Neural Tensor Network And Adaptive Graph Convolution For Sports

Lin Teng, Hang Li*, and Yuchang Si*

College of Artificial Intelligence, Shenyang Normal University, Shenyang 110034, 7 China

* *Corresponding author. E-mail: lihangsoft@163.com, ljnan127@163.com*

Received: Jun. 14, 2025; Accepted: Oct. 06, 2025

Current human bone action recognition algorithms have some problems such as insufficiently detailed description of the global relationship and insufficient mining of spatio-temporal features. Therefore, this paper proposes a novel sports action recognition based on neural tensor network and adaptive graph convolution. Firstly, the attention mechanism and neural tensor network (NTN) algorithm are used to solve the connection strength between each pair of joint nodes and construct the global adjacency matrix. Secondly, by using the topK strategy, the topK neighbor nodes are dynamically selected based on the connection strength to update the global adjacency matrix. Thirdly, the hybrid pooling model is adopted to extract the global context information and the temporal key frame features. By simultaneously modeling joint information, bone information, joint movement information and bone movement information, the representation ability of the features extracted by the model for movements is strengthened. The experimental results on the Something-Something V1&V2 and Kinetics-400 datasets show that the proposed model in this paper outperforms most other advanced action recognition methods, proving that this new model can effectively improve the performance of action recognition.

Keywords: sports action recognition, neural tensor network, adaptive graph convolution, topK strategy, hybrid pooling model

http://dx.doi.org/10.6180/jase.202606_29(6).0015

1. Introduction

Action recognition is a challenging task that requires identifying the behavioral semantics in videos from the action atoms with different semantics contained in video frames. It has been widely applied in many fields such as video understanding [1], video recommendation, and human-object interaction [2, 3] etc. The actions in the video contain a series of action atoms. The semantics of the action atoms are explained by analyzing the temporal changes of the actions (including moving objects or people), and objects (or people) constantly change their actions, resulting in different semantics of the action atoms at different times. Most of the existing behavior recognition methods determine the video behavior category by extracting the features of the entire video frame, but ignore learning the regional-level features in the video frame [4]. These methods are mainly divided into convolution-based methods and Transformer-based methods.

The convolution-based methods can describe the temporal changes through post-segmentation aggregation of time, temporal translation [5, 6], and temporal dilated convolution [7] to explain the behavioral semantics of the video. The temporal variation is manifested to a certain extent as the action features of the video [8], and many works have also attempted to learn the action features to improve the temporal expression ability of the model. References [9, 10] learned action features by designing a time difference module. Ma et al. [11] utilized the correlation of videos to extract features. In order to select the spatio-temporal regions related to the actions, some methods enabled the model to pay more

Submission Template to Journal of Applied Science and

Engineering attention to the action regions of objects in the video frames through spatio-temporal patch selection and foreground extraction [12]. Meanwhile, there are also some works [13, 14] that use different types of temporal attention to enhance action-related features. Most of these methods are based on two-dimensional convolution and still have difficulties in learning the features of long video frames. These methods based on three-dimensional convolution can capture the action features of multiple frames within local time intervals. They extend the convolutional neural network to the time dimension through methods such as expanding the two-dimensional convolution kernel [15] and multi-view three-dimensional convolution [16].

These methods obtain global temporal features by simply stacking the local features learned by multi-layer three-dimensional convolutional neural networks. However, the features will be lost layer by layer during the propagation process of the multi-layer network, making it difficult for them to pay attention to the minute temporal changes between regions.

The Transformer-based methods can capture the temporal relationship between frames in the global sequence. However, by calculating the self-attention between all regions in the frame, while generating a huge amount of computation, the model is more likely to focus on the redundant spatial regions rather than the behavioral semantic regions that change over time. In addition, there are some methods that are different from completely supervised training methods, attempting to explore new paradigms of behavior recognition learning in the ways of self-supervision [17] and weak supervision [18]. The above-mentioned methods ignore the region-level features in learning video frames and fail to analyze the changes of regional features over time. It may not be able to pay attention to all the action atoms in the video frames, thereby affecting the action recognition of the video.

The main human action recognition methods based on deep learning include convolutional neural network (CNN), recurrent neural network (RNN), Transformer and graph neural network (GCN). CNN, RNN and Transformer usually process the data into a lattice structure, while the skeletal sequence is more in line with the definition of the graph sequence, and thus the graph convolutional network emerged. Qiu et al. [19] introduced the concept of spatio-temporal graphs and constructed adjacency matrices that could reflect the topological structure of the human body. Batool et al. [20] proposed that the convolutional layer of the adaptive spatial graph adaptively learned the correlation between joints. Peng et al. [21] proposed the central connected graph convolutional network with significant image features (SIFE-CGCN) to capture the minute differences between similar actions.

