frontiers | Frontiers in Energy Research

TYPE Original Research
PUBLISHED 29 November 2023
DOI 10.3389/fenrg.2023.1321384



OPEN ACCESS

EDITED BY
Yumin Zhang,
Shandong University of Science and Technology, China

REVIEWED BY
Zhiyi Chen,
RMIT University, Australia
Sheng Chen,
Hohai University, China
Congyue Zhang,
Southeast University, China

*CORRESPONDENCE
Ning Ji,
ningji599@163.com

RECEIVED 14 October 2023
ACCEPTED 15 November 2023
PUBLISHED 29 November 2023

CITATION
Wu Y, Ren G, Jiang B, Dai W, Ji N and Chen X (2023), Attentive multi-scale aggregation based action recognition and its application in power substation operation training.
*Front. Energy Res.* 11:1321384.
doi: 10.3389/fenrg.2023.1321384



# Attentive multi-scale aggregation based action recognition and its application in power substation operation training

Yi Wu[1], Gang Ren[1], Bing Jiang[2], Wei Dai[3], Ning Ji[1]* and Xi Chen[1]

[1]Technican Training Center of State Grid Jiangsu Electric Power Co., Ltd., Suzhou, China, [2]College of Automation and College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing, China, [3]State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China

With the rapid development of the power system and increasing demand for intelligence, substation operation training has received more attention. Action recognition is a monitoring and analysis system based on computer vision and artificial intelligence technology that can automatically identify and track personnel actions in video frames. The system accurately identifies abnormal behaviors such as illegal operations and provides real-time feedback to trainers or surveillance systems. The commonly adopted strategy for action recognition is to first extract human skeletons from videos and then recognize the skeleton sequences. Although graph convolutional networks (GCN)-based skeleton-based recognition methods have achieved impressive performance, they operate in spatial dimensions and cannot accurately describe the dependence between different time intervals in the temporal dimension. Additionally, existing methods typically handle the temporal and spatial dimensions separately, lacking effective communication between them. To address these issues, we propose a skeleton-based method that aggregates convolutional information of different scales in the time dimension to form a new scale dimension. We also introduce a space-time-scale attention module that enables effective communication and weight generation between the three dimensions for prediction. Our proposed method is validated on public datasets NTU60 and NTU120, with experimental results verifying its effectiveness. For substation operation training, we built a real-time recognition system based on our proposed method. We collected over 400 videos for evaluation, including 5 categories of actions, and achieved an accuracy of over 98%.

KEYWORDS

**substation operation, skeleton-based action recognition, multi-scale aggregation, attention mechanism, spatio-temporal fusion**

## 1 Introduction

Substations are an essential part of power systems and their safe operation is crucial to ensure the reliability of power supply. The safety awareness and standardized operation of substation operators are important factors to ensure the safe operation of substations. Therefore, incorporating artificial intelligence especial action recognition technology into substation operation training can effectively improve the safety awareness and standardized operation level of substation operators, thereby ensuring the safe operation of substations.

For example, dangerous or erroneous action of substation operators can be identified and warned.

In order to recognize action, skeleton data can first be extracted from a video sequence and then recognized. This approach has the advantage of fast processing speed and avoidance of interference from changes in background and lighting in the video.

However, action recognition based on skeleton data remains a challenging task, as it not only requires modeling the spatial domain (between joint points) but also better describing temporal features. Early studies used manually designed features to process skeleton data, but these features had limited expressive power and could not describe complex actions. In recent years, deep learning methods, especially those based on graph convolutional networks (GCN), have achieved superior performance. The human skeleton can be considered a graph structure composed of joint points and natural connections between them, making skeleton data suitable for modeling in the spatial domain (between joint points). However, GCN cannot be used for time domain modeling. Existing methods for recognizing GCN-based classes typically use traditional one-dimensional convolution to describe temporal features, but due to the varying length of dependency between moments, the kernel size has a significant impact on recognition accuracy. Additionally, these methods often alternately process spatial and temporal information, resulting in insufficient interaction between the temporal and spatial dimensions and unable to fully explore the inherent connections between time and space.

To address these issues, we propose in this paper a time-domain multi-scale information aggregation method for human skeleton-based action recognition. In order to accurately capture the dependency between varying length moments, the convolution results of multiple time-domain convolutional kernels are aggregated at a new scale dimension, producing a four-dimensional tensor including time, space, feature channels, and scale. To enable the network to automatically select important features, this paper proposes a time-space-scale fusion attention mechanism that fully integrates information across different dimensions to produce a scale-sensitive attention weight to reweight the original feature tensor. The method is validated on two publicly available datasets: NTU60 and NTU120. We have deployed our method at substation operation training locations, building a real-time behavior recognition system. We collected more than 400 video sequences, including five different action categories, with an overall recognition rate of 98%.

