electronics

MDPI

Article

Global Correlation Enhanced Hand Action Recognition Based on NST-GCN

Shiqiang Yang *, Qi Li, Duo He, Jinhua Wang and Dexin Li

School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, Xi'an 710048, China
* Correspondence: yangsq@126.com

Abstract: Hand action recognition is an important part of intelligent monitoring, human–computer interaction, robotics and other fields. Compared with other methods, the hand action recognition method using skeleton information can ignore the error effects caused by complex background and movement speed changes, and the computational cost is relatively small. The spatial-temporal graph convolution networks (ST-GCN) model has excellent performance in the field of skeleton-based action recognition. In order to solve the problem of the root joint and the further joint not being closely connected, resulting in a poor hand-action-recognition effect, this paper firstly uses the dilated convolution to replace the standard convolution in the temporal dimension. This is in order to process the time series features of the hand action video, which increases the receptive field in the temporal dimension and enhances the connection between features. Then, by adding non-physical connections, the connection between the joints of the fingertip and the root of the finger is established, and a new partition strategy is adopted to strengthen the hand correlation of each joint point information. This helps to improve the network's ability to extract the spatial-temporal features of the hand. The improved model is tested on public datasets and real scenarios. The experimental results show that compared with the original model, the 14-category top-1 and 28-category top-1 evaluation indicators of the dataset have been improved by 4.82% and 6.96%. In the real scene, the recognition effect of the categories with large changes in hand movements is better, and the recognition results of the categories with similar trends of hand movements are poor, so there is still room for improvement.

Keywords: hand action recognition; ST-GCN; dilated convolution; non-physical connection; partition strategy

Citation: Yang, S.; Li, Q.; He, D.; Wang, J.; Li, D. Global Correlation Enhanced Hand Action Recognition Based on NST-GCN. *Electronics* **2022**, *11*, 2518. https://doi.org/10.3390/electronics11162518

Academic Editor: George A. Papakostas

Received: 18 July 2022
Accepted: 9 August 2022
Published: 11 August 2022

1. Introduction

Hand action recognition is an important research content in the field of computer vision, and it is also a cross-study subject in many subject fields such as machine vision, pattern recognition and artificial intelligence. It is widely used in video surveillance, human–computer interaction, intelligent robot, virtual reality and other fields [1].

Existing hand action recognition methods can be divided into two kinds of mainstream: image sequence-based methods [2–5] and hand skeleton sequence-based methods [6–8] according to input type. RGB or RGB-D image sequence is used as input in the image sequence-based method. Chen et al. [5] proposed a multi-scale attention 3D convolutional network for gesture recognition, with a multimodal fusion scheme to fuse the features of RGB and depth data. A 2D or 3D hand skeleton point coordinate sequence is used as an input in the hand skeleton sequence-based method. As opposed to these two mainstream methods, Jhaung et al. [9] use radar signals as input and output gesture categories after model discrimination.

Hand skeleton data is a kind of topological representation of hand joint and bone structure. It has congenital advantages when facing a complex background, hand scale, visual angle and motion speed. With the development of depth sensor and hand pose estimation technology, accurate hand skeleton structure data can be obtained.

Traditional skeletal structure-based methods, which typically extract motion patterns from specific skeletal structure sequences using hand-crafted features, perform well on some specific datasets but have poor generalization. In recent years, with the development of deep learning methods in other computer vision applications, models such as a convolutional neural network (CNN) [10], a recurrent neural network (RNN) [11] and a graph convolution network (GCN) [12] have emerged. The skeleton structure of hand action is composed of a natural time series of joints, and RNN is more suitable for processing time series data. Therefore, there are many skeleton structure action recognition methods based on RNN and its improved methods. Chen et al. [13] proposed a motion feature augmented recurrent neural network that firstly encodes the joints of each finger and then the joints of the whole hand. When CNN processes the skeleton data sequence, it usually needs to combine the RNN model. Nunez et al. [14] proposed a method to extract features of each frame using CNNs and to aggregate the output of CNNs with a LSTM [15]. The combination of the temporal context information of RNN and the rich spatial information of CNN can often achieve better results than the single structure model. In the last two years, many scholars began to apply GCN to the action recognition of skeleton structure. The hand skeleton sequence is a natural topology graph structure, while the GCN model is more suitable to describe the spatial and temporal topology information between skeletal joints, and has more advantages than RNN. ST-GCN [12] (Spatial Temporal Graph Convolutional Networks) used human topology to construct the adjacency matrix to describe human skeleton structure. This was successfully applied to human action recognition. The literature [16,17] applied ST-GCN to hand action recognition, however only a fixed hand topology was used. Without considering the connection between the root joint and the further joints, this may be not the best choice for hand recognition.

In this paper, based on ST-GCN, we aim to address the problem that the root joint is not closely connected with the more distant joint. which leads to the poor effect of hand movement recognition. A total of three improved modules are proposed using a dilated convolution in the temporal dimension. Adding a non-physical connection and a new partition strategy, it improves the perception ability of the model to the whole hand.

2. Hand Action Recognition Model Based on NST-GCN

ST-GCN network is the first network model to apply a graph convolution to action recognition. It no longer uses hand-crafted features, but instead uses a graph convolution network to extract the features of skeleton sequences. The process of hand action recognition based on ST-GCN model is as follows: Construct a skeleton spatial-temporal graph from given joint points, and then the skeleton spatial-temporal graph is input into the ST-GCN, the classification of hand movements is output after model discrimination. In this paper, the ST-GCN model is studied on the basis of two-dimensional joint coordinate of skeleton dataset.

2.1. Spatial–Temporal Graph Construction

In the recognition of hand action, the two-dimensional coordinate of the joint in the image is often used to construct the skeleton sequence. In the previous hand action recognition methods, the CNN is often used to extract the features of the hand information, and then classifiers are used to classify the extracted features. In the ST-GCN model, it is necessary to construct the skeleton spatial–temporal graph of the joint sequence, which is based on the following criteria: in the spatial domain, for the hand joint on a single picture, two adjacent hand joints are connected to form a skeleton in the spatial domain, and two adjacent frames of the same hand joint are connected to form a skeleton in the temporal domain. The 22 nodes in the DHG-14/28 dataset [18] are shown in Figure 1. A skeleton spatial–temporal graph can be constructed by selecting the hand joints in the DHG-14/28 dataset and the hand action sequence in T frames. The skeleton spatial–temporal graph is shown in Figure 2. The blue dots denote the joints of the hand, the solid lines represent

the natural physical connections of the hand, and the dotted lines represent the temporal connections of the same joint on adjacent frames.

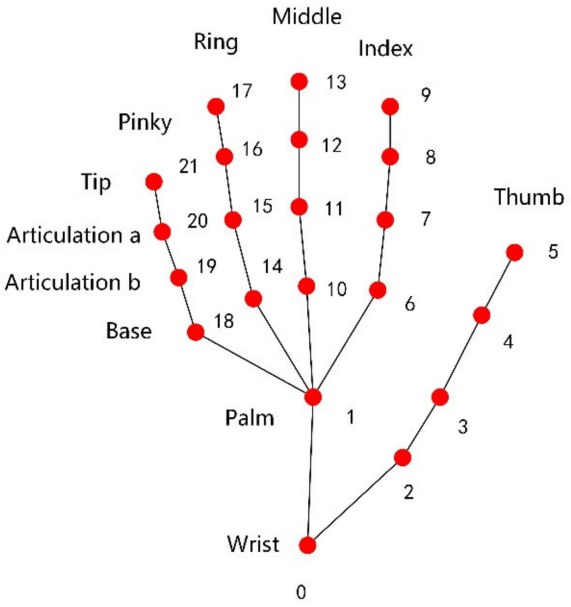

Figure 1. Location of key points in DHG-14/28.

