2025 年 第 25 卷 第 8 期
2025, 25(8): 03316-09

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2309528
引用格式:姬晓飞, 张薇, 冯雅迪. 改进时空图卷积模型的双人交互行为识别算法[J]. 科学技术与工程, 2025, 25(8): 3316-3324.
Ji Xiaofei, Zhang Wei, Feng Yadi. Improved spatial temporal graph convolutional model for two-person interaction recognition algorithm[J]. Science Technology and Engineering, 2025, 25(8): 3316-3324.

改进时空图卷积模型的双人交互行为识别算法

姬晓飞, 张薇, 冯雅迪
(沈阳航空航天大学自动化学院, 沈阳 110136)

摘 要 针对双人交互行为识别网络中存在忽略人体间的非自然连接关系和交互关系等突出问题,提出一种改进时空图卷积模型的双人交互行为识别算法。首先通过边卷积操作汇聚节点的边特征,以捕获人体的非自然连接关系;其次利用改进的关系网络,构建了双人之间的交互关系图;然后将边卷积操作分支及交互关系图嵌入时空图卷积网络块,分别构建为边-图卷积块和交互关系块;最后将两者高效融合,提出一个能同时捕捉非自然连接关系和交互关系的改进时空图卷积算法,从而实现双人交互行为识别。为验证网络的有效性,在国际公开大型标准数据集 NTU RGB+D 上进行测试。实验结果显示,该算法识别准确率达 97.77%,相比于基线时空图卷积模型提升了 4.28 个百分点,提高了双人交互行为特征的表现力,取得了比现有先进网络模型更好的识别效果。

关键词 双人交互行为识别; 关节点数据; 边卷积; 关系网络; 时空图卷积网络
中图法分类号 TP391.41; **文献标志码** A

Improved Spatial Temporal Graph Convolutional Model for Two-person Interaction Recognition Algorithm

JI Xiao-fei, ZHANG Wei, FENG Ya-di
(College of Automation, Shenyang Aerospace University, Shenyang 110136, China)

[**Abstract**] Aiming at the prominent problems of ignoring the unnatural connection relationship and interaction relationship between human bodies in two-person interaction recognition algorithm, a two-person interaction recognition network based on improved spatial temporal graph convolutional model was proposed. Firstly, the edge features of joint point data were aggregated by edge convolution to capture the unnatural connectivity relations inherent in the human body. Secondly, the interaction relationship graph between two people was constructed by using the improved relationship network. Furthermore, the branch of edge convolution and the interaction relationship graph were embedded into the spatial temporal graph convolutional network block, which were constructed as an edge-graph convolutional block and interaction relation graph convolutional block. Finally, an improved spatial temporal graph convolution algorithm was proposed to capture both the unnatural connection relationship and the interaction relationship, so as to realized the recognition of two-person interaction behavior. To verify the effectiveness of the network, it was tested on the international public large-scale standard dataset NTU RGB+D. The experimental results show that the network obtain a recognition accuracy of 97.77%, which is an improvement of 4.28 percentage points compared to the baseline spatial temporal graph convolutional network. It improves the expressiveness of two-person interaction behavioral features, and achieves a better recognition effect than the existing state-of-the-art network models.

[**Keywords**] two-person interaction recognition; joint point data; edge convolution; relational network; spatial temporal graph convolutional network

随着计算机视觉技术的不断发展,基于视频的人体交互行为识别广泛应用于在智能安防、人机交互、姿态评估等领域[1]。相比单人动作,双人交互行为在日常生活中更普遍,同时又是组成多人交互行为的基本单元,因此基于视频的双人交互行为识别研究具有重要的现实意义。普遍使用的 RGB 视频数据源能表征丰富的人体外观信息,但易受光照、遮挡及视频分辨率等多种因素的影响,因此基于 RGB 视频的复杂行为识别准确率较低[2]。而由 Kinect 智能体感相机获取的关节点数据具有简单、明确、语义性高且不易受外观环境影响的优点,因此,基于关节点数据的双人交互行为识别研究吸引

收稿日期: 2023-12-04; **修订日期**: 2024-12-15
基金项目: 辽宁省教育厅重点攻关项目(LJKZZ20220033)
第一作者: 姬晓飞(1978—),女,汉族,辽宁鞍山人,博士,副教授。研究方向:模式识别理论、视频分析。E-mail:jixiaofei7804@126.com。

了众多研究组的重视,且取得了一些突破性进展。

针对双人交互行为识别的研究,基于关节点数据,可以根据所采用的深度学习算法进行分类。主要的类别包括:基于卷积神经网络(convolutional neural networks, CNN)、循环神经网络(recurrent neural network, RNN)、图卷积神经网络(graph convolutional network, GCN)以及融合网络的方法。Ding 等[3]通过计算 5 种空间骨架特征并从中选择关键特征后,将其映射到彩色空间且按时间序列编码为彩色纹理图像,最后送入 CNN 中联合捕获时空信息进行动作识别。姬晓飞等[4]对 RGB 视频和关节点数据异构数据源提取浅层特征,然后分别将其送入 CNN 网络中进行深层特征提取,最后进行决策级融合两分支得到识别结果。武东辉等[5]采用 CNN 和长短期记忆网络(long short-term memory, LSTM)结合的方式获取了丰富的人体行为时空特征,并加入注意力机制对重要特征进行优化进一步提升识别精度。Wang 等[6]首先将关节点坐标逐帧表示形成图像矩阵作为 CNN 网络输入,然后由双向门控循环单元(bidirectional gated recurrent unit, Bi-GRU)构建注意力机制来学习帧的时间权重,得到深层时空特征完成动作识别。基于 CNN 和 RNN 的方法往往将关节点数据进行图像化编码后作为网络输入,不仅改变了关节点数据原始的拓扑结构,还弱化了不同关节点间的相关性及交互个体间的重要交互信息。