## 2 Related works

Computer vision has been increasingly applied in the power system due to its ability to analyze large amounts of data and detect anomalies. By analyzing video footage or sensor data, computer vision algorithms can identify potential issues in the power grid such as damaged equipment, broken wires, or other hazards that could lead to outages or safety concerns. For example, the system in (Chan et al., 2004) was able to conduct automatically intruder detection, fire alarm zone detection and substation meter reading in power substations. Automatic busbar detection from images can be conducted in (Chen et al., 2015). Mobile robots for electric power substation equipment's inspection was surveyed in (Allan and Beaudry, 2014; Lu et al., 2017; Dong et al., 2023). Automatic safety helmet detection for operators was achieved in (Li et al., 2017). In this paper, we focus on the actions of substation operators and develop algorithms to automatically identify their actions, providing a basis for subsequent analysis of the standardization and safety of their actions.

For skeleton-based action recognition, early methods employed manually designed features (Vemulapalli et al., 2014; Weng et al., 2017), with limited generalization ability and unable to extend to recognizing various complex actions. With the development of deep learning, methods based on recurrent neural networks (RNN), specifically long short-term memory networks (LSTM) were proposed to model the time domain (Du et al., 2015; Liu et al., 2016; Zhang et al., 2017). With the introduction of graph convolutional networks (GCN) and their superior performance, more and more research has been conducted based on GCN.

Graph neural networks (GNNs) (Wu et al., 2020) can handle graph data with arbitrary topology, and have been extensively studied in recent years. In these studies, graph convolutional networks (GCNs) were first introduced as the first-order approximation of local spectral convolutions (Kipf and Welling, 2016), due to their simple mean neighborhood aggregator, they are widely used for processing various graph data, including human skeleton data. However, existing methods for skeleton-based action recognition based on GCNs (Yan et al., 2018; Li et al., 2019; Cheng et al., 2020; Shi et al., 2020) tend to focus on improving the information processing in the spatial domain, while using a single one-dimensional convolution with a fixed receptive field in the temporal domain. This makes the network unable to model complex temporal dependencies and separate the time and spatial domains, resulting in limited exchange of information between them. To address these issues, this paper proposes a multi-scale time-domain information fusion network that effectively models complex relationships in the temporal domain, and a time-space-channel-scale fusion mechanism that fully communicates the four different data dimensions.

## 3 Proposed method

### 3.1 Method overview

The overall framework of the proposed method is shown in Figure 1. After three-dimensional (or two-dimensional) skeleton data goes through a series of spatial-temporal processing units, it passes through fully connected layers and obtains classification results by using the softmax function. Each spatial-temporal processing unit consists of two parts: a spatial processing unit and a temporal processing unit. The spatial processing unit is conducted by adaptive graph convolution (AGCN), while the temporal processing unit is the core of our method, which consists of multi-scale convolutional aggregation and space-time-scale fusion attention mechanism (STSA). In this method, the number of spatial-temporal processing units is set to 10.

### 3.2 Multi-scale aggregation

To overcome the problem of single receptive field in temporal convolution and difficulty in describing complex temporal

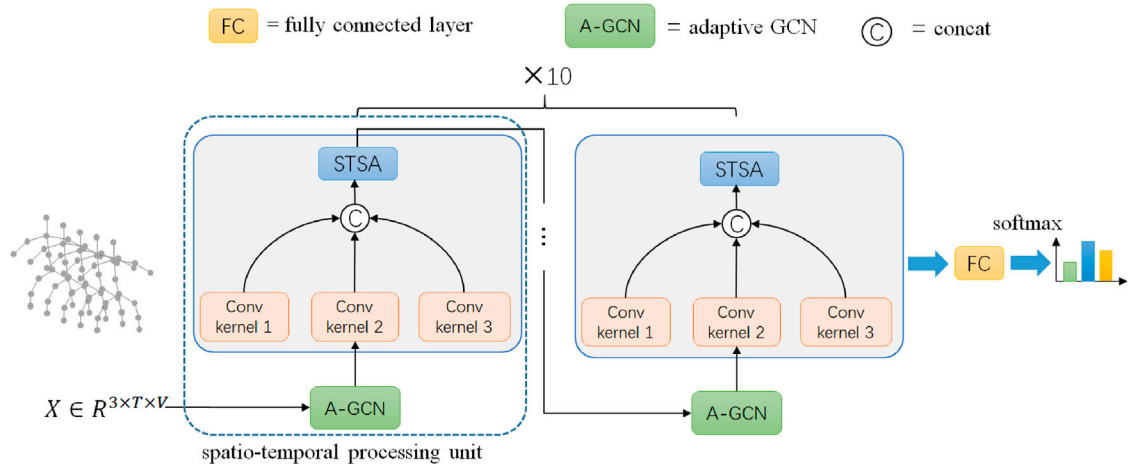


FIGURE 1
Overall architecture of the proposed method.

