Postprint: Action Recognition Algorithm Based on Adaptive Partitioning and Association of Keyframe Nodes

Authors: Liu Suolan, Tian Zhenzhen, Gu Jiahui, Zhou Yuejing

Date: 2022-06-06T00:00:00+00:00

Abstract

In video-based human action recognition tasks, the fact that most frames do not contain important discriminative information seriously interferes with the accuracy of recognition applications. Key pose frames can both represent the video content and reduce computational load, while skeleton data contains higher-dimensional information compared to image data. Therefore, we propose an action recognition algorithm based on adaptive partitioning and association of keyframe skeleton nodes. Firstly, an adaptive pooling deep network is constructed to evaluate frame importance and obtain key pose frame sequences. Secondly, inter-node associations in non-natural connection states are established through a node self-learning model. Finally, the improved spatio-temporal information is applied to STGCN and SoftMax classification is employed for recognition. Comparisons with several typical techniques on the open-source large-scale datasets NTU-RGB+D and Kinetics verify that the proposed method can retain key action information while reducing redundant data volume, and the action recognition accuracy is improved by an average of 0.63%~11.81%.

Full Text

Preamble

Vol. 39 No. 10
Application Research of Computers
ChinaXiv Partner Journal

Action Recognition Based on Adaptive Partition and Association of Key-Frame Nodes

Liu Suolan[1],[2], Tian Zhenzhen[1], Gu Jiahui[1], Zhou Yuejing[1]
[1] School of Computer & Artificial Intelligence, Changzhou University,

Changzhou, Jiangsu 213164, China
[2] Jiangsu Key Laboratory of Social Security Image & Video Understanding, Nanjing University of Science & Technology, Nanjing 210094, China

Abstract

In video-based human action recognition tasks, most frames do not contain important discriminative information, which severely interferes with recognition accuracy. Key pose frames can effectively represent videos while reducing computational load, and skeletal data contains richer multi-dimensional information compared to RGB images. Therefore, this paper proposes an action recognition algorithm based on adaptive partition and association of key-frame skeletal nodes. First, an adaptive pooling deep network is constructed to evaluate frame importance and obtain key pose sequences. Second, a node self-learning model establishes inter-node associations in non-natural connection states. Finally, the improved spatio-temporal information is applied to STGCN with SoftMax classification for recognition. Compared with several typical techniques on the open-source large-scale datasets NTU-RGB+D and Kinetics, the proposed method reduces redundant data while retaining key action information, achieving an average improvement in action recognition accuracy of 0.63%~11.81%.

Keywords: action recognition; key poses; adaptive; node association; STGCN

0 Introduction

In recent years, scholars have proposed numerous video action recognition methods and introduced several public action datasets [**?**]. Graph Convolutional Networks (GCN) originated from Convolutional Neural Networks (CNN) [**?**, **?**], generalizing neural networks for application to topological graph structures to replace traditional graph processing methods such as graph regularization and kernel methods. Their superior performance has garnered widespread attention in video applications [**?**]. Caramalau et al. [**?**] applied GCN to human action recognition tasks, using sequential graph convolutional networks for active learning training applied to training data and sampling processing to identify and discard redundant unlabeled data streams for obtaining effective annotated samples. This method demonstrates better performance than traditional methods in subsequent recognition, classification, and prediction tasks.

Compared with other modalities that exhibit insufficient robustness when facing complex backgrounds, multiple viewpoints, and occlusion while also consuming more computational resources, human skeletal information is clearer, more intuitive, and less susceptible to external interference, offering relatively good adaptability [**?**]. Typically represented using 2D or 3D coordinates, skeletal data