GCN 特别适用于处理非欧式空间数据,能有效应对如人体关节点这类分布不均匀的拓扑结构数据。Yan 等[7]将 GCN 与时序卷积神经网络(temporal convolutional networks, TCN)结合为时空图卷积网络(spatial temporal GCN, ST-GCN),联合捕获人体关节点的时空变化关系。Li 等[8]设计了推理动作潜在连接的动作模块和具有高阶骨架特征的结构模块,两者组合为广义骨架图,送入中 ST-GCN 提取时空特征。刘锁兰等[9]突破了 ST-GCN 中传统骨架建模规则,根据节点与根节点之间的距离为 0、1、2 划分为 3 个子集,建立了一种新型分区策略,加强了身体相对位置之间的关系。Song 等[10]将可分离卷积的 3 个输入分支进行早期融合,并应用复合缩放策略来同步扩展模型的宽度和深度,在增加特征多样性的同时,保证了特征提取的效率。目前,大部分改进后 ST-GCN 模型的整体性能均得到了进一步提升,但对于双人交互行为的特性缺乏全方位考虑,其仍然存在以下两个突出问题:一是由于其仅考虑人体的物理连接方式,忽略了物理距离较远但语义关系性强的节点之间的连接;二是此网络把独立个体数据送入网络进行特征提取,对双人间重要的交互关系缺乏考虑。为了捕获人体的非自然连接关系,Shi 等[11]提出了一种双流自适应图卷积神经网络(two-stream adaptive graph convolutional networks, 2S-AGCN),它以数据驱动的方法自适应学习关节点信息,突破了固定的人体物理连接限制,并结合关节点流和骨骼流形成双流结构进一步提升模型性能。张静亭等[12]构建距离人体重心位置较远的关节间的连接为非自然连接,根据关节点位置信息对双人之间的连接边赋予不同的权重。Li 等[13]为了同时捕获人体自然和非自然连接关系,设计了知识嵌入图卷积(knowledge embeded graph convolutional networks, K-GCN)网络,通过计算不同关节点间的相关性实现人体的非自然连接。曹毅等[14]引入图注意力机制来聚集邻域节点特征,并采用非局部网络将全局节点特征聚集到当前节点构建自适应邻接矩阵。Plizzari 等[15]提出了一种新的时空变压器网络,该网络构建了基于 Transformer 网络的空间和时间自注意模块,以动态建模帧内和帧间关节间的连接关系,使得网络关注于时空活跃关节点,降低特征冗杂性。Chen 等[16]通过学习一个共享拓扑作为所有通道的通用先验,然后利用每个通道的特定相关性对其进行细化来建立通道拓扑模型,有效聚合不同通道中的联合特征。此类自适应图均从点的嵌入中生成点特征,虽有效提取了点的局部特征,但由于每个点邻域不同,不满足点的置换不变性,网络灵活性和泛化性较差。

目前,大多数基于 ST-GCN 的双人交互行为识别研究大体采用将双人交互整体割裂为两个单独个体馈入单人 GCN 模型中,通过后期的特征融合进而完成双人交互动作识别。这样的处理方式未充分考虑双人交互动作间交互信息的有效利用,导致交互动作识别的准确率不高。为了有效建模交互个体间的交互信息,Li 等[17]将双人关节点数据视为一个整体作为网络输入,构建了保留基本交互关系的双人图,并提出 4 种双人个体间的手工标记策略来进一步构建双人间的交互关系,但不同动作间的连接关系有较大差异,算法缺乏适应性与灵活性。Zhu 等[18]结合交互中双人骨架的几何特征和相对注意力,构建表示两个骨架之间的关系链的动态关系图,进而嵌入时空图卷积块提取时空交互特征。Li 等[19]首先定义了双人交互区域,将两个交互个体关节点间的最大相对欧式距离和双人间的关键连接定义为双人交互连接,构建为知识给定图后送入 ST-GCN 进行识别。成科扬等[20]根据交互动作中节点的运动特性建立了对应连接和潜在连接,对两个

交互个体中运动轨迹相似的关节点构建对应连接,其它关节点间通过欧式距离衡量交互节点间的相关性构建潜在连接。此类方法提取的关系特征均为低级特征,特征提取不充分,表征能力不足。

综上所述,目前将 ST-GCN 应用于复杂的双人交互行为识别领域还面临着非自然连接和交互信息表示不足的挑战。因此,提出一个改进的时空图卷积算法,该算法利用具有非自然连接能力的边卷积操作和能表征交互关系的关系网络(relational network, RN)捕获双人非自然连接关系及交互连接关系,充分考虑了非自然连接在行为识别中能够增大网络感受野和提升模型灵活性的突出作用,以及交互关系对于表征双人间交互信息的必要性。所提算法沿用 ST-GCN 处理框架,在原自然连接图的基础上融入边卷积操作分支和交互关系图,在提高复杂交互行为识别精度的同时,以保证模型的泛化性和稳健性。

1 算法整体框架

所提出的算法模型如图 1 所示,将关节点数据作为网络输入,经由实例归一化(instance normalization, IN)和批归一化(batch normalization, BN)层进行数据预处理,其中深层时空特征提取模块采用 ST-GCN 中 9 个时空图卷积块残差连接的架构。算法具体实现步骤如下。

步骤 1 构建边-图卷积块(edge-graph convolutional block, E-GCB)。对人体关节点数据进行边卷积操作,将节点本身及其邻域边的特征进行聚合。所获得的边特征与由时空图卷积块生成的图特征通过拼接方式融合构建 E-GCB,从而赋予网络识别人体关节点间的自然及非自然连接关系的双重能力,增强了特征的表现力。

