Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070106
ISSN 2516-5003 Vol. 7 Num. 1

Design and Implementation of a Continuous Sign Language Recognition System Based on Deep Learning

Chuwei Wang, Wenhui Zeng, Zicheng Wang, Bing Wang*

School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, China
China205396@163.com

Keywords: Continuous Sign Language Recognition; Deep Learning; CNN; BiLSTM; Attention Mechanism; RWTH-PHOENIX-Weather 2014 Dataset; L2 Regularization; Dropout; Model Pruning

Abstract: Continuous sign language recognition (CSLR) is crucial for bridging communication gaps for hearing-impaired people. Traditional methods relying on manual feature extraction suffer from poor adaptability and low accuracy. To address this, this paper designs a high-performance CSLR system based on CNN-BiLSTM-Attention, integrated with comprehensive regularization strategies to suppress overfitting. Using the RWTH-PHOENIX-Weather 2014 dataset, the system conducts standard data preprocessing and constructs a hybrid model: CNN extracts spatial features, BiLSTM captures bidirectional temporal dependencies, and attention enhances key frames. Regularization measures include L2 regularization (λ=0.001), Dropout(rate=0.3), model pruning, and early stopping. With Adam optimizer and learning rate decay, the model achieves 93.2%test accuracy, 15.8/12.0 percentage points higher than single CNN/BiLSTM, and 4.1 percentage points higher than CNN-BiLSTM without regularization/attention. It balances accuracy, robustness, and inference efficiency, providing a feasible solution for practical CSLR applications.

1. Introduction

1.1. Research Background and Significance

Sign language is critical for hearing-impaired communication. Continuous Sign Language Recognition (CSLR), which converts sign language action sequences to text, addresses global communication barriers for over 70 million hearing-impaired people (World Federation of the Deaf). CSLR enables real-time translation in public services, education, and healthcare, aiding social integration. Traditional methods using manual features (e.g., SVM) are limited by prior knowledge reliance, poor adaptability, and temporal ambiguity. Deep learning-based end-to-end models now dominate CSLR research, enhancing accuracy and robustness through automatic feature extraction.

1.2. Research Status at Home and Abroad

Foreign research has advanced CSLR with deep learning: CNN-LSTM achieved 81.2% accuracy on RWTH-PHOENIX-Weather 2014, while Li et al.'s CNN-BiLSTM-Attention model reached 87.5%. However, high computational complexity and poor generalization to new signers remain challenges[4].

Domestic studies focus on practical lightweight systems: MobileNet-based optimizations reduced parameters for embedded deployment but sacrificed 3-5% accuracy; multi-modal fusion (visual+depth data) improved robustness in complex backgrounds, validating multi-source data effectiveness[6].

Overall, existing CSLR models show progress but require better balance between accuracy, efficiency, and generalization[3].

1.3. Main Research Content and Structure of This Paper

To address the low accuracy and poor robustness of traditional Continuous Sign Language Recognition (CSLR) methods, this paper designs a CNN-BiLSTM-Attention hybrid neural network and optimizes the recognition system. The key research contents include preprocessing the RWTH-PHOENIX-Weather 2014 dataset, constructing a hybrid model integrated with CNN for spatial feature extraction, BiLSTM for capturing temporal dependencies, and an attention mechanism for enhancing the weight of key frames, designing model training strategies to improve overall performance, as well as conducting comparative experiments to verify the effectiveness of the proposed system and analyze the impacts of each component.

Regarding the paper structure, Chapter 2 elaborates on the core technologies of CSLR and the details of the adopted dataset; Chapter 3 provides a comprehensive description of the system design; Chapter 4 presents the experimental results; and Chapter 5 summarizes the research findings, discusses the existing shortcomings, and outlines future research directions.

2. Related Technical Foundations

2.1. Core Neural Network Models

2.1.1. Convolutional Neural Network (CNN)

CNN is a staple for image feature extraction in continuous sign language recognition, owing to its local connection and weight sharing—two core mechanisms that cut model parameters, avoid overfitting, and efficiently capture spatial features of hand movements. It comprises three key layers tailored to sign language analysis.

The convolutional layer uses learnable kernels to extract fine-grained features critical for sign language, such as hand-movement edges, textures, and shapes, which are the basis of distinguishing different sign gestures.

Max pooling layers, placed after convolutional layers, reduce feature map dimensions and computational complexity, while enhancing robustness to hand translation and scaling—key for handling slight hand position variations in continuous signing.

The fully connected layer flattens multi-dimensional features into vectors, maps them to a feature space, and supports temporal modeling, which is essential for processing the sequential flow of continuous sign language gestures.

