DOI:10.20079/j.issn.1001-893x.240722001

基于时空特征融合与注意力机制的图卷积动作识别方法*

王晓路,谭永辉,李晓婷

(西安科技大学 通信与信息工程学院,西安 710054)

摘　要:为了进一步提高人体动作识别的精度和充分发掘动作序列的时空特征,提出了基于时空特征融合与注意力机制的图卷积动作识别方法。采用空间注意力图卷积对拓扑图进行通道级细化,捕捉不同运动类型下关节的相关性特征,并采用时域多尺度图卷积模块扩展时间卷积结构以捕获多尺度时间特征。构建多层次特征融合模块将初始特征与时域多尺度图卷积输出特征作为模块输入,采用双分支结构分别获取全局和局部通道特征,并在通道维度进行时空特征融合以增强模型特征提取能力;在此基础上,提出一种肢体注意力机制对人体拓扑结构进行划分并分别计算其在通道维度上的注意力权重,加强模型对局部动作特征的关注能力。实验结果表明,在 NTU RGB+D 数据集的 CS 和 CV 评估模式下分别达到了 93.0% 和 96.9% 的识别准确率,在 NTU RGB+D 120 数据集的 X-Sub 和 X-Set 评估模式下分别达到了 89.8% 和 91.1% 的识别准确率,均高于 ST-GCN、CTR-GCN 等模型的识别准确率。

关键词:动作识别;人体骨架;图卷积;时空特征融合;注意力机制

开放科学(资源服务)标识码(OSID):

中图分类号:TP183　　**文献标志码**:A　　**文章编号**:1001-893X(2025)11-1789-09

Graph Convolution Action Recognition Based on Spatiotemporal Feature Fusion and Attention Mechanism

WANG Xiaolu, TAN Yonghui, LI Xiaoting

(School of Communication and Information Engineering, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract: In order to further improve the accuracy of human action recognition and fully explore the spatiotemporal features of action sequences, a graph convolution action recognition method based on spatiotemporal feature fusion and attention mechanism is proposed. The spatial attention map convolution is used to refine the topology to capture the correlation features of the joints under different motion types, and the time convolution structure is extended by the time domain multi-scale convolution module to capture the multi-scale time features. A multi-level feature fusion module is constructed, which takes the initial feature and the convolution output feature of the time-domain multiscale graph as the module input, and uses a two-branch structure to obtain the global and local channel features respectively. On this basis, a limb attention mechanism is proposed to divide the human topological structure and calculate the attention weights in the channel dimension respectively to enhance the model's ability to pay attention to local action features. The experimental results show that the recognition accuracy is 93.0% and 96.9% in CS and CV evaluation mode of NTU RGB+D data set, and 89.8% and 91.1% in X-Sub and X-Set evaluation mode of NTU RGB+D 120 data set, respectively. The recognition accuracy is higher than that of ST-GCN, CTR-GCN and other models.

Key words: human skeleton; motion recognition; graph convolution; spatiotemporal feature fusion; attention mechanisms

* **收稿日期**:2024-07-22;**修回日期**:2024-10-14
基金项目:西安市科技计划项目(2020KJRC0070)
通信作者:谭永辉　　Email:tanyonghui2022@163.com

0 引 言

近年来,由于人体动作识别被广泛应用于人机交互和行为检测等实用领域,因此受到广泛的关注。现有人体动作识别的研究聚焦于 RGB 图像、骨架等多模态数据[1-2]。其中,因骨架数据仅涉及关节点的空间坐标,与传统的 RGB 视频识别方法相比,基于骨架数据的动作识别能够有效降低识别过程中光照变化、环境背景、遮挡等干扰因素的影响,具有更好地适应动态环境和复杂背景的优点。这一特性使得人体骨架数据在当前的多模态数据流中具备更优的识别性能。