步骤 2 构建交互关系块(interaction relation graph convolutional block, IR-GCB)。利用改进的 RN 网络捕获双人交互个体每对关节点间的交互关系,得到的关系向量经归一化和重构操作生成交互关系图,交互关系图与自然连接图以逐元素相加的形式融合构建 IR-GCB,进一步完善了双人深层次交互特征,实现了静态矩阵到动态矩阵的转变,增大了网络的感受野,提升了网络的特征提取能力。

步骤 3 将 E-GCB 与 IR-GCB 以逐块残差连接的方式进行堆叠,使得模型同时具有捕获非自然连接关系和交互关系的能力,且在保证双人交互行为识别速度的同时,提升了识别准确率。

步骤 4 动作识别。将得到的深层特征送入全局平均池化(global average pooling, GAP)和全连接

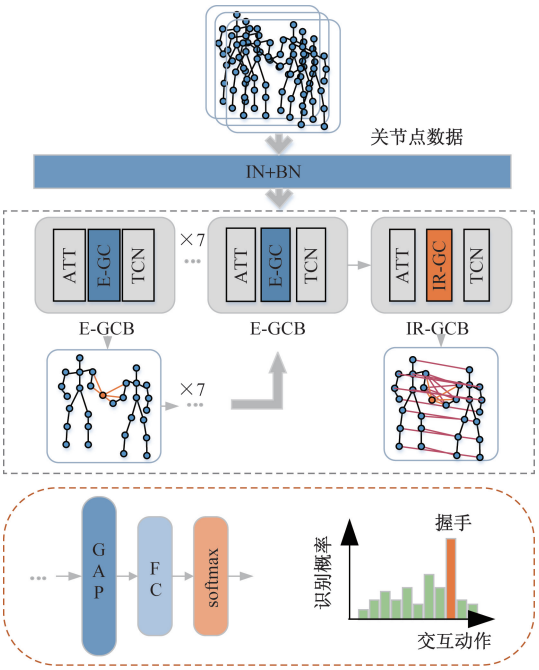


图 1 算法整体结构图

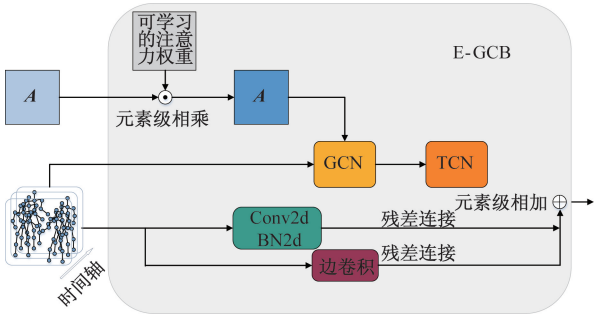
Fig. 1 Overall structure of the algorithm

(fully connected, FC)层得到特征向量,进而由 softmax生成分类概率实现双人交互行为识别。

2 边-图卷积块的构建

为了增加人体关节点间的非自然连接,考虑到三维点云中广泛使用的边卷积[21]可以直接生成描述点与其邻居节点间关系的边缘特征,在保持关节点置换不变性的同时捕获局部几何结构,将边卷积操作以残差连接的形式嵌入到 ST-GCN 中,得到的边特征与原始图特征以拼接的方式融合,构成边-图卷积块如图 2 所示。

边卷积采用 k 近邻算法(k-nearest neighbor, k-NN)固定邻居的数目,具有置换不变性;另外,由于边卷积不仅显式地构造局部图,而且深入分析边缘的嵌入,因此该模型能够在欧几里得空间和语义空间中聚合点与邻节点间的边缘特征,捕获了自然



A 为邻接矩阵

图 2 边-图卷积块结构框图

Fig. 2 Block diagram of edge-graph convolution block structure

和非自然连接关系,增强了局部特征的丰富性,使得模型对细微的行为差异更加敏感。其中,利用边卷积操作进行特征提取的过程如图 3 所示,具体步骤如下。

步骤 1　首先将输入数据按帧进行平均池化操作,得到代表帧。

步骤 2　基于欧氏距离的 k-NN 构造具有局部边的局部图,其中 k 为所取的相邻点个数,是可变参数。

步骤 3　通过可训练参数 $\boldsymbol{W}_{\text{edge}} \in \mathbf{R}^{C\times(2\times C)}$ 聚合局部边和物理边。

边卷积是对节点的所有相邻边特征执行通道的对称聚合,同时每个节点也指向自己。因此,在第 i 个顶点的输出由式(1)所示。

$$x_i' = \underset{j:(i,j)\in\varepsilon}{\Gamma} h_\theta(x_i,x_j) \tag{1}$$

式(1)中:i 为第 i 个顶点;j 为与顶点 i 存在相连边的邻接顶点;$(i,j)\in\varepsilon$ 为在图 $G=(V,E)$ 中,从顶点 i 出发的所有边 (i,j),其中 ε 为边的集合;同图像的卷积类似,选定 x_i 为中心像素,$\{x_j:(i,j)\in\varepsilon\}$ 作为其周围的一个贴片;Γ 为聚合操作,一般取 sum 或 max;$h_\theta(\cdot)$ 为边特征,每个点的边特征由非对称边缘函数 $\bar{h}_\theta$ 计算,$\bar{h}_\theta$ 为一个具有一组可学习参数 θ 的非线性函数,可表示为

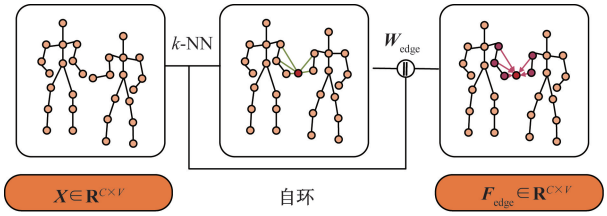
$$h_\theta(x_i,x_j) = \bar{h}_\theta(x_i,x_j - x_i) \tag{2}$$

