A Sign Language Recognition Based on Optimized Transformer Target Detection Model

Li Liu, Zhiwei Yang, Yuqi Liu, Xinyu Zhang, and Kai Yang(✉)

Nanjing Normal University of Special Education, Nanjing 210038, China
Yk@njts.edu.cn

Abstract. Sign language is the communication medium between deaf and hearing people and has unique grammatical rules. Compared with isolated word recognition, continuous sign language recognition is more context-dependent, semantically complex, and challenging to segment temporally. The current research still needs to be improved regarding recognition accuracy, background interference resistance, and overfitting resistance. The unique coding and decoding structure of the Transformer model can be used for sign language recognition. However, its position encoding method and multi-headed self-attentive mechanism still need to be improved. This paper proposes a sign language recognition algorithm based on the improved Transformer target detection network model (SL-OTT). The continuous sign language recognition method based on the improved Transformer model computes each word vector in a continuous sign language sentence in multiple cycles by multiplexed position encoding with parameters to accurately grasp the position information between each word; adds learnable memory key-value pairs to the attention module to form a persistent memory module, and expands the number of attention heads and embedding dimension by linear high-dimensional mapping in equal proportion. The proposed method achieves competitive recognition results on the most authoritative continuous sign language dataset.

Keywords: Sign Language Recognition · Target Detection Model · Neutral Network

1 Introduction

In recent years, sign language recognition technology has used intelligent devices like computers to convert sign language movements into messages that can be communicated with other social groups [1]. There are more than 5,500 commonly used words in Chinese sign language [2]. However, the number of people who understand the meaning of sign language is minimal, and most non-deaf people need to learn about sign language. Few of them are willing to spend time and effort to learn this skill, which is one of the reasons for the communication barrier between the deaf community and other social groups. Therefore, the study of sign language recognition technology can not only help deaf people to adapt to the social environment, promote the development of human-computer interaction and provide a better human-computer interaction experience for users [3].

B. Wang et al. (Eds.): ICMTEL 2023, LNICST 535, pp. 197–208, 2024.
https://doi.org/10.1007/978-3-031-50580-5_16

Research on sign language recognition in computer vision is progressing and evolving. Kollar et al. [4] proposed an end-to-end recognition method by embedding a convolutional neural network (CNN) into a hidden Markov model (HMM) and training the CNN using the frame state alignment generated by the HMM. Continuous sign language recognition can be considered a sequence-to-sequence learning problem. Several scholars have proposed methods based on extended short-term memory networks (LSTM) to capture temporal dependencies. Two typical alignment strategies are usually used to align input video sequences and target sentence sequences: connectionist temporal classification (CTC) sequence modeling approaches without predefined alignment and recurrent neural network encoder-decoder frameworks [5, 6]. However, sign language videos inherently contain complex feature information, including feature associations between hands, faces, and torsos in a single frame and feature variations of individual body parts between frames. Therefore, when dealing with the recognition task of longer sign language videos conforming to natural scenes, traditional recognition methods cannot extract enough useful features in end-to-end training and need more ability to model the video sequence.

In recent years, the Transformer model [7] has been gradually extended from the field of natural language understanding to the area of computer vision due to its remarkable long sequence modeling capability. The use of its global self-attention mechanism enables the parallel implementation of sequence-to-sequence recognition and translation tasks, which makes the model a new framework for many machine translation tasks, including continuous sign language recognition tasks. However, due to the weak position encoding capability of the Transformer model, it is difficult to accurately grasp the positions of individual word vectors in a sign language video sequence using only a simple sine cosine position encoding method. On the other hand, when dealing with continuous sign language recognition tasks, the traditional Transformer model uses only simple self-attention and multi-attention methods, which makes it challenging to realize the overall modeling of longer sign language sequences, resulting in poor recognition results [8].

Considering the above problems, we propose the SL-OTT algorithm to extract visual features more effectively and improve the accuracy and robustness of the model. The proposed SL-OTT improves the Transformer model to make it more suitable for the continuous sign language recognition task.

The main contributions of this paper are as follows.

