyday life — public services, education, healthcare, and social interaction. Human interpreters can bridge this gap, but they are often unavailable, costly, or impractical in remote or emergency situations. This has motivated more than a decade of active research into automated sign language recognition (SLR) systems [1], drawing on a wide range of hand-crafted and deep-learning-based feature extraction and classification techniques [2].

SignSei is proposed as a real-time interpreter of Indian Sign Language that bridges this communication gap using deep learning and computer vision. SignSei interprets gestures from a live video feed and converts them to text or speech, enabling two-way communication between hearing-impaired and hearing individuals. Python provides the core development environment, TensorFlow and Keras support model development and training, and OpenCV handles real-time video stream processing.

A core challenge in sign language recognition is the complexity of gestures themselves. Signs combine manual components — hand shape, orientation, and movement — with non-manual components such as facial expression and head or body movement, both of which must be captured accurately to preserve meaning [2]. The same hand movement, for instance, can carry a different meaning depending on facial expression. SignSei addresses this by feeding the live video stream to CNNs and RNNs for spatial and temporal feature recognition respectively. A second challenge is variability: the same sign is performed differently by different people due to variation in speed, intensity, and personal style, so the system is designed to be robust to this variation across users and environments. Such a framework could be applied in public service offices, hospitals, and schools, supporting hearing-

impaired individuals without requiring a human interpreter.

II. Related Works

A substantial body of work has looked specifically at recognizing continuous, sentence-level sign language rather than isolated signs. Aloysius et al. adapted the Conformer architecture, originally developed for speech recognition, to continuous sign language recognition (CSLR), proposing ConSignformer and a Cross-Modal Relative Attention mechanism to jointly model spatial and temporal cues [3]. The same authors later optimized this architecture with a Sign Query Attention module to improve efficiency and scalability [4], and in earlier work incorporated explicit relative position information into Transformer-based models for sign language recognition and translation [5]. These works motivate SignSei's use of attention mechanisms alongside CNN-RNN pipelines to capture long-range temporal dependencies in a gesture sequence.

Word-level and India-specific recognition systems have also progressed considerably. Sreemathy et al. built an 80-class ISL word database and combined YOLOv4 with an SVM-and-MediaPipe pipeline in an expert system that achieved high real-time accuracy [6]. Kothadiya et al. proposed DeepSign, which applies stacked LSTM and GRU layers to isolated ISL video frames using their own IISL2020 dataset, reporting close to 97% accuracy across 11 signs [8]. Sharma and Singh built a large, uncontrolled-environment ISL dataset from 65 signers and used augmentation together with a custom CNN for feature extraction and classification [13]. Sharma et al. benchmarked several deep neural network architectures specifically for ISL recognition [12], while Sharma and Anand compared different deep models and optimizer choices for the same task [14]. Zhang and Jiang's recent survey situates all of this work within the broader progression of CNN-, RNN-, and Transformer-based approaches to sign language recognition over the past five years [7].

Dataset availability has historically limited ISL research. Sridhar et al. addressed this with INCLUDE, a large-scale ISL dataset spanning 0.27 million frames across 4,287 videos and 263 word signs, and benchmarked multiple deep network configurations on it [11]. On the modeling side, Camgoz et al. established end-to-end neural sign language translation as a sequence-to-sequence problem, directly mapping continuous sign video to spoken-language text [9]. Singh applied 3D-CNNs to dynamic ISL gestures to jointly capture motion and depth information [15], while Kumar and Kumar used an Extreme Learning Machine for ISL finger-spelling (alphabet) recognition [16]. Gangrade and Bharti proposed a CNN-based vision system for ISL hand gesture recognition [17], and Likhar et al. compared several deep learning methods for the ISL recognition task [18].