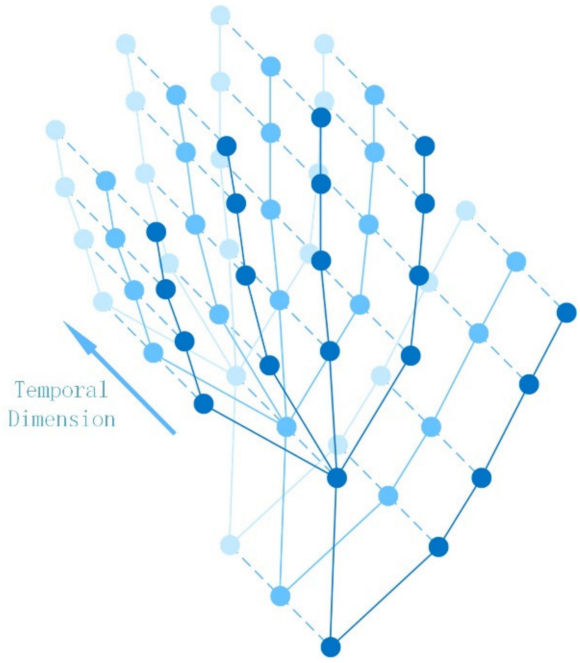

Figure 2. Skeleton spatial–temporal graph.

For the spatial–temporal graph of the hand skeleton, the joint of the hand is represented as the node set $V = \{v_{ti} | t = 1, \ldots, T, i = 1, \ldots N\}$. It consists of the hand joints coordinate information in the skeleton sequence. The edge of the skeleton graph is represented by the edge set E, which consists of two subsets. The first subset is the skeleton edge formed by two adjacent joints at same frame, denoted as $E_S = \{v_{ti}v_{tj}(i, j) \in R\}$. The other subset is the skeleton edge formed by connect the same joints in consecutive frames as $E_F = \{v_{ti}v_{(t+1)i}\}$. For any joint i, its trajectory is the connecting line of all E_F. The skeleton spatial–temporal

graph $G = (V, E)$ is the set of all joints and edges and contains all the changes of the joints in an action sequence.

2.2. Spatial–Temporal Graph Convolution Neural Network Construction

The graph convolution is different from a two-dimensional convolution, and the discrete feature points need to be extracted. In a traditional convolutional neural network, an image has a 2D grid structure and its convolution output is a 2D grid. The convolution operation of an image is similar to that of a normal convolutional neural network. For an input graph with the kernel size $K \times K$ and the number of channels C, the convolution operation output of the spatial position x can be defined as:

$$f_{out}(x) = \sum_{h=1}^{K} \sum_{\omega=1}^{K} f_{in}(\mathbf{p}(x, h, w)) \cdot \mathbf{w}(h, w) \tag{1}$$

where $\mathbf{p}$ is the sampling function representing the location x and its neighbors (h, w) get convoluted, $\mathbf{w}$ is the weight function, it provides a weight vector for inner product operation with the features of the sampled input vectors.

On the skeleton spatial–temporal graph, the sampling function $\mathbf{p}$ represents the graph convolution on the central pixel x and its adjacent pixels. On the graph, the adjacent set sampling function of the node v_{ti} is $B(v_{ti}) = \{v_{tj} | d(v_{tj}, v_{ti}) \leq D\}$. Where $d(v_{tj}, v_{ti})$ denotes the minimum length of any path from v_{tj} to v_{ti}. We select D = 1 to represent the collection of adjacent nodes for all root joints. Therefore, the sampling function $\mathbf{p}(v_{ti}, v_{tj})$ can be defined as follows:

$$\mathbf{p}(v_{ti}, v_{tj}) = v_{tj} \tag{2}$$

The representation of the weight function $\mathbf{w}$ is similar to that of the 2D convolution filter on the skeleton spatial–temporal graph. The weight function of the graph convolution can be constructed in the way of the 2D convolution, and each position provides a weight value. $B(v_{ti}) \rightarrow \{0, \ldots, K-1\}$ is the corresponding relationship, and the weight function can be constructed by mapping the nodes of adjacent sets to its subset label. The partitioning strategy can be used to simplify this mapping change by dividing the adjacent set $B(v_{ti})$ of joint nodes v_{ti} into K subsets. Thus, the constructed weight function $\mathbf{w}(v_{ti}, v_{tj})$ is defined as follows:

$$\mathbf{w}(v_{ti}, v_{tj}) = \mathbf{w}'(l_{ti}(v_{tj})) \tag{3}$$

The sampling function (2) and the weight function (3) are substituted for the function (1), and the function of the spatial graph convolution is obtained:

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(\mathbf{p}(v_{ti}, v_{tj})) \cdot \mathbf{w}(v_{ti}, v_{tj}) \tag{4}$$

where $Z_{ti}(v_{tj}) = |\{v_{tk} | l_{ti}(v_{tk}) = l_{ti}(v_{tj})\}|$ is a normalized term, it is equal to the cardinality of the corresponding subset in order to balance the contribution of different subsets to the output.

By substituting Equations (2) and (3) into Equation (4), the final convolution function of spatial graph is obtained:

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(v_{tj}) \cdot \mathbf{w}'(l_{ti}(v_{tj})) \tag{5}$$

The convolution operation mentioned above is the spatial graph CNN, it does not include convolution of time dimension. The convolution formula of the spatial graph of Equation (5) needs to be extended to time dimension. At the same time, the joint temporal

information between the two frames is included in the adjacent set of joint nodes, so the concept of neighborhood to include temporally connected joints as:

$$B(v_{ti}) = \left\{ v_{qj} \middle| d(v_{tj}, v_{ti}) \leq K, \left| q - t \right| \leq \lfloor \Gamma/2 \rfloor \right\} \tag{6}$$

where the parameter Γ is the time kernel.

In the temporal domain, the result L_{ST} of mapping the adjacent region of a node v_{tj} based on the sampling function and the weight function is:

$$l_{ST}(v_{qj}) = l_{ti}(v_{tj}) + (q - t + \lfloor \Gamma/2 \rfloor) \times K \tag{7}$$

where $l_{ti}(v_{tj})$ is the label map for the single frame case at v_{ti}.

This paper adopts the graph convolution proposed by the Kipf and Welling paper [19]. The skeleton graph for a single frame is represented by the adjacency matrix $\mathbf{A}$ and the identity matrix $\mathbf{I}$. The ST-GCN can be implemented with the following formula:

$$f_{out} = \mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{\Lambda}^{-\frac{1}{2}} f_{in}\mathbf{W} \tag{8}$$

where $\Lambda^{ii} = \sum_j (A^{ij} + I^{ij})$ denotes the degree matrix. Combining spatial–temporal dimension information, the input feature map can be expressed as (*C, V, T*) dimensions. For partitioning strategies with multiple subsets, such as the new partitioning strategy described in Section 2.3.2, the adjacency matrix can be decomposed into multiple matrices, that is, $\mathbf{A} + \mathbf{I} = \sum_j \mathbf{A}_j$, so the degree matrix will also become $\Lambda^{ii} = \sum_k A_j^{ik}$. Therefore, the above formula can be transformed into:

$$f_{out} = {\mathbf{\Lambda}_j}^{-\frac{1}{2}} \mathbf{A}_j {\mathbf{\Lambda}_j}^{-\frac{1}{2}} f_{in}\mathbf{W}_j \tag{9}$$

In accordance with the new partition strategy, j is set to 3, while weight vectors of multiple output channels are superimposed to form a weight matrix $\mathbf{W}$.