1) A continuous sign language recognition method based on the improved Transformer model named SL-OTT is proposed, which accurately grasps the position information between each word in a constant sign language sentence by multiplexed positional encoding with parameters for each word vector in multiple cycles.
2) To enhance the model's ability to model long sequences of sign language videos overall modeling capability, we add learnable memory key-value pairs to the attention module to form a persistent memory module. In addition, the number of attention heads and embedding dimensions is expanded proportionally through linear high-dimensional mapping. The Transformer model's multi-head attention mechanism is used to maximize the overall modeling ability of long sign language sequences to deeply dig into the critical information in each frame inside the video.

The rest of this article is arranged as follows: Sect. 2 describes the related work; Sect. 3 depicts the proposed algorithm in detail, including the overall architecture; Sect. 4 analyzes the experiment result; At last Sect. 5 concludes this paper.

2 Related Work

In the past three decades, the research on sign language recognition in computer vision has continued [9]. The ultimate goal of sign language recognition tasks is to achieve recognition of sign language actions through spatial-temporal modeling and to be able to translate sign language videos into spoken sentences. So far, most research has focused on sign language recognition of isolated words dedicated to application-specific datasets [3], thus limiting the applicability of these techniques. In recent years, there has been an increase in research on continuous sign language recognition tasks.

Yang et al. [10] used a threshold model based on conditional random field (CRF) to determine whether each frame in an utterance is a sign language word or a transition action, and then used CRF to recognize the segmented sign language words and finally achieved 87% recognition rate in an American sign language utterance database consisting of 48 words. The threshold model could be better for detecting sign language word boundaries in practical applications because of the large variability of sign language data from non-specific populations. Cui et al. [3] extracted the spatial features of each frame by CNN and then pulled the spatial-temporal characteristics of each sign language segment by superimposed temporal convolutional and temporal pooling layers. Ren et al. [11] proposed a machine learning method to classify cancers, as pattern recognition for medical images is widely used in computer-aided cases. Wang et al. [12] used the PSO-guided self-tuning CNN to diagnosis COVID-19. Deep learning-based method can help classify medical images and efficiently improve the accuracy of diagnosis. Currently, most of the sign language word boundary detection algorithms need to be more robust to non-specific people, which affects the recognition of sign language utterances to some extent.

With the continuous development in the field of neural machine translation NMT, many excellent coding and decoding networks have been proposed, the most important of which is the Transformer model. In 2017, Vaswani et al. [9] proposed a Transformer model based on an attention mechanism, which replaces recurrent neural networks (RNNs) with a full-attention structure to achieve parallel computation while using a multi-headed attention mechanism [14] to capture the dependencies between the preceding and following texts fully and to grasp the dependencies between long-interval word vectors accurately. Transformer is not only applicable to machine translation tasks [15] but has also been successful in various other challenging tasks, such as language modeling, sentence representation learning, speech recognition, etc. And the Transformer model entirely takes into account the contextual issues in language translation during the operation, and its end-to-end recognition method is also well suited for solving continuous sign language recognition tasks; therefore, the Transformer model is widely used in endless sign language recognition tasks. Huang et al. [16] proposed a sign language attention network SAN that models the context on an entire frame sequence, modeling hand sequences on cropped hand images and combining hand features with their corresponding spatiotemporal contextual features using a self-attentive mechanism. Inspired

by the above work, this paper proposes an improved Transformer model for continuous sign language recognition, effectively improving the position encoding and overall modeling ability of the network for longer sign language sequences.

3 Methodology

In this section, we mainly introduce the overall architecture of the proposed SL-OTT and related modules.

3.1 Overall Architecture

The continuous sign language recognition method based on the improved Transformer model proposed in this paper is shown in Fig. 1. The overall algorithm framework consists of two significant parts of codecs. The model mainly addresses the problems of insufficient location coding ability and weak modeling ability of long sequences in the original method and proposes three optimized modules of multiple reusable learnable location coding (LPE), durable memory module (SMM), and attention extension module (AEM).

In Fig. 1, the overall process of the model is as follows: first, the location encoding with parameters (LPE) is reused in front of each multi-headed attention module, and the location encoding weight of each word vector is continuously updated according to the training loss rate to achieve an accurate grasp of the location of each word vector in longer sentences; second, in response to the difficulty of modeling long sequences of sign language videos in the Transformer model, the depth and accuracy of the attention module are expanded by adding persistent memory vectors (SMM) to each multi-headed attention module. At the same time, the number of heads of the attention module and the number of dimensions assigned to each head (AEM) are expanded year-on-year using a high-dimensional linear mapping, which further enhances the overall modeling ability of the model for longer sequences without reducing the perceptual field assigned to each head.