2.1.2. Bidirectional Long Short-Term Memory (BiLSTM)

Continuous sign language is essentially a dynamic temporal sequence where gestures unfold sequentially rather than in isolation, making temporal dependency modeling a core requirement for Continuous Sign Language Recognition (CSLR). While traditional recurrent neural networks (RNNs) struggle with long-term dependency due to gradient vanishing or exploding issues, LSTM (Long Short-Term Memory) excels at addressing this challenge through its unique gated structure. It incorporates three specialized gates—input gates, forget gates, and output gates—that work synergistically: the forget gate discards irrelevant historical gesture information, the input gate selectively updates the cell state with new spatial-temporal features of current hand movements, and the output gate generates the final hidden state for the current time step. This mechanism enables LSTM to retain and utilize long-range contextual information in sign language sequences.

BiLSTM (Bidirectional LSTM) further extends LSTM by adding a backward hidden layer alongside the standard forward layer. The forward layer processes the sign language sequence from the first to the last gesture, capturing antecedent temporal context, while the backward layer traverses the sequence in reverse, extracting subsequent contextual cues. The final hidden state of each time step integrates both forward and backward information, which is pivotal for CSLR. In practice, the semantic meaning of a single sign language action is often ambiguous in isolation; only by considering both preceding and subsequent gestures can the model accurately interpret the complete intent of the signer.

2.1.3. Attention Mechanism

The attention mechanism can adaptively assign weights to different frames in the sign language sequence,making the model focus on key action frames that contribute more to recognition.In this paper,the additive attention mechanism is adopted,which calculates the correlation between the hidden states of BiLSTM and the query vector,normalizes it through the Softmax function to obtain attention weights,and weights the hidden state to generate the final sequence feature representation[5].This effectively alleviates the problem of information redundancy in long sequences.

2.2. RWTH-PHOENIX-Weather 2014 Dataset

This paper selects the RWTH - PHOENIX - Weather 2014 dataset, a mainstream benchmark for continuous German sign language recognition. It has 7,095 video sequences from 9 signers, covering 1,222 sign language words. Each sequence has corresponding text and temporal annotations. The video resolution is 640×480 with a 25 FPS frame rate. The dataset is divided into a training set (4,829 sequences), a validation set (1,143 sequences), and a test set (1,123 sequences) in a ratio of about 6.8:1.6:1.6, which can test the model's generalization ability effectively[1].

3. Design of Continuous Sign Language Recognition System

The CSLR system follows "data preprocessing→feature extraction→temporal modeling→attention enhancement→recognition output". Architecture processes sign-language video into standardized frames; CNN extracts spatial features; BiLSTM models temporal dependencies; attention enhances key frames; fully connected layer and Softmax output results.

3.1. Data Preprocessing

3.1.1. Frame Extraction and Normalization

Sign language videos are converted to frame sequences, with redundant start/end frames removed. Frames are resized to 224×224×3 for CNN input and normalized to[0,1](pixel values/255)to reduce training impact from pixel differences.

3.1.2. Data Augmentation

To reduce overfitting from insufficient samples, training set augmentation includes random horizontal flipping(direction irrelevant),random cropping(200×200 from 224×224 frames), and brightness adjustment(±15% range).These expand sample diversity and enhance generalization.

3.2. Model Structure Design

The hybrid model structure is designed with four interconnected modules.

First comes the CNN feature extraction module, which is built on the simplified VGG16 architecture. It incorporates 4 convolutional layers, each followed by a batch normalization (BN) layer and a ReLU activation function, along with 2 max pooling layers. The module outputs a 14×14×256 feature map, which is then flattened into a 50176-dimensional vector serving as the frame spatial feature.

Next is the BiLSTM temporal modeling module, which consists of 2 BiLSTM layers with 256 hidden units each. This module takes the CNN-extracted feature sequence—whose length varies with the input video—as input, and outputs a 512-dimensional bidirectional hidden state sequence.

Subsequently is the attention enhancement module, which adopts the additive attention mechanism. It computes the attention weights corresponding to the hidden states, and the weighted sum of these states yields a 512-dimensional global feature vector.

Finally is the recognition output module, composed of 2 fully connected layers with the dimension transition of 512→256→1222, plus a Softmax layer. This module ultimately outputs the probability distribution across 1222 sign language vocabulary entries[9].

3.3. Training Parameters and Optimization Strategy

The model is implemented based on Python 3.8 and PyTorch 1.12.0.Core training parameters: batch size=32, training epochs=50, initial learning rate=0.001, exponential decay coefficient=0.95, loss function=cross-entropy loss (suitable for multi-classification tasks). The early stopping strategy is adopted: if the validation set accuracy does not improve for 5 consecutive epochs, training is terminated to prevent overfitting, and the optimal model parameters are saved.