2.3. NST-GCN Hand Action Recognition Model

2.3.1. Dilated Convolution

The dilated convolution can enlarge the receptive field of the convolution without increasing the network parameters. Compared with standard convolution, the dilated convolution introduces a hyper-parameter d named dilatation rate, which refers to the spacing between the kernel points, and the standard convolution is when d is 1.

For the hand skeleton, if the computational complexity of the network is not increased, the original spatial–temporal convolution module can only extract the features of the adjacent nodes with distance 1. The nodes that are relatively far apart in the natural structure of the hand but contain important information, are most likely to be gradually decayed in convolution operations. In this paper, we apply the dilated convolution to the hand skeleton graph, and set the hyper-parameter d to 2 to obtain a larger range of features without increasing the computational complexity. In this paper, the hand skeleton graph with the step size D = 1 and dilated rate d = 2 is used for convolution, as shown in Figure 3. The green point is the root node of the graph convolution, and the neighborhood includes the yellow points. The receptive field is much larger than it was without the use of the dilated convolution.

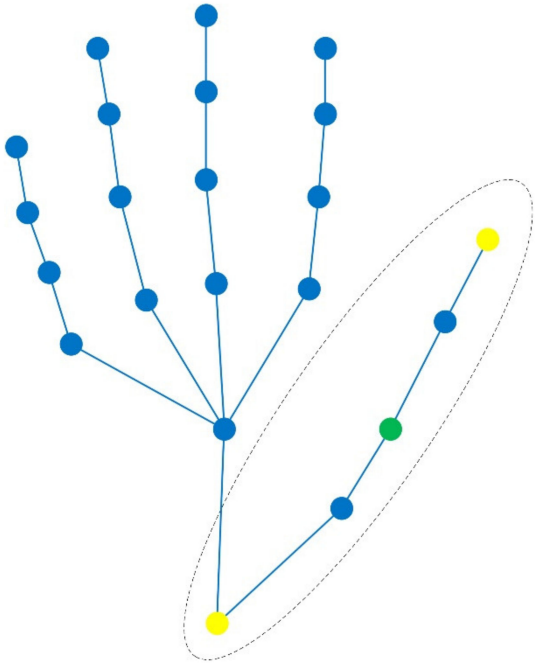

Figure 3. Dilated convolution form on hand skeleton map.

2.3.2. Add Non-Physical Connection

The ST-GCN model constructs the graph structure based on the hand model in the dataset. For the dataset with many kinds of hand movements and similar trends, the hand graph structure using only natural connections may not be effective in extracting critical feature information. The type of movement with high correlation with the top and end joints of five fingers may have a lower accuracy because of less correlation of information in the joint. Therefore, we manually add a non-physical connection as shown by the red dotted line in Figure 4 to the hand skeleton graph, connecting the tip joint of one finger to the root joint of the other finger.

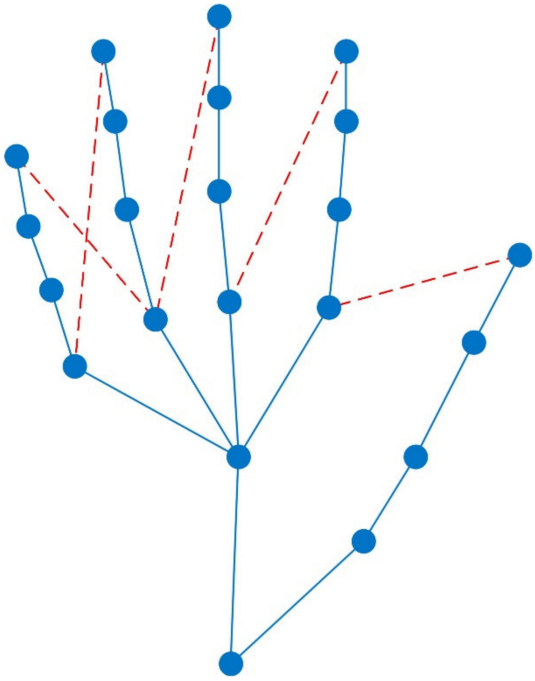

Figure 4. Add non-physical connections to the hand skeleton graph.

2.3.3. Build the New Partition Strategy

A good partitioning strategy can effectively enhance the model's ability to extract features. The ST-GCN model proposes three strategies for partitioning the set of neighborhood points, as shown in Figure 5.