contains rich spatial and temporal domain information that strengthens correlations between adjacent joints [?]. For instance, Vemulapalli et al. [?] expressed skeletons as a series of rigid body motions, using a special Euclidean group representation for 3D geometric relationships between rigid bodies mapped to points in Lie group space, achieving good classification results with SVM. With the successful application of deep learning, skeletal modeling methods based on deep learning have rapidly emerged, with increasing experts using skeletal modalities for human detection and action recognition [?]. Wang et al. [?] proposed a video-based action recognition model (Temporal Segment Network, TSN) that combines sparse temporal sampling strategies with video-level supervision, enabling effective learning using entire videos. Qiu et al. [?] proposed a 3D action recognition method using CNN and 3D information to construct 3D trajectories of joints from skeletal sequences, converting each skeleton sequence into three clips composed of several frames before using deep neural networks for spatio-temporal feature learning. Yan et al. [?] extended graph convolutional networks to spatio-temporal graph models (Spatial Temporal Graph Convolutional Networks, STGCN), designing a general representation for skeletal sequences in action recognition. The STGCN model maps nodes to human joints, constructing multi-layer spatio-temporal graph convolutions that integrate information along both spatial and temporal dimensions, achieving significant performance improvements on benchmark action recognition datasets and attracting widespread attention.

Building upon this, [?] represented skeletal data as a directed acyclic graph based on motion correlations between human joints and bones, designing a Directed Graph Neural Network (DGNN) to extract information about joints, bones, and their interrelationships for prediction. To better adapt to action recognition tasks, they also performed adaptive representation of graph topology during training, achieving significant performance improvements. Cheng et al. [?] combined the shift structure from CNNs with GCN, proposing an improved Shift Graph Convolutional Network (Shift-GCN) to overcome existing limitations. Shift-GCN replaces traditional graph convolution operations with new shift graph operations and lightweight pointwise convolutions, where shift operations provide flexible receptive fields for spatial and temporal graphs while performing CNN shift operations on temporal TCN, greatly reducing model parameters and computational complexity.

[?] described a new multi-stream attention-enhanced adaptive graph convolutional neural network (MS-AAGCN) for skeleton-based action recognition, where graph topology can be learned uniformly or individually based on end-to-end input data. This data-driven approach increases model flexibility and generality to adapt to different data samples. Additionally, spatio-temporal attention modules further enhance adaptive graph convolutional layers, enabling the model to focus on important joints, frames, and features. Synchronous modeling of joint and bone motion information under a multi-stream framework improves recognition accuracy. Zhao et al. [?] proposed a key frame extraction method based on node-weighted contributions combined with STGCN for

multi-feature fusion, effectively improving recognition accuracy. In 2021, Liu et al. [**?**] proposed an Adaptive Attention Memory Graph Convolutional Network (AAM-GCN) for human skeleton data, using attention mechanisms to extract key frames from skeleton sequences to obtain more discriminative temporal features.

In summary, research shows that combining temporal and spatial dimension information with GCN models is widely applied in skeleton-based human action recognition [**?**]. However, since most frame images in videos do not contain motion information (static), including them in network training can negatively impact the training process. Therefore, to address interference and information redundancy, key frame extraction has become an important preprocessing step in video-based action recognition.

---

1 Human Action Recognition Based on STGCN

STGCN is a model based on graph convolutional neural networks that enhances spatio-temporal connections and has achieved remarkable performance in skeleton-based action recognition. However, GCN models still have certain issues. For example, heuristic setting and fixing of graph topology across all model layers and input data may not be suitable for the hierarchical structure of GCN models and the complexity and diversity of data in action recognition tasks. Although two-stream or multi-stream networks learn spatial adjacency matrices or enhance spatial graph expression capabilities through incremental adaptive modules, their performance remains limited by the model structure itself. Moreover, GCN methods typically have considerably high computational complexity, often exceeding 15 GFLOPs per action sample [**?**]. The application of incremental modules, multi-stream fusion strategies, and directed acyclic graph networks further increases computational complexity dramatically, and extracting different features also requires substantial computational overhead.

