基于增强时空图卷积网络的骨架行为识别

姜 维,关孟怡,魏富鹏,孙浩宸,孟 尧,吴慧欣
(华北水利水电大学信息工程学院,河南郑州 450000)

摘 要: 图卷积网络(Graph Convolutional Network,GCN)被广泛应用在基于骨架序列的行为识别方法中,并取得显著效果. 然而,随着行为种类和场景复杂度的增加,现有方法在建模人体结构细节与时序依赖方面仍面临诸多挑战,具体表现为以下两个问题:其一,在提取关节间的关联特征时,往往未能充分反映边缘处关节(双手、双脚与头部)之间的相互作用以及边缘处关节与其他关节之间的协同效应;其二,在提取时间特征时,局限于短期时间特征的提取,未能有效捕获长期时序依赖关系. 针对以上问题,本文提出一种增强时空图卷积网络模型(Enhanced Spatial-Temporal Graph Convolutional Network,EST-GCN),它由多分支空间增强图卷积(Multi-branch Spatial Enhanced Graph Convolution,MSEGC)模块和多尺度时间增强卷积(Multi-scale Temporal Enhanced Convolution,MTEC)模块堆叠组成. MSEGC通过多阶段学习并传递双流图卷积下的特征,以增强边缘处关节的特征表达能力,从而捕获边缘处关节与其他关节之间的关系;MTEC通过多阶段学习并传递多尺度时间卷积下的时间特征,扩大时间跨度,从而捕获时间帧之间更广泛的时序依赖关系. 模型依次通过MSEGC与MTEC提取并融合空间与时间特征,协同建模关节结构关联与时序依赖,提升时空特征判别性. 为充分挖掘骨架数据的时空特征,在输入设计上,本文引入关节位置、运动速度与骨骼3类特征,并采用多流融合方式以增强特征表示能力. 本文所提出的方法,在NTU-RGB+D数据集的X-Sub与X-View基准上,分别实现了92.4%与96.2%的准确率;在NTU-RGB+D 120数据集的X-Sub与X-Setup基准上,分别达到了88.7%和90.0%的准确率,证明了该方法的有效性. 此外,为进一步验证模型在真实场景下的人体行为识别性能,本文基于NTU-RGB+D数据集的视频样本开展了骨架行为识别实验,并在多人交互及关节噪声干扰条件下进行了额外测试. 实验结果表明,即使在局部关节出现错乱分配的情况下,模型仍能实现准确识别,验证了所提方法的实用性与鲁棒性.

关键词: 行为识别;骨架序列;图卷积网络;多分支空间图卷积;多尺度时间卷积;时空特征;多流融合
基金项目: 国家自然科学基金(No.42371466);河南省高等学校重点科研项目计划(No.23A520031,No.24A520020,No.25B520012);河南省住房城乡建设科学技术计划项目(No.HNJS-2024-K35)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)10-3692-13
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250259

Enhanced Spatial-Temporal Graph Convolutional Network for Skeleton-Based Action Recognition

JIANG Wei, GUAN Meng-yi, WEI Fu-peng, SUN Hao-chen, MENG Yao, WU Hui-xin
(*School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou, Henan* 450000, *China*)

Abstract: Graph convolutional network (GCN) has been extensively applied to skeleton-based action recognition and have achieved remarkable performance. However, as the number of action categories and scene complexity increase, existing methods still face significant challenges in modeling detailed human body structures and temporal dependencies, which can be summarized as two main issues. Firstly, when extracting relational features among joints, these methods often inadequately capture the interactions between peripheral joints (such as hands, feet, and head) and their synergistic effects with other joints. Secondly, when extracting temporal features, these methods focus on short-term temporal feature extraction neglecting of long-term dependencies. To address these issues, this paper proposes an enhanced spatiotemporal graph convolutional network (EST-GCN), which consists of multi-branch spatial enhanced graph convolution (MSEGC) and multi-scale temporal enhanced convolution (MTEC) modules. The MSEGC module enhances the feature representation of peripheral joints by capturing relationships between peripheral joints and others through multi-stage learning and propagation within a two-stream graph convolution framework. Meanwhile, the MTEC module effectively captures long-term temporal depen-

收稿日期:2025-04-06;录用日期:2025-10-09;责任编辑:覃怀银

dencies across frames through multi-stage learning and propagation of temporal features from multi-scale convolutions, thereby expanding the temporal receptive field. The model sequentially extracts and fuses spatial and temporal features via MSEGC and MTEC, jointly modeling joint structural correlations and temporal dependencies to improve the discriminability of spatial-temporal features. To fully exploit the spatial-temporal information of skeleton data, three types of input features—joint positions, motion velocities, and bone features—are introduced and fused through a multi-stream strategy to enhance feature representation. The proposed method achieves accuracies of 92.4% and 96.2% on the X-Sub and X-View benchmarks of the NTU-RGB+D dataset, respectively; and 88.7% and 90.0% on the X-Sub and X-Setup benchmarks of the NTU-RGB+D 120 dataset, which validates its effectiveness. Furthermore, to validate the model's performance in real-world scenarios, additional skeleton-based action recognition experiments are conducted on video samples from the NTU-RGB+D dataset, including tests under multi-person interactions and joint noise interference. The results show that the model can still achieve accurate recognition even when local joint misassignments occur, further verifying the practicality and robustness of the proposed approach.

Key words: action recognition; skeleton sequence; graph convolutional network; multi-branch spatial enhanced graph convolution; multi-scale temporal enhanced convolution; spatial-temporal features; multi-stream strategy

Foundation Item(s): National Natural Science Foundation of China (No.42371466); Key Research Projects of Henan Higher Education Institutions (No.23A520031, No.24A520020, No.25B520012); Science and Technology Plan Project of Housing and Urban-Rural Development in Henan Province (No.HNJS-2024-K35)

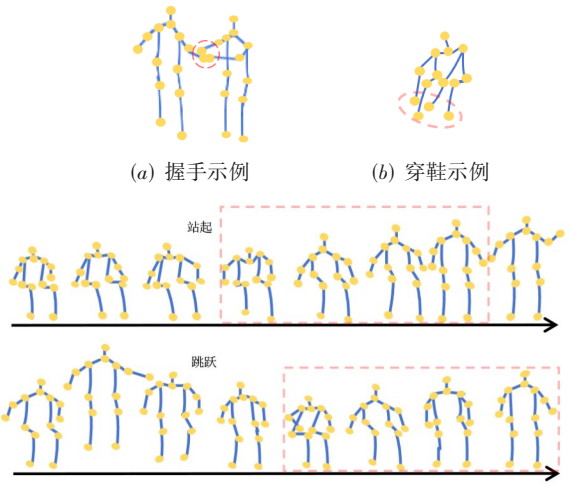
1　引言

人体行为识别是计算机视觉中一项被广泛研究但仍未解决的分类任务,在视频监控、人机交互等各种应用中受到了广泛的关注. 特别是基于骨架的人类行为识别,它不是从原始的RGB视频中预测,而是通过关节和骨骼表示人体结构、关节坐标表示人体位置,来预测行为. 近年来的一些工作结果[1~3]已经证明了它的有效性.

早期基于深度学习的方法[4]将人类关节作为一组独立的特征,并将其组织成特征向量或伪图像,然后将其输入循环神经网络(Recurrent Neural Network,RNN)或卷积神经网络(Convolutional Neural Network,CNN)来预测行为标签. 这两种方法忽略了人体骨架的自然拓扑结构,不能充分探索关节之间的连接,识别效果有限. 基于图卷积神经网络(Graph Convolutional Network,GCN)的方法将人体骨架简化为由顶点和边组成的拓扑图. 与基于RNN和CNN的方法相比,GCN以图结构更自然地对骨架数据进行建模,取得了更好的效果. 时空图卷积网络(Spatial Temporal Graph Convolutional Network,ST-GCN)[1]是第一种使用时空图卷积模块来建模骨架运动模式的方法. 每一个这样的模块由空间图卷积模块和时间卷积模块组成. 前者融合了相邻关节之间的特征,后者提取运动特征. 由于以上性质,ST-GCN具有更强的行为建模能力和更好的性能,之后基于图卷积模型的研究越来越多.

