A Local Spatial–Temporal Synchronous Network to Dynamic Gesture Recognition

Dongdong Zhao, Qinglian Yang, Xingwen Zhou, Hongli Li, and Shi Yan

Abstract—**Dynamic skeletal data contain advanced information represented by complex compound actions and have been extensively studied in human motion recognition. Previous works usually depend on designing the traversal rules made by hand or using two separate components to obtain spatial and temporal information based on skeleton graph, respectively, which still have generalization difficulties and significant challenges in learning effective spatial–temporal information. This article proposes a local spatial–temporal synchronous network (LSTSN) for skeleton-based dynamic gestures recognition, which can simultaneously deal with spatial and temporal information in gestures. Specifically, to fully obtain the spatial–temporal information of the skeleton data, a coupled position embedding and a three-neighbor local spatial–temporal graph (LSTG) are constructed. An adaptive threshold method is designed to describe the joint connections of LSTG, and then, an attention mechanism is used to efficiently capture complex local spatial–temporal correlations. Meanwhile, multiple spatial–temporal attention modules of different periods are designed to effectively capture the LSTG's heterogeneity. Besides, from the data aspect, a hand motion vector (HMV) is also introduced into the model to express subtle finger motions. Extensive experiment results on DHG-14/28 and SHREC'17 Track datasets demonstrate that the proposed LSTSN is competitive with state-of-the-art methods.**

Index Terms—**Attention, gesture recognition, hand motion vector (HMV), spatial–temporal graph.**

I. Introduction

DYNAMIC gesture recognition has been an active research field during the past few decades due to the potential value and extensive applications in the video game industry [1], food industry, television control [2], and machinery industry [3].

With the emergence of new depth sensors [4] (such as leap motion controller (LMC) [5]) and the significant progress in the accurate gesture pose estimation [7]–[9], a new upsurge of skeletal data research [10], [11] in the field of dynamic gesture recognition has also been set off. In the beginning, skeleton-based gestures recognition methods mainly rely on designing powerful feature descriptors to simulate hand movements [12]–[14]. However, these handmade features are not able to describe high semantic information and have limited generalization ability [15]. Inspired by the great promise of deep learning [16], [17], significant improvements have been made for dynamic gesture recognition. However, the spatial structure and temporal dynamics for hand skeletons are not effectively utilized in these deep learning methods.

It is well known that the spatial structure, temporal dynamics, and connectivity of joints play an essential role in dynamic gesture recognition, and the skeleton graph-based method has been applied to dynamic gestures [15], [18], [19]. However, all of these methods use two separate components to capture spatial and temporal dependencies, respectively, and feed the spatial representation into the module of time dimension modeling to capture the spatial–temporal correlation indirectly. In particular, to some extent, the connection characteristics of each joint in different time steps obscure the spatial–temporal correlations. In addition, they analyze the spatial–temporal relationship among joints of all frame; i.e., for a joint, the two modules consider the influence of all the other joints in the whole time series, which may weaken the influence of some important adjacent joints to a certain extent. Moreover, the heterogeneity in dynamic gesture data is also not considered. All these abovementioned problems may affect the extraction of joint features and reduce the recognition performance.

To solve the abovementioned problems, a local spatial–temporal synchronous network (LSTSN) for dynamic gesture recognition is proposed in this article, which can capture the spatial–temporal correlation and heterogeneity in skeletal data simultaneously. Specifically, LSTSN first decouples the spatial and temporal dimensions; then, the sine and cosine functions are used to encoding them sequentially, respectively. After that, the corresponding embedding matrices are generated, respectively. By adding these two matrices at the same time, the spatial–temporal dimension attributes of skeletal joints can be distinguished, which can help obtain the spatial and temporal information in dynamic gestures simultaneously. Besides, to highlight the interaction of the main joints, the proposed model mainly focuses on the relationship between joints of adjacent frame in terms of time, which is realized by constructing a local spatial–temporal graph (LSTG) connecting the spatial graphs of adjacent time steps through a sliding window. In a spatial graph, we first define the connection relation of skeletal joints according to the motion law of gesture and prove through experiments that it is necessary to highlight the connection relation of important joints. Then, to clarify the influence

Manuscript received 2 April 2022; revised 2 June 2022; accepted 10 June 2022. This work was supported in part by the Natural Science Foundation of Gansu Province, China, under Grant 20JR10RA638; and in part by the Fundamental Research Funds for the Central Universities under Grant lzujbky-2021-67 and Grant lzuxxxy-2019-tm17. *(Corresponding author: Shi Yan.)*

The authors are with the School of Information Science and Engineering, Lanzhou University, Gansu 730000, China (e-mail: ddzhao@lzu.edu.cn; yangql19@lzu.edu.cn; zhouxw20@lzu.edu.cn; lihl20@lzu.edu.cn; yanshi@lzu.edu.cn).

Digital Object Identifier 10.1109/TCSS.2022.3184420

of primary joints, we propose an adaptive threshold method to adjust the connection relationship of joints dynamically. After that, the attention mechanism is used to capture the complex local spatial–temporal correlations in the LSTGs, while multiple separate local spatial–temporal synchronous attention modules (LSTSAMs) are deployed in different time slices to capture the heterogeneity of spatial–temporal network data, after which each joint already contains local spatial–temporal correlation. Considering the subtle finger motion, the relative position vector of fingertips is also introduced to describe the finger motion accurately. The extensive experiments on two standard benchmarks, DHG-14/28 dataset and SHREC'17 Track dataset, demonstrate the superior performance of the proposed LSTSN representation to the current state-of-the-art approaches.

In conclusion, the main contributions of the proposed method can be summarized as follows.

1) A novel spatial–temporal position embedding (STE) is proposed to make the synchronous acquisition of spatial–temporal information possible.
2) An LSTG is constructed, and the spatial joints connection relationship is defined by an adaptive method to highlight the interaction of main joints.
3) Multiple separate spatial–temporal synchronous attention modules are deployed in different time slices to capture the heterogeneity in long-range spatial–temporal graphs. Moreover, considering the subtle finger motion, the relative position vector of fingertips is introduced into the model.

