DOI: 10.13382/j.jemi.B2407949

基于多尺度可变形图卷积的双人交互行为识别*

王 丽[1] 曹江涛[1] 谢 帅[1] 姬晓飞[2]
(1. 辽宁石油化工大学信息与控制工程学院 抚顺 113001;2. 沈阳航空航天大学自动化学院 沈阳 110136)

摘 要:基于骨架序列数据的双人交互行为识别具有广阔的应用前景,针对目前识别模型中存在双人交互特征表示不充分、动作类内特征表示冗余等问题,提出了一种基于多尺度可变形图卷积网络(multi-scale deformable graph convolutional network, MD-GCN)的双人交互行为识别方法。首先,构建双人交互超图,包括双人超图以及双人交互关系矩阵。与传统图不同,该超图能够更好地表达两人之间的交互关系,充分捕捉双人之间的交互特征。其次,将 3 流输入分支分别进行数据预处理和特征提取,然后将这些特征信息融合后送入以多尺度可变形图卷积网络为主的主分支中,最后进行动作分类。该网络能够多模态地学习可变形的采样位置,捕捉具有显著交互特征的关键信息,有效避免了特征冗余以及信息丢失。所提出的 MD-GCN,在 NTU RGB+D 60 和 NTU RGB+D 120 数据集中的 26 类交互动作的识别任务中,准确率最高达到 98.41%,有效地解决了双人交互行为识别中特征表示的挑战。实验结果表明,该方法在保持识别准确率的同时,显著减小了模型运算成本,模型推理性能达到了良好的平衡,具有较高的应用价值。

关键词: 交互行为识别;骨架序列;图卷积;可变形卷积;多流输入
中图分类号: TP18; TP391.41 **文献标识码**: A **国家标准学科分类代码**: 520.6040

Multi-scale deformable graph convolutional networks for two person interactive action recognition

Wang Li[1] Cao Jiangtao[1] Xie Shuai[1] Ji Xiaofei[2]
(1. School of Information and Control Engineering, Liaoning Petrochemical University, Fushun 113001, China;
2. School of Automation, Shenyang Aerospace University, Shenyang 110136, China)

Abstract: Two-person interaction action recognition based on skeleton sequence data has broad application prospects. To address the issues of insufficient interaction feature representation and redundant intra-class features in current recognition models, we propose a multi-scale deformable graph convolutional network (MD-GCN) for recognizing two-person interaction actions. First, we construct a two-person interaction hypergraph, including a person pair hypergraph and an interaction relationship matrix. Unlike traditional graphs, this hypergraph better captures the interaction between the two people, enabling a more comprehensive representation of the interaction features. Next, three input branches perform data preprocessing and feature extraction, and then the extracted features are fused and fed into the main branch, which is based on the multi-scale deformable graph convolutional network for action classification. This network learns deformable sampling positions in a multi-modal manner, effectively capturing key interaction features while avoiding feature redundancy and information loss. The proposed MD-GCN achieves a recognition accuracy of up to 98.41% on the 26 interaction action classes from the NTU RGB+D 60 and NTU RGB+D 120 datasets. This approach effectively addresses the challenges of feature representation in two-person interaction action recognition. Experimental results show that the method not only maintains high recognition accuracy but also significantly reduces the computational cost, achieving a good balance between inference performance and accuracy, making it highly valuable for practical applications.

Keywords: interaction action recognition; skeleton sequence; graph convolutional; deformable convolution; multi stream input

收稿日期:2024-11-08 Received Date: 2024-11-08
*基金项目:辽宁省人工智能领域科技创新项目(应用基础研究计划项目)(2023JH26/10300013)资助

0 引 言

理解和识别基于骨架序列数据的双人交互行为,对现实生活中的应用是至关重要的[1]。骨架序列数据具有数据量小、复杂背景对其干扰较弱等显著优势。随着先进的人体姿态估计算法[2]的发展,以及 Kinect 等传感器[3]的广泛应用,获取骨架序列变得愈加简单。将骨架序列与深度学习技术结合的方法受到的关注也日益增加。游伟等[4]提出了一种采用边缘计算的多时间尺度骨架特征融合行为识别方法,将骨架特征提取与识别任务部署至多个边缘节点,在各个边缘节点上分别提取不同时间尺度的骨架特征并进行识别。该方法能够根据准确率要求动态调节计算资源,减轻服务器计算压力,然而该模型没有充分考虑人体行为的关节结构以及时空关系。基于卷积神经网络(convolutional neural network, CNN)[5]的方法将骨架特征序列构建为二维伪图像进行处理,可以很好的将动作的空间特性进行建模。基于循环神经网络(recurrent neural network, RNN)[6]的方法将骨架序列作为特征向量序列进行处理,更好的捕获了动作的时间特性。为了充分表达双人交互信息并且有效建模时序关系,赵挺等[7]将 CNN 与加入注意力机制的双向长短时期记忆网络(attention-bidirectional long short-term memory network, A-BLSTM)结合起来,利用 CNN 提取空间交互特征,利用 A-BLSTM 考虑关键帧的时序信息。然而,将骨架序列进行图像化过程中仍然丢失了不少重要信息。大多数基于 CNN 和 RNN 方法的工作都忽略了骨架序列原始的拓扑图结构。基于图卷积神经网络(graph convolution neural network, GCN)[8]的方法将骨架序列表示为时间连续的特定图进行处理,充分考虑特定的节点和其邻居特征,具有结构学习强、鲁棒性高的优势,因此成为目前主流的行为识别研究方法。

