

基于全局分割图卷积网络的骨骼双人交互行为识别

徐寅虎

(南京邮电大学 计算机学院、软件学院、网络空间安全学院, 江苏 南京 210023)

摘要: 图卷积网络(GCN)在双人交互识别方面应用非常广泛,但是传统图卷积网络对于节点特征的学习不够充分,尤其在双人交互行为中,每个节点的特征往往包含多个人的信息,导致节点特征不够准确。为了有效提取双人交互行为关节点之间的相关性特征,聚合双人之间的特征信息,提出全局分割图卷积网络(GS-GCN),进行基于骨骼的双人交互行为识别。GS-GCN包含全局分割图卷积(GSGC)和层次聚合注意力(HAA)模块,GSGC将图卷积(GCN)和全局分割图(GS-Graph)相结合,将双人交互识别看成全局性识别,并具有多个邻接矩阵,可提取单人特征和双人全局特征。此外,单人的层次和双人层次的重要性各不相同,为了突出交互行为之间的交互信息,引入层次聚合的注意力模块(HAA),突出交互行为中更加明显的语义信息。在NTU-RGB+D、NTU-RGB+D 120和SBU双人交互数据集上进行实验,验证了模型有效性,与其他双人交互识别方法相比,该模型性能更优。

关键词: 双人交互识别;图卷积网络;邻接矩阵;注意力模块

DOI:10.11907/rjtk.231307

开放科学(资源服务)标识码(OSID):



中图分类号:TP183;TP391.41

文献标识码:A

文章编号:1672-7800(2024)006-0157-06

Skeleton-based Double Human Interaction Recognition Based on Global Segmentation Graph Convolutional Network

XU Yinhu

(School of Computer Science, School of Software, School of Cyberspace Security, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Graph convolutional network (GCN) is widely used in interaction recognition, but the traditional graph convolutional network is not enough to learn node features. Especially in interaction, the features of each node often contain information of multiple people, resulting in inaccurate node features. In order to effectively extract the correlation features between the nodes between interaction behaviors and aggregate the feature information, a global segmentation graph convolution network (GS-GCN) is proposed, which consists of global segmentation graph convolution (GSGC) and hierarchical aggregation attention (HAA) modules. GSGC combines graph convolution (GCN) and global segmentation graph (GS-Graph), treats two-person interaction recognition as global recognition, has multiple adjacency matrices, and extracts single-person feature and two-person global feature. In addition, the importance of single level and double level is different. In order to highlight the interactive information between interaction behaviors, the attention module of hierarchical aggregation (HAA) is introduced to highlight the more obvious semantic information in interaction behaviors. Through experiments on NTU-RGB+D, NTU-RGB+D 120 and SBU two-person interaction data sets, the validity of the model is verified, and it is obviously superior to other two-person interaction recognition methods.

Key Words: human interaction recognition; graph convolution network; adjacency matrix; attention module

0 引言

随着计算机视觉和人工智能技术的不断发展,双人交互行为研究越来越受到关注^[1-2]。双人交互行为是指两个

人之间的相互作用,例如握手、拥抱、击掌等。对于这些行为的识别和分析在人机交互、智能监控、虚拟现实等领域具有广泛应用^[3-4]。近年来,随着深度学习技术的兴起,基于骨骼识别的双人交互行为研究逐渐成为热门领域。骨骼识别是一种基于RGB^[5]或深度图像^[6]的技术,通过识别人体骨骼关节位置和运动轨迹推断人体行为。相比传统

收稿日期:2023-05-31

作者简介:徐寅虎(1998-),男,南京邮电大学计算机学院、软件学院、网络空间安全学院硕士研究生,研究方向为人体行为识别。本文通讯作者:徐寅虎。

的 RGB 图像和深度图像,骨骼图像具有更少的冗余信息,更能反映人体骨骼运动的特征,能够有效地提高双人交互行为识别准确度和鲁棒性。因此,基于骨骼的双人交互识别研究具有重要理论和实际意义。