在后续的研究中,许多基于GCN的模型被相继提出. 然而,在提取骨架数据的空间特征时,大多数模型往往忽略了人体拓扑结构中边缘处关节所提供的重要信息,而边缘处关节对于行为识别起着至关重要的作用. 例如,如图1(a)所示,在两人互动的"握手"行为中,两人将手合在一起,手与手之间有很强的联系. 同样,如图1(b)所示,在穿鞋过程中,手与脚之间的相互作用尤为紧密. 然而,传统方法所捕获的关节关系并不能充分反映两只手之间、手与脚之间的相互作用. 因此,构建一个高效的空间特征提取器,以增强边缘处关节的特征表达能力,并精确捕捉边缘处关节之间以及边缘处关节与其他关节之间的复杂依赖关系,显得尤为关键. 另外,不同行为所捕捉的局部时间特征存在相似性,时间分辨率的局限性往往会使模型作出错误判断. 例如,如图1(c)所示,在框定时间段内,站起和跳跃具有相似的肢体动作,极易混淆. 因此,捕捉长时间跨度的特征,以捕获更广泛的时序依赖关系,对于构建一个高效的行为识别模型至关重要. 目前,已有基于GCN的模型[5,6]考虑到其中一个或两个方面. 然而,这些问题还远未得到有效解决.

针对这些问题,本文设计一种增强时空图卷积网络(Enhanced Spatial-Temporal Graph Convolutional Network,EST-GCN). 该网络由多分支空间增强图卷积(Multi-branch Spatial Enhanced Graph Convolution,MSEGC)模块和多尺度时间增强卷积(Multi-scale Temporal Enhanced Convolution,MTEC)模块堆叠组成. 输入骨架序列依次被送入MSEGC和MTEC. 这两个模块专注于学习不同的特征,MSEGC负责捕捉空间维度上关节之间的特征,MTEC则专注于提取时间维度上的运动特征. MSEGC不仅捕捉人体自然拓扑结构中关节的连接模式,还以数据驱动的方式提取每帧中边缘处关节之间的连接关系. 通过融合这些特征并进行多阶段的信息传递,MSEGC逐步强化边缘处关节的特征表达



(a) 握手示例 (b) 穿鞋示例

(c) 站起与跳跃示例

图1 行为示例

能力,进而有效增强边缘处关节之间及其与其他关节之间复杂的相互作用,提高模型对细粒度活动的识别能力.MTEC具备捕获与融合多个时间感受野下特征的能力,并通过多阶段的信息传递机制有效扩展时间跨度,进而捕捉到更为广泛的时序依赖关系,提高模型对相似行为的辨识能力.两个模块分别专注于建模骨架关节之间的依赖性和时间帧之间的动作关联性,为行为识别提供更全面的特征表示.

本文在具有挑战性的大规模数据集NTU-RGB+D[7]与NTU-RGB+D 120[8]上进行了综合实验,结果表明,本文模型在准确率方面取得不错的效果.本文的贡献总结如下:

(1)本文提出多分支空间增强图卷积(MSEGC)模块,它通过多阶段地学习并传递边缘处关节之间及人体自然拓扑结构关节间连接关系的融合特征,从而增强边缘处关节间的相互作用以及边缘处关节与其他关节间的协同作用.

(2)本文提出多尺度时间增强卷积(MTEC)模块,它通过多阶段地学习并传递不同时间感受野内的特征,扩大时间跨度,捕获时间帧之间更广泛的时序依赖关系.

(3)基于以上二者,本文建立了EST-GCN模型.在两个数据集上的实验结果证明了该模型的有效性.

2 相关工作

2.1 图卷积网络

图卷积神经网络(GCN)被广泛用于处理非欧几里得数据,如社交网络[9]和人类骨架结构[1].现有的GCN模型主要分为两大类:基于光谱的方法和基于空间的方法.基于光谱的GCN方法[10]通过特征分解技术将图信号转换到频谱域,并在此基础上执行卷积操作.然而,特征分解速度较慢,限制了计算效率.相比之下,基于空间的GCN方法[11]是将卷积网络从欧几里得空间到非欧几里得空间的扩展.该方法基于空间域内图节点及其邻居节点的骨架数据构建时空图,并在邻域内应用卷积进行特征聚合.由于其能够灵活处理各种图结构,大多数基于骨架的行为识别方法都遵循空间的GCN思想.本文也采用基于空间的GCN方法,对行为识别展开研究.

2.2 基于骨架的行为识别

与RGB等其他图像数据相比,骨架数据对背景变化等具有鲁棒性,因此基于骨架的方法在行为识别中得到了广泛的关注.基于骨架的人体行为识别方法主要分为两种流程:基于手工制作的方法和基于深度学习的方法.早期的基于骨架的行为识别方法[12]通常使用手工制作的特征来捕捉人体动作.然而,由于人为地以非自然的形式建模骨骼数据,这些方法在提取语义特征以理解人体结构的相关性方面存在局限性,很难考虑到与人体运动相关的所有因素.随着深度学习的发展,数据驱动的自学习方法变得越来越重要.基于RNN的方法[13]将骨架数据视为一个时间序列,首先提取帧级骨架特征,然后对顺序依赖关系进行建模.然而,该方法存在训练困难和并行性差的问题.基于CNN的方法[14]需要将骨架序列转换为伪图像,这会导致一些时空信息的丢失.上述两类行为识别方法将骨架数据转化为向量或伪图像的方式忽略了人体骨架的自然拓扑结构,识别效果有限.

骨架数据具有天然的拓扑特性,GCN能够充分利用骨架数据中的时空信息,更有效地对其拓扑结构进行建模.ST-GCN[1]通过定义一个稀疏矩阵连接的时空图,同时考虑人体骨骼在时间域和空间域上的运动依赖性,从而实现对骨架结构的精准建模.ST-GCN不再依赖于手工制作的遍历规则,展现出更强的行为建模能力和更优的性能.在其基础上,诸多以GCN为基础的方法,如2s-AGCN[2]、FGCN[15]、HD-GCN[16]、MD^{2}TL-GCN[17]、MADT-GCN[18]和SelfGCN[19]都遵循ST-GCN的基本思想,探索更有效的基于骨架的行为识别方法.

为了增强关节的特征表达能力,前人已经探索了多种方法.其中,一些方法利用骨架邻接矩阵的高阶多项式来提取多尺度的结构特征.例如,Yang等人[15]引入反馈机制,将关节间的高层语义信息传递到低层,并通过多阶段时间采样策略逐步提取时空特征,实现对行为的准确识别.Xia等人[18]提出一种高效且有效的基于骨架的行为识别方法,它通过自适应图卷积网络计算关节之间的空间、时间和通道维度相似度,并构建相应的连接,有效地捕捉关节之间的复杂关系.Xie等

人[20]提出DS-GCN,该模型将关节和边的类型以隐式方式编码到骨架拓扑中,旨在通过优化节点之间的连接方式来提升基于骨架的动作识别性能.尽管上述方法在建模骨架时提升了关节的表达能力,但在增强边缘处关节的特征表达能力方面仍存在不足,未能充分理解边缘处关节之间的相互作用及其与其他关节之间的协同效应.此外,现有研究更多地关注骨架在空间维度上的特征建模,而往往未能充分提取时间维度上的信息.为丰富时间维度的特征,Liu等人[5]设计了MS-G3D网络,将不同阶的邻接矩阵和不同膨胀率的多分支时间卷积与三维卷积(3D)相结合,构建一个解耦和统一的时空多尺度图卷积网络.Zhu等人[21]提出了时间运动激励方法,利用时间差分的概念突出运动敏感特征,从而更有效地捕捉行为的动态变化.尽管这些方法在丰富时间特征方面取得了进展,但在捕捉长时间跨度的特征以及获取广泛的时序依赖关系方面,仍然存在一定的局限性.

为了解决这些问题,本文提出增强时空图卷积网络模型(EST-GCN),通过多阶段学习并传递特征信息,显著增强了边缘处关节的特征表达能力,丰富关节间连接关系信息,同时有效扩展了长时间跨度,捕捉行为的变化轨迹,获取更广泛的时序依赖关系,全面提升行为识别性能.