The remainder of this article is organized as follows. Some Related works are presented in Section II. In Section III, an LSTSN and the corresponding implementation process are developed. In Section IV, extensive experiment results on DHG-14/28 and SHREC'17 Track datasets are given to demonstrate the effectiveness of the proposed method. The conclusions are given in Section V.

II. Related Work

Self-attention mechanism is a variant of the attention mechanism, which mainly depends on internal information, so it is easier to capture the interaction between the interior features of data sequences. This article applies multiple independent self-attention modules to the local skeletal spatial–temporal graph of three adjacent frames to learn the spatial–temporal information contained in the map and capture additional node identities.

Skeleton-Based Hand Gesture Recognition: At present, the mainstream skeleton-based methods can be divided into two categories: First, method based on deep learning. Such methods usually connect the joint coordinates into a tensor; then, by feeding it into a neural network, the hand features can be directly learned, such as the gesture recognition system composed of LMC and bidirectional recurrent neural networks (BRNN) [20]. It points out that the output of the current unit is related to both the previous units and future units, which motivates us that the spatial–temporal information between adjacent frame joints might be the essential feature. Second, method based on skeleton graph. The skeleton graph is constructed according to the skeletal data, and then, the dependence of its joints is modeled, such as Dynamic Graph-Based Spatial-Temporal Attention (DG-STA) [21]. However, modeling spatial and temporal correlation and heterogeneity still lacks effective methods. In this article, we focus on designing a model to synchronously capture these correlations and improve the accuracy of spatial–temporal network data prediction by considering the heterogeneity.

III. Methods

In this section, we first construct a spatial–temporal network graph according to the skeleton of gesture; then, the LSTSN is established to capture spatial–temporal correlation simultaneously rather than using two modules to model spatial dependence and temporal correlation, respectively. As shown in Fig. 1, the network mainly has three core ideas: First, a new STE is proposed, in which the spatial and temporal information is embedded synchronously. Then, an LSTG is constructed, based on which the local spatial–temporal correlation is captured by the LSTSAM designed in this article. Finally, several LSTSAMs are deployed to model the heterogeneity of spatial–temporal network sequences.

A. Spatial–Temporal Network

Take the 22 skeletal joints of the hand shown in Fig. 2(a) as an example. Three kinds of dependency relationships can be defined to construct a spatial–temporal network. The first type of dependency is the actual spatial dependency based on spatial relationship, i.e., the relative position relationship between adjacent joints along the human hand skeleton in the same frame. The second type is based on the temporal correlation of the same joint in the time series, i.e., the difference of joints in different frames. The third type is the dependency between a joint and its neighbor joints in different frames. In contrast, the existing methods, such as DG-STA [21], Spatial Temporal Graph Convolutional Networks (ST-GCN) [18], and Hand Gesture Graph Convolutional Networks (HG-GCN) [15], usually use two separate components to capture the spatial and temporal dependencies, respectively. However, if these complex local spatial–temporal correlations can be captured at the same time, it will be very effective for spatial–temporal skeletal data forecasting, because this modeling approach reveals the primary way how spatial–temporal network data are generated.

B. Spatial–Temporal Position Embedding

Connecting joints of different time steps to the same LSTG blurs the spatial–temporal attributes of each joint. In other words, this LSTG places joints in the same environment at different time steps without distinguishing them. To make full use of the spatial and temporal information, a new STE is established before using the sliding window operation to construct the local graph, by which the spatial and temporal information of each joint can be distinguished.

Different from [21], for the input skeletal sample sequence $X \in \Re^{T \times N \times C}$, we decompose the space and position coding and sequentially encode the joints in the same frame, and the same joints in different frames are also encoded in order;

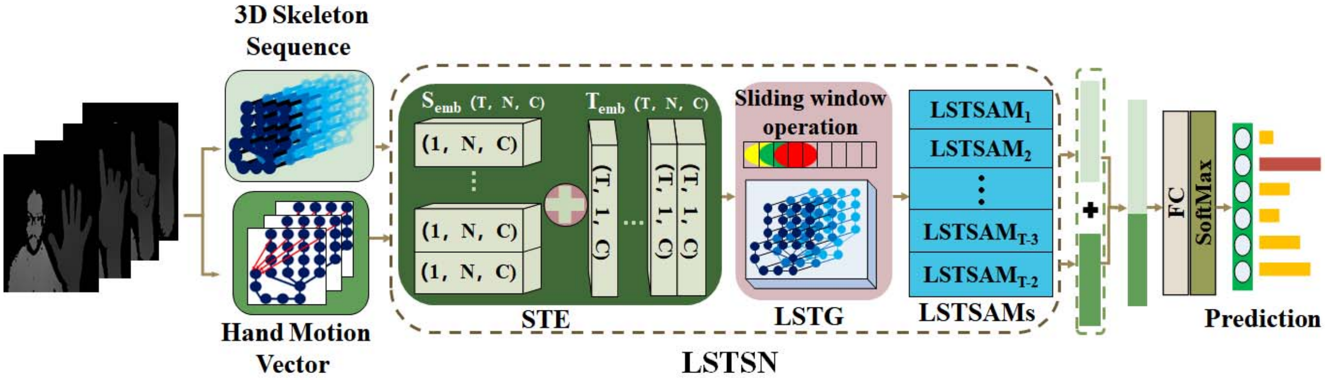
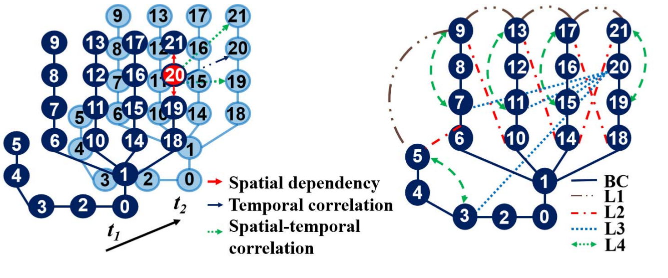
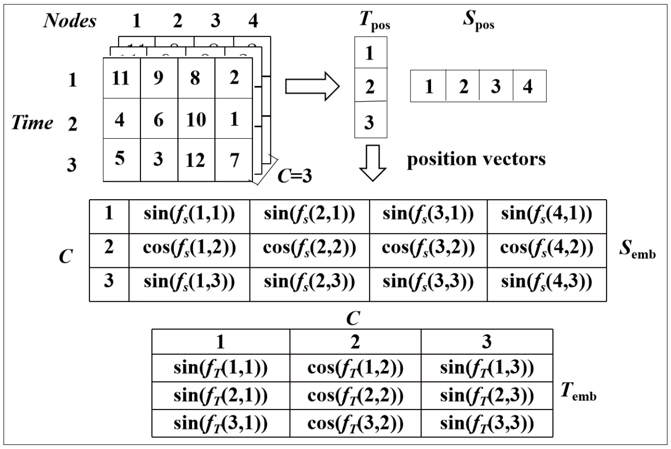