Efficiency for real-time and mobile deployment is a recurring theme that directly informs SignSei's design. Temkar et al. built a real-time ISL recognition system around MobileNetV2 and transfer learning to keep inference lightweight enough for everyday use [10]. Shashidhar et al. similarly used a CNN-based pipeline to convert ISL gestures directly to speech output [19]. SignSei draws on these efficiency-oriented designs by adopting a MobileNet-based feature extractor, while combining it with the attention and sequence-modeling techniques discussed above to balance accuracy against the low-latency requirements of a live video interpreter.

III. System Architecture

SignSei is designed to interpret ISL gestures in real time and output text or speech, providing deaf and hard-of-hearing users with an additional channel of communication. The architecture is built around a live video feed from a high-definition webcam and is designed to support multi-modal input options such as voice commands or touch gestures. A preprocessing stage built around data augmentation is central to the design, aimed at producing a diverse dataset that represents the range of ISL gestures the system needs to recognize. Figure 1 illustrates the overall recognition pipeline described in this section.

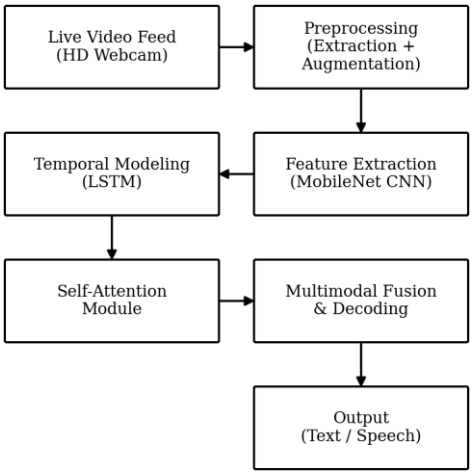

Figure 1. High-level SignSei recognition pipeline, from live video capture through CNN-based feature extraction, LSTM temporal modeling, attention, and fused decoding to text/speech output.

During preprocessing, OpenCV is used alongside GPU acceleration to extract video frames efficiently. Augmentation strategies operate on two levels: spatial augmentation (flipping, cropping, color jittering, noise addition) increases visual variety in the dataset, while

temporal augmentation (frame skipping, interpolation) is intended to simulate fluid transitions between gestures.

Once the data is prepared, spatial features are extracted using a convolutional backbone building on the design principles established by early CNN architectures for visual recognition [24] and refined by deeper networks such as VGG [25]. For real-time performance on modest hardware, SignSei adopts a lightweight MobileNet-style feature extractor that relies on depthwise separable convolutions [20], following the efficiency-oriented approach validated for ISL specifically by Temkar et al. [10]. Temporal dependencies across a gesture sequence are then captured using Long Short-Term Memory (LSTM) networks [23], which are well suited to modeling the variable-length, variable-speed nature of signing.

To help the model focus on the most informative frames in a sequence, SignSei incorporates a self-attention module inspired by the Transformer architecture [21] and its Conformer variant, which combines convolution with self-attention for sequence modeling [22]; this mirrors the attention-based CSLR designs of Aloysius et al. [3], [4]. This is complemented by ideas from fully convolutional and dense temporal convolution approaches to continuous sign language recognition [26], [27], which inform how SignSei aggregates features across a gesture sequence before decoding. Multimodal cues — combining 2D appearance features with 3D/depth-style motion information, in the spirit of Kothadiya et al.'s multi-stream design [8] and Camgoz et al.'s sequence-to-sequence translation framework [9] — are fused prior to decoding.

The recognition engine combines CNN-extracted spatial features with LSTM-derived temporal patterns and the attention module's frame weighting to produce real-time text or speech output. Combining OpenCV with CUDA is intended to keep frame analysis fast enough to minimize latency. Model compression techniques (quantization, pruning) are applied so the model can run on lower-resource devices, broadening the system's potential reach. The pipeline proceeds through: Dataset Creation, Data Augmentation, Video Segmentation, Feature Extraction, Model Training, Sign Recognition, Model Evaluation, Language Modeling, and finally Optimization and Integration, where the trained model is packaged into a user-facing application with text-to-speech output. This end-to-end development pipeline is illustrated in Figure 2.