3　方法

3.1节和3.2节分别阐述了基于图卷积的骨架行为识别的符号定义以及对骨架序列数据进行预处理的方法.在3.3节中,介绍了本文所提出的增强时空图卷积网络模型(EST-GCN)的总体架构,如图2所示.在3.4节和3.5节中,分别对提出的多分支空间增强图卷积(MSEGC)模块和多尺度时间增强卷积(MTEC)模块进行了详细阐释.

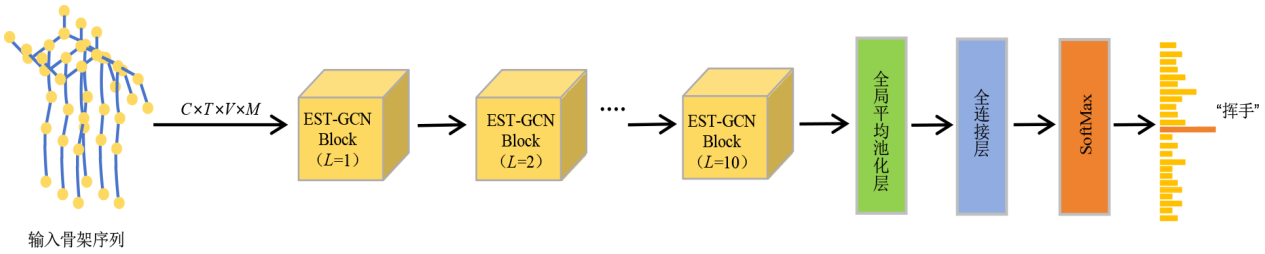


图2　EST-GCN全局网络结构

3.1　骨架图符号定义

一个人体骨架可以被建模为一个图$G=(V,E)$.其中,N个关节被表示为顶点集$V=\{v_1,v_2,\cdots,v_N\}$;骨骼形成的边被表示为集合E,它由邻接矩阵$\boldsymbol{A}\in\{0,1\}^{N\times N}$捕获.如果关节$i$和$j$直接连接,则元素$a_{ij}$为1,否则为0.节点$v_i$的邻域为$N(v_i)=\{v_j|a_{ij}\neq 0\}$.人类骨架行为序列表示为$\boldsymbol{X}\in\mathbf{R}^{C\times T\times V\times M}$长度的动作序列.数量$C$表示通道数,代表人体关节的坐标;数量$T$表示时间帧;数量$V$表示节点数,数量$M$表示人数.$\boldsymbol{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_T\}$是包含每一帧节点特征的集合,其中$\boldsymbol{X}_i\in\mathbf{R}^{C\times V\times M}$表示第$i$帧关节的特征.

对于每一帧的输入骨架数据$\boldsymbol{f}_{\text{in}}$,传统的共享拓扑图卷积使用权值$\boldsymbol{W}$来变换特征.然后,它使用邻接矩阵$\boldsymbol{A}$来聚合时间帧中每个节点的相邻节点的特征.这可以表示为

$$\boldsymbol{f}_{\text{out}}=\sigma\left(\sum_{i=1}^{3}\hat{\boldsymbol{A}}_i\boldsymbol{f}_{\text{in}}\boldsymbol{W}_i\right)\tag{1}$$

其中,$\hat{\boldsymbol{A}}_i=\boldsymbol{D}_i^{-\frac{1}{2}}\boldsymbol{A}_i\boldsymbol{D}_i^{\frac{1}{2}}$表示第$i$个子集的归一化邻接矩阵,$\boldsymbol{A}_i$表示关节间距离为$i$的邻接矩阵,$\boldsymbol{D}_i$表示第$i$个子集的度矩阵,$\sigma(\cdot)$表示激活函数.

3.2　数据预处理

根据先前的研究[17,18],数据预处理对于基于骨架的行为识别非常重要.本文采用与基准模型Effcient-GCN[22]相同的数据预处理方法,如图3所示,将输入特征分为以下3类:(1)关节位置特征Joint;(2)关节运动速度特征Velocity;(3)骨骼特征Bone.其处理过程如下:假设一个行为序列的原始三维(3D)坐标集:$\boldsymbol{X}=\{\boldsymbol{x}\in\mathbf{R}^{C_{\text{in}}\times T_{\text{in}}\times V_{\text{in}}\times M_{\text{in}}}\}$,其中$C_{\text{in}}$、$T_{\text{in}}$、$V_{\text{in}}$分别表示输入坐标、输入帧数和输入关节数.然后,得到相对位置集$\boldsymbol{R}=\{\boldsymbol{r}_i|i=1,2,\cdots,V_{\text{in}}\}$作为标准化位置特征.将原始关节位置集$\boldsymbol{X}$与相对关节位置集$\boldsymbol{R}$连接起来,作为关节位置输入特征Joint.定义如下:

$$\begin{aligned}\boldsymbol{r}_i&=\boldsymbol{x}[:,:,i,:]-\boldsymbol{x}[:,:,c,:]\\ \text{Joint}&=\boldsymbol{X}\oplus\boldsymbol{R}\end{aligned}\tag{2}$$

其中,c表示脊柱中心关节的索引.此外,很容易得到两组运动速度:快速运动$F=\{\boldsymbol{f}_t|t=1,2,\cdots,T_{\text{in}}\}$和慢速运动$S=\{\boldsymbol{s}_t|t=1,2,\cdots,T_{\text{in}}\}$,定义如下:

$$\begin{aligned}\boldsymbol{f}_t&=\boldsymbol{x}[:,t+2,:,:]-\boldsymbol{x}[:,t,:,:]\\ \boldsymbol{s}_t&=\boldsymbol{x}[:,t+1,:,:]-\boldsymbol{x}[:,t,:,:]\\ \text{Velocity}&=F\oplus S\end{aligned}\tag{3}$$

将关节快速运动集F与关节慢速运动集S连接起

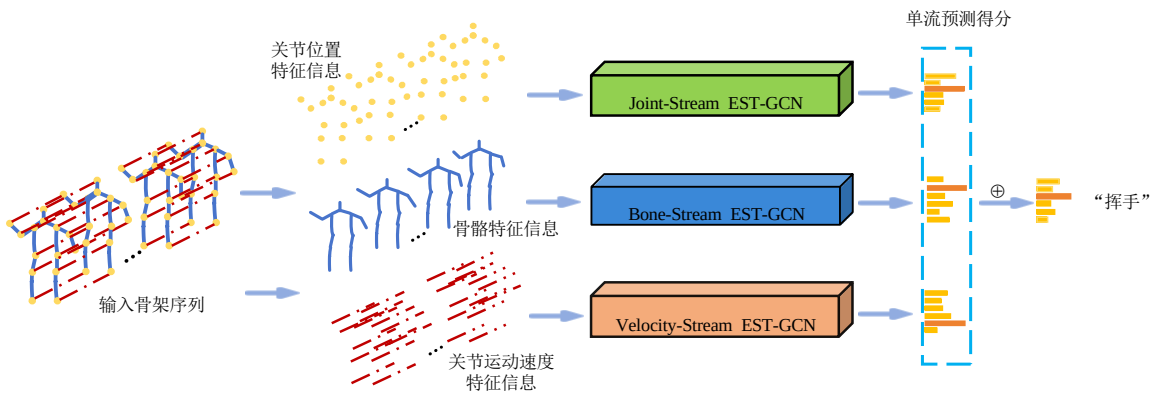


图3 EST-GCN网络的3流输入架构(⊕表示加权求和)

来,作为关节运动速度输入特征Velocity. 最后,骨骼输入特征Bone由骨骼长度集$L=\{\boldsymbol{l}_i|i=1,2,\cdots,V_{\text{in}}\}$与骨骼角度集$A=\{\boldsymbol{a}_i|i=1,2,\cdots,V_{\text{in}}\}$连接而成. 为了得到这两组骨骼数据,每根骨骼的长度和角度都被计算为

$$\boldsymbol{l}_i=\boldsymbol{x}[:,:,i,:]-\boldsymbol{x}[:,:,i_{\text{adj}},:]$$

$$\boldsymbol{a}_{i,w}=\arccos\left(\frac{\boldsymbol{l}_{i,w}}{\sqrt{\boldsymbol{l}_{i,x}{}^2+\boldsymbol{l}_{i,y}{}^2+\boldsymbol{l}_{i,z}{}^2}}\right) \tag{4}$$

$$\text{Bone}=L\oplus A$$

其中,i_{adj}为第i个关节的相邻关节,$w\in\{x,y,z\}$为三维坐标. 该骨骼输入特征被定义为由靠近骨架重心的源关节指向远离重心的目标关节的向量,因此对于连接两个关节的节点,统一将靠近重心的一端视为其相邻关节.