Fig. 1. Proposed network structure. The network consists of three major components, i.e., STE, LSTG, and LSTSAM.



(a) (b)

Fig. 2. (a) Three kinds of dependencies of the red joint 20 in the spatial–temporal network. The red arrows denote the influence in the spatial dimension (spatial dependency), the blue arrow indicates the temporal correlation of the joint 20, and the green arrows indicate the influence that across both the spatial and temporal dimensions: spatial–temporal correlation. (b) Connections of joints designed for testing, where BC, L1, L2, L3, and L4 represent the basic connection of hand skeleton and four newly defined edge connections, respectively.



Time	1	2	3	4
1	11	9	8	2
2	4	6	10	1
3	5	3	12	7

T_{pos}
1
2
3

S_{pos}			
1	2	3	4

C				
1	$\sin(f_s(1,1))$	$\sin(f_s(2,1))$	$\sin(f_s(3,1))$	$\sin(f_s(4,1))$
2	$\cos(f_s(1,2))$	$\cos(f_s(2,2))$	$\cos(f_s(3,2))$	$\cos(f_s(4,2))$
3	$\sin(f_s(1,3))$	$\sin(f_s(2,3))$	$\sin(f_s(3,3))$	$\sin(f_s(4,3))$

1	2	3
$\sin(f_T(1,1))$	$\cos(f_T(1,2))$	$\sin(f_T(1,3))$
$\sin(f_T(2,1))$	$\cos(f_T(2,2))$	$\sin(f_T(2,3))$
$\sin(f_T(3,1))$	$\cos(f_T(3,2))$	$\sin(f_T(3,3))$

Fig. 3. STE example with four 1-D joints and three frames.

i.e., the spatial position vector and temporal position vector are set to $S_{\text{pos}} \in [1, 2, \ldots, N]$ and $T_{\text{pos}} \in [1, 2, \ldots, T]$, respectively. Then, the temporal embedding matrix $T_{\text{emb}} \in \Re^{T \times C}$ and a spatial embedding matrix $S_{\text{emb}} \in \Re^{N \times C}$ can be set using sine and cosine functions of different frequencies, respectively. The calculation formula is as follows:

$$f(\text{pos}, i) = \frac{\text{pos}}{10\,000^{\frac{2i}{C}}}, \quad i = 1, 2, \ldots, C \tag{1}$$

$$\text{PE}(\text{pos}, i) = \begin{cases} \sin(f(\text{pos}, i)), & \text{if } i = 2j \\ \cos(f(\text{pos}, i)), & \text{if } i = 2j + 1. \end{cases} \tag{2}$$

Taking three frames with four joints in each frame as an example, the coding process is shown in Fig. 3.

The sizes of the two tensor matrices T_{emb} and S_{emb} are $\Re^{T \times C}$ and $\Re^{N \times C}$, respectively. To meet the conditions of the tensor addition rule, i.e., all tensors must be of the same order and type, tensor expansion is used to expand their dimensions, and finally, two tensor matrices, having the same size as $X \in \Re^{T \times N \times C}$, are obtained. Thus, for an input skeletal sample sequence X, we have

$$X_{\text{emb}} = X + T_{\text{emb}} + S_{\text{emb}} \in \Re^{T \times N \times C} \tag{3}$$

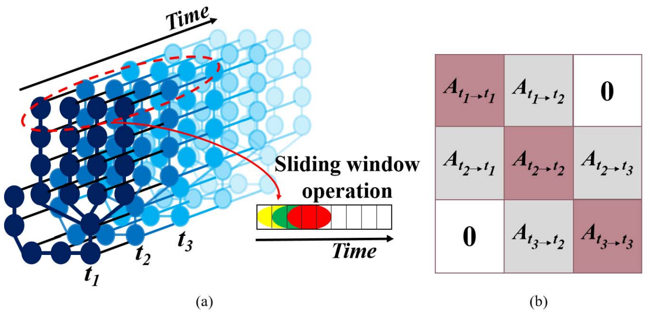
where the X_{emb} is referred to as the position embedded sequence of skeleton. By adding these two embedding matrices, the spatial and temporal information of each joint can be distinguished in the LSTG.

C. Local Spatial–Temporal Graph

To directly capture the influence of each joint on its neighbors that belongs to both the current and the adjacent time steps, we construct the LSTG consisting of three spatial graphs, as shown in Fig. 4(a) (t_1, t_2, t_3). Each time step has a spatial graph, which is constructed by the time-invariant spatial relationship between joints. By connecting each joint of adjacent time steps with itself, the space graph can be connected in terms of time to form an LSTG. Therefore, in an LSTG, the correlation of each joint with its neighbors in the established graph can be captured directly. It is noted that these neighbors of each joint include not only their neighbors in the same time step but also their own joints in the two adjacent time steps, which will have a certain correlation to each joint as its second-order neighbors.