早期的基于深度学习方法是手工将骨架数据转换为伪图像形式,并利用卷积神经网络[3-4]提取特征信息并预测结果,但用二维图像代替人体本身的自然拓扑结构不可避免地削弱了数据间原有的内在关联性,导致识别效果受到限制。因此,在处理拓扑结构方面具有显著优越性的图卷积神经网络(Graph Convolutional Neural Network,GCN)[5-7]逐渐成为基于骨架的动作识别最广泛的方法。Liu 等人[8]提出了 MS-G3D 网络,将 CNN 的三维卷积技术整合到 GCN 中,提高捕捉局部时空信息的能力,同时引入了多尺度聚合方法,以实现有效的长距离建模。Chen 等人[9]提出了通道拓扑细化图卷积网络(Channel-wise Topology Refinement Graph Convolution Network,CTR-GCN),将预设的共享拓扑作为通用先验并在通道维度对其进行细化来获取动态拓扑,从而获取更强大的建模能力。赵登阁等人[10]构造了多尺度时空图注意卷积网络,采用多尺度以及轻量级注意力机制的方式来增强空间特征的捕获和理解。文献[11-12]采用动态拓扑的方式表示每个样本内关节之间的复杂连接关系,为动作识别提供更丰富和有用的信息。Zhang 等人[13]提出了一种多流多尺度的扩展时空图卷积网络,采用多流数据作为输入以捕获更多特征信息,并设计一种多尺度膨胀时间图卷积层以获得更丰富的特征。这些方法在提取特征进行融合时常采用残差连接的方式来保留不同层次的时空特征,在聚合时空信息时可能造成局部上下文特征的丢失或过度关注某些区域的问题。

另一方面,文献[14-17]通过引入注意力机制的方式增强网络对关节的重要特征信息提取能力。Qiu 等人[18]提出了一种新的分段注意力方法,通过将骨架序列划分为多个片段并对各片段内连续帧进行编码,同时在此基础上采用自注意力机制捕捉连续帧中不同关节之间的关系,以增强模型识别性能。Wang 等人[19]提出了一种多流注意增强递归图卷积网络,通过引入时空通道注意力机制的方式来提高模型对重要关节、帧和通道特征的关注程度。上述方法采用的注意力机制是对全局重要关节权重的学习,但人体不同动作对各肢体部位的依赖程度有所差异,导致需要重点关注的部位不同。因此,需要构建一种注意力机制加强对各肢体部位的聚焦程度,扩大同种类动作的局部动作特征差异,以提高动作识别精度。

针对上述方法中存在的局部特征丢失和重要肢体部位聚焦程度不足的问题,本文提出了一种基于时空特征融合与注意力机制的图卷积动作识别网络,用于提高时空特征有效融合和局部动作特征的关注程度。采用双分支结构构建多层次特征融合模块,以不同层面的特征信息作为输入分别提取局部通道特征和全局通道特征并进行融合,增强网络模型对不同语义层次特征的鲁棒性。同时,为提升图卷积网络对骨架局部特征提取的能力,设计了一种集肢体分割、通道和空间注意力一体的肢体注意力模块,将人体拓扑结构按照不同肢体部位进行分割,并在不同通道维度上分别训练其权重。这种细化特征图的方式能够更充分提取骨架关节点的特征,提高图卷积网络对局部相似动作特征的识别性能。

1 动作识别的模型架构

本文提出一种基于时空特征融合与注意力机制的图卷积动作识别网络,整体网络结构如图 1 所示。受文献[20]的启发,数据预处理采用具有一定互补能力的全局和局部序列预处理方式将原始骨架数据转换成不同形式的骨架特征,分别是全局关节流、全局骨骼流、局部关节流和局部骨骼流,用以缓解单流数据提取特征信息局限性问题。首先,通过空间注意力图卷积模块对人体骨架的拓扑结构进行通道级拓扑细化;其次,采用时域多尺度图卷积模块设置不同大小的膨胀率和卷积核的方式来获取长距离时间信息;随后,使用多层次特征融合模块(Multilevel Feature Fusion,MFF)接收初始特征和时间图卷积的输出特征,并采用双分支结构融合不同层次的特征信息;接着,肢体注意力模块(Limb Attention,LA)将特征融合后的特征信息在关节维度进行分割,并分别在通道维度上训练其权重,提高网络对局部动作

