2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)
21-22 February 2024, Mazandaran University of Science and Technology, Babol, Iran

A two-stage sign language recognition method focusing on the semantic features of label text

Xuebin Xu
School of Computer Science and Technology
Xi'an University of Posts & Telecommunications
Xi'an, China
xuxuebin@xupt.edu.cn

Jun Fu
School of Computer Science and Technology
Xi'an University of Posts & Telecommunications
Xi'an, China
fujun_learner@stu.xupt.edu.cn

2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP) | 979-8-3503-8394-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/AISP61396.2024.10475205

***Abstract*— The ability to recognize sign language is an indispensable technology that plays a crucial role in facilitating communication between individuals who are deaf or hard of hearing. It is of utmost importance to comprehensively understand the nonverbal expressions employed by the hearing impaired. In order to enhance the efficacy of sign language recognition technology, it is imperative to focus on language modeling and improve the utilization of linguistic elements. At present, much attention in sign language recognition techniques that integrate language modeling is directed toward the translation of GLOSS to text in research related to Sign Language Translation (SLT). Our paper, however, proposes a creative approach that involves the linguistic modeling of the corresponding text of sign language during the process of converting signs to GLOSS. Specifically, we have implemented a text correction module that uses a front-mounted sign language recognition module to make preliminary predictions. The corrected GLOSS sequence is then used to obtain the final recognition result with higher accuracy. Our framework was tested on the RWTHPHOENIX-Weather-2014-T dataset and CSL dataset to evaluate its effectiveness in recognizing sign language on a large scale. The experimental results demonstrate that the proposed method significantly enhances the accuracy of the sign language recognition model.**

Keywords*— *Continuous Sign Language Recognition, Machine Learning, Language Modeling, Image Processing, Time Series Modeling

I. INTRODUCTION

For people who are hearing-impaired, sign language serves as a means of communicating their intentions. It involves a sequence of hand movements that convey messages through a series of continuous posture movements. The key distinguishing factor between sign language and common gestures is that sign language encompasses language functions through gestures. The app features a sign language presenter who demonstrates sign language actions, while the other party can observe and comprehend the meaning of these actions. The observer can then express these meanings in their natural language.

Sign language recognition tasks hold immense importance in facilitating communication between hearing-impaired individuals and those who do not use sign language. The accurate recognition and translation of sign language videos into corresponding language text enables a better understanding of the message conveyed by presenters. This technology has the potential to reduce the barriers to employment in the service industry for the hearing impaired, leading to increased inclusivity and equality. By meeting the service demands of the hearing-impaired population, this technology can help solve some of their life and survival difficulties. It is a crucial step towards creating a more accessible and equitable society.

Accurately recognizing sign language videos and converting them into corresponding linguistic texts is not only a feasible task but also demonstrates the potential of technology in the field of language translation. The unique features of sign language can be effectively captured and utilized through advanced modeling techniques, opening new possibilities for sign language recognition. Embedding or migrating these sign language recognition models into other automated applications or task models can expand their application scope and provide new directions and inspiration for further research or the development of automated intelligent devices. This cross-disciplinary integration offers vast opportunities for technological innovation, paving the way for numerous new research possibilities.

The goal of sign language recognition is precise identification, not only translating videos into written texts but also ensuring accuracy and smoothness in the process. Achieving the highest accuracy or lowest error rate necessitates continual optimization of recognition technology and algorithm refinement. Thus, the task demands a meticulous method, emphasizing the output's reliability and the system's overall efficiency and speed.

This paper presents a novel approach to improving the precision of sign language recognition systems. The proposed technique utilizes language models to guide the recognition process, incorporating a pre-sign language recognition module alongside a text correction module, an innovation not found in previous research. The text correction module serves to rectify the predictions made by the pre-sign language recognition module, culminating in the proposed sign language recognition model. Through the integration of these modules, the accuracy of the sign language recognition model is significantly enhanced.

II. RELATED WORK

When addressing the challenge of SLR, a frequently utilized approach involves employing a feature extraction

module for obtaining visual representations of videos, followed by the use of a sequence-modeling module to comprehend the long-term connections between these representations. This methodology has been documented in various literature.