在拓扑图构建方面,Yan 等[9]首次提出了时空图卷积网络(spatial temporal GCN, ST-GCN),该网络将关节点视为图的顶点,并以人体结构与时间的自然联系作为图的边,构建了一个时空图,取得了较好的识别效果,吸引了众多研究者的关注。为了获得更丰富的图信息并且利用骨架序列的二阶信息,Shi 等[10]提出了双流自适应图卷积网络(two-stream adaptive GCN, 2 s-AGCN),该网络采用更灵活的图结构,并结合双流输入(包括关节点、骨骼方向和长度),显著提升了模型的识别准确率和通用性。该方法主要解决骨架序列长期时间的依赖性,然而对于每一种模态的输入都单独训练完整的网络最终进行后期融合,大大增加了训练参数量,并且时间边缘仅绑定相邻帧中的相同关节。在此基础上,为了获取局部时空关节的物理依赖性,Li 等[11]提出了有向扩散图卷积网络(directed diffusion graph convolutional network, DD-GCN),构建有向扩散图并引入活动划分策略优化图卷积核的权重机制,反映了动作节点间的相对运动方式。该方法的图构建将双人看作离散的个体进行操作,忽略了双人之间丰富的交互信息并且增加了空间人体结构的冗余建模。为了结合时空交互信息以及骨架几何信息,Zhu 等[12]提出了二元关系图卷积网络(dyadic relational graph convolutional network, DR-GCN),利用一个表示动态关系图的关系邻接矩阵,将骨架序列的几何特征和相对注意力结合起来。该方法考虑了双人的自然交互连接,但同时增加了不相关关节的冗余连接。为了解决这个问题,Li 等[13]提出了二人图卷积网络(two-person GCN, 2p-GCN),利用双人图结构、几何表示策略以及四流输入特征(关节点、骨骼、关节运动和骨骼运动),学习双人动作中的关键时空交互信息。该方法一定程度上解决的连接冗余的问题,但其只表达交互关节点间物理关系,未能有效利用相关性强的特征,而且随着输入特征的不断增多,模型计算成本也逐渐增加。为了有效捕捉远距离关节之间的关系并且降低计算成本,Jiang 等[14]提出了轻量化的多尺度自适应图卷积网络(lighter and faster: a multi-scale adaptive graph convolutional network, LMA-GCN),通过双门控自适应图卷积以及时空分区门控注意力等模块调整层结构的权重分配,增强了模型提取判别特征的整体能力。然而该方法的图构建难以捕捉微小动作之间的连接关系。为了实现多个非自然关节点间的连接信息互通,代金利等[15]提出了基于交互关系超图卷积模型的双人交互行为识别方法,为每一帧分别构建单人超图和交互关系图,强调交互节点间的相关性。该方法充分利用了非自然连接节点间的结构关系,然而未能有效利用四肢的运动特征,模型感受野有限难以灵活适应连续动作类内变化。

综上所述,基于图卷积的双人交互行为识别方法在实际应用中仍面临一些挑战,包括交互特征表示不足、动作类内特征冗余以及细微动作的识别精度不理想等问题。为了解决这些问题,本文提出了一种基于多尺度可变形图卷积网络的双人交互行为识别方法。首先构建了一个双人交互超图,包括双人超图和双人交互关系矩阵,有效地捕捉双人之间潜在的交互关系,并提供更加相关的空间结构与交互特征;其次,设计了一个由空间图卷积(spatial graph network, SGC)、多尺度可变形时间卷积(multi-scale deformable time convolution, MDTC)以及时空部分注意力机制(spatial temporal part attention, STPA)等模块构成的主分支,能够获得动态变化的感受野,有效地提取帧内与帧间的时空信息;然后,通过部分处理和串联操作对输入特征(关节位置、加速度和骨骼特征)进行早期融合,减少模型参数量并提升关键特征的表

达能力;最后,通过全连接层进行动作分类。实验结果表明,所提出的模型在双人交互行为识别任务中,能够显著提升模型的性能,并且解决了现有方法在交互特征表示方面的不足。

1　算法框架

针对现有的基于关节点数据的方法中双人交互特征表示不充分以及模型复杂的问题,提出算法的整体框架如图 1 所示。该网络分为输入分支和主分支。输入分支对输入的 3 流特征进行预处理,并且使用级联操作进行融合,将融合后的特征送入主分支得到时空特征,最后进行动作分类。具体实现步骤如下:

步骤 1) 双人超图构建。将双人看作一个整体并对人体四肢设计物理连接和超连接,构建双人超图。

步骤 2) 双人交互关系矩阵构建。首先对每个人的重要互动关节点(如手和脚),建立个体连接。其次建立两个人之间的交互连接。最后通过测量三维空间中关节之间的距离来评估交互程度。

步骤 3) 结合步骤 1) 和 2) 构成双人交互超图作为网络的图结构,表示双人关节之间的个体、交互相关性。

步骤 4) 将输入数据(关节位置、加速度、骨骼特征)的 3 个分支进行部分处理,不再使用完整的骨架,然后使用串联操作进行早期融合,并构建多尺度可变形时间卷积模块以及时空部分注意力机制。