3.3 EST-GCN模型总体架构

本文设计了多分支空间增强图卷积(MSEGC)模块和多尺度时间增强卷积(MTEC)模块,它们的结构分别如图4②和图4③所示. 将这两个模块堆叠组合,构建一个带有注意力机制[22]的增强时空图卷积网络基础层(EST-GCN Block),其结构如图4①所示. 进一步通过叠加L个网络基础层,构建用于基于骨架的人体行为识别模型EST-GCN,其整体架构如图2所示. 为了达到更精确的分类效果,本文采用了与基准模型[22]相同的3流输入框架来处理数据,并最终选取3流求和的结果作为最终的决策依据,其方法框架如图3所示.

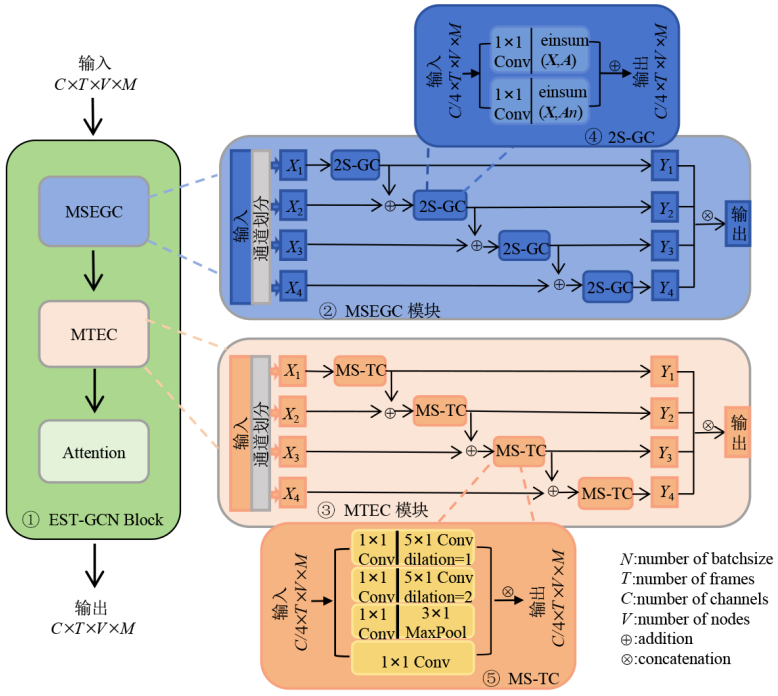


图4 网络基础层(EST-GCN Block)的组成和结构

本文首先通过MSEGC对T帧输入骨架序列的空间特征进行逐级提取与传递,以增强关节间关联特征的表达能力. 随后,利用MTEC对空间建模结果进行逐级增强的时间特征建模,以捕捉广泛的时间序列依赖性. 经L层时空建模处理后的特征将依次通过全局平均池化层和全连接层,最终输出行为分类得分. 该模型在提高对细微动作识别能力的同时,也增强了对相似动作的区分精确度. 同样,即便在关节数据中存在噪声或数据缺失的情况下,模型仍能通过多阶段的信息交流和补偿机制,有效地进行应对,这确保了行为识别的精确性和鲁棒性.

3.4　多分支空间增强图卷积模块

当采用传统图卷积网络对预定义邻接矩阵进行人体骨架数据的空间特征提取时,往往难以深入挖掘关节之间的复杂关联性. 因此,为了充分提取输入特征的空间信息,增强边缘处关节的特征表达能力,本文提出了一种多分支空间增强图卷积(MSEGC)模块,其整体结构如图4②所示.

为了减少计算复杂度,在同样的通道数中表达更加多样的空间结构特征,MSEGC首先将输入特征$\boldsymbol{X}\in\mathbf{R}^{C\times T\times V\times M}$在通道维度上均匀划分为4个子集,每个子集$\boldsymbol{X}_i\in\mathbf{R}^{C'\times T\times \mathrm{V}\times M}$作为一个独立的分支,逐级输入到图卷积模块中,捕捉各通道中关节之间的相关性信息. MSEGC中的图卷积模块采用双流图卷积结构(Two-Stream Graph Convolution,2S-GC),其结构如图4④所示. 该图卷积模块不仅基于人体自然拓扑结构预定义的邻接矩阵$\boldsymbol{A}$,以提取全局的关节拓扑关系,还以数据驱动的方式,通过FastDTW函数计算每帧中边缘处关节之间的相关性,从而生成边缘邻接矩阵$\boldsymbol{A}_n$,以捕捉边缘处关节之间的关联特征. 随后,通过加权融合双流特征,综合表征边缘处关节之间的依赖关系以及边缘处关节与其他关节之间的拓扑关系. 其计算过程为

$$\boldsymbol{Y}_i=\left(\sum_{i=0}^{d}\boldsymbol{W}\boldsymbol{X}_i\left(\hat{\boldsymbol{A}}_i\odot\boldsymbol{M}_i+\boldsymbol{B}_j\right)\right)+\left(\boldsymbol{W}\boldsymbol{X}_i\boldsymbol{A}_n\odot\boldsymbol{M}_n\right)\tag{5}$$

其中,$\boldsymbol{Y}_i$为双流特征的综合性特征表示;$\boldsymbol{W}$是用于执行卷积操作的可学习参数;d是预定义的最大图距离;$\hat{\boldsymbol{A}}_i$是预定义的第i个子集的归一化邻接矩阵;⊙表示点乘;$\boldsymbol{M}_i$与$\boldsymbol{M}_n$、$\boldsymbol{B}_j$是与邻接矩阵$\boldsymbol{A}$具有相同维度的矩阵,它们用于调整每条边的重要性,并与训练过程中的其他参数共同进行优化.

为了有效扩展空间维度的等效感受野,MSEGC将当前分支所提取的综合性特征表示$\boldsymbol{Y}_i$以级联的方式与下一个子集$\boldsymbol{X}_{i+1}$进行融合,并将其作为输入再一次传入2S-GC模块中,进行下一分支的特征学习. 此过程可表示为

$$\boldsymbol{Y}_{i+1}=\mathrm{Conv}_{i+1}\left(\boldsymbol{X}_{i+1}+\boldsymbol{Y}_i\right)\tag{6}$$

其中,$\boldsymbol{Y}_{i+1}$为下一分支上的输出特征. 这一设计使MSEGC能够捕捉到远距离关节之间的关系. 例如,特征$\boldsymbol{Y}_i$包含关节点的两跳邻居信息,当它与$\boldsymbol{X}_{i+1}$融合后,通过图卷积操作聚合生成$\boldsymbol{Y}_{i+1}$,则$\boldsymbol{Y}_{i+1}$实际上可获得4跳邻居节点的特征信息.

最后,为了获得更具整体结构感知能力的空间表示,本文将4个分支上的输出特征沿通道维度拼接,构建一个全局的空间特征表示$\boldsymbol{Y}$:

$$\boldsymbol{Y}=\mathrm{act}\left(\left[\boldsymbol{Y}_1\|\boldsymbol{Y}_2\|\boldsymbol{Y}_3\|\boldsymbol{Y}_4\right]+\boldsymbol{X}\right)\tag{7}$$

其中,$\boldsymbol{X}$表示对经过多层图卷积处理后原始信息中丢失特征的补偿. 随着信息的逐级融合,MSEGC逐步实现了从局部到远距离关节依赖关系的联合建模,有效增强了边缘处关节之间以及边缘处关节与其他关节之间关联特征的表达能力.

