electronics

MDPI

Article

Traffic Police Gesture Recognition Based on Gesture Skeleton Extractor and Multichannel Dilated Graph Convolution Network

Xin Xiong [1], Haoyuan Wu [1], Weidong Min [2,3,*], Jianqiang Xu [1], Qiyan Fu [1] and Chunjiang Peng [1]

[1] School of Information Engineering, Nanchang University, Nanchang 330031, China; 351032718002@email.ncu.edu.cn (X.X.); 401030919020@email.ncu.edu.cn (H.W.); xjq@ncu.edu.cn (J.X.); 351029019003@email.ncu.edu.cn (Q.F.); ppcj@sohu.com (C.P.)
[2] School of Software, Nanchang University, Nanchang 330047, China
[3] Jiangxi Key Laboratory of Smart City, Nanchang 330047, China
* Correspondence: minweidong@ncu.edu.cn

Abstract: Traffic police gesture recognition is important in automatic driving. Most existing traffic police gesture recognition methods extract pixel-level features from RGB images which are uninterpretable because of a lack of gesture skeleton features and may result in inaccurate recognition due to background noise. Existing deep learning methods are not suitable for handling gesture skeleton features because they ignore the inevitable connection between skeleton joint coordinate information and gestures. To alleviate the aforementioned issues, a traffic police gesture recognition method based on a gesture skeleton extractor (GSE) and a multichannel dilated graph convolution network (MD-GCN) is proposed. To extract discriminative and interpretable gesture skeleton coordinate information, a GSE is proposed to extract skeleton coordinate information and remove redundant skeleton joints and bones. In the gesture discrimination stage, GSE-based features are introduced into the proposed MD-GCN. The MD-GCN constructs a graph convolution with a multichannel dilated to enlarge the receptive field, which extracts body topological and spatiotemporal action features from skeleton coordinates. Comparison experiments with state-of-the-art methods were conducted on a public dataset. The results show that the proposed method achieves an accuracy rate of 98.95%, which is the best and at least 6% higher than that of the other methods.

Keywords: traffic police gesture recognition; gesture skeleton extractor; multichannel dilated; graph convolution network

Citation: Xiong, X.; Wu, H.; Min, W.; Xu, J.; Fu, Q.; Peng, C. Traffic Police Gesture Recognition Based on Gesture Skeleton Extractor and Multichannel Dilated Graph Convolution Net-work. *Electronics* **2021**, *10*, 551. https://doi.org/10.3390/electronics10050551

Academic Editor: Jungong Han; Guiguang Ding

Received: 21 January 2021
Accepted: 21 February 2021
Published: 26 February 2021

1. Introduction

In daily traffic, traffic signals are important for ensuring the smooth flow of road traffic and increasing roadway traffic security. Traffic signals include not only signal lamps, signs, and markings but also traffic police commands. In the event of special situations, such as traffic light failure, bad weather, traffic congestion, and so on, traffic police typically control traffic and guide drivers using command gestures. In recent years, self-driving cars have gradually entered people's field of vision. Therefore, driverless cars must be able to not only recognize traffic lights but also quickly and correctly respond to and process traffic police's flexible gestures. Thus, traffic police gesture recognition is crucial in driver assistance systems and intelligent vehicles. In recent years, deep learning achieved considerable success in computer vision [1–3]. It excels in image-level image classification, object detection, image segmentation, and other fields and achieved breakthroughs [4–7] in video-level action recognition and action detection. Therefore, recent works [8–11] used deep learning methods to solve traffic police gesture

recognition problems. However, current traffic police gesture recognition methods pose certain difficulties, and the recognition task generally faces two challenges. First, most existing traffic police gesture recognition methods extract pixel-level features from RGB images which are uninterpretable because of the lack of gesture skeleton features and may result in inaccurate recognition due to background noise. Appropriate and effective features representing traffic police gestures should be chosen and extracted. However, traffic police typically work in complex and unpredictable environments, which can introduce interference and render features uninterpretable. A gesture skeleton extractor (GSE) is proposed in this study, which can extract and improve interpretable skeleton coordinate information. Compared with extracted pixel-level features, skeleton information can eliminate background interference and make features interpretable through coordinates and the proposed attention mechanism. Second, existing deep learning methods are not suitable for handling gesture skeleton features. These methods ignore the inevitable connection between skeleton joint coordinate feature and gestures. Several works [8,9,12] extracted traffic police skeleton data and proved that this method is effective. However, some problems exist in prior works. For example, previous studies generally relied on handcrafted components or pixel-level information to extract skeleton features. This method cannot determine the relationship between skeleton joint coordinates and gestures, misses interpretable topologic features, and demonstrates weak generalization capability and poor recognition performance. Recently, graph convolution achieved considerable success in such tasks [13–17]. Reference [13] proposed a graph-based scalable semi-supervised learning method, and Yan et al. [14] employed a graph convolution in skeleton data for action recognition. Inspired by previous works, a new graph convolution network, namely the multichannel dilated graph convolution network (MD-GCN), is proposed to address the traffic police gesture task in this study. Compared with traditional action recognition methods, the proposed method focuses on specific traffic police gesture features and the classification network and demonstrates better performance.

In this work, a traffic police gesture recognition method based on a GSE and the MD-GCN is proposed. A GSE is proposed in this study which can extract discriminative and interpretable gesture features. In the gesture discrimination stage, GSE-based features are introduced into a new framework called the MD-GCN which extracts body topological and spatiotemporal action features. The MD-GCN can enrich the representation of gesture features and make the features interpretable. Moreover, the proposed method demonstrates superior performance in the Police Gesture Dataset [9]. The contributions of the proposed method are summarized below:

1. A GSE is proposed to extract skeleton coordinate information and remove redundant skeleton joints and bones, which improves skeleton data representation to strengthen the attention of the network.
2. The MD-GCN is proposed to extract body topological and spatiotemporal action features from skeleton coordinates, which can enrich the representation of gesture features and make features interpretable.

