Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.021784

ARTICLE

Continuous Sign Language Recognition Based on Spatial-Temporal Graph Attention Network

Qi Guo, Shujun Zhang* and Hui Li

College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, 266061, China

*Corresponding Author: Shujun Zhang. Email: zhangsj@qust.edu.cn

Received: 04 February 2022 Accepted: 05 May 2022

ABSTRACT

Continuous sign language recognition (CSLR) is challenging due to the complexity of video background, hand gesture variability, and temporal modeling difficulties. This work proposes a CSLR method based on a spatial-temporal graph attention network to focus on essential features of video series. The method considers local details of sign language movements by taking the information on joints and bones as inputs and constructing a spatial-temporal graph to reflect inter-frame relevance and physical connections between nodes. The graph-based multi-head attention mechanism is utilized with adjacent matrix calculation for better local-feature exploration, and short-term motion correlation modeling is completed via a temporal convolutional network. We adopted BLSTM to learn the long-term dependence and connectionist temporal classification to align the word-level sequences. The proposed method achieves competitive results regarding word error rates (1.59%) on the Chinese Sign Language dataset and the mean Jaccard Index (65.78%) on the ChaLearn LAP Continuous Gesture Dataset.

KEYWORDS

Continuous sign language recognition; graph attention network; bidirectional long short-term memory; connectionist temporal classification

1 Introduction

Sign language represents one of the primary forms of communication for the deaf and hard of hearing community and acts as an essential bridge between deaf (and hard of hearing) and hearing people. According to a survey by the World Health Organization, there are currently about 489 million people worldwide who have hearing disabilities and impairments.

Sign language utilizes hand and body movements to convey information. Therefore, the major task of sign language recognition (SLR) is to use computers to capture and understand the features of different behavioral sequences, translating them into texts or speeches. SLR is a challenging research topic encompassing computer vision, psychology, pattern recognition, and other research fields. It promotes and supports the integration of the deaf and hard of hearing into society. In addition, the research results in the SLR domain can be applied to gesture-related fields, including the game industry, military command, sign language teaching, and home automation, thus improving people's daily lives.

The current SLR methods can be divided into traditional machine learning methods and deep neural network-based methods. The former is difficult to extract representative semantic information and has poor real-time performance, while the latter can obtain higher accuracy and recognition speed, and has better real-time performance. It has become the mainstream of SLR methods [1]. Based on different research objects, SLR can be classified into isolated SLR (ISLR) and continuous SLR (CSLR). While the former aims to recognize single words or phrases, the latter translates sign language videos into corresponding spoken-language sentences. CSLR has important social significance in promoting communication between deaf and hearing people. Therefore, this work proposes a novel CSLR method based on spatial-temporal graph attention network (ST-GAT). The method aims to focus on local details and prevent the complex background in sign language datasets from interfering with the SLR. More precisely, OpenPose is utilized to detect the joints and bones. The two-stream information from joints and bones is then fused to build a spatial-temporal graph based on the physical connection of the human body and the same key points in different frames. The ST-GAT module is composed of the graph attention network (GAT) and temporal convolutional network (TCN), used to extract spatial and temporal feature sequences. These sequences are inputted into bidirectional long short-term memory (BLSTM) to obtain sign language word-level sequences. In addition, connectionist temporal classification (CTC) is used to align the sequences with no need for temporal segmentation.

The remainder of this paper is organized as follows. Section 2 introduces the related work on CSLR and graph neural networks (GNNs). Then, the overall framework and principle are described in Section 3. Section 4 discusses the implementation details and experimental results. Finally, the fifth section summarizes the work and outlines several directions for future research.

2 Related Work

In recent years, SLR has attracted significant attention due to its dependence on unique grammatical rules and rich visual information. As noted previously, SLR can be divided into ISLR and CSLR. However, ISLR [2–5] is a classification task unsuitable for real communication contexts. Thus, CSLR has become a dominant area of SLR research. CSLR typically relies on a feature extraction module to obtain visual representations from a sign language video and then utilizes a sequence learning module to learn long-term dependencies on visual representations. This section first briefly describes the sign language recognition methods with and without temporal segmentation; then, the research of graph neural network is discussed.

2.1 Temporal Segmentation Based Methods

Due to its powerful representation capabilities, many recent SLR methods are based on deep learning. For example, convolutional neural networks (CNNs) [6,7] and 3D CNNs [8–10] are used to model visual features and actions in sign language videos. Cui et al. [11] proposed an RGB and optical flow multi-mode fusion framework to segment sign language videos into ordered gloss label sequences. Further, the authors implemented an iterative optimization method to improve recognition performance. Zhang et al. [12] proposed determining a threshold matrix for coarse segmentation and rate thresholds for fine segmentation in the offline training stage. Then, the threshold matrix is used in the online recognition stage to perform coarse segmentation, and dynamic time warping is employed to determine the segmentation points.

Nevertheless, temporal segmentation involved in CSLR is a complex problem. If the segmentation is inaccurate, errors are propagated to subsequent operations. In addition, this process requires labeling

each isolated sign language vocabulary in a sentence, which is not only time-consuming and labor-intensive but also limits the dataset size.

2.2 Sign Language Recognition without Temporal Segmentation

With the advances in deep learning, direct recognition methods have emerged that do not require temporal segmentation. For example, Xiao et al. [13] proposed a bidirectional spatial-temporal long short-term memory (LSTM) fusion attention network to avoid sentence segmentation, word alignment, and manual labeling. Zhang et al. [14] proposed using a multimodal CNN to extract video features, LSTM to model temporal dependence, and CTC to bypass temporal segmentation and achieve end-to-end CSLR. In [15], a new transformer-based method was proposed to jointly learn CSLR and translation in an end-to-end manner. Zhou et al. [16] proposed a spatial-temporal multi-cue network that learns spatial-temporal correlations between visual cues in an end-to-end manner. This network uses BLSTM and CTC for sequence learning and interference. To facilitate feature extraction, Zhou et al. [17] proposed self-attention-based fully-inception networks with CTC loss and aggregation cross-entropy loss for end-to-end CSLR. In [18], the authors argued that overfitting based on the CTC method in CSLR stems from insufficient training of the feature extractor. Therefore, a visual alignment constraint was proposed to enhance feature extraction with alignment supervision. Cheng et al. [19] proposed an end-to-end full convolution CSLR framework, which learns video sequences' spatial and temporal features simultaneously given only sentence-level annotations. In [20], the authors proposed stochastic modeling of CSLR components, which generate random fine-grained labels for training the CTC decoder. The model utilizes ResNet18 as the visual model, transformer encoder as the context model, and CTC alignment model. To overcome the problems arising due to complex backgrounds and inconsistent illumination, Xiao et al. [21] proposed an SLR method for the Chinese language. The multimodal fusion method utilizes LSTM and a coupled hidden Markov model to fuse the hand and 3D skeleton sequence information.