Finally, the modeled sign language sequences are translated, and the CTC model outputs the final recognition results. Finally, the modeled sign language sequences are translated by the CTC model, and the final recognition results are output. The overall modeling ability of the model for extended sequences of sign language videos is enhanced by the joint improvement method of multiple modules, which effectively improves the recognition accuracy of the model for sign language videos.

3.1.1 Learnable Location Coding

Due to the unique self-attentive computation of the Transformer model, the position of the input sequence needs to be encoded to prevent information loss, generally by adding position information to the input sequence using a sine and cosine function before the encoder. However, the semantic relationships between individual words of longer continuous sign language video sequences are more complex. Using simple cosine position encoding methods, capturing the position relationships between sign language

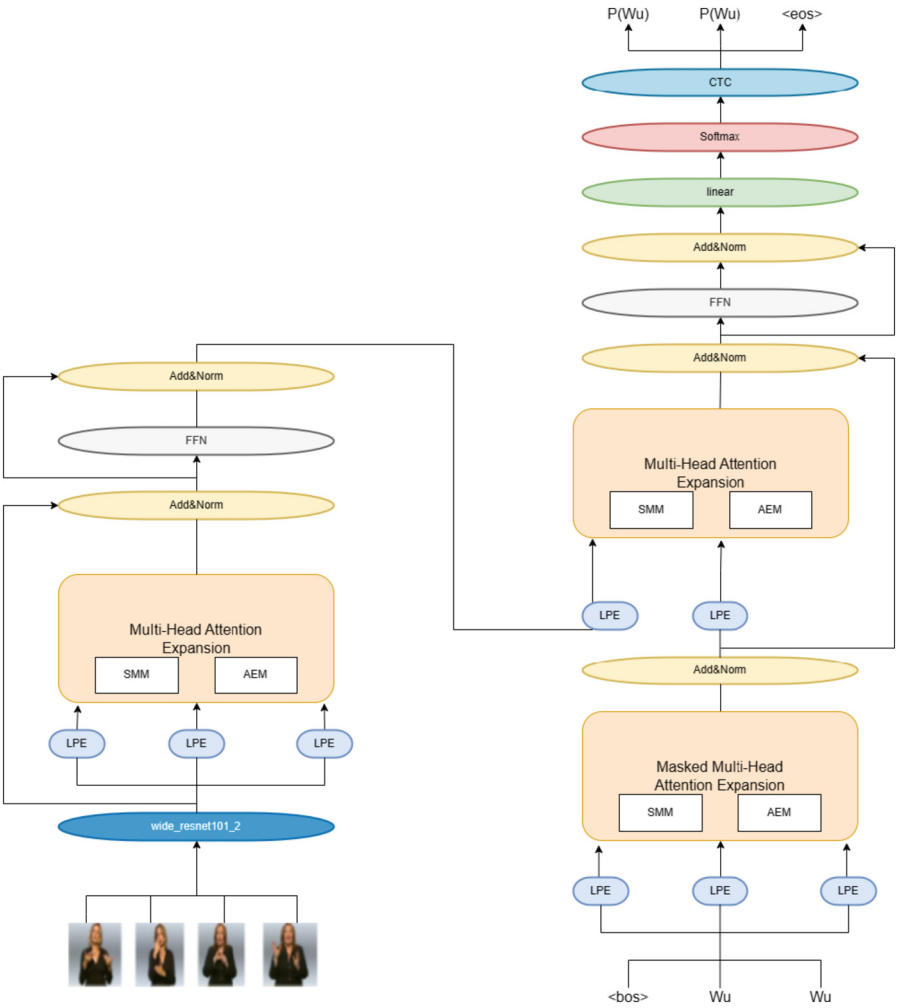

Fig. 1. The overall structure of the SL-OTT.

context word vectors in long sequences is difficult. Therefore, in this paper, we adopt a multiple reuse learnable position coding method to grasp better the general semantic information of the sign language video and make the syntactic relationships between word vectors more reasonable.

