面向骨架行为识别的多语义动态图卷积网络

宋 忱，钱惠敏†，吴大伟

(河海大学 人工智能与自动化学院，江苏 常州 213000)

摘 要: 近年来, 图卷积网络在人体骨架行为识别领域展现出卓越性能. 针对现有基于图卷积的方法存在节点复杂相关性建模的局限, 以及模态间互补信息利用不足的问题, 提出一种多语义动态图卷积网络 (MSD-GCN). 该网络为关节-骨骼融合双流架构, 并行处理关节与骨骼模态数据. 双流网络由多个多语义动态图卷积算子 (MSD-GC)、多尺度时间卷积算子 (MS-TC) 和关节-骨骼跨模态对比学习模块 (JB-CMCL) 组成. 具体而言, MSD-GC 算子通过语义感知分层图 (SH-Graph) 重构高语义粒度分区, 并行执行跨语义空间建模模块 (CSSM) 捕获全局关节相关性, 以及局部几何建模模块 (LGM) 捕捉细微运动特征, 实现多尺度动态特征提取. JB-CMCL 则通过跨模态特征对齐和混淆样本辨别机制, 引导双流网络中关节与骨骼模态的特征融合与增强, 提升模型细粒度识别能力. 在 NTU RGB+D、NTU RGB+D 120 和 Northwestern-UCLA 数据集进行广泛的实验. 结果表明, 所提出的组件与整体网络具有极强的性能, 能够较好地识别混淆动作. 与最先进的方法相比, 该模型具有极强的竞争力.

关键词: 人体骨架；行为识别；图卷积神经网络；注意力机制；对比学习机制；双流网络；多路集成

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0793

引用格式: 宋忱, 钱惠敏, 吴大伟. 面向骨架行为识别的多语义动态图卷积网络 [J]. 控制与决策.

Multi-semantic dynamic graph convolutional networks for skeleton-based action recognition

SONG Chen，QIAN Hui-min†，WU Da-wei

(College of Artificial Intelligence and Automation，Hohai University，Changzhou 213000，China)

Abstract: In recent years, graph convolutional networks have exhibited outstanding performance in the field of skeleton-based action recognition. Nevertheless, existing GCN-based methods suffer from limitations in modeling complex node correlations and insufficient utilization of complementary information between modalities. To address these issues, this paper proposes a multi-semantic dynamic graph convolutional network (MSD-GCN). This network adopts a joint-bone fused dual-stream architecture, processing joint and bone modality data in parallel. The dual-stream network consists of multiple multi-semantic dynamic graph convolution (MSD-GC) operators, multiple multi-scale temporal convolution (MS-TC) operators, and a joint-bone cross-modal contrastive learning (JB-CMCL) module. Specifically, the MSD-GC operator reconstructs high semantic granularity partitions through a semantic-aware hierarchical graph (SH-Graph) and executes in parallel a cross-semantic space modeling (CSSM) module to capture global joint correlations and a local geometry modeling (LGM) module to capture subtle motion features. The JB-CMCL module guides feature fusion and enhancement between joint and bone modalities within the dual-stream network through cross-modal feature alignment and hard sample discrimination mechanisms, thereby improving the model's fine-grained recognition capability. Extensive experiments were conducted on NTU RGB + D, NTU RGB + D 120, and northwestern-UCLA datasets. The results demonstrate that the proposed components and the overall network exhibit superior performance, effectively recognizing ambiguous actions. Compared with state-of-the-art methods, our model shows strong competitiveness.

Keywords: human skeleton； action recognition； graph convolutional network； attention mechanism； contrastive learning mechanism； dual-stream network； multi-way ensemble

收稿日期: 2025-07-29；**录用日期:** 2025-11-20.

基金项目: 中国航空科学基金项目 (2024M034108001).

责任编委: 贲晛烨.

†通信作者. E-mail: am_hohai@163.com.

0 引 言

人体行为识别, 即计算机接受视频等数据实现人类行为的识别与理解. 该技术可用于视频监控系统识别异常行为[1-2], 如预防犯罪或检测司机危险驾驶; 用于虚拟现实系统[3] 中实现情景理解、人机交互等功能.

近年来, 得益于 LiDAR、Kinect、多普勒雷达等传感器的发展, 人体行为识别技术已从传统的视觉数据拓展到深度、骨架、点云、雷达等[4-5] 多模态数据. 其中, 基于骨架数据的方法因其对背景环境、光照不均匀、人体重叠等干扰具有更强的鲁棒性[6], 受到计算机视觉研究人员的广泛关注与青睐. 这类方法通常使用 OpenPose 工具箱[7] 预测视频中的二维关节坐标或者使用 Kinect V2 深度传感器[8-9] 采集三维关节坐标. 不同于传统视觉数据, 骨架数据在非欧几里得空间内表征为具有时间信息的关节坐标序列. 为有效建模骨架数据的时空特征, 研究人员将人体骨架抽象为图形结构, 图中节点表示人体关节, 边缘表征骨骼连接关系. 卷积神经网络 (Convolutional Neural Network, CNN) 无法将卷积核直接应用到图结构, 选择将骨架数据处理成伪图像或者视频式体积序列[10], 导致空间关系混乱, 破坏骨架的图结构信息. 循环神经网络 (Recurrent Neural Network, RNN) 及其变体将骨架数据强制排列成有序坐标序列[8, 11], 同样未利用骨架的先验知识, 因其串行处理机制忽略了关节间的并行协同关系.

这些限制促使研究者们寻找自然匹配骨架图本质的建模工具, 从而推动图卷积网络 (Graph Convolutional Network, GCN) 的应用, 该模型能够基于任意图结构实现高效的信息传递. Yan 等[12] 根据人体动力学知识, 首次将 GCN 应用到面向骨架的人体行为识别中, 提出 ST-GCN. 该方法沿着人体的物理连通性聚合来自关节子集合的空间信息, 结合时间卷积获取人体关节之间的时间相关性. 相较于先前 CNN 和 RNN 的方法, ST-GCN 取得令人印象深刻的性能, 但是它依赖手工设计的拓扑图, 导致节点只能在指定分区内共享信息, 限制了模型捕捉远距离节点间相关性的能力.

为突破固定物理拓扑的局限性, 国内外研究主要聚焦于使用注意力或其他方式自适应捕获关节的潜在相关性. 例如, Ye 等[13] 提出动态图卷积, 为每一维度所有关节的上下文特征学习一个动态拓扑图, 融合动态图与静态拓扑图的特征, 获得全局相关性表达. Chen 等[14] 提出通道拓扑细化图卷积, 通过学习节点相关性来实现通道级动态拓扑的建模. Chi 等[15] 使用自注意力机制推断捕捉动作行为的关节关系的内在拓扑, 增强静态拓扑图信息. Wang 等[16] 使用由可分离参数图卷积算子处理后的高维特征生成新的拓扑图, 再次送入可分离参数图卷积算子获得优质边缘特征. Liu 等[17] 提出一种多尺度图卷积, 使每个节点仅与其最小 K 跳数邻居以及自身相关联, 消除不同邻域之间的冗余依赖性. 这些方法虽在一定程度突破固定拓扑的限制, 但仍然继承 ST-GCN 的拓扑图设计, 不仅在节点间复杂相关性的建模能力上存在不足, 也未能够有效平衡节点全局关系依赖建模与局部特征保持. 此外, 现有方法多依赖单一骨架模态进行特征提取, 仅在决策层进行浅层融合, 未能充分利用和挖掘骨架模态之间的互补信息. 尤其是在处理手部等复杂动作, 需要同时捕获局部关节的细微运动和肢体的大范围协调, 这些局限更加明显.