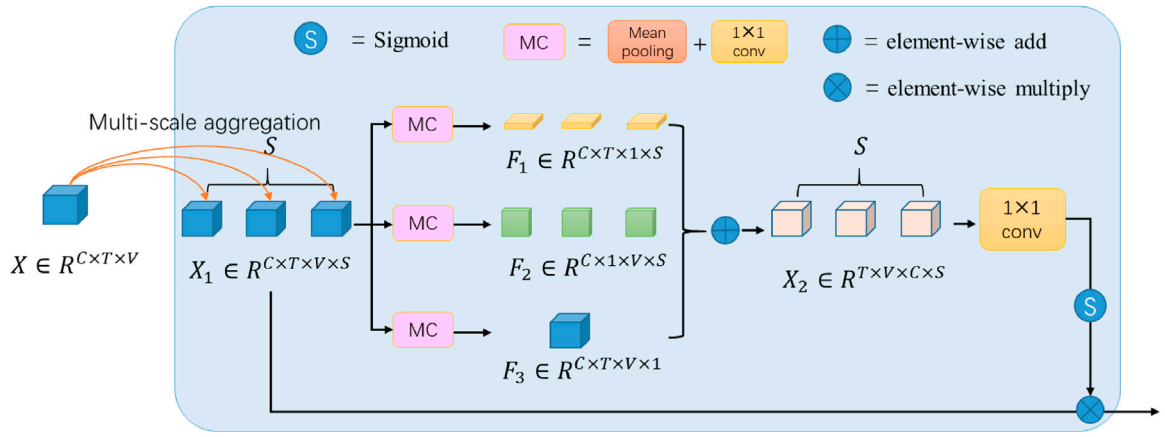


FIGURE 2
Diagram of space-time-scale fusion attention.

**TABLE 1** Performance evaluation on NTU RGB + D dataset compared with other methods.

| Method (year) | Cross-sub (%) | Cross-view (%) |
|---|---|---|
| ST-GCN (Yan et al., 2018) | 81.5 | 88.3 |
| 2s-AGCN (Shi et al., 2019) | 88.5 | 95.1 |
| Dynamic-GCN (Ye et al., 2020) | 91.5 | 96.0 |
| Adaptive-ST-GCN (Chen et al., 2021a) | 91.5 | 96.0 |
| MSTGCN (Chen et al., 2021b) | 91.5 | 96.6 |
| EfficientGCN-B4 (Song et al., 2022) | 90.8 | **96.7** |
| GSTLN (Dai et al., 2023) | 91.9 | 96.6 |
| Proposed | **92.1** | 96.5 |

The bold values are the maximum values.

**TABLE 2** Performance evaluation on NTU RGB + D 120 dataset compared with other methods.

| Method (year) | Cross-sub (%) | Cross-set (%) |
|---|---|---|
| ST-GCN (Yan et al., 2018) | 70.7 | 73.2 |
| 2s-AGCN (Shi et al., 2019) | 82.9 | 84.9 |
| Dynamic-GCN (Ye et al., 2020) | 87.3 | 88.6 |
| Adaptive-ST-GCN (Chen et al., 2021a) | 88.4 | 88.3 |
| MSTGCN (Chen et al., 2021b) | 87.5 | 88.8 |
| EfficientGCN-B4 (Song et al., 2022) | 88.7 | 88.9 |
| GSTLN (Dai et al., 2023) | 88.1 | 89.3 |
| Proposed | **88.9** | **89.7** |

The bold values are the maximum values.

**TABLE 3 Results of ablation experiments.**

| Method | Cross-sub (%) |
|---|---|
| Kernels: 3, 5, 7 | 88.6 |
| Kernels: 3, 5, 7, 9 | 88.7 |
| Kernels: 5, 7, 9, 11 | 88.8 |
| Kernels: 5, 7, 9 | **88.9** |
| Kernels: 7, 9, 11 | 88.7 |
| Kernels: 7, 9, 11 | 88.7 |
| w/o STSA | 88.3 |
| Fusion time-scale | 88.3 |
| Fusion space-time | 88.6 |
| Fusion space-scale | 88.5 |

The bold values are the maximum values.

dependencies, we propose in this paper a multi-scale convolutional aggregation method. The effectiveness of using multiple convolutional kernels to obtain different receptive fields has been validated in previous works. However, in these works, the results of multiple convolutional kernels are usually added or connected to achieve the purpose of multi-scale information aggregation. In this way, the importance of information at multiple scales is the same, making it difficult for the network to adaptively select scale information and have poor flexibility. This paper proposes to aggregate multi-scale information into a new scale dimension and then combine it with subsequent space-time-scale fusion attention mechanism to enable the network to fully fuse different dimensions of information and re-weight features based on the principle of adaptively selecting important information at time-space-scale dimensions.