Typically, video action recognition treats all frames in a sequence as equally important, which prevents focusing on the most representative frames and results in high computational costs. Extensive research has demonstrated that extracting key frames from videos for human behavior analysis can effectively reduce the impact of redundant data on computation while still utilizing pose information to describe motion information and express action categories. In daily actions such as "walking," the process contains multiple states where the human body presents different postures at each moment. In frame sequences, the target object is sometimes upright, while in other frames, even with pose changes, the contribution of consecutive frames to action recognition is redundant due to action continuity. Therefore, effectively measuring the contribution of frame images to action recognition to extract more informative key pose frames would help reduce computational load. Based on this, this paper designs a deep reinforcement sub-network to automatically learn and determine the importance of

different frames in a sequence, enabling important frames to play a more active role in classification and reducing computational complexity.

Furthermore, research shows that not all joints are equally important for action discrimination in skeleton-based action recognition. Some actions are closely related to certain sets of joints, while others are associated with different sets. For example, "making a phone call" is mainly related to head, shoulder, elbow, and wrist joints, with minimal relationship to leg joints. However, discriminating actions like "walking" or "kicking" primarily requires association calculations through leg nodes. Based on this observation, this paper optimizes the node partition model according to current key frame spatio-temporal information and historical information, establishing associations in non-natural connection states through multi-level node partitioning, enabling node associations in sequences to be adaptively optimized over time to improve model robustness and recognition accuracy.

2 Proposed Method

This paper adopts an STGCN-based approach for human action recognition. To reduce the impact of redundant information on recognition efficiency, a video key frame extraction and skeleton node association construction method is proposed based on the STGCN model. First, video data is fed into a deep reinforcement sub-network to learn and determine the importance of different frames in the sequence to obtain key information frames, and skeletal information is extracted through pose estimation. Second, a node self-learning model associates different nodes in the sequence to measure the impact of node motion changes on recognition. In the STGCN module, higher-level feature maps are generated by combining key node temporal and spatial information, and finally, action classification is performed using a SoftMax classifier.

2.1 Key Frame Discrimination and Extraction

Videos contain richer information than static images, but in practice, most frames in a video segment may not contain important action discrimination information (e.g., static frames). Including these frames in network training can negatively affect the training process. Therefore, effectively discriminating redundant information and extracting key frames is an important research topic in this field. However, existing key frame extraction algorithms lack ground truth, requiring key frames to be automatically generated based on inter-sequence correlations.

The key pose frame discrimination and skeleton extraction process proposed in this paper is shown in Figure 1 [Figure 1: see original paper]. (a) Sample M frames from the original RGB video sequence through sampling; (b) Obtain spatio-temporal features of candidate frames and feed them into a multilayer perceptron network module to compute inter-frame feature differences and

predict their importance for action recognition; (c) Compute deep features of candidate frames, using both current pooled features and inter-frame feature differences as inputs in the pooling stage, enabling the network to focus on previously unnoticed features to determine whether they are key frames, obtaining an adaptive pooling vector pooledF containing only key frames; (d) Output key image frames through the deep network and obtain key frame skeletons using OpenPose pose estimation. The proposed model aims to utilize spatio-temporal features of frames and express frame importance through reinforcement learning predictions of inter-frame differences, effectively enhancing frame discriminability and removing redundant information.

Figure 1 Key pose discrimination model based on adaptive pooling network

Let the training sample be denoted as $\{X_i, y_i\}$, where X_i is the i-th training image and y_i is the corresponding action class label. The initial candidate frame set obtained from the training sequence through sampling is represented as $\{a_j\}$, with j and i being sampling-related indices. Frame features are obtained through deep networks $\phi(a_j)$. Since sampling tends toward "more rather than less," the convolutional layer output may still contain substantial redundant information. Introducing weight parameters in the pooling stage can effectively highlight key frames while weakening the influence of secondary frames. Therefore, a multilayer perceptron network model with attention mechanism is designed to predict the importance of each initial candidate frame, defined as follows:

In Equation (1), $\Phi(\cdot)$ is the attention mechanism computation function, $f_{pre}(\cdot)$ represents the feature of the previous frame image in the original sequence, and the learning process is as follows: λ_i represents the importance weight value of candidate frame a_i.