为解决这些问题, 本文提出了一种新颖的多语义动态图卷积网络 (Multi-Semantic Dynamic Graph Convolutional Networks, MSD-GCN) 模型, 主要贡献包括如下:

1) 本文提出了多语义动态图卷积算子 (MSD-GC). MSD-GC 结合语义感知分层图 (SH-Graph), 在单个语义空间内, 使用跨语义空间建模模块 (CSSM) 结合自注意力机制, 数据驱动地学习任意两个关节之间的初始关联强度, 并通过可导的拓扑精炼函数进一步聚焦于最相关的连接, 从而突破物理拓扑限制以捕获节点间复杂相关性. 自注意力机制倾向于平滑化细节, 可能弱化局部关节间微小的相对运动, 局部几何建模模块 (LGM) 借用 EdgeConv[18], 几何驱动地捕捉局部上下文特征, 以保持细微的运动特征.

2) 本文提出了一种关节-骨骼跨模态对比学习模块 (JB-CMCL), 结合关节-骨骼融合双流网络架构, 同时处理关节和骨骼输入特征, 能够实现关节模态和骨骼模态特征的早期特征融合与模态间互补增强. JB-CMCL 模块在共享特征空间中显式建模关节、骨骼模态特征一致性, 利用混淆样本辨别机制, 促使同类动作的相互对齐, 同时分离异类样本的潜在表征, 从而深度挖掘并利用模态间的互补信息.

3) 为了验证 MSD-GCN 的有效性和识别精度, 本文在NTU-RGB+D, NTU-RGB+D 120 和NW-UCLA 流行动作识别数据集进行广泛的实验. 实验结果表明, 所提出的 MSD-GCN 以较低的计算代价在这 3 个数据集上实现了强大的性能.

1 面向骨架行为识别的多语义动态图卷积网络

本节详细阐述了提出的面向骨架行为识别的多语义动态图卷积网络 MSD-GCN, 结构如图 1(a) 所示. 在 1.1 节中, 提供了基于 GCN 的骨架行为识别相关的预备知识. 在 1.2 节中, 介绍了多语义动态图卷积算子 (Multi-Semantic Dynamic Graph Convolution, MSD-GC), 它结合语义感知分层图 (Semantic-Aware Hierarchical Graph, SH-Graph), 并行使用跨语义空间建模模块 (Cross-Semantic Spatial Modeling, CSSM) 和局部几何建模模块 (Local Geometric Modeling, LGM), 建模全局特征和局部几何特征, 如图 1(b) 所示. 在 1.3 节中, 描述了在双流网络中, 用于细化特征表示的关节-骨骼跨模态对比学习模块 (Joint-Bone Cross-Modal Contrastive Learning, JB-CMCL), 如图 1(c) 所示. 在 1.4 节中, 总结了完整的识别网络.

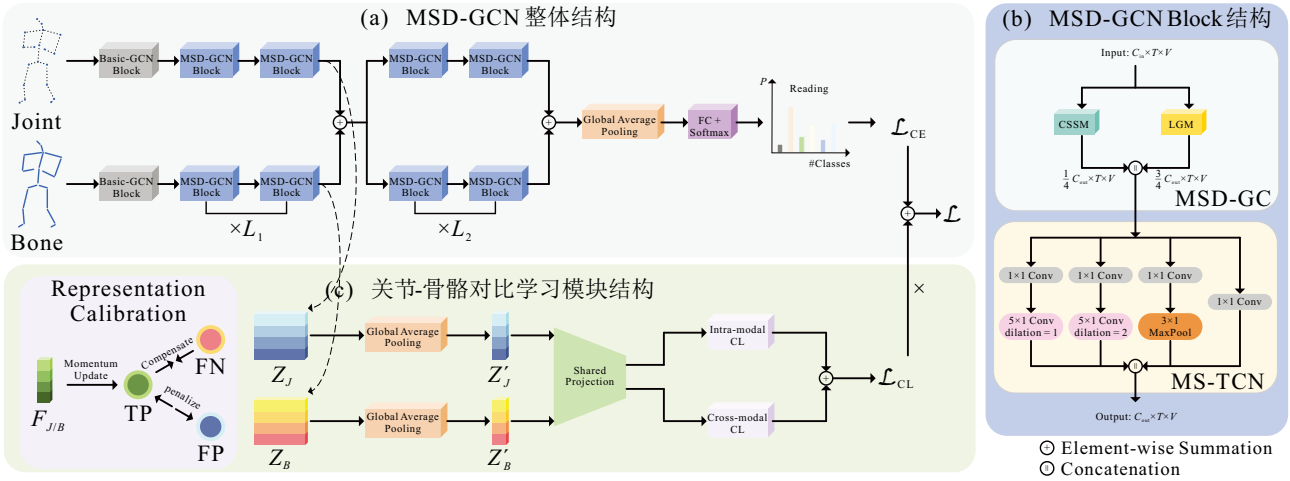


图1　面向骨架行为识别的多语义动态图卷积网络结构.

1.1 预备知识

人体骨架序列拓扑图可以被建模为$\mathcal{G}=\{\mathcal{V},\mathcal{E}\}$, 其中$\mathcal{V}=\{V_{ti}|t=1,\ldots,T;i=1,\ldots,N\}$表示$T$帧骨架序列中关节点的集合, 且每帧包含$N$个关节点. $\mathcal{E}$表征$\mathcal{V}$中关节点之间的连接边缘集合. 单帧内, 由人体关节的物理连通确定关节点之间的连接关系, 并采用邻接矩阵$A\in\mathbb{R}^{N\times N}$表示. 若关节点$i$和关节点$j$存在连接, $A_{ij}=1$; 否则, $A_{ij}=0$. 相邻帧间, 相同关节点之间存在连接.

图卷积算子聚合邻居关节点的特征更新自身关节点特征. 为了提升模型的空间建模能力, ST-GCN[12] 将邻接矩阵A的节点分为三个子集: 根节点子集s_{id}; 向心节点子集s_{cp}; 离心节点子集s_{cf}. 给定输入图卷积算子的人体骨架序列为$X\in\mathbb{R}^{T\times N\times C_{in}}$, 输出序列为$Y\in\mathbb{R}^{T\times N\times C_{out}}$, 则在第$t$帧上的空间图卷积定义为:

$$Y_t=\sum_{s=1}^{S}\bar{A}_s X_t W_s. \tag{1}$$

式中$\bar{A}_s=D_s^{-\frac{1}{2}}A_sD_s^{\frac{1}{2}}\in\mathbb{R}^{N\times N}$表征第$i$个子集的归一化邻接矩阵, D_s是A_s的度矩阵; $W_s\in\mathbb{R}^{C_{in}\times C_{out}}$是$X_t$中第$s$个分区的特征变换权重矩阵; S为邻接矩阵A的子集个数.

