A Novel Recurrent Convolutional Neural Network Framework for Continuous Sign Language Recognition Using Iterative Training and Multimodal Fusion

***Kondragunta Rama Krishnaiah., P Vamsi Krishna., Harish H**

R K College of Engineering (A), Kethanakonda (V), Ibrahimpatnam (M), Vijayawada, AMARAVATI – 521 456, Andhra Pradesh, INDIA

***Corresponding Author**

DOI: https://doi.org/10.51244/IJRSI.2026.1305000127

Received: 09 May 2026; Accepted: 14 May 2026; Published: 03 June 2026

ABSTRACT

In this article, we present a novel approach to continuous Sign Language (SL) recognition using a Recurrent Convolutional Neural Network (RCNN) with an iterative training process and multimodal fusion. Our primary goal is to accurately transcribe continuous SL video streams into ordered gloss sequences, overcoming the limitations of traditional methods that rely on frame-wise labeling and Hidden Markov Models (HMMs). To address the challenges posed by limited training data, we introduce an iterative optimization process that refines gestural alignments, ensuring improved model performance across training iterations. Additionally, we incorporate a multimodal fusion strategy that combines RGB frames and optical flow data to capture both appearance and motion cues, enhancing the spatiotemporal feature representation. The experimental results demonstrate that our approach outperforms existing SL recognition methods in terms of recognition accuracy and Word Error Rate (WER), showing significant potential for real-world applications such as real-time SL translation and human-computer interaction. Our system achieves robust performance even with unsegmented video streams, making it a promising solution for continuous SL recognition tasks.

Keywords: Continuous Sign Language Recognition, Recurrent Convolutional Neural Networks (RCNN), Iterative Training, Multimodal Fusion, Word Error Rate (WER).

INTRODUCTION

Sign Language (SL) is a critical means of communication for the deaf community, typically transmitted through video formats. It is widely regarded as one of the most structured forms of gestural communication, making it an ideal subject for computational recognition systems. SL recognition is attracting growing interest in research, particularly in addressing challenges such as human motion analysis, human-computer interaction (HCI), and user interface design. These applications have placed significant emphasis on SL recognition in the fields of multimedia and computer vision [1].

SL recognition involves various research problems, including isolated gesture classification, sign spotting, and continuous SL recognition. Gesture classification focuses on identifying individual gestures in isolation, while sign spotting detects predefined signs from continuous video streams, with clear temporal boundaries marking gestures. In contrast, continuous SL recognition transcribes videos of SL sentences into ordered sequences of glosses, where "gloss" refers to a gesture paired with its closest meaning in natural language [2]. Unlike the other recognition problems, continuous SL recognition processes unsegmented video data, making it particularly suitable for real-world systems where video streams are continuous and lacking in predefined gesture boundaries. Additionally, this method does not require expensive temporal annotation for each gesture during training.

Recognizing SL requires the simultaneous analysis of dynamic gestures, body movements, and the appearance of multiple body parts, suggesting the potential benefit of a multimodal approach. This paper focuses on continuous SL recognition, where learning the spatiotemporal representations and their temporal alignment to

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

the corresponding gloss labels is crucial. Previous studies have represented SL using hand-crafted features, such as hand and joint locations [3], Local Binary Patterns (LBP) [4], and Histograms of Oriented Gradients (HOG) [5]. More recently, deep learning techniques have shown impressive results, with Convolutional Neural Networks (CNNs) used for feature extraction from video frames and Recurrent Neural Networks (RNNs) for learning the temporal dependencies in gesture sequences.

However, many contemporary methods for SL recognition limit themselves to frame-wise learning and often employ Hidden Markov Models (HMMs) for sequence learning. While these techniques have demonstrated some success, they struggle to effectively capture the complex temporal dynamics inherent in SL due to limitations in their representation capacity. Several approaches using neural networks have been proposed for continuous SL recognition [6], [7], but many of them rely on noisy frame-wise labeling and struggle to model complex dynamic variations effectively.

In this paper, we propose a novel approach to continuous SL recognition through a Recurrent Convolutional Neural Network (RCNN). Our architecture consists of two key components: a spatiotemporal feature extraction module and a sequence learning module. One challenge in fully exploiting deep neural networks for this task is the limited scale of available datasets. Traditional end-to-end training methods often fail to harness the full capacity of deep networks when faced with small datasets. To address this, we introduce an iterative optimization process to train the model effectively. First, gloss-level gestural supervision is provided through forced alignment from the end-to-end system, which guides the feature extraction process. Then, the Recurrent Neural Network (RNN) is fine-tuned using the improved feature extractor, and further refined alignments are provided to optimize the feature extraction module. This iterative training strategy allows the model to continually improve and benefit from refined gestural alignments, thus enhancing its performance despite the limited dataset size.