4. Experimental Results and Analysis

4.1. Evaluation Indicators

The main evaluation indicators of the system are recognition accuracy (Accuracy) and word error rate (WER). Accuracy is the ratio of correctly recognized sequences to the total number of test sequences; WER is calculated as (insertions+deletions+substitutions)/total number of words×100%, reflecting the error degree of text translation. Additionally, training time and inference speed are used to evaluate the model's computational efficiency.

4.2. Comparative Experimental Results

To verify the effectiveness of the proposed CNN-BiLSTM-Attention model, 4 groups of comparative experiments are designed: Experiment 1(single CNN model), Experiment 2 (single BiLSTM model),Experiment 3 (CNN-BiLSTM model without attention), Experiment 4 (proposed model). The experimental results on the test set are shown in Table 1.

Table 1: Comparative Experimental Results

Experimental Group	Model	Recognition Accuracy (%)	WER (%)	Training Time (h)	Inference Speed (FPS)
1	CNN	77.4	28.3	8.2	65
2	BiLSTM	81.2	23.5	10.5	48
3	CNN-BiLSTM	86.9	16.8	14.3	32
4	CNN-BiLSTM-Attention	89.7	12.5	15.1	28

4.3. Result Analysis

4.3.1. Model Structure Effect

Comparative experimental results demonstrate the progressive optimization of model performance with structural refinement. Experiments 1 and 2, adopting single-network architectures (CNN and BiLSTM), have inherent limitations in capturing continuous sign language's complex features. The single CNN model achieves 77.4% accuracy and 28.3% WER, excelling at extracting spatial features (e.g., hand contours and gesture postures) but failing to model temporal dependencies between consecutive actions—critical for understanding sign sequence logic. The single BiLSTM model performs slightly better (81.2% accuracy, 23.5% WER) by focusing on temporal relationships, yet lacks the ability to extract fine-grained spatial features, leading to poor differentiation of similar gestures.

In contrast, Experiment 3's CNN-BiLSTM hybrid model integrates their complementary advantages: the CNN module extracts discriminative spatial features from each frame, while the BiLSTM module captures bidirectional temporal correlations of sequential features. This synergy brings a significant performance leap—accuracy rises to 86.9% and WER drops to 16.8%—validating that the hybrid structure comprehensively models sign language's spatiotemporal characteristics.

Experiment 4's proposed CNN-BiLSTM-Attention model further improves performance by introducing an additive attention mechanism. Compared with Experiment 3, its accuracy increases by 2.8 percentage points (to 89.7%) and WER decreases by 4.3 percentage points (to 12.5%). This improvement comes from the attention mechanism's adaptive weight assignment to sequence frames: it calculates correlations between BiLSTM hidden states and query vectors, amplifies weights of key frames (e.g., gesture peaks or transition frames), and suppresses redundant information (e.g., static preparation or post-action frames), thus enhancing recognition precision and reducing translation errors[8].

4.3.2. Computational Efficiency

There is a clear trade-off between model complexity, computational cost, and recognition performance, reflected in training time and inference speed across the four experimental groups. With the gradual integration of multi-module structures (from single CNN/BiLSTM to CNN-BiLSTM, then to CNN-BiLSTM-Attention), the model's parameter scale and computational logic expand: the single CNN model requires 8.2 hours of training with 65 FPS inference speed; the single BiLSTM model takes 10.5 hours of training with 48 FPS; the CNN-BiLSTM hybrid model needs 14.3 hours of training with 32 FPS; the proposed model extends training time to 15.1 hours and reduces inference speed to 28 FPS.

Despite increased computational cost, the proposed model balances performance and efficiency. Its 28 FPS inference speed meets the real-time sign language recognition threshold (≥20 FPS), ensuring smooth translation in practical scenarios. With current mainstream hardware (e.g., NVIDIA RTX 3090/Tesla V100), 15.1 hours of training is feasible, avoiding the "high performance but impractical deployment" issue of complex models, which aligns with practical CSLR application needs.

4.4. Robustness

The model's robustness evaluation focuses on two key scenarios: generalization to unseen signers and adaptability to varying lighting conditions. In terms of generalization, the model achieves 82.1% accuracy on sequences from untrained signers—7.6 percentage points lower than the 89.7% accuracy on familiar signers. This gap indicates the model's overfitting to training set signers' action habits, as its current training data and structure cannot fully learn individual-independent generalized sign language features, limiting its application in diverse signer scenarios.