式(2)中,x_i 维护了全局结构信息,$x_j - x_i$ 维护了局部邻域信息,二者结合提取的特征更为完整和有效。

边卷积操作分支嵌入时空图卷积块中,以通道拼接的方式将输出的边特征与时空图特征进行融合,进而得到边-图卷积块,其输出公式为

$$F_{\text{Eout}} = \left\{ \sum_j \left[\boldsymbol{X}(\boldsymbol{A}_j \otimes \boldsymbol{M})\boldsymbol{W}_j \right] \right\} \left\| \left[e\left(\sum_{t=1}^{T} \boldsymbol{X}_t \right) \right] \right\| \tag{3}$$

式(3)中:$\otimes$为拼接操作;$\boldsymbol{X}$ 为输入特征;等号右侧第一项{ }内为使用空间结构分区策略的 ST-GCN 的输出特征,其中,j 为分区数目;$\boldsymbol{M}$ 为可学习的权重矩阵,其初始值为全 1 矩阵,与邻接矩阵 $\boldsymbol{A}_j$ 进行逐点相乘;$\boldsymbol{W}_j$ 为权重矩阵;等号右侧第二项[ ]内为边卷积操作的输出特征,其中,e 为边卷积操作;$\boldsymbol{X}_t$ 为输入特征 $\boldsymbol{X}$ 在第 t 帧的特征分量。



C 为通道数;V 为关节点数

图 3　边卷积实现过程

Fig. 3　Edge convolutional implementation process

3　交互关系块的构建

与单人行为识别研究相比,双人交互行为识别研究蕴含着丰富的交互关系,为了更好地利用其深层交互关系,首先将网络输入的独立单人图改进为整体双人图,将双人作为一个整体而不是独立为交互个体,更加适用于双人交互行为识别领域研究。其次,对可训练 RN 网络进行改进作为关系推理模块,用以提取更深层次的交互关系,并以此构建交互关系图。Santoro 等[22]提出 RN 网络作为一种简单的架构来处理需要进行关系推理的问题,在理解包括运动数据的动态物理系统具有先进的性能,在推理对象对之间的关系尤其有效,因此,其适用于更深层次的动态双人交互关系的推理,它可以由动作激活,且可以存在于任意一对关节间。以逐元素相加的形式融合交互关系图与自然连接图,进而生成 IR-GCB,以表征交互个体间重要的交互关系。在最简单的形式中,RN 是一个复合函数,可表示为

$$\text{RN}(O) = f_\phi\left[\sum_{i,j} g_\theta(o_i,o_j) \right] \tag{4}$$

式(4)中:RN 推理模型包括关系模型 f_ϕ 和特征提取模型 g_θ,其中,下标 ϕ 和 θ 为模型参数;o_i 和 o_j 分别为两个交互个体第 i 和 j 个关节点。

使用其特征提取模型并对其进行改进,得到新的关系网络,以表示两个交互个体每对关节点之间的关系。IR-GCB 中交互网络的输入是状态描述矩阵,将每个关节 i 定义为一个对象,使用其沿帧的坐标作为它们的低级特征:$j_i = (x_1,y_1,x_2,y_2,\cdots,x_T,y_T,i,b)$,其中,$x_t$ 和 y_t 为第 t 帧属于身体部位 b 的关节 i 的二维坐标,T 为所需要使用的帧的采样。考虑躯干、左手、右手、左腿、右腿共 5 个身体部位。因此,每个人 p 对于每个视频都有一组关节,可以定义为:$P_p = \{j_p^1,j_p^2,\cdots,j_p^N\}$,其中 N 为姿势数据提供的关节总数。RN 将此状态描述矩阵中的每一行视为一个对象,因此一个对象描述包含了其属性随时间演变的信息。改进 RN 网络的特征提取过程如图 4所示。

本文设计的 $p_1,p_2 = \{j_1,j_2,\cdots,j_N\}$,其中 p_1 和 p_2 分别为双人交互个体低级特征集合,j_N 为第 N 个关节点沿帧坐标表示的低级特征,g_θ 是一个 4 层多层感知机(multi-layer perceptron, MLP),包括 3 层 1 000 个单元和 1 层 625 个单元,用来计算来自不同交互个体的每对关节点之间的关系,将输出的每对

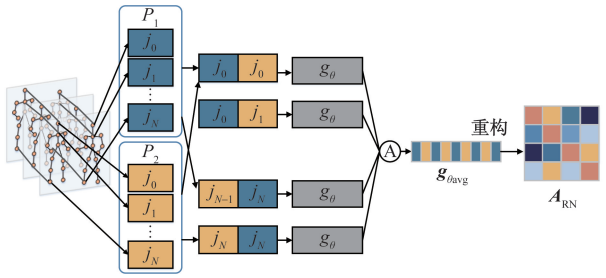


图 4 改进关系网络的特征提取过程

Fig. 4 Improved feature extraction implementation process for relational networks

关系向量相加,然后取其均值,得到两个交互个体之间的整体关系表示 $\boldsymbol{g}_{\theta\text{avg}}$,即关系向量。为了与 ST-GCN 网络进行融合,需要将关系向量进行归一化操作,并进一步重构为关系矩阵 $\boldsymbol{A}_{\text{RN}}$,归一化操作如式(5)所示。

$$\boldsymbol{A}_{\theta\text{avg}} = \text{softmax}\left(\frac{\boldsymbol{g}_{\theta\text{avg}} + \boldsymbol{r}}{\tau}\right) \in \mathbf{R}^{C} \tag{5}$$

式(5)中:$\boldsymbol{r}$ 为一个随机向量;从 Gumbel(0, 1)分布中采样,τ 控制了 $\boldsymbol{A}_{\theta\text{avg}}$ 的离散化,设置 τ =0.5。