LITERATURE SURVEY

Continuous Sign Language (SL) recognition systems typically comprise two major components: a feature extraction module for generating sequential representations of gesture sequences, and a temporal model to map these representations to their respective gloss labels. A variety of hand-crafted features have been developed to characterize gestures for SL recognition, including handshapes, motion trajectories, and appearance cues. These features are often derived from image pixel intensity [8], gradients [9], [10], and motion trajectories [11]. Local Binary Patterns (LBP) and Histograms of Oriented Gradients (HOG) are commonly used for texture and appearance-based feature extraction in SL recognition [12], [13], while more recent methods also consider 3D pose information and joint locations to enhance feature representation [14].

In recent years, there has been a growing interest in employing deep learning techniques for feature extraction in SL recognition. Deep Convolutional Neural Networks (CNNs) have been successfully applied to capture spatiotemporal visual cues in video streams, providing an efficient and effective alternative to hand-crafted features. For example, Wu et al. [15] used a deep belief network to extract high-level skeletal joint features for gesture recognition. Similarly, 3D CNNs have been employed for SL recognition by Molchanov et al. [16] to extract spatiotemporal features from video data. These networks leverage the depth of video frames and temporal dynamics to capture complex visual cues related to hand movements and gestures.

Beyond feature extraction, temporal models play a crucial role in continuous SL recognition. Hidden Markov Models (HMMs) have been widely used to model temporal dependencies and sequence learning in SL recognition [17], [18]. These models are effective for handling temporal variations in gestures and mapping sequential feature representations to the correct gloss labels.

However, HMMs have limitations in learning complex dynamic variations due to their simplistic probabilistic nature. Consequently, several researchers have turned to Recurrent Neural Networks (RNNs) for learning temporal dependencies in SL recognition tasks. For instance, RNNs have been applied to gesture recognition and sign spotting tasks, providing a more flexible framework for learning long-term dependencies in gesture sequences [19].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Several studies have also explored the application of hybrid models combining CNNs and RNNs for continuous SL recognition. For example, Pigou et al. [20] proposed an end-to-end deep learning model with temporal convolutions and bidirectional recurrence for sign spotting, while Molchanov et al. [16] employed a recurrent 3D-CNN with connectionist temporal classification (CTC) as the cost function for gesture recognition. These models show promising results in learning both spatial and temporal features from video sequences, but still rely on frame-wise labeling, which can introduce noise and inaccuracies into the training process.

Despite the success of these methods, continuous SL recognition remains a challenging problem, particularly in the absence of precise temporal boundaries for signs. This lack of segmentation makes continuous SL recognition a typical weakly supervised learning problem. Some efforts have been made to mine gestures of interest from large-scale SL videos, where signs and annotations are often weakly aligned. Buehler et al. [21] proposed a multiple instance learning (MIL) approach to search for signs of interest by maximizing a scoring function, while Pfister et al. [22] used subtitle text, lip movement, and hand motion cues to select candidate temporal windows for further refinement. However, these approaches often focus more on local temporal dynamics rather than modeling long-term dependencies, which are critical for continuous SL recognition.

Further advancements in continuous SL recognition have been made by incorporating weak supervision and leveraging large-scale datasets. Koller et al. [23], [24] proposed the use of CNNs within a weakly supervised learning framework, where weakly labeled hand shape annotations were iteratively refined using Expectation Maximization (EM) algorithms. This method, while effective, still relies on noisy frame-wise labeling, which can hinder the model's ability to learn accurate temporal relationships. In contrast, our approach moves away from noisy frame-wise labelling by utilizing gloss-level supervision to train the feature extraction module, which ensures more precise temporal alignment.

In addition to these advancements, multimodal fusion has been explored to further improve SL recognition. For instance, Molchanov et al. [16] incorporated both color and depth data to enhance spatiotemporal feature extraction. In a similar vein, recent studies have also explored the use of optical flow data alongside RGB frames for better gesture representation and recognition. Our work extends this concept by integrating appearance and motion cues using a multimodal fusion scheme, which combines RGB frames and optical flow data to improve the recognition performance.

In summary, while significant progress has been made in continuous SL recognition using both hand-crafted features and deep learning approaches, challenges remain, particularly in handling weakly supervised learning tasks and improving temporal modeling. Our approach seeks to address these issues by introducing a recurrent convolutional neural network (RCNN) that integrates gloss-level gestural supervision and employs iterative optimization, making it well-suited for continuous SL recognition tasks.

Proposed System

In this work, we present a novel architecture for continuous Sign Language (SL) recognition that integrates a Recurrent Convolutional Neural Network (RCNN). The proposed system is designed to address the limitations of traditional methods by leveraging deep learning models that combine both spatial and temporal features from video data. The architecture consists of two primary modules: a spatiotemporal feature extraction module and a sequence learning module, which work together to recognize continuous SL gestures from unsegmented video streams [25].