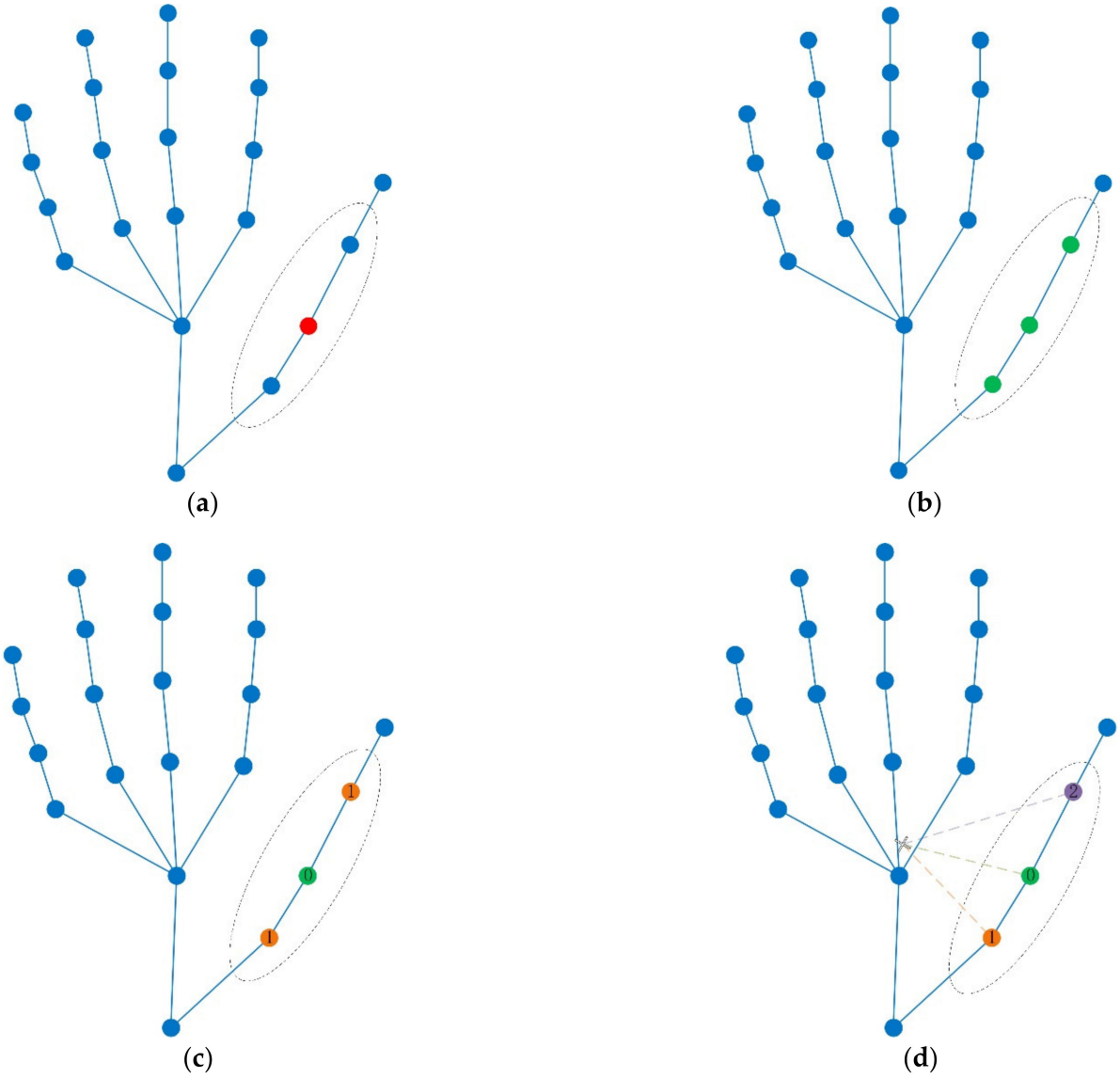

Figure 5. Original partition strategy for constructing convolution operations. (**a**) Example frame of input skeleton; (**b**) Uni-labeling partitioning; (**c**) Distance partitioning; (**d**) Spatial configuration partitioning.

Figure 5a is an example frame of an input skeleton, with a blue dot representing the hand joints. When the parameter D of the filter is 1, the receptive field is the region surrounded by the dotted gray ellipse. Figure 5b is the uni-labeling partition strategy, using a filter with a parameter D of 1. The root node and its adjacent nodes are divided into a subset and assigned the same weight parameters, which are the green dots in Figure 5b. Figure 5c configures the partitioning strategy for the distance using a filter with a parameter D of 1 in order to partition the distances between the nodes and their neighbors. At a distance of 0, the root node itself is represented as a subset. that is, the green points in Figure 5c, and at a distance of 1. The point with a distance of 1 to the root node is represented as a subset, shown as the orange dots in Figure 5c. Figure 5d is a spatial

configuration of the partitioning strategy, using a filter with a parameter D of 1. The spatial-based configuration can divide the node set into three subsets: the root node is a subset, that is, the green dots in Figure 5d; the adjacent nodes closer to the center of gravity of the hand skeleton (black cross) than the root nodes themselves are a subset, that is, the orange dots in Figure 5d; the adjacent joints that are farther from the center of gravity of the hand skeleton (black cross) than the root joints themselves are a subset, shown as the purple dots in Figure 5d.

According to the partition strategy of the spatial configuration, the motion of the hand joint can be divided into centripetal motion and centrifugal motion. The adjacent region of the root joint is divided into three sub-regions: (1) the root node itself; (2) the centripetal set: the neighborhood nodes closer to the skeleton's center of gravity than the root node; and (3) the centrifugal set: the neighborhood nodes farther away from the skeleton's center of gravity than the root node. This strategy can be expressed as follows:

$$l_{ti}(v_{ti}) = \begin{cases} 0 & if \quad r_j = r_i \\ 1 & if \quad r_j < r_i \\ 2 & if \quad r_j > r_i \end{cases} \tag{10}$$

where r_j is the distance from the joint to the center of gravity, r_i is the average distance from the gravity center to joint i overall frames in the training set.

In order to make full use of ST-GCN integrated spatial–temporal features, a new partitioning strategy is used in this paper. The strategy is first used in the spatial dimension and then extended to the spatial–temporal dimension. This paper classifies each joint by the length of the distance between the root joint and other joints, then divides each joint into different subsets.

The new partitioning strategy shown in Figure 6 uses a filter with a parameter D of 2, according to the length of the distance between the root node and the other nodes. It is divided into three subsets: the root node itself a subset, shown as the green point in Figure 6, and the distance from the root node is 1 is a subset, that is, the orange point in Figure 6; the point with a distance of 2 to the root node is divided into a subset, which is the yellow point in Figure 6. The new partition strategy mainly focuses on the movement of the joints in the hand movement to the local joint components of the unit and the physical composition of the hand. The filter parameter D is set to 2, which extends the whole subset and enhances the association of the hand nodes by associating the root node with the further nodes. It makes the model more profound to the local information perception of the hand, thus further improving the accuracy of hand action recognition. Figure 6 shows a new partition strategy proposed in this chapter for constructing convolution operations. The adjacent regions of the joints are divided into three subregions: (1) the root joint (green); (2) the adjacent joint (orange) with a distance of 1; and (3) the remaining adjacent joint (yellow) with a distance of 2.

The new partition strategy not only considers the local motion of the hand, but also considers the connection between the local motion. By associating the root joint with the more distant joint, the information of each joint of the hand is strengthened, and the relationship between global motion and local motion is closer. The perception ability of the model to the whole motion is enhanced, therefore, the accuracy of hand action recognition can be improved.

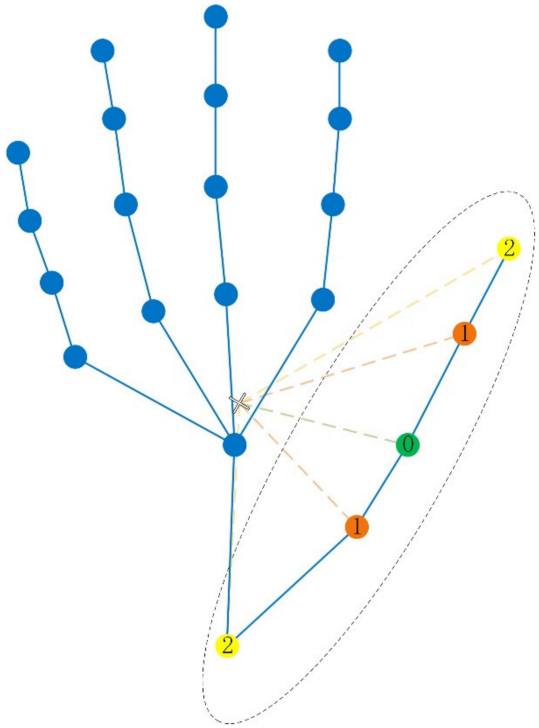

Figure 6. New partition strategy for constructing convolution operations.

2.3.4. Hand Action Recognition Based on NST-GCN Model

This study aims to solve the problem that the ST-GCN temporal convolution kernel samples the hand motion video in a small range of frames most of the time, which results in the loss of temporal feature information. In this paper, the temporal features of the hand motion video are processed by using the dilated convolution. This can enlarge the receptive field in the temporal dimension without reducing the resolution and increasing the cost. The use of multiple dilated convolution layers can effectively extract high-dimensional temporal features, and adding new non-physical connections can improve the model's ability to extract spatial and temporal features. The improved model is called NST-GCN (New ST-GCN), which is shown in Figure 7. Compared with the original network model, the NST-GCN model has a larger sampling area, and the specific structure of NST-GCN is shown in Figure 8.

