基于图注意与有向图神经网络的人体动作识别

詹 源[1],明山水[2],田 元[2]

(1. 武汉广播电视台播送中心,湖北 武汉 430022;2. 华中师范大学 人工智能教育学部,湖北 武汉 430079)

摘 要: 基于人体骨骼数据的图卷积神经网络不易受背景环境噪声影响且鲁棒性较强,已成为现阶段人体动作识别领域的研究重点,但该网络对同阶邻域中不同邻域赋予相同权值,限制了其捕捉空间信息相关性的能力。为此,引入图注意网络加权和求和相邻节点的特征,允许每个节点根据其相邻特征分配不同权重,以增强特征提取和学习能力。同时,为解决将骨架表示为无向图时只能确定相邻节点或边之间的关系,从而限制了捕获节点或边之间依赖关系能力这一问题。引入有向图卷积,利用一阶和二阶相邻节点的特征信息进行图卷积,既保留了有向图的方向性特征,又扩展了图卷积的感知域,从而能够提取更多特征。实验表明,所提方法能有效提升动作识别的精度。

关键词: 动作识别;图神经网络;图注意;有向图

DOI: 10. 11907/rjdk. 231708 **开放科学(资源服务)标识码(OSID):**

中图分类号: TP391.14 **文献标识码:** A **文章编号:** 1672-7800(2024)009-0176-05

Human Action Recognition Method Based on Graph Attention Network and Directed Graph Neural Network

ZHAN Yuan[1], MING Shanshui[2], TIAN Yuan[2]

(1. *Wuhan Broadcasting and Television Station*, *Wuhan* 430022, *China*;
2. *Faculty of Artificial Intelligence in Education*, *Central China Normal University*, *Wuhan* 430079, *China*)

Abstract: The graph convolutional neural network based on human skeleton data is not easily affected by background environmental noise and has strong robustness, which has become a research focus in the field of human action recognition at present. However, this network assigns the same weight to different neighborhoods in the same order, which limits its ability to capture spatial information correlations. To this end, a graph attention network weighted sum is introduced to sum the features of adjacent nodes, allowing each node to assign different weights based on its adjacent features to enhance feature extraction and learning effectiveness. At the same time, in order to solve the problem of representing the skeleton as an undirected graph, only the relationship between adjacent nodes or edges can be determined, which limits the ability to capture dependency relationships between nodes or edges. Introducing directed graph convolution, utilizing the feature information of first-order and second-order adjacent nodes for graph convolution, not only preserves the directional features of the directed graph, but also expands the perceptual domain of graph convolution, thereby extracting more features. The experiment shows that the proposed method can effectively improve the accuracy of action recognition.

Key Words: action recognition; graph neural network; graph attention; directed graph

0 引言

人体动作识别在人机交互、犯罪监控、虚拟现实和医疗保健等领域具有巨大的探索和应用潜力。因此,许多研究者一直在尝试提出更准确、更高效的动作识别方法,人体动作识别是指对视频片段中人类行为的识别和分类[1-3]。传统方法将动态视频分解为静态帧,然后人工提取特征进行识别和分类。随着计算机硬件飞速发展,深度网络特征学习策略已成为计算机视觉领域的主流研究方

收稿日期: 2023-07-03
扫描二维码阅读全文:
基金项目: 信息化与基础教育均衡发展省部共建协同创新中心研究项目(xtzd2022-008)

作者简介: 詹源(1977-),男,武汉广播电视台播送中心工程师,研究方向为视音频制作及传播;明山水(1998-),女,华中师范大学人工智能教育学部硕士研究生,研究方向为深度学习、行为识别;田元(1982-),女,博士,华中师范大学人工智能教育学部副教授、硕士生导师,研究方向为人工智能及教育应用。本文通讯作者:田元。

向,基于该策略的方法大致可分为基于 RGB 图像的方法[4-6]、基于骨骼数据的方法[7-9]。

由于骨骼数据容易从高精度深度相机和姿态估计算法中获得,因此有许多基于骨骼的动作识别模型和方法用于动作识别任务,其中最常用的方法是基于卷积神经网络(Convolutional Neural Network, CNN)的骨骼动作识别方法和基于循环神经网络(Recurrent Neural Network, RNN)的骨骼动作识别方法。前者主要思想是通过设计转换规则将骨架数据转换为类似图像的特征图后作为输入,然后利用 CNN 进行特征提取,以达到动作识别的目的。Ke 等[10]将每帧中任意两个骨骼连接点之间的距离映射到图像中,并将图像馈送到基于 ImageNet 的 CNN 模型中实现动作分类。Li 等[11]结合注意力机制,通过线性变换对重要骨骼关节自动重排和选择,对动作进行分类。