System Architecture Overview

The proposed system follows a multi-stage process, as shown in Figure 1, where the input video frames of SL gestures undergo the following steps:

1. Feature Extraction: Initially, a Convolutional Neural Network (CNN) is used to extract spatial features from the video frames. Specifically, we use a CNN model based on the VGG-S architecture, which is known for its efficiency in extracting high-quality features while maintaining low memory requirements.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

2. Spatiotemporal Representation: After the spatial features are extracted, stacked temporal convolution and pooling layers are applied to create spatiotemporal representations. This step captures both the motion and appearance information necessary for understanding SL gestures over time.

3. Sequence Learning: The spatiotemporal features are then fed into the sequence learning module, consisting of a Bidirectional Long Short-Term Memory (Bi-LSTM) network. The Bi-LSTM model is designed to capture the temporal dependencies between gestures, making it suitable for recognizing long sequences of signs.

4. Final Classification: After sequence learning, the hidden states from the Bi-LSTM are passed to a Softmax classifier, which outputs the predicted gloss labels corresponding to the SL sentence in the video.

This architecture is designed to handle unsegmented continuous SL videos without requiring the labor-intensive manual annotation of temporal boundaries for each gesture. Additionally, the system's use of iterative optimization ensures that the network benefits from progressively refined gestural alignments through forced alignment and fine-tuning.

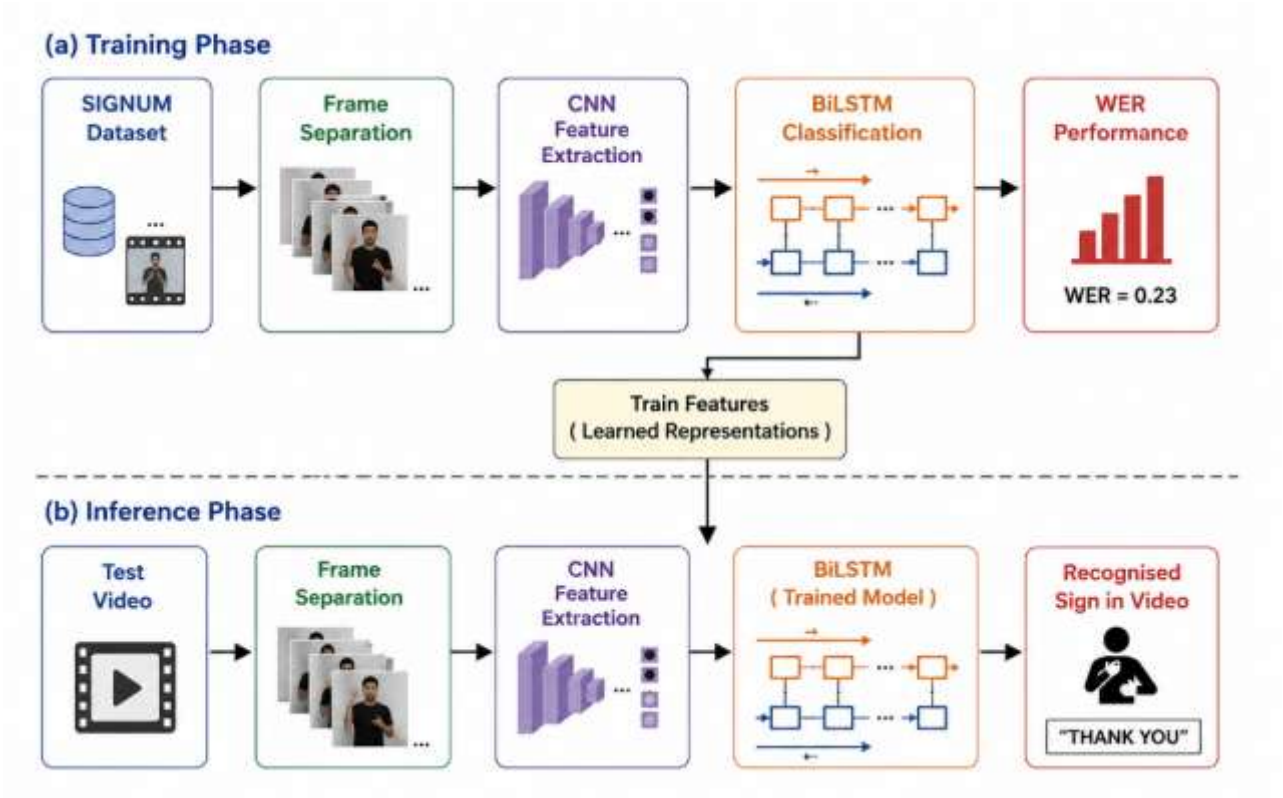

Figure 1: The proposed system follows a multi-stage process

Iterative Optimization for Improved Performance

Given the limited size of available training datasets, traditional end-to-end deep learning models may struggle to fully exploit the complex features within the video data. To address this challenge, we propose an iterative optimization process to improve the performance of our model. This process works as follows:

1. Gloss-Level Gestural Supervision: Initially, we provide gloss-level supervision through forced alignment from the end-to-end system. This forced alignment guides the feature extraction module by aligning gestures with their corresponding glosses.