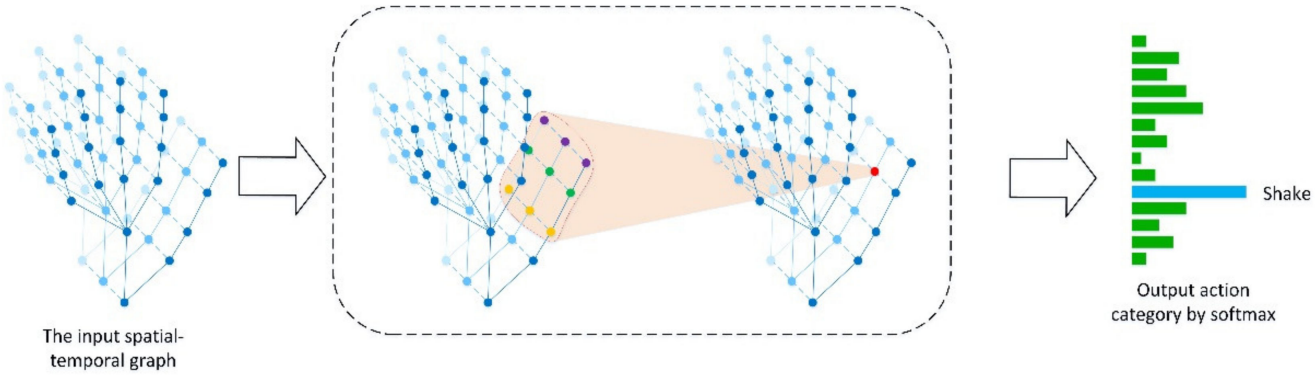

Figure 7. NST-GCN model.

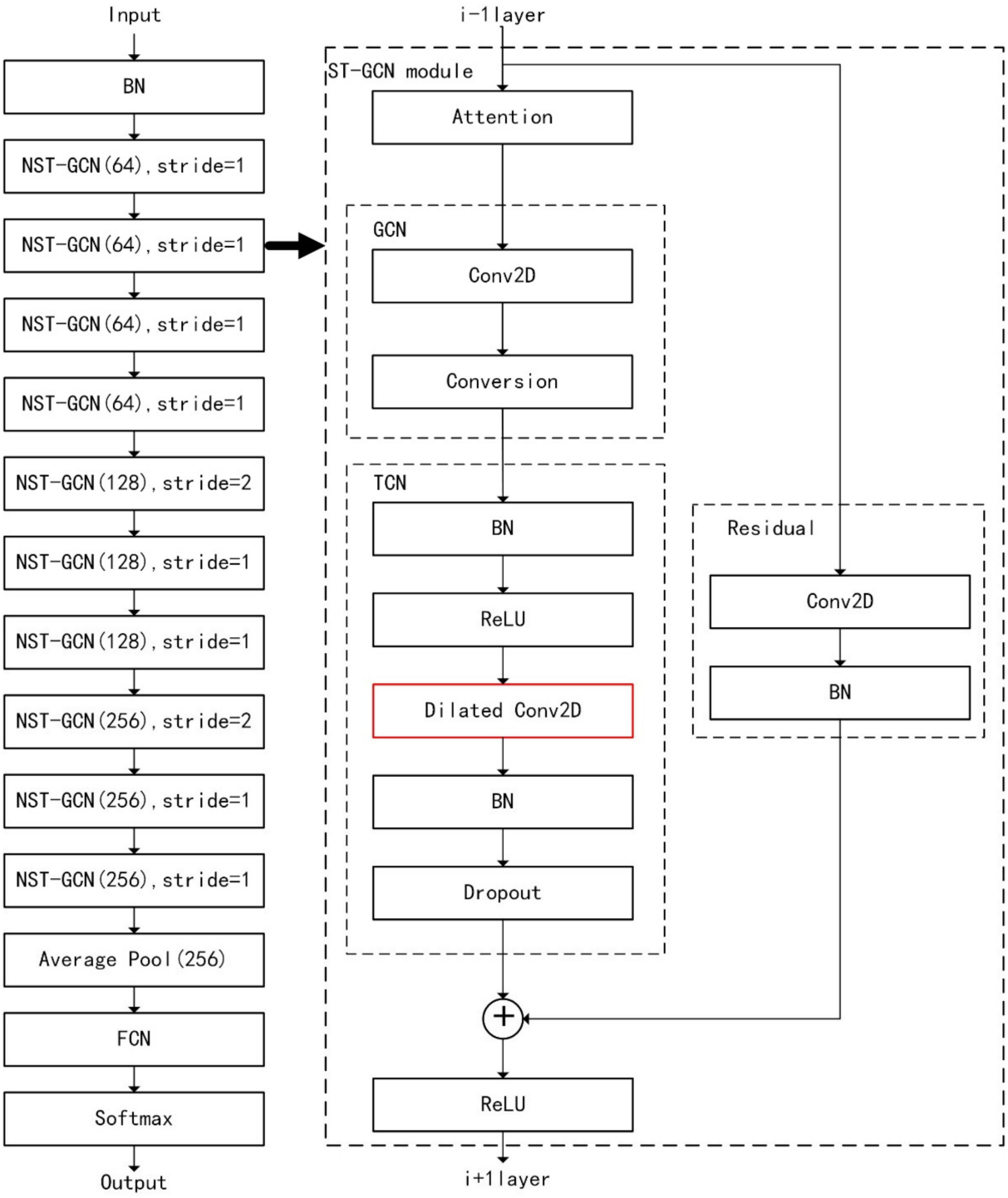

Figure 8. NST-GCN network structure.

Firstly, the hand motion video obtains the hand joint coordinates by the NSRM (Non-parametric Structure Regularization Machine) [20] with a hand pose estimation algorithm. It then forms the data format according to the skeleton graph construction strategy, and inputs into NST-GCN network. Input data from the NST-GCN network were first normalized, and then spatial–temporal features of the hand were extracted through 10 NST-GCN modules. The first fourth NST-GCN modules of the NST-GCN model contains 64 output channels, the fifth seventh NST-GCN modules contains 128 output channels, and the eighth tenth NST-GCN modules contains 25 output channels. Finally, the results of the hand

classification are obtained by a softmax classifier. The input data of the NST-GCN is the coordinate of the hand joints, and its dimension is (*B, C, T, V, N*), where *B* is the batch size, *C* is the coordinate data of the hand joints (x, y, score), and *T* is the frame number of the hand action video, *V* is the number of joints in the hand, and *N* is the number of hands involved in the video. If the batch size *B* is 16, the coordinate characteristic number of the hand joint *C* is 3, the frame number of the hand action video *T* is 70, and the number of the hand joint *V* is 22. The dimensions of network input data are (16, 3, 70, 22, 1) when the number of hands participating in the video $n = 1$.

3. Experiment

To test the effect of different improvement strategies on the accuracy of hand action recognition, in this paper the ST-GCN model is tested from three aspects: using only dilated convolution, using the dilated convolution and adding non-physical connections, using the dilated convolution and a new partitioning strategy and adding non-physical connections. The robustness and advancement of the NST-GCN model are verified in the DHG-14/28 large open hand action dataset.