基于 RNN 的方法利用其在处理时间序列上的优势进行动作识别。Du 等[12]提出分层递归神经网络将不同身体部位的特征分层结合。在浅层中每个子网络提取单个节点上的特征后在深层融合,当所有节点信息融合后进行最终的动作识别。Du 等[8]基于长短期记忆 RNN 构建模型,采用注意机制对不同框架和关节分配不同权重的注意力,构建两个子网络实现端到端的动作为识别。虽然,基于 CNN 和 RNN 的方法都被广泛应用,但由于以矢量或类似图像的二维网格形式描述骨架序列无法准确表示骨架数据的拓扑结构,所以骨架空间结构信息无法得到较好的保留。

近年来,图卷积网络得到了迅速发展,由于骨节点数据是一种可被表示为图的自然发生的拓扑结构,因此图卷积网(Graph Convolutional Network, GCN)相较于卷积神经网络(Convolutional Neural Networks, CNN)更适合处理这类数据。Velickovic 等[13]提出基于注意机制的图注意网络(Graph Attention Networks, GAT),利用叠加隐藏自注意层对邻域相应节点赋予不同权重,实现了图结构数据上的节点分类。Yan 等[14]提出时空图卷积网络(Spatial-Temporal Graph Convolutional Network, ST-GCN),通过 2D 或 3D 关节坐标表示人体骨骼信息,并采用时空图处理视频内的动态动作信息。Thakkar 等[15]在此基础上将骨架图划分为 4 个子图,分别对 4 个子图进行图卷积以提升模型识别精度。

虽然,ST-GCN 能有效捕获图中节点间的时空依赖关系且动作识别效果较好,但其使用的图卷积网络对同阶邻域中的不同邻域赋予相同的权值,限制了其捕捉空间信息的能力。另外,由于 ST-GCN 将骨骼表示为无向图,关节点和骨骼分别表示为节点和边,而无向图只能确定相邻节点或边之间的关系,限制了其捕获不相邻节点或边之间依赖关系的能力。

为此,本文引入 GAT 对相邻节点特征进行加权和求和,允许每个节点根据其相邻特征分配不同权重,以增强模型的特征提取和学习能力。同时,通过有向图神经网络(Directed Graph Neural Network, DGNN),利用有向图表示骨骼,结合关节和骨骼信息进行特征提取,以充分提取动作过程中关节点之间的依赖关系[16]。

1　本文方法

1.1　网络结构

本文提出了一种基于图注意网络和有向图神经网络的人体动作识别方法,其本质为基于多流数据的人体动作识别框架。

通过在骨骼识别任务中获取关节信息、骨骼信息和运动信息,然后捕获图中节点和边缘的依赖性,最后结合注意机制提升人体动作识别精度。整体网络构架分为多流数据输入、GAT、DGNN、数据融合与输出 4 个部分,如图 1 所示。

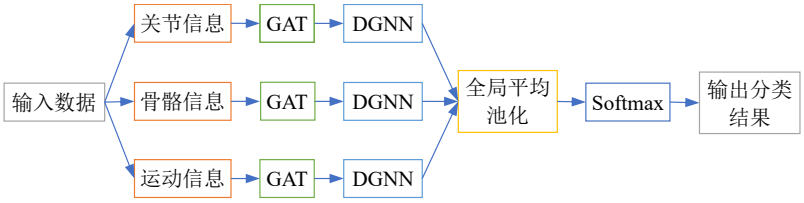


Fig. 1　Overall network architecture

图 1　整体网络构架

首先,提取运动过程中的关节信息、骨骼信息和运动信息。关节信息指 ST-GCN 中的原始拓扑图;骨骼信息指骨骼的长度和方向;运动信息则是同一关节点在两个相邻时间帧之间的位置差值。其次,通过有向图表示人体骨架,并对有向图进行图卷积。模型还可自主学习网络参数优化网络,引入图注意力网络 GAT 使模型获取节点的邻域特征,为不同节点赋予不同权重,在网络不同层次内获得不一样的拓扑图结构,通过迭代获得适合不同种类动作的拓扑图,进而提升模型的动作识别准确率。最后,通过 Softmax 融合 3 个输入流的数据得到识别结果。