Let the input features be $X \in R^{C\times T\times V}$, and after passing through $S$ different sizes of convolutional kernels, we get $S$ equally sized features. We aggregate them in the new scale dimension into a feature tensor: $X_1 \in R^{C\times T\times V\times S}$.

## 3.3 Space-time-scale fusion attention

The output tensor of the multi-scale aggregation has four dimensions: space, time, scale, and feature channel. As shown in Figure 2, we then perform feature reduction along space, time and scale dimension respectively, the reduction operation named MC module consists of a mean pooling layer (M) and a $1 \times 1$ convolution block (C). The resulted feature tensors are: $F_1 \in R^{C\times T\times 1\times S}$, $F_2 \in R^{C\times 1\times V\times S}$, and $F_3 \in R^{C\times T\times V\times 1}$, which are then expanded to $C \times T \times V \times S$ respectively and added as $X_2 \in R^{C\times T\times V\times S}$. This process can be written as:

$$X_2 = R_V(F_1) + R_T(F_2) + R_S(F_3)$$

in which

$$F_1 = MC_V(X)$$
$$F_2 = MC_T(X)$$
$$F_3 = MC_S(X)$$

where $MC(.)$ the MC module, $R(.)$ is the repeat operation.

After reduction along a certain dimension, the information in the remaining dimensions can be fully fused without interference from the reduced dimension. The final addition operation will further merge the fusion results of each dimension. In the implementation, replication can be completed by the automatic expansion function of the addition operation (most deep learning frameworks such as PyTorch, Tensorflow, etc., support this function).

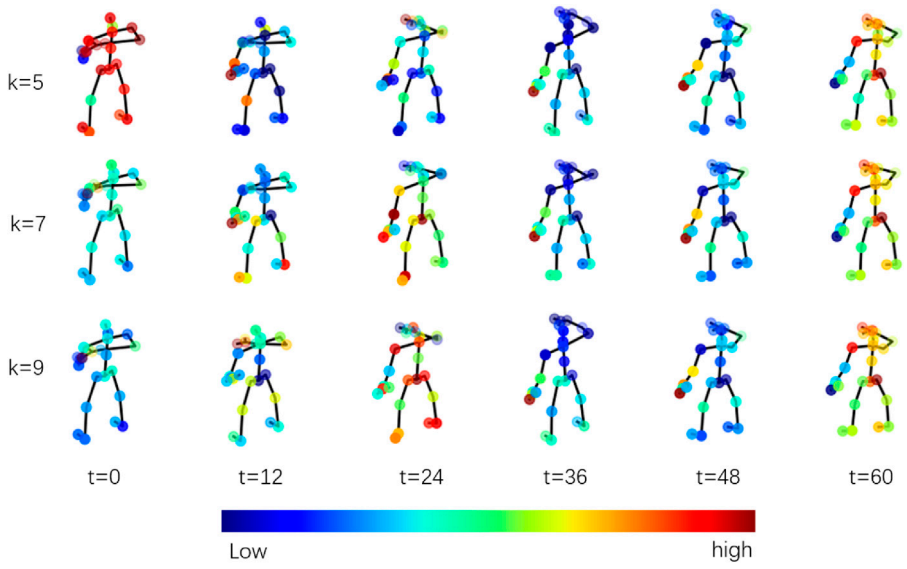


FIGURE 3
Visualization result of the learned attention by STSA.