MSEGC提供了一种级联式空间特征聚合机制,有效增强其对复杂空间结构的建模能力. 在完成空间特征提取后,MSEGC的输出特征$\boldsymbol{Y}$将会传递给多尺度时间增强卷积(MTEC)模块,以进行时间维度上的信息提取.

3.5　多尺度时间增强卷积模块

尽管空间建模对行为识别至关重要,但它仅能提供有限的特征信息. 为了应对行为识别领域的挑战,建模时间信息也是必不可少的. Alsarhan等人[23]与Ding等人[24]对这一方法进行了研究. 然而,在大多数方法中,时间依赖关系通常使用固定的kt × 1时间卷积直接建模,其表述为

$$\boldsymbol{Y}_{\mathrm{out}}=\boldsymbol{W}_{\mathrm{fix}}\boldsymbol{X}_{\mathrm{in}}\tag{8}$$

其中,$\boldsymbol{X}_{\mathrm{in}}\in\mathbf{R}^{C_{\mathrm{in}}\times T\times V\times M}$为输入特征,$\boldsymbol{Y}_{\mathrm{out}}\in\mathbf{R}^{C_{\mathrm{out}}\times T\times V\times M}$为输出特征,固定卷积的权重为$\boldsymbol{W}_{\mathrm{fix}}\in\mathbf{R}^{C_{\mathrm{in}}\times C_{\mathrm{out}}\times \mathrm{kt}\times 1}$. 这种建模方法使学习高级的时间特征变得困难,从而降低了描述不同行为的能力. 为了丰富时间维度上的特征,并捕获更广泛的时序依赖关系,提高相似行为的区分度,本文提出了多尺度时间增强卷积(MTEC)模块,其整体结构如图4③所示.

同样,为了在通道数不变的情况下捕获更加多样的运动模式特征,MTEC首先将输入特征$\boldsymbol{X}\in\mathbf{R}^{C\times T\times V\times M}$在通道维度上均匀划分为4个子集,并采用逐级结构将各子集$\boldsymbol{X}_i\in\mathbf{R}^{C'\times T\times \mathrm{V}\times M}$依次输入时间卷积模块. MTEC采用的时间卷积模块是MS-G3D[5]网络中时间卷积结构的简化版本(MS-TC),其具体结构如图4⑤所示. MTEC中的MS-TC通过精简分支数量和增大时间核尺寸,在降低参数计算量的同时有效提升了时间建模能力. 该模块由4个并联的卷积分支组成,每个分支采用不同的卷积核大小与膨胀率组合,从而拓展了MTEC在时间维度上的感知范围,增强了对多尺度时间特征的建模能力. 最终,MS-TC中各个分支的输出将在通道维

度上进行拼接,以形成融合性的输出特征 $\boldsymbol{Y}_i$:

$$\boldsymbol{Y}_i=\mathrm{act}\left(\left[\boldsymbol{Y}_{i1}\|\boldsymbol{Y}_{i2}\|\boldsymbol{Y}_{i3}\|\boldsymbol{Y}_{i4}\right]+\boldsymbol{X}\right) \tag{9}$$

其中,$\boldsymbol{X}$表示用于残差补偿的输入特征,用以保留原始信息.

为了增强时间上下文的连贯性,MTEC将当前阶段的输出特征 $\boldsymbol{Y}_i$与下一个子集$\boldsymbol{X}_{i+1}$进行融合,并将该融合特征进一步输入MS-TC模块,用于下一阶段时序建模. 该过程可表示为

$$\boldsymbol{Y}_{i+1}=\mathrm{Conv}_{i+1}\left(\boldsymbol{X}_{i+1}+\boldsymbol{Y}_i\right) \tag{10}$$

其中,$\boldsymbol{Y}_{i+1}$为下一阶段的输出特征. 这种设计能够进一步扩展MTEC在时间维度上的等效感受野,使其在覆盖更长时间序列的同时,能够整合跨帧的互补信息. 例如,特征 $\boldsymbol{Y}_i$表征以每5帧为单位的时序信息,当其与$\boldsymbol{X}_{i+1}$融合并经时间卷积操作生成$\boldsymbol{Y}_{i+1}$后,实际上$\boldsymbol{Y}_{i+1}$可感知超过5帧的时序特征,从而实现更长时间依赖的建模.

最后,为了融合各阶段多尺度的时序信息,MTEC将4个阶段上的输出特征沿通道维度拼接,形成全局时间特征表示$\boldsymbol{Y}$,该过程可表示为

$$\boldsymbol{Y}=\mathrm{act}\left(\left[\boldsymbol{Y}_1\|\boldsymbol{Y}_2\|\boldsymbol{Y}_3\|\boldsymbol{Y}_4\right]\right) \tag{11}$$

这种逐级增强时间建模机制强化了MTEC对时间特征的学习能力,使其能够有效捕捉到时间帧之间长跨度的特征,从而在建模动作持续性方面展现出更强的表现力.

通过多个阶段的信息交流与融合,MTEC能够捕获到更广泛的时序依赖关系,从而更加全面地捕捉整个行为的轨迹特性,增强模型对相似行为的判别能力,有效提升类别内的统一性和类别间的区分度.

4 实验

为了评估增强时空图卷积网络(EST-GCN)在基于骨架数据的行为识别任务中的有效性,本文在NTU-RGB+D[7]和NTU-RGB+D 120[8]两个公开数据集上,将其与基线模型EfficientGCN-B0[22]进行了直接比较. 本文通过消融实验验证各模块的有效性,并将EST-GCN与当前先进的基于骨架的行为识别方法进行性能对比.

4.1 数据集

NTU-RGB+D[7]数据集包含60个行为类别和56 880个样本. 这些样本涵盖了40个日常行为类别、9个与健康相关的行为类别和11个交互行为类别,由年龄在10~35岁之间的40名受试者执行. 该数据集是使用Microsoft Kinect v2传感器从3个不同的摄像头角度捕捉的,数据以3D骨骼信息的形式表示. 数据集被按照两种标准分为训练集和测试集. (1)跨主体(X-Sub):指定20名受试者用于训练,其余受试者用于测试. (2)跨视角(X-View):一个摄像头捕捉的数据用于训练,另外两个摄像头捕捉的数据用于测试.

NTU-RGB+D 120[8]是目前拥有3D关节注释的最大数据集. 它通过增加57 367个骨骼序列和60个额外的行为类别,扩展了NTU-RGB+D数据集,从而获得了由106名志愿者通过3个摄像头捕捉的113 945个样本,涵盖120个行为类别. 该数据集再次采用两种不同的方法分为训练集和测试集. (1)跨主体(X-Sub):训练数据来自53名志愿者的行为,测试数据来自其余志愿者的行为. (2)跨设置(X-Set):训练数据来自具有偶数ID的样本,测试数据来自具有奇数ID的样本.

4.2 实验设置

本文使用PyTorch框架在NVIDIA GeForce RTX 3090 GPU上进行了所有实验. 表1列出了实验环境参数. 批量大小均设置为16. 本文采用了随机梯度下降算法(Stochastic Gradient Descent,SGD)来进行优化,动量设为0.9,权重衰减设为0.000 1,初始学习率设为0.1,并在前10个周期中采用了预热策略,即逐渐将学习率从0增加到初始值,以确保训练过程的稳定性,之后开始按照余弦规律衰减. 本文选择交叉熵损失函数,并在全局平均池化层(Global Average Pooling,GAP)之后和最终的全连接层(Fully Connected Layer,FCL)之前添加了一个丢弃层,以防止过拟合,丢弃概率为0.25. 另外,在X-view基准测试的实验中,本文执行了一种特殊的数据变换[2]以实现视图对齐. 对于NTU-RGB+D数据集,迭代次数设为60. 对于NTU-RGB+D 120数据集,迭代次数设为70. 对于这两个数据集,每个样本中的最大帧数设为288帧. 如果样本帧数少于288帧,则重复该样本直到达到288帧.