2. Related Work

Traffic police gesture recognition belongs to the field of action recognition. Compared with traditional action recognition methods, the proposed method focuses on specific traffic police gesture features and the classification network and demonstrates better performance. Owing to its important application value, numerous researchers conducted studies on traffic police gesture recognition, and their methods can be divided into on-body sensor-based methods and visual-based methods. On-body sensor-based methods use microelectromechanical systems and inertial sensors (e.g., accelerometers and gyroscopes) to collect posture and movement data. For example, Wang et al. [18] employed two three-axis accelerators fixed on the back of the hand of a traffic police of-

ficer to collect hand motion features and then recognized the gestures by comparing them with predefined templates. Yuan et al. [19] adopted a similar method to design a Chinese traffic police gesture recognition system, which achieved a satisfactory recognition rate. However, the operation of such methods is relatively complex and requires traffic police to wear sensor devices, which can increase their work burden. In addition, the relatively expensive equipment limits the wide application of sensor-based methods. Compared with on-body sensor-based methods, visual sensor-based methods are more convenient and affordable. Such methods use only a vision sensor to locate and capture the gestures of traffic police in a video and then extract poses, motion, and other features from the video and classify the traffic police gestures according to the features. In recent years, numerous studies on visual methods for traffic police gesture recognition emerged. For instance, Le et al. [20] captured depth images using a Kinect sensor and then created feature vectors according to the relative angles between the joints of the skeleton model extracted from the depth images. Support vector machine classifiers were employed in reference [20] for traffic police gesture recognition. Guo et al. [21] used a single Kinect camera to obtain depth information and RGB image data and then constructed static descriptors from the depth information and estimated dynamic descriptors from the RGB image sequences. The authors generated a fusion descriptor by combining a static descriptor with a dynamic descriptor and then used the mean structural similarity index to spot Chinese traffic police gestures. Zhang et al. [12] utilized Kinect to obtain the fingertip position and then recognize the gesture by an improved DTW algorithm. Similarly, Ma et al. [8] used Kinect 2.0 to build the Traffic Police Command Gesture Skeleton dataset and proposed a spatiotemporal convolution neural network to identify traffic police command gestures. The method can capture and provide a human skeleton with Kinect sensors for identifying gestures and demonstrates improved recognition accuracy. However, when the distance between the Kinect sensor and the traffic police is wide, the accuracy rate of the traffic police hand recognition method declines. The above vision-based methods are based on depth images; however, some methods only use RGB images as input. For example, Guo [22] utilized the Gabor feature-based two-dimensional (2D) principal component analysis. Cai et al. [23] extracted the body and arm of a traffic police officer in the foreground from an image retrieved by an RGB camera and then proposed a max-covering scheme to determine the coordinates of the pixels in the upper arm and forearm and recognized traffic police gestures through the rotation joint angle. However, these methods could not easily process a situation where the arm of the traffic police officer is perpendicular to the image plane. With the success of 2D human posture estimation [24–29], reference [9] employed a modified convolutional pose machine to extract 2D skeletal data from an RGB video, utilized the data to construct two types of handcrafted features as spatial features, extracted temporal features from the handcrafted features of the consecutive images using a long short-term memory (LSTM) network, and used a dense connected network to classify Chinese traffic police gestures. Although this network demonstrates fast real-time performance, its recognition accuracy rate in the Police Gesture Dataset is insufficient at only 91.18%. To improve recognition accuracy, the researchers constructed handcrafted features for supplementation similar to [9,22,23], which is relatively cumbersome. In contrast to [9], the proposed method does not require handcrafted features and achieves a high recognition accuracy rate of 98.95% in the Police Gesture Dataset.

The above methods present several issues. Some methods extract pixel-level features from RGB images which are uninterpretable because of the lack of gesture skeleton features and may result in inaccurate recognition due to background noise. Some methods are not suitable for handling gesture skeleton features because they ignore the inevitable connection between skeleton joint coordinate information and gestures. To alleviate the aforementioned issues, a traffic police gesture recognition method based on a GSE and the MD-GCN is proposed. The extraction of traffic police skeleton data through pose estimation can reduce the influence of the background on recognition. A GSE is

proposed to extract skeleton coordinate information and remove redundant skeleton joints as well as bones, which improves skeleton data representation to strengthen the attention of the network. An MD-GCN is proposed to extract body topological and spatiotemporal action features from skeleton coordinates, which can enrich the representation of gesture features and make features interpretable. Then, the coordinates and confidence of the joints of the skeleton sequences construct a tensor as the input of the proposed MD-GCN. Finally, we realize the classification of the traffic police gestures.

3. Overview of the Proposed Method

Human body motion can be described by the movement of certain joints. In recent years, dynamic human skeletons were widely used in the field of action recognition. Compared with RGB information, skeleton information has several advantages. Specifically, skeleton information makes features interpretable and simple and is not disturbed easily by the background. The gesture sequence of a traffic police skeleton is important for gesture recognition. This study is based on the gesture sequence features of a traffic police skeleton instead of RGB features. As shown in Figure 1, the proposed traffic police gesture recognition method has two components, that is, a GSE and MD-GCN. The first component extracts the traffic police skeleton sequence from a video using a pose estimation algorithm: OpenPose [24]. Next, the gesture skeleton sequence eliminates the redundant joints to improve the attention of the network on traffic police gestures. The other component is the MD-GCN, with the gesture skeleton sequence as the input. The MD-GCN constructs a graph convolution using a multichannel dilated to enlarge the receptive field, which extracts body topological and spatiotemporal action features from the skeleton coordinates. The 2D skeleton coordinates and the confidence of the joints construct a 3D tensor which is the input of the proposed MD-GCN. The network significantly enriches the representation of gesture features and improves gesture recognition accuracy by fusing the features of the different channel graph convolutions.

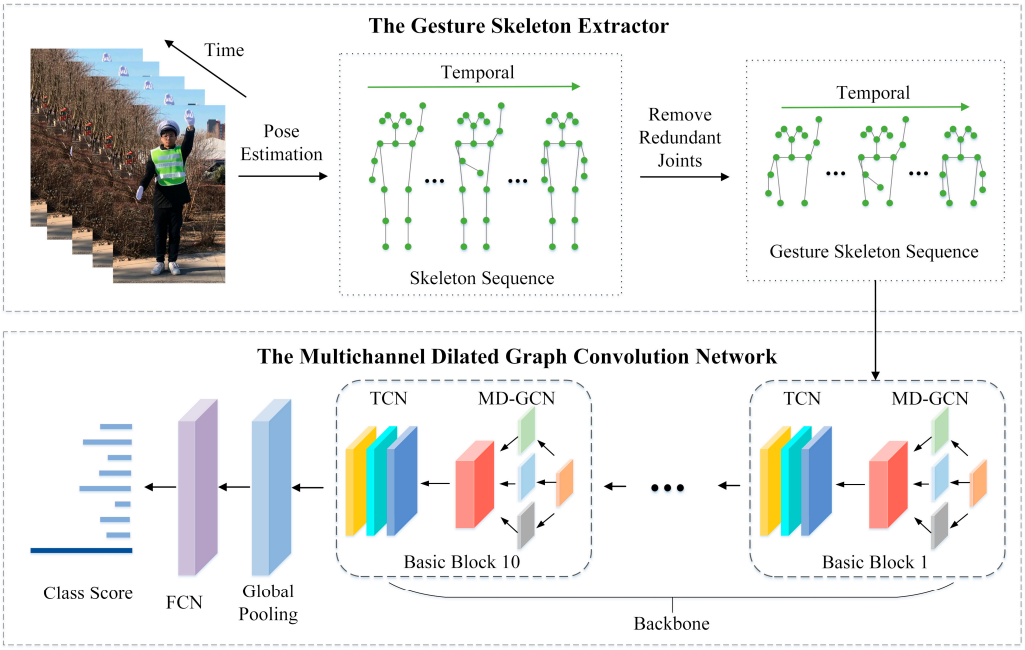

Figure 1. Overview of the proposed gesture recognition method. The FCN means the fully connected network, the TCN means the temporal convolution network.

