WEP Warwick Evans Publishing

Transactions on Engineering and Technology Research
ISSN: 2960-1797, eISSN: 2960-2246 | Volume 3
ICMEIT 2024

A drill pipe counting method based on skeleton action recognition

Jingyi Du [1], Haohao Chen [1, *], Rui Gao [1]

[1] College of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China

*1275551093@qq.com

Abstract. A drill pipe counting method based on human skeleton sequence recognition is proposed to address the issues of high workload, high environmental requirements, and large counting errors in existing drill pipe counting methods. Firstly, obtain human skeleton information through YOLOv7 and FastPose, and create a human skeleton dataset; Secondly, based on ST-GCN, an ML-ASTGNet action recognition network is designed, which captures more global contextual information through adaptive graph convolution (A-GCN) and improves the spatial modeling ability of the action recognition network; Introducing the Time Motion Excitation Module (TME), which utilizes the idea of time difference to characterize motion information, generates larger weights for frames with larger motion amplitudes to highlight sensitive features of motion, and designs a Multi Scale Multi Fine Granular Time Convolution (DM-TCN) to learn the final time feature information from different scales, improving the time modeling ability of action recognition networks. The experimental results show that the accuracy of ML-ASTGNet on the self built human skeleton dataset reaches 92.1%, which is 8.2% higher than ST-GCN and shows better action recognition performance. Finally, this method has achieved good counting performance in actual drill pipe counting tests with small counting errors.

Keywords: Action recognition; Drill pipe count; Human skeleton; Spatial-temporal graph convolution.

1. Introduction

Gas is the first hidden danger to the safety of coal mines underground, and it is the main disaster that causes major and above accidents in coal mines. Preventing gas exceeding limits is a key link and important means to curb gas accidents [1, 3]. At present, drilling and extraction are generally used underground in coal mines to reduce gas concentration and reduce the occurrence of gas disasters. The measurement of drilling depth directly affects the quality of drilling operations, and the drilling depth is directly proportional to the number of drilling operations. Therefore, the accuracy of drill pipe counting becomes a crucial factor in drilling operations.

With the rapid development of computer vision, artificial intelligence technologies represented by neural networks have been widely applied underground. Chaoxiu Yao et al [4] used object detection algorithm (YOLOv5) to detect workers, drill pipes, and drilling rigs in videos, generating candidate regions for the three types of targets mentioned above. The trend of unloading drill pipes depends on whether the candidate areas for workers and drill pipes overlap. If there is overlap, the count is determined by whether the candidate areas of the drill pipe and drilling rig are separated. It is unreasonable to judge whether there is a unloading trend based on whether the two candidate areas overlap due to the obstruction of workers to the drill pipe during the unloading process. Rui Gao et al [5] proposed a downhole drill pipe counting method based on an improved ResNet network. This method identifies the unloading and non unloading behaviors of workers by establishing an image classification network, and integrates the recognition results to obtain a confidence curve. Finally, determine the number of drill rods based on the number of falling edges. This method only considers the spatial features of the image and does not take into account the temporal features of the image, which may lead to counting errors.

As one of the current research hotspots, action recognition based on graph convolution has shown that human skeleton action recognition based on graph convolution is a more effective way[6]. Jingyi Du et al [7] introduced the spatial-temporal graph convolutional action recognition network into the

drill pipe counting method. This method uses the ST-GCN [8] network to classify the loading and unloading drill pipe actions of workers and achieve drill pipe counting. Although this method achieves insensitivity to ambient light, the network structure is simple and the spatial-temporal modeling ability is insufficient. It cannot fully explore the temporal and spatial connections of human behavior, and is easily influenced by irrelevant behaviors, leading to misidentification or misidentification of behavior.

In response to the above issues, the paper proposes a drill pipe counting method based on a multi-level adaptive spatial-temporal graph convolutional network. Firstly, the human body is detected using the object detection algorithm YOLOV7 [9] . Secondly, FastPost [10] is used to obtain the key point information of the human body and concatenate it into a sequence of human skeleton. Finally, the human skeleton information is passed into a multi-level adaptive spatial-temporal graph convolutional network (ML-ASTGNet). The final fully connected layer of the network classifies the loading and unloading drill pipe actions, achieving drill pipe counting. The experiment shows that the drill pipe counting method greatly overcomes the interference of the actual underground environment, has good action recognition ability, and improves the accuracy of drill pipe counting.