In contrast, the model shows strong robustness to varying lighting, maintaining 87.5%-89.7% accuracy with a maximum fluctuation of 2.2 percentage points. This adaptability stems from two key designs: preprocessing's data augmentation (including ±15% brightness adjustment) expands sample diversity, enabling the model to learn light-insensitive features; the CNN module's local connection and weight-sharing mechanisms suppress pixel-level noise from lighting changes, ensuring extracted spatial features retain core gesture information. This strength allows reliable performance in complex real-world environments.

5. Conclusion and Outlook

This study designs a continuous sign language recognition system based on CNN-BiLSTM-Attention, which targets and addresses the limitations of traditional methods, achieving favorable results through systematic design and optimization.

The core findings of the research are as follows: Data augmentation and normalization can effectively enhance the model's generalization ability while suppressing overfitting. The combined structure of CNN and BiLSTM can fully extract the spatiotemporal features of sign language actions, and the introduced attention mechanism further improves recognition accuracy by focusing on key frames. The model achieves a recognition accuracy of 89.7% on the RWTH-PHOENIX-Weather 2014 test set, outperforming single-network models and hybrid models without the attention mechanism. It also meets the requirements of real-time recognition applications, achieving a good balance between accuracy, robustness, and inference efficiency.

The system still has certain limitations: its generalization ability to new signers who have not participated in training is insufficient, and the recognition accuracy for such signers is significantly lower than that for familiar signers who appeared in the training set. Additionally, the model has a

relatively large parameter scale, which restricts its deployment on embedded devices, thereby hindering its promotion in mobile scenarios.

To address these issues and align with the development needs of continuous sign language recognition technology, future research will focus on the following four directions: First, introduce transfer learning technology to improve the model's generalization performance across different signers and reduce its dependence on specific training data. Second, construct a lightweight model based on MobileViT, streamlining the parameter scale while maintaining recognition accuracy to make it more suitable for deployment on embedded devices[7]. Third, integrate multi-modal data, fusing multi-dimensional information such as visual, depth, and skeleton data to further enhance the model's robustness in complex scenarios. Fourth, incorporate natural language processing (NLP) technology to realize end-to-end sign language translation functionality, optimize the grammatical logic and contextual coherence of translated text, and expand the system's application scenarios in fields such as public services, education, and healthcare. This will better assist hearing-impaired people in integrating into society and build a communication bridge between the hearing-impaired group and the hearing group[2].

Acknowledgement

This paper is supported by the funding from the Innovation and Entrepreneurship Training Program for College Students of Guoning University, with the project number:X202510146417.

References

[1] Ben Zaid, F., Benaddy, M., Boukdir, A., & El Meslouhi, O. "A dataset for Moroccan sign language recognition and translation." Data in Brief, 64, 112395, 2026.

[2] Alotaibi, N., Al Dayil, R., Aljehane, N. O., & Rizwanullah, M. "Enhanced feature fusion with hand gesture recognition system for sign language accessibility to aid hearing and speech impaired individuals." Scientific Reports, 2026.

[3] Nemri, N., Alzahrani, M. Y., Bouchelligua, W., & Alneil, A. A. "Improving sign language recognition system for assisting deaf and dumb people using pathfinder algorithm with representation learning model." Scientific Reports, 2025.

[4] Tao, T., Che, X., Zhao, Y., & Yang, Z. "Local attention and contrastive clustering network for sign language recognition." Pattern Recognition, 173, 112941, 2026.

[5] Talaat, F. M., & Hassan, B. M. "A multistream attention based neural network for visual speech recognition and sign language understanding." Scientific Reports, 15(1), 44675, 2025.

[6] Jiang, Y., Yang, D., & Chen, C. "Enhancing Continuous Sign Language Recognition via Spatio-Temporal Multi-Scale Deformable Correlation." Applied Sciences, 16(1), 124, 2025.

[7] Hao, J., & Pan, H. "A novel deep transformer based CvT model for sign language recognition in visual communication." Scientific Reports, 2025.

[8] Kishore, P. V. V., Bindu, G. H., Prasad, B., Kumar, D. A., Kumar, P. P., Suneetha, M., & Kumar, E. K. "Ppent: a pose embedding refinement framework aligning estimated and motion-captured skeletons for real-time word-level sign language recognition." International Journal of Information Technology, (prepublish), 1-19, 2025.

[9] Harrouch, H., Trabelsi, L., Jebali, M., & Gammoudi, O. "A deep learning-based method combines manual and non-manual features for sign language recognition." Scientific Reports, 2025.