The adaptive pooling process considers not only the deep features of candidate frames but also utilizes inter-frame information to measure the importance of contained information for recognition, obtaining the pooling vector pooledF containing only key frames for subsequent behavior recognition tasks.

2.2 Node Association Model

After pose estimation processing to obtain key frame skeleton sequences, the graph $G = (V, E)$ is constructed from video data, where V represents the set of nodes and E represents the set of edges. The node set $s \in \mathbb{R}^{N\times 3}$ contains three-dimensional coordinates of N joints (e.g., in the Kinetics dataset using OpenPose to extract 18 joints, $N = 18$). Natural human node associations and fixed partition schemes often cannot effectively express all actions occurring in a sequence. This paper designs a node association learning model that dynamically optimizes partitions based on sequence content, which can not only enhance/weaken connections between nodes within certain neighborhoods but also create associations between nodes without direct connections.

A node association learning model is established on the GCN basis to select

partitions for nodes and enhance model adaptability. The adjacency matrix is defined as $A_k = S_k + T_k + Q_k$, where S_k is the base adjacency matrix representing natural connections between joints, T_k represents temporal constraints, and Q_k represents the attention matrix. The attention mechanism measures the actual dependency relationships between a node and other nodes in the current state graph, determining whether to connect and the connection strength. The key to algorithm implementation is obtaining dependency weight values for node pairs (s_i, s_j) that discriminate actions, computed as follows:

$\theta(\cdot)$ and $\gamma(\cdot)$ respectively compute the angle and positional information of the current node relative to a reference source point. To reduce parameters, computational complexity, and prevent overfitting, attention values between any two nodes can be further computed using node trajectories, with temporal dimensions fused into Q_k. During training, multi-level partitions and associations are typically predefined based on action class labels. For example, with 18 joints extracted, associations can reach up to 7 levels. Figure 2 [Figure 2: see original paper] shows a schematic diagram of level-2 partitioning and association using the elbow node as an example.

Figure 2 Node level-2 zoning and association diagram

---

3 Experiments

3.1 Datasets and Settings

1) NTU-RGB+D Dataset [**?**]: This dataset contains 56,880 video samples covering 40 daily action categories (including drink water, throw, clapping, phone call, shaking hands), 9 health-related actions, and 11 two-person interaction actions. Data was collected from 40 subjects aged 10-35 performing each action twice using three Microsoft Kinect v2 sensors placed at different horizontal angles (-45°, 0°, and 45°). Data modalities include depth information, 3D skeleton information, RGB frames, and infrared sequences. The dataset provides two evaluation protocols: (a) Cross-Subject, using videos from 20 subjects (IDs: 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38) totaling 40,320 videos as the training set and the remaining 16,560 samples as the test set; (b) Cross-View, partitioning training and test sets by camera ID, with 18,960 samples from camera 1 as the test set and 37,920 samples from cameras 2 and 3 as the training set.

2) Kinetics-400 Dataset [**?**]: This dataset contains approximately 300,000 videos collected from YouTube, covering 400 action categories including person-object interaction actions (abseiling, applauding, feeding fish, opening bottle, playing piano, yoga) and person-person interaction actions, with at least 400 video samples per category and each video lasting approximately 10 seconds.

To reduce the impact of video parameter differences on subsequent processing, this paper adjusts all video frames to a resolution of 340$×$256 and converts

the frame rate to 30 fps, as shown in Figure 3 [Figure 3: see original paper].

Figure 3 Samples from NTU-RGB+D and Kinetics-400 datasets

3.2.1 Key Frame Extraction Experimental Settings and Results Analysis

In the key frame discrimination and extraction stage, the MLP is configured as a four-layer perceptron network using hyperbolic tangent (tanh) and sigmoid activation functions in the final layer. Adding nonlinear functions as hidden layers effectively addresses gradient vanishing issues. The pooling vector initial state is set to match the first frame' s feature vector, with subsequent inputs to the adaptive module using the feature vector difference between the current frame and its previous frame.