表1 NTU-RGB+D和NTU-RGB+D 120数据集的实验参数设置

Hyperparameter	NTU-RGB+D	NTU-RGB+D 120
Batch size	16	16
Frames	288	288
Learning Rate(LR)	0.1	0.1
Weight decay	0.000 1	0.000 1
Epochs	60	70
Loss function	Cross entropy	Cross entropy
Gradient descent	SGD	SGD
Activation	Swish	Swish

4.3 消融实验

如图3所示,本文采用多流输入数据融合策略,由以下3种数据流组成:(1)关节位置流(Joint);(2)关节运动速度流(Velocity);(3)骨骼特征流(Bone). 最终的预测结果通过不同输入数据流的分数进行加权求和得出.

4.3.1　模块的有效性

为了验证EST-GCN模型中每个模块的关键作用和效能,本文分别引入其中一个模块来验证其有效性. 如表2所示,相较于基线模型EfficientGCN-B0[22],随机更换一个模块均能够提升模型性能. 当两个模块联合使用时,表现最为出色. 以下是得出的结论:(1)当引入MSEGC模块时,模型在NTU-RGB+D数据集的X-Sub与X-View基准上性能分别提升1.7个百分点和1.0个百分点,在NTU-RGB+D 120数据集的X-Sub与X-Set基准上性能分别提升1.7个百分点和4.3个百分点;(2)当引入MTEC模块时,模型在NTU-RGB+D数据集的X-Sub与X-View基准上性能分别提升1.4个百分点和1.0个百分点,在NTU-RGB+D 120数据集的X-Sub与X-Set基准上性能分别提升1.3个百分点和4.3个百分点;(3)当MSEGC和MTEC同时引入时,模型在NTU-RGB+D数据集的X-Sub与X-View基准上性能分别提升2.2个百分点和1.3个百分点,在NTU-RGB+D 120数据集的X-Sub与X-Set基准上性能分别提升2.1个百分点和5.0个百分点. 综上所述,两个模块在两个数据集上均展现出积极作用. 具体而言,在NTU-RGB+D数据集上,二者均在X-Sub基准上提升更为显著. 这是因为X-Sub按受试者划分,使两个模块能够更专注于同一受试者在不同视角下的骨架空间结构与时序动态特征建模. 相较之下,在NTU-RGB+D 120数据集上,两个模块则在X-Set基准上表现出更优性能. 这是由于NTU-RGB+D 120是NTU-RGB+D的扩展版本,包含更多相似行为实例,而X-Set按数据ID划分,具有更强的随机性与覆盖性,因此两个模块能够学习到更加全面和鲁棒的特征表征.

表2　在NTU-RGB+D[1]与NTU-RGB+D 120[2]数据集上引入不同模块后的识别准确率

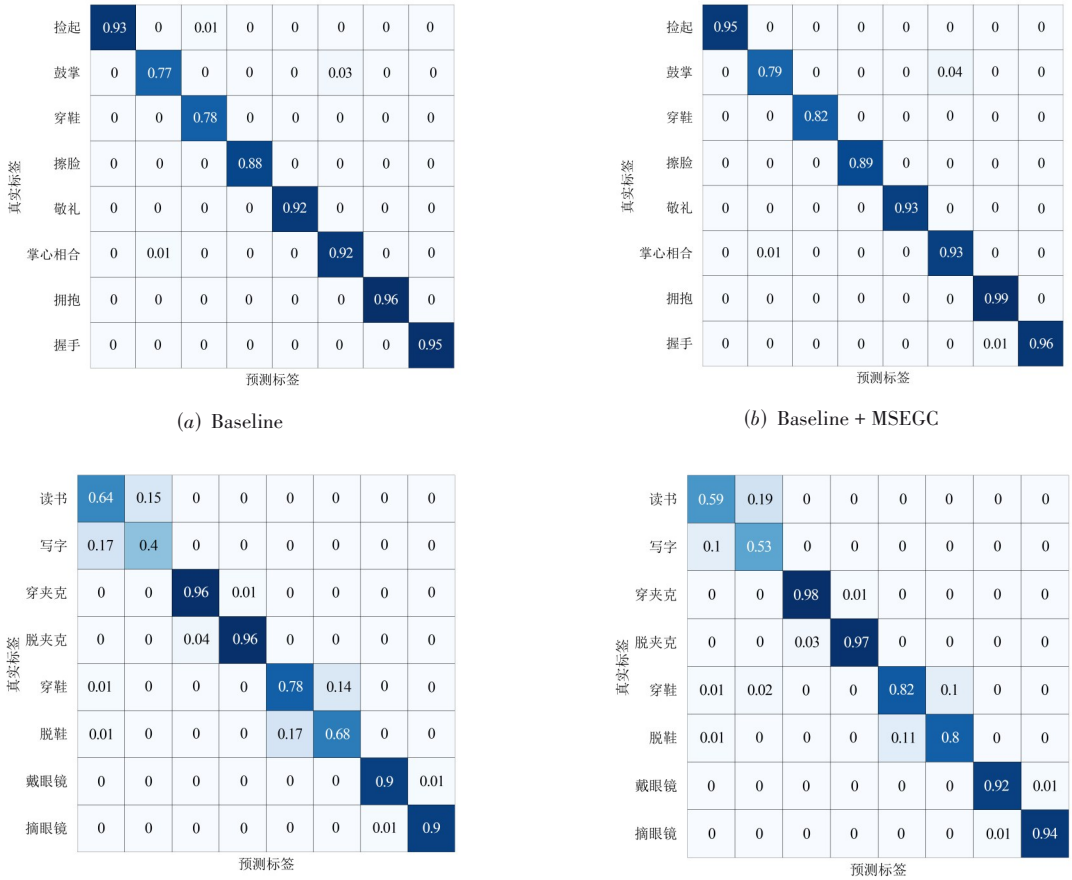
Method	X-Sub[1]/%	X-View[1]/%	X-Sub[2]/%	X-Set[2]/%
Baseline	90.2	94.9	86.6	85.0
Baseline+MSEGC	91.9	95.9	88.3	89.3
Baseline+MTEC	91.6	95.9	87.9	89.3
EST-GCN	92.4	96.2	88.7	90.0

为了更直观地展示各模块的有效性,图5给出关节位置输入数据流在NTU RGB+D数据集上识别结果的部分混淆矩阵. 在边缘处关节表达能力方面,如图5(a)和图5(b)所示,相较于基准模型,MSEGC模块在所示行为类别中识别准确率更高,表明其有效增强了边缘处关节的特征表达能力,能够更精准地捕捉边缘关节之间的关联特征. 而在时序建模能力方面,如图5(c)与图5(d)所示,MTEC模块在易混类别中提升了模型的区分能力. 这类行为在短时间内具有相似动作,但从更长时序跨度来看存在显著差异,因此捕捉更广泛的时序依赖关系至关重要. MTEC在这一点上发挥了重要作用,特别是在“穿鞋”与“脱鞋”相似动态动作中,有效降低了组内混淆度. 然而,对于“读书”和“写字”这组相似动作,随着时序跨度的扩大,写字被误判为读书的概率降低,而读书被误判为写字的概率则升高. 这是因为在短时序跨度下,模型难以捕捉写字过程中手部的周期性运动,容易将其误判为读书,而在较长跨度下,这类特征逐渐显现,从而提升了区分度;对于以静态手部姿态为主的读书动作而言,其骨架序列在长时段内与写字的静态片段高度相似,从而更易产生混淆. 综上所述,可以看出MSEGC模块在细粒度活动识别中发挥了重要作用,而MTEC模块则有效增强了模型对相似行为的判别能力,二者可共同促进整体识别性能的提升.

4.3.2　输入数据流的有效性

为验证所提出的方法在不同输入数据流上的有效性,本文基于NTU-RGB+D数据集的X-Sub基准进行了实验分析. 如表3所示,每个数据流都起到有效作用. 以下是得出的结论:(1)引入MSEGC时,模型在关节位置流和关节运动速度流的性能均提升1.28个百分点,分别从87.85%提升至89.13%和从86.81%提升至88.09%;(2)引入MTEC模块时,模型在关节位置流的性能提升1.09个百分点,从87.85%增长至88.94%;(3)尤为显著的是,EST-GCN在关节运动速度流上性能显著提升2.24个百分点,从86.81%提升至89.05%. 这一结果表明,EST-GCN在处理关节运动速度流信息时具有更强的表征能力. 相较于基准,每个输入数据流的性能均有提升,这充分证明所提出方法在各个输入数据流上的有效性,展示了其在多模态行为识别中的广泛适用性和优越性.