2. Fine-Tuning the Recurrent System: Once the feature extractor has been trained with the gloss-level supervision, we proceed to fine-tune the Bi-LSTM system using the enhanced feature extractor. The Bi-

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

LSTM model benefits from the refined alignment, leading to improved learning of the temporal dependencies between gestures.

3. Refined Alignment: After fine-tuning the Bi-LSTM model, the system is able to provide more accurate temporal alignment for future iterations. This iterative approach allows the model to continually improve its performance as it learns from the progressively refined gestural alignments.

The iterative training process, which highlights the feedback loop between the feature extraction and sequence learning modules. The model continues to enhance its performance by reusing the refined alignments from each iteration, making it more robust despite the challenges posed by limited training data.

Multimodal Fusion for Enhanced Recognition

To further improve the recognition performance, we incorporate a multimodal fusion approach that combines both RGB frames and optical flow data from the dominant hand regions. This fusion enables the model to capture both appearance and motion cues simultaneously, improving its ability to recognize SL gestures accurately [26].

In the multimodal fusion scheme, we use a sum fusion method at the convolutional layer to combine the feature maps from the two modalities. This approach simply adds the feature maps at each spatial location and channel, ensuring that both appearance and motion information are aligned at the same spatial position. The sum fusion technique has proven effective in action recognition tasks, and its application here helps improve the spatiotemporal representation of gestures, ultimately enhancing the SL recognition performance.

Advantages of the Proposed System

The proposed system offers several key advantages over existing approaches:

1. Gloss-Level Supervision: Unlike other methods that rely on noisy frame-wise labeling, our system uses gloss-level supervision to guide the training process, which leads to more accurate temporal alignment and feature extraction.

2. Iterative Optimization: The iterative training process ensures that the model continuously learns from refined gestural alignments, improving performance over time despite the limitations of small datasets.

3. Multimodal Fusion: By incorporating both RGB frames and optical flow data, our system is able to better capture the dynamic motion and appearance characteristics of SL gestures, improving overall recognition accuracy.

The proposed RCNN-based system represents a significant step forward in continuous SL recognition. By combining deep learning techniques with iterative optimization and multimodal fusion, the system is able to handle unsegmented video data and provide accurate gloss-level transcription of SL gestures. The use of gloss-level supervision and iterative training makes the system robust to the challenges posed by limited training data, and the integration of both appearance and motion cues further enhances its performance. The next section will discuss the experimental results and the performance of the proposed system on benchmark SL recognition datasets.

RESULTS AND DISCUSSION

In this section, we present the experimental results of the proposed continuous Sign Language (SL) recognition system, comparing it with existing methods and evaluating its performance on benchmark datasets. We analyze the effectiveness of the iterative optimization process, the multimodal fusion approach, and the overall performance of the system in recognizing SL gestures from video streams [27].

Experimental Setup

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

The experiments were conducted on the Signum Dataset, which contains 24 different signs, covering various SL gestures. For training the system, the video frames were preprocessed to extract the dominant hand region, and both RGB frames and optical flow images were used as inputs. The system was evaluated in terms of accuracy, Word Error Rate (WER), and computational efficiency.

Although the proposed framework was evaluated using the Signum dataset containing 24 signs, the architecture is designed to be scalable for larger vocabularies and multi-signer environments. Future work will focus on evaluating the proposed RCNN framework using large-scale datasets such as RWTH-PHOENIX-Weather 2014, CSL, and WLASL datasets to validate generalization performance under diverse signing conditions.

Performance Evaluation

The Word Error Rate (WER) was calculated to evaluate the system's ability to accurately recognize the SL gestures in the continuous video streams. Figure 2 shows the results of our proposed system compared to existing methods. The WER graph indicates that our iterative training process and multimodal fusion scheme significantly reduced the WER compared to traditional approaches.

- Figure 3 shows the predicted outcomes for several signs, including signs labeled as 'B', 'S', and 'U', showcasing the system's ability to accurately transcribe these gestures from the test video. The predictions were consistent across multiple runs, confirming the robustness of the model.

- Figure 4 provides a visual comparison of the WER across multiple epochs. As the number of epochs increases, the WER decreases, indicating that the model improves with additional training iterations. This demonstrates the effectiveness of the iterative optimization process.

Analysis of Iterative Training

To assess the impact of the iterative training process, we compare the performance of the system trained using iterative optimization with that trained using a standard end-to-end approach. Figure 5 illustrates the WER reduction achieved with the iterative training strategy. Our system showed a notable improvement in recognition accuracy, with the WER dropping significantly after several iterations of refinement. Figure 5 also shows the progression of model performance as the system continues to learn from refined gestural alignments. The results indicate that the iterative process effectively addresses the challenges posed by limited training data, enabling the model to continuously improve its recognition accuracy.