2.3 Graph Neural Network-Based Methods

Islam et al. [22] performed hand detection and tracking, and then uses 2D CNN for gesture recognition. Due to the complexity of the environment, the influence of lighting and the color of clothes, the hand detection may be inaccurate, which affects the accuracy of gesture recognition. In recent years, methods utilizing skeleton data have attracted significant attention due to their robustness to background complexity and variations in illumination, body scale, and camera perspective. However, traditional CNN cannot process skeleton data. Therefore, researchers combined graph and deep learning, generating GNNs, including graph convolution network (GCN) and GAT. GCN and GAT have been widely used in various research fields [23–29], such as action recognition, text classification, and traffic forecasting. Yan et al. [23] proposed a skeleton-based spatial-temporal GCN. The spatial-temporal graph was constructed according to skeleton key points, and multi-layer spatial-temporal graph convolution was used to integrate spatial-temporal features. Huang et al. [24] developed a view-transformed graph attention recurrent network for view-invariant action recognition. The method uses GAT to automatically calculate the attention coefficient and extract the spatial features from skeleton data. Shi et al. [25] proposed a method that automatically learns the entire network's topology end-to-end by improving the adjacency matrix. The joints and bones information serves to form a two-stream network structure. Yao et al. [26] transformed the text classification problem into a graph node classification problem. The authors proposed a text GCN that captures the global word co-occurrence information from documents and makes full use of the limited document labels. Tackling the problem of temporal forecasting in traffic, Yu et al. [27] proposed a spatial-temporal GCN that uses

a pure convolution structure to build models with fewer parameters and fast training speed. Similarly, Sun et al. [28] proposed an end-to-end traffic forecasting method that dynamically models spatial and temporal correlation. It considers only the traffic conditions around the focus, simplifying the impact on the traffic network. To focus on inter-image dependencies, Zhang et al. [29] combined CNN and GCN for the purpose of extracting relation-awareness features using GCN, which achieved high accuracy in breast cancer classification tasks.

Inspired by the discussed research, this paper describes a new CSLR method based on ST-GAT. In order to avoid the interference of the complex background of the sign language behaviors, the two-stream information of joints and bones is extracted and constructed into a spatial temporal graph. The ST-GAT module composed of GAT and TCN deeply extracts the spatial-temporal correlation of skeleton data, and BLSTM and CTC perform sequence learning to obtain the final natural language sentences corresponding to sign language videos.

3 Proposed Method

The overall framework of the proposed method includes three stages: data preprocessing, feature extraction and sequence learning, as shown in Fig. 1.

Figure 1: Flowchart of the proposed framework

3.1 Framework Overview

Given a video containing T frames (denoted with $x = \{x_t\}_{t=1}^{T}$), the goal of CSLR is to extract the semantics of the sign language expressed in the video. Let the semantics be a sentence composed of L words $\ell = \{\mathrm{l_i}\}_{\mathrm{i=1}}^{\mathrm{L}}$. The proposed framework is shown in Fig. 2. The framework aims to avoid the complex background's influence and enable the focus on hand movements. Thus, OpenPose is utilized to estimate the human pose from the original video, obtaining the joints and bones information representation. Combining the joints and bones information enables constructing the spatial-temporal skeleton graph. Taking into account the localization of sign language and the feature correlation between skeleton points, the framework uses GAT to extract spatial features between key points and TCN to learn the short-term temporal correlation between frames. ST-GAT is composed of nine spatial-temporal graph attention layers (ST-GAT layer), including GAT and TCN. Here, "$\oplus$" represents residual connection. The spatial-temporal feature sequences extracted by ST-GAT are input to BLSTM to learn the long-term dependence, and the CTC is used to solve the problem of the alignment of input and output labels.

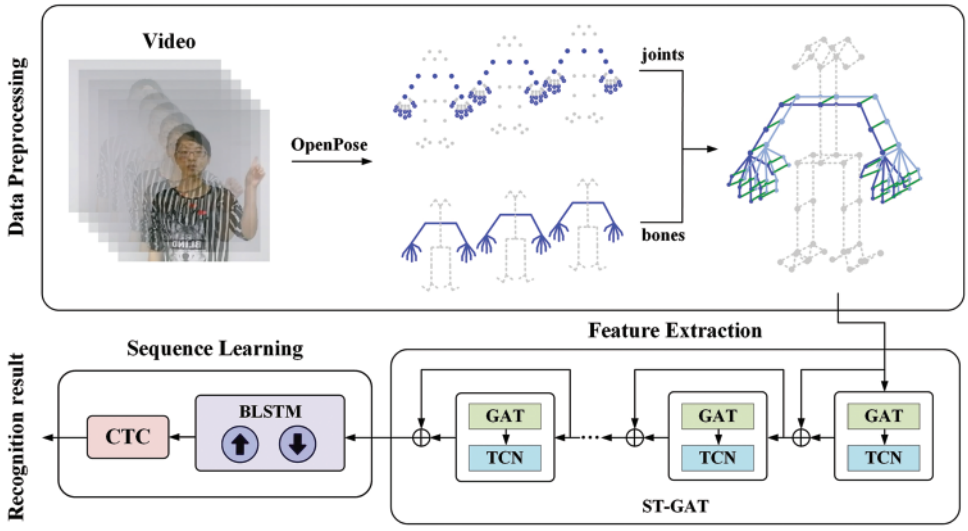


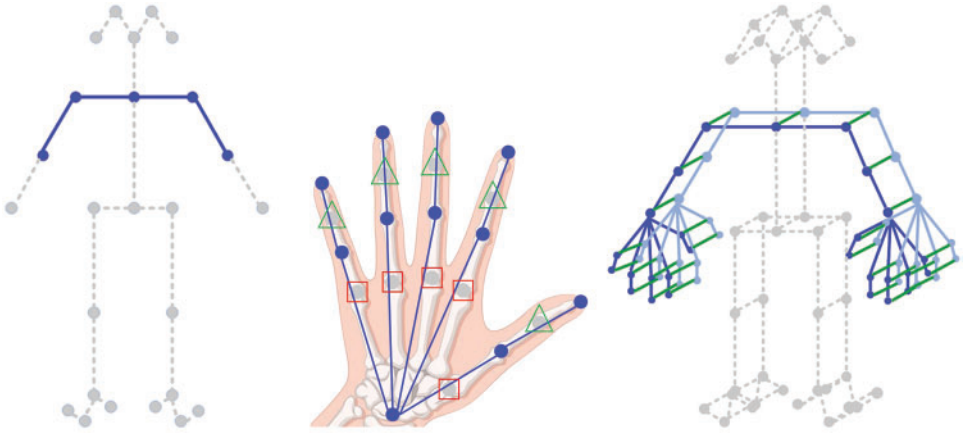
Figure 2: Overview of the proposed framework