利用 Gumbel Softmax[23]得到近似范畴形式的连接概率 $\boldsymbol{A}_{\theta\text{avg}}$,即维度为 625 的关系向量,接着将此向量重构为 50×50 大小的双人交互关系矩阵 $\boldsymbol{A}_{\text{RN}}$,数值依次重新排列在矩阵四分区的左下和右上交互分区部分,其余区域置 0。关系矩阵与邻接矩阵逐元素相加融合为新的动态邻接矩阵 $\boldsymbol{A}_{\text{IR}} = \boldsymbol{A} + \boldsymbol{A}_{\text{RN}}$,进而构建为能够表征双人间的深层次交互关系的 IR-GCB,其输出特征表示为

$$\boldsymbol{F}_{\text{IRout}} = \sum_{j} \{\boldsymbol{X}[(\boldsymbol{A}_j + \boldsymbol{A}_{\text{RN}}) \otimes \boldsymbol{M}]\boldsymbol{W}_j\} \tag{6}$$

4 融合网络的构建

ST-GCN 的输入为大小为(N,C,T,V,M)的 5 维矩阵,其中,N 为视频数,C 为通道数,T 为时间帧数,V 为关节点数,M 为一帧中的人数。首先,对网络的输入数据进行 BN 和 IN 数据预处理操作,统一一个关节在不同帧中的位置特征,然后送入 ST-GCN 中进行时空特征提取。其中,引入一个可学习的权重掩码作为注意力调整邻接矩阵中对应边的连接强度,接着交替使用 GCN 和 TCN,完成对时空维度的变换,并设计残差结构计算获得残差值,将其与 TCN 的输出按位相加得到时空图卷积块的输出,9 个时空图卷积块进行堆叠构造深层神经网络,得到高级时空语义特征,最后送入 GAP 和 FC 层经由 softmax 激活函数输出最终预测结果。

将时空特征提取块改进为具有捕获非自然连接能力的 E-GCB 和能够提取交互关系的 IR-GCB,二者同时融合到 ST-GCN 网络中,能够突破 ST-GCN 网络仅考虑人体固定物理连接的局限性,同时挖掘双人间交互连接的价值,得到更高级和有效的判别特征。融合网络整体结构为前 8 个块为边-图卷积块,最后一个块为交互关系块。其最后输出的特征可表示为

$$\boldsymbol{F}_{\text{out}} = \sum_{j} \{\boldsymbol{F}'_{\text{Eout}}[(\boldsymbol{A}_j + \boldsymbol{A}_{\text{RN}}) \otimes \boldsymbol{M}]\boldsymbol{W}_j\} \tag{7}$$

式(7)中:$\boldsymbol{F}'_{\text{Eout}}$ 为第 8 层 E-GCB 的输出特征。

设计的改进 ST-GCN 网络共有 9 个时空特征提取单元组成,分别为 8 个 E-GCB 和一个 IR-GCB,输出通道数前 3 层为 64,中间 3 层为 128,最后 3 层为 256,在第 4 层和第 7 层时域的步长设为 2 作为池化层,其它层均为 1,其中每个层的时间卷积核大小为 9,最后经过 MLP 和 FC 操作得到 256 维特征向量,进而送入 softmax 激活函数进行分类与识别,实现边-图卷积块与交互卷积块的有效融合。

5 实验结果与分析

5.1 数据集介绍

在 NTU RGB + D 数据集上进行训练与测试。NTU RGB + D 数据集[24]是一个包含广泛通用动作的数据集,包含 60 个种类的动作,共 56 880 个样本,其中有 40 类为日常行为动作,9 类为与健康相关的动作,11 类为双人交互动作,即打/拍、踢、推、拍背、用手指、拥抱、递东西、摸口袋、握手、走向、分开,该数据集由精确的 Kinect v2 深度体感相机采集得到,包括深度信息、3D 骨骼信息、RGB 帧以及红外序列 4 种数据类型,包含每个人所有帧的 25 个关节的 3D 坐标,具有 40 个不同的主题和较大的视点变化,通过使用 3 个摄像头同时记录。另外,为了评估模型性能,提出两种数据集划分标准:交叉主体(cross-subject, CS),在此模式下,使用预先定义的 20 个人物进行模型训练,而其他人物的数据用于测试;交叉视图(cross-view, CV),选用 3 个不同角度的相机中的相机 1 采集的样本作为测试集,相机 2 和 3 保留为训练集。此数据集采集的数据标准、规模大,条件极具挑战性,选用数据集中的 11 类交互动作,将其三维关节点数据作为数据源,选择 CV 划分标准划分数据集,以适应复杂多变的环境。数据库中基于 RGB 视频帧和关节点数据的不同视角不同主体的部分双人交互行为示例如图 5 所示。

5.2 实验环境和实验设置

网络输入大小为(N,C,T,V,M)的 5 维矩阵,其中 N 设置为 16,C 为 3,T 设置为 300,小于 T 帧的序列填充空白帧在视频末端,由于输入由单人图改进为双人交互图,因此双人视为一个整体输入,V 为

图 5　NTU RGB + D 双人交互动作示意图

Fig. 5　Schematic diagram of NTU RGB + D two-player interaction action

50,M 为 1,考虑到双人的位置信息,在文献[7]数据预处理的基础上,增加了镜像处理操作,保证了数据集的多样性。

实验在 Windows11 操作系统下进行,采用基于 Python3.8 的深度学习框架 Pytorch,GPU 为一个 NVIDIA GeForce RTX 3060,处理器为 12th Gen Intel(R) Core(TM) i5-12400F 的深度学习环境。时间卷积窗口设为 9,最大图采样长度设为 2,训练集和测试集的批量大小为 16,选择交叉熵损失作为损失函数,选用动量为 0.9,权值衰减为 0.000 1 的随机梯度下降(stochastic gradient descent, SGD)优化器,学习率初始化为 0.1。实验均通过 100 轮训练进行训练与测试,由测试精度、测试损失及响应速度为评估指标进行结果分析。