Although graph convolutional neural networks have made significant progress in many fields, they still face some challenges: (1) Most graph convolutional networks use the topological structure of the human body to construct the corresponding adjacency matrix, which cannot fully explore the characteristics of the human body structure in the spatial dimension; (2) The interaction range of each limb in human movement is limited, and it is unreliable to directly define the global relationship using the fully connected matrix; (3) Time convolution only uses $K_t \times 1$ convolution kernels to extract features; (4) Most networks only model joint flows and do not fully consider other data flows, lacking certain data support.

Therefore, this paper designs a novel sports action recognition model (NTN-AGC) based on neural tensor network and adaptive graph convolution. The main contributions of this article are as follows. (1) By combining the global context-aware attention with the NTN network, an innovative construction method with the adaptive global adjacency matrix is proposed. On the basis of improving the accuracy of node representation, the ability to extract the structural features of the human spatial dimension is enhanced. (2) We adopt the topK global adjacency relationship calculation model to eliminate some unnecessary cooperative relationships among nodes. (3) We introduce a hybrid pooling model in the time dimension and combine it with the convolution extraction method in the time dimension to enhance the accuracy of time feature extraction. (4) The NTN-AGC model simultaneously models joint flow, limb flow, joint motion flow, and limb motion flow for action recognition.

2. Materials and methods

Since the traditional CNN model can only be used to handle regular one-dimensional sequence or two-dimensional matrix data, many data models in practice do not have such a regular structure. Researchers extend the traditional convolution to the graph structure and propose the graph convolutional network. Graph convolution can be simply described as superimposing all the neighbors of a node according to different weights. The calculation formula of the graph convolutional network is as follows.

$$H^{(l+1)} = \sigma\left(\widetilde{D}^{-0.5}(\widetilde{A}+M)\right)\widetilde{D}^{-0.5}H^{l}W^{l} \tag{1}$$

Where, H^l represents the node feature representation of the $l-$ th layer. $A = A + I$. I is the identity matrix. D is the degree matrix of A.M is a learnable mask. W is the weight

parameter matrix of network learning. σ is the activation function.

The two-stream adaptive graph convolutional network (2 s − AGCN) designed by Xie et al. [22] adopted a two-stream structure and modeled both joint flow and bone flow simultaneously. This model learned the topology of the graph in an end-to-end manner through the BP algorithm. The architecture was composed of 9 layers of adaptive graph convolutional blocks stacked together. Each module included a spatial graph convolutional layer, a batch normalization layer, ReLU nonlinear activation, Dropout regularization, a temporal graph convolutional layer, and residual connections. Here, the spatial graph convolution layer combined the adjacency matrix representing the physical structure of the human body with the data-dependent graph determined by the Gaussian function to form a new adjacency matrix for graph convolution and extract spatial feature information. Although this model performed well in capturing spatio-temporal features, it had high computational complexity and a long training time.

The multimodal graph self-attention network (MGSAN) utilized the self-attention mechanism to aggregate local and global features and enhanced the modeling ability of the dependency relationships between different skeletal joints through multimodal learning. However, the model could still face the problem of high computational complexity, especially when dealing with large-scale data. The computational overhead of graph self-attention could lead to an efficiency bottleneck. Furthermore, the model still had certain deficiencies in aspects such as data dependence, cross-modal fusion, and training efficiency.

The salient image feature enhanced central connected graph convolutional network (SIFE-CGCN) proposed by Bai et al. [23] was used to identify similar actions. This model introduced the central connection strategy to capture the connection relationship between joints and fused salient image features to improve motion discrimination. Although this model had achieved excellent results in recognizing similar actions, it faced high computational overhead. The extraction of image features was highly dependent on the recognition accuracy, and the fusion of bone data and image features still needed further optimization.