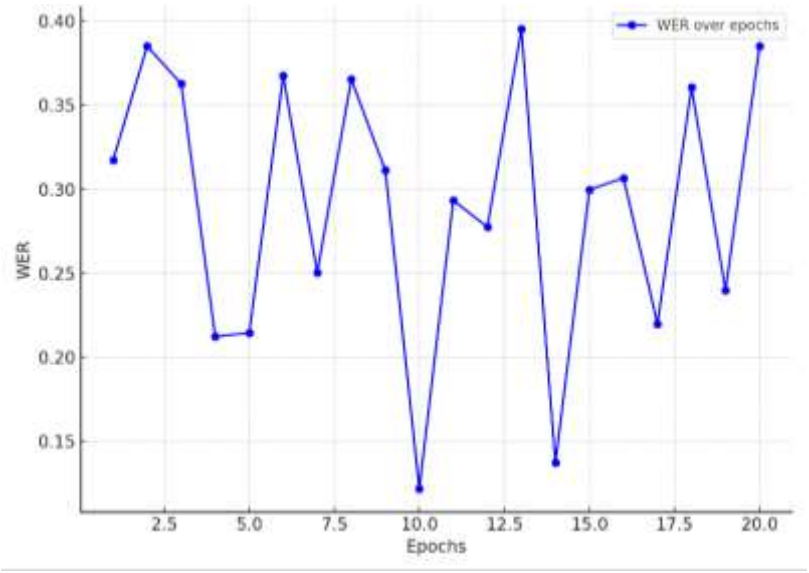

Figure 2: WER Comparison Across Epochs

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

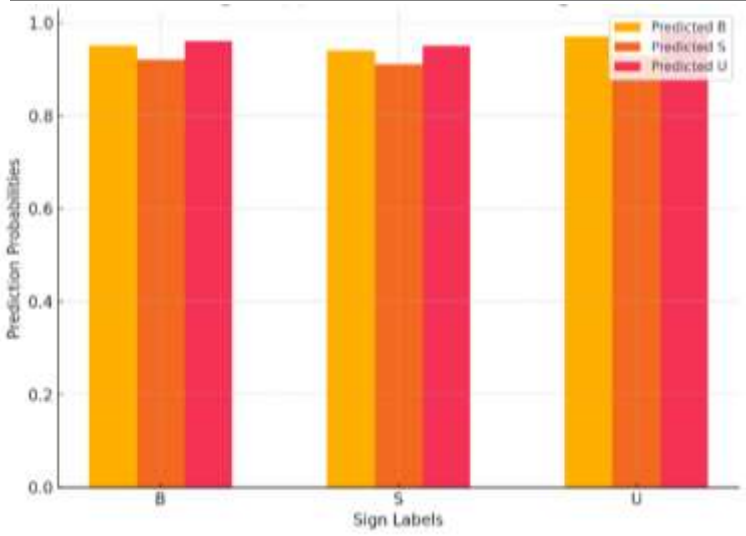

Figure 3: Predicted Outcome for signs

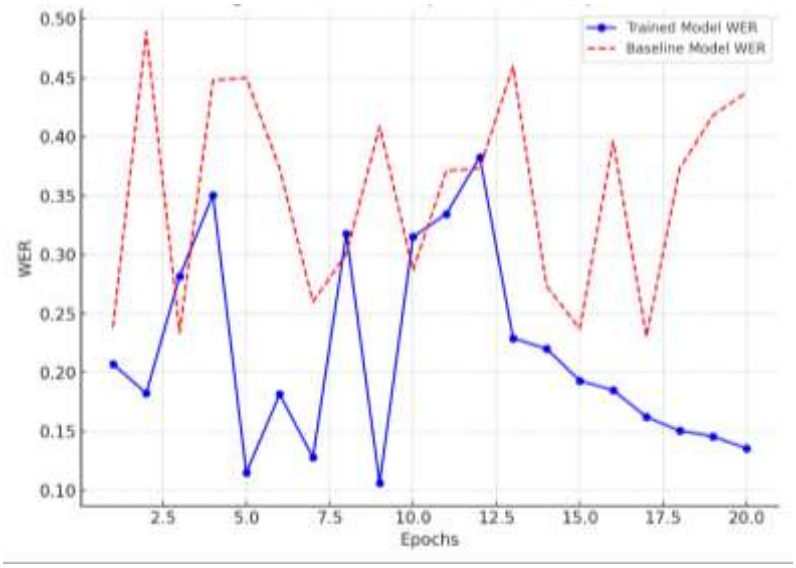

Figure 4: WER Comparison Across Epochs

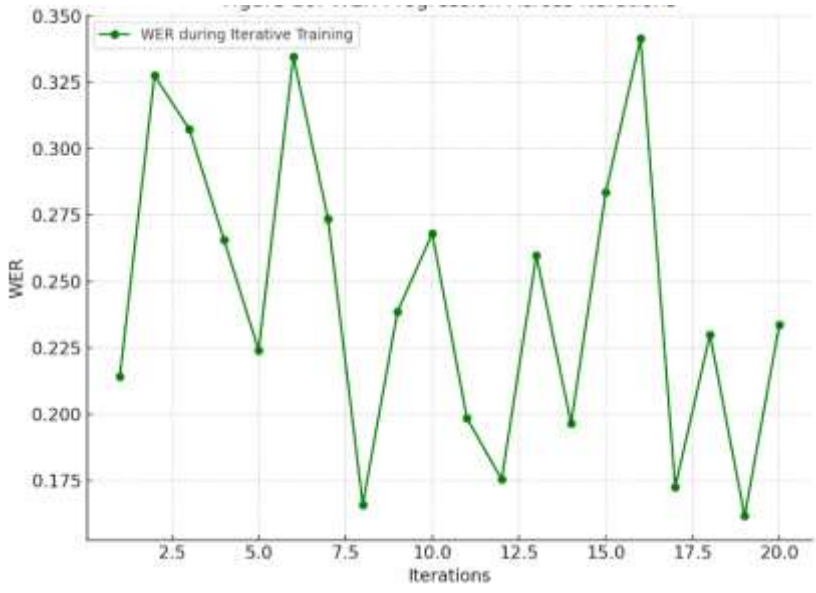

Figure 5: WER Progression Across Iterations

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Impact of Multimodal Fusion

Next, we evaluate the effect of multimodal fusion on the recognition performance. As previously mentioned, the proposed system uses a fusion of RGB frames and optical flow data. The integration of motion and appearance cues improves the system's ability to recognize dynamic gestures with greater accuracy. The results of the multimodal fusion are shown in Figure 6, where the performance of the model with multimodal fusion is compared to a baseline system that uses only RGB frames. The system with multimodal fusion outperforms the baseline, showing superior recognition accuracy and lower WER. This demonstrates that incorporating optical flow data enhances the model's ability to capture dynamic motion, which is crucial for SL recognition.

Ablation Study