The location encoding layer directly inherits a matrix nn.Embedding, the length of which is the size of the dictionary and the width, is used to represent the attribute vector of each element in the dictionary, which is used to realize the mapping of words to word vectors. Then, the weight matrix in Embedding is randomly initialized, and the weight values are updated iteratively during the training process. The model can automatically learn the location information that matches the current word vector better.

As shown in Fig. 1, the SL-OTT incorporates learnable position encoding in each encoder, with the first encoder input coming from the image features and the subsequent encoder input coming from the output of the previous encoder, to achieve accurate encoding of each key frame in the phrase sequence by multiplexing them in multiple places.

3.2 Deepening Attention to the Durable Memory Module

To solve long sequential sign language recognition tasks, it is of utmost importance to make the attention module entirely, profoundly and accurately mine all the input sequences' feature information and capture the model's long-term dependencies. In the literature [17], it is proposed that adding a memory key value vector to the multi-headed self-attentive model can achieve a similar effect to the feedforward layer, thus removing the feedforward network layer and reducing the parameter computation; to improve the modeling ability of the sign language sequences, this paper adds a memory key value vector to the multi-headed attention module, but retains the feedforward layer and expands the attention depth and breadth of the self-attentive module to make it more suitable for continuous This paper refers to the persistent memory module (SMM) as a sign language recognition task.

In terms of mathematical interpretation, for a Transformer module i with input X, the final output of the module Y is:

$$y_j = \sum_{i=0}^{M} w_{ij} x_j \quad j = 1, 2, \cdots, n \tag{1}$$

In time series prediction, the autoregressive decoding of standard self-attentive models inevitably introduces huge cumulative errors. Different time series data usually have strong spatial dependence. Using the self-attentive model, multilayer perceptron and convolution module as encoders can eliminate the cumulative error to a certain extent, which can focus on both global and local information, and achieve more accurate and efficient modeling in the time domain.

3.3 Attention Extension Module

The Transformer model projects the input sequences to different subspaces of the self-attention layer to extract features, and the literature [14] indicates that increasing the number of heads in the multi-headed attention module can improve the model performance and increase the diversity of the attention graph when training the Transformer model in extension. Therefore, in this paper, the idea is carried over to the continuous sign language recognition task for long sequences, and the depth of attention of the model is enhanced by increasing the number of attention heads so that the model can better focus on the overall feature information of the sequence. However, for models with fixed embedding dimensions, an immediate increase in attention heads reduces the dimensionality assigned to each lead. The dimensionality reduction likewise affects the diversity of the attention maps. This paper adopts the direct expansion method to solve this problem.

The attentional map is mapped to a linear transformation matrix $\tilde{A} = [\tilde{A}^1, \cdots, \tilde{A}^{H'}]$. Through the linear transformation matrix W_A, the following equation is satisfied:

$$\tilde{A}^h = \sum_{i=1}^{H} W_A(h, i) * A^i, h = 1, \cdots H' \tag{2}$$

This method linearly maps the multi-headed self-attentive model into the high-dimensional space. It can ensure that the number of dimensions of each head is constant while appropriately increasing the number of attention heads, so that the model can enjoy both the benefits of more attention heads and the advantages of high embedding dimensions.

Since sign language behavior is naturally localized, the detailed movements of the palm and fingers are crucial for semantic information. Convolutional modules with local feature extraction performance are added after multi-headed self-attentive to form the core components of the Transformer. They learn they shared position-based kernel functions over a local window that maintains translational variance and can capture features such as edges and shapes, allowing the model to focus more on regional characteristics of the sign language. The convolution module consists of three convolution layers from shallow to deep: point-by-point convolution, deep convolution and a third layer of point-by-point convolution. The convolution module starts with a gating mechanism, point-by-point convolution and a gated linear unit, followed by a 1D depth convolution. After a 1D depth convolution, batch normalization and the Swish activation function train the depth model.

The Swish function is unsaturated, smooth and non-monotonic, and using the Swish activation function to regulate the network can speed up the convergence of the model.

4 Experimental Analysis

This section describes the datasets used for the experiments and the metrics used to evaluate the model performance, the experimental details, recognition results and analysis.