传统的基于骨架识别的交互行为识别方法通常将人体关节视为独立的特征,并使用基于循环神经网络(RNN)^[7]或卷积神经网络(CNN)^[8]的模型进行动作预测。然而,这些方法忽略了交叉关节之间的相关特征,不利于交互行为识别准确性和鲁棒性。为了克服这一限制,Li 等^[9]提出一种基于骨架特征的双人互动行为识别方法,该方法通过使用图卷积网络(GCN)和长短时记忆网络(LSTM)对人体骨架序列进行建模,从而有效地捕捉人体关节之间的距离和语义信息,进而提高了交互行为识别准确性和鲁棒性。此外,Zhou 等^[10]提出一种基于骨架和视觉信息的双人互动行为识别方法,该方法将 RGB 图像和人体骨架序列进行融合,使用多层卷积神经网络(CNN)和长短时记忆网络(LSTM)对数据进行建模,从而实现了对人体动作的预测。然而,这些方法都没有考虑身体部位之间的距离和语义信息,很难学习到交互行为之间的相关特征。

目前,存在许多单人动作识别方法^[11-14],其中 Zhang 等^[14]提出了一种层次分解的图卷积网络方法,用于骨架动作识别。该方法将关节层次结构和局部相对位置信息结合起来,并利用多尺度分解和上下文编码机制捕捉不同级别的语义信息。通过引入一个层次化的拓扑结构,可以更好地对行为进行建模。然而,当处理双人交互行为时,该方法需要将双人看成两个单独的个体进行处理,从而无法捕捉双人之间的交互信息。因此,现有的单人动作识别方法在双人场景下存在一定局限性。如图 1 所示,在“握手”这个行为中交叉关节之间的行为特征尤为明显,需要关注双人之间的交互特征,“推”动作交互信息明显聚集在第一个人的手部。因此,需要更加关注双人交互行为之间的特征聚集。

为了解决上述问题,本文提出基于全局分割图卷积网络进行骨骼双人交互行为识别。该网络包含一个全局分

割图卷积和层次聚合注意力模块,以有效提取双人交互行为之间的相关性特征。全局分割图卷积将图卷积和全局分割图相结合,全局分割图将双人交互识别视为全局性识别,将双人特征信息放入同一个语义空间,在语义空间中将其划分 3 个特征空间,即 2 个单人的特征空间和 1 个双人特征空间,以提取有效提取双人交互之间的节点特征。对于不同的交互行为,交互的特征聚集不在不同的部位,比如“握手”,交互信息明显聚集在双手手部的语义空间,而对于“推”这个交互行为,交互信息明显更加聚集在单个人的手部语义空间。因此,层次聚合注意力模块将对 3 个特征空间分配自适应权重,突出更加具有相关性的交互部位。

1 相关工作

1.1 基于图卷积网络的行为识别

在基于骨架的动作识别中,人体骨骼被构造为具有关节节点的图结构。为了处理这种非欧几里得数据,图卷积被引入。ST-GCN^[15]是首个使用拓扑图提取和聚合骨骼时空信息的方法,但它使用手工制作的骨架图,无法提取非自然连接关节之间的相关特征。为此,Chen 等^[16]提出基于骨骼动作识别的通道拓扑优化图卷积 CTR-GCN 方法,同时学习共享拓扑和通道相关性,学习联合节点之间的非自然连接,增强了图卷积的特征提取和聚集能力。Zhang 等^[14]提出一种层次分解的图卷积网络方法,用于骨架动作识别。该方法将关节层次结构和局部相对位置信息结合起来,并利用多尺度分解和上下文编码机制捕捉不同级别的语义信息。通过引入一个层次化的拓扑结构,可以更好地对行为进行建模。图卷积可以对骨架数据进行端到端的建模和处理,在动作识别等领域具有良好的建模能力、鲁棒性和有效性。但是,传统图卷积都是对单人进行建模处理,忽略了双人之间交互特征的重要性。本文提出了全局分割图卷积,用于提取单人特征和双人全局特征,并捕捉双人之间的交互信息。

1.2 基于骨骼的双人交互识别

人体交互识别是动作识别的一个子领域。相较于单人动作识别,人体交互方法不仅要能够建模每个个体的行为,还要捕捉它们之间的交互信息。在此背景下,Yun 等^[17]评估了几种几何关系身体姿态特征,包括关节特征、平面特征和速度特征,以用于交互建模。结果显示,关节特征在交互建模方面的表现优于其他特征,而速度特征则对噪声敏感。此外, Ji 等^[18]通过将属于每个个体相同身体部位的关节进行分组,建立了姿势集以描述每个身体部位的交互。Perez 等^[19]提出了一种基于双流长短时记忆网络(LSTM)的交互关系网络 LSTM-IRN,用于模拟来自同一人身体关节的内部关系和来自不同人关节的间接关系。然而,LSTM-IRN 忽略了身体部位的距离演化,这在人体交互