步骤 5) 动作分类。

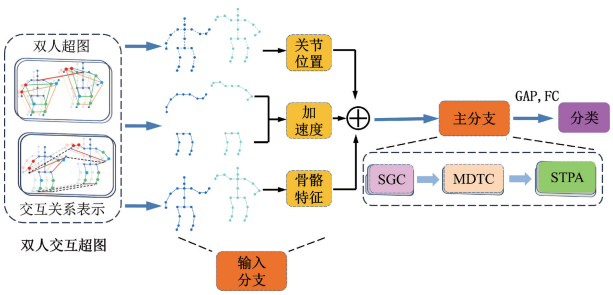


图 1　算法框架

Fig. 1　The overall framework of the network

2　双人交互超图

为了获得更丰富的双人交互信息以及建模更有效的人体空间结构表示,构建双人交互超图,包括双人超图构建和双人交互关系矩阵构建两部分。

2.1　双人超图构建

为了捕获丰富的交互关系,在普通图物理连接的基础上,采用了超图连接。首先把两人看作一个整体,重新对关节点顺序进行排列。NTU RGB+D 数据集[16]的关节点标记如图 2 所示,具有 25 个标记点。其次对所有关节点进行物理连接和超边连接,如图 3(a)所示,红线将双人连接为一个整体,蓝线以及黄线表示双人之间的超连接。超图可以被视为近似的图学习问题,由超节点连接的成对超边构成,如图 3(b)所示,可以用线性数量的超边将任意数量的超节点相连[17]。

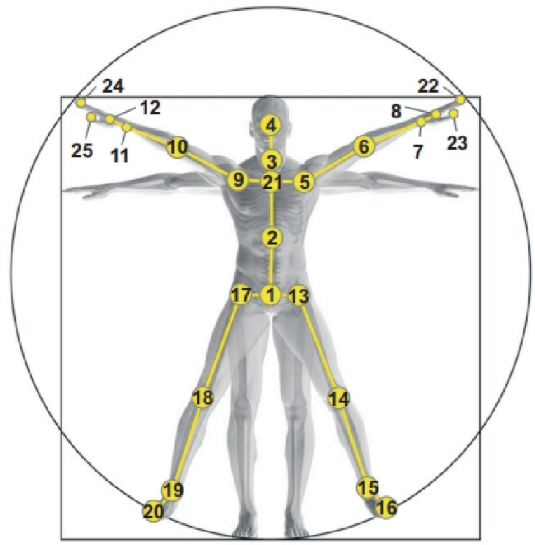
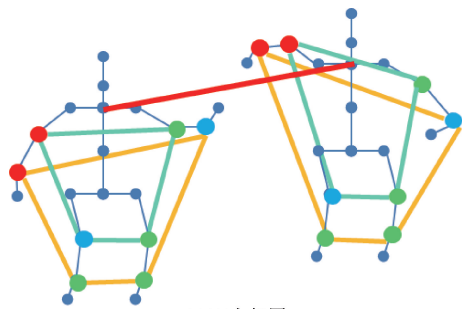


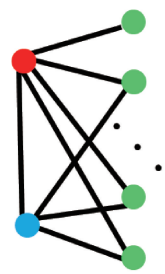
图 2　NTU RGB+D 数据集人体关节点

Fig. 2　NTU RGB+D data set human node



(a) 双人超图

(a) Two person hypergraph



(b) 超边连接

(b) Hyergraph operation

图 3　双人超图结构的创建

Fig. 3　The creation of two person hypergraph structure

双人骨架序列可以表示为以关节为顶点,骨骼为边的图拓扑。因此,将双人超图表示为 $G_H = (V, E_H)$。其中, $V = \{v_i \mid i = 1,2,\cdots,2N\}$ 表示一帧中的所有超顶点集合, $E_H = \{e_{ij} \mid i,j = 1,2,\cdots,2N\}$ 表示顶点连接的所有超边集合。顶点 v_i 与 v_j 之间的连接计算,如式(1)所示,如果存在超边连接为 1,否则为 0[18]。

$$H_{v_i,e_{ij}} = \begin{cases} 1, & v_i \in e_{ij} \\ 0, & v_i \notin e_{ij} \end{cases} \tag{1}$$

使用超图的邻接矩阵 $\boldsymbol{A}_H \in \boldsymbol{R}^{2N\times 2N}$ 构建双人连接关系,相较于普通图,每条边不局限于相邻两点之间的连接关系,能够实现远距离的四肢权重互通。

2.2 双人交互关系矩阵构建

在双人超图构建完成以后,利用交互动作相关性最强的关节点分别建立双人之间手、脚对应关节点的交互连接。为了表示关节点之间交互的强度,利用空间中的关节距离作为双人相互作用的度量[19]。如图 4 所示,紫色实线为个体连接,黑色虚线为交互连接。

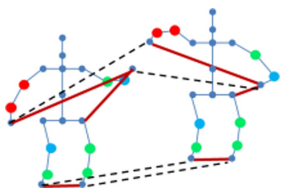


图 4 双人交互关系矩阵的创建

Fig. 4 The creation of interaction relationship matrix

使用 $\hat{A}$ 表示几何关节的相关性,双人交互关节之间的强度表示如式(2)所示。v_i^t、v_j^t 之间的欧氏距离越小, $\hat{A}$ 的值越大,关节之间的交互越强。

$$\hat{A}[t,i,j] = \frac{1}{T}\sum_t \exp\left(-\frac{\| \boldsymbol{F}_{input}(v_i^t) - \boldsymbol{F}_{input}(v_j^t) \|^2}{C}\right) \tag{2}$$