2. Multi-level adaptive spatial-temporal graph convolutional network

2.1 Overall network structure

As a natural graph data structure, ST-GCN has the ability to model the human skeleton in both temporal and spatial dimensions simultaneously. Unlike action recognition models such as CNN and RNN, the ST-GCN action recognition model can reduce the influence of factors such as human appearance clothing, background lighting, and changes in shooting angles. However, the original ST-GCN action recognition network lacks spatial and temporal feature extraction capabilities, resulting in poor action recognition performance.

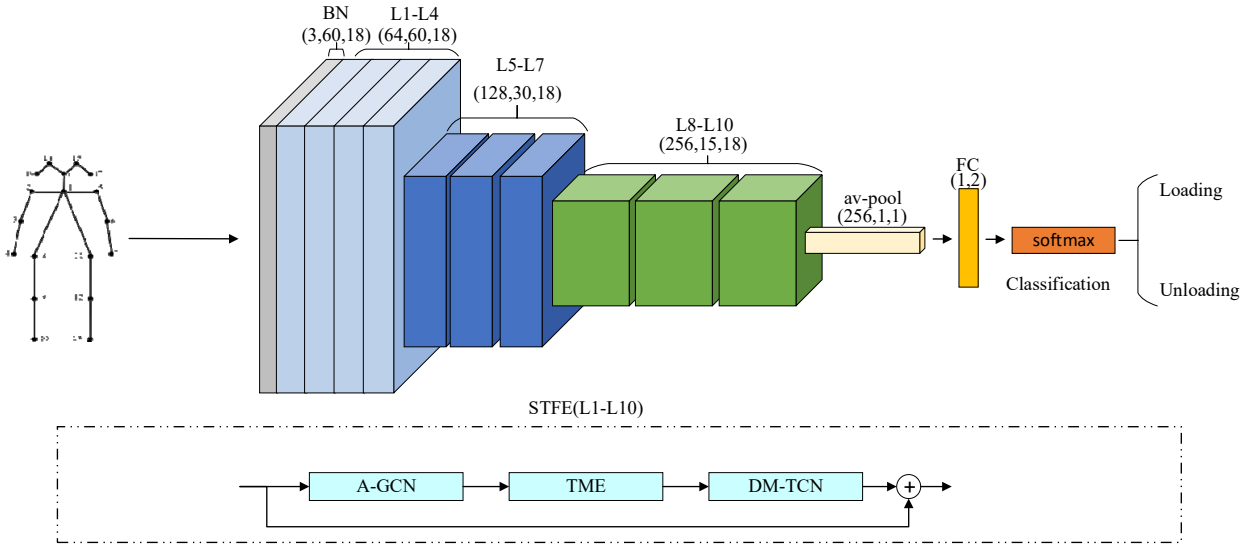

Figure 1. Structure of ML-ASTGNet network

The paper proposes a new ML-ASTGNet action recognition network based on the ST-GCN network, as shown in Fig. 1 It consists of a batch normalization layer (BN), three sets of ten a spatial-temporal feature extraction module(STFE), an average pooling layer(av pool), a fully connected layer (FC), and a softmax layer. Among them, BatchNormal is used to normalize skeleton data and improve the network's generalization ability; A spatial-temporal l feature extraction unit mainly consists of three parts: A-GCN, TME, and DM-TCN, which are sequentially connected, and residual connections are introduced to stabilize training. Among them, A-GCN is an adaptive graph convolution module used to comprehensively capture the spatial information of the human skeleton; TME is a motion time excitation module [11], used to highlight sensitive features of motion, and combined with multi-scale

multi fine-grained time convolution (DM-TCN) for time dimension modeling, to improve the network's information extraction ability in the time dimension. Strengthening the spatial-temporal modeling of skeleton data through the spatial-temporal feature extraction module to extract effective spatial-temporal feature information from human skeleton sequence data and improve the accuracy of action recognition networks; Reduce the computational complexity of the model by averaging pooling layers; Classify actions through fully connected layers and softmax functions.