Fig. 1 "Hand shake" and "push" interaction

图 1 “握手”和“推”的交互行为

识别中是重要的先验知识。相比上述方法,本文将双人交互识别视为全局性识别,从全局上提取交互行为的相关性特征。

1.3 基于层次聚合注意力的交互行为识别

注意力机制最初是在自然语言处理(NLP)领域被提出,是深度学习的基本组成部分,被广泛应用于计算机视觉、语音识别等领域^[20,21]。最近, Hu 等^[22]提出的 SE-net 将注意力机制引入到计算机视觉中,使用面向空间的注意力和多层感知器以提取通道信息。Woo 等^[23]进一步提出了一种卷积块注意模块 CBAM,它结合了通道注意力和空间注意力机制,通过输入中间特征图,沿着通道和空间两个独立的维度计算注意力图,再与输入特征图相乘进行自适应特征优化。在图卷积网络的背景下,对相邻节点特征求和的注意力机制被提出^[34-35],该机制的权重完全由节点特征决定,与底层图结构无关。本文提出了层次聚合注意力模块,对每个层次结构分配自适应权重,突出交互行为中的重点交互信息。

2 全局分割图卷积网络

本文提出的全局分割图卷积网络架构如图 2 所示,其输入是一个交互动作的骨骼序列。首先,以单人—单人—双人的形式进行构建 3 个邻接矩阵,并提取复杂的节点特征;然后,全局分割图卷积以提取每个人的单人特征和双人的全局特征;接着通过层次聚合注意力(HAA 模块,突出双人交互行为之间更具代表性的特征;最后,通过一个全连接层进行预测。