式中: $V = \{v_i^t, v_j^t \mid i,j = 1,2,\cdots,2N\}$ 表示双人之间第 t 帧的第 i 和第 j 个关节; C 为通道数; $\boldsymbol{F}_{input}(v_i^t)$ 和 $\boldsymbol{F}_{input}(v_j^t)$ 分别表示帧 $t \in \{1,\cdots,T\}$ 中第 i 和第 j 个关节的特征向量。

3 双人交互行为识别算法

3.1 数据预处理

大多数方法对输入数据的各个模态使用统一的预处理方式,并且在网络中单独、完整地训练一条实例,不仅增加了模型的总参数,而且得到了多余的特征。为了解决以上问题,针对交互动作生成能够获取更多信息的对称图,针对动作特征明显的身体部位进行部分处理,最后进行多模态数据的早期融合。

1) 多模态输入

为了从原始骨架序列中获得更多输入特征,预处理后将输入数据特征分为 3 类,分别是关节位置、加速度以及骨骼特征,如图 5 所示。将输入通道扩展为每个分支的 $2C$ 通道,则输入特征表示为 $\boldsymbol{f}_{input} \in \boldsymbol{R}^{2C\times T\times 2N}$,其中 C、T、N 分别表示输入通道数、帧数以及关节点数。

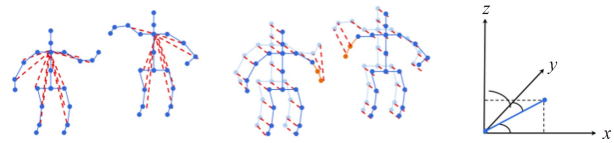


图 5 输入数据表示

Fig. 5 The demonstration of input data

关节位置表示各个关节点相对中心关节点的关节位置,计算如式(3)所示, c 表示中心标识关节点的索引。

$$s_i = \{x[:,:,i] - x[:,:,c], i = [1,2,\cdots,2N]\} \tag{3}$$

加速度表示相邻帧速度的差,计算如式(4)所示。

$$a_i = \{v_i^t - v_i^{t-1}, i = [1,2,\cdots,2N], t = [1,\cdots,T]\} \tag{4}$$

骨骼特征表示骨骼长度和沿轴角度,计算如式(5)所示, i_l 表示相邻关节的骨骼长度, x,y,z 表示三维坐标。

$$\begin{cases} l_i = x[:,:,i] - x[:,:,i_l] \\ a_{i,xyz} = \arccos\left(\dfrac{l_i^t}{\sqrt{(l_{i,x}^t)^2 + (l_{i,y}^t)^2 + (l_{i,z}^t)^2}}\right) \end{cases} \tag{5}$$

2) 对称处理和部分处理

交互动作的个体分别使用主动和被动表示。根据文献[13]的启发,对输入样本采用对称处理。对样本交换彼此的关节标签创建一个对称序列,即为每个样本创建额外的主动和被动交换的对称图来增强输入数据,如图 6 所示。

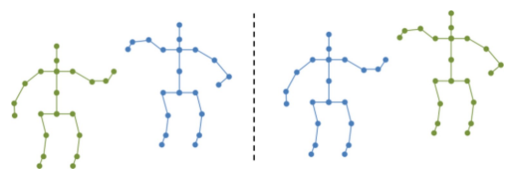


图 6 对称处理

Fig. 6 Symmetric processing

许多交互动作通常不会涉及到人体所有关节点,仅与相关性高的肢体有关。例如,“握手”只需对上肢关节进行识别。因此,提出对 3 类输入特征进行部分处理。将每帧中的完整输入根据动作强度分为不同部分。关节位置类和骨骼特征类的关节点数为 25 个,加速度类只涉及四肢的关节点数为 22 个。

3.2 MD-GCN 整体架构

本文所提出的 MD-GCN 整体架构如图 7 所示。在输入分支中,3 流输入分别通过 BatchNorm 和 Initial Block

层进行数据到特征转换,然后通过 GCN 相关模块进行特征提取,之后采用串联操作融合。融合后的特征送入主分支中,从而确定最终的动作类别。块中的数字表示输入、输出通道,⊕表示级联运算,⊙表示逐元素积运算。