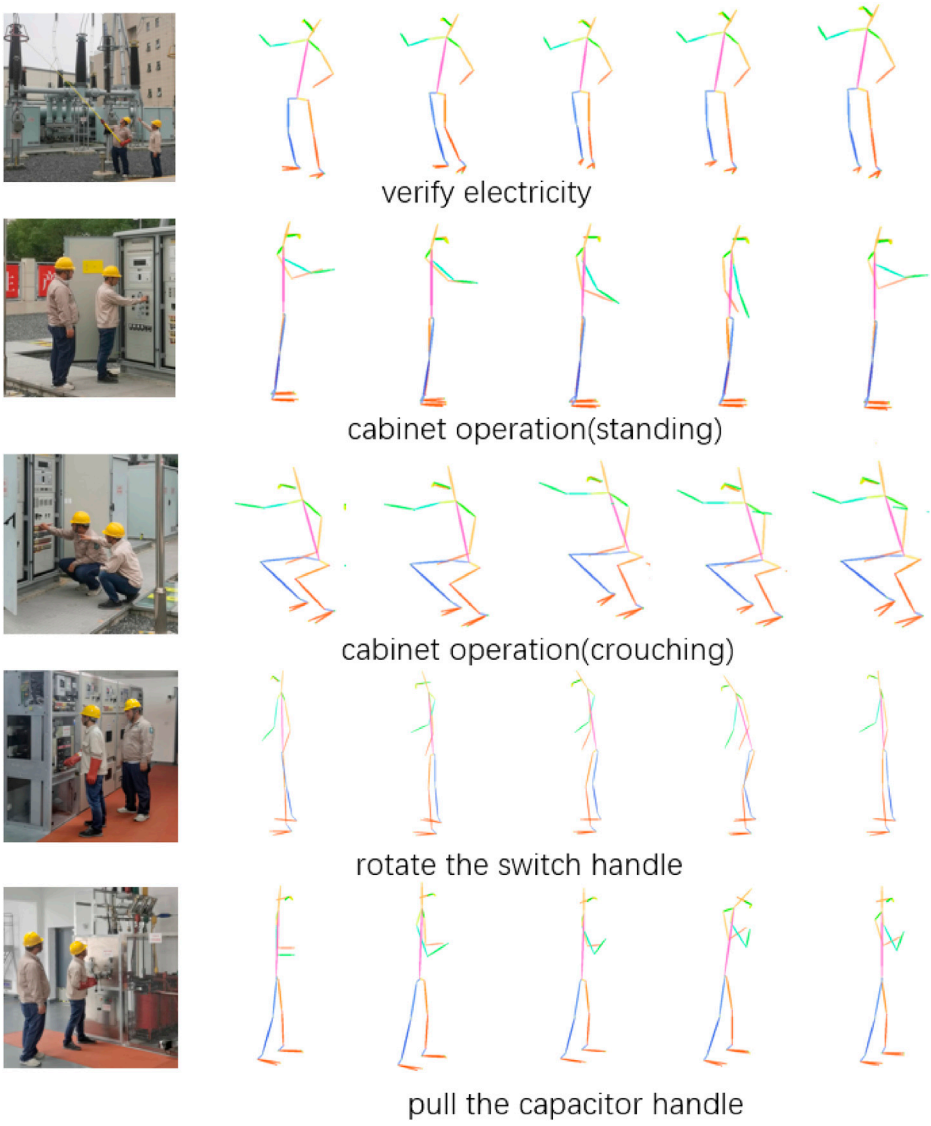


FIGURE 4
Five categories of substation operation actions.

**TABLE 4 Number of samples in each action category.**

| Category | Num. Of samples |
|---|---|
| 1. Verify electricity | 86 |
| 2. Cabinet operation (standing) | 92 |
| 3. Cabinet operation (crouching) | 84 |
| 4. Rotate the switch handle | 78 |
| 5. Pull the capacitor switch | 81 |
| Total | 432 |

## 3.4 Adaptive GCN

The skeleton can be represented as a graph structure, with joint points as vertices and connections between joint points as edges. Let the set of joint feature vectors be denoted by $P = \{P_i\}_{i=1}^{V}$, where $V$ is the number of joint points. The set of edges can be represented by an adjacency matrix $A$. By obtaining the adjacent points of each vertex, a neighborhood can be obtained for performing convolution operations similar to those used in image data:

$$p_i' = \sum_{p_j \in N(p_i)} p_j w_{ij}$$

Where $N(P_i)$ is the neighborhood of $P_i$. A linear approximation to the above convolution operator was proposed in (Kipf and Welling, 2016):

$$P' = \Lambda^{-\frac{1}{2}}(A + I)\Lambda^{-\frac{1}{2}} PW$$

Where $P$ is the matrix of the combination of all the vertex features, and

$$\Lambda_{ij} = \sum_j \left(A_{ij} + I_{ij}\right)$$

**TABLE 5** Action recognition results of the five action categories and overall results.

| Category | Acc. (%) (proposed) | Acc. (%) (Chen et al., 2021a) | Acc. (%) (Yan et al., 2018) | Acc. (%) (Shi et al., 2019) |
|---|---|---|---|---|
| 1. Verify electricity | 100 | 100 | 98.1 | 96.2 |
| 2. Cabinet operation (standing) | 97.3 | 94.6 | 92.4 | 89 |
| 3. Cabinet operation (crouching) | 100 | 100 | 94 | 92 |
| 4. Rotate the switch handle | 93.5 | 93.5 | 89.4 | 87.2 |
| 5. Pull the capacitor switch | 100 | 94 | 91.8 | 89.8 |
| Overall | 98.3 | 96.5 | 92.9 | 90.9 |

In this paper, an adaptive topology structure similar to (Chen Y. et al., 2021) is used, where $A$ is considered as a trainable parameter while the adjacency matrix serves as the initial values for $A$. This allows the network to go beyond the natural connections in topological structure and better describe the complex relationships between joint points.

# 4 Experiments

## 4.1 Evaluation on public datasets

The effectiveness of the proposed method was evaluated on two publicly available datasets: NTU-RGB + D (Shahroudy et al., 2016) and NTU-RGB + D 120 (Liu et al., 2019).