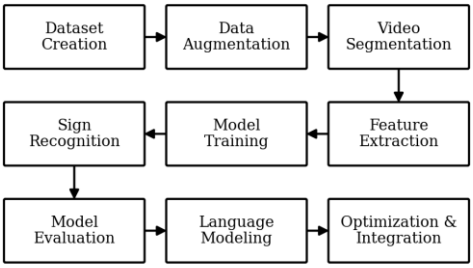

FIGURE 2. END-TO-END DEVELOPMENT PIPELINE OF SIGNSEI, FROM DATASET CREATION THROUGH MODEL TRAINING, EVALUATION, AND FINAL OPTIMIZATION AND INTEGRATION.

IV. RESULTS AND EVALUATION

Note: this section currently describes the intended evaluation methodology rather than completed results. Before submission, the authors should replace the discussion below with actual measured outcomes — accuracy, precision, recall, and F1-score on a held-out test set, along with a confusion matrix and error analysis — from a trained and tested version of the system.

The evaluation of SignSei is intended to combine quantitative and qualitative measures. Quantitatively, accuracy, precision, recall, and F1-score would be calculated on a held-out test dataset, following the benchmarking methodology used for the INCLUDE ISL dataset [11] and for other ISL deep-learning benchmarks [12], to assess how well the model recognizes ISL gestures. Qualitatively, structured user surveys and interviews would assess usability, translation quality, and overall satisfaction, ideally across users from different linguistic and geographic backgrounds. Real-world testing in school and public-service settings would help validate the system's practical performance outside controlled conditions.

As a point of reference, comparable ISL systems combining classical and deep models have reported real-time accuracies above 98% on constrained vocabularies [6], while lightweight MobileNet-based ISL recognizers have demonstrated that accuracy can be maintained even under the latency constraints of live, on-device inference [10], [13]. If SignSei performs as intended, the expected outcomes are: real-time gesture-to-text/speech translation with low latency; robustness to variation in signing speed, style, and environment across different users; and translation output that incorporates contextual understanding and punctuation rather than a literal sign-by-sign mapping. The system is also intended to be extensible to other sign languages beyond ISL in future work.

V. Conclusion

SignSei is proposed as an AI- and computer-vision-based platform for accurate, real-time, and accessible Indian Sign Language interpretation. The goal is a system that can interpret a wide range of ISL gestures, allowing deaf individuals to communicate more independently across a variety of settings, with an interface designed to integrate across devices and platforms. The system's design builds on a decade of progress in sign language recognition research [1] and on recent advances in deep-learning-based recognition techniques specifically [7].

Beyond personal communication, such a system could support learning environments, doctor-patient interaction, and public services for the deaf community — pending validation through real experimental results and user testing. Future work could extend SignSei toward multi-scene, large-vocabulary continuous recognition using newer benchmark datasets such as Isharah [28], and toward additional Indian and international sign languages beyond ISL.

References

[1] A. Wadhawan and P. Kumar, "Sign language recognition systems: A decade systematic literature review," *Archives of Computational Methods in Engineering*, vol. 28, no. 3, pp. 785–813, 2021.

[2] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 1, pp. 131–153, 2019.

[3] N. Aloysius, M. Geetha, and P. Nedungadi, "Continuous sign language recognition with adapted conformer via unsupervised pretraining," *arXiv preprint arXiv:2405.12018*, 2024.

[4] N. Aloysius, M. Geetha, and P. Nedungadi, "Optimized multi-modal conformer-based framework for continuous sign language recognition," *IEEE Open Journal of the Computer Society*, vol. 6, pp. 739–749, 2025.

[5] N. Aloysius, M. Geetha, and P. Nedungadi, "Incorporating relative position information in transformer-based sign language recognition and translation," *IEEE Access*, vol. 9, pp. 145929–145942, 2021.