1.2　图注意网络

有向图和无向图最大的区别是邻接矩阵为不对称矩阵。本文为了在有向图上进行图卷积引入图注意网络,修改空间图的卷积层不仅能是模型学习网络参数,还能优化连通性图,以获得更适合描述动作的图结构,提高动作识别准确性。

假设一个图有 n 个节点,节点的 $\boldsymbol{F}$ 维特征集表示为 $h = \vec{h}_1, \vec{h}_2, \cdots, \vec{h}_n, \vec{h}_i \in \boldsymbol{R}^F$。采用图注意机制后,输出新节点特

征集表示为 $h' = \vec{h}'_1, \vec{h}'_2, \cdots \vec{h}'_n, \vec{h}'_i \in \boldsymbol{R}^{F'}$。将输入特征通过线性变换转化为高阶特征后，利用自注意（self-attention）为每个节点分配注意权值。对于目标节点，只计算其邻域节点与目标节点之间的相关性。

为了更好地在不同节点间分配权重，使用 Softmax 归一化目标节点与所有相邻节点之间的相关性。

$$\alpha_{ij} = Softmax\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k \in N_i} \exp\left(e_{ik}\right)} \tag{1}$$

式中：e_{ij} 表示两个节点之间的边向量。

GAT 的引入可帮助网络模型更好地对每个样本的动作进行建模，增加模型个性化。本文使用两个卷积层，将节点 v_{ti} 的输入特征 $h(v_{ti})$ 映射到 $\boldsymbol{X}$ 向量和 $\boldsymbol{Y}$ 向量。

$$\begin{cases} X_{ti} = \boldsymbol{W}_X f(v_{ti}) \\ Y_{ti} = \boldsymbol{W}_Y f(v_{ti}) \end{cases} \tag{2}$$

式中：$\boldsymbol{W}_X$、$\boldsymbol{W}_Y$ 为两个卷积层的权重矩阵；$\boldsymbol{X}_{ti}$、$\boldsymbol{Y}_{ti}$ 的内积 $\boldsymbol{P}_{(\mathrm{t,i})\to(\mathrm{t,j})} = \left\langle \boldsymbol{X}_{ti}, \boldsymbol{Y}_{ti} \right\rangle$ 表示 v_{ti} 与 v_{tj} 之间的相似性，然后使用 Softmax 函数将其归一化。

$$\alpha_{(\mathrm{t,i})\to(\mathrm{t,j})} = \frac{\exp\left(\boldsymbol{P}_{(\mathrm{t,i})\to(\mathrm{t,j})}\right)}{\sum_{n=1}^{N} exp\left(\boldsymbol{P}_{(\mathrm{t,i})\to(\mathrm{t,n})}\right)} \tag{3}$$

在训练过程中，利用初始的物理连接关系实时更新边的权值，以优化连通图拓扑结构。通过学习训练数据适应不同的动作样本，一方面模型通过学习不同动作样本可有效学习到不同动作中任意两个身体关节点的权值，以增加模型多功能性，使模型能在面对不同数据时有效预测动作；另一方面加入图注意机制能让网络在训练过程中不断优化图结构，适应多个样本的变化，形成最适合描述动作的拓扑结构，以提升模型性能，使动作预测结果更准确。

1.3 有向图卷积网络

本文引入有向图卷积，利用一阶、二阶相邻节点的特征信息进行图卷积，既保留了有向图的方向性特征，又扩展了图卷积的感知域，从而提取到更多特征。有向图卷积利用节点的一阶邻接关系和二阶邻接关系得到的扩展卷积计算所得，在有向图卷积中节点 v_i、v_j 如果存在连接边 $e_ij \in \boldsymbol{E}$ 则 e_ij 是一阶边；节点 v_i、v_j、v_k 如果存在边 $e_ik \in \boldsymbol{E}$ 和 $e_jk \in \boldsymbol{E}$ 则 $e_jk \in \boldsymbol{E}$ 是一条二阶边。一阶最邻近节点表示直接相邻的节点，二阶最邻近节点划分为一阶邻近节点和远邻节点。一阶邻近矩阵的公式表示为：

$$\boldsymbol{A}_F\left(v_i, v_j\right) = \boldsymbol{A}^{sym}\left(v_i, v_j\right) \tag{4}$$