4. Gesture Skeleton Extractor

Most existing traffic police gesture recognition methods extract pixel-level features from RGB images, which are uninterpretable because of the lack of gesture skeleton fea-

tures and may result in inaccurate recognition due to background noise. Such methods do not consider skeleton features and the influence of redundant skeleton parts. Moreover, extracted features are not adequately robust and result in poor recognition performance. To extract discriminative and interpretable gesture skeleton coordinate information, a GSE is proposed in this study which can extract skeleton coordinate information using the pose estimation method and remove redundant skeleton joints as well as bones. The GSE can improve skeleton data representation to strengthen the attention of the network. Gesture skeleton sequences can reduce the influence of the background and improve action recognition. Redundant joints, which are not useful or helpful for recognition, exist in skeleton data for traffic police gesture recognition. Redundant joints increase the computational cost of the network and reduce action recognition accuracy. The Police Gesture dataset contains eight Chinese Traffic Police Gestures, namely (1) stop, (2) move straight, (3) left turn, (4) left turn waiting, (5) right turn, (6) lane changing, (7) slow down, and (8) pull over. Actions such as arm movements and head turning mainly involve traffic police officers' upper limbs, whereas their lower limbs move very slightly. The head turning and moving is a part of the standard traffic police gesture. Hence, it is necessary and reasonable to keep the head of the traffic police officer in the analysis. The joint features of lower limbs are not helpful for recognition but influence action features. Thus, the gesture classification network must pay attention only to the changes in the upper body. A GSE is proposed to improve the skeleton sequence which can strengthen feature interpretation and the attention of the network. Figure 2 shows a gesture skeleton sequence that is processed by the GSE. For a skeleton sequence obtained by the OpenPose algorithm, the GSE removes the left and right knee joints and left and right ankle joints of each police skeleton in the sequence, shown in Figure 2a, and simultaneously removes the bones associated with these joints. Finally, the skeleton sequence in Figure 2a is transformed into the gesture skeleton sequence, as shown in Figure 2b. The number of police skeleton joints is reduced from 18 to 14, and the number of police skeleton bones is reduced by four. Therefore, the GSE eliminates the background, strengthens the attention of the network on the upper body of the traffic police skeleton, and reduces the input scale of the skeleton sequence data. The experiments show that the GSE can improve traffic police gesture recognition performance.

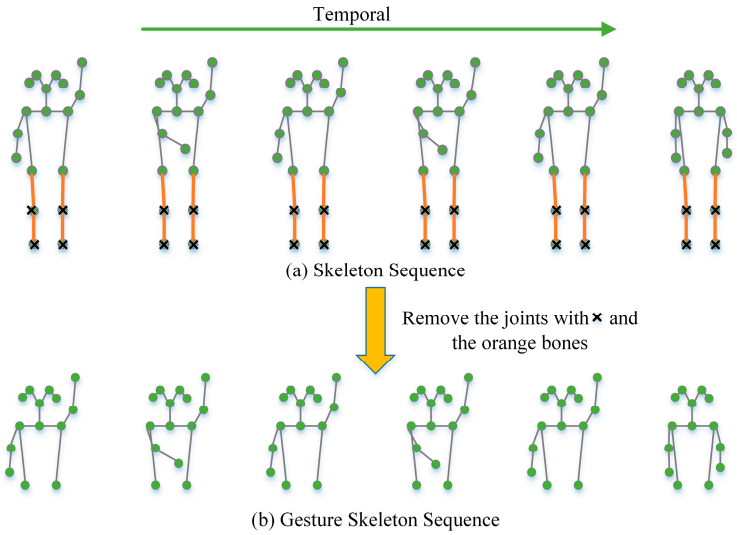

Figure 2. A skeleton sequence is transformed into a gesture skeleton sequence.

5. Multichannel Dilated Graph Convolution Network

Existing deep learning methods are not suitable for handling gesture skeleton features, which neglect the inevitable connection between skeleton joint coordinate infor-

mation and gestures. The shortcomings result in the inaccurate extraction of gesture features and poor recognition performance. To alleviate these problems, in the gesture discrimination stage, GSE-based features are introduced into a new framework called the MD-GCN. The proposed MD-GCN constructs a graph convolution with a multichannel dilated to enlarge the receptive field, which extracts body topological and spatiotemporal action features from skeleton coordinates. The MD-GCN improves the fixed adjacent matrix in the multichannel dilated connection, which can enrich the representation of gesture features and make features interpretable. The gesture skeleton sequence has a natural graph structure, and the graph convolution demonstrates superior performance in processing graph data. Existing GCN methods adopt the graph convolution theory proposed by Kipf et al. [12]. In a human skeleton graph, $\mathcal{G} = (\mathcal{V}, \mathrm{A}, \mathrm{X})$, where $\mathcal{V} = \{\mathrm{v_i}\}_{i=1}^{\mathcal{N}}$ is a set of $\mathcal{N}$ nodes representing skeletal joints; $\mathrm{A} \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ is an adjacency matrix, which represents the connection between the joints in the skeleton; and $\mathrm{X} = [\mathrm{x_{1,}} \mathrm{x_{2,}} \ldots, \mathrm{x_{N,}}] \in \mathbb{R}^{\mathcal{N}\times\mathcal{D}}$ is a feature matrix, which contains the $\mathcal{D}$-dimensional feature vectors of the $\mathcal{N}$ nodes in set $\mathcal{V}$. For a dynamic human skeleton sequence of length T, the layer-wise update formula for a single skeleton at the t-th time is used to extract spatial features as per Formula (1):

$$\mathrm{X_t^{(l+1)}} = \sigma\left(\overline{\mathrm{D}}^{-1}\overline{\mathrm{A}}\,\overline{\mathrm{D}}^{-1}\mathrm{X_t^{(l+1)}}\mathrm{W^l}\right) \quad (1)$$

where l represents the number of convolution layers, $\overline{\mathrm{A}} = \mathrm{A} + \mathrm{I}$ is an undirected skeleton graph adjacency matrix with self-connection, $\overline{\mathrm{D}}$ is the diagonal degree matrix of $\overline{\mathrm{A}}$ and $\overline{\mathrm{D}}^{-1}\overline{\mathrm{A}}\,\overline{\mathrm{D}}^{-1}$ is the normalized adjacency matrix of $\overline{\mathrm{A}}$, and W is a learnable weight matrix.