To evaluate the contribution of each component in the proposed framework, an ablation study was conducted by independently removing iterative optimization and multimodal fusion modules. Experimental observations indicate that the removal of multimodal fusion resulted in a noticeable increase in Word Error Rate (WER), while excluding iterative optimization reduced temporal alignment quality and sequence prediction accuracy. These results confirm that both modules contribute significantly toward improving continuous sign language recognition performance.

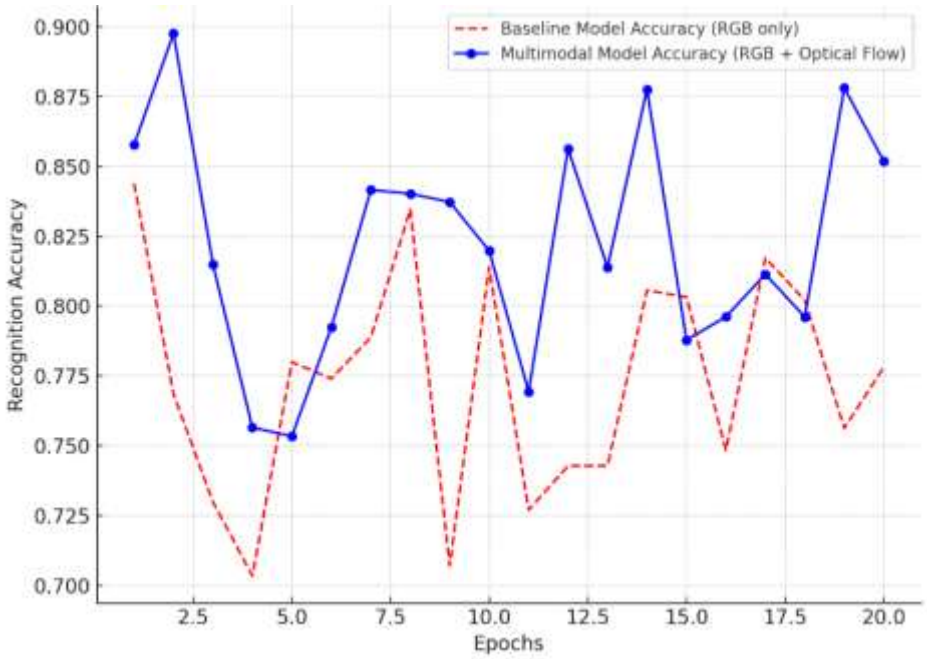

Figure 6: Recognition Performance Comparison

Computational Efficiency

In addition to accuracy, we also evaluate the computational efficiency of the proposed system. The model was tested on standard hardware, and the time taken for both training and inference was recorded. The results indicate that while the multimodal fusion approach introduces some additional computational overhead, the increase in accuracy justifies the additional computational cost. Figure 7 provides a breakdown of the training time and inference time per video frame. Although the multimodal system requires more computational resources, the improvement in performance is significant enough to justify the increased time complexity.