To construct the LSTGs, sliding window with width 3 and sliding step 1 is used to clip the skeletal position embedded sequence $X_{\text{emb}} \in \Re^{T \times N \times C}$, and then, $T - 2$ time series slices corresponding to $T - 2$ LSTGs are obtained. The clipped sequence is expressed as $X_{\text{clip}} = [X_1, X_2, \ldots, X_{T-2}] \in \Re^{(T-2) \times 3N \times C}$, where $X_t \in \Re^{3N \times C}$, $t = 1, 2, \ldots, T - 2$.

For the local sample sequence X_t, its LSTG adjacency matrix is $A \in \Re^{3N \times 3N}$, and the adjacency matrix of the spatial



$A_{t_1\to t_1}$	$A_{t_1\to t_2}$	0
$A_{t_2\to t_1}$	$A_{t_2\to t_2}$	$A_{t_2\to t_3}$
0	$A_{t_3\to t_2}$	$A_{t_3\to t_3}$

(a) (b)

Fig. 4. (a) (t_1, t_2, t_3) denotes an LSTG, which is realized by sliding window operation with window width 3 and sliding step 1. (b) Adjacency matrix of the localized spatial–temporal graph in (a).

graph $A_{t_i\to t_j} \in \Re^{N\times N}$, $0 < i, j \le 3$ is formulated as follows:

$$A_{t_i\to t_j} = \begin{cases} \text{score}(v_i, v_j), & \text{if } v_i \to v_j \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where v_i denotes the ith joint in LSTG, $v_i \to v_j$ denotes that the joint v_i connects to the joint v_j, and the calculation of $\text{score}(v_i, v_j)$ is based on multihead attention [22]. Fig. 4(b) illustrates the adjacency matrix of the LSTG.

Assuming that the input feature of the joint v_i is $f_i \in \Re^C$, the hth attention head first maps f_i to the query, key, and value vectors: $q, k, v \in \Re^d$ by applying three fully connected (FC) layers, respectively. Finally, the connection weights between joints in the LSTG are obtained by calculating the "scaled dot-product" [22] between q and k of the joints within the same time step scaling the point product, and then, the results are normalized by the following softmax function:

$$\text{s}(v_i, v_j) = \frac{\langle q, k\rangle}{\sqrt{d}} \tag{5}$$

$$\text{score}(v_i, v_j) = \text{softmax}(\text{s}(v_i, v_j)) = \frac{\exp(\text{s}(v_i, v_j))}{\sum_{k=1}^{n} \exp(\text{s}(v_i, v_k))} \tag{6}$$

where d is the dimension of q, k, v; $\langle , \rangle$ represents the inner product operation; n represents the number of joints connected to the joint v_i; $\text{score}(v_i, v_j)$ measures the weight of attention from the joint v_i and v_j.

D. Joint Connection

By observing the law of hand movement, we define the joint connections, as shown in Fig. 2(b), in which L1 can measure the change in the distance between fingertips during the movement of the hand. L2 allows the information of one finger to be spread to another and helps to encode the relationship between them, e.g., overlap or separate. L3 takes the second joint point of the little finger as the new original point and considers the connection with the third joint of each finger, which helps to measure the open degree of the hand better. Moreover, in the process of finger bending, the distal phalanx has the largest range of change, while the change of distance between the intermediate phalanx and distal phalanx is always small. The fingertip is connected with the third joint of the same finger to obtain information about some actions in gesture, as shown in Fig. 2(b) (L4). Through experiments, we prove that it is necessary to highlight the connection relation of important joints. To further remove the connection of weak relationship joints and highlight the connection function of strong relationship joints, we design an adaptive method to dynamically adjust the connection relationship of joints.

For an LSTG, it contains skeletal joints of three frame gestures. For each joint, the correlation score between it and all the other joints will be calculated, and the highest values $S_{\max}$ will be selected, and then, a threshold will be set as: $p = S_{\max}/\Theta$, where Θ is the adjustment parameter of the threshold p. When $|\text{score}| < p$, it can be considered that the joint has no connection relationship with the current joint, and the value of the current score will be set to 0. On the contrary, it is considered that these two joints have a connection relationship and retain the value of the current score.

E. Local Spatial–Temporal Synchronous Attention Module

In the proposed network structure, the LSTSAM is built to capture local spatial–temporal correlations, and the core of LSTSAM is multihead attention [22], [23], as shown in Fig. 5. To consider large-range spatial–temporal correlations, a sliding window is used to cut out LSTG at different periods. In view of the heterogeneity in the spatial–temporal data, multiple LSTSAMs are used to model different periods' LSTG, which can achieve better performance than sharing only one for all periods.

In the hth attention head, the local sample sequence $X_t \in \Re^{3N\times C}$ can finally be mapped into the query, key, and value matrices: $Q, K, V \in \Re^{3N\times D}$. Finally, the hth attention feature of $X_t \in \Re^{3N\times C}$ is obtained, which is expressed as follows:

$$\left(h^h_{\text{MDA}}\right)_t = A \cdot V \in \Re^{3N\times D}. \tag{7}$$

Cropping Operation: If we stack directly multiple LSTSAMs and retain the features of all the adjacent time steps, a large amount of redundant information will exist in the model, which will seriously affect the performance of the model. Therefore, before aggregating data, it is necessary to take the cropping operation, as shown in Fig. 5(b), to remove all the features of the joints at the previous and the next time steps, and only the joints in the current moment are retained. The multihead attention operations have already aggregated the information from the previous and next time steps. Each joint contains the localized spatial–temporal correlations even though we crop the two adjacent time steps. The cropped data are expressed as $Y^h_t \in \Re^{N\times D}$.