However, the existing GCN extracts features with fixed adjacent matrix representation. Therefore, the GCN samples only the information of the nearest neighbor nodes, thereby ignoring the information of the k-hop nearest neighbor nodes. To solve the issue mentioned above, the MD-GCN is proposed in the gesture discrimination stage, which can improve the fixed adjacent matrix in the multichannel dilated connection. The MD-GCN extracts features through K paths. Different dilated graph convolutions are deployed on the K paths to extract the features of different channels. Next, the MD-GCN fuses the different channel features from the K paths. This process can enrich the representation of gesture features and make features interpretable.

5.1. MD-GCN Operator

As shown in Figure 3a–c, there are 3×3 convolution kernels with dilated rates of 1, 2, and 3 extract pixel-level features of different scales. A dilated convolution [30], which is successful in semantic segmentation tasks, is proposed to aggregate the context information. We expanded it to a GCN and propose a dilated graph convolution operator to extract different scale features in the graph structured data. The dilated graph convolution operator alleviates the issue of the graph convolution extracting only single-scale information from the nearest node. Figure 3d,c,e show that a 1-dilated graph convolution, 2-dilated graph convolution, and 3-dilated graph convolution extract the information of the 1-hop neighbor nodes, 2-hop nearest neighbor nodes, and 3-hop nearest neighbor nodes, respectively. The k-dilated graph convolution denotes that the feature of each node in the next-layer graph is obtained by sampling and aggregating the features of the k-hop nearest neighbor nodes and self-connection.

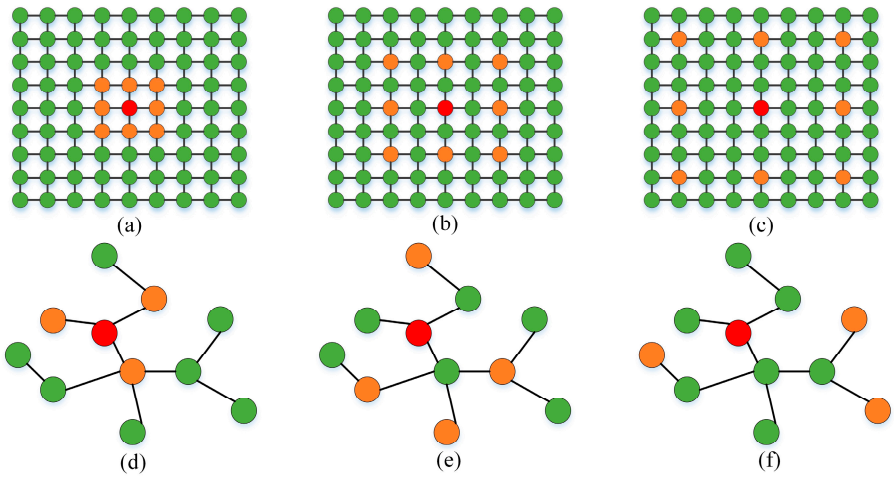

Figure 3. The red node means the central node. The orange node represents the node aggregated. (**a**), (**b**), (**c**): 3 × 3 dilated convolution, with dilated rates of 1, 2, and 3 from left to right; (**d**), (**e**), (**f**): k-dilated graph convolution, with k values of 1, 2, and 3 from left to right. The dilated graph convolution enlarges the receptive field.

However, a single k-dilated graph convolution extracts only single-scale features with the k-hop adjacency matrix and leads to missing representative features. To learn additional representative gesture features, a multichannel graph convolution operator with a dilated strategy is proposed to extract the features of different channels with different k values. As shown in Figure 4, different dilated graph convolutions are deployed on the K paths to extract the features of different scales from the previous layer. Next, the features from the different channels obtained by the different dilated graph convolutions are fused. For a graph, the layer-wise update formula is used to extract spatial features as per Formula (2):

$$X_t^{(l+1)} = \sigma\left(\sum \bar{D}^{-1}\bar{A}_k\bar{D}^{-1}X_t^l W_k^l\right)\ k \in \text{ScaleSet} \tag{2}$$

where l represents the number of layers, $\bar{A} = A + I$ is an undirected k-hop graph adjacency matrix with self-connection, $\bar{D}$ is the diagonal degree matrix of $\bar{A}$ and $\bar{D}^{-1}\bar{A}\,\bar{D}^{-1}$ is the normalized adjacency matrix of $\bar{A}$, W is a learnable weight matrix, and ScaleSet denotes the fusion scale set.

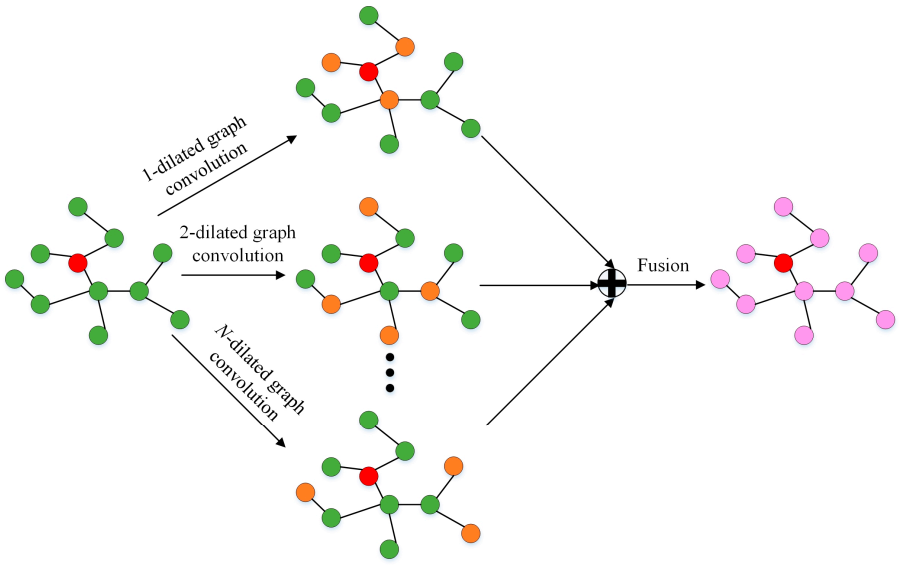

Figure 4. The process of the multichannel dilated graph convolution network (MD-GCN) operator. The leftmost graph is the original graph and the rightmost is a new graph processed by different dilated convolution operations and fusion.