The number of hidden layer neurons directly affects perceptron network performance and key frame extraction effectiveness. Too few neurons lead to suboptimal accuracy, while too many cause overfitting and poor convergence. Considering differences in action category numbers and samples across datasets, this paper dynamically estimates the number of hidden layer neurons N_{hidden} as $N_{hidden} = \sqrt{N_{input} + N_{output}} + \alpha$, where N_{input} and N_{output} represent input and output layer neuron counts, $N_{training}$ is the number of training samples, and the adjustment parameter α ranges from 1 to 10.

To validate the effectiveness of the proposed key frame extraction algorithm, comparisons were made with two commonly used algorithms: optical flow method based on motion analysis and video clustering algorithm. The optical flow method selects key frames as those with minimum local motion optical flow, while clustering uses k-means to select frames with minimum distance to cluster centers. A random video clip "robot dancing" from the Kinetics dataset was selected for demonstration, containing 301 frames over approximately 10 seconds. The clustering algorithm extracted 76 key frames (2, 19, 34, ⋯), the optical flow method extracted 16 key frames (4, 48, 79, 107, ⋯), and the proposed algorithm extracted 5 key frames (18, 91, 199, 263, 268). The clustering method performed poorly with high redundancy despite a 25.1% compression rate, primarily due to severe human interference in initializing cluster centers and threshold selection directly affecting the number of selected key frames. The optical flow method reflects static states in video data by computing motion magnitude, making it highly dependent on video shot structure selection. The proposed algorithm effectively identifies and removes repetitions and redundancies, compressing the video to 5 key frames while completely reflecting two key actions in the video, with pose estimation effectively detecting key joints of moving targets.

Figure 4 Key frame extraction of three algorithms and pose estimation results

3.2.2 NTU-RGB+D(Cross-Subject) Action Recognition Results and Analysis

This work compares with the STGCN model [**?**] under three partition strategies (uniform, distance, and spatial) and our previously reported research [**?**]. In the node partition and association algorithm, parameters θ and ρ are set, with optimal performance parameters obtained by varying these values. Experiments show that setting $\theta = 3$ and $\rho = 1$ strengthens intrinsic associations while appropriately enhancing extrinsic associations and reducing computation from connections. Model performance is evaluated using top-1 and top-5 metrics. Comparison methods use an initial learning rate of 0.01, reduced to 0.1× at epoch 80. The proposed algorithm' s recognition rates are obtained through repeated experiments with dynamically adjusted learning rates per iteration. All experiments run on Ubuntu 16.04 with a 1060-6GB GPU using the PyTorch deep learning framework.

Figures 5 [Figure 5: see original paper] and 6 [Figure 6: see original paper] show top-1 and top-5 recognition rate comparisons on NTU-RGB+D(Cross-Subject). The model gradually optimizes during training, with the proposed method achieving improvements over [**?**] and [**?**] in both top-1 and top-5 performance. The maximum top-1 accuracy improvement reaches 3.84% at epoch 50, while top-5 accuracy improvement peaks at 1.34% at epoch 45. Recognition rates stabilize at approximately 82% and 97% between epochs 50-80, proving that improving human skeleton node partition and association in STGCN effectively enhances model recognition performance. Particularly after setting appropriate learning rates per epoch, the results outperform our previous work, further demonstrating that suitable learning rate parameters positively impact model performance.