Connecting all attention features learned by the attention heads together, the features of the local sample sequence are obtained by

$$Y_t = \text{concate}\left(\left[Y^1_t, Y^2_t, \ldots, Y^H_t\right]\right) \in \Re^{N\times C'}, \quad C' = H * D \tag{8}$$

where H is the number of attention heads, and Y_t $(t = 1, 2, \ldots, T-2)$ is the output of the tth LSTSAM.

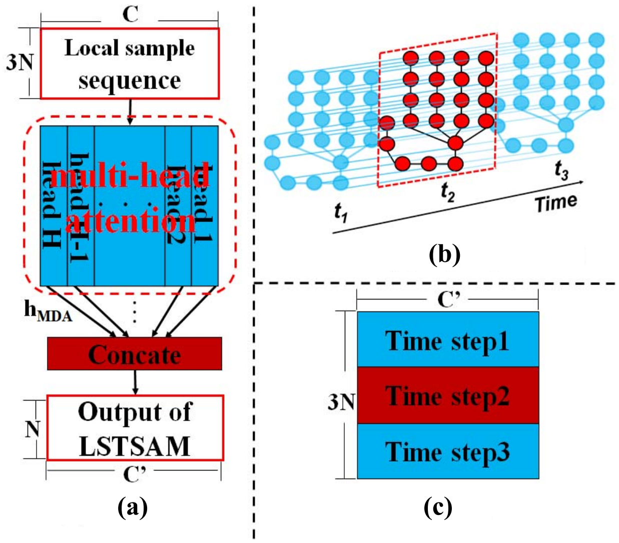


Fig. 5. (a) Example of the architecture of LSTSAM. C and C' denote the numbers of features of the input and the output, respectively; $h_{\text{MDA}} \in \Re^{3N\times C}$ denotes one attention feature. (b) Example of a cropping operation. (c) Output of the aggregating operation.

After that, the local spatial–temporal correlations in these $T-2$ local sample sequences can be captured, respectively. All these $T-2$ LSTSAMs' outputs are concatenated into one matrix $Y \in \Re^{T-2\times N\times C'}$.

Finally, we average-pool the features Y into a vector and feed it into an FC layer for classification.

F. Hand Motion Vector

The spatial–temporal configurations of joints are obtained naturally by training in the LSTSN model using 22 joints to model the hand skeleton. However, it is still difficult to express subtle finger motions. To this end, a hand motion vector (HMV) is introduced in this article, which can describe finger motion precisely. Fingertip movement is related to the other three joints on one finger. In addition, gestures always involve the interaction of the thumb and four fingers. Therefore, the small motion of a hand can be expressed as the relative positions of the four fingers and thumbs in each frame. For the HMV of the nth hand gesture sample V_n, it can be represented as follows:

$$v_t = \left(f^n_{I,t}, f^n_{M,t}, f^n_{R,t}, f^n_{L,t}\right) - \left(f^n_{o,t}, f^n_{o,t}, f^n_{o,t}, f^n_{o,t}\right) \quad (9)$$

$$V_n = \{v_t \mid t = 1, 2, \ldots, T\} \quad (10)$$

where $f^n_{o,t}$, $f^n_{I,t}$, $f^n_{M,t}$, $f^n_{R,t}$, and $f^n_{L,t}$ are the fingertip coordinates of thumb, index, middle, ring, and little at the tth frame of the nth hand gesture sample, respectively.

The overall algorithm flow is shown in Fig. 6.

IV. Experiments

In this section, exploratory experiments are conducted to evaluate the effectiveness of the proposed method. Furthermore, we perform comparisons with the representative gesture recognition methods, including Shape of Connected Joints (SoCJ) + Histogram of Hand Directions (HoHD) + Histogram of Wrist Rotations (HoWR) [24], Chen's method [16], Convolutional Neural Network (CNN) + Long Short-Term Memory (LSTM) [17], Handwriting-Inspired Features for 3D skeleton-based action recognition (HIF3D) [25], Residual Temporal Convolutional Network (Res-TCN) [19], ST-GCN [18], Spatial Temporal Attention Residual Temporal Convolutional Network (STA-Res-TCN) [19], HG-GCN [15], DG-STA [21], and Decoupled Spatial-Temporal Attention Network (DSTA-Net) [28], based on the analysis on the performance of the proposed method.

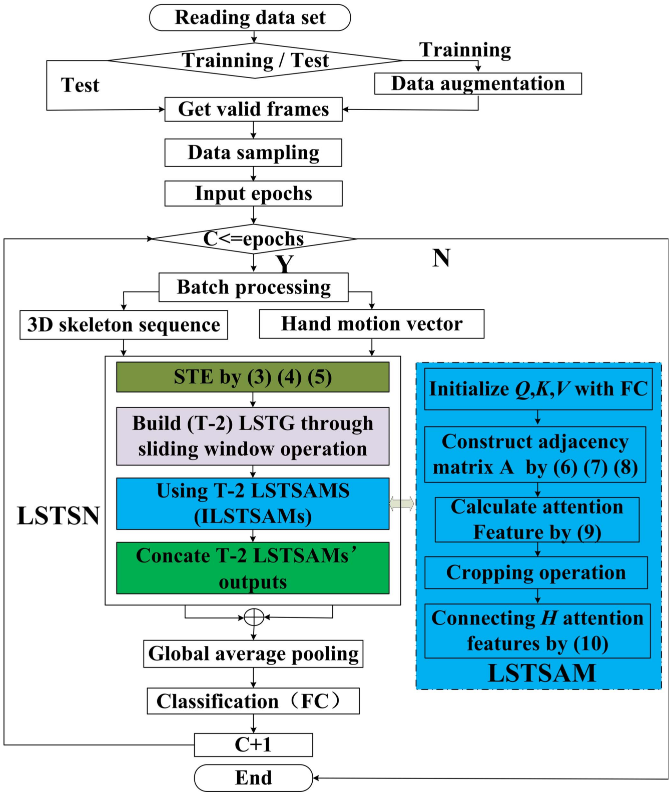


Fig. 6. Algorithm flowchart.