5.2. MD-GCN Classification Network

A GCN-based method is proposed to extract skeleton joint coordinate information features. However, the existing GCN method for action recognition extracts single-scale features with fixed adjacent matrix representation, which ignores the different scale representations of a gesture skeleton graph. The shortcoming results in the inaccurate extraction of gesture features and poor recognition performance. In the gesture discrimination stage, GSE-based features are introduced into a new framework called the MD-GCN, which constructs a graph convolution with a multichannel dilated to extract body topological and spatiotemporal action features from skeleton coordinates. The dilated graph convolution enlarges the receptive field. The MD-GCN captures the spatial relationships in k-hop nearest neighbor nodes and extracts spatiotemporal and multichannel features from a skeleton sequence.

The architecture of the MD-GCN classification network is illustrated in Figure 5. The input of the network is a gesture skeleton sequence that can be represented by a $C \times T \times V$ tensor, where T denotes the sequence length or number of video frames, V represents the number of nodes in each skeleton graph, and C is the dimension of each node. In the traffic police gesture recognition task, a skeleton sequence without the GSE is a $3 \times 80 \times 18$ tensor, but a skeleton sequence with the GSE is a $3 \times 80 \times 14$ tensor. The backbone of the entire network is composed of 10 basic blocks (B1 − B10) followed by a global pooling layer, a fully connected layer, and a SoftMax layer. The basic blocks extract body topological and spatiotemporal action features from skeleton coordinates. Figure 6 shows that the basic blocks in the network are composed of two parts. The left part is used to extract interpretable human topology and spatial gesture features, and the right part is used to extract temporal gesture features. Each basic block contains an MD-GCN operator and a temporal convolution network (TCN). Each MD-GCN operator contains multichannel dilated graph convolution. The MD-GCN operator mentioned above is deployed on the left part to extract multichannel features, and the orange convolution, gray convolution, and light blue convolution represent the three types of k-dilated graph convolutions of different scales in Figure 6 (k = 1, 2, and 4 in our method). The MD-GCN operator is followed by a batch normalization (BN) layer and ReLU layer. In the right part, the gold Convt with a $t \times 1$ convolution kernel (t = 9 in the experiment) extracts temporal features in the dimension of T and is also followed by a BN layer and ReLU layer. As for the extraction of temporal information, the TCN in Figure 6 uses the convolution kernel of (T × 1) to aggregate features from the corresponding joints in continuous sequences. Then, it realizes feature aggregation on the temporal dimension. As for the LSTM, it can only process one frame at a time, which means that it is difficult to achieve a parallel computing for LSTM. Generally, the training time and the efficiency of LSTM-based methods are longer and lower. Compared with LSTM, our TCN can process multiple skeleton sequences at a time. Owing to the feature of the MD-GCN with different dilated rates having the same dimension which is C × T × V, the fusion operator just adds the features up for fusion.

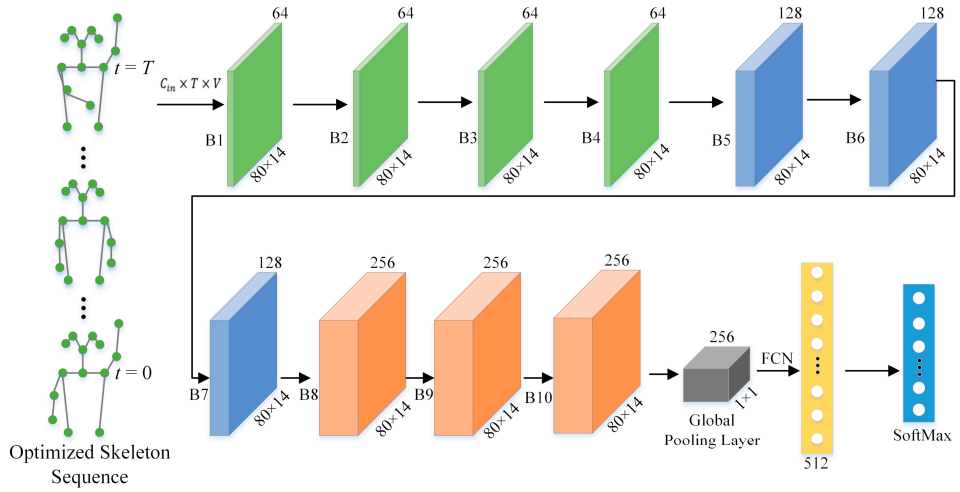

Figure 5. The architecture of the MD-GCN classification network.

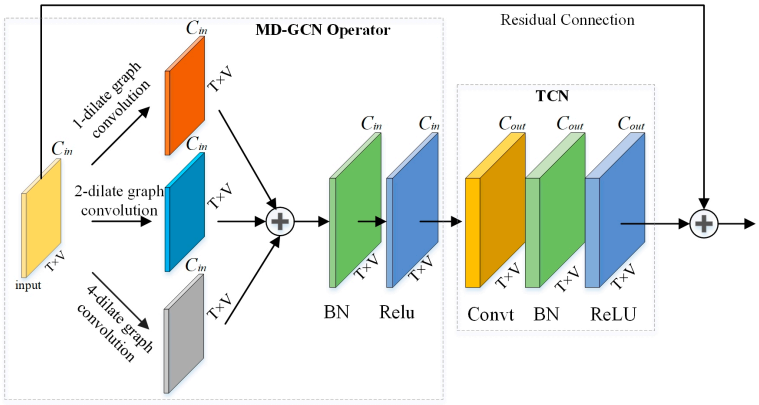

Figure 6. The basic block of the MD-GCN classification network.

6. Experiments

6.1. Dataset and Training Details

Figure 7 shows that the Chinese traffic police gestures are composed of eight types of gestures, namely: (1) stop, (2) move straight, (3) left turn, (4) left turn waiting, (5) right turn, (6) lane changing, (7) slow down, and (8) pull over. A "stand in attention" gesture was added to expand the Police Gesture Dataset [9]. The "stand in attention" gesture means that the police did not change the current traffic situation. The Police Gesture Dataset includes 20 videos of traffic police making gestures. Each video is approximately 5 to 11 minutes long, and a total of 3354 gestures are shown in the videos. Among them, 11 videos in the training set show 1789 gestures and 9 videos in the testing set show 1565 gestures. The gesture dataset provides only the 20 raw videos without skeleton data and ground truth gesture labels for each video frame. Thus, the videos were cut according to the labels to obtain 3354 video clips of traffic police gestures. Next, the skeleton data of the video clips were extracted by the OpenPose [24] algorithm. The experiments were implemented on the Pytorch deep learning framework, and the GPU was a Quadro RTX 4000 with 8G of memory (NVIDIA, Santa Clara, CA, USA). All the MD-GCN models were trained with the same batch size (40), training epochs (60), and optimizer (SGD with Nesterov momentum 0.9). In the models, the basic learning rate was set to 0.1. The basic learning rate was used in the first 20 epochs and set to 0.01 in epochs 20 to 40, then divided by 10 after every 10 epochs between 40 and 60 epochs. Cross-entropy was chosen as the backpropagation loss function. Two challenging action recognition datasets are experimented in this work, NTU-RGB+D [31] and Kinetics [32]. The NTU-RGB+D

dataset contains 60 classes with more than 56,000 video samples. It contains 56,880 sample actions and the provided annotations of 3D joint locations were used in the experiment. The dataset is established by two benchmarks, cross-subject (CS) and cross-view (CV). The Kinetics dataset has 400 action classes and more than 400 video clips for each action. The released Kinetics-Skeleton data are used in this work, which contain a training set with 240,000 clips and a validation set with 20,000 clips. The same data process as reference [14] is what we used for the two datasets.