Figure 5 Experimental results of TOP-1 on NTU-RGB+D(Cross-Subject)

Figure 6 Experimental results of TOP-5 on NTU-RGB+D(Cross-Subject)

3.2.3 NTU-RGB+D(Cross-View) Action Recognition Results and Analysis

Table 1 shows experimental comparison results on NTU-RGB+D(Cross-View). The proposed method improves recognition accuracy by approximately 2.82% and 0.63% in top-1 and top-5 respectively compared to the spatial method in [**?**] under identical conditions, and by 5.21% and 1.07% compared to the best results in [**?**]. The improvement is more pronounced than on NTU-RGB+D(Cross-Subject), primarily because Cross-View only changes data collection angles while training and test data come from all action performers, greatly reducing intra-class differences in action recognition. Consequently, recognition rates on NTU-RGB+D(Cross-View) are comparable to comparison methods.

Table 1 Comparison results on NTU-RGB+D(Cross-View)

Methods	top-1	top-5
STGCN19	-	-
文献 [25]	-	-
Ours	-	-

3.2.4 Comparison with Other Algorithms on NTU-RGB+D

The proposed algorithm is compared with several typical methods: Lie Group [**?**], Deep-LSTM [**?**], TSN (Temporal Segment Networks) [**?**], and Clips+CNN+MTLN [**?**]. Lie Group expresses human actions as feature vectors in manifold space, modeling spatio-temporal correlations between frames to form Lie Group feature sequences for classification. Deep-LSTM consists of three LSTM layers and fully connected layers. TSN' s main advantage over common two-stream networks lies in addressing long-term video action discrimination and overfitting from small samples, using sparse frame sampling to remove redundancy and reduce computation. Clips+CNN+MTLN first divides skeleton sequences into three clips, then uses CNN to learn skeleton information in sequential frameworks and processes generated parallel clips using a Multi-Task Learning Network (MTLN) to merge spatial structure information for video action recognition.

As shown in Table 2 , the proposed algorithm improves recognition accuracy by 3.92% and 2.47% on Cross-View and Cross-Subject protocols respectively compared to the best-performing Clips+CNN+MTLN. Comparison reveals that recognition accuracy under different viewpoints generally outperforms that under different subjects, with a maximum difference of 7.13%, primarily because different performers exhibit large action variations due to behavioral habits even when performing the same actions, directly affecting recognition accuracy. Lie Group focuses more on spatial information while weakening temporal information impact, leading to suboptimal results. Deep-LSTM excels in processing time series but key frame extraction weakens temporal features, resulting in poor recognition rates. Although Deep-LSTM, TSN, and Clips+CNN+MTLN consider temporal node information during motion, they fail to effectively establish spatial node associations and ignore translation and scale variation impacts. The proposed method compresses video frames while retaining behavioral temporal features through key frames and constructs logical spatial changes for non-associated nodes, better expressing spatio-temporal action characteristics and demonstrating superior recognition performance on this dataset.

Table 2 Best results of top-1 with other compared methods

Methods	X-View	X-Subject
Lie Group[14]	-	-
Deep LSTM[4]	-	-
TSN[15]	-	-

Methods	X-View	X-Subject
Clips+CNN+MTLN[16]	-	-
Ours	-	-

3.2.5 Kinetics Action Recognition Results and Analysis

Table 3 shows experimental results on the Kinetics dataset after model improvement. With appropriately set iterative learning rates, the proposed training results achieve improvements of 2.71% and 2.34% in top-1 and top-5 respectively compared to the original STGCN model. However, test results are far below expectations compared to NTU-RGB+D, primarily because NTU-RGB+D has fixed camera positions and controlled environments, while Kinetics videos from YouTube have non-fixed acquisition patterns with large camera motions causing poor data stability, increasing difficulty for pose estimation and inter-node information association. Nevertheless, compared to [**?**], key frame processing and strengthened human topology associations show more obvious improvements for such challenging video data.