Experiments are conducted on two large publicly available continuous sign language benchmark datasets, RPW (RWTH-PHOENIX-Weather) 2014 [19] and RPW 2014 T [20]. The above two datasets are the classic benchmark for current domestic and international sign language video recognition research.

4.1 Experiment Settings

The network model was implemented on Ubuntu 14.04 with the following configuration: Intel(R) Pentium(R) CPU G3260 @3.30 GHz, NVIDIA GTX 3090, and 1 TB hard disk.

An 18-layer two-dimensional ResNet [12] is used for the CNN module. The final fully-connected layer is removed. A two-layer CM-Transformer encoder with four multi-headed attentions is used, with a model dimension of 512 and a position feedforward layer dimension of 2,048. CTC as decoder to generate complete sign language sentences. The complexity of the whole network is calculated: the number of parameters is 25828294

and the number of floating-point operations per second is 1.88 GMac; the complexity of the modified Transformer model is 0.06. The weights of the whole network are initialized at the beginning of the training phase.

In this paper, all persistent memory vectors are reparametrized according to the number of dimensions and thresholds of the original vectors and embedded in all head-shared locations so that the added constant memory vectors have the same unit variance as the initial context vectors.

We use *Word Error Rate (WER)*, the most commonly used word error rate for continuous sign language recognition tasks, as the evaluation index. The lower the *WER* value, the better the model effect and the higher the accuracy rate. The lower the *WER* value, the better the model and the higher the accuracy. Each experiment's *del/ins* value was recorded as an additional reference for evaluating the model.

$$WER = \frac{\#substitutions + \#insertions + \#deletions}{\#glosses\ in\ reference} \tag{3}$$

4.2 Comparison with Other Algorithms

The proposed SL-OTT is compared with the comparison method regarding word error rate and recognition time, where the lower word error rate indicates better recognition. Among the comparison methods, the sign language self-attentive model (SLT, sign language transformer) method [4] uses a self-attentive model for modeling; the stochastic frame loss + stochastic gradient stopping + stochastic fine-grained labeling (SFD + SGS + SFL) method [5] uses a continuous sign language recognition method with a self-attentive model encoder and a CTC decoder, while using The iterative re-aligned (Re-aligned) method [6] proposes an algorithm to process the provided training labels and dynamically refine the label-to-image alignment in a weakly supervised manner; the staged-Opt method [3] solves the video mapping problem by introducing a recursive convolutional neural network for spatial-temporal feature extraction and sequence learning to recursive convolutional neural network for spatial-temporal feature extraction and sequence teaching to solve the mapping of video clips to labels. The iterative alignment network (Align-iOpt) method [17] uses a 3D convolutional residual network and an encoder-decoder network with CTC for sequence modeling, which is optimized in an alternating manner; the CNN-LSTM-HMM method [12] embeds a powerful CNN-LSTM model in each HMM stream according to a hybrid approach.

The comparison results of recognition results between the proposed method and the comparison method on the two datasets are shown in Table 1 and Fig. 2, respectively.

As shown in Table 1 and Fig. 2, the proposed SL-OTT achieves better accuracy on the RPW 2014 dataset, reduces the word error rate by 1.5% compared to the currently available state-of-the-art SFD + SGS + SFL method, and achieves competitive results on the RPW 2014 T dataset. The proposed method outperforms the CNN-LSTM-HMM method in terms of recognition time. The proposed method outperforms SFD + SGS + SFL in comparison with the recognition method using the self-attentive model. In contrast, SLT uses a complete self-attentive model codec model with high model complexity, extensive memory usage, and extended training time. Both methods require

50,000 rounds of training before the model converges under the same hardware and software conditions. At the same time, the proposed method only requires 30 games.