5.3　实验结果

为了验证网络改进模块的有效性,将单独加入非自然连接或交互连接的模型分别在 NTU RGB + D 数据库下进行测试,对比分析分别加入两种连接对网络性能的影响,并对改进 ST-GCN 网络最终结果进行分析与评测。

5.3.1　加入非自然连接的有效性验证

在原始 ST-GCN 网络结构基础上,将所有时空图卷积单元替换为能够捕获非自然连接关系的 E-GCB,9 个 E-GCB 进行堆叠组成完整网络,将数据集数据送入此网络进行训练与测试。加入边卷积操作前,稳定后的测试识别精度为 93.49%,损失函数稳定在约 0.3;而加入边卷积操作后,稳定识别率为 95.69%,提高了 2.2 个百分点,损失函数稳定在约 0.2,网络整体性能得到提升。测试前后得到的混淆矩阵如图 6 所示。

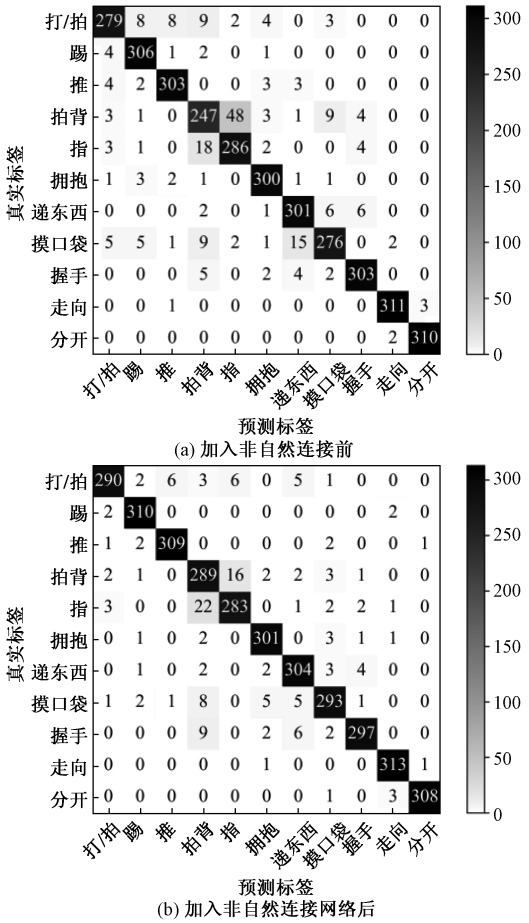


图 6　ST-GCN 加入非自然连接前后的混淆矩阵

Fig. 6　Confusion matrix of ST-GCN before and after the inclusion of unnatural connections

从混淆矩阵的角度分析,加入边卷积操作前,如图 6(a)所示,“拍背”“指”和“递物品”“摸口袋”动作混淆较为明显,两种混淆动作中交互双方执行

过程相似的很高。由于 ST-GCN 网络中仅依赖人体骨骼的物理连接方式进行推理,因此捕获时空变化特征对于相似动作的辨识性不高,因此造成混淆。加入边卷积操作后,如图 6(b)所示,由于在自然连接的基础上构建非自然连接关系,加强了远距离节点之间的联系,以上两类易混淆动作得识别准确率明显提升,从而验证了加入非自然连接的有效性。

5.3.2 加入交互关系的有效性验证

保留 ST-GCN 的网络结构,将其所有时空图卷积单元改进为交互关系块 IR-GCB,为了充分证明算法的有效性和合理性,对加入关系连接的双人交互行为识别网络进行结果测试。与加入交互关系前的 ST-GCN 相比,本网络的精度和损失都更具优势,趋于平稳后的精度达 96.30%,提升了 2.81 个百分点,损失维持在约 0.15。其测试生成的混淆矩阵如图 7 所示。

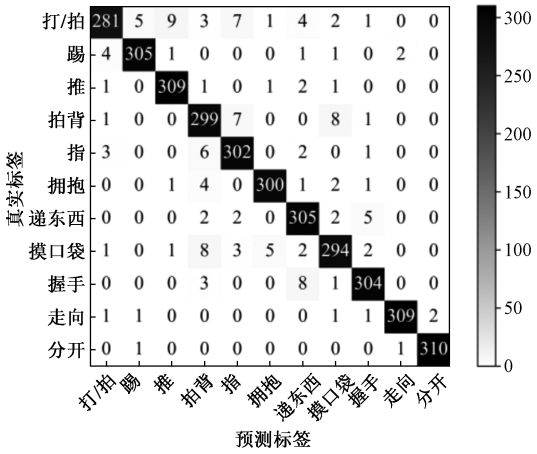


图 7 ST-GCN 加入交互关系后的混淆矩阵

Fig. 7 Confusion matrix of ST-GCN after adding interactions

分析混淆矩阵可知,加入交互关系后的ST-GCN网络善于捕获双人间的交互关系,明显提升了 ST-GCN 对于相似动作“拍背”“指”和“递东西”“摸口袋”的识别准确性。充分证明加入交互关系后网络的特征辨别能力得到明显提升。

5.3.3 最优融合方式测试

为了使所提出的网络结构最优,设计 3 种方式将 IR-GCB 和 E-GCB 与 ST-GCN 进行有效融合。一是将 IR-GCB 和 E-GCB 两种块融合为一体,结合为新的特征提取块;二是交互关系块置于边-图卷积块的前侧;三是交互关系块置于边-图卷积块的后侧,测试结果如表 1 所示,测试结果均为每类方案多种组合方式中的最高测试精度。