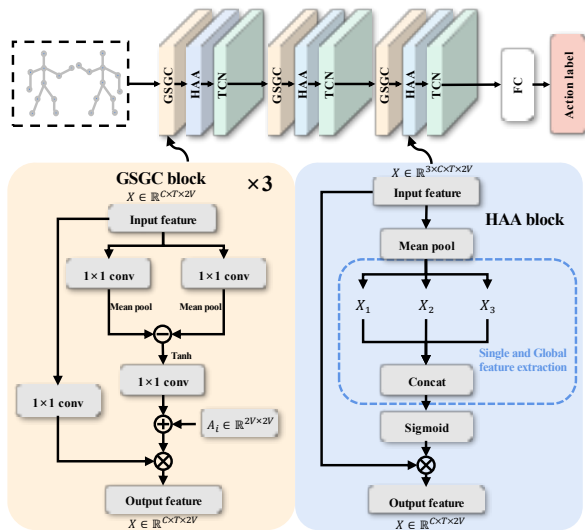


Fig. 2 Architecture of global segmentation graph convolution

图 2 全局分割图卷积架构

2.1 全局分割图

传统的人体骨骼图构造方法是使用自然连接作为边,以关节点作为点进行连接。然而,在处理双人交互时,仅仅构建单人图是不足以描述全局信息的。为了解决该问

题,本文提出了一种新的方法(见图 3),将双人交互行为转化为 3 个图:第一个人的图、第二个人的图和全局图。

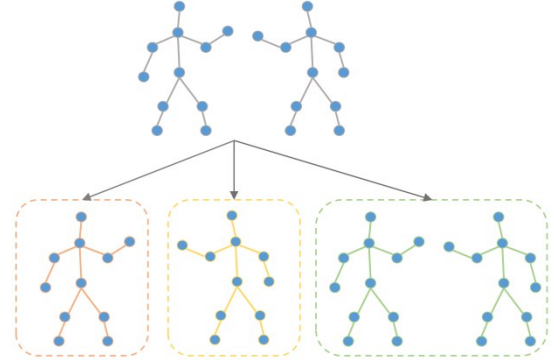


Fig. 3 Global segmentation graph

图 3 全局分割图

全局分割图构建过程包括两个关键步骤:①将两个人的关节合并为一个整体,并共享相同的特征空间;②在特征空间中将其划分为单人—单人—双人的语义空间,从而构建包含单人和全局信息的图。图的邻接矩阵代表人体关节点之间边的连接关系,它是图的拓扑结构的表示。定义 $A_1, A_2 \in \mathbb{R}^{V \times 2V}$ 为第一个人和第二个人的邻接矩阵,并在全局特征空间中对单人进行邻接矩阵构建。邻接矩阵中的节点数量 $2V$ 代表双人的关节点数量,而不是单人。传统图卷积的邻接矩阵定义中通常仅包含了单人节点,本文将另一个人的连接信息填充为零,在双人的语义空间中提取单人特征信息。为了构建全局邻接矩阵 $A_3 \in \mathbb{R}^{V \times 2V}$,在特征空间中将双人的边连接放置在一个语义空间中。将所有的邻接矩阵放置在同一个层次维度上,定义为层次邻接矩阵 $A \in \mathbb{R}^{3 \times 2V \times 2V}$, A 就包含了单人和全局特征。将 3 个邻接矩阵拼接到同一个层次维度上以提取交互特征,就完成了全局分割图构造。拼接计算公式如式(1)所示。

$$A = [A_1 // A_2 // A_3] \quad (1)$$

邻接矩阵用度矩阵进行归一化,以训练稳定性,并且将矩阵的所有元素作为可学习参数,用于训练适应性。

2.2 全局分割图卷积

使用线性变换将输入特征 $X \in \mathbb{R}^{C \times T \times 2V}$ 转化为高级特征。为了防止数据冗余,全局分割图卷积(Global Segmentation Graph Convolution, GSGC)对时间维度进行平均池化,重点关注空间特征,然后使用相关建模函数 $\psi(\cdot)$,提取不同运动下双人关节点之间的空间分布特征,并使用 ϕ 和 Φ 降低通道维度。GSGC 将全局分割图构造的层次邻接矩阵 $A_i \in \mathbb{R}^{2V \times 2V}$ 加入到关节特征中,得到关节节点 x_i 和 x_j 之间的成对相关性。随后, GSGC 将输入特征和加入了成对相关性的关节特征进行特征聚合,得到输出特征 Z 。各计算公式如式(2)一式(5)所示。

$$X^* = XW \quad (2)$$

$$\psi_{ij} = \tanh(\phi(x_i) - \phi(x_j)) \quad (3)$$

$$Z_i = X \otimes (\psi(X^*) + A_i) \quad (4)$$

$$Z = [Z_1 // Z_2 // Z_3] \quad (5)$$

其中, GSGC 通过 1×1 的卷积对输入特征进行线性变换, W 是可学习的权重矩阵。通过减法操作计算通道维度中成对关节点之间的空间分布特征, 将单层的邻接矩阵 A_i 加入到关节特征空间中以获取每个关节点之间的连接信息。GSGC 运用爱因斯坦求和约定 $\otimes$ 在层次维度上对输入特征和每个层次的输出特征进行聚合, 最终得到输出 $Z \in \mathbb{R}^{3 \times C \times T \times 2V}$ 。

2.3 层次聚合注意力

GSGC 使用一种聚合策略添加所有按层次结构的输出。然而, 由于每个数据样本在不同层次中的重要性各不相同, 本文提出了层次聚合注意力(HAA)模块。该模块对层次结构维度应用加权策略, 并对层次结构输出分配自适应权重。

HAA 模块应用于 GSGC 后的特征 $Z \in \mathbb{R}^{3 \times C \times T \times 2V}$ 。将特征 Z 在时间维度上进行平均池化, 提取空间特征。为了有效地对每个层次结构分配合适的权重, HAA 提取每个层次的代表性节点特征, 提取后对于按层次结构的输出进行空间平均池化。空间池化函数 $\mathcal{M}$ 如式(6)所示。

$$\mathcal{M}(Z_i) = \frac{1}{N_k + N_{k+1}} \sum_{v \in (V_k, V_{k+1})} \frac{1}{T} (Z_i(v)) \quad (6)$$

其中, N_k 表示第 k 个层次结构集中的顶点数。 V_k 代表每个层次结构中的代表性节点。

在提取代表性节点的层次信息时, 各层次特征并不共享信息。通过欧几里得距离学习层次特征空间中的相似度, 将带有所有层次信息的特征映射与 GSGC 输出特征映射相乘, 得到层次维度加权后的输出特征映射 H 。

$$H = \sum_{i=1}^3 X \cdot \text{sigmoid}(\alpha(\mathcal{M}_i)) \quad (7)$$

其中, 欧几里得距离 α 学习层次特征空间中的相似度。层次结构通过 sigmoid 激活函数传递, 该激活函数为每个层次结构集的特征赋予自适应权重, 并与输入特征映射相乘, 提出每个动作样本之间的层次相关性, 最终得到输出特征 $H \in \mathbb{R}^{C \times T \times 2V}$ 。