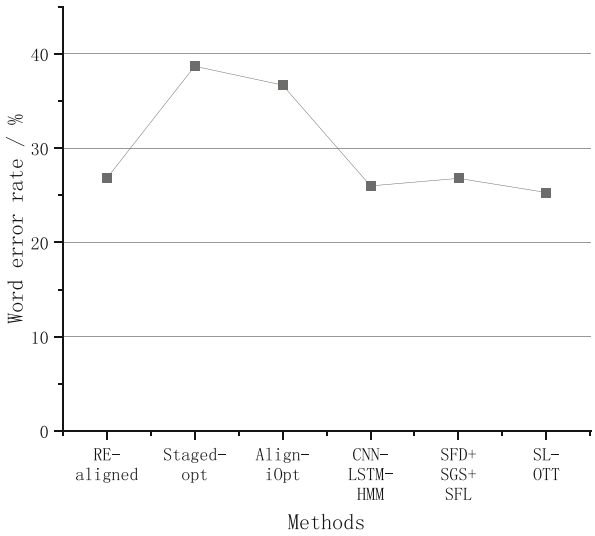

Fig. 2. Comparison of identification results of various methods on the RPW 2014 dataset.

Table 1. Comparison of identification results of various methods on RPW 2014 T dataset.

Methods	Word error rate / %	Time/h
CNN-LSTM-HMM	20.1	178.1
SLT	20.2	129234.7
SFD + SGS + SFL	22	5.6
SL-OTT	21.3	4.9

At the same time, the proposed SL-OTT can make up for this by combining the self-attentive model with the convolutional model to extract features for local gesture changes of the hand, which can bring out the global nature of the self-attentive model while keeping the regional characteristics of the convolutional model. The proposed model can fully exploit the topology and shape of the hand pose, and finally obtain the desired sign language recognition results.

To verify the impact of choosing different modules in the network structure on the overall performance, the results of the ablation experiments conducted on the PHOENIX-Weather 2014 dataset in this paper are listed in Table 2 and Fig. 3.

From Fig. 3, we can see that wide_resnet101_2 can extract the features of sign language video more accurately due to its broader feature map, more significant number of channels and deeper layers, and obtains the best recognition results.

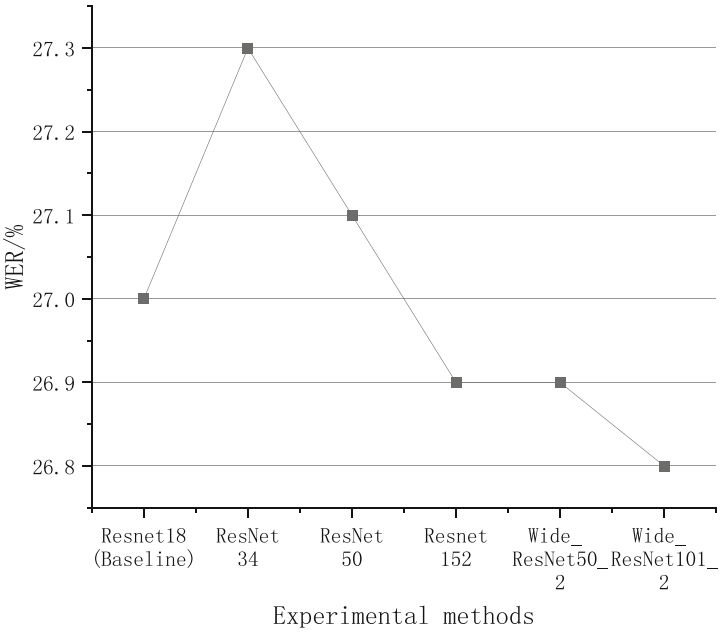

Fig. 3. Comparison of the effect of different types of CNN.

As shown in Table 2, the LPE module achieves accurate position encoding of continuous sign language video by multiplexing the learnable position encoding and reduces the model's error rate by 0.2%; the AEM module also proves to be effective in expanding the depth and breadth of attention, and the increase in the number of attention heads and embedding dimension effectively reduces the error rate by 0.3% when processing the long sequence of continuous sign language video. Due to its unique persistent memory mechanism, the SMM module reduces the model's error rate by 1.2%. The SMM module significantly reduced the model misspelling rate by 1.2% due to its unique ongoing memory mechanism, further demonstrating the importance of the attention mechanism in handling long sequence tasks such as continuous sign language recognition, and the addition of the persistent memory module enhanced the overall modeling capability of the model.

Table 2. Results of ablation experiments with different modules of LPE, SMM and AEM.