[6] R. Sreemathy, M. P. Turuk, S. Chaudhary, K. Lavate, A. Ushire, and S. Khurana, "Continuous word level sign language recognition using an expert system based on machine learning," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 170–178, 2023.

[7] Y. Zhang and X. Jiang, "Recent advances on deep learning for sign language recognition," *Computer Modeling in Engineering & Sciences*, vol. 139, no. 3, pp. 2399–2450, 2024.

[8] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A.-B. Gil-González, and J. M. Corchado, "Deepsign: Sign language detection and recognition using deep learning," *Electronics*, vol. 11, no. 11, p. 1780, 2022.

[9] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7784–7793.

[10] R. Temkar, S. Jagtap, K. Jadhav, and M. Deshmukh, "Real-time sign language recognition using MobileNetV2 and transfer learning," *arXiv preprint arXiv:2412.07486*, 2024.

[11] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, "INCLUDE: A large scale dataset for Indian sign language recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1366–1375.

[12] A. Sharma, N. Sharma, Y. Saxena, A. Singh, and D. Sadhya, "Benchmarking deep neural network approaches for Indian Sign Language recognition," *Neural Computing and Applications*, vol. 33, pp. 6685–6696, 2021.

[13] S. Sharma and S. Singh, "Recognition of Indian Sign Language (ISL) using deep learning model," *Wireless Personal Communications*, 2022.

[14] P. Sharma and R. S. Anand, "A comprehensive evaluation of deep models and optimizers for Indian sign language recognition," *Graphics and Visual Computing*, vol. 5, p. 200032, 2021.

[15] D. K. Singh, "3D-CNN based dynamic gesture recognition for Indian Sign Language modeling," *Procedia Computer Science*, vol. 189, pp. 76–83, 2021.

[16] A. Kumar and R. Kumar, "A novel approach for ISL alphabet recognition using Extreme Learning Machine," *International Journal of Information Technology*, vol. 13, no. 1, pp. 349–357, 2021.

[17] J. Gangrade and J. Bharti, "Vision-based hand gesture recognition for Indian sign language using convolution neural network," *IETE Journal of Research*, 2020.

[18] P. Likhar, N. K. Bhagat, and G. N. Rathna, "Deep learning methods for Indian Sign Language recognition," in *Proc. IEEE 10th Int. Conf. Consumer Electronics (ICCE-Berlin)*, 2020.

[19] R. Shashidhar, S. R. Hegde, K. Chinmaya, A. Priyesh, A. S. Manjunath, and B. N. Arunakumari, "Indian Sign Language to speech conversion using convolutional neural network," in *Proc. IEEE MysuruCon*, 2022.

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6000–6010.

[22] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, and Y. Wu, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech 2020*, 2020.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] D. Guo, S. Wang, Q. Tian, and M. Wang, "Dense temporal convolution network for sign language translation," in *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2019, pp. 744–750.

[27] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai, "Fully convolutional networks for continuous sign language recognition," in *Proc. European Conf. Computer Vision (ECCV)*, 2020, pp. 697–714.

[28] S. Alyami, H. Luqman, S. Al-Azani, M. Alowaifeer, Y. Alharbi, and Y. Alonaizan, "Isharah: A large-scale multi-scene dataset for continuous sign language recognition," *arXiv preprint*, 2025.

About the Authors

Mrs. Ilakkia A is an Assistant Professor in the Department of Artificial Intelligence and Data Science at Sri Manakula Vinayagar Engineering College, Puducherry, India, and served as project advisor for this work.

Yuvanthikhaa A is an undergraduate student in the Department of Artificial Intelligence and Data Science at Sri Manakula Vinayagar Engineering College, Puducherry, India.

Hariynie C is an undergraduate student in the Department of Artificial Intelligence and Data Science at Sri Manakula Vinayagar Engineering College, Puducherry, India.