Simultaneously using convolutional neural networks and recurrent neural networks to model sequences is also a classic approach. Many studies have shown that this method can obtain better sign language recognition performance. In addition, Connectionist Temporal Classification (CTC) is often used in the training of sign language recognition models. This approach was originally inspired by other sequence-to-sequence tasks such as speech recognition and handwriting recognition. Reference [1] proposed MC-LSTM by introducing adaptive fusion mechanism based on MV-LSTM, and used it together with spatio-temporal convolutional network for sign language recognition. Reference [2] proposed a sign language recognition method ODTL-SLRC combining EfficientNet, HGSO algorithm and BiLSTM. Reference [3] designed a system for automatic recognition of Indian Sign Language. The system uses CNN features and handcrafted features to be fused into a stacked BiLSTM network for modeling. Reference [4] designed a real-time dynamic sign language recognition system including gesture spotting, gesture sequence compression and gesture recognition. The system uses LSTMs to model sequences during both gesture discovery and gesture recognition. Reference [5] designed a spatio-temporal multi-cued network called STMC. The study used the network for spatial representation and temporal modeling of sign language videos, after which sequence learning continued with an LSTM-based encoder-decoder model. In order to realize the Indian sign language recognition task. Reference [6] first preprocessed the data and extracted features, and then used a module called MOPGRU for gesture recognition. MOPGRU is a module that integrates MediaPipe and GRU with improved update gates. After [7] performed principal component analysis on the data, they first used a Deep Restricted Boltzmann Machine (DBM) for feature fusion, and then used a gated recurrent unit to finally complete the classification and recognition of sign language.

Reference [8] proposed a dynamic sign language recognition method based on MediaPipe. In order to learn the dependencies between moving symbols, this method uses RNNs such as LSTM and GRU to learn sequence features. Reference [9] implemented a sign language recognition system including the InceptionResNetV2 model by using a combination of LSTM and GRU.

Machine translation and grammatical error correction are two typical text-to-text type tasks. Because they have important application scenarios in life and production, they have become the focus of many researchers. On the one hand, machine translation technology is also widely used to complete grammatical error correction tasks. Reference [10] regards College English grammar correction, which has an input granularity of character level and output granularity of sentence level, as a special text translation task. Based on RNN and attention mechanism, he designed a neural machine translation for grammatical error correction of English text. Model. On the other hand, the idea of grammatical error correction can also guide the improvement of translation algorithm performance. In order to obtain higher quality pseudo-parallel data in the detranslation task, [11] introduced a grammatical and spelling error correction model to process the input data of the model. There are also some studies that do not construct a grammar correction model based on an off-the-shelf neural machine translation model, but redesign the encoder-decoder network to output the target sequence. Reference [12] constructed a model dedicated to grammatical error correction based on multi-head attention mechanism and GRU. To implement error correction in Ukrainian text, [13] designed a pre-trained BERT (Bidirectional Encoder Representation from Transformers) type neural network and examined its performance using machine translation evaluation metrics.

III. METHODOLOGY

A. Overview

The proposed model for sign language recognition, as depicted in Fig. 1, consists of two distinct modules - the pre-installed sign language recognition module (SLRF) and the correction module (Corrector).