2.2 Spatial feature extraction module

ST-GCN uses predefined skeleton maps, which only represent the natural structure of the human body connected by joints. In complex action recognition tasks, this fixed skeleton map is not optimal. For example, when workers are loading and unloading drill pipes, the correlation between their left and right arms and their waist should be stronger, indicating that the graphic structure should be related to the data rather than fixed and unchanging. Secondly, in human skeleton data, the connection relationship and weight between nodes may change with time or spatial position, and this dynamism is crucial for modeling spatial-temporal relationships. In response to the above issues, the paper introduces adaptive graph convolution on the basis of ST-GCN, parameterizes the graph and participates in the training and updating of convolutional parameters, improves the accuracy of action recognition networks, and reduces misidentification of actions. The network structure of A-GCN is shown in Fig. 2, with orange modules indicating learnable parameters. The principle is shown in equation (1) and equation (2).

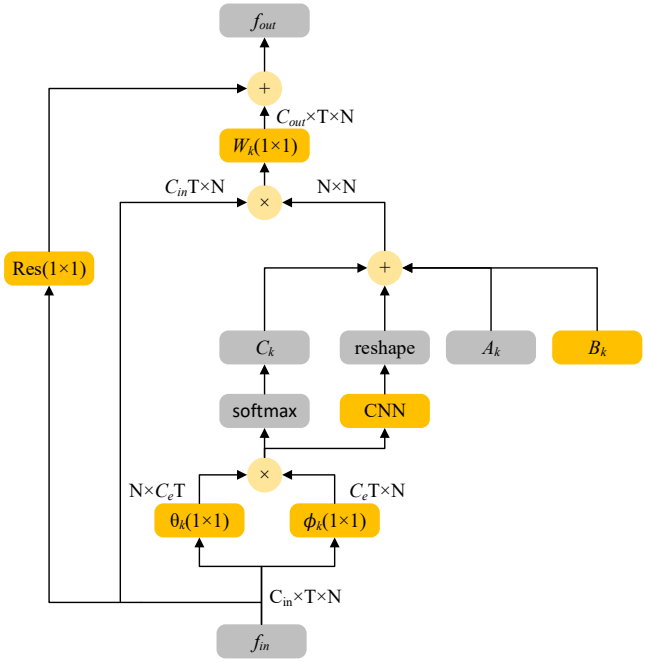

Figure 2. Structure of Adaptive Graph Convolutional Network

$$f_{out} = \sum_k^{K_v} W_k f_{in} (A_k + B_k + C_k + CNN) \tag{1}$$

$$A_k = \Lambda_k^{-\frac{1}{2}} \overline{A}_k \Lambda_k^{-\frac{1}{2}} \tag{2}$$

In the above two equations, f_{in} is the input of graph convolution, f_{out} is the output of graph convolution, K_v is the region partitioning strategy, and ST-GCN divides the region into three parts: the root node set, the centripetal point set, and the centrifugal point set (K_v=3). k represents the k-th region, and $\overline{A}_k$ is an N-order matrix formed by adding the adjacency matrix A and the identity matrix

I, which can represent the natural connections of various joints in the human body in a single frame.B_k also represents the adjacency matrix, but the values in B_k are trained and optimized along with other parameters, completely learned from data training. The model can learn autonomously based on specific tasks, and the learned parameters can not only represent node connectivity, but also represent strength. C_k is a data dependent matrix that learns a unique graph for each sample. It uses a normalized embedded Gaussian function to calculate the similarity between two vertices and determine whether there will be a connection relationship and strength between them. The effect is similar to B_k. The CNN branch utilizes convolution to aggregate features from different channels and capture global contextual information between joints.