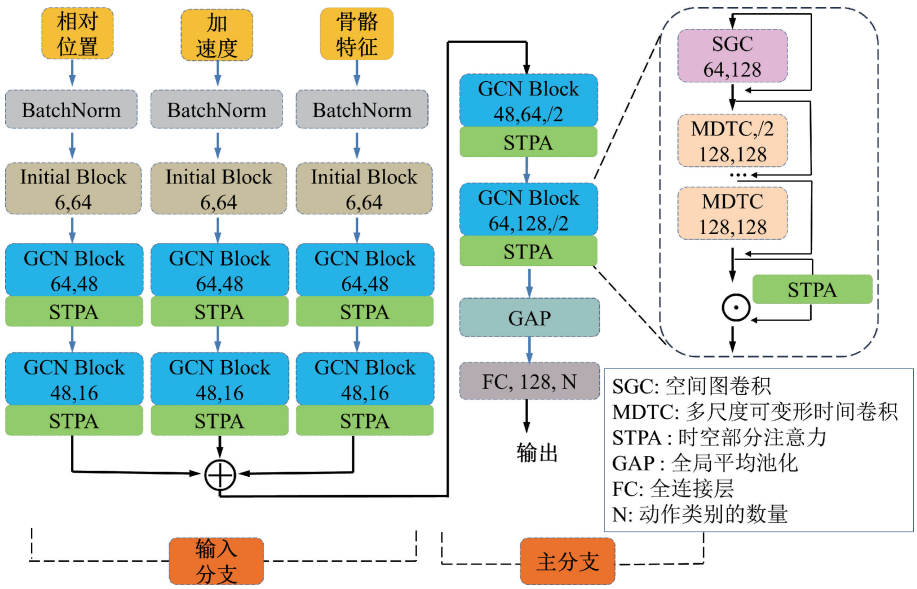


图 7　MD-GCN 网络结构

Fig. 7　MD-GCN network structure

利用双人交互超图构建空间图卷积模块,以获取双人动作之间丰富的交互特征以及空间特征;再引入可变形时间采样和多尺度卷积模块[20]构建多尺度可变形时间卷积模块,以学习时间图上的可变形采样位置,使模型能够感知连续动态的感受野;然后使用时空部分注意力聚焦最相关的关节,将其与整个骨架序列区分开来。

空间和时间层的叠加具有不同的输出通道和时间跨度。输入分支中初始层是 6-64,2 个 GCN 层分别是 64-48、48-16;主分支中 2 个 GCN 层分别是 48-64、64-128。主分支中的步长是 2,其他模块都是 1。

3.3　多尺度可变形时间卷积模块

为了获得时间结构上动态和连续的感受野,将采样位置点设置一个数据驱动的可学习超参数,通过插值提取相应的帧特征。具体包括时间建模、时间采样以及图卷积聚合特征。

1) 时间建模

为了提取具有不同持续帧的动作表示,采用多尺度时间建模。如图 8 所示,每个分支包含一个 1×1 卷积以减少通道数,前 2 个分支包含可变形时间卷积模块(deformable time convolution, DTC),第 3 个分支中包含 1 个 3×1 的 max-pooling 层,⊕表示级联运算,该模块可以自适应地学习判别感受野。

2) 可变形时间采样

骨架序列以特定的帧速率离散采样连续的动作会导致帧间信息丢失,而且随着多个池化层减少帧数,这种情况会变得更糟[21]。为了解决以上问题,对连续的动作序列数据来进行可变形和连续采样。

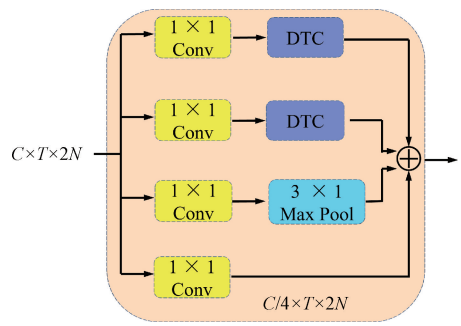


图 8　多尺度可变形时间卷积模块

Fig. 8　Multi-scale deformable time convolution module

首先定义一个超参数,采样的位置数 e。在感受野 R 中对其采样,并设置为可学习参数 $\hat{t}$。参考普通的时间卷积模块[9],得到新的采样策略,如式(6)所示。

$$\hat{t} = t_c + \hat{t}_r \tag{6}$$

3) 时间图卷积

遵循普通时间卷积模块的特征聚集方法,即沿着时间维度进行核大小为 e 的单个一维卷积,如式(7)所示。

$$Z = \sum_{r=1}^{e} w_r \Gamma(Y, \hat{t}) \tag{7}$$

式中:$\Gamma(\cdot,\cdot)$表示双线性插值的采样函数;w_r 表示卷积的第 r 个元素。

MDTC 的核大小是一个超参数,只是一个用来初始

化采样位置的因子。该模块可以使用端到端的方式进行训练,最终获得动态、连续的感受野,能够摆脱根据输入序列长度选择核大小的限制,减轻了池化造成的信息损失。

3.4 时空部分注意力机制

为了使模型更适合捕捉帧间和人体肢体中的信息。考虑时间关系,受到文献[22]的启发,设计时空部分注意力机制,如图 9 所示。

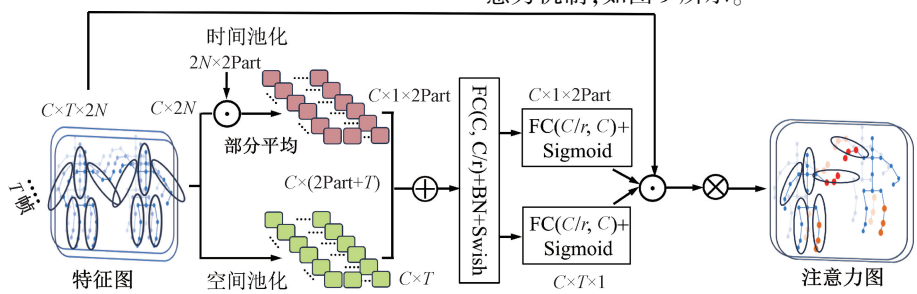


图 9 时空部分注意力模块

Fig. 9 Spatial temporal part attention modules

首先利用池化操作将特征向量串联并通过全连接层,然后采用 2 个独立的全连接层获得帧和部分维度的 2 组注意力分数,最后将 2 组分数相乘计算得到最终的时空注意力图。该注意力模块的定义如式(8)和(9)所示。

$$f_{input}=\theta((pool_t(f_{in})\oplus pool_p(f_{in}))\cdot\omega) \tag{8}$$

$$f_{out}=f_{in}\odot(\sigma(f_{input}\cdot\omega_t)\otimes\sigma(f_{input}\cdot\omega_p)) \tag{9}$$