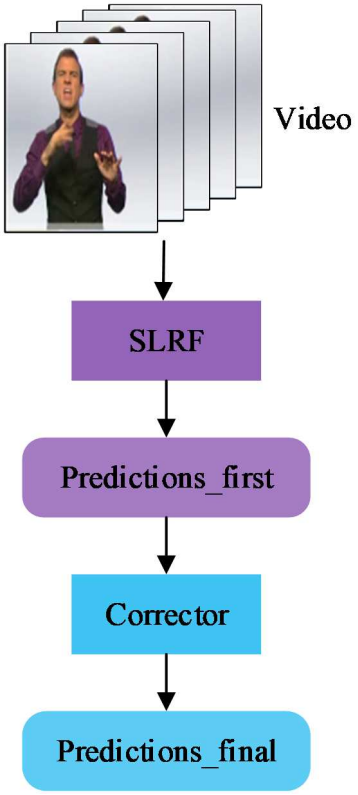

Fig. 1. Two-stage sign language recognition model with correction module

The SLRF module is responsible for extracting spatial visual features, short-term time features, and long-term time features from the sign language video. It then aligns and annotates the sign language frame image and label vocabulary using the CTC decoder, resulting in initial prediction results known as Predictions_first. This comprehensive approach enables the model to accurately recognize and interpret sign language gestures with high precision and efficiency. Recent studies have demonstrated that incorporating semantic information from label text can significantly enhance the accuracy of sign language recognition systems. This paper presents a novel method that utilizes an encoder-decoder model-based Corrector to refine the mapping relationship between Predictions_first and actual labels. Specifically, the encoder encodes the Predictions_first, while the decoder leverages an understanding of the semantic connections

within the label to decode the encoded information and generate the final prediction, Predictions_final.

B. SLRF

One commonly employed approach to recognizing continuous sign language is the use of convolutional networks for extracting visual and short-term time features, RNNs for capturing long-distance time features, and CTC for aligning the signs. This technique has shown promising results and has the potential to contribute to the development of more efficient and accurate sign language recognition systems.

The proposed pre-sign language recognition module, as illustrated in Fig. 2, utilizes a 2D-CNN and pooling layer to extract spatial visual features from the input frame image. Following this, short-term time features undergo processing by the BLSTM layer to learn long-term dependencies across all time steps. The results of the prediction are then processed by a Softmax-based classifier, CTC decoder, and beam search algorithm.

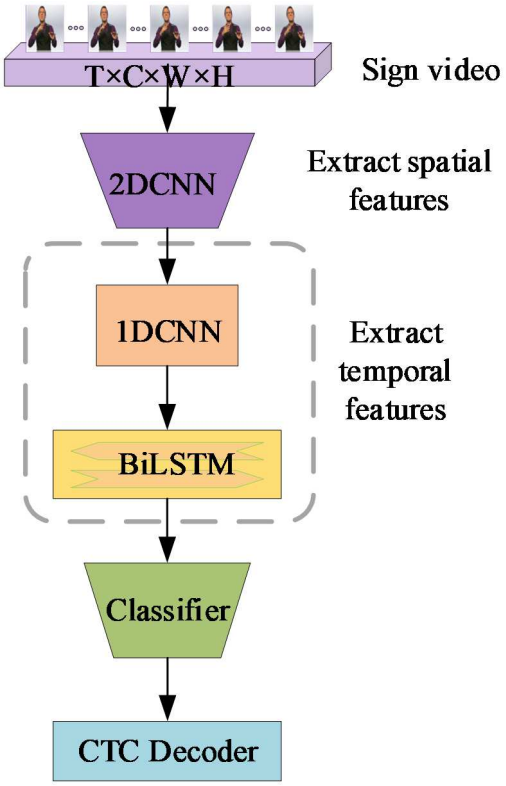

Fig. 2. SLRF

C. Corrector

When undertaking machine learning tasks that involve sequence-to-sequence processing, encoder-decoders are widely regarded as the standard method. Within the realm of natural language processing, encoder-decoders are commonly utilized to extract text meaning and generate sequences. Moreover, they are a classical and highly effective solution for machine translation and Grammatical Error Correction tasks. Our research endeavors to enhance this approach by designing a Corrector module that is founded upon the encoder-decoder architecture. This module functions to rectify the initial predictions of SLRF prediction (Preddictions_first), ultimately resulting in a final prediction (Preddictions_final) that exhibits improved accuracy.