1.2 多语义动态图卷积算子

在面向骨架的行为识别研究中, 模型捕捉关节间复杂相互作用是提升性能的关键. 本文提出的 SH-Graph 扩大了图卷积的感受野, 为模型提供了捕获远程依赖的可能. 在此基础上, MSD-GC 算子通过 CSSM 模块优化信息聚合路径, 并借助 LGM 模块捕获邻近关节间细微运动模式, 以实现关节复杂相关性的多尺度表达.

1.2.1 语义感知分层图

ST-GCN 提出的典型骨架拓扑结构图是根据人体自然结构预先定义的, 结构缺乏灵活性, 难以捕获远距离依赖关系. 为此, 本文提出 SH-Graph, 通过层次化分解骨架关节点构建语义空间, 将关节点划分为脊椎、躯干和四肢三个高语义粒度分区.

以 NTU-RGB+D 系列数据集为例, 按照表 1 的策略, 将人体骨架图的 25 个关节点划分为 4 层. 例如, 以胸膛 (节点 21) 作为起源节点时, 其余 24 个关节点分别被划入脊椎层、躯干层和四肢层. 并基于此, 在相邻层级间构建 3 个语义空间: 脊椎语义空间、躯干语义空间和四肢语义空间. 当起源节点分别为臀部 (节点 1)、腹部 (节点 2) 时, 可得到不同的脊椎层、躯干层和四肢层, 生成不同的语义空间. 为了定义语义感知分层邻接矩阵, 在第l个语义空间中,

定义其较低层级$l-1$的节点子集为源关节点子集o^{l-1}, 其较高层级l的节点子集为目标关节点子集t^l, 并且$o^{l+1}=t^l$(即构成第$l+1$个语义空间的源关节点子集是第l个语义空间的目标关节点子集). 例如, 在脊椎语义空间中, 起源层的关节点构成源关节点子集o^0, 脊椎层的所有关节点则构成目标关节点子集t^1.

表1 语义感知分层图分层策略

起源层 l=0	脊椎层 l=1	躯干层 l=2	四肢层 l=3
1	2, 13, 14, 17, 18, 21	3, 5, 9, 15, 19	4, 6 - 8, 10 - 12, 16, 20, 22 - 25
2	1, 3, 5, 9, 13, 17, 21	4, 6, 10, 14, 18	7, 8, 11, 12, 15, 16, 19, 20, 22 - 25
21	1 - 6, 9, 10	7, 8, 11 - 14, 17, 18	15, 16, 19, 20, 22 - 25

定义语义感知分层邻接矩阵$A_{SH}\in\mathbb{R}^{3\times3\times N\times N}$的3个图子集为$s_{id}^l$、$s_{cf}^l$和$s_{cp}^l$, 且:

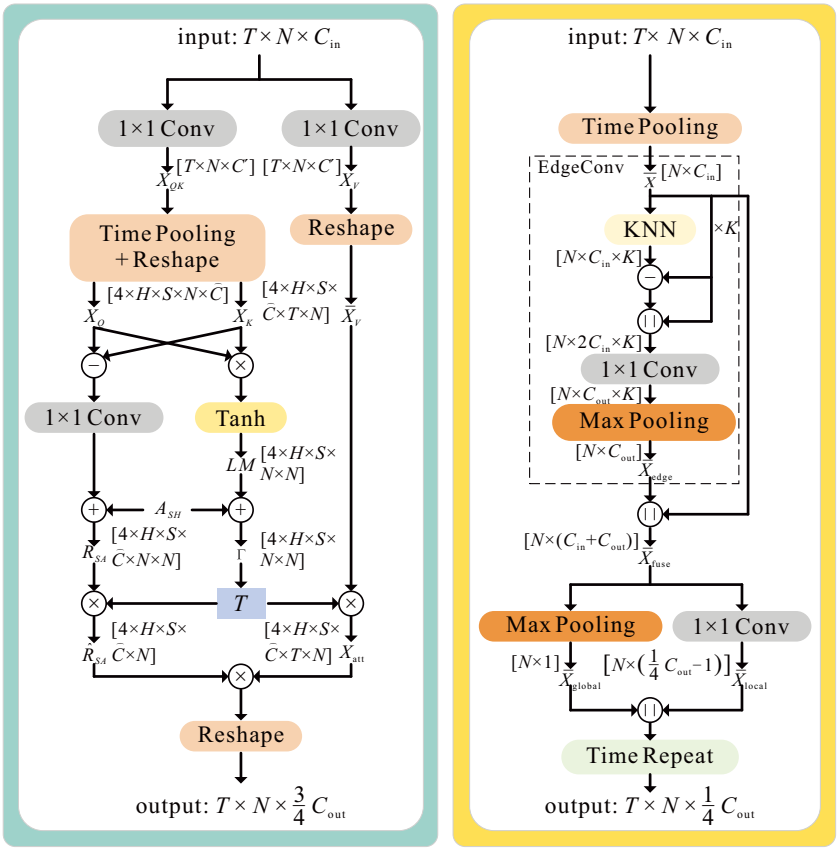
$$\begin{aligned} s_{id}^l &= o^{l-1}\cup t^l,\\ s_{cp}^l &= o^{l-1}\longleftarrow t^l,\ l=1,2,3\\ s_{cf}^l &= o^{l-1}\longrightarrow t^l, \end{aligned}\tag{2}$$

其中, $o^{l-1}\cup t^l$表示o^{l-1}与t^l的关节点之间存在连接; $o^{l-1}\longleftarrow t^l$表示$t^l$的关节点沿着向心方向全连接$o^{l-1}$内所有关节点; $o^{l-1}\longrightarrow t^l$代表$o^{l-1}$的关节点沿着离心方向全连接$t^l$内所有关节点.

语义感知分层邻接矩阵A_{SH}包含了人体自然连接的真实边缘和全连接创造的远距离虚拟边缘. 由于虚拟边缘突破物理骨架的约束, 直接建立远距离节点间的关联, 从而实现了比传统方法更大的感受野. 在MSD-GC算子中, A_{SH}作为可学习的共享拓扑结构, 能帮助模型从输入序列中稳健地提取特征.

1.2.2 跨语义空间建模模块

CTR-GCN模型[14]提出通道拓扑细化图卷积算子, 通过独立处理每个通道, 在特征空间中建模输入特征. 具体而言, 该算子使用相关系数建模函数$\mathcal{M}(\cdot)$学习不同类型的运动特征, 同时仅利用简单的特征提取函数$\mathcal{T}(\cdot)$实现浅层输入特征到深层特征的映射. 这种设计存在明显局限性: (1) 共享的邻接矩阵因其固定拓扑结构限制了图卷积的感受野, 使得空间特征主要集中于人体自然连接的相邻节点, 潜在的远程依赖关系主要依赖$\mathcal{M}$获得; (2) 简单的特征提取方式难以充分挖掘数据中的语义信息. 为突破此局限, CSSM采用多头自注意力机制, 捕获关节之间的全局相关性. 结构如图2(a)所示. 对于输入特征$X\in\mathbb{R}^{T\times N\times C_{in}}$, 使用2个逐点卷积层$W_{QK}$和$W_V\in\mathbb{R}^{C_{in}\times C'}$, 用来生成对应的查询-键向量