3.1. Dataset & Evaluation Metrics

The DHG-14/28 dataset is a publicly available hand action dataset. The dataset contains a sequence of 2800 hands with 14 hand categories, including Grab, Tap, Expand, Pinch, Rotation CW, Rotation CCW, Swipe Right, Swipe Left, Swipe Up, Swipe Down, Swipe X, Swipe V, Swipe + and Shake, performed as 1-finger or 5-fingers configurations (thus also as 28 classes). For DHG-14/28, each hand configuration was performed 5 times by 20 participants.

In this paper, the performance of the NST-GCN model for hand action recognition is verified by using the accuracy indexes of Top-1 and Top-5. Top-1 is the probability that the first-ranked category in the inferred probability vector is the correct category, also known as the accuracy of the classification; Top-5 is the probability that the top five-ranked categories in the inferred probability vector contains the correct category. Top-1 and Top-5 can be calculated using the following formula:

$$top-1 = \frac{\sum_i^N \delta(class_i^{true} = rank_1(class_i^{pred}))}{N} \tag{11}$$

$$top-5 = \frac{\sum_i^N \delta(class_i^{true} \in rank_5(class_i^{pred}))}{N} \tag{12}$$

where δ is the judgment function, if the condition is false, then take the value of 0, otherwise take the value of 1; $class_i^{true}$ is the correct category for the *i*-th action; $rank_1(class_i^{pred})$ is the inference categories with the highest score of the *i*-th action probability; $rank_5(class_i^{pred})$ is the top five inference categories for the *i*-th action; *N* is the number of hands.

3.2. Results and Analysis of Model Training and Hand Action Recognition

This paper uses 2D joint point coordinates to train, validate and test a hand motion recognition model on the DHG-14/28 dataset. The model is trained on the training set, and the model is tested on the validation set and test set. The environment configuration for network training is shown in Table 1.

Table 1. Experimental training environment.

Environment Configuration	
System	windows10
GPU	P106-100
Memory size	6GB
CPU	Intel(R) Core(TM) i5-4460 CPU @ 3.20GHz
Python	3.7
Torch	1.2.0
CUDA	10.0

The model is trained using the SGD optimizer, with a momentum of 0.9 and weight decay of 0.0001. The batch size is set to 16 and the number of epochs is 100. The initial learning rate is 0.1, and when the number of epochs reaches 60, 80 and 90, the learning rate is attenuated to 1/10 of the original.

This paper tests the effect of each module's improvement on the overall performance of the NST-GCN model through experiments. The basic method is to directly use the ST-GCN model to recognize the hand movements of the two-dimensional joint based on the DHG-14/28 dataset. DC is the dilated convolution module, NPC is the non-physical connection module, and NPS is the new partition strategy module. Table 2 details the performance improvement of the model by adding an additional improved module at each stage based on ST-GCN. Using ST-GCN directly to recognize the hand movements, the evaluation indexes of 14 Top-1, 14 Top-5, 28 Top-1 and 28 Top-5 were 73.04%, 90.54%, 74.11% and 92.50%, respectively. On the basis of ST-GCN, compared with the original model, the evaluation indexes of 14 categories of Top-1 and 28 categories of Top-1 were improved by 1.07% and 4.46%. On the basis of ST-GCN, using dilated convolution at the temporal dimension and adding non-physical connectivity, it has improved by 3.75% and 5.71% compared with the 14 Top-1 and 28 Top-1 evaluation indexes of the original model. Based on ST-GCN, using dilated convolution and new partitioning strategy, it has improved by 3.21% and 5.71%, compared with the 14 Top-1 and 28 Top-1 evaluation indexes of the original model. Compared with the original ST-GCN model, the NST-GCN model has an improvement of 4.82% and 6.96% in the evaluation indexes of 14 Top-1 and 28 Top-1. Therefore, the three improved modules proposed in this paper are effective, and the NST-GCN model proposed in this paper is effective for hand action recognition.

Table 2. Comparison of accuracy of different improvement modules.

Method	DC	NPC	NPS	14 Top-1	14 Top-5	28 Top-1	28 Top-5
ST-GCN	-	-	-	73.04%	90.54%	74.11%	92.50%
+DC	√	-	-	74.11%	90.54%	78.57%	93.39%
+NPC	√	√	-	76.79%	91.61%	79.82%	93.57%
+NPS	√	-	√	76.25%	91.25%	79.82%	93.57%
Ours	√	√	√	77.86%	91.61%	81.07%	93.57%

3.3. Real Scene Model Test Results

In order to verify the recognition effect of NST-GCN in the real scene, the mobile phone camera was used to collect the hand motion video in the laboratory scene, and the resolution of the video was 720 × 1280. The 14 categories of hands in the DHG-14/28 dataset was used: grab, tap, expand, pinch, rotation CW, rotation CCW, swipe right, swipe left, swipe up, swipe down, swipe X, swipe V, swipe + and shake. Each type of hand movement is completed by 2 different testers. Each tester repeats the movement 5 times, and each category of hand action contains a total of 10 sequences.

Figures 9 and 10 show the changing process of hand joint point information for shake and grab actions in the DHG-14/28 dataset, respectively. Firstly, the specific position of the hand is obtained through the hand target detection model, and then the intercepted

hand image is sent to the hand pose estimation model to obtain the coordinates of the hand joints. Finally, the change process of the hand joint information is obtained. From Figures 9 and 10, it can be observed that the change of hand joint information can effectively show the change trend of hand actions.

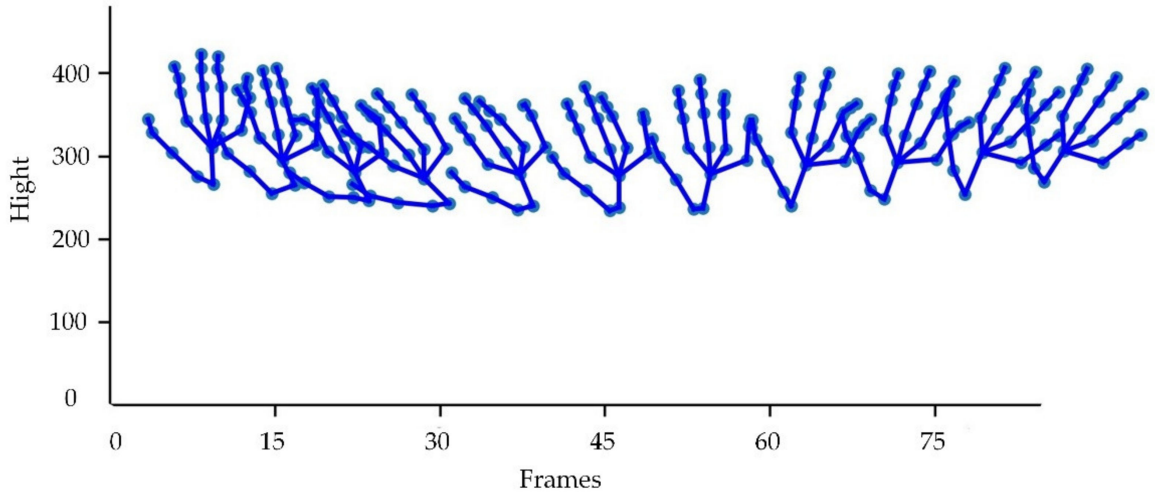

Figure 9. The change process of shake action.