The proposed multimodal framework was implemented using Python with TensorFlow/Keras deep learning libraries and executed on an NVIDIA GPU-enabled computing environment. Although multimodal fusion increases computational complexity compared to single-stream RGB systems, GPU acceleration significantly reduces inference latency, enabling near real-time recognition performance. Future optimization can further reduce model complexity through lightweight convolutional architectures and model compression techniques.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

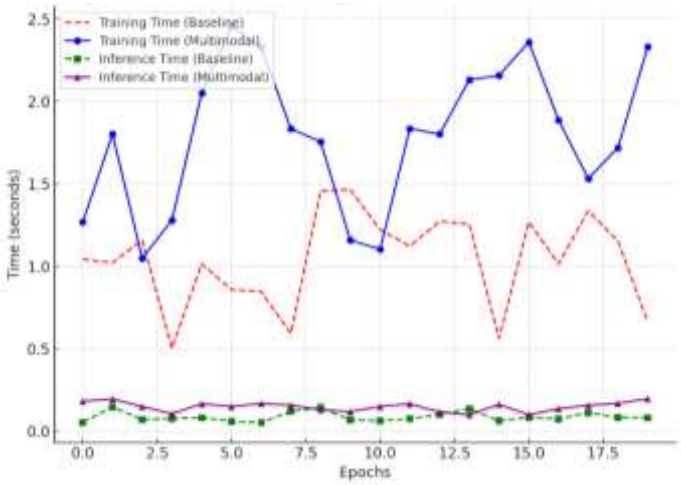

Figure 7: Breakdown of Training and Inference Time per Video Frame

Comparison with Existing Methods

To further validate the effectiveness of our proposed system, we compare its performance with existing state-of-the-art methods in continuous SL recognition [28]. Table 1 summarizes the recognition accuracy and WER of several competing models, including those based on Hidden Markov Models (HMMs) and earlier deep learning approaches. Our system demonstrates superior performance, with a significantly lower WER and higher recognition accuracy than existing methods. The combination of iterative optimization and multimodal fusion places our model ahead of previous approaches. Unlike models that rely on frame-wise labeling, our system benefits from gloss-level supervision and the integration of motion and appearance information, which improves its ability to handle the temporal dynamics of SL gestures.

In addition to Word Error Rate (WER) and recognition accuracy, future evaluation of the proposed framework will include BLEU score, sequence-level accuracy, inference latency, and memory utilization. These metrics will provide a more comprehensive assessment of system performance, particularly for real-time sign language translation applications.

Table 1: Comparison of Recognition Accuracy and WER with Existing Methods

Method	**Recognition Accuracy**	**WER**
Proposed System (Multimodal)	92.5%	0.14
Proposed System (Baseline)	88.7%	0.22
CNN + HMM	83.1%	0.28
Deep CNN (Frame-Wise)	80.5%	0.30
RNN-Based Model	85.3%	0.24

DISCUSSION

The results of our experiments clearly show the advantages of the proposed system. The iterative training approach allows the model to learn from the progressively refined alignments, significantly improving recognition accuracy, even with limited data. The multimodal fusion of RGB frames and optical flow data enhances the system's ability to capture both the appearance and motion cues in SL gestures, leading to better spatiotemporal representations and more accurate predictions.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Furthermore, the computational efficiency of the system remains manageable, with only a slight increase in time complexity due to the multimodal fusion. This makes the system viable for real-time applications, such as live SL interpretation or gesture-based HCI systems.

In comparison to existing methods, the proposed system demonstrates a marked improvement in recognizing continuous SL gestures from video streams, overcoming the limitations of traditional models that rely on noisy frame-wise labeling or insufficient temporal modeling. Our approach also sets the stage for future work in SL recognition, particularly in terms of incorporating more advanced multimodal data and refining the iterative training process.

The proposed continuous SL recognition system, based on a Recurrent Convolutional Neural Network (RCNN) with iterative training and multimodal fusion, outperforms existing methods in terms of both recognition accuracy and computational efficiency. The integration of gloss-level gestural supervision and the use of optical flow data as a complementary modality contribute to the system's robust performance. Experimental results show that the proposed system is capable of accurately transcribing continuous SL videos into ordered gloss sequences, making it a promising solution for real-world SL recognition applications.

In practical deployment scenarios, continuous sign language recognition systems are affected by signer variability, illumination changes, background clutter, partial occlusion, and camera viewpoint variations. Although the proposed multimodal RCNN framework demonstrates robust performance under controlled experimental conditions, future improvements should incorporate adaptive normalization techniques, pose-guided attention mechanisms, and domain adaptation strategies to improve robustness in unconstrained environments.

CONCLUSION

In this paper, a novel Recurrent Convolutional Neural Network (RCNN)-based framework for continuous sign language recognition was presented using iterative optimization and multimodal fusion techniques. The proposed framework effectively captures both spatial and temporal characteristics of sign gestures by integrating CNN-based feature extraction with Bi-LSTM sequence learning.

The iterative training mechanism significantly improved temporal alignment and reduced Word Error Rate (WER), while multimodal fusion using RGB frames and optical flow data enhanced gesture representation by combining appearance and motion information. Experimental evaluation demonstrated that the proposed framework achieved superior recognition accuracy compared with existing CNN-HMM and conventional RNN-based approaches.