(a) CSSM (b) LGM

图2 CSSM和LGM模块结构

X_{QK}和值向量$X_V \in \mathbb{R}^{T\times N\times C'}$:

$$X_{QK} = XW_{QK}, X_V = XW_V. \tag{3}$$

为了降低计算代价, 本文采用分组卷积来生成查询-键向量X_{QK}和值向量X_V, 其中分组数设置为4. 随后, 对查询-键向量X_{QK}使用时间平均池化来聚合全局时间特征. 接着, 通过维度重塑操作将X_{QK}重新划分为查询向量X_Q和键向量$X_K \in \mathbb{R}^{4\times H\times S\times N\times \widehat{C}}$. 其中, H为多头注意力数, S为邻接矩阵的图子集数量. 自注意力机制分别学习 SH-Graph 的 3 个语义空间内关节的局部自注意力图$LM \in \mathbb{R}^{4\times H\times S\times N\times N}$, 定义如下:

$$LM = Tanh(X_q X_k^{'}). \tag{4}$$

公式 (4) 中$'$表示矩阵转置. 将反映语义空间内的关节相关性的LM与参数化的共享拓扑A_{SH}结合, 得到跨语义空间相似性$\Gamma \in \mathbb{R}^{4\times H\times S\times N\times N}$:

$$\Gamma = A_{SH} + \alpha \cdot LM. \tag{5}$$

其中, α是可学习参数, 用于自适应地调整语义空间相似性强度. 这种数据驱动的方法建模空间拓扑, 能够灵活地表达关节之间的全局相关性.

为了进一步优化这种全局相关性, 需要筛选出与Γ的中心关节γ_i相似度最高的k个邻居关节. 并且, 为了使Γ参与模型的梯度下降算法, 本文并没有采用传统的离散选择函数, 设计了一种连续的拓扑精炼函数$T(\cdot)$, 以保证梯度的连续性:

$$T(\Gamma_{ij}) = softmax(\Gamma_{ij} + \delta \cdot M_{ij}). \tag{6}$$

其中, δ是一个超参数常值, $M \in \mathbb{R}^{4\times H\times S\times N\times N}$为基于相似度的邻居索引掩码矩阵. 具体的来说, 对于γ_i的k个最似邻居关节 $\gamma_{j1}, \ldots, \gamma_{jk}$, $M_{ij} = 1$, 其他关节则为 0.

精炼后的Γ的概率分布与经过特征通道调整的值向量$\bar{X}_V \in \mathbb{R}^{4\times H\times S\times \widehat{C}\times T\times N}$相乘, 得到多头自注意力细化后的空间特征$X_{att} \in \mathbb{R}^{4\times H\times S\times \widehat{C}\times T\times N}$:

$$X_{att} = \sum_{j=1}^{N} \bar{X}_V^{(j)} T(\Gamma)^{(j)}. \tag{7}$$

与特征提取函数$\mathcal{T}(\cdot)$相比, CSSM 基于自注意力机制与拓扑精炼函数$T(\cdot)$实现了更全面的空间特征表征. CTR-GCN 模型认为不同通道编码的运动特征存在差异, 导致关节之间的相关性呈现动态变化. 强制所有通道共享一个先验拓扑结构会限制图卷积网络的特征提取灵活性, 反而忽略了通道间的关联. 本文保留相关系数建模函数$\mathcal{M}(\cdot)$, 获取关节点之间的通道相关性. 按照该思想, 结合参数化的共享拓扑A_{SH}, 得到语义空间级的拓扑关系$R_{SA} \in \mathbb{R}^{4\times H\times S\times \widehat{C}\times N\times N}$:

$$R_{SA} = A_{SH} + \beta \cdot \mathcal{M}(X_Q, X_K) = A_{SH} + \beta \cdot (\sigma(X_Q - X_K)). \tag{8}$$

其中, σ是一个线性变化函数, 用于调整特征通道和改变维度, β是一个可学习参数. 进一步的, 为了增强不同语义空间之间的相互关系, 将跨语义空间的多头注意力结果应用于拓扑关系R_{SA}, 得到$\hat{R}_{SA} \in \mathbb{R}^{4\times H\times S\times \widehat{C}\times N}$:

$$\hat{R}_{SA} = \sum_{j=1}^{N} T(\Gamma)^{(j)} {R_{SA}}^{(j)}. \tag{9}$$

最后, CSSM 重构典型图卷积公式 (1), 在空间图卷积中求和所有语义特征空间的特征, 得到跨语义空间建模结果$Z_{CSSM} \in \mathbb{R}^{\frac{3}{4}C_{out}\times T\times N}$, 如下所示:

$$Z_{CSSM} = Reshape(\sum_{l=1}^{H} \hat{R}_{SA}^{(l)} X_{att}^{(l)}). \tag{10}$$

其中, $Reshape$表示维度调整操作, 并且$\frac{3}{4}C_{out} = 4 \times S \times \widehat{C}$.

1.2.3 局部几何建模模块

尽管在 CSSM 里, 通过多头注意力机制计算全局相似性, 捕获了关节的语义关系, 但是可能忽略了局部几何结构. 例如, 在“打响指”, “数钱”等动作中, 拇指、指尖、手腕的局部运动具有几何相关性, 但是多头注意力可能忽略这种局部依赖性. 为了解决这个问题, 本文设计了 LGM, 结构如图 2(b) 所示, 用于获取关节的局部几何相关性.

EdgeConv 在点云分割和分类任务中, 能动态捕捉中心节点与其邻居节点之间的上下文信息. 在高维特征空间中, EdgeConv 能为中心节点i, 基于特征欧氏距离确定n个最近邻节点. 对于节点特征x_i和$x_{j_n} \in \mathbb{R}^{1\times C_{in}}$, EdgeConv 操作可以被表示为:

$$EdgeConv(x_i) = \underset{j_n\in\mathcal{N}(i)}{Maxpool}[(x_i \parallel x_i - x_{j_n})W_E]. \tag{11}$$

式中, $Maxpool(\cdot)$为最大池化操作, $\mathcal{N}(i)$表示中心节点的邻居节点集合$\{j_1, j_2, \ldots, j_n\}$, $\parallel$表示按照通道维度连接特征, 逐点卷积层$W_E \in \mathbb{R}^{2C_{in}\times C_{in}}$用于降低结果的特征维度.

首先, LGM 对输入特征X使用平均时间池化, 以获得反映输入的长期趋势特征$\bar{X} \in \mathbb{R}^{N\times C_{in}}$. 通过$EdgeConv$算子, 获得深层边缘特征$\bar{X}_{Edge} \in \mathbb{R}^{N\times C_{out}}$:

$$\bar{X}_{Edge} = EdgeConv(\bar{X}). \tag{12}$$

随后, 为了保留浅层特征空间中的局部几何特

征和深层特征空间中的语义特征, 按照通道维度拼接浅层特征$\bar{X}$和深层边缘特征$\bar{X}_{Edge}$, 得到融合特征$\bar{X}_{fuse} \in \mathbb{R}^{N\times(C_{in}+C_{out})}$. 简单地解耦$\bar{X}_{fuse}$为聚焦于动作的整体趋势的全局分量$\bar{X}_{global} \in \mathbb{R}^{N\times 1}$和捕捉细微几何关系的局部分量$\bar{X}_{local} \in \mathbb{R}^{N\times(\frac{1}{4}C_{out}-1)}$. 最后, 为了维护时间维度, 将$\bar{X}_{global}$和$\bar{X}_{local}$沿时间维度复制$T$次. LGM 的最终输出$Z_{LGM} \in \mathbb{R}^{\frac{1}{4}C_{out}\times T\times N}$:

$$Z_{LGM} = \underset{T}{Repeat}[MaxPool(\bar{X}_{Edge}) \parallel \bar{X}_{Edge}W_{local}]. \tag{13}$$

式中$Repeat(\cdot)$将特征沿时间维度复制T次, 逐点卷积层$W_{local} \in \mathbb{R}^{(C_{in}+C_{out})\times(\frac{1}{4}C_{out}-1)}$用于调整通道特征.