进一步深入分析发现,任意两个不同输入数据流的结合效果各异.(1)引入MSEGC时,模型在以关节位置流与关节运动速度流共同作为输入时,效果最佳,性能提升1.65个百分点. 这表明这两类数据流在捕捉空间和时间信息方面具有较强的互补性,能够共同提升模型的表征能力.(2)引入MTEC时,模型在关节位置流与骨骼特征流的协同作用下得到1.41个百分点的性能提升,这一结果表明,骨骼特征与关节位置信息相结合,能够更好地捕捉人体结构的动态变化,提高行为识别的准确性.(3)在EST-GCN框架下,关节位置流与骨骼特征流的组合实现了2.06个百分点的性能提升. 这进一步验证了骨骼特征与关节位置信息之间存在显著的协同作用. 这些结果不仅揭示了不同输入特征组合在提升模型性能方面的重要作用,还指出了特征间潜在的互补性和协同效应.



(*a*) Baseline (*b*) Baseline + MSEGC

(*c*) Baseline (*d*) Baseline + MTEC

图5 混淆矩阵对比图

表3 不同输入数据流在NTU-RGB+D的X-Sub基准上的识别准确率

Inputs	Baseline/%	Baseline+MSEGC/%	Baseline+MTEC/%	EST-GCN/%
Joint	87.85	89.13	88.94	89.70
Velocity	86.81	88.09	87.65	89.05
Bone	88.54	89.09	89.19	89.92
Joint+Velocity	89.62	91.27	90.87	91.65
Joint+Bone	89.12	90.74	90.53	91.18
Velocity+Bone	89.83	91.20	91.08	91.76
Joint+Velocity+Bone	90.20	91.92	91.59	92.35

表4 NTU-RGB+D数据集上的比较

Method	Publisher	X-Sub/%	X-View/%
ST-GCN[1]	AAAI18	81.5	88.3
2s-SGCN[2]	CVPR19	88.5	95.1
Shift-GCN(2s)[25]	CVPR20	89.7	96.0
MS-G3D Net[5]	CVPR20	91.5	96.2
SGN[26]	CVPR20	89.0	94.5
ST-TR[27]	CVIU21	89.9	96.1
MST-GCN(2s)[28]	AAAI21	91.1	96.4
FGCN[15]	TIP21	90.2	96.3
MS&TA-HGCN-FC[29]	TCSVT22	90.8	96.4
ML-STGNet[21]	TIP22	91.9	96.2
SMotif-GCN+TBs[30]	TPAMI23	90.5	96.1
EfficientGCN-B0[22]	TPAMI23	90.2	94.9
MADT-GCN(2s)[18]	EAAI24	89.9	96.1
LMA-GCN[31]	EAAI24	88.2	95.0
MTGCN[32]	IJON25	90.5	95.2
EST-GCN(2s)		91.8	96.0
EST-GCN(3s)		92.4	96.2

4.4 比较与分析

为验证EST-GCN的有效性,本文在NTU-RGB+D和NTU RGB+D 120数据集上将其与同领域先进方法进行比较,结果如表4和表5所示,其中“2s”指“Bone”与“Velocity”两流结合,“3s”指本文所有数据流结合.

可以看出,EST-GCN在所有数据流结合下性能最佳. 以下是得出的结论.(1)对于NTU-RGB+D数据集,如表4所示,EST-GCN在X-Sub基准上实现92.4%的准

表5　NTU-RGB+D 120数据集上的比较

Method	Publisher	X-Sub/%	X-Set/%
ST-GCN[1]	AAAI18	70.7	73.2
2s-SGCN[2]	CVPR19	82.5	84.2
Shift-GCN(4s)[25]	CVPR20	85.9	97.6
MS-G3D Net[5]	CVPR20	86.9	88.4
MST-GCN(4s)[28]	AAAI21	87.5	88.8
ST-TR[27]	CVIU21	82.7	84.7
FGCN[15]	TIP21	85.4	87.4
MS&TA-HGCN-FC[29]	TCSVT22	87.0	88.4
ML-STGNet[21]	TIP22	88.6	90.0
InfoGCN(2s)[33]	CVPR22	88.5	89.7
SMotif-GCN+TBs[30]	TPAMI23	87.1	87.7
EfficientGCN-B0[22]	TPAMI23	86.6	85.0
MADT-GCN(4s)[18]	EAAI24	86.5	88.2
LMA-GCN[31]	EAAI24	83.2	84.1
MTGCN[32]	IJON25	80.8	81.7
EST-GCN(2s)		87.9	89.6
EST-GCN(3s)		88.7	90.0

确率,在X-View基准上的准确率达到96.2%. 与FGCN[15]、MST-GCN(2s)[28]、MS&TA-HGCN-F[29]这3个方法相比,EST-GCN在X-View基准上的准确率仅低了0.1%和0.2%. (2)对于规模更大、更具挑战性的NTU-RGB+D 120数据集,表5中结果表明,EST-GCN在X-Sub基准上实现了88.7%的准确率,在X-Set基准上则达到了显著的90.0%的准确率.

通过深入分析,本文发现NTU-RGB+D 120数据集作为NTU-RGB+D的扩展,包含更多相似行为的实例. 例如,戴帽子、脱帽子、戴耳机和摘耳机等行为在特定的时间序列中表现出高度相似的肢体运动,因此模型需要捕捉更长时间跨度内的信息以实现准确区分. EST-GCN方法能够有效捕捉时间帧之间更广泛的时序依赖关系,从而降低因时间分辨率不足而产生的干扰信息. 此外,与其他划分基准不同,X-Set根据样本的奇偶性进行数据集划分,这种方式不受实验人员影响,具备更高的随机性和公正性. 因此,在NTU-RGB+D 120数据集的X-Set基准上,EST-GCN的识别性能得到显著提升. 这表明EST-GCN在处理大规模、复杂行为识别任务方面具有明显的优越性,尤其是在面对具有相似动作模式的数据集时,展现出更强的鲁棒性和更高的准确性.

根据表4和表5中的比较结果,可以得出以下结论:在所有数据集上,本文所提出的EST-GCN模型已经接近或达到了现有方法的性能水平,证明了该方法的优越性. 为更直观地体现模型的识别性能,图6给出了类别识别效果示例. 同时,本文进一步考察了多人交互过程中骨架点可能出现错乱分配的情况. 具体而言,本

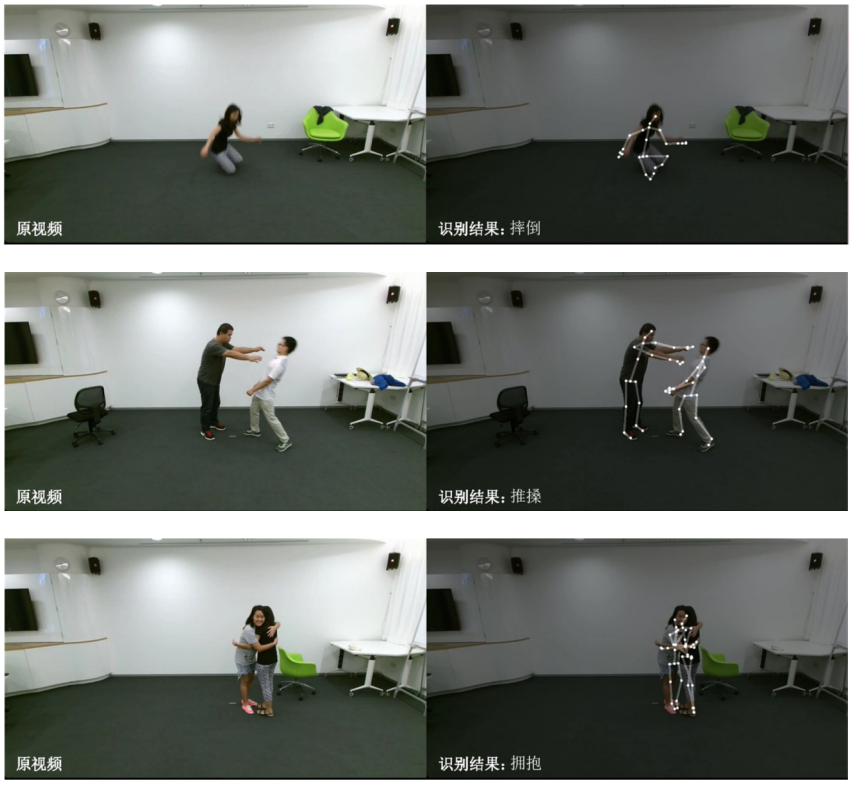


图6　类别识别效果示例

文对所用数据集的原始视频采用OpenPose技术重新提取骨架数据,并在此基础上进行识别实验. 如图7所示,以“推搡”和“拥抱”行为作为示例,尽管生成的骨架序列中存在关节分配错误,但模型最终仍能实现准确分类. 这表明,在识别过程中局部且短时的骨架错乱并不足以对整体判别造成实质影响. 这一结果验证了本文模型在面对局部噪声和跨人关节错乱时的鲁棒性.