2.1. Spatio-temporal Graph Construction

In actual situations, the relationship between the joint points is constantly changing during the motion process. Therefore, the fixed topology diagram A in ST-GCN cannot represent the constantly changing connection mode between the joint points. Therefore, this paper designs an adaptive spatio-temporal graph convolutional layer.

$$f_{\text{out}} = \sum_{k}^{K_v} W_k f_{\text{in}} \left(A_k + B_k\right) \tag{2}$$

In the formula, f_{in} and f_{out} are the input and output of the convolutional layer respectively. W_k is the weight matrix. Specifically, the adjacency matrix of a graph is composed of two parts: A_k and $B_k \cdot A_k$ represents the predefined topology diagram, which is used to depict the physical connections of the human body. B_k is a graph dynamically generated based on sample data, reflecting the connection strength between specific time points. In order to determine the connection relationship and strength between the joint points, this paper applies the neural tensor network (NTN) [24] to calculate the similarity between the joint points and screens out the more important connections through the topK strategy. Through this adaptive design, the adjacency matrix can be dynamically adjusted according to the time-domain characteristics of the action to ensure that key dynamic changes are captured at different motion stages.

In the task of human behavior recognition, the movements of the same joint point in different frames show a certain correlation. However, the importance of the actions of each frame to the entire motion sequence varies. This requires assigning differentiated weight allocations to each joint point among different frames. The global context-aware attention mechanism is calculated by:

$$h = \sum_{t=1}^{T} f\left(v_t^T c\right) v_t = \sum_{t=1}^{T} f\left(v_t^T \tanh\left(\frac{1}{T}\sum_{t=1}^{T} v_t\right) W\right) v_t \tag{3}$$

Where, v_t is the feature of the joint point on the t frame. ^{c} represents the global context feature of the joint point. f represents the sigmoid activation operation. T is the number of frames of this action sample. W is a learnable weight matrix. $\mathrm{c} = \tanh\left(\frac{1}{T}\sum_{t=1}^{T} v_t\right) W$ is a simple mean that undergoes nonlinear changes. Through the weight matrix W, c provides the temporal global structure and characteristic information of this joint point. According to c, the attention weight a_t can be calculated for each frame of this joint point. For the joint point v_t, the attention weight of this joint point is obtained by calculating the inner product of c and the joint point, which can ensure that joint points similar to the global context obtain higher weights. Then, we apply the sigmoid function to ensure that the value is always between $(0, 1)$. Finally, the joint feature $h = \sum_{t=1}^{T} a_t v_t \in R^c$ is the weighting of the frame.

The simplest way to model the relationship between two joints based on the previously generated joint features

is the inner product. However, the inner product will lead to insufficient interaction between the two. Therefore, NTN is adopted to simulate the relationship between two joint points in this paper.

$$g\left(h_i h_j\right) = f\left(h_i^T W^{[1:k]} h_j + V\left[h_i h_j\right]^T + b\right) \tag{4}$$

Where h_i and h_j are the joint point features generated by the attention layer. $W^{[1:k]}$ is the weight tensor. V is also a weight vector. b is an offset vector of length K.f represents the activation function. K is a hyperparameter that controls the number of correlation scores generated by the model for each pair ofjoints. The schematic diagram of the neural tensor network is shown in Fig. 1.

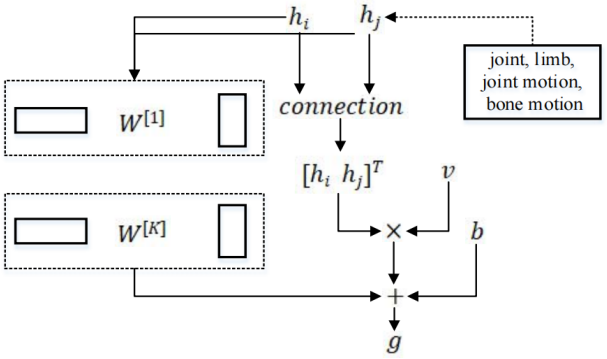

Fig. 1. NTN network

Submission Template to Journal of Applied Science and Engineering In actual movement, a single limb cannot interact with all the limbs of the human body simultaneously, and at different stages of movement, the interaction object of the same limb may change over time. To solve this problem, a global adjacency relation update strategy based on topK is proposed.

The topK strategy adaptively adjusts the adjacency relationship by selecting the K most relevant adjacent limbs of each limb at a specific time node. As a sorting algorithm, the main function of $f_{\text{topk}}(.)istofilteroutthepositionindexofthefirst$K elements in the sorted vector and return it. The K value is set based on specific experimental results.

In the generation process of the dynamic adjacency matrix, the similarity between each pair of joint points is calculated first through the NTN network. These similarities reflect the connection strength between the joint points. Then, the topK strategy is used to select the K larger values from the similarity values to construct a sparse dynamic adjacency matrix. For each frame, the dynamic adjacency matrix only retains the K connections with the highest interaction intensity with the target joint points. This can effectively reduce the computational complexity and focus on the important joint point interactions.