通过拼接 CSSM 和 LGM 的输出结果, 调整通道顺序后, 得到了多语义空间动态图卷积网络的输出$Z_{MSD} \in \mathbb{R}^{T\times N\times C_{out}}$.

1.3 关节-骨骼跨模态对比学习模块

骨骼和关节模态是骨架数据的两种常见表示, 两者特征本质相关但表达视角各有不同. Shi 等[19]首次引入了关节模态$V_{joint} = \{v_{t1}, \ldots, v_{tn}\}$的二阶信息, 即骨骼模态$V_{bone} = \{v_{ti} - v_{tj} \mid (i,j) \in V_b\}$, 其中$V_b$是根据人体关节连通性设计的标签对集合. 针对行为识别任务中关节与骨骼模态的互补性建模问题, 本文提出了 JB-CMCL, 用于引导关节与骨骼特征的高层语义对齐, 获得准确的特征表示, 促进 MSD-GCN 网络的早期特征融合.

JB-CMCL 参考 FR-Head[20]提出的细粒度的样本分类机制, 引入高置信度 (TP) 样本、假阳性 (FP) 样本和假阴性 (FN) 样本的三元辨别机制. 具体而言, 一个真实标签或者预测标签为m的动作样本特征F_{in}^m会被分为三类: (1) 模型预测标签与真实标签均为m则视为高置信度正样本特征F_{TP}^m; (2) 模型预测为m而真实标签非m则归为假阳性样本特征F_{FP}^m; (3) 真实标签为m但模型预测为非m则判定为假阴性样本特征F_{FN}^m; 这种三元辨别机制通过精准区分不同置信度的样本, 为对比学习提供更精确的监督信号.

在实践中, JB-CMCL 分别对关节流、骨骼流的早期特征Z_J、$Z_B \in \mathbb{R}^{T\times N\times C_{out}}$使用平均池化层, 在时间维度$T$、节点维度$N$上压缩成一维向量$Z_J'$、$Z_B' \in \mathbb{R}^{C_{out}}$. 以关节模态为例, 首先将骨骼和关节一维向量通过统一投影头映射到共享空间, 将Z_J'和Z_B'转化为高级特征$F_J, F_B \in \mathbb{R}^d$, 作为后续对比机制的输入. 随后采用动量更新策略, 随机初始化和更新动作m的骨骼模态原型P_B^m:

$$P_B^m = \pi \cdot P_B^m + (1-\pi)\cdot \bar{F}_{TP}^m. \tag{14}$$

式中π是动量项, 经验上设置成 0.9; $\bar{F}_{TP}^m$是F_B中所有 TP 特征的平均值.

为了修正模型的漏检和误检, JB-CMCL 鼓励所有的 FN 样本靠近同类的 TP 样本, 并使所有 FP 样本远离异类的 TP 样本. 为此, 引入动作类别m的奖励项ϕ_B^m和惩罚项φ_B^m:

$$\begin{cases} \phi_B^m = 1 - \cos(F_{in}^m, \bar{F}_{FN}^m) \\ \varphi_B^m = 1 + \cos(F_{in}^m, \bar{F}_{FP}^m) \end{cases} \tag{15}$$

式中$\bar{F}_{FN}^m$、$\bar{F}_{FP}^m$分别为F_B中所有 FN/FP 特征的平均值, $\cos(\cdot,\cdot)$为余弦相似性函数.

最终, 关节模态下的对比损失函数为:

$$\mathcal{L}_{CL}(F_J) = -\log \frac{e^{\cos(F_J,P_B^m)/\tau-(1-p^m)\phi_B^m}}{e^{\cos(F_J,P_B^m)/\tau-(1-p^m)\phi_B^m} + \sum\limits_{l\neq m} e^{\cos(F_J,P_B^l)/\tau}} - \log \frac{e^{\cos(F_J,P_B^m)/\tau-(1-p^m)\varphi_B^m}}{e^{\cos(F_J,P_B^m)/\tau-(1-p^m)\varphi_B^m} + \sum\limits_{l\neq m} e^{\cos(F_J,P_B^l)/\tau}} \tag{16}$$

式中, p^m为动作m的预测置信度, $(1-p^m)$项为低置信度样本赋予更大的惩罚或奖励权重, 从而促使模型关注那些模糊或难以判别的样本, 提升整体识别性能. τ为温度系数.

MSD-GCN 的完整损失函数为:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda(\mathcal{L}_{CL}(F_J) + \mathcal{L}_{CL}(F_B)). \tag{17}$$

其中, $\mathcal{L}_{CE}$是行为识别网络的交叉熵损失, λ是对比学习损失的平衡参数.

1.4 网络结构

MSD-GCN 网络采用关节-骨骼融合的双流架构设计, 分别以关节特征和骨骼特征作为 2 个单流网络的输入. 具体而言, 单流网络中堆叠 1 个基本图卷积 Basic-GCN 块和 9 个 MSD-GCN 块, 它们的输出通道数分别为 64、64、64、64、128、128、128、256、256、256, 形成层次化的特征提取网络. Basic-GCN 块依据公式 (1) 实现空间特征提取, MSD-GCN 块除利用 1.2 节所描述的 MSD-GC 算子增强模型的空间特征表达能力, 还使用 CTR-GCN 的多尺度时间卷积算子 (Multi-Scale Temporal Convolution, MS-TC) 聚合多帧中的时间特征. MS-TC 由 4 个包含逐点卷积的分支网络组成, 前 2 个分支包含的卷积核大小为 5, dilation 分别为 1 和 2, 第三个分支额外包含卷积核大小为 3 的最大池化. 在经过第 5 个和第 8 个 MSD-GCN 块后, 时间维度减半. 这两个模块都添加

了残差连接以稳定训练, 为了方便起见, 在图 1(b) 中省略这一部分.

在网络训练过程中, 原始输入特征经过 1 个 Basic-GCN 块和L_1个 MSD-GCN 块后, 使用一组可学习参数对双流网络的中间特征进行加权逐元素求和, 实现早期特征融合. 其融合结果再次经过L_2个 MSD-GCN 基本块, 逐元素求和得到高维融合模态表示. 最后, 经过全局平局池化将不同样本特征图压缩到同一尺寸, 使用全连接层输出动作分类的分数, 求和后得到最后的交叉熵损失$\mathcal{L}_{CE}$. 对比学习损失$\mathcal{L}_{CL}$与$\mathcal{L}_{CE}$求和后, 通过 softmax 函数对动作样本进行分类. 在测试阶段, 本文并未遵循最近的研究[14, 20]使用的关节、骨骼、关节速度、骨骼速度四种模态, 而是提出了一种简单且新颖的三路集成方法. 该方法同时使用关节和骨骼模态, 分别以胸膛、腹部、臀部三个关节作为语义感知分层图起源节点, 独立训练关节-骨骼融合双流网络. 由于不同的起源节点, 语义感知分层图表达不同的骨骼语义空间, 模型能够学习同一样本的不同特征. 在推理阶段, 分别从每一路网络末端的全连接层提取输出结果作为分类分数. 三路网络的分类分数以一组权重参数进行加权求和. 以产生最终的识别结果.