式中:f_{in} 和 f_{out} 表示输入和特征图;⊕表示级联运算;⊗和⊙表示平均通道外积和逐元素积;$pool_t$ 和 $pool_p$ 分别表示帧级和部分级的平均池化;$\omega\in R^{C\times\frac{C}{r}}$、$\omega_t\in R^{\frac{C}{r}\times C}$ 和 $\omega_p\in R^{\frac{C}{r}\times C}$ 表示可训练的参数;σ 和 θ 分别表示 Sigmoid 和 Swish 激活函数。

4 实验结果与分析

4.1 数据集

NTU RGB+D 60 以及 NTU RGB+D 120 数据集是在室内环境中收集的大规模人体动作识别数据集。分别由 3 台 Kinect v2 摄像机从不同视角采集,每帧包含一个或两个演员。其中,NTU RGB+D 60 的训练集与验证集的比例为 2∶1,即 5 606 个训练集和 2 802 个验证集。为了对本文所提算法进行准确率评估以及多视角应用验证,选择 NTU RGB+D 60 中的 11 类以及 NTU RGB+D 120 中的 26 类双人行为作为测试对象,并且选择 X-view、X-set 作为基准。

4.2 实验细节

所有实验均在 Ubuntu 操作系统上进行,使用基于 PyTorch 的深度学习框架,并利用 NVIDIA 1080 Ti GPU。批处理大小(batch size)为 16,样本大小调整为 64 帧,最大训练次数(max epoch)为 70,初始学习率为 0.1,预热 10 个 epoch 后随着余弦 schedule 衰减。将随机梯度下降法(stochastic gradient descent, SGD)的动量参数设置为 0.9,权重衰减调整为 0.000 1。

4.3 实验结果及分析

1) 输入数据预处理的有效性测试

对多模态输入数据进行预处理的对比如表 1 所示。

表 1 输入数据处理前后对比

Table 1 Comparison of input data before and after processing

模型	浮点数/GFLOPs	Params/($\times10^6$)	准确率/%
原始	8.45	1.10	-
预处理	**3.76**	**1.10**	**+0.09**

由表 1 可知,浮点数明显降低约 5 GFLOPs,识别准确率提升了 0.09%。充分表明,将输入数据进行对称处理和部分处理,能够节约模型推理时间并且得到更多行为特征。多模态输入分支的对比如表 2 所示,其中,***s*** 表示关节位置,***a*** 表示加速度,***b*** 表示骨骼特征。

表 2 输入数据结果对比

Table 2 Comparison of input data results

模型	浮点数/GFLOPs	Params/($\times10^6$)	形式	准确率/%
s	2.19	0.82	1 流	97.44
s + ***b***	2.98	0.96	2 流	97.79
s + ***b*** + ***a*** (本文)	**3.76**	**1.10**	**3 流**	**98.00**

随着输入分支的增多,识别准确率越来越高,浮点数和参数量(Params)并没有显著增加。选用 3 流输入送到输入分支中进行早期融合的准确率最高,表明多流输入特征为动作识别提供了关键的特征信息。

2) 双人交互超图构建的有效性测试

为了评估提出的双人交互超图对于双人交互行为识别的有效性,将双人超图结构和双人交互关系矩阵结构

分别在模型中配置,识别结果如表 3 所示,H 表示双人超图结构,I 表示交互关系矩阵结构。

表 3　消融实验对比

Table 3　Comparison of ablation experiments

模型	H	I	准确率/%
原始图	×	×	97.46
H	√	×	98.00
I	×	√	97.95
HI	√	√	**98.15**

与原始图相比,只使用双人超图,识别准确率提高了 0.54%。只使用交互关系矩阵,准确率提高了 0.49%。同时使用双人超图与交互关系矩阵,准确率提高了 0.69%。结果表明,双人交互超图能够有效地表示个体和交互的连接特征。

3）MD-GCN 模块构建的有效性测试

构成 MD-GCN 的模块分别有双人交互超图、STPA 以及 MDTC 等,分别进行消融实验,并从模型参数量以及识别准确率方面进行分析,实验结果如表 4 所示。

表 4　消融实验对比

Table 4　Comparison of ablation experiments

模型	浮点数/GFLOPs	Params/($\times10^6$)	准确率/%
原始	8.45	1.10	97.46
+STPA	8.44	1.10	97.57
+HI	3.76	1.10	98.15
+MDTC	4.14	1.28	98.41
HI+STPA+MDTC(本文)	**4.14**	**1.28**	**98.41**

在数据预处理之后配置 STPA、双人交互超图以及 MDTC,识别准确率分别提高了 0.11%、0.69% 以及 0.26%,然而模型参数量和浮点数并没有显著增加。实验表明,模型在时间尺度上学习空间特征是有效的。