2.2. Temporal Hybrid Pooling

When dealing with complex action patterns, the traditional $K_t \times 1$ convolutional kernel is difficult to comprehensively capture the dynamic changes at different time scales, which may lead to the neglect of key time dimension features. In order to effectively improve the extraction ability of temporal features, this paper introduces a hybrid pooling model. By combining different types of pooling operations, the capture ability of temporal features is enhanced through multi-level feature fusion.

The implementation process of this model includes the following several steps. The first step is to input the features with a size of $N \times C \times T \times V$. The second step is temporal convolution, which uses a $k_t \times 1$ convolution kernel to convolve the features and then performs batch normalization processing. The third step is mixed pooling. After the output of the convolutional layer, the dimension permutation operation is carried out accordingly, and then global average pooling and maximum pooling are performed respectively, as shown in equation (5). f_2 is the normalized feature. f_T is the feature after mixed pooling. Step 4 is feature enhancement. The feature f_T is processed using 1×1 convolution, normalization is implemented with softmax, and the feature size is restored to the state before pooling through expand. Then, the permutation operation is performed again. The process is shown as equation (6). f_3 is the feature after the re-displacement operation. Step 5 is the output. It introduces the residual Res, adds f_4 and f_1 to obtain the feature output f_{out} , as shown in equation (7).

$$\begin{aligned} f_T &= \text{Concat}\left(\text{Maxpool}\left(f_2\right), \text{Avgpool}\left(f_2\right)\right) \\ &= \frac{2}{c \times V} \sum_{k=1}^{c} \sum_{j=1}^{V} f_2(k, j) \end{aligned} \tag{5}$$

$$\begin{aligned} f_4 &= f_1 \times f_3 \\ &= \sum_{k=1}^{C} \sum_{i=1}^{T} \sum_{j=1}^{V} f_1^k(i, j) f_3^k(i, j) \end{aligned} \tag{6}$$

$$f_{\text{out}} = \text{Res}\left(f_1 f_4\right) = f_1 + f_4 \tag{7}$$

In step 3, by exchanging the channel dimension and the time dimension, the sensitivity of the model to the dynamic changes of the time series is enhanced. The above content reveals the advantages of the hybrid pooling model, which can simultaneously extract the key frame action features and global context features in the time dimension. When

these features are combined with the features obtained through temporal convolution, the capture efficiency of the temporal dimension characteristics of action information can be significantly improved.

2.3. Overall Multi-stream Architecture

This paper uses a multi-stream structure framework to model the joint information, skeletal information, joint motion information and skeletal motion information respectively, and enhances the action recognition effect by extracting more features. Fig. 2 describes the overall structure of the multi-stream adaptive spatio-temporal graph convolutional network. First, it preprocesses the bone data to obtain the original data, and then extracts the features and inputs them into the model. The model achieves full-process end-to-end training from input to output through the spatio-temporal graph convolution module of the L layers. Supposing the coordinates of the two joint points of the bone are $v_1 = (x_1, y_1, z_1)$ and $v_2 = (x_2, y_2, z_2)$, then the coordinates of the bone are $\text{bone}_{(v_1,v_2)} = (x_2 - x_1, y_2 - y_1, z_2 - z_1)$.

Motion information refers to the variation of the same joint point between different frames.

Suppose the coordinate of the $i-$ th joint point of the $t-$ th frame is $v_{i,t}$, then the motion information is motion $_{i,t} = v_{i,t} - 0.5v_{i,t-1} + v_{i,t+1}$. The calculation of bone movement information is the same. Finally, the classification results of the multi-stream network are weighted and summed. The specific weight ratio is determined by the experiment.

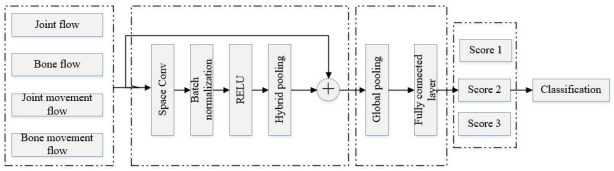

Fig. 2. The overall structural framework of the proposed model

3. Results and discussion

3.1. Experiment data sets and experiment settings

This paper conducts experiments on three public datasets for behavior recognition. The

Something-Something V1 dataset contains 108499 video clips, covering 174 action categories. The Something-Something V2 dataset is an extension of the Something-Something V1 dataset. It contains 220847 video clips, covering 174 action categories, and the average duration of the video clips is 4 s . The Something-Something dataset collects the same actions performed on different objects and identifies action categories more by learning the temporal characteristics of the objects. Kinetics-400 contains 300k video clips, covering 400 action categories, and the average duration of the video clips is 10 s . The Kinetics-400 dataset is collected from YouTube videos related to daily life, and the action categories are highly relevant to the features of objects and scenes.