2　实验与结果分析

2.1　数据集与实验环境

为了验证本文所提出的行为识别方法的有效性和性能, 在 NTU-RGB+D[8]、NTU-RGB+D 120[9] 和 Northwestern-UCLA[21] 进行了广泛的实验.

NTU-RGB+D 是一个用于行为识别的大规模数据集, 由 40 位志愿者制作的 56 880 个骨架序列组成. 每个骨架序列由 3 个 Kinect v2 相机提供 25 个关节点的 3 维坐标, 所有序列被分为 60 个动作类别. 该数据集设计跨对象 (Cross-subject, X-sub) 和跨视角 (Cross-view, X-view) 两个基准评估识别精度. 其中 X-sub 将 20 位志愿者的动作作为训练集, 其余 20 位志愿者的动作用于测试集. 而 X-view 则将相机 1 收集的数据用于测试集, 相机 2 和 3 收集的样本为训练集.

NTU-RGB+D 120 是 NTU-RGB+D 60 的扩展, 额外添加 60 个动作类别, 由 106 位志愿者执行 113 945 个样本组成. 与 NTU-RGB+D 不同的是, 该数据集重新设计了跨对象 (Cross-subject, X-sub) 和跨设置 (Cross-set, X-set) 两种基准. 具体而言, X-sub 将 53 位志愿者的 63 026 个动作样本用于训练集, 其余 53 位的 50 919 个动作样本用于验证集. X-set 将偶数设置 ID 的 54 468 个动作样本用于训练集, 奇数设置 ID 的 59 477 个动作样本用于测试集.

Northwestern-UCLA 由 3 个 Kinect 相机捕获 10 位演员执行的 10 个动作类别的 1 494 个骨骼序列组成. 训练集使用前两个相机视图, 测试集使用后另一个相机视图.

本文在具有 PyTorch 深度学习框架的单个 RTX 3090Ti GPU 上进行实验. 训练中, 使用 CTR-GCN 的数据预处理方法, 采用混合精度训练加速计算并减少显存占用, 使用动量为 0.9 的随机梯度下降 (SGD) 优化器更新参数, 通过余弦退火衰减算法调整学习率, 在前 5 个迭代周期使用预热方法提升模型收敛性能. 对于 NTU-RGB+D 系列数据集, 具体设置为: 批量大小为 64, 初始学习率为 0.1, 权重衰减为 0.000 4, 迭代次数为 80, 在第 35、55 和 75 轮时学习率衰减 0.1 倍. 对于 Northwestern-UCLA 数据集, 则设置: 批量大小为 16, 初始学习率为 0.05, 权重衰减为 0.000 5, 迭代次数为 65, 在第 30、50 轮时学习率衰减 0.1 倍.

最后, 本文给出 MSD-GCN 模型训练中的超参数的设置情况. CSSM 模块的多头注意力数H为 3, 拓扑精炼函数T中最近邻数k为 8, 掩码强度系数δ为 10. LGM 模块的 EdgeConv 邻居数n设置为 3. JB-CMCL 模块中按照经验将温度系数τ设为 0.125, 预测置信度p为 0.125. MSD-GCN 损失函数中, 平衡参数λ设置为 0.01. MSD-GCN 基本块数量L_1和L_2分别为 2 和 7.

2.2　消融实验与分析

在本节中, 本文探讨 MSD-GCN 模型各个组件的有效性以及相应的不同设计方法, 同时分析每个模块的配置及其组合的有效性. 选择 CTR-GCN 的关节模态作为基准模型, 实验均是在 NTU RGB+D 120 的 X-Sub 基准下进行性能比较.

本文对 MSD-GCN 各组件的贡献进行了消融实验, 并且从准确率、参数量、计算复杂度进行分析. 结果如表 2 所示, 可以看出随着组件的增加, 性能逐渐提升. 首先, 仅使用 CSSM 代替基线的图卷积模块, 在仅消耗基线 0.6 倍参数和计算复杂度的情况下, 性能略微提升 0.1%, 这表明 CSSM 通过多头注意力机制捕获全局依赖关系, 并且显著提高了模型效率. 随后, 并行添加了 LGM, 以较低的复杂度和 0.25M 参数量, 性能从 84.90% 提升到 85.47%. 由于 LGM 几何驱动捕捉关节局部特征, 协同 CSSM 实现局部几何特征与全局语义特征的互补. 接着, 引入关节-骨骼双流网络, 以基线 1.5 倍参数量和 1.1 倍计算复杂

度, 性能提升了 4.0%, 充分表明双流架构通过融合关节和骨骼信息, 显著提升了模型的识别能力. 最后, 在双流网络中添加 JB-JCML 进一步优化特征表示, 在保持参数量和计算量不变的情况下, 准确率再度提升 0.15%.

表2 组件有效性的实验结果

方法	参数量	FLOPs	X-Sub(%)
基准模型	1.44M	1.79G	84.90
CSSM	0.80M	0.99G	85.04
CSSM+LGM	1.05M	1.01G	85.47
CSSM+LGM+Two-Stream	2.11M	2.02G	88.93
CSSM+LGM+JB-JMCL	2.11M	2.02G	89.08

本文继续探讨了公式 (6) 超参数k和公式 (11) 超参数n对模型性能的影响, 结果如图 3 所示. 首先分析 CSSM 模块的超参数k, 当$k=8$时模型取得 85.04% 的最优性能, 此时模型能选择有效局部信息以提升性能. k值过小无法充分建模远程依赖, 而k值过大, 则引入不相关节点干扰, 使准确率下降. 随后, 分析 LGM 模块中 EdgeConv 算子的邻居节点数n提取局部几何特征的影响, n=3 时模型以 85.47% 的准确率表现最佳, 表明适度扩展局部邻域能有效捕捉细微关节运动模式, 而n过大则因过多的邻居节点引入冗余几何噪声, 导致性能震荡下降.

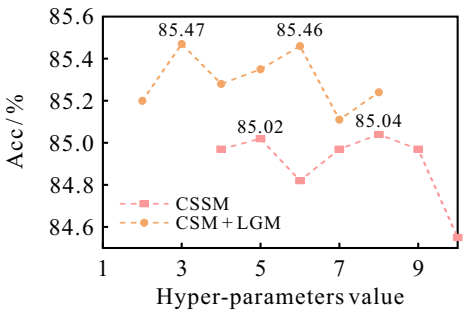


图3 超参数k和n对识别精度的影响

本文还研究了公式 (6) 拓扑精炼函数的掩码强度系数δ的影响, 结果如表 3 所示. 当δ从 0 增加到 10 时, 模型性能显著提升. δ=10 时出现了最佳性能 85.04%, 此时能有效强化重要连接的权重. 相反, δ继续增加到 15 时, 性能开始下降, 这表明过度强化局部连接, 会削弱模型对整体结构的感知, 反而破坏全局特征平衡.