分析可知,由于边图卷积块中动态图每层更新一次,而交互关系块中以帧为基准进行更新,且边-图卷积模块的边卷积操作在构建局部邻域图时,部分节点的边特征也包含双人间的交互特征,与交互关系块中的 RN 提取的交互特征缺乏归一化操作后的有效补充,因此方案 1 结果不理想,从而确定两个块融合为一个块的方案不可行;分析融合方案 2、3 的测试结果可知,由于进行深层次网络卷积过程中会存在特征削弱的问题,交互特征在解决此类问题中表现更突出,因此确定边-图卷积块在前,交互关系块在后,且最佳匹配结果对应融合网络结构为前 8 块为边-图卷积块,最后一块设置为交互关系块。实验得到的训练集和测试集的损失及精度曲线如图 8 所示。

由图 8 可知,经过 30 个 epoch 后,精度和损失函数均已趋于平稳,平稳精度为 97.77%,损失函数大致趋于 0.1,表明改进的时空图卷积网络性能稳定且良好,测试生成的混淆矩阵如图 9 所示。

观察图 9 可知,改进 ST-GCN 网络的混淆矩阵中各类动作混淆程度均不高,虽然对“指”和“拍背”

表 1 E-GCB 与 IR-GCB 融合方式测试结果

Table 1 Test results of E-GCB and IR-GCB fusion mode

算法	精度/%
ST-GCN 基线	93.49
融合方案 1	97.42
融合方案 2	97.66
融合方案 3	97.77

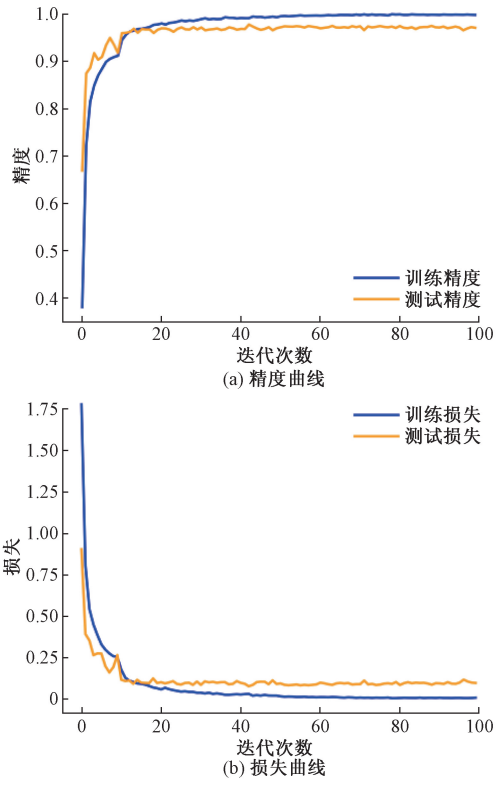


图 8 改进 ST-GCN 的测试结果

Fig. 8 Improvement of ST-GCN test results

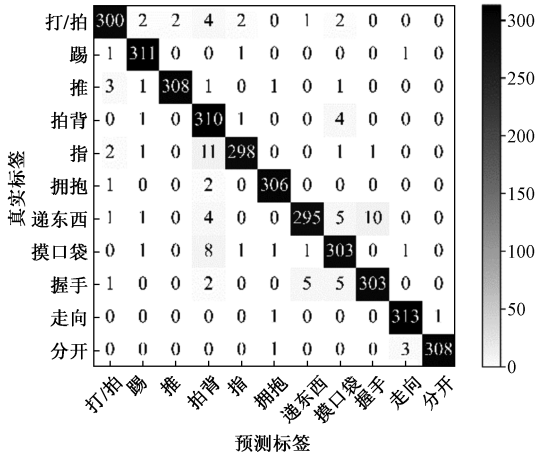


图 9　改进 ST-GCN 后的混淆矩阵

Fig. 9　Confusion matrix after improvement of ST-GCN

相似度高的动作仍存在轻微混淆,但整体达到预期识别效果且均优于融合前网络,证明非自然连接和交互连接间存在特征互补性,验证了融合网络的有效性。

5.4　与先进网络模型对比

为了进一步验证本文算法的有效性,将其实验结果同基于关节点的其他算法在 NTU RGB + D 交互动作数据集下进行结果对比,整理为表 2 所示。

表 2　本文算法与其他先进算法的比较

Table 2　Comparison of the proposed algorithm with other state-of-the-art algorithms

算法	精度/%
ST-GCN[7](基线)	93.49
AS-GCN[8]	93.00
K-GCN[13]	96.01
2S-AGCN[11]	96.67
CTR-GCN[16]	96.80
DR-GCN[18]	97.19
本文算法	97.77

文献[8,11,13,16]算法均构建了动态邻接矩阵改进 ST-GCN 网络,但以图卷积操作为基础的特征提取,每个节点邻节点数目不同,节点间不满足位置置换不变性,且忽略了双人间重要的交互信息,特征提取不够充分;文献[18]算法实现过程复杂,且仅提取了双人间的浅层交互特征,又缺乏对人体潜在非自然连接的关注,模型缺乏鲁棒性。本文模型利用边卷积操作来构建非自然连接,其邻节点数目固定,既能高效提取局部特征,又满足置换不变性,且引入善于捕获个体交互关系的 RN 网络来构建双人交互连接关系,使得网络不仅具有强大的特征提取能力,平衡精度和速度,且有较高的灵活性和泛化性。

6　结论

提出一种新的改进时空图卷积模型的双人交互行为识别网络。模型基于关节点数据,通过引入边卷积操作构建边-图卷积块来获得远距离节点相关性信息,改进 RN 网络生成交互关系图嵌入 ST-GCN 更新静态邻接矩阵为动态邻接矩阵,灵活获取有价值的交互信息,突破了以往模型的局限。实验结果表明,本文模型能够有效捕捉不同关节点间潜在的相关性,与先进算法相比具有突出优势。但考虑到关节点数据缺乏捕捉外观信息的能力,因此今后在此研究基础上计划加入 RGB 数据源丰富外观信息,且拓宽识别领域到人-物交互和多人交互行为识别,并试图构造一个通用的人类行为识别模型。

参　考　文　献

[1] Zhang H B, Zhang Y X, Zhang B N, et al. A comprehensive survey of vision-based human action recognition methods[J]. Sensors, 2019, 19(5): 1005.