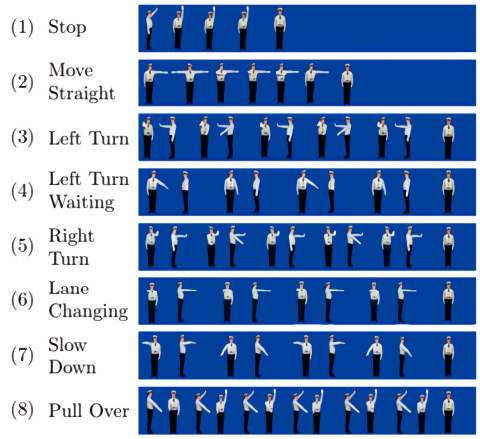

Figure 7. The eight gestures of Chinese traffic police [9].

6.2. Evaluation of the Proposed Method

A traffic police gesture recognition method based on a GSE and the MD-GCN is proposed in this study. To evaluate the effectiveness of the proposed method, experiments were conducted on the Police Gesture Dataset. Appropriate and effective features representing traffic police gestures should be chosen and extracted. However, traffic police typically work in complex and unpredictable environments, which can introduce interference and render features uninterpretable. A GSE is proposed to extract gesture skeleton sequences which contain discriminative and interpretable skeleton coordinate information. Compared with extracted pixel-level features, gesture skeleton sequences can eliminate background interference and make features interpretable through coordinates. To verify the performance of the basic blocks with different MD-GCN operators in gesture recognition in the network, an MD-GCN operator with dilated graph convolutions with different dilated rates was deployed for comparison. The results in Table 1 show that the MD-GCN operator with fused dilated rates of 1, 2, and 4 demonstrates the most impressive performance in the Police Gesture Dataset. The MD-GCN network with basic blocks (dilated rate = 2, 3, and 4) exhibits the worse recognition accuracy (96.71%), because the model misses the 1-dilated graph convolution, which contains the features of the nearest neighbor nodes. However, although the MD-GCN (dilated rate = 1, 2, 3, and 4) fuses the features of more channels, recognition performance declines. The more aggregated channels there are, the higher the complexity of the model and the time. The GCN aggregates features for classifying, and the more information the network aggregates, the easier it is for nodes to be consistent. We conclude that unlimited multichannel features result in problems of overfitting and oversmoothing.

Table 1. The accuracy comparison of MD-GCN with different operators.

MD-GCN Operator	Accuracy (%)	Model Size	Training Time (s)	Inference Time (s)
With dilated rate = 1	97.32	10 M	532	2.5
With dilated rate = 2	96.35	10 M	537	2.8
With dilated rate = 3	96.13	10 M	534	2.5
With dilated rate = 4	95.38	10 M	528	2.3
With dilated rate = 1, 2, 3	98.05	11.7 M	603	3.0
With dilated rate = 1, 2, 4	98.95	11.7 M	602	3.0
With dilated rate = 1, 3, 4	97.76	11.7 M	620	3.3
With dilated rate = 2, 3, 4	96.71	11.7 M	617	3.1
With dilated rate = 1, 2, 3, 4	98.08	12.5 M	718	3.9

To prove the effectiveness of the GSE, experiments were conducted on the Police Gesture Dataset. As shown in Table 2, with the same MD-GCN operator configuration (dilated rate = 1, 2, and 4), the network with the GSE demonstrates better recognition performance than the network without the GSE. With GSE means the network using gesture skeleton sequences which remove the redundant lower limbs as the input. Without GSE means the skeleton sequence as the input. Experiments show that the lower limbs lead to a lower accuracy. As the results show, ignoring the lower limbs would lead to a satisfactory accuracy and efficiency. The main reason is that the lower limb movements in different public traffic gestures are similar, so the lower limb movements are a redundant and meaningless feature for recognition.

Table 2. The accuracy and time assessments of gesture skeleton extractor (GSE).

Method	Accuracy (%)	Training Time (s)	Inference Time (s)
With GSE	98.95	603	3.0
Without GSE	98.01	691	3.3

A new graph convolution network, namely the MD-GCN, is proposed to solve the traffic police gesture task in this study. The MD-GCN constructs a graph convolution with a multichannel dilated to enlarge the receptive field, which extracts body topological and spatiotemporal action features from skeleton coordinates. The MD-GCN captures the spatial relationships in k-hop nearest neighbor nodes and extracts spatiotemporal and multichannel features from a skeleton sequence. Compared with a typical GCN [13], the MD-GCN demonstrates better performance for traffic police gestures. Various experiments demonstrated the effectiveness of the MD-GCN. Table 1 shows the experiment results comparing the accuracy of the MD-GCN with that of a typical GCN (dilated rate = 1). The accuracy of the MD-GCN (dilated rate = 1, 2, and 4) is 1.63% higher than that of the GCN.