模型在 NTU RGB+D 60 数据集的 11 类交互动作上的混淆矩阵如图 10 所示。

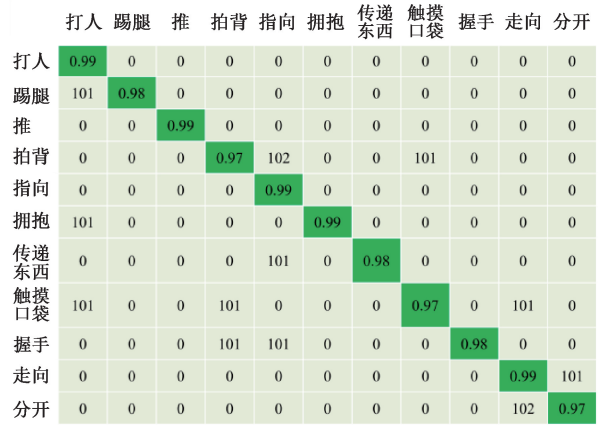


图 10　混淆矩阵

Fig. 10　Confusion matrix

所提出的 MD-GCN 在“触摸口袋”和“分开”这两个动作的识别准确率最低为 97%,“打人”、“推人”、“指向”、“拥抱”以及“走向”这几类涉及四肢部位活动较多的动作都达到了 99%的准确率,充分表明了本文所提方法能够探索与交互行为最相关的人体关节,其获取时空特征的优越性。

4.4　与现有算法对比

将其他主流网络中的图结构替换为提出的双人交互超图,识别结果如表 5 所示,A 表示传统图,H 和 I 分别表示双人超图以及交互关系矩阵。

表 5　与其他模型的图结构对比

Table 5　Comparison of ablation results

模型	浮点数/GFLOPs	Params/($\times10^6$)	准确率/%
ST-GCN[9](A→HI)	16.32	3.10	**+0.43**
2S-AGCN[10](A→HI)	37.32	6.94	**+0.94**
Efficient-GCN[22](A→HI)	8.45	1.10	**+0.69**
De-GCN[21](A→HI)	—	5.56	**+0.52**
ST-GAT[23](A→HI)	14.60	2.40	**+0.17**

由表 5 可知,替换图结构之后的识别准确率都明显提高,进一步证明了双人交互超图的有效性以及泛化性。

TC 表示传统时间卷积,将其替换为 MDTC,实验结果如表 6 所示。

表 6　各模块的实验结果对比

Table 6　Comparison of experimental results about the modules

模型	浮点数/GFLOPs	Params/($\times10^6$)	准确率/%
ST-GCN[9](TC→MDTC)	16.32	3.10	**+1.91**
CTR-GCN[20](TC→MDTC)	4.68	2.01	**+1.01**
Efficient-GCN[22](TC→MDTC))	8.45	1.10	**+0.26**

由表 6 可知,替换时间卷积模块的识别准确率都有不同程度的提高,取得了良好的表示效果。

与其他方法的进一步对比分析如表 7 所示,证实了本文所提出的 MD-GCN 模型的有效性。

表 7　与其他模型的实验结果对比

Table 7　Comparison of ablation results

模型	浮点数/GFLOPs	Params/($\times10^6$)	NTU 60 准确率/%	NTU 120 准确率/%
ST-GCN[9](1 流)	16.32	3.10	93.70	80.07
2S-AGCN[10](2 流)	37.32	6.94	96.36	89.11
Efficient-GCN[22](3 流)	8.45	1.10	97.46	89.90
2P-GCN[13](4 流)	**3.76**	**1.67**	**98.73**	**92.31**
CTR-GCN[20](4 流)	7.16	5.68	97.60	91.80
ST-GAT[23](4 流)	14.60	2.40	98.00	91.53
De-GCN[21](4 流)	—	5.56	97.41	91.89
本文(3 流)	**4.14**	**1.28**	**98.41**	**91.85**

和其他模型相比,本文所提出的 MD-GCN 的参数量明显减少并且浮点数最低,在保持较高准确率的同时,很大程度上减小了模型的计算成本。尽管 MD-GCN 在识别准确率上略低于 2P-GCN,但值得注意的是,2P-GCN 采用了 4 流输入,其模型参数量比 MD-GCN 多 0.39×10^6,模型推理时间则更长。在模型效率方面,MD-GCN 展现了更为优越的性能。

5 结 论

本文提出了一种基于多尺度可变形图卷积网络的双人交互行为识别方法。该模型充分利用多流输入样本的独特性以及双人交互超图,能够有效提取双人动作之间的空间结构特征和交互特征。此外,MD-GCN 结合了空间图卷积、多尺度可变形时间卷积和注意力机制,充分挖掘帧间信息和与人体动作最相关的时空特征,从而在保证模型准确率的同时,减少了模型的参数量。在 NTU RGB+D 60 和 NTU RGB+D 120 交互数据集上的实验结果验证了该方法的优异性能。未来的研究将进一步聚焦于人物遮挡等特殊场景下的双人交互行为识别问题。

参考文献

[1] 边存灵, 吕伟刚, 冯伟. 骨架人体行为识别研究回顾、现状及展望[J]. 计算机工程与应用, 2024, 60(20): 1-29.
BIAN C L, LYU W G, FENG W. Skeleton-based human action recognition: History, status and prospects[J]. Computer Engineering and Applications, 2024, 60(20): 1-29.

[2] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7291-7299.