3 实验与结果比较

3.1 实验数据集

NTU-RGB+D 是一个大型动作数据集, 包含来自 60 个动作类的 56 578 个骨架序列^[27]。每一个动作都由 3 个摄像机在相同的高度从不同的水平角度拍摄。每个人体骨骼包含 25 个身体关节的三维坐标。该数据集包含 11 种人类互动类型, 包括打/拍、拍背、给予、走向、踢、指、摸口袋、分开走、推、拥抱和握手。本文将这部分交互数据集定义为 NTU60*。

NTU-RGB+D120^[28] 扩展了 NTU-RGB+D, 从 60 个额外的动作类中增加了 57 367 个样本。它总共包含 120 个动作类的 113 945 个骨架序列。除 NTU-RGB+D 中的 11 个交互类外, 该数据集还包含了 15 个人机交互类, 包括物体撞

击、挥刀、撞倒、抢东西、开枪、踩脚、击掌、欢呼和喝酒、携带物体、拍照、跟随、窃窃、交换物品、支持某人和石头剪刀布, 共 26 个交互类。本文将这部分交互数据集定义为 NTU120*。

SBU^[29] 是一个两人交互数据集, 包含了接近、离开、推、踢、拳、交换物品、拥抱和握手等 8 类人类交互。7 名参与者(配对多达 21 种不同的排列)完成了所有 8 种互动。该数据集总共包含 282 个短视频。每个视频包含每个人每帧 15 个关节的 3D 坐标。

Table 1 Ablation experimental results of global segmentation graph convolution

表 1 全局分割图卷积的消融实验结果	
方法	NTU120*的准确率/%
Baseline	87.67
Single Segmentation Graph Convolution	88.02
+ GSGC	88.52
+ GSGC + HAA	89.11

Table 2 Experimental comparison among global segmentation graph convolutional network and advanced methods

表 2 全局分割图卷积网络与先进方法的实验比较 (%)					
方法	SBU	NTU60*	NTU60*	NTU120*	NTU120*
		X-Sub	X-View	X-Sub	X-Set
FSNET ^[30]	—	74.0	80.5	61.2	69.7
ST-LSTM ^[31]	93.3	83.0	87.3	63.0	66.6
ST-GCN ^[16]	—	83.3	87.1	78.9	76.1
GCA ^[32]	—	85.9	89.0	70.6	73.7
2s-GCA ^[33]	94.9	87.2	89.9	73.0	73.3
AS-GCN ^[12]	—	89.3	93.0	82.9	83.7
ViT ^[34]	93.1	89.7	92.5	81.5	82.5
LSTM-IRN ^[35]	98.2	90.5	83.5	77.7	79.6
IGFormer ^[36]	98.3	93.6	96.5	85.4	86.5
CTR-GCN ^[17]	94.7	94.3	96.1	87.7	89.0
Our GS-GCN	98.4	95.1	97.2	89.1	90.1

3.2 实验参数设定

实验中, 采用 CTR-GCN 作为基线模型, 该模型提出了一种新颖的通道拓扑细化图卷积, 动态学习不同的拓扑结构, 并有效地聚合不同通道中的关节特征, 用于基于骨骼的动作识别。实验将 SGD 优化器的动量设置为 0.9, 权值衰减为 0.000 4。学习阶段的数量设置为 65 个, 并将热身策略应用于前 5 个阶段, 以实现更稳定的学习。将学习率设置为余弦退火衰减, 最大学习率为 0.1, 最小学习率为 0.000 1。对于交互数据集, 将批量大小设置为 16。本文所有的实验都是在 RTX 5000 GPU 上进行。

3.3 消融实验

本文对 NTU-120* 数据集进行了广泛的消融研究, 以验证所提出的全局分割图卷积和层次聚合注意力模块(HAA)的有效性。在消融研究中, 仅使用关节模态数据, 消融实验结果如表 1 所示。

首先, 本文验证了合并分裂图卷积(GSGC)的有效性。将两个人的交互视为两个单独的人的交互, 并且在构造图时只使用本文分割思想。可以发现, 本文分割思想在处理

两人交互时也很有效,精度提高了0.35%。与单一分割图卷积相比,提出的全局分割图卷积在准确性上提高了0.85%,与使用单人分割图卷积相比提高了0.5%。可以发现,本文“全局分割”思想比“单人分割”思想更容易提取交互信息。