As shown in Fig. 3, the proposed connector consists of an encoder (Encoder) and a decoder (Decoder). The Encoder encodes the output of SLRF, which is passed to the Decoder in the form of Hidden_encoder and Output_encoder. One prevalent technique for extracting and transmitting information from time series data is via the utilization of Gated Recurrent Unit (GRU) models. The Encoder utilizes BiGRU to extract time-related features from the data and pass them on, while the Decoder uses GRU to do the same. The Hidden_encoder denotes the output of the GRU shadow unit in the encoder, whereas Output_encoder represents the output of the output unit in the encoder. The bidirectional GRU employs the GRU as its underlying unit. Unlike the GRU, which only transmits past information to future nodes, the bidirectional GRU takes into account future information, thus providing a more accurate representation of time-related contextual information.

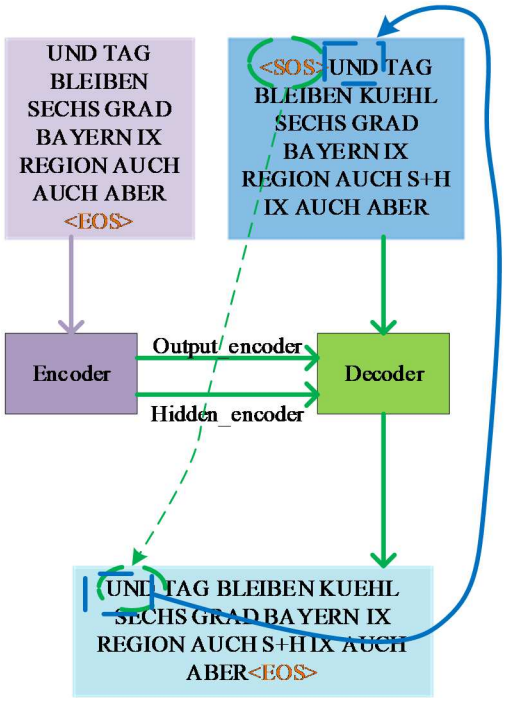

Fig. 3. Corrector

IV. Experiments and Results

A. Overview

In this study, we evaluated the effectiveness of our proposed method on two distinct datasets: RWTH-Phoenix-Weather-2014 and the Chinese Sign Language (CSL) Dataset. The experimental results demonstrated significant improvements when our method was applied to both datasets, further confirming its feasibility and effectiveness. Additionally, we conducted ablation experiments on the introduced modification module to determine its specific contribution to the overall performance. The results of these ablation studies indicated that the modification module consistently enhanced performance, regardless of whether it was applied to the RWTH-Phoenix-Weather-2014 or the Chinese Sign Language (CSL) Dataset, thereby verifying its efficacy and robustness.

B. Metrics

To evaluate the performance of our proposed method for continuous sign language recognition, we employed the word

error rate (WER) as a standard metric. WER is a widely accepted means of measuring the similarity between predicted and actual gloss, by calculating the minimum number of operations required to transform a predicted sign sequence into the ground truth. This calculation is determined by the following formula:

$$WER = \frac{S + I + D}{N} \times 100\% \quad (1)$$

The calculation of WER (Word Error Rate) involves four variables: S for the total number of replacements, D for the total number of deletions, I for the total number of inserts, and N for the total number of comments.

C. Dataset

The most widely used CSLR dataset is RWTH-Phoenix-Weather-2014, consisting of weather forecast records. The videos were filmed at 25 frames per second by 9 signers, each featuring a German sentence annotated in sign language. The dataset encompasses 5672, 540, and 629 videos for training, validation, and testing, respectively, and boasts a vocabulary size of 1295.

The CSL dataset is widely used in research for continuous sign language recognition. It consists of 100 pre-defined sentences in Chinese sign language, with a total of 25,000 videos performed by 50 sign language presenters. The dataset is divided into a training set and a test set using a method described in the literature.

There are some differences between RWTH-Phoenix-Weather-2014 and CSL. First of all, the natural language corresponding to RWTH-Phoenix-Weather-2014 is German, while the corresponding language of CSL is Chinese. Second, the vocabulary in RWTH-Phoenix-Weather-2014 focuses on the field of weather forecasting, and the vocabulary in CSL is Chinese words that are often used in daily conversations. In addition, the number of vocabularies involved in CSL is less than that of RWTH-Phoenix-Weather-2014.