[3] 朱大勇, 郭星, 吴建国. 基于 kinect 三维骨骼节点的动作识别方法[J]. 计算机工程与应用, 2018, 54(20): 152-158.
ZHU D Y, GUO X, WU J G. Action recognition method using kinect 3D skeleton data[J]. Computer Engineering and Applications, 2018, 54(20): 152-158.

[4] 游伟, 王雪. 人行为骨架特征识别边缘计算方法研究[J]. 仪器仪表学报, 2020, 41(10): 156-164.
YOU W, WANG X. Study on the edge computing method for skeleton-based human action feature recognition[J]. Chinese Journal of Scientific Instrument, 2020, 41(10): 156-164.

[5] DING W W, DING C Y, LI G, et al. Skeleton-based square grid for human action recognition with 3D convolutional neural network[J]. IEEE Access, 2021, 9: 54078-54089.

[6] GAO Y, LI C, LI S, et al. Variable rate independently recurrent neural network (IndRNN) for action recognition[J]. Applied Sciences, 2022, 12(7): 3281.

[7] 赵挺, 曹江涛, 姬晓飞. CNN A-BLSTM network 的双人交互行为识别[J]. 电子测量与仪器学报, 2021, 35(11): 100-107.
ZHAO T, CAO J T, JI X F. 2CNN A-BLSTM network for two-person interaction behavior recognition [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(11): 100-107.

[8] 张德, 王怡婷, 甄昊宇. 结合解耦注意力图卷积与时态建模的骨架动作识别[J]. 国外电子测量技术, 2023, 42(9): 91-98.
ZHANG D, WANG Y T, ZHEN H Y. Combining decoupling attention graph convolution and temporal modeling for skeleton-based action recognition [J]. Foreign Electronic Measurement Technology, 2023, 42(9): 91-98.

[9] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 7444-7452.

[10] SHI L, ZHANG Y F, 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.

[11] LI C, HUANG Q, MAO Y. DD-GCN: Directed diffusion graph convolutional network for skeleton-based human action recognition [C]. 2023 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2023: 786-791.

[12] 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.

[13] LI Z, LI Y, TANG L, et al. Two-person graph convolutional network for skeleton-based human interaction recognition [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(7): 3333-3342.

[14] JIANG Y, DENG H. Lighter and faster: A multi-scale adaptive graph convolutional network for skeleton-based action recognition [J]. Engineering Applications of Artificial Intelligence, 2024, 132: 107957.

[15] 代金利, 曹江涛, 姬晓飞. 交互关系超图卷积模型的双人交互行为识别[J]. 智能系统学报, 2024, 19(2):

316-324.
DAI J L, CAO J T, JI X F. Two-person interaction recognition based on the interactive relationship hypergraph convolution network model [J]. CAAI Transactions on Intelligent Systems, 2024, 19(2): 316-324.
[16] 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.
[17] YADATI N, NIMISHAKAVI M, YADAV P, et al. HyperGCN: A new method of training graph convolutional networks on hypergraphs [C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019: 1511-1522.
[18] FENG Y, YOU H, ZHANG Z, et al. Hypergraph neural networks [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 3558-3565.
[19] 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.
[20] 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.
[21] MYUNG W, SU N, XUE J H, et al. DeGCN: Deformable graph convolutional networks for skeleton-based action recognition [J]. IEEE Transactions on Image Processing, 2024, 33: 2477-2490.
[22] SONG Y F, ZHANG Z, SHAN C, et al. Constructing stronger and faster baselines for skeleton-based action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 1474-1488.
[23] HU L Y, LIU S, FENG W. Spatial temporal graph attention network for skeleton-based action recognition [J]. ArXiv preprint arXiv: 2208.08599, 2022.

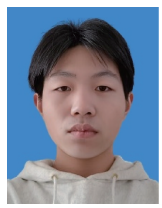
作者简介

王丽，2022 年于宝鸡文理学院获得学士学位，现为辽宁石油化工大学硕士研究生，主要研究方向为计算机视觉。
E-mail：1848426781@qq.com

Wang Li received her B. Sc. degree from Baoji University of Arts and Sciences in 2022. Now she is a M. Sc. candidate at Liaoning Petrochemical University. Her main research interest includes computer vision.

曹江涛，2009 年于英国普茨茅斯大学获得博士学位，现为辽宁石油化工大学教授、硕士生导师，主要研究领域为智能方法及其应用、视频分析与处理等。
E-mail：cigroup@126.com

Cao Jiangtao received his Ph. D. degree from the University of Portsmouth in 2009. Now he is a professor and M. Sc. supervisor at Liaoning Petrochemical University. His main research interests include intelligent methods and their applications, video analysis and processing, etc.



谢帅，2023 年于湖北师范大学获得学士学位，现为辽宁石油化工大学硕士研究生，主要研究方向为计算机视觉。
E-mail：3111699529@qq.com

Xie Shuai received his B. Sc. degree from Hubei Normal University in 2023. Now he is a M. Sc. candidate at Liaoning Petrochemical University. His main research interest includes computer vision.

姬晓飞（通信作者），现为沈阳航空航天大学副教授，主要研究方向为视频分析与处理、模式识别理论等。
E-mail：jixiaofei7804@126.com

Ji Xiaofei (Corresponding author) now she is an associate professor and M. Sc. supervisor at Shenyang Aerospace University. Her main research interests include video analysis and pattern recognition theory, etc.