表3 不同掩码强度系数δ的实验结果

δ	0	5	10	15	20
X-Sub(%)	84.04	84.68	85.04	84.92	84.30

表 4 详细考察了三种对比学习方法在不同起源节点下的的表现差异, 分别为使用双模态特征的跨模态对比学习, 仅使用单模态特征的模态内对比学习和结合跨模态与模态内对比学习机制的混合方法. 结果表明, 跨模态对比策略在三种起源节点配置下均取得最优性能, 相比无对比学习方法提升 0.2%-0.4%. 值得注意的是, 模态内对比和混合对比策略不仅未能带来预期增益, 反而导致模型性能下降 0.2%-0.4%. 这可能因为模态内对比过度强化单模态特征的一致性, 导致模型忽略关节与骨骼模态间的互补信息, 削弱了多模态融合的优势, 而混合对比策略中可能是因为跨模态和模态内目标存在优化方向冲突, 使得模型难以收敛到最优解.

表4 不同对比学习方法的实验结果

方法	X-Sub(%)		
	胸膛	腹部	臀部
无对比学习	88.93	88.82	88.51
模态内对比学习	88.67	88.71	88.42
模态间对比学习	89.08	88.92	88.90
混合对比学习	88.72	88.61	88.69

表 5 验证了对比学习损失与交叉熵损失的协同关系, 并表明对比学习平衡参数λ对模型识别性能有显著影响. 当λ为 0.01 时, 模型取得最高准确率 89.08%, 这表明适当的对比学习监督信号能够有效引导模型在共享特征空间中对齐关节与骨骼模态信息, 为交叉熵分类任务学习出更紧凑的判别性特征. λ过小, 对比学习损失权重不足, 难以充分引导跨模态特征对齐; 随着λ增大, 模型会过度强化对比学习目标, 干扰分类网络的主要任务学习, 导致模型性能逐渐降低.

表5 对比学习平衡参数λ的实验结果

λ	0.001	0.01	0.02	0.05	0.08	0.1	1
X-Sub(%)	88.92	89.08	88.90	88.84	88.57	88.57	86.65

本文分析了几组在基线上准确率均小于 70% 的极易混淆动作, 并使用 t-SNE 可视化它们在特征空间中的分布, 如图 4 所示, 左列为基线结果, 右列为本文结果. 红色、绿色、蓝色、黄色、紫色分别表示阅读、写作、玩手机/平板电脑、竖起大拇指和玩魔方动作, 下方青色、橙色、粉色则表示制作确定标志、制作胜利标志和数钱动作. 这些动作因为涉及手部和手指精细操作, 并且幅度和姿态相近, 导致基线模型在特征空间中对这些动作的表示出现混淆和重叠. 相反, MSD-GCN 在特征空间中实现了更加明显的聚类, 并且扩大了相似动作的特征距离. 这证实了本文有效地提取了全局和局部特征表示, 并基于捕获

的详细运动特征区分混淆动作.

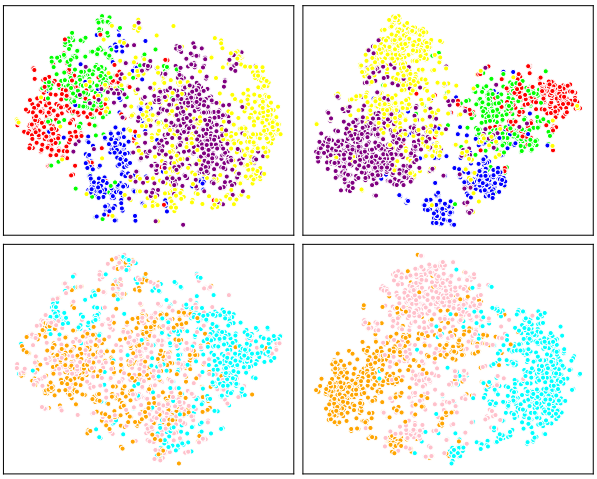


图4　t-SNE 可视化结果

最后, 本文探索了关节-骨骼融合双流网络中早期融合阶段的 MSD-GCN 基本块数L_1对模型性能的影响, 结果如表 6 所示. 实验结果表明当L_1=2 时模型取得 89.08% 的最高准确率, 这说明在第二个 MSD-GCN 块后进行跨模态特征融合能够最有效地平衡单模态特征学习与多模态信息交互. 本文认为过早融合会限制单模态特征的充分提取, 而过晚融合则会导致模态间协同效应减弱. 实验验证了在双流架构中需要在模态特征充分提取和跨模态有效交互之间找到最佳平衡点.

2.3　与先进方法对比

为了进行公平评估, 本文仅比较四种模态的结果, 对于 InfoGCN[15] 和 HD-GCN[22] 引用他们的四流集成结果. 本文将所提出的 MSD-GCN 在 NTU RGB+D、NTU RGB+D 120 和 Northwestern-UCLA 数据集上与多个基于 GCN 的最新工作进行比较, 具体的比较结果如表 7 所示, 用黑体标注最好的结果, 斜体为次好的结果. 以臀部作为起源节点的单路 MSD-GCN 网络展现强大性能. 与流行的主干网络[12] 相比, 相同参数量情况下, 在两个数据集上分别提升约 10% 和 20%. 使用臀部和胸部的双路集成 MSD-GCN 网络, 融合了上肢协调和下肢驱动两种互补的运动模式. 以 70% 的参数量在三个数据集上全面超过基准模型 CTR-GCN, 在 NTU RGB+D 120 上领先 InfoGCN 0.5%. 对于额外结合腹部的三路集成 MSD-GCN 网络结果, 识别性能再次提升. 与基于相对距离编码且最先进的 BlockGCN[23] 相比, 本文方法在 NTU RGB+D 60 X-View 和 NTU RGB+D 120 X-Set 基准上略微高出 0.1%, 在其他基准上则略低 0.1%. 综上所述, 随着集成路数增加, MSD-GCN 的识别性能也逐步增强. 这证明不同节点构建的语义感知分层图能够引导模型捕获差异性的特征, 各路结果都提供了互补性信息, 共同提升了模型的最终性能. 这些实验结果充分验证了本文方法的先进性和有效性.