3.2 Data Preprocessing

As noted previously, working with skeleton data successfully bypasses the problem of background complexity. Thus, using such data for SLR can significantly improve recognition accuracy. This work extracts skeleton data from sign language videos via OpenPose [30], an algorithm that marks joints and connects bones to estimate human pose.

Since the lower part of the body and face convey little information for SLR, these joints are removed from the set of detected joints, leaving only five body joints shown in Fig. 3a. To capture the flexibility of sign language and detailed hand movements, hand joints are detected in addition to body joints. OpenPose detects 21 joints of the hand by default. Due to the dense distribution of 21 hand joints, it will not only consume excessive memory, but also may yield redundant information. The metacarpophalangeal points (MCP) and distal interphalangeal points (DIP) have poor flexibility according to the activity of the hand itself. Thus, the MCP and DIP are omitted in this study, and only 11 hand joints are retained. Points in green triangles and red rectangles in Fig. 3b respectively represent the DIP and MCP that have been filtered out.

Next, an undirected spatial-temporal skeleton graph ($G = \{\mathcal{V}, \mathrm{E}\}$) is constructed from a skeleton sequence with V joints and T frames. As shown in Fig. 3c, the skeleton sequence contains both intra-body and inter-frame connections. The set of nodes of the spatial-temporal skeleton graph ($\mathcal{V} = \{v_{ti} | t = 1, \ldots, T; i = 1, \ldots, V\}$) includes all key points in the skeleton sequence. The edge set (E) consists of two subsets. The first subset, $\mathrm{E}_s = \{v_{ti}v_{tj} | (i,j) \in H\}$, contains the connections between nodes in each frame based on natural connections between human joints (represented with set H). These edges are presented in Fig. 3c by the solid blue lines. The second subset, $\mathrm{E}_F = \{v_{ti}v_{(t+1)j}\}$, includes inter-frame connections, connecting the same node in successive frames, as shown by the solid green lines in Fig. 3c. Finally, the dashed lines represent filtered (i.e., redundant) information.



(a) (b) (c)

Figure 3: Skeleton sample diagram. (a) Key points on a body; (b) Key points on a hand; (c) Spatial-temporal skeleton graph

3.3 Feature Extraction

The spatial-temporal skeleton graph obtained after data preprocessing is a non-Euclidean structure data. Therefore, it is irregular and can only be processed via GNNs. Many studies [23–29] have fully extracted spatial information using GNN.

Unlike other behavioral recognition tasks, SLR has evident local and detailed characteristics. The change in the two hands' position in frames and their interactive relationship play an important role in conveying signers' intention. Therefore, these aspects require special attention. To this end, this work proposes an ST-GAT module for extracting spatial features and modeling short-term temporal sequences of continuous sign language videos. The ST-GAT module consists of nine ST-GAT layers, and each layer comprises a GAT and a TCN. GAT uses an attention mechanism to aggregate features on neighboring nodes so that each node has a different weight.

The ST-GAT layer's input is denoted with $\mathbb{X}_{in} \in \mathbb{R}^{N\times C\times T\times V}$, where N is the batch size, C is the number of features of each node, T is the number of frames in the video, and V is the number of nodes in each frame. In an ST-GAT layer, GAT comprises a graph attention layer. Each frame $X = \{X_1, X_2, \ldots, X_V\}$, $X_i \in \mathbb{R}^C$ is processed separately in the graph attention layer to obtain new node features $X' = \left\{X'_1, X'_2, \ldots, X'_V\right\}$, $X'_i \in \mathbb{R}^{C'}$, where C' is the dimension of the new node feature vector.

To calculate the neighboring nodes' weight, the graph attention layer applies a linear transformation weight matrix(denoted $W \in \mathbb{R}^{C'\times C}$) to each node and executes the self-attention mechanism $a \in \mathbb{R}^{2C'}$, obtaining the attention coefficient. Attention coefficient e_{ij} represents the importance of node j relative to node i. It is calculated as

$$e_{ij} = a\left(WX_{ti}, WX_{tj}\right) \tag{1}$$

The attention mechanism in the graph structure allocates the attention only to the set of node i' neighbors (V_i). To promote the comparison of attention coefficients between different neighboring nodes, regularization is performed using Softmax and Leaky ReLU. Formally,

$$\alpha_{ij} = \frac{exp(LeakyReLU(a^T[WX_i||WX_j]))}{\sum_{k \in V_i} exp(LeakyReLU(a^T[WX_i||WX_k]))} \tag{2}$$

where $||$ is the concatenation operation and $\cdot^T$ represents transposition. These operations yield regularized attention coefficients for different nodes, enabling the prediction of each node's output feature. Namely, let σ denote the nonlinear activation function. Then,

$$X'_i = \sigma\left(\sum_{j \in V_i} \alpha_{ij} W X_j\right) \tag{3}$$

The graph attention layer uses a multi-head attention mechanism to stabilize the learning process. In other words, Eq. (3) uses K independent attention mechanisms. Their features are then joined (or averaged), yielding the following two output representations:

$$X'_i = ||_{k=1}^{K} \sigma\left(\sum_{j \in V_i} \alpha_{ij}^k W^k X_j\right) \tag{4}$$

$$X'_i = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \in V_i} \alpha_{ij}^k W^k X_j\right) \tag{5}$$

Once spatial features $\mathbb{X}' \in \mathbb{R}^{N \times C' \times T \times V}$ are obtained, temporal features are extracted from the spatial-temporal skeleton graph. Since the temporal convolutional operation considers the relationship between the same key point in different frames, and the spatial-temporal skeleton graph has a fixed shape, the traditional convolution network can extract temporal features. For T frames, the utilized TCN is a 2D convolutional network. The size of the convolution kernel is $K_t \times 1$, meaning that the convolution of one key point and K_t frames is completed each time. The stride is set to one to carry out the next node's convolution upon completing that of the current node. Finally, the spatial-temporal feature $\mathbb{X}_{out} \in \mathbb{R}^{N \times C' \times T' \times V}$ is obtained through representational learning.