A. Datasets

1) DHG-14/28 [24]: This dataset includes 2800 sample sequences of 14 hand gestures. The hand gestures are executed in two ways: using one single finger, which is regarded as 14 classes, and the whole hand, which is regarded as 28 classes. Each gesture is performed five times by 20 participants in both two ways, resulting in 2800 sequences.

2) SHREC'2017 Track [26]: This dataset includes 14 gestures with 2800 sample sequences collected by 28 participants performing gestures between one and ten times. The SHREC'2017 Track dataset has the same data collection method as the DHG-14/28 dataset. Both of them contain depth images and 3-D hand skeletal data with 22 hand joints per frame.

B. Experiment Settings

Before entering the LSTSN, we need to preprocess the dataset as follows: first, we extract eight frames from each

TABLE I

RECOGNITION ACCURACIES (%) OF OUR METHOD FOR DIFFERENT COMPONENTS COMBINATIONS ON THE SHREC'17 TRACK DATASET

Setting	Gestures	
	14	28
LSTG+ILSTSAMs	91.3	86.2
LSTG+SLSTSAM+STE	93.7	89.6
FSG+ILSTSAMs+STE	94.1	89.9
LSTG+ILSTSAMs+STE (LSTSN)	**94.8**	**90.9**

video as input samples and then enhance them by scaling, shifting, time interpolation, and noise adding. Finally, the initial normalization of the data is also performed by subtracting the palm position of the first frame from each skeleton sequence.

The experiments are carried out based on Python 3.6 and Pytorch 1.6. In our network training, we set the number of attention heads of multiple heads as 8, and the dimension d of query, key, and value vector as 32. The Rectified Linear Unit (ReLU) function is selected as the activation function, and layer normalization [27] is used in the training process to standardize the intermediate output of the network. The training of data samples is carried out by batch processing, whose size is set to 32. The other parameters of the network are set as follows: learning rate $= 0.001$, dropout rate $= 0.2$, and the Adam optimizer is used to train the model.

For the DHG-14/28 dataset, we use leave-one-out cross-validation to test and evaluate the model. For the SHREC'2017 Track dataset, we divide the dataset into 1960 training sequences and 840 test sequences, respectively. For the purpose of comparison, we use the same method mentioned in [21] to report the accuracies of both 14 and 28 gestures.

C. Ablation Study

The proposed network consists of three major components, i.e., STE, LSTG, and LSTSAM. In this section, we validate the effectiveness of these components and explore the influence of different input combinations and settings of the proposed method. From Table I, it can be seen that the proposed full model (LSTSN) achieves the best performance, where SLSTSAM denotes sharing only one LSTSAM for all periods and ILSTSAMs represents multiple independent LSTSAMs that are used to model different periods' LSTG.

1) Evaluation of LSTG and STE: To investigate the influence of the LSTG structure, we compare the proposed LSTG with the FC skeleton graph (FSG) structure introduced by [21]. The results in Table I show that the LSTG structure outperforms the one trained on FSG. As FSG may be suboptimal for some hand gestures, while LSTG can describe the "linkage" motion information of different joints more accurately, the attention model can learn more semantic features and get a better gesture recognition. For the proposed STE, we test its effectiveness by training a variant of the proposed method where STE is removed. Table I shows that the STE can improve the performance of the model.

2) Evaluation of LSTSAM: In this article, we mainly use ILSTSAMs to obtain spatial–temporal features. To test the effectiveness of this independent module, we compare the ILSTSAMs with the SLSTSAM. From Table I, we can see that ILSTSAMs significantly outperform SLSTSAM, which shows the necessity of modeling the heterogeneity of spatial–temporal network data.

TABLE II

RECOGNITION ACCURACIES (%) OF OUR METHOD FOR DIFFERENT INPUT COMBINATIONS ON THE SHREC'17 TRACK DATASET

Input data	14 Gestures	28 Gestures
X (DG-STA)	93.7	89.4
HMV (DG-STA)	70.2	65.1
HMV+X (DG-STA)	94.1	90.7
X (LSTSN)	94.8	90.3
HMV (LSTSN)	68.6	61.9
HMV+X (LSTSN)	**95.6**	**92.2**

TABLE III

RECOGNITION ACCURACIES (%) OF OUR METHOD FOR DIFFERENT NODE CONNECTIONS ON THE SHREC'17 TRACK DATASET WITH 14 GESTURES PROTOCOLS

Connections	14 Gestures	parameters Θ	14 Gestures
BC	92.8	6	94.4
BC+L1	93.5	7	94.5
BC+L2	93.5	8	94.7
BC+L3	93.8	9	**95.0**
BC+L4	93.3	10	94.5
BC+L2+L3+L4	94.4	11	94.2
BC+L1+L2+L4	94.2		
AC	94.6		
BC+L1+L2+L3+L4	94.8		

3) Different Input Combinations: To examine the effects of different input descriptor combinations (the 3-D skeletal sample sequence X and HMV), ablation experiments are conducted on the SHREC'17 Track dataset. The results in Table II confirm that HMV is critical indeed for gesture recognition. Combining all the hand skeletons, X and HMV achieve the best performance consistently. Note that HMV contributes more in DG-STA than LSTSN, while X contributes more to LSTSN. Moreover, using both X and HMV in LSTSN performs better than the one in DG-STA, which confirms that LSTSN has the ability to obtain spatial–temporal information.

4) Different Joint Connections: To explore the influence of different joint connections and different threshold parameters, we conduct the ablation study on the SHREC'17 Track dataset with 14 gestures protocols. The results are shown in Table III, where BC is the basic spatial connection of a hand, AC denotes the connections between all joints, and L1, L2, L3, and L4 represent the first, second, third, and fourth connections of a hand in Fig. 2(b), respectively. These results confirm that finding and confirming appropriate joint connections are critical for gesture recognition indeed, while using all the joints does not work very well. When using adaptive methods, setting $\Theta = 9$ is sufficient to obtain good performance.