表6　不同 MSD-GCN 基本块数量的实验结果

L_1	1	2	3	4	5	6	7	8
X-Sub(%)	89.05	89.08	88.61	88.85	88.95	88.90	88.28	88.06

表7　本文方法与其他先进方法的对比结果

方法	出版社	参数量	NTU RGB+D		NTU RGB+D 120		NW-UCLA
			X-Sub(%)	X-View(%)	X-Sub(%)	X-Set(%)	
ST-GCN[12]	AAAI2018	2.1M	81.5	88.3	70.7	73.2	—
2s-AGCN[19] (2-ensemble)	CVPR2019	7.6M	88.5	95.1	82.5	84.2	—
MS-G3D[17](2-ensemble)	CVPR2020	5.6M	91.5	96.2	86.9	88.4	—
MST-GCN[24](4-ensemble)	AAAI2021	12.0M	91.5	96.6	86.5	88.4	—
CTR-GCN[14](4-ensemble)	ICCV2021	6.0M	92.4	96.8	88.9	90.6	96.5
InfoGCN[15](4-ensemble)	CVPR2022	6.4M	92.7	96.9	89.4	90.7	96.6
FR-Head[20](4-ensemble)	CVPR2023	8.0M	92.8	96.8	89.5	90.9	96.8
HD-GCN[22](4-ensemble)	ICCV2023	6.7M	*93.0*	97.0	89.8	91.2	**96.9**
BlockGCN[23](4-ensemble)	CVPR2024	5.2M	**93.1**	97.0	**90.3**	91.5	**96.9**
MSD-GCN	—	2.1M	92.2	96.4	89.1	90.4	95.5
MSD-GCN(2-ensemble)	—	4.2M	92.7	96.9	89.9	91.2	96.1
MSD-GCN(3-ensemble)	—	6.3M	*93.0*	**97.1**	*90.1*	**91.6**	*96.8*

3　结　论

在这篇研究中, 本文提出面向骨架行为识别的多语义动态图卷积网络, 一个简单且性能强大的行为识别网络, 整合了图卷积、时间卷积和对比学习. 首先, 定义语义感知分层图, 在精简分区数量的同时显著扩展单分区语义覆盖范围, 为全局关系建模提供高效拓扑基础. 其次, 提出了多语义动态图卷积算子, 它并行结合跨语义空间建模方法与局部几何建模方法利用多头注意力机制捕获语义空间内关节的全局依赖, 结合可导拓扑精炼函数聚焦最相关连接,

并且动态捕捉局部关节的几何上下文特征. 此外, 提出了关节-骨骼跨模态对比学习机制, 通过显式建模关节与骨骼模态的语义一致性并且引入混淆样本辨别机制, 增强了模型对难分类样本的判别能力. 最后, 将上述模块与进行时序建模的多尺度时间卷积结合, 构成了多语义动态图卷积基本块, 用于提取骨架序列时空特征. 通过在三个大型骨架数据集上的实验结果和全面分析表明, MSD-GCN 在多个指标接近了最先进的方法, 其余指标则达到了最先进精度. 实验表明, MSD-GCN 在处理细粒度动作识别时表现尤其出色. 本文方法在模态融合研究仍显不足, 仅围绕基于对比学习的关节、骨骼模态融合方法, 并未探索更广泛的特征融合机制和异构模态的潜力, 限制了模型在复杂场景下的判别能力. 未来工作将致力于研究更高效的跨模态融合机制, 以及探索如何将提供动力学信息的关节速度模态、骨骼速度模态与本文现有模型相结合, 构建更全面的行为理解网络. 另一方面, RGB 视觉信息蕴含人与外界交互的上下文信息, 这一互补特性对区分高度混淆动作至关重要. 本文将进一步研究骨架与 RGB 信息的深度融合, 以提升模型在复杂场景下的鲁棒性.

参考文献 (References)

[1] 朱红蕾, 卫鹏娟, 徐志刚. 基于骨架的人体异常行为识别与检测研究进展[J]. 控制与决策, 2024, 39(8): 2484-2501.
(Zhu H L, Wei P J, Xu Z G. Research progress on skeleton-based human abnormal behavior recognition and detection[J]. Control and Decision, 2024, 39(8): 2484-2501.)

[2] 张晓平, 纪佳慧, 王力, 等. 基于视频的人体异常行为识别与检测方法综述[J]. 控制与决策, 2022, 37(1): 14-27.
(Zhang X P, Ji J H, Wang L, et al. Overview of video based human abnormal behavior recognition and detection methods[J]. Control and Decision, 2022, 37(1): 14-27.)

[3] Dallel M, Havard V, Dupuis Y, et al. Digital twin of an industrial workstation: A novel method of an auto-labeled data generator using virtual reality for human action recognition in the context of human–robot collaboration[J]. Engineering Applications of Artificial Intelligence, 2023, 118: 105655.

[4] Sun Z H, Ke Q H, Rahmani H, et al. Human action recognition from various data modalities: A review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3200-3225.

[5] 刘光辉, 王秦蒙, 孟月波, 等. 特征引导的多模态聚合低光环境行为识别方法[J]. 控制与决策, 2024, 39(7): 2305-2314.
(Liu G H, Wang Q M, Meng Y B, et al. Night behavior recognition based on multi-mode feature fusion[J]. Control and Decision, 2024, 39(7): 2305-2314.)

[6] 孔玮, 刘云, 李辉, 等. 基于图卷积网络的行为识别方法综述[J]. 控制与决策, 2021, 36(7): 1537-1546.
(Kong W, Liu Y, Li H, et al. A survey of action recognition methods based on graph convolutional network[J]. Control and Decision, 2021, 36(7): 1537-1546.)

[7] Cao Z, Hidalgo G, Simon T, et al. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 172-186.

[8] Shahroudy A, Liu J, Ng T T, et al. NTU RGB + D: A large scale dataset for 3D human activity analysis[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 1010-1019.

[9] 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.

[10] Hou Y H, Li Z Y, Wang P C, et al. Skeleton optical spectra-based action recognition using convolutional neural networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 28(3): 807-811.

[11] Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015: 1110-1118.

[12] 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.

[13] Ye F F, Pu S L, Zhong Q Y, et al. Dynamic GCN: Context-enriched topology learning for skeleton-based action recognition[C]. Proceedings of the 28th ACM International Conference on Multimedia. Seattle, 2020: 55-63.

[14] Chen Y X, Zhang Z Q, Yuan C F, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]. IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 13339-13348.

[15] Chi H G, Ha M H, Chi S, et al. InfoGCN: Representation learning for human skeleton-based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 20154-20164.

[16] Wang P, Wen J, Si C Y, et al. Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition[J]. IEEE Transactions on Image Processing, 2022, 31: 6224-6238.

[17] Liu Z Y, Zhang H W, Chen Z H, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 143-152.

[18] Wang Y, Sun Y B, Liu Z W, et al. Dynamic graph CNN

for learning on point clouds[J]. ACM Transactions on Graphics, 2019, 38(5): 1-12.

[19] Shi L, Zhang Y F, Cheng J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 12018-12027.

[20] Zhou H Y, Liu Q J, Wang Y H. Learning discriminative representations for skeleton based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 10608-10617.

[21] Wang J, Nie X H, Xia Y, et al. Cross-view action modeling, learning, and recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Columbus, 2014: 2649-2656.

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

[23] Zhou Y X, Yan X D, Cheng Z Q, et al. BlockGCN: Redefine topology awareness for skeleton-based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 2049-2058.

[24] Chen Z, Li S C, 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. Virtually, 2021: 1113-1122.

作者简介

宋忱 (2000−), 男, 硕士生, 主要研究方向为计算机视觉、目标检测、视频分析与理解, E-mail: sc_ahut@163.com;

钱惠敏 (1980−), 女, 副教授, 博士, 主要研究方向为目标检测、视频分析与理解, E-mail: am_hohai@163.com;

吴大伟 (1989−), 男, 副教授, 博士, 主要研究方向为先进飞行控制与协同控制、无人机智能巡检, E-mail: wudawei_hhu@hhu.edu.cn.