Table 3 Comparison results on Kinetics

Methods	top-1	top-5
STGCN19	-	-
文献 [25]	-	-
Ours	-	-

3.2.6 Comparison on Kinetics Dataset

Comparison methods include feature coding [**?**], Deep LSTM [**?**], and TSN [**?**]. Feature coding extracts human motion information from skeleton sequences and uses sparse coding to obtain feature vectors for action recognition. Table 4 compares recognition performance in terms of top-1 and top-5 accuracy. The proposed method improves recognition rates by 9.08% and 11.81% in top-1 and top-5 respectively compared to Deep LSTM. However, average recognition rates on this dataset are overall low (below 50%), primarily because most videos are captured in open environments with handheld devices causing high instability and large camera motions resulting in high action blur, making skeleton extraction difficult. Therefore, even with node adaptive partition and association models, recognition accuracy remains low.

Table 4 Comparison results with other three methods on Kinetics

Methods	Top-1	Top-5
Deep LSTM[4]	-	-
Feature Coding[26]	-	-

Methods	Top-1	Top-5
TSN[15]	-	-
Ours	-	-

4 Conclusion

In video-based action recognition tasks, redundant information often leads to long training times and high resource demands, affecting effective recognition accuracy. Practice has proven that using key pose frame images can effectively discriminate action categories. Based on this, this paper introduces a deep reinforcement sub-network into the STGCN model to learn and evaluate the importance of different frames in sequences, extracting key frames to represent videos and obtaining skeletal information through pose estimation. Node adaptive models learn associations between nodes in non-natural connection states to expand the impact of node motion changes on recognition. Testing on the challenging large-scale datasets NTU-RGB+D and Kinetics shows performance improvements over several mainstream recognition techniques including Feature Coding, Lie Group, Deep LSTM, TSN, and Clips+CNN+MTLN. Notably, although the node association model can establish connections between nodes under non-natural partitions, parameters θ and ρ should not be too large to reduce computational complexity. This paper selects optimal parameters through repeated experiments; therefore, establishing an appropriate objective function to further automatically optimize parameter selection is an important direction for future research.

References

[1] Zhang Hongbo, Zhang Yixiang, Zhong Bineng, et al. A comprehensive survey of vision-based human action recognition methods [J]. Sensors 2019, 19: 1005-1025.

[2] Wang Zhengwei, She Qi, Smolic A. ACTION-NET: Multipath excitation for action recognition [C]. //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2021. https://doi.org/10.48550/arXiv.2103.07372.

[3] Wang Limin, Tong Zhan, Ji Bin Ji, et al. TDN: Temporal Difference Networks for Efficient Action Recognition [C]. //Proc of IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ: IEEE Press, 2021. https://doi.org/10.48550/arXiv.2012.10071.

[4] Shahroudy A, Liu Jun, Tsong N T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis [C]. //Proc of IEEE Conference on Computer

Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 1010-1019.

[5] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset [EB/OL]. (2017). https://doi.org/10.48550/arXiv.1705.06950.

[6] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]. //Proc of International Conference on Neural Information Processing Systems, 2012, 1097-1105.

[7] Ji Shuiwang, Xu Wei, Yang Ming, et al. 3D Convolutional neural networks for human action recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence (PAMI), 2012, 35(1): 221-231.

[8] 徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述 [J]. 计算机学报, 2020, 43(5): 755-780. (Xu Bingbing, Cen Keting, Huang Junjie, et al. A survey on graph convolutional neural network [J]. Chinese Journal of Computers, 2020, 43(5): 755-780.)

[9] Duvenaud D, Maclaurin D, Aguilera J, et al. Convolutional networks on graphs for learning molecular fingerprints [J]. Advances in Neural Information Processing Systems, 2015: 2224-2232.

[10] 谢昭, 周义, 吴克伟, 等. 基于时空关注度 LSTM 的行为识别 [J]. 计算机学报, 2021, 44(2): 261-274. (Xie Zhao, Zhou Yi, Wu Kewei, et al. Activity Recognition Based on Spatial-Temporal Attention LSTM [J]. Chinese Journal of Computers, 2021, 44(2): 261-274.)