The pseudo-code describing the ST-GAT layer is outlined in Algorithm 1.

3.4 Sequence Learning

Sign language recognition is a sequence learning task whose inputs are image frames and outputs are spoken-language sentences. After feature extraction, long-term dependency needs to be explored for the final recognition.

Recurrent neural networks (RNNs) represent the preferred approach for processing sequential data. However, the sequence length can hamper the propagation of the latter sequence's gradient back to the previous sequence, giving rise to the vanishing gradient problem. In addition, the hidden layer's input includes the outputs of both the input layer and the preceding hidden layer, and RNN has no memory function.

Algorithm 1: Pseudo-code of the ST-GAT layer

Input: $\mathbb{X}_{in} \in \mathbb{R}^{N\times C\times T\times V}$

Output: Spatial-temporal features $\mathbb{X}_{out} \in \mathbb{R}^{N\times C'\times T'\times V}$

Step 1 to Step 13: Spatial modeling on each frame

For each frame of data $X = \{X_1, X_2, \ldots, X_V\},\ X_i \in \mathbb{R}^C$ **do:**
  Compute the attention coefficient between node *i* and node *j* $e_{ij} = a(WX_{ti}, WX_{tj})$
  Perform regularization $\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k\in V_i} exp(e_{ik})}$
  If nheads != true (When nheads is true, it indicates the use of multi-head attention mechanism) **then:**
    $X'_i = \sigma(\sum_{j\in V_i} \alpha_{ij} W X_j)$
  Else
    If graph attention layer is not the last layer **then:**
      $X'_i = ||_{k=1}^{K} \sigma(\sum_{j\in V_i} \alpha_{ij}^k W^k X_j)$
    Else
      $X'_i = \sigma(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in V_i} \alpha_{ij}^k W^k X_j)$
    End
  End
End # Spatial features $\mathbb{X}' \in \mathbb{R}^{N\times C'\times T\times V}$ are obtained

Step 14 to Step 20: Temporal modeling on *T* frames in one batch

For each node in *T* frames **do:**
  Set the size of the convolution kernel to $K_t \times 1$
  If *T* frames do not complete the convolution **then:**
    The convolution of K_t frames is completed
    The convolution kernel moves between frames according to the preset stride
  End
End # Spatial-temporal features $\mathbb{X}_{out} \in \mathbb{R}^{N\times C'\times T'\times V}$ are obtained

The problems of vanishing gradient and the absence of memory function in RNN can be solved using LSTM. However, LSTM can perform only one-way transmission, meaning that it considers only the correlation between the current input and the preceding time frame. In the CSLR, the sign language video is translated as a sentence with grammatical rules, where each word depends on both the preceding and following words of the video sequence. Therefore, the proposed method utilizes BLSTM to learn the semantic association among the former and latter actions in sign language videos. BLSTM encodes the preceding and the following frames, computing both forward and backward hidden sequences and saving the past and future input information.

The ST-GAT module generates feature sequence $o = \{o_t\}_{t=1}^{T'}$, where T' is the temporal length of the TCN module's final output. Subsequently, BLSTM and CTC are used for sequence learning. The spatial-temporal features extracted by ST-GAT are sent to BLSTM to learn the long-term dependence and align the input and output via CTC to improve the recognition accuracy.

3.4.1 BLSTM

As shown in Fig. 4, the BLSTM network structure consists of two LSTM stacked in opposite directions. It contains six shared weights (denoted w1–w6). The forward and the backward layers jointly connect to the output layer to transmit the information.

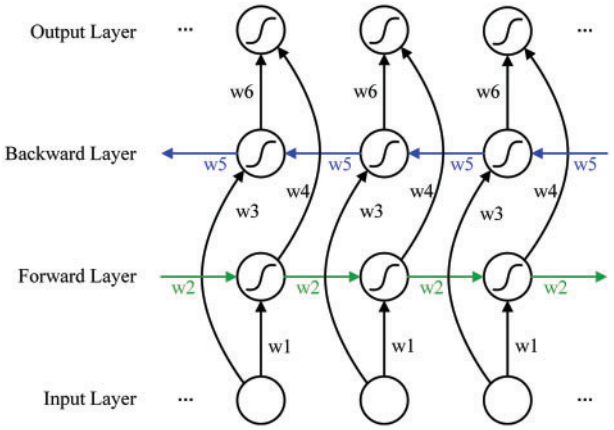


Figure 4: BLSTM network structure diagram. The green and blue lines represent forward and backward transmission, respectively

At each moment, the calculation is carried out in both the forward and the backward layers simultaneously to obtain and save the hidden layer's output. Then, the output is fed into Softmax.

3.4.2 CTC

CTC is mainly used to align the input and output. Therefore, the proposed method uses CTC [31] to map video sequences $o = \{o_t\}_{t=1}^{T'}$ to ordered gloss sequences $\ell = \{\ell_i\}_{i=1}^{L}$.

CTC uses an extended vocabulary $\mathbb{V}$ with a blank label "−" representing silence and transition. This label has no clear meaning. $\mathbb{V}$ is defined as $\mathbb{V} = \mathbb{V}_{\text{origin}} \cup \{-\}$. Now, the alignment path between the input sequence and the target gloss sequence is defined as $\pi = \{\pi_t\}_{t=1}^{T'}$, where $\pi_t \in \mathbb{V}$. Given an input sequence o, the probability of alignment path π is defined as:

$$p(\pi|o) = \prod_{t=1}^{T'} p(\pi_t|o) = \prod_{t=1}^{T'} y_{t,\pi_t} \tag{6}$$

where y_{t,π_t} is the probability of the label being π_t at time step t. Then, many-to-one mapping operation $\mathcal{B}$ that maps the aligned path π to the target sequence ℓ is defined, removing all blank and duplicate labels from the alignment path (e.g., $\mathcal{B}$(II−love−−you) = I, love, you). The conditional probability of sign gloss sequence ℓ is defined as the sum of the probabilities of all corresponding paths π:

$$p(\ell|\mathrm{o}) = \sum_{\pi \in \mathcal{B}^{-1}(\ell)} p(\pi|\mathrm{o}) \tag{7}$$

where $\mathcal{B}^{-1}(\ell) = \{\pi | \mathcal{B}(\pi) = \ell\}$ is the inverse operation of $\mathcal{B}$, representing all possible alignments corresponding to the sign gloss sequence ℓ. Finally, the CTC loss is formulated as:

$$\mathcal{L}_{CTC} = -ln(p(\ell|o)) \tag{8}$$

4 Experiments

This section describes the experiments conducted to evaluate the effectiveness of the proposed CSLR method based on ST-GAT. The datasets and evaluation indicators are described first. Then, the implementation details are introduced, and the experiments on the two datasets are discussed.