式中：$\boldsymbol{A}^{sym}$ 为邻接矩阵 $\boldsymbol{A}$ 的对称矩阵，相当于将一个有向图转换为对应的无向图的前导矩阵。

然而，上述转换方法会导致信息丢失，因此本文通过二阶相似度来保存信息。二阶邻近矩阵连接相似的节点计算公式为：

$$\boldsymbol{A}_{S_{in}}\left(v_i, v_j\right) = \sum\nolimits_k \frac{\boldsymbol{A}_{(v_k, v_i)} \boldsymbol{A}_{(v_k, v_j)}}{\sum_n \boldsymbol{A}_{(n,k)}} \tag{5}$$

$$\boldsymbol{A}_{S_{out}}\left(v_i, v_j\right) = \sum\nolimits_k \frac{\boldsymbol{A}_{(v_i, v_k)} \boldsymbol{A}_{(v_j, v_k)}}{\sum_n \boldsymbol{A}_{(n,k)}} \tag{6}$$

式中：$\boldsymbol{A}$ 为加权邻接矩阵，节点之间的二阶相似度由连接到其共享邻近节点的边的归一化权值之和决定，反映了节点 v_i、v_j 之间的相似性；$\boldsymbol{A}_{S_{in}}(i,j)$ 越大表示两个节点间的二阶相似度越高。

在有向图中，一阶和二阶邻近矩阵具有与无向图卷积相似的函数。根据无向图的卷积公式可得：

$$\boldsymbol{Z}_F = \left(\boldsymbol{D}_F^{-\frac{1}{2}} \boldsymbol{A}_F \boldsymbol{D}_F^{-\frac{1}{2}}\right) \boldsymbol{X\Theta} \tag{7}$$

$$\boldsymbol{Z}_{S_{in}} = \left(\boldsymbol{D}_{S_{in}}^{-\frac{1}{2}} \boldsymbol{A}_{S_{in}} \boldsymbol{D}_{S_{in}}^{-\frac{1}{2}}\right) \boldsymbol{X\Theta} \tag{8}$$

$$\boldsymbol{Z}_{S_{out}} = \left(\boldsymbol{D}_{S_{out}}^{-\frac{1}{2}} \boldsymbol{A}_{S_{out}} \boldsymbol{D}_{S_{out}}^{-\frac{1}{2}}\right) \boldsymbol{X\Theta} \tag{9}$$

式中：$\boldsymbol{X}$ 为节点特征向量；$\boldsymbol{\Theta}$ 为滤波器参数矩阵；$\boldsymbol{Z}$ 为卷积结果。

然后，采用信息融合方法，在获取信息的同时保持有向结构。

$$\boldsymbol{Z} = \mathrm{Concat}(\boldsymbol{Z}_F, \alpha \boldsymbol{Z}_{S_{in}}, \beta \boldsymbol{Z}_{S_{out}}) \tag{10}$$

式中：Concat(·)表示矩阵的连通性；α、β 为权重系数，反映不同邻近节点的重要性。当二阶邻近节点较少时可减少二阶邻近节点权重，使用更多的一阶邻近节点信息进行计算。

在有向图卷积网络的实现中，有向图每个节点的输入数据是一个维数为 $C\times T\times N_v$ 的张量 f_v，C 表示通道数，T 表示动作视频帧数，N_v 表示骨架数据中的节点数。同样，有向图每条边的输入数据是一个维数为 $C\times T\times N_e$ 的张量 f_e，N_e 表示骨架数据的边数。本文在传统卷积层的每个有向图网络块后增加一个批归一化（Batch Normalization，BN）层和一个整流线性单元（Rectified Linear Unit，ReLU）层，通过有向图卷积网络处理可有效捕获关节与骨骼之间的依赖关系，从而提升模型识别性能。

2 实验结果与分析

本文实验系统为 Windows 10，GPU 为 NVIDIA GeForce RTX 3080，在 PyCharm 平台进行开发，通过 PyTorch 深度学习训练框架评估模型。首先使用 NTU-RGB+D 数据集在 X-Sub 和 X-View 基准测试集上对本文模型进行测试，结果如表 1 所示。由此可见，本文方法相较于 DGNN 方法在 X-Sub 和 X-View 上分别提高 1.3%、1%。在 Kinetics-Skeleton 数据集中，将基于 RNN 的 Deep LSTM[22]、基于 CNN 的 TCN[20]、基于 GCN 的 ST-GCN[14]、基于 GNN 的 DGNN[16] 与本文模型进行比较，实验结果如表 2 所示。