1) NTU-RGB + D: This dataset is a large-scale three-dimensional human skeleton action recognition dataset. It contains 56,880 skeleton motion clips. These actions were performed by 40 volunteers using three different perspectives of the Kinect v2 camera, categorized into 60 classes. Two common benchmarks used on this dataset are: 1) Cross-subject (cross-subject): training samples are from 20 volunteers, while testing samples are from the remaining 20 volunteers. 2) Cross-view (cross-view): training samples are from two camera perspectives, while testing data comes from a different perspective.
2) NTU RGB + D 120: This dataset is currently the largest three-dimensional human skeleton based action recognition dataset. It was created by adding an additional 57,367 skeleton motion clips to the NTU-RGB + D dataset, surpassing the number of categories to over 60. As a result, the dataset includes a total of 113,945 samples with more than 120 categories. Likewise, the newly added samples were also captured using three different perspectives of the Kinect v2 camera. Two common benchmarks used on this dataset are: 1) Cross-subject (cross-subject): training samples are from 53 volunteers, while testing samples come from the remaining 53 volunteers. 2) Cross-setup (cross-setup): training and testing samples are split based on the camera setup number.

The proposed method is implemented using the PyTorch deep learning framework, and training is completed on an RTX 3090 GPU. Stochastic gradient descent (SGD) algorithm with a learning rate of 0.1 and momentum of 0.9 is adopted as the optimizer. In all experiments, the number of training epochs is 65, with the first 5 rounds serving as warm-up to make training more stable.

As shown in Tables 1, 2, the proposed method is compared against existing methods on NTU RGB + D dataset and NTU RGB + D 120 dataset. These comparative methods are all of relatively high performance in recent years. As can be seen from Table 1, the proposed method achieved the best performance on the NTU RGB + D dataset in the Cross-sub benchmark. On the other benchmark: Cross-view, although performance is not best, it also had a small gap with the best performance. From Table 2, we can see that the proposed method achieved the best performance on two benchmarks of NTU RGB + D 120 (Cross-sub and Cross-set). These results demonstrate the effectiveness of proposed method. All the results are recognition top-1 accuracy, which is computed as the number of corrected predicted samples divided by the total number of samples.

In order to evaluate the impact of different combination of temporal convolution kernels, we conduct a series of experiments on cross-sub benchmark of NTU RGB + D 120, the evaluation results are shown in Table 3. The method with kernels of sizes 5, 7, 9 achieved the best performance. Adopting larger kernels or smaller kernels will not boost the performance. We also tried adding more convolutional kernels, but this did not lead to an improvement in performance.

In order to evaluate the role of the space-time-scale fusion attention (STSA) mechanism, we conducted an experiment with STSA removed. As shown in Table 3, the performance drops significantly, which demonstrate the effectiveness of the STSA. We also evaluate the role of different dimensions in STSA by removing one of the dimension branches. As shown in Table 3, we evaluate time-scale, space-time and space-scale fusion respectively. Among them, space-time fusion yields relatively better result, which is yet lower than space-time-scale fusion.

To visualize the content learned by our network, especially the STSA attention mechanism proposed in this paper, we present the results of the STSA attention mechanism in the first temporal-spatial processing unit in Figure 3. The example is a “drinking water” scenario. Red indicates high weights, and blue indicates low weights. The weights are normalized using the following equation:

$$w_{normalized} = \frac{w - \min}{\max - \min}$$

In terms of the scale dimension (kernel size k), we can see that at different scales, the network focuses on different contents, which means that the network has the ability to adaptively select the scale.

### 4.2 Application in power substation operation training

The proposed method was applied to operation training in power substation. We collected the operation videos of trainees during the training process and used body posture estimation algorithm Alphapose (Fang et al., 2022) to extract 2D skeleton data of human bodies for action recognition. The model was trained on Halpe dataset (Fang et al., 2022) which is able to extract 2D skeleton including 26 joints. In order to make the network's structure unchanged, the 2D data is treated the same way as 3D skeleton data. The action categories of trainees were divided into five types: 1) verify electricity; 2) cabinet operation (standing); 3) cabinet operation (crouching); 4) rotate the switch handle; 5) pull the capacitor switch. Examples of the five categories of actions are shown in Figure 4. The colors in Figure 4 represent different parts of human skeleton. We collected multiple videos from multiple perspectives of operators for each category. Similarly to NTU60, we resampled the skeleton data extracted from each video in time, and all the resampled skeleton data had the same dimension in time. We selected 60% of them as training samples and the remaining as testing samples. The number of samples in each action category is listed in Table 4.

The recognition system runs on a PC with Intel i9 CPU, 32 GB RAM and RTX3090 GPU. The system is able to achieve real-time recognition. For each frame, the computation time of posture estimation is 41 ms, and the computation time of action recognition is 12 ms, so the system runs at 18.9 FPS which is sufficient for most of the real-time applications.