2.3 Time feature extraction module

Time modeling is a key issue in action recognition. A high-performance spatial-temporal graph convolutional network not only requires a GCN module with strong spatial information extraction capabilities, but also a robust temporal feature extraction module. In the ST-GCN action recognition network, its TCN module only uses simple 2D temporal convolution, and its ability to extract information in the temporal dimension is limited. In response to this issue, the paper introduces a Time Motion Excitation (TME) module and designs a Multi Scale Multi Fine Granular Time Convolution (DM-TCN), which combines the two to solve the problem of insufficient time modeling ability.

Real motion is the measurement of displacement between two consecutive frames. The TME module calculates the feature differences between adjacent frames and uses these motion feature differences to modulate weights. This assigns larger weights to frames with larger motion amplitudes, highlighting sensitive motion information. The network structure of the TME module is shown in Fig. 3. Firstly, a 1*1 convolution is used to reduce the number of channels and improve computational efficiency. Next, calculate the feature difference between adjacent frames and stack the feature difference to restore T steps. After global pooling, restore the number of channels through 1*1 convolution, and finally output the weight score through sigmoid activation function.

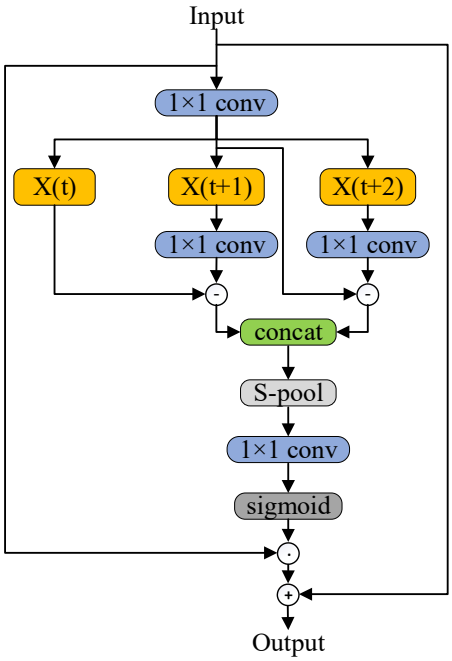

Figure 3. Structure of TME module network

Simple 2D time convolution is far from sufficient to extract the final time information. If multiple dilated convolutions are used to aggregate time information of different scales, replacing the time convolution in the original ST-GCN, but as the dilation scale increases, the ability of time convolution to capture detailed information decreases. To address this issue, the paper designs a multi-scale multi

fine-grained time convolution (DM-TCN), which will learn the final time feature information from different scales and fine-grained, improving the long-term modeling ability of the network. The structure of the DM-TCN network is shown in Fig. 4. By stacking four sets of dilated convolutions with different dilated rates, pooling a single global average, and a multi fine-grained one-dimensional convolution with a total of six branches, different scales and fine-grained temporal features are extracted. Finally, residual connections are used to improve the network's generalization ability.

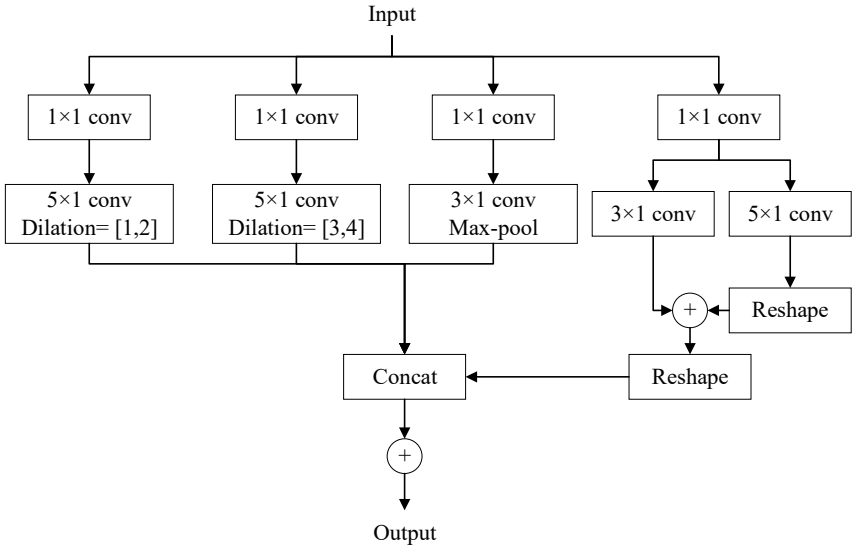

Figure 4. Structure of DM-TCN module network

3. Experiment

3.1 Environment of experiment

The experiment was conducted on a 64 bit Windows 11 operating system with AMD Ryzen 7 5800H processor and NVIDIA GeForce RTX 3070 graphics card; The deep learning framework is Python 1.13.1, the programming language is Python 3.9, the image processing accelerator is CUDA12.0, and cudnn is 8.3.2.