(*a*) 推搡:关节错乱分配示例(上)、识别结果(下)

(*b*) 拥抱:关节错误分配示例(上)、识别结果(下)

图7 类别识别效果示例(关节错乱情形下)

5 结论

本文提出了一种包含MSEGC和MTEC两个模块的基于骨架行为识别任务的EST-GCN模型,该网络模型可以提取丰富的空间特征和多尺度时间特征. 本文首先通过MSEGC模块捕捉空间维度上关节之间的特征. MSEGC通过多分支学习并逐级传递边缘处关节之间及其与其他关节间连接关系的融合特征,有效捕捉边缘处关节间以及边缘处关节与人体自然拓扑结构中其他关节之间的协同作用关系. 然后,MSEGC模块的输出结果传至MTEC模块,MTEC则通过捕获与融合多个时间感受野下的时间特征,并通过逐级增强的时间信息传递机制有效扩展了时间跨度,捕获到时间帧之间更广泛的时序依赖关系. 在NTU-RGB+D和NTU-RGB+D 120两个广泛使用的骨架数据集上的实验结果证明了EST-GCN模型的有效性. 在未来的工作中,本文将进一步挖掘骨架拓扑结构中关节之间潜在的依赖关系,并研究如何在提升模型性能的同时简化网络框架和降低计算成本.

参考文献

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

[2] SHI L, ZHANG Y F, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 12018-12027.

[3] 赵俊男, 佘青山, 孟明, 等. 基于多流空间注意力图卷积SRU网络的骨架动作识别[J]. 电子学报, 2022, 50(7): 1579-1585.
ZHAO J N, SHE Q S, MENG M, et al. Skeleton action recognition based on multi-stream spatial attention graph convolutional SRU network[J]. Acta Electronica Sinica, 2022, 50(7): 1579-1585. (in Chinese)

[4] LI C, ZHONG Q Y, XIE D, et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[EB/OL]. (2018-04-17)[2025-05-25]. https://arXiv.org/abs/1804.06055.

[5] LIU Z Y, ZHANG H W, CHEN Z H, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 143-152.

[6] LI M S, CHEN S H, CHEN X, et al. Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3316-3333.

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

[8] 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, 2020, 42(10): 2684-2701.

[9] ZHUO J M, CUI C, FU K, et al. Propagation is all you need: A new framework for representation learning and classifier training on graphs[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 481-489.

[10] MONDAL A, SHASHANT R, GIRALDO J H, et al. Moving object detection for event-based vision using graph spectral clustering[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops. Piscataway: IEEE, 2021: 876-884.

[11] YING R, YOU J X, MORRIS C, et al. Hierarchical graph representation learning with differentiable pooling[C]// Advances in Neural Information Processing Systems 31. San Diego: NeurIPS, 2018: 4800-4810.

[12] HU J F, ZHENG W S, LAI J H, et al. Jointly learning heterogeneous features for RGB-D activity recognition[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 5344-5352.

[13] ZHANG S Y, LIU X M, XIAO J. On geometric features for skeleton-based action recognition using multilayer LSTM networks[C]//2017 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2017: 148-157.

[14] SOO K, REITER A. Interpretable 3D human action analysis with temporal convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2017: 1623-1631.

[15] YANG H, YAN D, ZHANG L, et al. Feedback graph convolutional network for skeleton-based action recognition[J]. IEEE Transactions on Image Processing, 2022, 31: 164-175.

[16] LEE J, LEE M, LEE D, et al. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 10410-10419.

[17] 罗会兰, 曹立京. 基于多维动态拓扑学习图卷积的骨架动作识别[J]. 电子学报, 2024, 52(3): 991-1001.
LUO H L, CAO L J. Multi-dimensional dynamic topology learning graph convolution for skeleton-based action recognition[J]. Acta Electronica Sinica, 2024, 52(3): 991-1001. (in Chinese)

[18] XIA Y, GAO Q Y, WU W G, et al. Skeleton-based action recognition based on multidimensional adaptive dynamic temporal graph convolutional network[J]. Engineering Applications of Artificial Intelligence, 2024, 127: 107210.

[19] WU Z Z, SUN P P, CHEN X, et al. SelfGCN: Graph convolution network with self-attention for skeleton-based action recognition[J]. IEEE Transactions on Image Processing, 2024, 33: 4391-4403.

[20] XIE J Y, MENG Y D, ZHAO Y T, et al. Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition[J]. Proceedings of the

AAAI Conference on Artificial Intelligence, 2024, 38(6): 6225-6233.
[21] ZHU Y S, SHUAI H, LIU G C, et al. Multilevel spatial-temporal excited graph network for skeleton-based action recognition[J]. IEEE Transactions on Image Processing, 2023, 32: 496-508.
[22] 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, 2023, 45(2): 1474-1488.
[23] ALSARHAN T, ALI U, LU H T. Enhanced discriminative graph convolutional network with adaptive temporal modelling for skeleton-based action recognition[J]. Computer Vision and Image Understanding, 2022, 216: 103348.
[24] DING C Y, WEN S, DING W W, et al. Temporal segment graph convolutional networks for skeleton-based action recognition[J]. Engineering Applications of Artificial Intelligence, 2022, 110: 104675.
[25] CHENG K, ZHANG Y F, HE X Y, et al. Skeleton-based action recognition with shift graph convolutional network[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 183-192.
[26] ZHANG P F, LAN C L, ZENG W J, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 1112-1121.
[27] PLIZZARI C, CANNICI M, MATTEUCCI M. Skeleton-based action recognition via spatial and temporal transformer networks[J]. Computer Vision and Image Understanding, 2021, 208/209: 103219.
[28] CHEN Z, LI S C, YANG B, et al. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1113-1122.
[29] HUANG Z X, QIN Y S, LIN X B, et al. Motion-driven spatial and temporal adaptive high-resolution graph convolutional networks for skeleton-based action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(4): 1868-1883.
[30] WEN Y H, GAO L, FU H B, et al. Motif-GCNs with local and non-local temporal blocks for skeleton-based action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2009-2023.
[31] JIANG Y J, DENG H M. Lighter and faster: A multi-scale adaptive graph convolutional network for skeleton-based action recognition[J]. Engineering Applications of Artificial Intelligence, 2024, 132: 107957.
[32] CHEN H, SHEN Y H, ZHANG Y X, et al. Skeleton-based action recognition through dual-granularity feature fusion with self-adapting graph convolution and multi-scale temporal convolution[J]. Neurocomputing, 2025, 639: 130261.
[33] CHI H G, HA M H, CHI S, et al. InfoGCN: Representation learning for human skeleton-based action recognition[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 20154-20164.

作者简介

姜 维 男,1981年出生于河南省郑州市. 现为华北水利水电大学信息工程学院副教授. 研究方向为计算机视觉与机器学习.
E-mail: jiangwei@ncwu.edu.cn

关孟怡 女,1999年出生于河南省漯河市. 现为华北水利水电大学信息工程学院硕士研究生. 研究方向为计算机视觉.
E-mail: Z20231090840@stu.ncwu.edu.cn