4.1 Datasets and Evaluation

Experiments were performed on two open SLR datasets, namely the Chinese Sign Language dataset (CSL) [32] and the ChaLearn LAP Continuous Gesture Dataset (ConGD) [33].

4.1.1 CSL

The CSL dataset was collected by researchers from the University of Science and Technology of China. The dataset contains 100 sign language sentences related to daily life. There are 178 Chinese words in total. Overall, 50 signers were recorded, each performing all the sentences five times. Thus, there are 25,000 videos with 100+ h in this dataset. The videos are about 10−14 s long, with a resolution of 1280 × 720 and a frame rate of 30 fps.

This work utilizes Word Error Rate (WER) as a measure of similarity between two sentences to evaluate the performance of the proposed method. In the identified sequence, several words are inserted, substituted, or deleted to improve the consistency with the correct sequence. The percentage of the total number of words inserted, substituted, or deleted divided by the total number of words in the standard sequence is WER, i.e.,

$$\text{WER} = \frac{S + D + I}{L} \tag{9}$$

where L is the total number of words in the standard sequence, and I, D, and S are the total number of insertions, deletions, and substitutions, respectively. The smaller the WER, the better the recognition performance.

4.1.2 ConGD

ConGD is a dynamic gesture dataset containing continuous and isolated word gestures within a complex real-life context. The dataset has 249 gesture labels performed by 21 signers, yielding 22,535 RGB+D videos. Each video represents one or more gestures.

Building on the ChaLearn LAP 2017 challenge [34], this work uses the Mean Jaccard Index (MJI). MJI is based on the Jaccard coefficient. The Jaccard index for gesture category i in sequence s is defined as:

$$J_{s,i} = \frac{A_{s,i} \cap B_{s,i}}{A_{s,i} \cup B_{s,i}} \tag{10}$$

where $A_{s,i}$ and $B_{s,i}$ represent the ground truth and predicted label for gesture category i in sequence s. Thus, $J_{s,i}$ can be seen as the overlap rate between $A_{s,i}$ and $B_{s,i}$.

Let L denotes the number of gesture categories. Then, the Jaccard index J_s for sequences with l_s true labels is calculated as:

$$J_s = \frac{1}{l_s} \sum_{i=1}^{L} J_{s,i} \tag{11}$$

Finally, for all sequences $S = s_1, \ldots, s_n$ with n gestures, the MJI (denoted $\overline{J_S}$) is used as the evaluation criterion (the higher, the better). It is computed as:

$$\overline{J_S} = \frac{1}{n} \sum_{j=1}^{n} J_{s_j} \tag{12}$$

4.2 Implementation Details

The experiment first uses batch normalization on the skeleton data. The ST-GAT module is composed of nine ST-GAT layers, where the first three layers, the middle three layers, and the following two layers have 64, 128, and 256 output channels, respectively. The last layer has 512 output channels. The size of the temporal convolution kernel is nine, and the stride of the second, fourth, sixth, and eighth convolution layers is set to two as the pooling layer. Residual connections are applied to each ST-GAT layer. Further, a random dropout ratio of 0.4 is used for each ST-GAT layer to avoid overfitting. In addition, the multi-head attention mechanism is used to stabilize the learning process. Then, the extracted spatial-temporal features are forwarded to the sequence learning module, and the long-term dependence information is learned via BLSTM with hidden layers of size 256. Finally, the BLSTM's output is fed into CTC for input and output alignment to obtain the final prediction sentence. Within this work, the RMSProp optimizer is used with a learning rate of 1×10^{-4} and the batch size is set to four.

In the initial implementation stage, we used all 65 joints extracted by OpenPose as input, the experiment runed slowly and occupied a high memory. Therefore, we reduced the unnecessary joints afterwards using the key 27 joints and solved the problem of high memory and slow speed. In addition, only using ST-GAT module to extract spatial-temporal features will affect the recognition accuracy, because TCN focuses on short-term temporal relationships and cannot capture long-term dependencies in sign language videos. Therefore, the proposed method added BLSTM layer after ST-GAT to deeply explore the features of sign language videos, leading to a higher recognition performance.

4.3 Experimental Results on CSL

Table 1 compares the proposed method (abbreviated as ST-GAT-SL) with different recognition methods using CSL. ST-GAT-SL without the multi-head attention mechanism achieves a WER of 1.82%. Utilizing the multi-head attention mechanism in ST-GAL-SL (heads = 2) yields a WER of 1.59%. Compared with LS-HAN and STMC using multi-cue, the proposed skeleton-based recognition method reduces WER by 15.71% and 0.51%, respectively. In addition, compared with the recent works SLRGAN and VAC, ST-GAT-SL reduces WER by 0.51% and 0.01%, respectively. In Table 1, the state-of-the-art results such as VAC uses Visual Alignment Constraint to enhance feature extraction. Compared with VAC, ST-GAT-SL uses the filtered skeleton data as input, so that the model only focuses on the data that is intrinsic for sign language recognition, and removes the interference of redundant background. Since motion-information is extremely important for videos, ST-GAT-SL uses TCN to learn the short-term temporal correlation and BLSTM to learn long-term dependencies of frames for exact sequence learning.

Table 1: Method comparison on CSL (the lower, the better)

Year	Method	WER (%)
2018	LS-HAN [32]	17.3
2018	CTF [35]	11.2
2019	DenseTCN [36]	14.3
2019	Align-iOpt [37]	6.1
2019	DPD [38]	4.7
2020	STMC [16]	2.1
2021	SLRGAN [39]	2.1
2021	VAC [20]	1.6
2022	**ST-GAT-SL (ours, heads = 1)**	**1.82**
2022	**ST-GAT-SL (ours, heads = 2)**	**1.59**

The ST-GAT-SL's performance was analyzed on 100 categories of continuous sign language videos on CSL dataset. Fig. 5 and Table 2 show the result with and without the multi-head attention mechanism. As shown in Table 2, since the extracted joints are filtered and the kernel of GCN is lightweight, the parameters and complexity of ST-GAT-SL on the CSL dataset is reasonable in both memory and computational time. It can be seen from Fig. 5 that after 20 epochs of training, the Loss value of the multi-head attention mechanism model with heads = 2 tends to be stable. Overall, using a multi-head attention mechanism will increase the complexity of the model, but it is still lightweight, and the Loss and WER values are more stable, and the WER value is lower which means more precise recognition result.