In these five action categories, the first and third categories are relatively easy to identify, while the remaining three categories are more similar, with the main difference being hand movements. The proposed method is compared against (Yan et al., 2018; Shi et al., 2019; Chen Y. et al., 2021). The proposed method and (Chen Y. et al., 2021) achieved 100% accuracy in the first and third categories. And the other two methods (Yan et al., 2018; Shi et al., 2019) made wrong predictions in these two categories. The proposed method also achieved all correct classification results in the fifth category which outperforms other three methods. The overall recognition rate of the method in this paper exceeded 98% which outperformed (Chen Y. et al., 2021) by 1.8% (Yan et al., 2018), by 5.4% and (Shi et al., 2019) by 7.4%. See Table 5 for comparison results. From the results, we can learn that the action of "Rotate the switch handle" is most prone to misclassification. Though the proposed method achieved an accuracy of 93.5% ranking first alongside (Chen Y. et al., 2021), in our future work, we will conduct further research on this category of action.

## 5 Conclusion

In this paper, we propose a skeleton-based action recognition method that aggregates convolutional information of different scales in the time dimension and a space-time-scale attention module that enables effective communication and weight generation between dimensions. Our proposed method is validated on public datasets NTU60 and NTU120, with experimental results demonstrated its effectiveness. For substation operation training, we built a recognition system and collected hundreds of videos for evaluation, including 5 categories of actions, and achieved satisfactory recognition accuracy.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://rose1.ntu.edu.sg/dataset/actionRecognition/.

## Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

## Author contributions

YW: Conceptualization, Visualization, Writing–original draft. GR: Data curation, Investigation, Methodology, Validation, Visualization, Writing–review and editing. BJ: Investigation, Validation, Writing–original draft. WD: Investigation, Methodology, Visualization, Writing–review and editing. NJ: Data curation, Formal Analysis, Investigation, Methodology, Validation, Writing–original draft. XC: Data curation, Formal Analysis, Methodology, Resources, Visualization, Writing–review and editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research is funded by the Science and Technology Project of State Grid Jiangsu Electric Power Company, grant number J2021215.

## Acknowledgments

The authors would like to thank the editors and reviewers for improving this paper.

## Conflict of interest

Authors YW, GR, NJ, and XC were employed by Technican Training Center of State Grid Jiangsu Electric Power Co., Ltd. Author WD was employed by State Grid Jiangsu Electric Power Co., Ltd.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors declare that this study received funding from State Grid Jiangsu Electric Power Company. The funder had the following involvement in the study: design, collection, analysis, interpretation of data, the writing of this article and the decision to submit it for publication.

## Publisher's note



## References

Allan, J. F., and Beaudry, J. (2014). "Robotic systems applied to power substations-a state-of-the-art survey," in Proceedings of the 2014 International Conference on Applied Robotics for the Power Industry (CARPI), Foz do Iguassu, Brazil, October 2014 (IEEE), 1–6. doi:10.1109/CARPI.2014.7030049

Chan, W. L., Leo, P. S. L., and Ma, C. F. (2004). "Computer vision applications in power substations," in Proceedings of the 2004 International Conference on Electric Utility Deregulation, Restructuring and Power Technologies (DRPT), Hong Kong, China, April 2004 (IEEE), 383–388. doi:10.1109/DRPT.2004.1338526

Chen, H., Sun, S., Wang, T., Zhao, X., and Tan, M. (2015). "Automatic busbar detection in substation: using directional Gaussian filter, gradient density, Hough transform and adaptive dynamic K-means clustering," in Proceedings of 2015 Chinese Control Conference (CCC), Qing Dao, China, July 2015 (IEEE), 4668–4672. doi:10.1109/ChiCC.2015.7260360

Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., and Hu, W. (2021a). "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in Proceedings of the 2021 international conference on computer vision (ICCV), Montreal, QC, Canada, October 2021 (IEEE), 13359–13368. doi:10.1109/iccv48922.2021.01311

Chen, Z., Li, S., Yang, B., Li, Q., and Liu, H. (2021b). Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. *Proc. 2021 Conf. Artif. Intell. (AAAI)* 35 (2), 1113–1122. doi:10.1609/aaai.v35i2.16197

Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., and Lu, H. (2020). "Decoupling gcn with dropgraph module for skeleton-based action recognition," in Proceedings of the Computer Vision–ECCV 2020: 16th European Conference (ECCV), Berlin, Heidelberg, August 2020 (Springer), 536–553. doi:10.1007/978-3-030-58586-0_32

Dai, M., Sun, Z., Wang, T., Fen, J., and Jia, K. (2023). Global spatio-temporal synergistic topology learning for skeleton-based action recognition. *Pattern Recognit.* 140, 109540. doi:10.1016/j.patcog.2023.109540