D. Experiments on the dataset RWTH-Phoenix-Weather-2014

We compared the proposed method with previous methods and the results are recorded in the table I. The results of other methods are collected from their original papers or dataset release papers. We can see that our proposed SLRF_Corrector achieves the best results.

TABLE I. COMPARISON WITH METHODS ON RWTH-PHOENIX-WEATHER-2014 DATASET

Methods	WER	
	Dev	*Test*
Align-iOpt[14]	37.1	36.7
SF-Net(ResNet-18)[15]	35.6	34.9
Fully-Conv-Net [16]	23.7	23.9
DPD[17]	36.5	35.4
SFD+SGS+SFL[18]	26.2	26.8
CR3D[19]	——	24.4
MCSign-C[20]	35.2	35.3
SLRF_Corrector(this work)	**23.04**	**23.76**

E. Experiments on dataset CSL

We compared the proposed method with previous methods and the results are recorded in the table II. The results of other methods are collected from their original papers or dataset release papers. We can see that our proposed SLRF_Corrector achieves the best results.

TABLE II. COMPARISON WITH METHODS ON CSL DATASET

Method	WER
Align-iOpt[14]	6.1
DPD[17]	4.7
SF-Net[15]	3.8
Fully-Conv-Net[16]	3.0
SFD+SGS+SFL[18]	2.4
CrossModal[21]	2.4
SLRF_Corrector(this work)	**2.3**

F. Ablation Study

In this section, We conducted ablation studies to demonstrate the effectiveness of the correction module. As shown in Table III. The experimental results show that the sign language recognition model with the correction module achieves a lower word error rate on both datasets than the sign language recognition model without the correction module. This shows that the correction module can effectively improve the recognition accuracy of the sign language recognition system.

TABLE III. ABLATION STUDY ON CORRECTION MODULE

Method	RWTH-Phoenix-Weather-2014		CSLR
	DEV WER	*TEST WER*	*TEST WER*
SLRF	25.33	26.36	3.1
SLRF_Corrector	23.04	23.76	2.3

V. CONCLUSIONS

The recognition of sign language through technology is crucial for public welfare, industry, and academic research. To enhance the efficiency of machine learning algorithms, it is essential to incorporate more meaningful features. However, there is a lack of research on the internal feature information of the target sequence for sign language recognition. To address this issue, this study adopts the concept of grammatical error correction tasks and employs the encoder-decoder model to develop a sign language recognition model named SLRF_Corrector. The model includes an error correction module known as Corrector, which extracts these features and enhances the accuracy of sign language recognition.

We test the proposed method on the widely used RWTH-PHOENIX-Weather-2014 dataset and CSL dataset. The experimental results show that the SLRF_Corrector introduced into the Corrector module has achieved higher accuracy. This reflects the positive significance of the internal characteristics of the target sequence of sign language recognition for the realization of sign language recognition tasks.

The sign language recognition method proposed in this study has achieved commendable results on experimental datasets, yet it exhibits several aspects that require substantial improvements for broader application. Primarily, the method depends on the analysis of sign language videos, inadvertently overlooking key features typically found in other modalities, which limits its overall effectiveness. Additionally, this approach lacks a specific focus on optimizing processing speed, a crucial element for real-time applications. Therefore, to render this method more practical and effective in real-world scenarios, extensive optimization and deliberate

adaptation efforts are essential, particularly in terms of enhancing processing speed and incorporating a wider range of modalities.

VI. Future Work

Our future research will focus on two key directions. First, we plan to incorporate multimodal methods into our sign language recognition algorithms, enhancing accuracy in diverse environments. This approach integrates visual and sensory data, such as motion capture, for more effective recognition of complex sign gestures.

Second, we aim to improve the real-time performance of our system, making it more suitable for practical use. By employing parallel computing and algorithm optimization, we seek to increase processing speed and efficiency for smoother recognition experiences.

Ultimately, our goal is to develop an effective sign language recognition system, apt for real-world applications, to facilitate communication within the deaf community.