6.3. Comparison with State-of-the-art Methods

The proposed method was compared with several typical network architectures for the police gesture recognition task, and the results are shown in Table 3. Among the methods, the proposed method demonstrates the best performance in the Police Gesture Dataset. Hara et al. [33] and Tran et al. [34] used three-dimensional (3D) convolution to extract temporal and spatial features from RGB videos simultaneously. The 3D networks involve huge computational costs and extract uninterpretable features. Qiu et al. [35] proposed the P3D network, suggesting that a 3D convolution can be divided into a 2D convolution and one-dimensional convolution for action recognition. Reference [33–35] fine-tuned the Police Gesture Dataset. The batch size is set as four in training. SGD is utilized to converge the model, where the learning rate and momentum are set as 0.001 and 0.9, respectively. The decay period is set as 10, and the decay parameter is set as

0.005. Finally, the cross-entropy function is utilized as the loss function. Although this network reduces parameters and increases speed, it uses RGB videos as input; thus, background noise will considerably affect the performance of the network. Similarly, the convolutional LSTM of [36] uses image features directly, which leads to a low accuracy rate. The aforementioned methods show that eliminating the influence of the background and illumination are important for traffic police gesture recognition. Zhang et al. [9] extracted skeleton data from videos using the Police Keypoint Extraction Network and then converted the skeletal data into a sequence of handcrafted features. Subsequently, the authors used the sequence of handcrafted features to distinguish police gestures in an LSTM network or a bidirectional LSTM network [37]. This method requires the designing of handcrafted features, but handcrafted features can lose some important skeleton information. For example, in a traffic police gesture action, spatial dependence exists between the moving joints, but handcrafted features will destroy this information. Compared with the other methods, the proposed method utilizes the gesture skeleton sequences to avoid background noise interference and redundant joints. In the gesture discrimination stage, GSE-based features are introduced into the new MD-GCN framework, which constructs a graph convolution with a multichannel dilated to extract body topological and spatiotemporal action features from skeleton coordinates. The MD-GCN improves the fixed adjacent matrix in the multichannel dilated connection, which can enrich the representation of gesture features and make features interpretable. As shown in the confusion matrix in Figure 8, the proposed method demonstrates excellent performance in gesture action recognition. The proposed method can recognize the gestures of traffic police for most cases. In rare cases, due to the differences and changes in the angle of view, the "left turn waiting" might be misclassified as "stop" and "slow down" might be misclassified as "left turn". Further research to improve the robustness of the angle of view and strengthen the temporal feature will be conducted in the future to solve this issue.

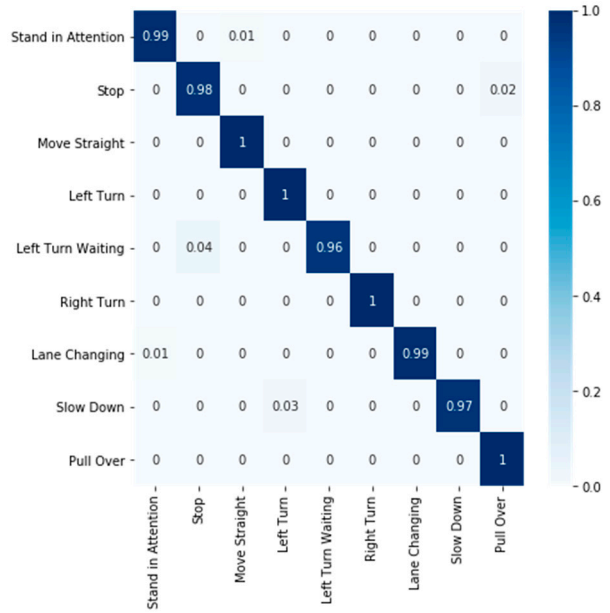

Figure 8. The confusion matrix of the proposed method on the Police Gesture Dataset.

Table 3. The accuracy comparison with some typical network architectures on the Chinese Traffic Police Gesture Dataset.

Method	Accuracy (%)
Hara [33]	81.16
Tran [34]	87.12
Qiu [35]	92.21
Xing [36]	82.40
Zhang [9]	91.18
Pigou [37]	91.04
The proposed method	98.95

In order to evaluate the proposed method conducted on complex datasets, the MD-GCN processed two challenging action recognition datasets: NTU RGB+D [31] and Kinetics [32]. The proposed method was also compared with some state-of-the-art methods in action recognition. The good performance of the MD-GCN in the task demonstrates its competitiveness and the proposed method can extract discriminative action feature. The accuracy results are shown in Tables 4 and 5.

Table 4. The accuracy performance with the NTU RGB+D dataset (%).

	CS	CV
Kim [38]	74.3	83.1
Ke [39]	79.6	84.8
Yan [14]	81.5	88.3
Tang [40]	83.5	89.8
Wen [41]	84.2	90.2
The proposed method	81.9	88.1

Table 5. The accuracy performance with the Kinetics dataset (%).

	Top-1	Top-2
Kim [38]	20.3	40.0
Yan [14]	30.7	52.8
Li [42]	34.8	56.5
The proposed method	30.8	53.3

7. Conclusions

In this work, a traffic police gesture recognition system based on a gesture skeleton extractor and a multichannel dilated graph convolution network was proposed in order to extract discriminative and interpretable gesture skeleton coordinate information. A GSE is proposed to extract skeleton coordinate information and remove redundant skeleton joints as well as bones. The GSE-based features are introduced into the proposed MD-GCN which constructs a graph convolution with multichannel dilated to extract features of body topological and spatiotemporal action from skeleton coordinates. The experiments showed that the proposed method has achieved the most advanced performance on the Police Gesture Dataset.

Author Contributions: Conceptualization, X.X., H.W. and W.M.; methodology, X.X., H.W. and W.M.; software, H.W., J.X., Q.F and C.P.; writing—original draft preparation, X.X., H.W., J.X., Q.F. and C.P.; writing—review and editing, X.X., W.M. and Q.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 62076117 and 61762061), the Natural Science Foundation of Jiangxi Province, China (Grant No. 20161ACB20004) and Jiangxi Key Laboratory of Smart City (Grant No. 20192BCD40002).

Acknowledgments: We gratefully acknowledge the assistance of Neurocomputing, 390, He J, Zhang C, He X, Dong R, Visual Recognition of traffic police gestures with convolutional pose machine and handcrafted features, 248–259.; Copyright (2020), with permission from Elsevier. The dataset was used in our experiments and the Figure 1 is reused as Figure 7 in this paper.