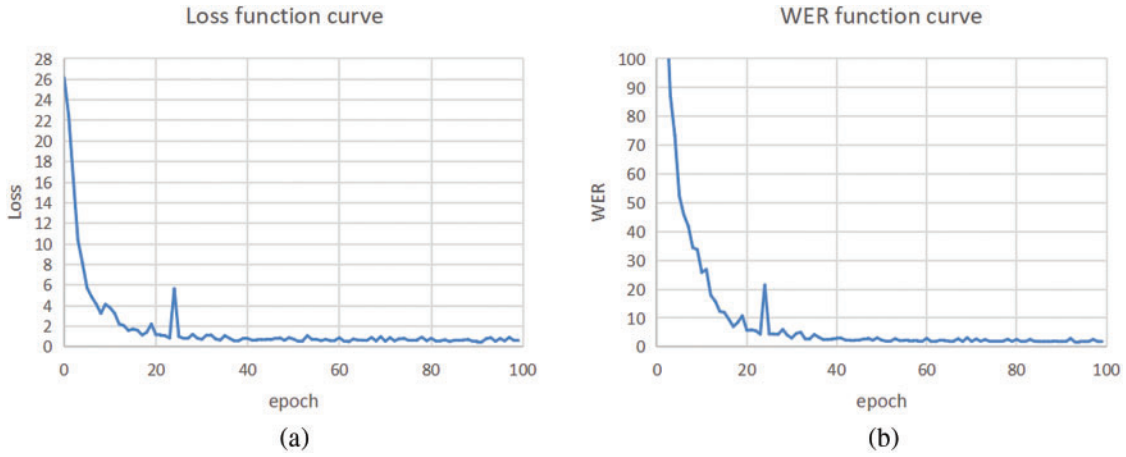


Figure 5: (Continued)

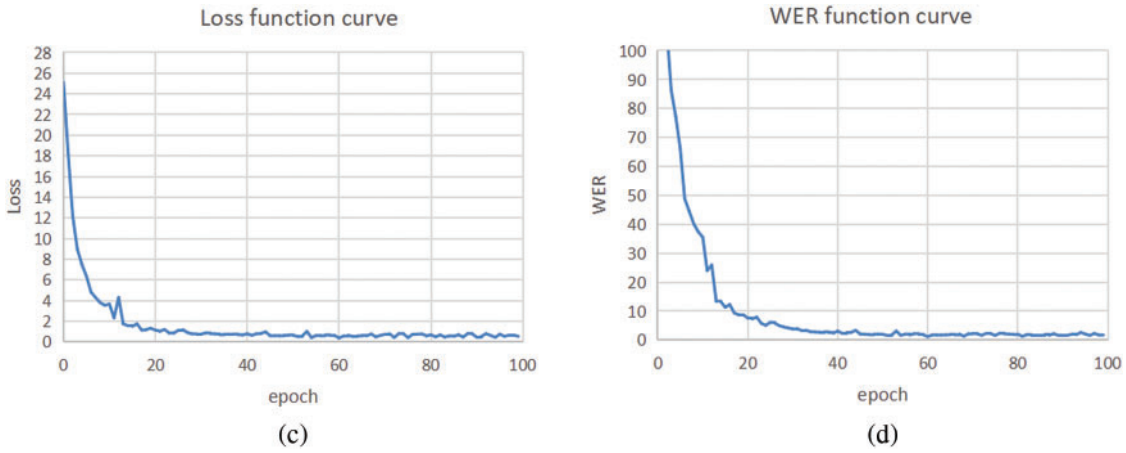
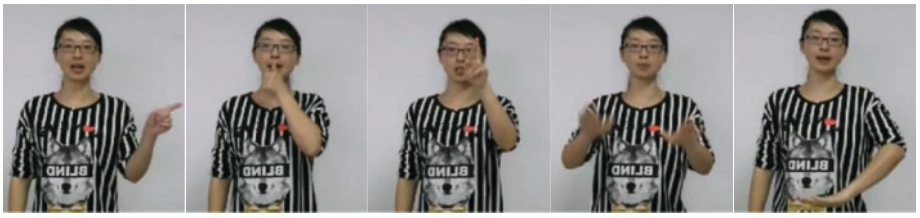


Figure 5: Loss and WER curves for CSL. Subfigures (a) and (b) show the result of the experiment without the multi-head attention mechanism, whereas (c) and (d) depict the results when the multi-head attention mechanism (heads = 2) was utilized

Table 2: The comparison of parameters and computational complexity, 'M' represents million, 'G' denotes gillion (thousand million)

Method	WER (%)	Parameters (M)	Complexity (G)
ST-GAT-SL (ours, heads = 1)	1.82	7.58 M	10.72GMac
ST-GAT-SL (ours, heads = 2)	1.59	12.0 M	18.34GMac

The ST-GAT-SL's performance regarding spatial-temporal modeling and recognition was evaluated using pairs of sign language videos with similar CSL actions, as shown in Fig. 6.



(a) The video frames showing a person signing "His father is a security guard."

(b) The video frames showing a person signing "Your father is an editor."

Figure 6: Sample frames from two CSL videos

In Fig. 6, the last two sample frames (representing the two occupations, namely "security guard" and "editor") have very similar hand movements. Nevertheless, the proposed method successfully distinguishes the cases, generating accurate recognition results. Meanwhile, the same semantic behaviors, such as that in second and third sample frames in both (a) and (b), are classified as the same words.

Table 3: Method comparison on ConGD (the higher, the better)

Year	Method	Accuracy(%)
2017	Chai et al. (2S-RNN) [40]	26.55
2017	Pigou et al. (Res-Block, BLSTM) [41]	31.90
2017	Wang et al. (ConvNets, 3D ConvLSTM) [42]	59.57
2018	Zhu et al. (TD-Res3D + Average fusion) [43]	71.63
2019	Hoang et al. (M-3DCNN-LSTM) [44]	55.23
2020	Mahmoud et al. (Im+DeepSig.) [45]	50.11
2021	Wang et al. (3D CNN, convLSTM, SPP) [46]	69.04
2022	ST-GAT-SL (ours, heads = 1)	63.72
2022	ST-GAT-SL (ours, heads = 2)	65.78

4.4 Experimental Results on ConGD

The ConGD dataset was selected for the experiments to verify the ST-GAT-SL's effectiveness on videos with complex backgrounds.