图 2（a）、图 2（b）分别为 DGNN 和本文方法的混淆矩阵。由此可见，本文方法的动作识别精度均有所提升，其中动作“sit down”的准确率提升幅度最大，达到 1.6%。

Table 1　Comparison of accuracy of various methods on the NTU-RGB+D dataset

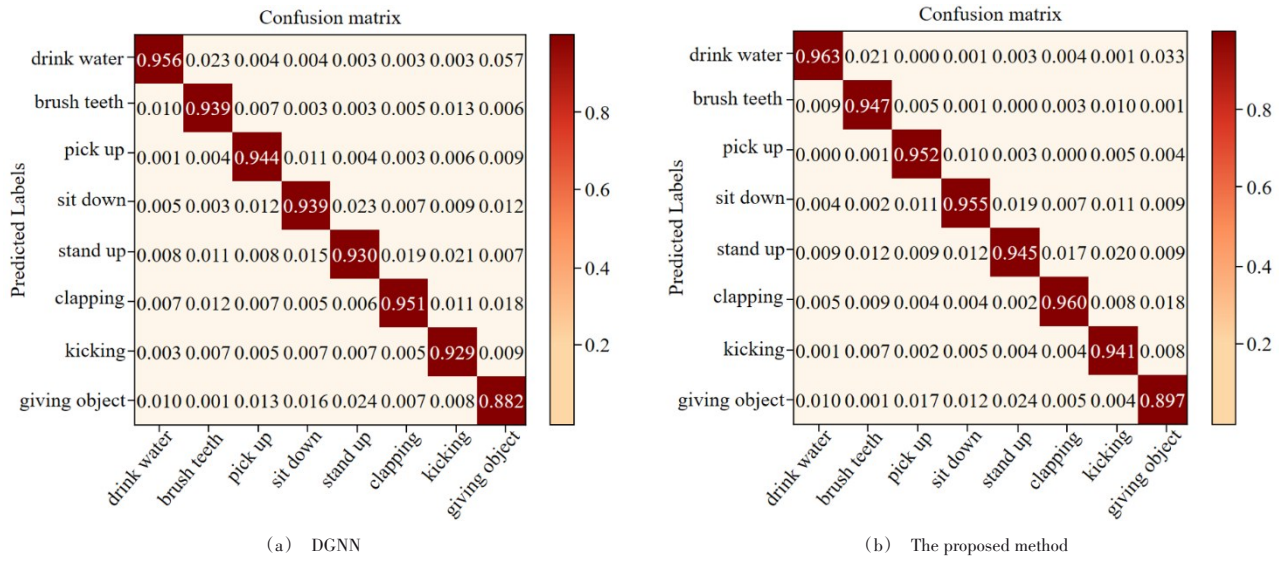
表 1　NTU-RGB+D 数据集上各方法精度比较

方法	X-Sub	X-View
Lie Group[17]	50.1	82.8
STA-LSTM[18]	73.4	81.2
GCA-LSTM[19]	76.1	84.0
TCN[20]	74.3	83.1
ST-GCN[14]	81.5	88.3
2s-AGCN[21]	88.5	95.1
DGNN[16]	89.9	96.1
本文方法	**91.2**	**97.1**

Table 2　Comparison of accuracy of various methods on the Kinetics-Skeleton dataset

表 2　在 Kinetics-Skeleton 数据集上各方法精度比较

方法	Top-1	Top-5
Deep LSTM[22]	16.4	35.3
TCN[20]	20.3	40.0
ST-GCN[14]	30.7	52.8
2s-AGCN[21]	36.1	58.7
DGNN[16]	36.9	59.6
本文方法	**37.3**	**60.3**



（a）　DGNN　　（b）　The proposed method

Fig. 2　Confusion matrices of DGNN and the proposed method

图 2　DGNN 与本文方法的混淆矩阵

同时,本文为了进一步验证方法的有效性,在 NTU-RGB+D 数据集上进行消融实验,以 Top-1 模型的识别精度为评价标准,使用多流框架关节信息、骨骼信息和运动信息作为模型输入数据。