HAA模块相比于仅加入GSGC的精度提高了0.59%,整个模型精度提高了1.44%。对于不同的动作,需要注意的层次信息也不同。如图4所示,“握手”的交互动作更加聚集在双人的手部,因而全局层次得分明显高于单人层次得分,网络可以更加关注双人之间的全局特征。对于“推”这个行为,第一个人没有明显的行为特征,第二个人的手部交互信息更加突出,故需要更加关注第二个人的相关性特征。更加关注具有代表性的特征,并不意味着不关注其他行为特征,HAA模块为单人—单人—双人的层次信息分配自适应权重,学习每个层次之间的依赖关系。

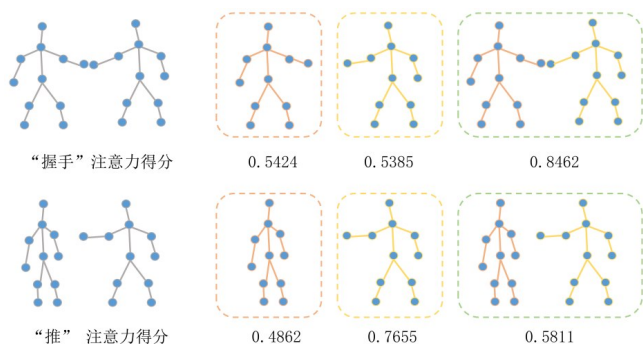


图4 层次聚合注意力分层得分

图4 层次聚合注意力分层得分

3.4 与先进方法比较

GS-GCN在NTU-RGB+D和NTU-RGB+D 120和SBU双人交互数据集上的实验结果如表2所示。本文提出的GS-GCN优于其他基于人体骨骼的交互识别方法,得益于所提出的GSGC和HAA模块,GS-GCN比基于CNN和RNN的方法获得了更好的性能。在NTU-RGB+D的X-Sub和X-View上,GS-GCN比CTR-GCN提高了1.2%和1.1%,具有更高的交互识别精度。在NTU-RGB+D120的X-Sub和X-Set上,MS-GCN比CTR-GCN的性能分别提高了1.4%和1.1%。在SBU数据集上,本文提出的GS-GCN达到了98.4%的更佳效果。

4 结语

为有效提升双人交互行为识别准确性,本文基于全局分割图卷积网络(GS-GCN)进行骨骼双人交互行为识别。该网络由全局分割图卷积(GSGC)和层次聚合注意力(HAA)模块组成,可提取单人特征和双人全局特征,并突出交互行为中更加明显的语义信息。此外,对于交互识别,本文提出了全局分割图以捕获和学习交互信息。大量实验结果表明,该模型在NTU-60和NTU-120交互数据集

上具有较好的性能。该模型有效地捕捉了动作之间的时空依赖关系和交互信息,从而提高了交互识别准确性。但是,其在处理相似度较高的交互行为时,网络很难提取更加精细的特征,在相似行为之间,特征聚集的部位更加位于个体之间的身体部位。网络提取的是全局特征,对于更加精细的特征,有待进一步改进优化。

参考文献:

- [1] JIANG Q L, DUAN K K, LIU J, et al. Skeleton interaction graph convolutional networks for dual-person action recognition [J]. Journal of Image and Graphics, 2021, 26(5): 817-827.
姜琪琳, 段珂珂, 刘杰, 等. 双人动作识别的骨骼交互图卷积网络[J]. 中国图象图形学报, 2021, 26(5): 817-827.
- [2] TANG Q, HUANG Q, MA X D, et al. Research on human interaction recognition based on deep learning [J]. Computer Engineering and Applications, 2020, 56(3): 112-119.
唐茜, 黄青, 马旭东, 等. 基于深度学习的人体交互识别研究[J]. 计算机工程与应用, 2020, 56(3): 112-119.
- [3] WANG J H, XIONG J, YANG H, et al. Research progress on human action and interaction recognition based on deep learning [J]. Computer Applications Research, 2021, 38(2): 333-339.
王俊豪, 熊津, 杨辉, 等. 基于深度学习的人体动作与交互识别研究进展[J]. 计算机应用研究, 2021, 38(2): 333-339.
- [4] LI J, ZHOU W B, ZHAO F. A survey of human action recognition methods based on deep learning [J]. Microelectronics and Computer, 2018, 35(9): 1-9.
李菊, 周文彬, 赵峰. 基于深度学习的人体动作识别方法综述[J]. 微电子学与计算机, 2018, 35(9): 1-9.
- [5] VEERIAH V, ZHUANG N, QI G J. Differential recurrent neural networks for action recognition [C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 4041-4049.
- [6] VAHDAT A, GAO B, RANJBAR M, et al. A discriminative key pose sequence model for recognizing human interactions [C]//2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011: 1729-1736.
- [7] ZHANG P, LAN C, XING J, et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data [C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2117-2126.
- [8] JIA J G, ZHOU Y F, HAO X W, et al. Two-stream temporal convolutional networks for skeleton-based human action recognition [J]. Journal of Computer Science and Technology, 2020, 35: 538-550.
- [9] ZHU L, WAN B, LI C, et al. Dyadic relational graph convolutional networks for skeleton-based human interaction recognition [J]. Pattern Recognition, 2021, 115: 107920.
- [10] XU W Y, WU M Q, ZHAO M, et al. Fusion of skeleton and RGB features for RGB-D human action recognition [J]. IEEE Sensors Journal, 2021, 21(17): 19157-19164.
- [11] ZHANG P, LAN C, ZENG W, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1112-1121.
- [12] CHEN Z, LI S, YANG B, et al. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 1113-1122.

- [13] CHENG K, ZHANG Y, CAO C, et al. Decoupling gcw with dropgraph module for skeleton-based action recognition [C]//Proceedings of 16th European Conference on Computer Vision, 2020: 536–553.
- [14] LEE J, LEE M, LEE D, et al. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition [DB/OL]. <https://arxiv.org/abs/2208.10741>, 2022.
- [15] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 7444–7452.
- [16] CHEN Y, ZHANG Z, YUAN C, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 13359–13368.
- [17] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12026–12035.
- [18] MUNEA T L, JEMBRE Y Z, WELDEGEGBRIEL H T, et al. The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation [J]. *IEEE Access*, 2020, 8: 133330–133348.
- [19] AHMAD T, JIN L, LIN L, et al. Skeleton-based action recognition using sparse spatio-temporal GCN with edge effective resistance [J]. *Neurocomputing*, 2021, 423: 389–398.
- [20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [DB/OL]. <https://arxiv.org/abs/1706.03762>, 2017.
- [21] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132–7141.
- [22] ZADENH A, LIANG P P, PORIA S, et al. Multi-attention recurrent network for human communication comprehension [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 5642–5649.
- [23] WOO S, PARK J, LEE J Y, et al. Cbam: convolutional block attention module [C]//Proceedings of the European Conference on Computer Vision, 2018: 3–19.
- [24] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12026–12035.
- [25] HU B, GUO K, WANG X, et al. RRL-GAT: graph attention network-driven multilabel image robust representation learning [J]. *IEEE Internet of Things Journal*, 2021, 9(12): 9167–9178.
- [26] 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.
- [27] LIU J, SHAHROUDY A, PEREZ M, et al. Ntu rgb+d 120: a large-scale benchmark for 3d human activity understanding [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 42(10): 2684–2701.
- [28] LIU J, SHAHROUDY A, WANG G, et al. Skeleton-based online action prediction using scale selection network [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 42(6): 1453–1467.
- [29] YUN K, HONORIO J, CHATTOPADHYAY D, et al. Two-person interaction detection using body-pose features and multiple instance learning [C]//2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012: 28–35.
- [30] LIU J, SHAHROUDY A, XU D, et al. Spatio-temporal lstm with trust gates for 3d human action recognition [C]//Proceedings of 14th European Conference on Computer Vision, 2016: 816–833.
- [31] LIU J, WANG G, HU P, et al. Global context-aware attention lstm networks for 3d action recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1647–1656.
- [32] 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*, 2017, 27(4): 1586–1599.
- [33] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [DB/OL]. <https://arxiv.org/abs/2010.11929>, 2020.
- [34] PEREZ M, LIU J, KOT A C. Interaction relational network for mutual action recognition [J]. *IEEE Transactions on Multimedia*, 2021, 24: 366–376.
- [35] PANG Y S, KE Q H, RAHMANI H, et al. IGFormer: interaction graph transformer for skeleton-based human interaction recognition [C]//Proceedings of 17th European Conference on Computer Vision, 2022: 605–622.

(责任编辑:孙娟)