The comparison between our method and the state-of-the-art work on ConGD is shown in Table 3. As seen in the table, ST-GAT-SL achieves higher accuracy than that reported in most of the work but lower than in [43,46]. This is because Zhu et al. [43,46] adopted complicated temporal segmentation algorithms. In the segmentation module, Zhu et al. [43] used RGB and depth data as input and Wang [46] used hand motion information to divide video frames into gesture frames and transition frames. In the recognition module, Zhu et al. [43] used RGB, depth, and optical flow information as input, whereas Wang [46] used only RGB and depth information. Although these two methods have higher accuracy, they require multiple data inputs and need large-capacity memory support. In comparison, our method achieved close to the best performance in a light weight way, using only skeleton data as input from RGB without requiring temporal segmentation. The network model is simple with strong robustness to datasets with complex backgrounds. To further emphasize this point, several frame samples were manually selected from ConGD, yielding the skeleton extraction results shown in Fig. 7.

The frame rate of sign language videos in ConGD is about 10 fps. The lower the frame rate, the blurrier the picture. As a result, the recognition accuracy for this benchmark dataset is still insufficient. Figs. 7a–7c show three kinds of scene complexity: video frame blurring, flexible interaction between the two hands, and various interactions between hands and face. All these situations hamper exact recognition. Fig. 7a shows that the proposed method successfully extracted bones and joints even though the picture is blurry. Furthermore, Fig. 7b demonstrates that hand skeleton information is correctly captured when there is an interaction between the left and right hands. A human pose can be correctly estimated when there is an occlusion or interaction between hands and face, as shown in Fig. 7c. Based on the satisfactory feature extraction and graph modeling, one can conclude that the proposed method achieves competitive SLR results in real-life, daily contexts.

(a) (b) (c)

Figure 7: Sample of extracted skeleton graphs

5 Conclusion

This work proposes a CSLR method based on ST-GAT. In order to eliminate the influence of complex background, the pose estimation algorithm OpenPose was used to preprocess the video to extract the information about joints and bones. On this basis, the spatial-temporal skeleton graph is constructed. The proposed ST-GAT module calculates the importance of different joints and builds a higher spatial and temporal feature map with multi-head attention mechanism. We use BLSTM and CTC to learn the bidirectional long-term series dependence and align the sequences for final recognition. Our proposed method achieves a WER of 1.59% on the CSL dataset and MJI of 65.78% on the ConGD dataset, which demonstrates the ST-GAT-SL's effectiveness. This work proves the feasibility of skeleton data in SLR tasks. So far, the method is not trained in an end-to-end manner, and we need to preprocess the videos to get the joints and skeleton information. Future studies will focus on end-to-end sign language translation methods and multi-channel input including RGB videos, optical-flow and skeleton data for better recognition performance.

Funding Statement: This work was supported by the Key Research & Development Plan Project of Shandong Province, China (No. 2017GGX10127).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Jiang, X., Satapathy, S. C., Yang, L., Wang, S. H., Zhang, Y. D. (2020). A survey on artificial intelligence in Chinese sign language recognition. *Arabian Journal for Science and Engineering, 45(12),* 9859–9894. DOI 10.1007/s13369-020-04758-2.
2. Zhang, S., Meng, W., Li, H., Cui, X. (2019). Multimodal spatiotemporal networks for sign language recognition. *IEEE Access, 7,* 180270–180280. DOI 10.1109/ACCESS.2019.2959206.
3. Huang, J., Zhou, W., Li, H., Li, W. (2018). Attention-based 3D-CNNs for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 29(9),* 2822–2832. DOI 10.1109/TCSVT.2018.2870740.
4. Hu, H., Zhou, W., Pu, J., Li, H. (2021). Global-local enhancement network for NMF-aware sign language recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications, 17(3),* 1–19. DOI 10.1145/3436754.
5. Chao, H., Wang, F. H., Ran, Z. (2019). Sign language recognition based on cbam-resnet. *Proceedings of the 2019 International Conference on ArtificialIntelligence and Advanced Manufacturing*, pp. 1–6. Dublin Ireland. DOI 10.1145/3358331.3358379.
6. Pigou, L., Dieleman, S., Kindermans, P. J., Schrauwen, B. (2014). Sign language recognition using convolutional neural networks. *European Conference on Computer Vision*, pp. 572–578. Cham: Springer. DOI 10.1007/978-3-319-16178-5_40.
7. Koller, O., Bowden, R., Ney, H. (2016). *Automatic alignment of hamnosys subunits for continuous sign language recognition*, pp. 121–128. University of Surrey.
8. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S. et al. (2016). Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA. DOI 10.1109/CVPR.2016.456.
9. Qiu, Z., Yao, T., Mei, T. (2017). Learning spatio-temporal representation with pseudo-3D residual networks. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5533–5541. Venice, Italy. DOI 10.1109/ICCV.2017.590.
10. Wu, D., Pigou, L., Kindermans, P. J., Le, N. D. H., Shao, L. et al. (2016). Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8),* 1583–1597. DOI 10.1109/TPAMI.2016.2537340.
11. Cui, R., Liu, H., Zhang, C. (2019). A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia, 21(7),* 1880–1891. DOI 10.1109/TMM.2018.28-89563.
12. Zhang, J., Zhou, W., Li, H. (2014). A threshold-based HMM-DTW approach for continuous sign language recognition. *Proceedings of International Conference on Internet Multimedia Computing and Service*, pp. 237–240. Xiamen, China. DOI 10.1145/2632856.2632931.
13. Xiao, Q., Chang, X., Zhang, X., Liu, X. (2020). Multi-information spatial–temporal LSTM fusion continuous sign language neural machine translation. *IEEE Access, 8,* 216718–216728. DOI 10.1109/A-CCESS.2020.3039539.
14. Zhang, Q., Wang, D., Zhao, R., Yu, Y. (2019). MyoSign: Enabling end-to-end sign language recognition with wearables. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 650–660. Marina del Ray, CA, USA. DOI 10.1145/3301275.3302296.
15. Camgoz, N. C., Koller, O., Hadfield, S., Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10023–10033. Seattle, WA, USA. DOI 10.1109/CVPR42600.2020.01004.
16. Zhou, H., Zhou, W., Zhou, Y., Li, H. (2020). Spatial-temporal multi-cue network for continuous sign language recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 13009–13016. New York, NY, USA. DOI 10.1609/aaai.v34i07.7001.