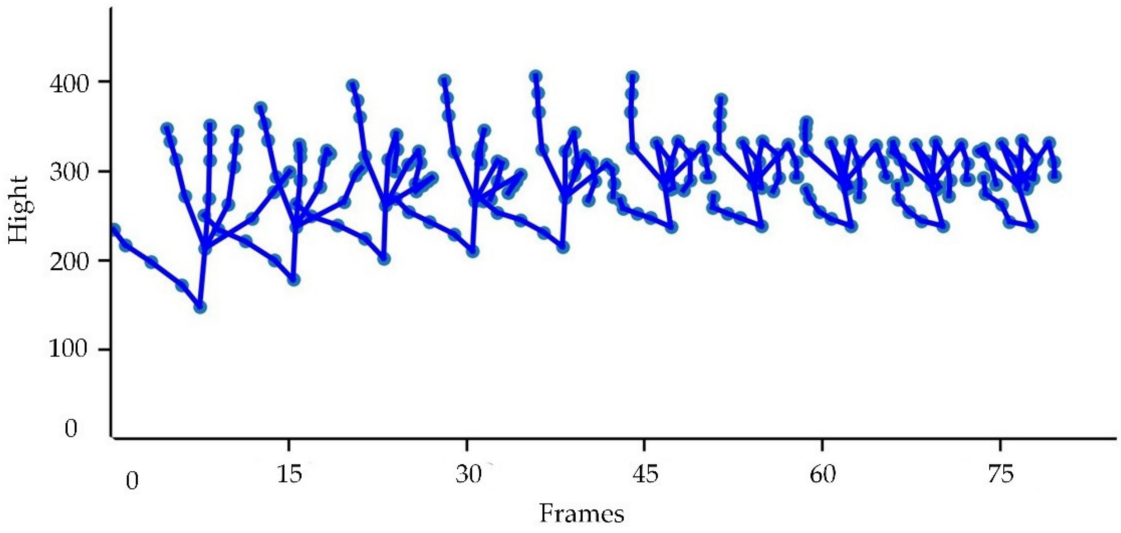

Figure 10. The changing process of grab action.

The change trend of the whole hand in the shaking action can be seen from Figure 9, and the change trend of the five fingers in the grab action can be seen from Figure 10. Due to the errors in hand detection and hand pose estimation, some joint points may be missing, but it does not affect the recognition of hand actions.

For each category of hand, 8 key frames are given to show the results of hand action recognition, and the experimental results are analyzed.

Figure 11 is the hand action recognition result of the grab action. Based on the coordinate information of the hand joints, the hand action recognition is realized by using the hand action recognition model based on the NST-GCN network.

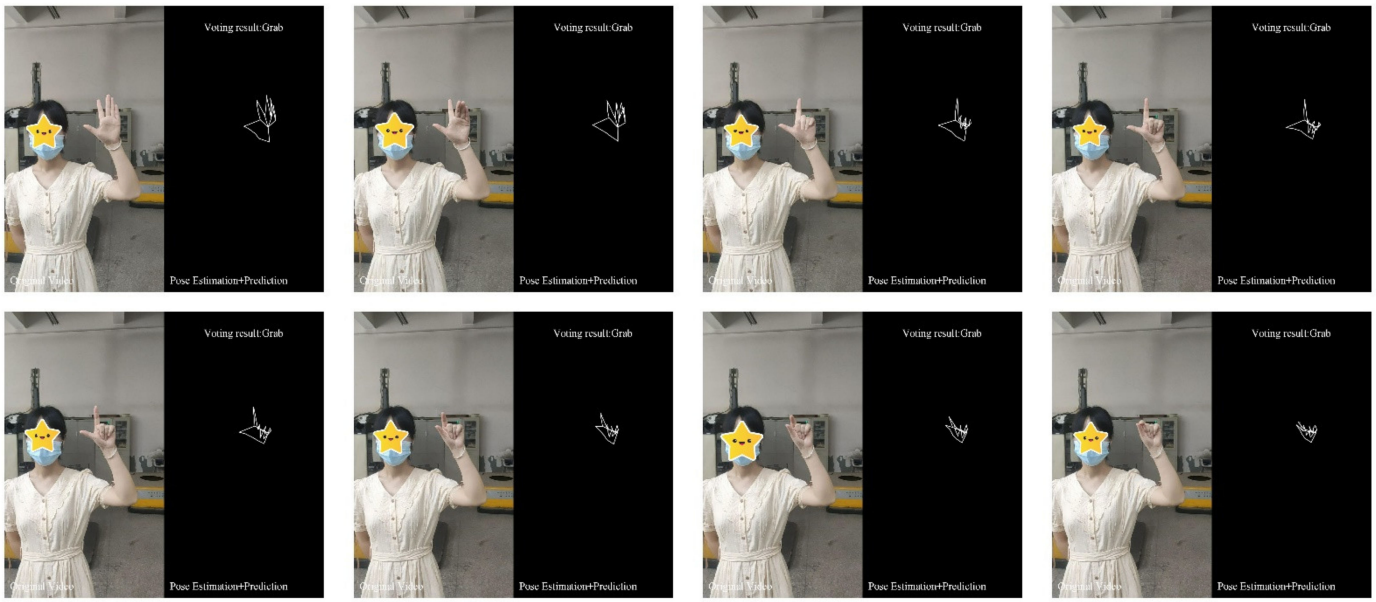

Figure 11. Hand action recognition results (Grab).

Figure 12 is the process of hand action recognition of the swipe V action. Based on the coordinate information of hand joints, hand action recognition is realized by using the hand action recognition model based on the NST-GCN network.

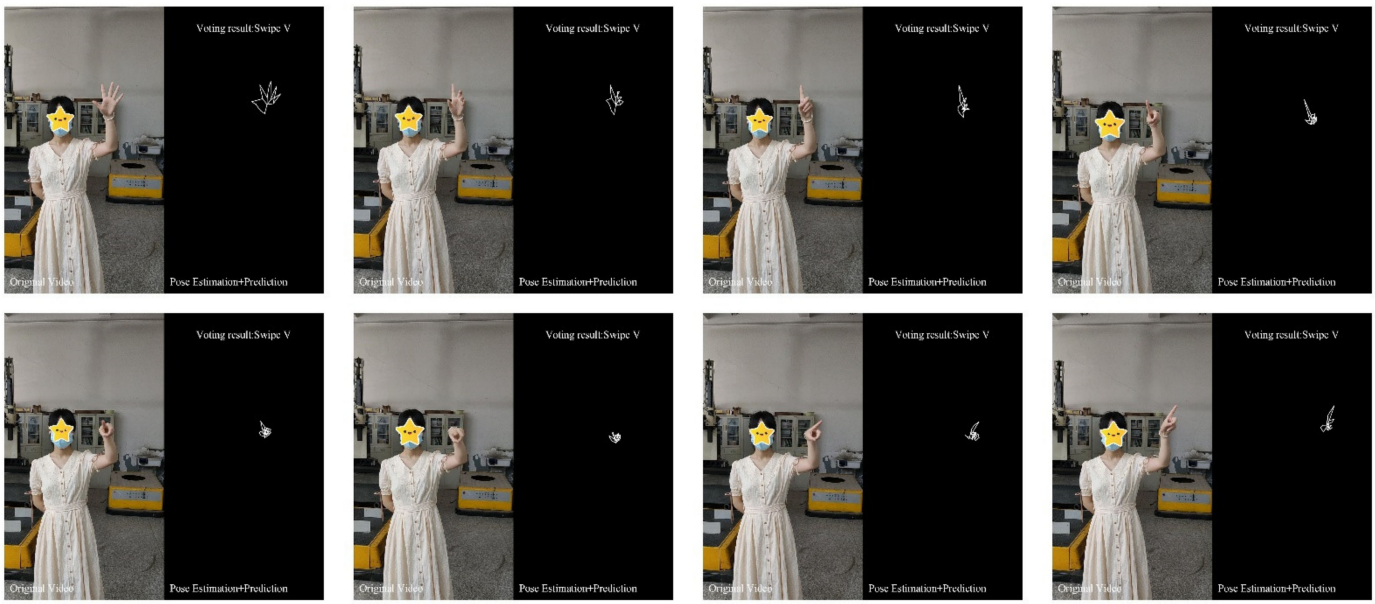

Figure 12. Hand action recognition results (Swipe V).