[2] Baradel F, Wolf C, Mille J, et al. Glimpse clouds: human activity recognition from unstructured feature points[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, D. C.: IEEE Computer Society, 2018: 469-478.

[3] Ding Z W, Wang P C, Ogunbona P O, et al. Investigation of different skeleton features for CNN-based 3D action recognition[EB/OL]. (2017-05-02)[2023-08-15]. https://arxiv.org/pdf/1705.00835v1.pdf.

[4] 姬晓飞, 秦琳琳, 王扬扬. 基于 RGB 和关节点数据融合模型的双人交互行为识别[J]. 计算机应用, 2019, 39(11): 3349-3354.
Ji Xiaofei, Qin Linlin, Wang Yangyang. Two-person interaction behavior recognition based on RGB and joint point data fusion model[J]. Computer Applications, 2019, 39(11): 3349-3354.

[5] 武东辉, 许静, 陈继斌, 等. 基于融合注意力机制与 CNN-LSTM 的人体行为识别算法[J]. 科学技术与工程, 2023, 23(2): 681-689.
Wu Donghui, Xu Jing, Chen Jibin, et al. Humanactivity recognition algorithm based on CNN-LSTM with attention mechanism[J]. Science Technology and Engineering, 2023, 23(2): 681-689.

[6] Wang X H, Deng H M. A multi-feature representation of skeleton sequences for human interaction recognition[J]. Electronics, 2020, 9(1): 187.

[7] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutionalnetworks for skeleton-based action recognition[C]//Proceedings of the 2018 AAAI Conference on Artificial Intelligence. New Orleans, LA: AAAI, 2018: 7444-7452.

[8] Li M S, Chen S H, Chen X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 3595-3603.

[9] 刘锁兰, 周岳靖, 王洪元, 等. 基于全局图遍历的 ST-GCN 人

体行为识别算法[J]. 扬州大学学报(自然科学版), 2022, 25(2): 62-68.
Liu Suolan, Zhou Yuejing, Wang Hongyuan, et al. Human behavior recognition based on global graph traversal and ST-GCN[J]. Journal of Yangzhou University (Natural Science Edition), 2022, 25(2): 62-68.

[10] Song Y F, Zhang Z, Shan C F, 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.

[11] Shi L, Zhang Y F, Cheng J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington, D. C.: IEEE Computer Society, 2019: 12018-12027.

[12] 张静亭, 曹江涛, 姬晓飞. 基于图卷积的3D骨架数据的双人交互行为识别[J]. 辽宁石油化工大学学报, 2023, 43(3): 86-90.
Zhang Jingting, Cao Jiangtao, Ji Xiaofei. Recognition of two-person interaction behavior based on graph convolution of 3D skeleton data[J]. Journal of Liaoning University of Petrochemical Technology, 2023, 43(3): 86-90.

[13] Li J N, Xie X M, Cao Y H, et al. Knowledge embedded GCN for skeleton-based two-person interaction recognition[J]. Neurocomputing, 2021, 444: 338-348.

[14] 曹毅, 吴伟官, 李平, 等. 基于时空特征增强图卷积网络的骨架行为识别[J]. 电子与信息学报, 2023, 45(8): 3022-3031.
Cao Yi, Wu Weiguan, Li Ping, et al. Skeletonbehavior recognition based on spatial temporal feature-enhanced graph convolutional networks[J]. Journal of Electronics and Information, 2023, 45(8): 3022-3031.

[15] Plizzari C, Cannici M, Matteucci M. Spatial temporal transformer network for skeleton-based action recognition[C]//Proceedings of the 2021 International Conference on Pattern Recognition International Workshops and Challenges. Berlin: Springer International Publishing, 2021: 694-701.

[16] Chen Y X, Zhang Z Q, Yuan C F, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//Proceedings of the 2021 IEEE International Conference on Computer Vision. (ICCV). Online: Institute of Electrical and Electronics Engineers Inc., 2021: 13339-13348.

[17] Li Z C, Li Y R, Tang L L, et al. Two-person graph convolutional network for skeleton-based human interaction recognition[J]. IEEE Transactions on Circuits and Systems for Video Technvology, 2023, 33(7): 3333-3342.

[18] Zhu L P, Wan B H, Li C Y, et al. Dyadic relational graph convolutional networks for skeleton-based human interaction recognition[J]. Pattern Recognition, 2021, 115(1): 107920.

[19] Li J N, Xie X M, Cao Y H, et al. SGM-Net: skeleton-guided multimodal network for action recognition[J]. Pattern Recognition, 2020, 104(13): 1073-1105.

[20] 成科扬, 吴金霞, 王文杉, 等. 融合时空图卷积的多人交互行为识别[J]. 中国图象图形学报, 2021, 26(7): 1681-1691.
Cheng Keyang, Wu Jinxia, Wang Wenshan, et al. Multi-person interaction action recognition based on spatial temporal graph convolution[J]. Journal of Image and Graphics, 2021, 26(7): 1681-1691.

[21] Wang Y, Sun Y B, Liu Z W, et al. Dynamic graph CNN for learning on point clouds[J]. ACM Transactions on Graphics, 2018, 38(5): 1-12.

[22] Santoro A, Raposo D, Barrett D G T, et al. A simple neural network module for relational reasoning[EB/OL]. (2017-06-05)[2023-08-15]. https://arxiv.org/pdf/1706.01427.pdf.

[23] Jang E, Gu S X, Poole B. Categorical reparameterization with gumbel-softmax[C]//Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon: International Conference on Learning Representations, 2017: 324-346.

[24] 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV: IEEE Computer Society, 2016: 1010-1019.