17. Zhou, M., Ng, M., Cai, Z., Cheung, K. C. (2020). Self-attention-based fully-inception networks for continuous sign language recognition. *24th European Conference on Artificial Intelligence*, pp. 2832–2839. IOS Press. DOI 10.3233/FAIA200-425.
18. Min, Y., Hao, A., Chai, X., Chen, X. (2021). Visual alignment constraint for continuous sign language recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11542–11551. Montreal, QC, Canada. https://arxiv.org/abs/2104.02330.
19. Cheng, K. L., Yang, Z., Chen, Q., Tai, Y. W. (2020). Fully convolutional networks for continuous sign language recognition. *European Conference on Computer Vision*, pp. 697–714. Springer, Cham. DOI 10.1007/978-3-030-58586-0_41.
20. Niu, Z., Mak, B. (2020). Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. *European Conference on Computer Vision*, pp. 172–186. Springer, Cham. DOI 10.1007/978-3-030-58517-4_11.
21. Xiao, Q., Qin, M., Guo, P., Zhao, Y. (2019). Multimodal fusion based on LSTM and a couple conditional hidden markov model for Chinese sign language recognition. *IEEE Access, 7,* 112258–112268. DOI 10.1109/ACCESS.2019.2925654.
22. Islam, M. (2020). An efficient human computer interaction through hand gesture using deep convolutional neural network. *SN Computer Science, 1(4),* 1–9. DOI 10.1007/s42979-020-00223-x.
23. Yan, S., Xiong, Y., Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA.
24. Huang, Q., Zhou, F., Qin, R. (2021). View transform graph attention recurrent networks for skeleton-based action recognition. *Signal, Image and Video Processing, 15(3),* 599–606. DOI 10.1007/11760-020-01781-6.
25. Shi, L., Zhang, Y., Cheng, J., Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035. Long Beach, CA, USA. DOI 10.1109/CVPR.2019.01230.
26. Yao, L., Mao, C., Luo, Y. (2019). Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 7370–7377. Honolulu, Hawaii, USA. DOI 10.1609/aaai.v33i01.33017370.
27. Yu, B., Yin, H., Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875.
28. Sun, B., Zhao, D., Shi, X., He, Y. (2021). Modeling global spatial–Temporal graph attention network for traffic prediction. *IEEE Access, 9,* 8581–8594. DOI 10.1109/ACCESS.2021.3049556.
29. Zhang, Y. D., Satapathy, S. C., Guttery, D. S., Górriz, J. M., Wang, S. H. (2021). Improved breast cancer classification through combining graph convolutional network and convolutional neural network. *Information Processing & Management, 58(2),* 102439. DOI 10.1016/j.ipm.2020.102439.
30. Cao, Z., Simon, T., Wei, S. E., Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299. Honolulu, HI, USA. DOI 10.1109/TPAMI.2019.2929257.
31. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376. Pittsburgh, Pennsylvania, USA. DOI 10.1145/1143844.1143891.
32. Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W. (2018). Video-based sign language recognition without temporal segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1. New Orleans, LA, USA.
33. Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S. et al. (2016). Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 56–64. Las Vegas, NV, USA. DOI 10.1109/CVPRW.2016.100.

34. Wan, J., Escalera, S., Anbarjafari, G., Jair Escalante, H., Baró, X. et al. (2017). Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 3189–3197. Venice, Italy. DOI 10.1109/ICCVW.2017.377.
35. Wang, S., Guo, D., Zhou, W. G., Zha, Z. J., Wang, M. (2018). Connectionist temporal fusion for sign language translation. *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 1483–1491. Korea. DOI 10.1145/3240508.3240671.
36. Guo, D., Wang, S., Tian, Q., Wang, M. (2019). Dense temporal convolution network for sign language translation. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 744–750. Macao, China. DOI 10.24963/ijcai.2019/105.
37. Pu, J., Zhou, W., Li, H. (2019). Iterative alignment network for continuous sign language recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4165–4174. Long Beach, CA, USA. DOI 10.1109/CVPR.2019.00429.
38. Zhou, H., Zhou, W., Li, H. (2019). Dynamic pseudo label decoding for continuous sign language recognition. *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1282–1287. Shanghai, China. DOI 10.1109/ICME.2019.00223.
39. Papastratis, I., Dimitropoulos, K., Daras, P. (2021). Continuous sign language recognition through a context-aware generative adversarial network. *Sensors, 21(7),* 2437. DOI 10.3390/s21072437.
40. Chai, X., Liu, Z., Yin, F., Liu, Z., Chen, X. (2016). Two streams recurrent neural networks for large-scale continuous gesture recognition. *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 31–36. Cancún, Mexico. DOI 10.1109/ICPR.2016.7899603.
41. Pigou, L., van Herreweghe, M., Dambre, J. (2017). Gesture and sign language recognition with temporal residual networks. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 3086–3093. Venice, Italy. DOI 10.1109/ICCVW.2017.365.
42. Wang, H., Wang, P., Song, Z., Li, W. (2017). Large-scale multimodal gesture recognition using heterogeneous networks. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 3129–3137. Venice, Italy. DOI 10.1109/ICCVW.2017.370.
43. Zhu, G., Zhang, L., Shen, P., Song, J., Shah, S. A. A. et al. (2018). Continuous gesture segmentation and recognition using 3DCNN and convolutional LSTM. *IEEE Transactions on Multimedia, 21(4),* 1011–1021. DOI 10.1109/TMM.2018.2869278.
44. Hoang, N. N., Lee, G. S., Kim, S. H., Yang, H. J. (2019). Continuous hand gesture spotting and classification using 3D finger joints information. *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 539–543. Taipei, Taiwan. DOI 10.1109/ICIP.2019.8803813.
45. Mahmoud, R., Belgacem, S., Omri, M. N. (2020). Deep signature-based isolated and large scale continuous gesture recognition approach. *Journal of King Saud University-Computer and Information Sciences, 34(5),* 1793–1807. DOI 10.1016/j.jksuci.2020.08.017.
46. Wang, H. (2021). Two stage continuous gesture recognition based on deep learning. *Electronics, 10(5),* 534. DOI 10.3390/electronics10050534.