Experimental methods	Del/ins	WER/%
Baseline (Transformer)	5.2/5.0	22.2
LPE + SMM	8.2/2.1	21.1
LPE + AEM	7.1/3.0	22
SMM + AEM	6.2/2.4	24.4
SL-OTT	6.1/3.2	21.6

5 Conclusion

The study of sign language recognition technology not only enables deaf people to better adapt to the social environment but also promotes the development of human-computer interaction and provides a better human-computer interaction experience for users. This paper proposes an optimized Transformer method for continuous sign language recognition. The Transformer model is optimized and improved by multiplexing the learnable location encoding, the persistent memory module for deepening attention, and the attention extension module for long sequences, to solve the problem that the traditional Transformer model is weak in location encoding and challenging to model long lines of sign language videos. The proposed method can be applied to an extensive benchmark data set. The proposed method achieves significant recognition progress on a large benchmark dataset.

References

1. Koller, O., Zargaran, O., Ney, H., et al.: Deep sign: hybrid CNN-HMM for continuous sign language recognition. In: British Machine Vision Conference 2016, pp. 1–2. British Machine Vision Association, York (2016)
2. Zhang, Z., Pu, J., Zhuang, L., Zhou, W., et al.: Continuous sign language recognition via reinforcement learning. In: International Conference on Image Processing (ICIP), pp. 285–289. IEEE Computer Society, Piscataway, NJ (2019)
3. Cui, R., Liu, H., Zhang, C.: Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 7361–7369. IEEE Computer Society, Piscataway, NJ (2017)
4. Camgoz, N.C., Koller, O., Hadfield, S., et al.: Sign language Transformers: joint end-to-end sign language recognition and translation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10023–10033. IEEE Computer Society, Piscataway, NJ (2020)
5. Niu, Z., Mak, B.: Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI, pp. 172–186. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58517-4_11
6. Culati, A., Chiu, C.C., Qin, J., et al.: Conformer: convolution-augmented 'Transformer for speech recognition. In: Proceedings of the INTERSPEECH 2020, pp. 5036–5040. International Speech Communication Association (ISCA), Baixas (2020)
7. Koller, O., Zargaran, S., Ney, H.: Re-sign: re-aligned end-to-end sequence modeling with deep recurrent CNN-HMMs. InL l/IEEE International Conference on Computer Vision and Pattern Recognition, pp. 4297–4305. IEEE Computer Society, Piscataway, NJ (2017)
8. Graves, A., Fernandezs, G.F., et al.: Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 369–376 (2006)
9. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Adv. Neural Inform. Process. Syst. 5998–6008 (2017)
10. Molchanov, P., Yang, X., Gupta, S., et al.: Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4207–4215 (2016)

11. Ren, Z., Zhang, Y., Wang, S.: A hybrid framework for lung cancer classification. Electronics **11**(10), 1614 (2022)
12. Wang, W., Pei, Y., Wang, S.H., Gorrz, J.M., Zhang, Y.D.: PSTCNN: Explainable COVID-19 diagnosis using PSO-guided self-tuning CNN. Biocell **47**, 373–384 (2023)
13. Kollar, O., Camgoz, N.C., Ney, H., et al.: Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. IEEE Trans. Pattern Anal. Mach. Intell. **42**(9), 2306–2320 (2019)
14. Hartigan, J.A., Wong, M.A.A.: K-means clustering algorithm. J. R. Stat. Soc. Ser. C (Appl. Stat.) 28(1), 100–108 (1979)
15. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
16. Huang, J., Zhou, W.G., Zhang, Q.L., et al.: Video-based sign language recognition without temporal segmentation. In: AAAI Conference on Artificial Intelligence, pp. 2–7. AAAI, New Orleans (2018)
17. Camgoz, N.C., Hadfield, S., Koller, O., et al.: SubUNets end-to-end hand shape and continuous sign language recognition. In: IEEE International Conference on Computer Vision (ICCV), pp. 3075–3084. IEEE, Venice (2017)
18. Pu, J., Zhou, W., Li, H.: Iterative alignment network for continuous sign language recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4165–4174. IEEE Computer Society, Piscataway, NJ (2019)
19. Slimane, F., Bouguessa, M.: Context matters: self-attention for sign language recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 7884–7891. Milan, Italy (2021)
20. Camgoz, N.C., Hadfield, S., Koller, O., et al.: Neural sign language translation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7784–7793. IEEE Computer Society, Piscataway (2018)