In the training stage, this paper uses the sparse sampling method in reference [25] to extract T frames from each video clip (in the experiment, T $=$ 8 or T $=$ 16). It adjusts the shorter side of the video frame to 256 , utilizes center cropping and zoom jitter, and finally inputs the cropped image with a shape of $224 \times 224 \times 3$ into the network. For the global time memory module in the local-global time feature learning module, we set the number of feature channels after compression $C' = C/16$. Setting the channel ratio parameter α to 0.5. In the regional semantic learning module, we set the number of regional semantic convolution kernels C_h to be learned as 8 . The learning rate and weight of the final classification layer of the network decay to five times that of stage 2 to stage 5 . On the Something-Something V1 and Something-Something V2 datasets, the batch size, initial learning rate, weight decay and dropout are set to $64, 0.02, 5e^{-4}$ and 0.5 respectively. On the Kinetics400 dataset, these hyperparameters are set to $64, 0.01, 1e^{-4}$ and 0.5 respectively. Small-batch stochastic gradient descent is used as the optimizer on all three datasets for a total of 50 epochs, and the learning rate is reduced to one-tenth of the original in the 30th, 40th, and 45th epochs.

3.2. Comparison experiments

Table 1 shows the comparative experimental results of the proposed method in this paper with other existing advanced methods on the Something-Something V1&V2 dataset. The comparison contents include the computational cost of the model (FLOPS), the number of model parameters (Param), and the classification accuracy rates of top- 1 and top- 5 under different reasoning strategies. SAM-STI uses three-dimensional spatio-temporal attention to enhance temporal features. ST-Adapter and TPS are Transformer-based methods, which adopt self-attention for spatio-temporal modeling of long sequences. STDN and GSF enhance the spatio-temporal expression ability by improving the two-dimensional convolutional neural network. SIFA and FMENet seek to improve the recognition accuracy through the inter-frame relationship. STDN introduces spatio-temporal hybrid adaptive convolution, which aggregates local and global features using channel,

time, space, and spatio-temporal joint attention respectively. After decomposing the characteristics of spatio-temporal interaction through the spatio-temporal gating mechanism, GSF models time and space respectively. They lack attention to the areas where time changes. SIFA studies the differences between adjacent frames, obtains the temporal attention within the local deformation area from it, thereby estimating the offsets of objects at different times and achieving the alignment of the behavioral semantics of adjacent frames. FMENet increases the focus on the relevant action regions through inter-frame differences and encodes the semantics of video behaviors using multi-layer receptive fields. However, they lack the learning of semantic changes in the global regions within video frames. The proposed NTN-AGC method enhances the local temporal features by using local temporal attention and further aggregates the global temporal regional features. It uses regional semantic learning to construct variable convolution kernels. The variable convolution kernel can learn the behavioral semantic features that change over time, and finally fuse the variable regional features with the global temporal regional features for feature enhancement. From the experimental results, the NTN-AGC has achieved significant performance gains. Compared with the existing methods, it has surpassed the classification accuracy of most methods on the Something-Something V1&V2 datasets.

Table 2 presents the results of the comparative experiments on the Kinetics-400 dataset. When using the model with 8 -frame input, the NTN-AGC method in this paper is superior to other methods such as FEXNet, T-STFT, and GSF. From the perspective of the characteristics of the dataset, the Kinetics-400 is human-centered. Among the contained videos, most are human-related behaviors, including daily life behaviors and some common activities, such as brushing teeth and catching fish. The action types in this dataset are highly correlated with the scenes. The network model may be able to infer the action types merely from the appearance features of the spatial background of the video frames, and the behavioral semantic changes between frame sequences are relatively small. The NTN-AGC method mainly improves the performance of the model in the behavior recognition task by focusing on the regional semantic changes in video frames. Therefore, the advantages of the NTN-AGC method cannot be fully reflected on the Kinetics-400 dataset. However, from the experimental results, the NTN-AGC still outperforms most existing methods on the Kinetics-400 dataset. In terms of the top-1, it has improved by 0.4% compared to the best method in Table 2.