Dong, L., Chen, N., Liang, J., Li, L., Yan, Z., and Zhang, B. (2023). "A review of indoor-orbital electrical inspection robots in substations. *Industrial Robot Int. J. robotics Res. Appl.* 50 (2), 337–352. doi:10.1108/IR-06-2022-0162

Du, Y., Wang, W., and Wang, L. (2015). "Hierarchical recurrent neural network for skeleton based action recognition," in Proceedings of the 2015 conference on computer vision and pattern recognition (CVPR), Boston, USA, June 2015 (IEEE), 1110–1118. doi:10.1109/cvpr.2015.7298714

Fang, H. S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., et al. (2022). Alphapose: whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans. Pattern Analysis Mach. Intell.* 2022, 7157–7173. doi:10.1109/tpami.2022.3222784

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. Available at: https://arxiv.org/abs/1609.02907.

Li, J., Liu, H., Wang, T., Jiang, M., Li, K., and Zhao, X. (2017). "Safety helmet wearing detection based on image processing and machine learning," in Proceedings of the 2017 international conference on advanced computational intelligence (ICACI), Doha, Qatar, February 2017 (IEEE), 201–205. doi:10.1109/ICACI.2017.7974509

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., and Tian, Q. (2019). "Actional-structural graph convolutional networks for skeleton-based action recognition," in Proceedings of the 2019 conference on computer vision and pattern recognition (CVPR), Long Beach, USA, June 2019 (IEEE), 3595–3603. doi:10.1109/cvpr.2019.00371

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L., and Kot, A. C. (2019). Ntu RGB+ d 120: a large-scale benchmark for 3d human activity understanding. *IEEE Trans. pattern analysis Mach. Intell.* 42 (10), 2684–2701. doi:10.1109/tpami.2019.2916873

Liu, J., Shahroudy, A., Xu, D., and Wang, G. (2016). "Spatio-temporal lstm with trust gates for 3d human action recognition," in Proceedings of the Computer Vision–ECCV 2016: 14th European Conference (ECCV), Amsterdam, Netherlands, October 2016 (Springer), 816–833. doi:10.1007/978-3-319-46487-9_50

Lu, S., Zhang, Y., and Su, J. (2017). "Mobile robot for power substation inspection: a survey,", 830–847. doi:10.1109/JAS.2017.7510364*IEEE/CAA J. Automatica Sinica*44

Shahroudy, A., Liu, J., Ng, T. T., and Wang, G. (2016). "NTU RGB+D: a large scale dataset for 3D human activity analysis," in Proceedings of the 2016 conference on computer vision and pattern recognition (CVPR), Las Vegas, USA, June 2016 (IEEE), 1010–1019. doi:10.1109/cvpr.2016.115

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2019). "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in Proceedings of the 2019 conference on computer vision and pattern recognition (CVPR), Long Beach, USA, June 2019 (IEEE), 12026–12035. doi:10.1109/cvpr.2019.01230

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2020). Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Trans. Image Process.* 29, 9532–9545. doi:10.1109/tip.2020.3028207

Song, Y. F., Zhang, Z., Shan, C., and Wang, L. (2022). Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Trans. pattern analysis Mach. Intell.* 45 (2), 1474–1488. doi:10.1109/tpami.2022.3157033

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). "Human action recognition by representing 3d skeletons as points in a lie group," in Proceedings of the 2014 conference on computer vision and pattern recognition (CVPR), Columbus, USA, June 2014 (IEEE), 588–595. doi:10.1109/cvpr.2014.82

Weng, J., Weng, C., and Yuan, J. (2017). "Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition," in Proceedings of the 2017 conference on computer vision and pattern recognition (CVPR), Hawaii, USA, July 2017 (IEEE), 4171–4180. doi:10.1109/cvpr.2017.55

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE Trans. neural Netw. Learn. Syst.* 32 (1), 4–24. doi:10.1109/TNNLS.2020.2978386

Yan, S., Xiong, Y., and Lin, D. (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Proceedings of the 2018 conference on artificial intelligence (AAAI), Louisiana, USA, February 2018, 7444–7452. doi:10.1609/aaai.v32i1.12328

Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D., and Tang, H. (2020). "Dynamic GCN: context-enriched topology learning for skeleton-based action recognition," in Proceedings of the 2020 ACM international conference on multimedia, Seattle, USA, October 2020 (ACM), 55–63. doi:10.1145/3394171.3413941

Zhang, S., Liu, X., and Xiao, J. (2017). "On geometric features for skeleton-based action recognition using multilayer lstm networks," in Proceedings of the 2017 winter conference on applications of computer vision (WACV), California, USA, March 2017 (IEEE), 148–157. doi:10.1109/wacv.2017.24