Using a mobile phone to collect 14 kinds of hand action in the laboratory, each action contains 10 action sequences. The overall recognition rate statistics are shown in Table 3. As can be seen from Table 3, the recognition model of hand action based on the NST-GCN network has a good effect on the categories of hand action with great changes, such as swipe X, swipe V and swipe +, etc. The recognition accuracy was 100%, and was poor for hand categories with small amplitude or similar trend of hand changes, such as grab and pinch actions. The recognition model based on the NST-GCN network still has room for improvement for the similar trend of hand action.

Table 3. Results of hand action recognition accuracy.

Categories	Accuracy
Grab	60%
Tap	80%
Expand	80%
Pinch	70%
Rotation CW	60%
Rotation CCW	80%
Swipe right	90%
Swipe left	90%
Swipe up	90%
Swipe down	90%
Swipe X	100%
Swipe V	100%
Swipe +	100%
Shake	50%

4. Conclusions

In this paper, we present the NST-GCN model for hand action recognition based on joint information. Firstly, the spatial–temporal graph convolution is constructed, and the sampling function and weight function are set up. Then, three improved modules are introduced to solve the problem of the root joint and the further joint not being closely connected, resulting in the poor hand action recognition effect. We used dilated convolution in the temporal dimension to increase the receptive field in the time domain, adding a non-physical connection and using a new partition strategy to strengthen the hand correlation of each joint point information. The ablation experiments show the validity of the three improved modules, and the NST-GCN model is established. The results show that the NST-GCN model is more accurate than the original model, and the performance index of Top-1 is 77.86% on DHG-14/28. This is 4.82% higher than that of the ST-GCN model, which shows that it has a good recognition effect. In the real scene, the recognition accuracy of the hand categories with great changes in hand movements, and the recognition accuracy of the hand categories with similar trends in hand movements is poor, so the performance of the model still has room for improvement. In the future, the problem of low accuracy of hand motion recognition with a similar motion trend will be further studied.

Author Contributions: Conceptualization, J.W.; methodology, J.W.; resources, S.Y. and D.L.; software, Q.L.; supervision, S.Y.; writing—review and editing, Q.L. and D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Natural Science Foundation of Shaanxi Province, China] grant number [2021SF-422].

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to [In order to adapt to our study, we processed the dataset].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dabwan, B.A.; Jadhav, M.E. A review of sign language and hand motion recognition techniques. *Int. J. Adv. Sci. Technol.* **2020**, *29*, 4621–4635.
2. Fan, D.; Lu, H.; Xu, S.; Cao, S. Multi-Task and Multi-Modal Learning for RGB Dynamic Gesture Recognition. *IEEE Sens. J.* **2021**, *21*, 27026–27036. [CrossRef]
3. Zhang, W.J.; Wang, J.C.; Lan, F.P. Dynamic Hand Gesture Recognition Based on Short-Term Sampling Neural Networks. *IEEE/CAA J. Autom. Sin.* **2021**, *8*, 110–120. [CrossRef]
4. Zhang, X.; Yun, T.; Lin, Q. Dynamic Gesture Recognition Based on 3D Separable Convolutional LSTM Networks. In Proceedings of the IEEE 11th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 16–18 October 2020.
5. Chen, H.; Li, Y.; Fang, H.; Xin, W.; Lu, Z.; Miao, Q. Multi-Scale Attention 3D Convolutional Network for Multimodal Gesture Recognition. *Sensors* **2022**, *22*, 2405. [CrossRef]
6. Ma, C.; Zhang, S.; Wang, A. Skeleton-Based Dynamic Hand Gesture Recognition Using an Enhanced Network with One-Shot Learning. *Appl. Sci.* **2020**, *10*, 3680. [CrossRef]
7. Zhang, W.; Lin, Z.; Cheng, J. STA-GCN: Two-stream graph convolutional network with spatial–temporal attention for hand gesture recognition. *Vis. Comput.* **2020**, *36*, 2433–2444. [CrossRef]
8. Nguyen, N.H.; Phan, T.D.; Kim, S.H. 3D Skeletal Joints-Based Hand Gesture Spotting and Classification. *Appl. Sci.* **2021**, *11*, 4689. [CrossRef]
9. Jhaung, Y.-C.; Lin, Y.-M.; Zha, C.; Leu, J.-S.; Köppen, M. Implementing a Hand Gesture Recognition System Based on Range-Doppler Map. *Sensors* **2022**, *22*, 4260. [CrossRef] [PubMed]
10. Santos, C.; Samatelo, J.; Vassallo, R.F. Dynamic Gesture Recognition by Using CNNs and star RGB: A Temporal Information Condensation. *Neurocomputing* **2020**, *400*, 238–254. [CrossRef]
11. Chen, W.; Fan, Y.; Zhang, Y. Dynamic Gesture Recognition Based on iCPM and RNN. *J.Phys. Conf. Ser.* **2020**, *1684*, 012066.
12. Yan, S.J.; Xiong, Y.J.; Lin, D.H. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
13. Chen, X.; Guo, H.; Wang, G.; Zhang, L. Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017.
14. Nunez, J.C.; Cabido, R.; Pantrigo, J.J.; Montemyaor, A.S.; Velez, J.F. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognit.* **2018**, *76*, 80–96. [CrossRef]
15. Graves, A. Long Short-Term Memory. Studies in Computational Intelligence. In *Supervised Sequence Labelling Recurrent Neural Network*; Springer: Berlin, Germany, 2012; Volume 385, pp. 37–45.
16. Nguyen, X.S.; Brun, L.; Lezoray, O.; Bougleux, S. A neural network based on SPD manifoldlearning for skeleton-based hand gesture recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
17. Chen, Y.; Zhao, L.; Peng, X.; Yuan, J.; Metaxas, D.N. Construct dynamic graphs for hand gesture recognition via spatial-temporal attention. In Proceedings of the British Machine Vision Conference, Cardiff, UK, 9–12 September 2019.
18. Smedt, Q.D.; Wannous, H.V.; Borre, J.P. Skeleton-Based Dynamic Hand Gesture Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1206–1214.
19. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
20. Chen, Y.; Ma, H.; Kong, D.; Yan, X.; Wu, J.; Fan, W.; Xie, X. Nonparametric Structure Regularization Machine for 2D Hand Pose Estimation. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020.