In this section, a series of ablation experiments are conducted on the proposed model to verify the validity of each part of the model. The ablation experiments use the Something-Something V1 dataset. This paper studies the effectiveness of the local-global temporal feature learning module (LGTFL), the regional semantic learning module (RSL), and the regional semantic fusion module (RSF). As shown in Table 3, the first row is the benchmark model of this paper. Firstly, ablation studies are conducted on the two parts in the local-global time feature learning module. When only local time enhancement (LTE) is carried out, due to the lack of attention to global features, the final effect gain is not obvious. It can be seen from row 3 and row 4 that global time memory (GTM) significantly improves the classification accuracy. This is because videos are composed of a series of frame sequence features, and global modeling is very necessary for sequence features. Ignoring global time features will have a great impact on the determination of the final video behavior category. After adding the regional semantic learning module, the model can learn the temporal variations in the video frame sequence more fully from the global temporal regional features. Therefore, the accuracy rate has also been greatly improved. The final regional semantic fusion further enhances the variable regional features and the global temporal features. Experiments show that each module proposed in this paper has improved the accuracy of behavior recognition.

4. Conclusions

This paper proposes a human bone action recognition model based on a multi-stream adaptive spatio-temporal graph convolutional network. The connection strength between the joint points is calculated using the attention mechanism and the NTN algorithm to construct the global adjacency matrix. The topK strategy is adopted to dynamically select the topK neighbor nodes, update the global adjacency matrix, and further describe the global relationship in detail. The hybrid pooling model is adopted to extract the global context features and the time key frame features, and the time features are further fully extracted. Modeling the information of joints, bones, joint movements and bone movements simultaneously further enhances the representation ability of the features extracted by the model for movements. The experimental results show that the proposed model has achieved good performance in the task of human bone action recognition and effectively improved the accuracy of action recognition. The experimental results not only prove the advantages of multi-stream input in capturing action features, but also provide a reference for subsequent research. Future work can further explore the fusion methods of other information flows and their

Table 1. Comparison experiments on Something-Something V1&V2

Model	FLOPs(G)	Param(M)	V1		V2	
			top1/%	top5/%	top1/%	top5/%
FEXNet [26]	$37 \times 1 \times 1$	35	48.2	76.3	60.2	66.5
SAM-STI [27]	$52 \times 1 \times 1$	33	49.8	78.5	60.7	87.1
ST-Adapter [28]	$71 \times 1 \times 1$	31	51.6	79.6	61.9	87.9
TPS [29]	$72 \times 1 \times 1$	31	52.3	80.1	62.8	88.6
SIFA-Net [30]	$25 \times 3 \times 1$	30	53.1	80.3	64.9	88.9
STDN [31]	$43 \times 3 \times 2$	29	54.8	81.6	65.2	89.5
GSF [32]	$67 \times 3 \times 2$	24	54.9	81.9	65.9	89.7
FMENet [33]	$71 \times 3 \times 2$	26	55.6	82.5	71.3	90.5
NTN-AGC	$73 \times 3 \times 2$	28	58.7	86.4	73.8	91.2

Table 2. Comparison experiments on Kinetics-400

Method	top1/%	top5/%
SAM-STI	71.9	89.9
T-STFT	75.1	90.9
FEXNet	75.5	91.2
GSF	74.9	92.3
NTN-AGC	75.9	93.1

Table 3. The influence of different modules on the model

LGTFL		RSL	RSF	top-1/%	difference value/%
LTE	GTM				
×	×	×	×	46.2	0
√	×	×	×	46.9	0.7
×	√	×	×	48.3	2.1
√	√	×	×	48.7	2.5
√	√	√	×	49.4	3.2
√	√	√	√	49.8	3.6

impact on the model performance, while expanding the diversity of datasets and scenarios to further enhance the generalization ability and application scope of the model.

References

[1] N. Manakitsa, G. S. Maraslidis, L. Moysis, and G. F. Fragulis, (2024) *"A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision"* **Technologies *12***(2): 15. DOI: 10.3390/technologies12020015.

[2] D. Guo, K. Li, B. Hu, Y. Zhang, and M. Wang, (2024) *"Benchmarking micro-action recognition: Dataset, methods, and applications"* **IEEE Transactions on Circuits and Systems for Video Technology *34***(7): 6238–6252. DOI: 10.1109/TCSVT.2024.3358415.

[3] S. Yin, H. Li, A. A. Laghari, T. R. Gadekallu, G. A. Sampedro, and A. Almadhor, (2024) *"An anomaly detection model based on deep auto-encoder and capsule graph convolution via sparrow search algorithm in 6G Internet of Everything"* **IEEE Internet of Things Journal *11***(18): 29402–29411. DOI: 10.1109/JIOT.2024.3353337.