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

References

1. Yang, H.; Liu, L.; Min, W.; Yang, X.; Xiong, X. Driver Yawning Detection Based on Subtle Facial Action Recognition. *IEEE Trans. Multimed.* **2021**, *23*, 572–583, doi:10.1109/tmm.2020.2985536.
2. Zhou, L.; Min, W.; Lin, D.; Han, Q.; Liu, L. Detecting Motion Blurred Vehicle Logo in IoV Using Filter-DeblurGAN and VL-YOLO. *IEEE Technol*. **2020**, *69*, 3604–3614.
3. Xiong, X.; Min, W.; Zheng, W.S.; Liao, P.; Yang, H.; Wang, S. S3D-CNN: Skeleton-based 3D Consecutive-low-pooling Neu-ral Network for Fall Detection. *Appl. Intell.* **2020**, *50*, 3521–3534.
4. Sun, S.-W.; Liu, B.-Y.; Chang, P.-C. Deep Learning-Based Violin Bowing Action Recognition. *Sensors* **2020**, *20*, 5732, doi:10.3390/s20205732.
5. Li, F.; Li, J.; Zhu, A.; Xu, Y.; Yin, H.; Hua, G. Enhanced Spatial and Extended Temporal Graph Convolutional Network for Skeleton-Based Action Recognition. *Sensors* **2020**, *20*, 5260, doi:10.3390/s20185260.
6. Liu, Q.; Chen, E.; Gao, L.; Liang, C.; Liu, H. Energy-Guided Temporal Segmentation Network for Multimodal Human Action Recognition. *Sensors* **2020**, *20*, 4673, doi:10.3390/s20174673.
7. Tsai, J.-K.; Hsu, C.-C.; Wang, W.-Y.; Huang, S.-K. Deep Learning-Based Real-Time Multiple-Person Action Recognition System. *Sensors* **2020**, *20*, 4758, doi:10.3390/s20174758.
8. Ma, C.; Zhang, Y.; Wang, A.; Wang, Y.; Chen, G. Traffic Command Gesture Recognition for Virtual Urban Scenes Based on a Spatiotemporal Convolution Neural Network. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 37, doi:10.3390/ijgi7010037.
9. He, J.; Zhang, C.; He, X.; Dong, R. Visual Recognition of traffic police gestures with convolutional pose machine and hand-crafted features. *Neurocomputing* **2020**, *390*, 248–259, doi:10.1016/j.neucom.2019.07.103.
10. Li, C.; Yang, S. Traffic Police Gesture Recognition for Autonomous Driving. In Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 7–10 December 2018; pp. 1413–1418.
11. Guan, W.; Ma, X. Traffic Police Gesture Recognition using RGB-D and Faster R-CNN. In Proceedings of the International Con-ference on Intelligent Informatics and Biomedical Sciences, Bangkok, Thailand, 21–24 October. 2018; Volume 3, pp. 78–81.
12. Hang, C.; Zhang, R.; Chen, Z.; Li, C.; Li, Z. Dynamic Gesture Recognition Method Based on Improved DTW Algorithm. In Proceedings of the 2017 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), Wuhan, China, 2–3 December 2017; Volume 1, pp. 71–74.
13. Kipf, T.N.; Welling, M. Semi-supervised Classification with Graph Convolutional Networks. Available online: https://openreview.net/forum?id=SJU4ayYgl (accessed on 22 February 2017).
14. 14. Yan, S.; Xiong, Y.; Lin, D. 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; arXiv:1801.07455.
15. Wu, J.; Zhong, S.-H.; Liu, Y. Dynamic graph convolutional network for multi-video summarization. *Pattern Recognit.* **2020**, *107*, 107382, doi:10.1016/j.patcog.2020.107382.
16. Yang, L.; Guo, Y.; Gu, J.; Jin, D.; Yang, B.; Cao, X. Probabilistic Graph Convolutional Network via Topology-Constrained Latent Space Model. *IEEE Trans. Cybern.* **2021**, 1–14, doi:10.1109/tcyb.2020.3005938.
17. Cui, Z.; Henrickson, K.; Ke, R.; Wang, Y. Traffic Graph Convolutional Recurrent Neural Network: A Deep Learning Frame-work for Network-Scale Traffic Learning and Forecasting. *IEEE Intell. Transp*. **2020**, *21*, 4883–4894.
18. Wang, B.; Yuan, T. Traffic Police Gesture Recognition using Accelerometer. In Proceedings of the IEEE Sensors Conference, Lecce, Italy, 26–29 October. 2008; pp. 1080–1083.
19. Tao, Y.; Ben, W. Accelerometer-based Chinese Traffic Police Gesture Recognition System. *Chin. J. Electron*. **2010**, *19*, 270–274.
20. Le, Q.K.; Pham, C.H.; Le, T.H. Road Traffic Control Gesture Recognition using Depth Images. *IEIE Trans. Smart Process. Comput.* **2020**, *1*, 1–7.
21. Guo, F.; Tang, J.; Wang, X. Gesture recognition of traffic police based on static and dynamic descriptor fusion. *Multimed. Tools Appl.* **2017**, *76*, 8915–8936, doi:10.1007/s11042-016-3497-9.
22. Guo, F.; Tang, J.; Cai, Z. Automatic Recognition of Chinese Traffic Police Gesture Based on Max-Covering Scheme. *Int. J. Adv. Inf. Sci. Serv. Sci.* **2013**, *5*, 428–436, doi:10.4156/aiss.vol5.issue1.53.
23. Cai, Z.; Guo, F. Max-covering scheme for gesture recognition of Chinese traffic police. *Pattern Anal. Appl.* **2014**, *18*, 403–418, doi:10.1007/s10044-014-0383-9.
24. Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1302–1310.
25. Wei, S.-E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional Pose Machines. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; Volume 1, pp. 4724–4732.

26. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded Pyramid Network for Multi-person Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103-7112.
27. Su, K.; Yu, D.; Xu, Z.; Geng, X.; Wang, C. Multi-Person Pose Estimation with Enhanced Channel-Wise and Spatial Information. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 9 May 2019; pp. 5667–5675.
28. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5686–5696
29. Nie, X.; Feng, J.; Zhang, J.; Yan, S. Single-Stage Multi-Person Pose Machines. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019; pp. 6950–6959.
30. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, PR, USA, 2–4 May 2016.
31. Shahroudy, A.; Liu, J.; Ng, T.-T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
32. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The Kinetics Human Action Video Dataset; Available online: https://arxiv.org/abs/1705.06950 (accessed on 19 May 2017).
33. Hara, K.; Kataoka, H.; Satoh, Y. Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 3154–3160.
34. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Net-works. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489-4497.
35. Qiu, Z.; Yao, T.; Mei, T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5534–5542.
36. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K. Convolutional Lstm Network: A Machine Learning Approach for Precipitation Nowcasting. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, Quebec, Canada, 7–12 December 2015; pp. 802–810.
37. Pigou, L.; Oord, A.V.D.; Dieleman, S.; Van Herreweghe, M.; Dambre, J. Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. *Int. J. Comput. Vis.* **2018**, *126*, 430–439, doi:10.1007/s11263-016-0957-7.
38. Kim, T.S.; Reiter, A. Interpretable 3D Human Action Analysis with Temporal Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1623–1631.
39. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A New Representation of Skeleton Sequences for 3D Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4570–4579.
40. 40. Tang, Y.; Tian, Y.; Lu, J.; Li, P.; Zhou, J. Deep Progressive Reinforcement Learning for Skeleton-Based Action Recognition. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5323–5332.
41. Wen, Y.-H.; Gao, L.; Fu, H.; Zhang, F.-L.; Xia, S. Graph CNNs with Motif and Variable Temporal Block for Skeleton-Based Action Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence; Association for the Advancement of Artificial Intelligence (AAAI), Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8989–8996.
42. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-Structural Graph Convolutional Networks for Skele-ton-based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3590–3598.