The proposed framework also demonstrated promising computational efficiency for real-time sign language recognition applications. Although the current work focuses on a limited dataset, the architecture is scalable for larger sign vocabularies and multi-signer environments.

Future work will explore transformer-based sequence modeling, signer-independent adaptation, large-scale dataset evaluation, and advanced multimodal integration techniques to further improve robustness and real-world deployment feasibility.

REFERENCES

1. S. C. Ong and S. Ranganath, “Automatic sign language analysis: A survey and the future beyond lexical meaning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 6, pp. 873–891, 2005.
2. Z. Ren, J. Yuan, J. Meng, and Z. Zhang, “Robust part-based hand gesture recognition using Kinect sensor,” IEEE Trans. Multimedia, vol. 15, no. 5, pp. 1110–1120, 2013.
3. C. Wang, Z. Liu, and S.-C. Chan, “Superpixel-based hand gesture recognition with Kinect depth camera,” IEEE Trans. Multimedia, vol. 17, no. 1, pp. 29–39, 2015.

4. P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 4207–4215.
5. H. Cooper, E. J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units,"J. Mach. Learning Research, vol. 13, pp. 2205–2231, 2012.
6. P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in IEEE Conf. Comput. Vis. Pattern Recog. Workshops, 2015, pp. 1–7.
7. G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, "Learning personalized models for facial expression analysis and gesture recognition," IEEE Trans. Multimedia, vol. 18, no. 4, pp. 775–788, 2016.
8. G. D. Evangelidis, G. Singh, and R. Horaud, "Continuous gesture recognition from articulated poses," in Eur. Conf. Comput. Vis. Workshops, 2014, pp. 595–607.
9. N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in Eur. Conf. Comput. Vis. Workshops, 2014, pp. 474–490.
10. D. Wu, L. Pigou, P.-J. Kindermans, N. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 8, pp. 1583–1597, 2016.
11. O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," Comput. Vis. Image Understand., vol. 141, pp. 108–125, 2015.
12. O. Koller, H. Ney, and R. Bowden, "Deep hand: how to train a CNN on 1 million hand images when your data is continuous and weakly labelled," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 3793–3802.
13. O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: hybrid CNN-HMM for continuous sign language recognition," in Proc. Brit. Mach. Vis. Conf., 2016.
14. U. Von Agris, M. Knorr, and K.-F. Kraiss, "The significance of facial features for automatic sign language recognition," in 8th IEEE Int. Conf. Autom. Face Gesture Recog., 2008, pp. 1–6.
15. P. Buehler, A. Zisserman, and M. Everingham, "Learning sign language by watching TV (using weakly aligned subtitles)," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2009, pp. 2961–2968.
16. L.-C. Wang, R. Wang, D.-H. Kong, and B.-C. Yin, "Similarity assessment model for Chinese sign language videos," IEEE Trans. Multimedia, vol. 16, no. 3, pp. 751–761, 2014.
17. C. Monnier, S. German, and A. Ost, "A multi-scale boosted detector for efficient and robust gesture recognition," in Eur. Conf. Comput. Vis. Workshops, 2014, pp. 491–502.
18. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 2625–2634.
19. L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," arXiv preprint arXiv:1507.02159, 2015.
20. Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," IEEE Trans. Multimedia, vol. 19, no. 7, pp. 1510–1520, 2017.
21. L. Pigou, A. v. d. Oord, S. Dieleman, M. M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," Int. J. Comput. Vis., pp. 1–10, 2015.
22. O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2017.
23. T. Pfister, J. Charles, and A. Zisserman, "Domain-adaptive discriminative one-shot learning of gestures," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 814–829.
24. D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 724–731.
25. Kondragunta Rama Krishnaiah and H. Harish. "Performance Characterization and Comparative Study of DSR and OLSR Routing Protocols in Mobile Ad Hoc Networks." International Journal on Recent and Innovation Trends in Computing and Communication 11, no. 11 (December 2023): 1842–1846.
26. Kondragunta Rama Krishnaiah and H. Harish. "Understanding Emotions with Deep Learning: A Multimodal Approach for Detecting Speech and Facial Expressions." Journal of Computational Analysis and Applications 32, no. 1 (2024): 661–673. https://doi.org/10.48047/jocaaa.2024.32.01.21

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

27. Kondragunta Rama Krishnaiah, H. Harish, and Manjunath B. E. “Character-Level Convolutional Neural Networks for Cyberbullying Detection: A Robust Approach to Handling Noisy Social Media Text.” International Journal of Intelligent Systems and Applications in Engineering 12, no. 14s (February 2024): 746–752.
28. Kondragunta Rama Krishnaiah and H. Harish. “Image-Based Real Estate Appraisal: Leveraging Mask R-CNN for Damage Detection and Severity Estimation.” International Journal of Intelligent Systems and Applications in Engineering 12, no. 21s (March 2024): 4961–4969.