[4] M. Antoun and D. Asmar, (2023) *"Human object interaction detection: Design and survey"* **Image and Vision Computing *130***: 104617. DOI: 10.1016/j.imavis.2022.104617.

[5] H. Zhou, W. Zhou, Y. Zhou, and H. Li, (2021) *"Spatial-temporal multi-cue network for sign language recognition and translation"* **IEEE Transactions on Multimedia *24***: 768–779. DOI: 10.1109/TMM.2021.3059098.

[6] P. Ravbar, K. Branson, and J. H. Simpson, (2019) *"An automatic behavior recognition system classifies animal behaviors using movements and their temporal context"* **Journal of neuroscience methods *326***: 108352. DOI: 10.1016/j.jneumeth.2019.108352.

[7] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, (2018) *"Beyond temporal pooling: Recurrence and temporal convolutions for gesture*

recognition in video" **International Journal of Computer Vision *126***(2): 430–439. DOI: 10.1007/s11263-016-0957-7.

[8] Y. Jiang and S. Yin, (2023) *"Heterogenous-view occluded expression data recognition based on cycle-consistent adversarial network and K-SVD dictionary learning under intelligent cooperative robot environment"* **Computer Science and Information Systems *20***(4): 1869–1883. DOI: 10.2298/CSIS221228034J.

[9] Y. Zhou, X. Yan, Z.-Q. Cheng, Y. Yan, Q. Dai, and X.-S. Hua. "Blockgcn: Redefine topology awareness for skeleton-based action recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 2049–2058. DOI: 10.1109/CVPR52733.2024.00200.

[10] Y. Abbas and A. Jalal. "Drone-based human action recognition for surveillance: a multi-feature approach". In: *2024 International Conference on Engineering & Computing Technologies (ICECT)*. IEEE. 2024, 1–6. DOI: 10.1109/ICECT61618.2024.10581378.

[11] Y. Ma and R. Wang, (2024) *"Relative-position embedding based spatially and temporally decoupled Transformer for action recognition"* **Pattern Recognition *145***: 109905. DOI: 10.1016/j.patcog.2023.109905.

[12] M. Munsif, N. Khan, A. Hussain, M. J. Kim, and S. W. Baik, (2024) *"Darkness-adaptive action recognition: Leveraging efficient tubelet slow-fast network for industrial applications"* **IEEE Transactions on Industrial Informatics *20***(12): 13676–13686. DOI: 10.1109/TII.2024.3431070.

[13] I. A. Abro and A. Jalal. "Multi-Modal Sensors Fusion for Fall Detection and Action Recognition in Indoor Environment". In: *2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE)*. IEEE. 2024, 1–6. DOI: 10.1109/ETECTE63967.2024.10823705.

[14] N. Siddiqui, P. Tirupattur, and M. Shah. "DVANet: Disentangling view and action features for multi-view action recognition". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. ***38***. 5. 2024, 4873–4881. DOI: 10.1609/aaai.v38i5.28290.

[15] S. Rakshit, T. Davies, M. M. Moradi, G. McSwiggan, G. Nair, J. Mateu, and A. Baddeley, (2019) *"Fast kernel smoothing of point patterns on a large network using two-dimensional convolution"* **International Statistical Review *87***(3): 531–556. DOI: 10.1111/insr.12327.

[16] X. Ma, W. Xu, H. Guan, and X. Zhang, (2024) *"Three-dimensional image recognition of soybean canopy based on improved multi-view network"* **Industrial Crops and Products *222***: 119544. DOI: 10.1016/j.indcrop.2024.119544.

[17] Y. Zhang, J. Li, N. Jiang, G. Wu, H. Zhang, Z. Shi, Z. Liu, Z. Wu, and X. Liu, (2023) *"Temporal transformer networks with self-supervision for action recognition"* **IEEE Internet of Things Journal *10***(14): 12999–13011. DOI: 10.1109/JIOT.2023.3257992.

[18] S. Basak, P. Corcoran, R. McDonnell, and M. Schukat, (2022) *"3D face-model reconstruction from a single image: A feature aggregation approach using hierarchical transformer with weak supervision"* **Neural Networks *156***: 108–122. DOI: 10.1016/j.neunet.2022.09.019.

[19] H. Qiu and B. Hou, (2024) *"Multi-grained clip focus for skeleton-based action recognition"* **Pattern Recognition *148***: 110188. DOI: 10.1016/j.patcog.2023.110188.