References

[1] O. Özdemir, İ. M. Baytaş, and L. Akarun, "Multi-cue temporal modeling for skeleton-based sign language recognition," *Frontiers in Neuroscience,* vol. 17, p. 1148191, 2023.

[2] F. Alrowais, S. S. Alotaibi, S. Dhahbi, R. Marzouk, A. Mohamed, and A. M. Hilal, "Sign Language Recognition and Classification Model to Enhance Quality of Disabled People," *networks,* vol. 9, p. 10, 2022.

[3] S. Das, S. K. Biswas, and B. Purkayastha, "Automated Indian sign language recognition system by fusing deep and handcrafted feature," *Multimedia Tools and Applications,* vol. 82, no. 11, pp. 16905-16927, 2023.

[4] M. Lee and J. Bae, "Real-time gesture recognition in the view of repeating characteristics of sign languages," *IEEE Transactions on Industrial Informatics,* vol. 18, no. 12, pp. 8818-8828, 2022.

[5] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for sign language recognition and translation," *IEEE Transactions on Multimedia,* vol. 24, pp. 768-779, 2021.

[6] B. Subramanian, B. Olimov, S. M. Naik, S. Kim, K.-H. Park, and J. Kim, "An integrated mediapipe-optimized GRU model for Indian sign language recognition," *Scientific Reports,* vol. 12, no. 1, p. 11964, 2022.

[7] H. Chen, D. Feng, Z. Hao, X. Dang, J. Niu, and Z. Qiao, "Air-CSL: Chinese Sign Language Recognition Based on the Commercial WiFi Devices," *Wireless Communications & Mobile Computing (Online),* vol. 2022, 2022.

[8] G. H. Samaan *et al.*, "Mediapipe's landmarks with rnn for dynamic sign language recognition," *Electronics,* vol. 11, no. 19, p. 3228, 2022.

[9] 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.

[10] X. Wu, "A Computational Neural Network Model for College English Grammar Correction," *Computational Intelligence and Neuroscience,* vol. 2022, 2022.

[11] N. L. Pham and T. V. Pham, "A Data Augmentation Method for English-Vietnamese Neural Machine Translation," *IEEE Access,* vol. 11, pp. 28034-28044, 2023.

[12] B. Chen and J. Zhang, "Pre-Training-Based Grammatical Error Correction Model for the Written Language of Chinese Hearing Impaired Students," *IEEE Access,* vol. 10, pp. 35061-35072, 2022.

[13] V. Lytvyn, P. Pukach, V. Vysotska, M. Vovk, and N. Kholodna, "Identification and Correction of Grammatical Errors in Ukrainian Texts Based on Machine Learning Technology," *Mathematics,* vol. 11, no. 4, p. 904, 2023.

[14] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4165-4174.

[15] Z. Yang, Z. Shi, X. Shen, and Y.-W. Tai, "Sf-net: Structured feature network for continuous sign language recognition," *arXiv preprint arXiv:1908.01341,* 2019.

[16] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai, "Fully convolutional networks for continuous sign language recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, 2020: Springer, pp. 697-714.

[17] H. Zhou, W. Zhou, and H. Li, "Dynamic pseudo label decoding for continuous sign language recognition," in *2019 IEEE International conference on multimedia and expo (ICME)*, 2019: IEEE, pp. 1282-1287.

[18] Z. Niu and B. Mak, "Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, 2020: Springer, pp. 172-186.

[19] C. Xue, M. Yu, G. Yan, M. Qin, Y. Liu, and J. Jia, "A multi-modal fusion framework for continuous sign language recognition based on multi-layer self-attention mechanism," *Journal of Intelligent & Fuzzy Systems,* vol. 43, no. 4, pp. 4303-4316, 2022.

[20] A. Akandeh, "Sentence-Level Sign Language Recognition Framework," *arXiv preprint arXiv:2211.14447,* 2022.

[21] I. Papastratis, K. Dimitropoulos, D. Konstantinidis, and P. Daras, "Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space," *IEEE Access,* vol. 8, pp. 91170-91180, 2020.