[11] Caramalau R, Bhattarai B, Kim T K. Sequential graph convolutional network for active learning [C]. //Proc of IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ: IEEE Press, 2021. https://doi.org/10.48550/arXiv.2006.10219.

[12] Zhao Hang, Torralba A, Torresani L, et al. Hacs: Human action clips and segments dataset for recognition and temporal localization [C]. //Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 8668-8678.

[13] Li Kunchang, Li Xianhang, Wang Yali, et al. CT-net: Channel tensorization network for video classification [C]. //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018. https://doi.org/10.48550/arXiv.2106.01603.

[14] Vemulapalli R, Arrate F, Chellappa R, et al. Human action recognition by representing 3D skeletons as points in a lie group [C]. //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 588-595.

[15] Wang Limin, Xiong Yuanjun, Wang Zhe, et al. Temporal Segment Networks: Towards good practices for deep action recognition [C]. //Proc of IEEE Conference on European Conference on Computer Vision, 2016. https://doi.org/10.48550/arXiv.1608.00859.

[16] Qiu Hongke, Mohammed B, Senjian A, et al. A new representation of skeleton sequences for 3D action recognition [C]. //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017. https://doi.org/10.48550/arXiv.1703.03492.

[17] Christoph F, Hao Qifan, Jitendra M, et al. Slow fast networks for video recognition [C]. //Proc of IEEE International Conference on Computer Vision, 2019: 6201-6210.

[18] Cheng Xiaopeng, Feng Dapeng. Skeleton embedded motion partition for human action recognition using depth sequences [J]. Signal Processing, 2018, 143: 56-68.

[19] Yan Sijie, Xiong Yuanjun, Lin Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition [C]. //Proc of the Thirty-second AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2018, 32(01): 7444-7452.

[20] Shi Lei, Zhang Yifang, Cheng Jian, et al. Skeleton-Based Action Recognition with Directed Graph Neural Networks [C]. //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 7912-7921.

[21] Cheng Ke, Zhang Xifan, He Xiangyu, et al. Skeleton-based action recognition with shift graph convolutional network [C]. //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 183-192.

[22] Shi Lei, Zhang Yifang, Cheng Jian, et al. Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks [J]. IEEE Trans on Image Processing, 2020, 29: 9532-9545.

[23] Zhao Yuerong, Guo Hongbo, Gao Ling, et al. Multifeature fusion action recognition based on key frames [J]. Concurrency & Computation Practice & Experience, 2021, 2021(3). https://doi.org/10.1002/cpe.

[24] Liu Di, Xu Hui, Wang Jianzhong, et al. Adaptive Attention Memory Graph Convolutional Networks for Skeleton-Based Action Recognition [J]. Sensors, 2021, 21: 6761-6780.

[25] 刘锁兰, 顾嘉晖, 王洪元, 等. 基于关联分区和 ST-GCN 的人体行为识别 [J]. 计算机工程与应用, 2021, 57(3): 168-178. (Liu Suolan, Gu Jiahui, Wang Hongyuan, et al. Human behavior recognition based on associative partition and ST-GCN [J]. Computer Engineering and Application, 2021, 57(3): 168-178.)

[26] Jian Meng, Zhang Shuai, Wu Lifang, et al. Deep key frame extraction for sport training [J]. Neurocomputing, 2019, 328: 147-156.

[27] 陈煜平, 邱卫根. 基于视觉的人体行为识别算法研究综述 [J]. 计算机应用研究, 2019, 36(7): 1927-1934. (Chen Yuping, Qiu Weigen. Survey of human action recognition

algorithm based on vision [J]. Application Research of Computers, 2019, 36(7): 1927-1934.)

[28] Zhang Yixiang, Zhang Hongbo, Du Jixiang, et al. RGB+2D skeleton: local hand-crafted and 3D convolution feature coding for action recognition [J]. Signal, Image and Video Processing, 2021, 2021(15): Article ID 1386.

Note: Figure translations are in progress. See original paper for figures.