[20] M. Batool, M. Alotaibi, S. R. Alotaibi, D. A. AlHammadi, M. A. Jamal, A. Jalal, and B. Lee, (2024) *"Multimodal human action recognition framework using an improved CNNGRU classifier"* **IEEE Access *12***: 158388–158406. DOI: 10.1109/ACCESS.2024.3481631.

[21] K. Peng, C. Yin, J. Zheng, R. Liu, D. Schneider, J. Zhang, K. Yang, M. S. Sarfraz, R. Stiefelhagen, and A. Roitberg. "Navigating open set scenarios for skeleton-based action recognition". In: *Proceedings of the AAAI conference on artificial intelligence*. ***38***. 5. 2024, 4487–4496. DOI: 10.1609/aaai.v38i5.28247.

[22] R. Xie, Y. Jiang, and J. Yu. "Two-stream adaptive graph convolutional network with multi-head attention mechanism for industrial safety detection". In: *Fifth International Conference on Telecommunications, Optics, and Computer Science (TOCS 2024)*. ***13629***. SPIE. 2025, 710–716. DOI: 10.1117/12.3067995.

[23] Z. Bai, Q. Ding, H. Xu, J. Chi, X. Zhang, and T. Sun, (2022) *"Skeleton-based similar action recognition through integrating the salient image feature into a center-connected graph convolutional network"* **Neurocomputing *507***: 40–53. DOI: 10.1016/j.neucom.2022.07.080.

[24] W. Hong, W. Xu, J. Qi, and Y. Weng, (2019) *"Neural tensor network for multi-label classification"* **IEEE Access *7***: 96936–96941. DOI: 10.1109/ACCESS.2019.2930206.

[25] L. Teng, Y. Qiao, M. Shafiq, G. Srivastava, A. R. Javed, T. R. Gadekallu, and S. Yin, (2023) *"FLPK-BiSeNet: Federated learning based on priori knowledge and bilateral segmentation network for image edge extraction"* **IEEE Transactions on Network and Service Management *20***(2): 1529–1542. DOI: 10.1109/TNSM.2023.3273991.

[26] S. Jang, H. Lee, W. J. Kim, J. Lee, S. Woo, and S. Lee, (2024) *"Multi-scale structural graph convolutional network for skeleton-based action recognition"* **IEEE Transactions on Circuits and Systems for Video Technology *34***(8): 7244–7258. DOI: 10.1109/TCSVT.2024.3375512.

[27] E. Dastbaravardeh, S. Askarpour, M. Saberi Anari, and K. Rezaee, (2024) *"Channel attention-based approach with autoencoder network for human action recognition in low-resolution frames"* **International Journal of Intelligent Systems *2024***(1): 1052344. DOI: 10.1155/2024/1052344.

[28] H. Xu, Y. Gao, Z. Hui, J. Li, and X. Gao, (2025) *"Language knowledge-assisted representation learning for skeleton-based action recognition"* **IEEE Transactions on Multimedia *27***: 5784–5799. DOI: 10.1109/TMM.2025.3543034.

[29] S. B. Khobdeh, M. R. Yamaghani, and S. K. Sareshkeh, (2024) *"Basketball action recognition based on the combination of YOLO and a deep fuzzy LSTM network: SB Khobdeh et al."* **The Journal of Supercomputing *80***(3): 3528–3553. DOI: 10.1007/s11227-023-05611-7.

[30] A. O. Kolawole, M. E. Irhebhude, and P. O. Odion, (2025) *"Human Action Recognition in Military Obstacle Crossing Using HOG and Region-Based Descriptors"* **Journal of Computing Theories and Applications *2***(3): 410–426. DOI: 10.62411/jcta.12195.

[31] M. A. Khan, K. Javed, S. A. Khan, T. Saba, U. Habib, J. A. Khan, and A. A. Abbasi, (2024) *"Human action recognition using fusion of multiview and deep features: an application to video surveillance"* **Multimedia tools and applications *83***(5): 14885–14911. DOI: 10.1007/s11042-020-08806-9.

[32] A. C. Cob-Parro, C. Losada-Gutiérrez, M. Marrón-Romera, A. Gardel-Vicente, and I. Bravo-Munoz, (2024) *"A new framework for deep learning video based Human Action Recognition on the edge"* **Expert Systems with Applications *238***: 122220. DOI: 10.1016/j.eswa.2023.122220.

[33] X. Wang, S. Zhang, J. Cen, C. Gao, Y. Zhang, D. Zhao, and N. Sang, (2024) *"Clip-guided prototype modulating for few-shot action recognition"* **International Journal of Computer Vision *132***(6): 1899–1912. DOI: 10.1007/s11263-023-01917-4.