3.2 Dataset production

The data used in the experiment were all from drilling videos taken underground in a coal mine in Henan Province, China. Divide the drilling video in the coal mine into several 6-10 second loading and unloading drill pipe video clips, obtain the human skeleton sequence data in each video clip through YOLOv7 and FastPose, and divide it into two types: drilling action and unloading action. This dataset will be used to train the ML ASTGNet model.

3.3 Model training

To verify the effectiveness of the action recognition network proposed in the paper on the loading and unloading drill pipe actions, the ML-ASTGNet action recognition network was trained on the loading and unloading drill pipe human skeleton dataset, and the training results are shown in Fig. 5. It can be seen that the convergence speed of ML-ASTGNet network is not as fast as ST-GCN network, but with the increase of iteration times, the convergence accuracy of ML-ASTGNet network reaches 92.1%, which is 8.2% higher than the original ST-GCN network.

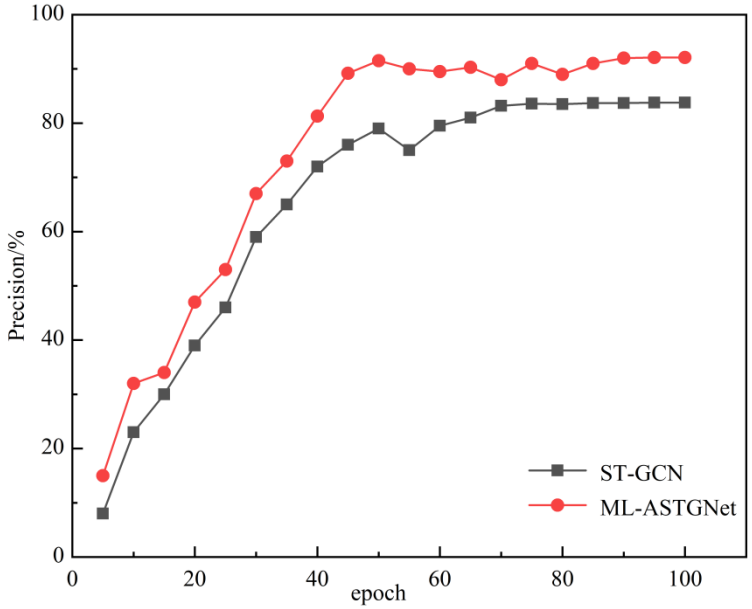

Figure 5. Precision of ST-GCN and ML-ASTGNet networks

The ablation experiment was conducted on a self built skeleton dataset, and the accuracy of the model was used as the evaluation indicator. The results are shown in Table 1. It can be seen that using A-GCN as the spatial feature extraction module, TME and DM-TCN as the temporal feature extraction modules has the highest recognition accuracy; The accuracy of using the spatial and temporal feature extraction modules mentioned above is 4.8% and 1.6% higher than ST-GCN, respectively. Further explanation shows that the spatial and temporal feature extraction ability of the original network has been improved, thereby improving the recognition accuracy of the network and meeting the expected improvement effect.

Table 1. Comparison of ablation experiment results

Original model	A-GCN	TME	DM-TCN	Precision/%
ST-GCN	×	×	×	83.9
	×	√	√	85.5
	√	×	×	88.7
	√	√	√	92.1

The action recognition network based on human bones identifies action types and performs drill pipe counting by comparing the differences between consecutive frames of human bone sequences. Fig. 6 shows the action recognition of loading and unloading drill pipes in a certain frame during the same shift.