In addition, we also expand an experimental analysis of the head number H of the multihead attention mechanism. From the experimental results in Fig. 7, it can be seen that the best effect is achieved when $H = 8$.

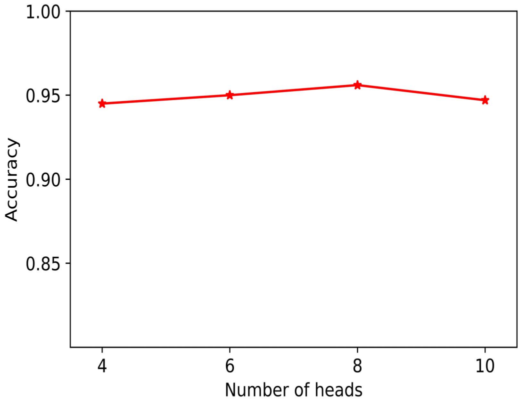


Fig. 7. Performance analysis of attention head numbers on 14 gesture setting in the SHREC'17 dataset.

TABLE IV

COMPUTATION AND PARAMETER NUMBER OF DIFFERENT NETWORK MODELS

Method	Params	FLOPs (G)	Data stream
DG-STA	32 / 0.3M	1.5	1
LSTSN (X)	**22 / 0.1M**	**0.9**	1
LSTSN (X+HMV)	36 / 0.4M	2.1	2
DSTA-Net	377 /3.4M	22.2	4

TABLE V

COMPARISONS OF ACCURACIES (%) ON THE SHREC'17 TRACK DATASET. THE RESULTS OF THE STATE-OF-THE-ART METHODS REPORTED ARE TAKEN FROM CORRESPONDING PAPERS DIRECTLY

Method	**14 Gestures**	**28 Gestures**
SoCJ+HoHD+HoWR	88.2	81.9
HIF3D	90.4	80.4
Res-TCN	91.1	87.7
ST-GCN	92.7	87.7
HG-GCN	92.8	88.3
STA-Res-TCN	93.6	90.7
DG-STA	94.4	90.7
DSTA-Net	**97.0**	**93.9**
LSTSN	95.6	92.2

D. Comparisons With Existing Methods

To see more details on the effectiveness of the LSTSN, in this section, we will give some further comparisons with the representative gesture recognition methods.

The computation amount [floating point operations per second (FLOPs)] and parameter number (Params) of several network models with better effects are compared and analyzed. It can be seen from Table IV that the presented method has the lowest number of parameters and computation under single data stream mode while using multidata stream mode, our method has almost the same number of parameters and computation as DG-STA. However, DSTA-Net is about ten times higher than ours.

The results on the SHREC'17 Track dataset are shown in Table V. LSTSN is superior to the other methods under both 14 gestures and 28 gestures settings except DSTA-Net. Although the recognition accuracy of LSTSN is slightly lower than that of DSTA-Net, our model is far superior to it in terms of network parameters. Moreover, the data features required by the model in this article are far less than those required by DSTA-Net, where only eight frames of sampled data are used in this article, while DSTA-Net requires 128 frames of data.

TABLE VI

COMPARISONS OF ACCURACIES (%) ON THE DHG-14/28 DATASET. THE RESULTS OF THE STATE-OF-THE-ART METHODS REPORTED ARE TAKEN FROM PREVIOUS PAPERS DIRECTLY

Method	**14 Gestures**	**28 Gestures**
SoCJ+HoHD+HoWR	83.1	80.0
Chen et al.	84.7	80.3
CNN+LSTM	85.6	81.1
Res-TCN	86.9	83.6
STA-Res-TCN	89.2	85.0
HG-GCN	89.2	85.3
ST-GCN	91.2	87.1
DG-STA	91.9	88.0
DSTA-Net	**93.8**	90.9
LSTSN	**93.8**	**91.3**

In addition, in the test task of the DHG-14/28 dataset with the same total number of data samples as the SHREC'17 Track dataset, when the training sample was slightly increased (1960 → 2660), the proposed LSTSN achieves the best classification effect under both 14 gestures and 28 gestures settings, as shown in Table VI. Especially, in 28 gestures, LSTSN's recognition rate is 0.4% higher than DSTA-Net. This further proves that the presented method has great advantages in terms of the number of parameters and the amount of computation, and the effectiveness of model construction.

V. CONCLUSION

A local spatial–temporal synchronous network has been proposed for dynamic gesture recognition based on the skeleton in this article. In particular, a new STE has been proposed to achieve the synchronous acquisition of spatial–temporal information. An LSTG is established, and an adaptive method is designed to adjust joint spatial connections to highlight the interactions of major joints. Besides, the attention model has been used to capture local spatial–temporal correlation, and the heterogeneity of spatial–temporal data is also considered using multiple separate LSTSAM. Considering the subtle finger motion, HMV has been further introduced into the model to improve the accuracy. The experimental results have shown that the new method achieves high accuracy on two public datasets at a very fast speed, which provides a general framework that can be further applied to other tasks.

REFERENCES

[1] S. S. Rautaray and A. Agrawal, "Interaction with virtual game through hand gesture recognition," in *Proc. Int. Conf. Multimedia, Signal Process. Commun. Technol.*, Dec. 2011, pp. 244–247.

[2] O. Al-Jarrah and A. Halawani, "Recognition of gestures in Arabic sign language using neuro-fuzzy systems," *Artif. Intell.*, vol. 133, nos. 1–2, pp. 117–138, 2001.

[3] M. Kim, J. Cho, S. Lee, and Y. Jung, "IMU sensor-based hand gesture recognition for human-machine interfaces," *Sensors*, vol. 19, no. 18, p. 3827, Sep. 2019.

[4] S. B. Reed, T. R. C. Reed, and S. M. Dascalu, "Spatiotemporal recursive hyperspheric classification with an application to dynamic gesture recognition," *Artif. Intell.*, vol. 270, pp. 41–66, May 2019.