表 3 中 MS-DGCN-J 表示只输入关节信息,MS-DGCN-B 表示只输入骨骼信息,MS-DGCN-M 表示只输入运动信息,MS-DGCN-J+B 表示输入关节信息和骨骼信息,MS-DGCN-J+M 表示输入关节信息和运动信息,MS-DGCN-B+M 表示输入骨骼信息和运动信息。实验表明,使用两个输入信息流相较于一个信息流的识别性能更好,使用 3 个信息流相较于两个信息流的识别性能更好。

Table 3　Ablation experiments

表 3　消融实验

方法	Top-1	Top-5
ST-GCN[11]	81.5	88.3
2s-AGCN[18]	88.5	95.1
DGNN[13]	89.9	96.1
MS-DGCN-J	90.8	96.7
MS-DGCN-B	88.9	95.1
MS-DGCN-M	91.0	96.9
MS-DGCN-J+B	90.7	96.7
MS-DGCN-J+M	91.1	97.0
MS-DGCN-B+M	89.5	96.4
本文方法	**91.2**	**97.1**

3 结语

本文提出一种基于图注意网络和有向图神经网络的人体动作识别方法。首先采用基于多流数据输入的框架将人体骨骼中的关节信息、骨骼信息和运动信息作为网络输入;其次利用GAT机制获取网络各层关键节点的时空信息,提升了模型动作识别精度;最后将骨架图表示为有向图,以捕获关节和骨骼之间的依赖关系。

在NTU-RGB+D和Kinetics-Skeleton两个大型数据集上的实验表明,本文方法可有效提升动作识别的精度。未来,将进一步优化模型各个部分,以使模型在各种数据集上都具备较高的识别精度。

参考文献:

[1] KONG Y, FU Y. Human action recognition and prediction: a survey[J]. International Journal of Computer Vision, 2022, 130(5): 1366-1401.

[2] JEGHAM I, KHALIFA A B, ALOUANI I, et al. Vision-based human action recognition: an overview and real world challenges[J]. Forensic Science International: Digital Investigation, 2020, 32: 200901.

[3] HERATH S, HARANDI M, PORIKLI F. Going deeper into action recognition: a survey[J]. Image and Vision Computing, 2017, 60: 4-21.

[4] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299-6308.

[5] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision, 2015: 4489-4497.

[6] WANG L, XIONG Y J, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]// European Conference on Computer Vision, 2016: 20-36.

[7] WEINZAEPFEL P, ROGEZ G. Mimetics: towards understanding human actions out of context[J]. International Journal of Computer Vision, 2021, 129(5): 1675-1690.

[8] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1110-1118.

[9] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human action recognition by representing 3D skeletons as points in a lie group[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 588-595.

[10] KE Q, BENNAMOUN M, AN S, et al. A new representation of skeleton sequences for 3D action recognition[DB/OL]. https://arxiv.org/pdf/1703.03492v2.pdf.

[11] LI C, ZHONG Q, XIE D, et al. Skeleton-based action recognition with convolutional neural networks[C]// Proceedings of the IEEE International Conference on Multimedia Expo Workshop, 2017: 597-600.

[12] DU W, WANG Y, YU Q. RPAN: an end-to-end recurrent pose-attention network for action recognition in videos[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 3725-3734.

[13] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[DB/OL]. https://arxiv.org/pdf/1710.10903.pdf

[14] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[DB/OL]. https://arxiv.org/pdf/1801.07455.pdf.

[15] THAKKAR K, NARAYANAN P J. Part-based graph convolutional network for action recognition[DB/OL]. https://arxiv.org/pdf/1809.04983.pdf.

[16] SHI L, ZHANG Y, CHENG J, et al. Skeleton-based action recognition with directed graph neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 7912-7921.

[17] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human action recognition by representing 3D skeletons as points in a lie group[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 588-595.

[18] SONG S, LAN C, XING J, et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017: 4263-4270.

[19] LIU J, WANG G, DUAN L Y, et al. Skeleton based human action recognition with global context-aware attention LSTM networks[J]. IEEE Transactions on Image Processing, 2018, 27(4): 1586-1599.

[20] KIM T S, REITER A. Interpretable 3D human action analysis with temporal convolutional networks[C]// IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017: 20-28.

[21] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 12026-12035.

[22] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+D: a large scale dataset for 3D human activity analysis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1010-1019.

(责任编辑:刘嘉文)