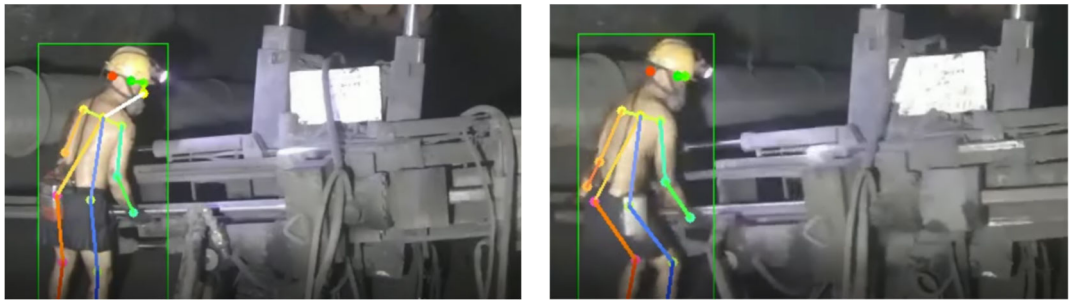

Figure 6. Action recognition of loading(left) and unloading(right) drill pipes

4. Summary

A drill pipe counting method based on human skeleton motion recognition has been proposed. Design an ML-ASTGNet action recognition network to classify the extracted human skeleton sequence, identify the loading and unloading drill pipe actions, and perform drill pipe counting. To verify the effectiveness of the ML-ASTGNet network, experiments were conducted on a self built dataset in this paper. The experimental results show that the training accuracy of ML-ASTGNet on the self built skeleton dataset reaches 92.1%, which is 8.2% higher than ST-GCN. It has a high action recognition accuracy and can demonstrate good drill pipe counting performance.

References

[1] YANG Bo. Talking about the management of gasblowout prevention in the whole process of mining and drilling in coal mine[J]. Coal Mine Modernization, 2022,31(02):94-97.DOI:10.13606/j.cnki.37-1205/td.2022.02.020.

[2] YE Junliang, QIN Qinglin. An application of directed hydraulic fracturing antireflection technology in underground mine gas control. China mine engineering, 2021,50(03):33-35+39.DOI:10.19607/j.cnki.cn11-5068/tf.2021.03.009.

[3] DENG Chengjun. Analysis on the practice of gas comprehensive treatment in outburst coal face.China mine engineering. ,2021,50(02):59-61.DOI:10.19607/j.cnki.cn11-5068/tf.2021.02.017.

[4] YAO Chaoxiu, HU Yalei. Drilling Pipe CountingAlgorithm Based on Video Analysis in Coal Mine[J]. Coal Technology, 2023,42(08):203-206.DOI:10.13301/j.cnki.ct.2023.08.044.

[5] GAO Rui, HAO Le, LIU Bao, et al. Research onunderground drill pipe counting method based on improved ResNet network[J]. Industry and Mine Automation, 2020, 46(10):6.DOI:10. 13272/j.issn.1671-251x.2020040054.

[6] Alsawadi M ,Kenawy E S E ,Rio M .Using BlazePose on Spatial Temporal Graph Convolutional Networks for Action Recognition[J].Computers, Materials & Continua,2022,74(1):19-36.

[7] DU Jingyi, DANG Mengke, QIAO Lei, et al. Drill pipe counting method based on improved spatial-temporal graph convolution neural network[J]. Industry and Mine Automation, 2023, 49(1):90-98.

[8] Yan S, Xiong Y, Lin D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition[J]. 2018

[9] Wang C Y , Bochkovskiy A , Liao H Y M .YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//arXiv.arXiv, 2022.DOI:10.48550/arXiv.2207.02696.

[10] Fang H S , Xie S , Tai Y W ,et al.RMPE: Regional Multi-person Pose Estimation[J]. 2016.DOI:10.48550/arXiv.1612.00137.

[11] Zhu Y, Shuai H, Liu G, et al. Multilevel Spatial-Temporal Excited Graph Network for Skeleton-Based Action Recognition[J]. IEEE Transactions on Image Processing, 2022, 32: 496-508.