[5] I. U. Rehman *et al.*, "Fingertip gestures recognition using leap motion and camera for interaction with virtual environment," *Electronics*, vol. 9, no. 12, p. 1986, Nov. 2020.

[6] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, Apr. 2012.

[7] G. Wang, X. Chen, H. Guo, and C. Zhang, "Region ensemble network: Towards good practices for deep 3D hand pose estimation," *J. Vis. Commun. Image Represent.*, vol. 55, pp. 404–414, Sep. 2018.

[8] J. S. Supančič, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, "Depth-based hand pose estimation: Methods, data, and challenges," *Int. J. Comput. Vis.*, vol. 126, pp. 1180–1198, Nov. 2018.

[9] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell, "Head gestures for perceptual interfaces: The role of context in improving recognition," *Artif. Intell.*, vol. 171, nos. 8–9, pp. 568–585, Jun. 2007.

[10] X. Chen, G. Wang, H. Guo, C. Zhang, H. Wang, and L. Zhang, "MFA-Net: Motion feature augmented network for dynamic hand gesture recognition from skeletal data," *Sensors*, vol. 19, no. 2, p. 239, Jan. 2019.

[11] C. Ma, S. Zhang, A. Wang, Y. Qi, and G. Chen, "Skeleton-based dynamic hand gesture recognition using an enhanced network with one-shot learning," *Appl. Sci.*, vol. 10, no. 11, p. 3680, May 2020.

[12] X. Yang and Y. L. Tian, "EigenJoints-based action recognition using nave-bayes-nearest-Neighbor," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 14–19.

[13] H. Chen, G. Wang, J.-H. Xue, and L. He, "A novel hierarchical framework for human action recognition," *Pattern Recognit.*, vol. 55, pp. 148–159, Jul. 2016.

[14] Q. D. Smedt, H. Wannous, and J. P. Vandeborre, "3D hand gesture recognition by analysing set-of-joints trajectories," in *Understanding Human Activities through 3D Sensors*. Cham, Switzerland: Springer, 2018.

[15] Y. Li, Z. He, X. Ye, Z. He, and K. Han, "Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, pp. 1–7, Dec. 2019.

[16] X. Chen, H. Guo, G. Wang, and L. Zhang, "Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2881–2885.

[17] J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Vélez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognit.*, vol. 76, pp. 80–94, Apr. 2018.

[18] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.

[19] J. Hou, G. Wang, X. Chen, J. H. Xue, Z. Rui, and H. Yang, "Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition," *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 273–286.

[20] L. Yang, J. Chen, and W. Zhu, "Dynamic hand gesture recognition based on a leap motion controller and two-layer bidirectional recurrent neural network," *Sensors*, vol. 20, no. 7, p. 2106, Apr. 2020.

[21] Y. Chen, L. Zhao, X. Peng, J. Yuan, and D. N. Metaxas, "Construct dynamic graphs for hand gesture recognition via spatial-temporal attention," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–13.

[22] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[23] W. Hu, X. Cai, J. Hou, S. Yi, and Z. Lin, "GTC: Guided training of CTC towards efficient and accurate scene text recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11005–11012.

[24] Q. De Smedt, H. Wannous, and J.-P. Vandeborre, "Skeleton-based dynamic hand gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 1206–1214.

[25] S. Y. Boulahia, E. Anquetil, F. Multon, and R. Kulpa, "Dynamic hand gesture recognition based on 3D pattern assembled trajectories," in *Proc. 7th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2017, pp. 1–6.

[26] Q. D. Smedt, H. Wannous, J. P. Vandeborre, J. Guerry, B. L. Saux, and D. Filliat, "SHREC-17 track: 3D hand gesture recognition using a depth and skeletal dataset," in *Proc. Eurographics Workshop 3D Object Retr.*, 2017, pp. 1–6.

[27] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[28] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition," in *Computer Vision–(ACCV)*, vol. 12626. Cham, Switzerland: Springer, 2021, pp. 38–53.

Dongdong Zhao received the B.E. degree from Zhengzhou University, Zhengzhou, China, in 2012, the M.E. degree from Lanzhou University, Lanzhou, China, in 2016, and the Dr.Eng. degree from Akita Prefectural University, Akita, Japan, in 2019.

He is currently an Associate Professor with the School of Information Science and Engineering, Lanzhou University. His research interests include multidimensional system theory, signal processing, and machine learning.

Qinglian Yang received the B.E. degree from Lanzhou University, Lanzhou, China, in 2019, where she is currently pursuing the master's degree with the School of Information Science and Engineering.

Her current research interests include machine learning and pattern recognition.

Xingwen Zhou received the B.E. degree from the Harbin Institute of Technology, Harbin, China, in 2017, and the M.E. degree from Lanzhou University, Lanzhou, China, in 2020, where he is currently pursuing the Ph.D. degree with the School of Information Science and Engineering.

His current research interests include computer application.

Hongli Li received the B.E. degree from Lanzhou University, Lanzhou, China, in 2020, where he is currently pursuing the master's degree with the School of Information Science and Engineering.

His current research interests include machine learning and pattern recognition.

Shi Yan received the B.S. and M.E. degrees from Lanzhou University, Lanzhou, China, in 2001 and 2004, respectively, and the Dr.Eng. degree from Akita Prefectural University, Akita, Japan, in 2010.

From 2004 to 2007, he was an Assistant Professor with the School of Information Science and Engineering, Lanzhou University. He was with the Department of Electronics and Information Systems, Akita Prefectural University, as a Visiting Researcher in 2007 and a Visiting Research Fellow in 2010. He was a Lecturer from 2007 to 2012, an Associate Professor from 2012 to 2020, and is currently a Professor with the School of Information Science and Engineering, Lanzhou University. His research interests include multidimensional system theory, signal processing, and machine learning.