重要关节的关注程度;最后,将获得的四流预测分数整合以获得综合分类结果。

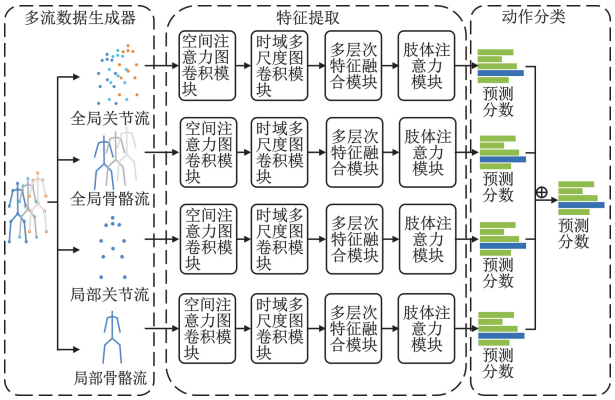


图 1　模型整体架构

2　多流数据生成器

2.1　全局序列预处理

在原始数据集中,由于不同动作序列的初始位置不同,导致不同动作序列的关节坐标存在一定的偏差。在对不同动作序列的识别时,这种偏差对模型的判别具有一定影响。为了改善这种负面影响,本文使用全局序列[20]的处理方式对原始数据进行处理。全局序列预处理是指一组动作序列中的所有帧引用同一个坐标系,坐标原点选取第一帧的中心关节。这种方式能够更好地处理长时间序列的依赖关系,提高模型捕捉全局信息的能力。

2.2　局部序列预处理

局部序列预处理[20]是对不同动作序列中每一帧单独使用一个坐标系,以每一帧的中心关节作为坐标原点,通过将脊柱设置为 Z 轴,将平行于肩膀方向设置为 X 轴,使得不同帧之间的身体方向相同,这有助于忽略方向偏差,更关注于不同帧之间动作的变化。因篇幅所限,全局和局部序列预处理公式及示意图略,读者可扫描本文 OSID 码在“本文开放的科学数据与内容”中查看。

3　基于时空特征融合与注意力机制的特征提取

3.1　空间注意力图卷积模块

在空间建模中,采用空间注意力图卷积模块来提取人体关节以及骨骼内部之间的相关性,图 2 为空间注意力图卷积示意图。

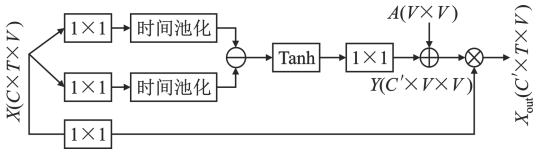


图 2　空间注意力图卷积模块

首先,将预处理后的骨架数据馈送到两个并行分支中,采用一维卷积和时间池化对特征信息进行降维处理并特征提取;其次,对两个分支执行逐元素减法,计算不同关节点沿通道维度之间的距离并将其作为通道维度特定的拓扑关系;最后,将生成的特征图与预定义邻接矩阵执行广播加法运算,以获得最终的通道级拓扑,实现更灵活有效地聚合空间信息。公式如下:

$$Y=\beta Q(X)+A \tag{1}$$

式中:$X\in\mathbb{R}^{C\times T\times V}$ 表示模块输入特征,其中,C 表示通道维度,T 表示样本的帧数,V 表示单个骨架中关节的个数;+表示广播加法运算;$A\in\mathbb{R}^{V\times V}$ 表示通道之间共享的先验拓扑;$Y\in\mathbb{R}^{C'\times V\times V}$ 表示经过通道细化后的通道级拓扑;β 表示用于调整细化强度的可学习参数;Q 表示不同通道下特定的拓扑关系,其公式如下:

$$Q(X)=\psi(\sigma(\mathrm{TP}(\phi(X))-\mathrm{TP}(\varphi(X)))) \tag{2}$$

式中:ϕ 和 φ 表示两个一维卷积作为嵌入函数,用于对特征矩阵映射重塑并降低通道维度以减少计算量;TP 表示时间池化;-表示广播减法运算,用于计算不同关节特征之间的距离;σ 表示激活函数 tanh,用于对关节特征距离进行非线性变换,将其作为特定通道下不同关节之间的拓扑关系;ψ 表示一维卷积,用于增加通道维度以实现通道级细化拓扑。

在获得通道级拓扑之后,本文采用一维卷积调整初始骨架特征的通道维度,并将输出与通道级拓扑相乘,以聚合空间维度信息。模块输出计算公式如下:

$$X_{\mathrm{out}}=Y\times(\theta(X)) \tag{3}$$

式中:$X_{\mathrm{out}}\in\mathbb{R}^{C'\times V\times V}$ 表示空间注意力图卷积模块的输出特征;θ 表示与 ψ 输出通道数不同的一维卷积;×表示矩阵乘法运算。

3.2　时域多尺度图卷积模块

考虑到传统的时间图卷积采用固定大小的卷积核无法有效地捕捉节点特征随时间的变化,因此本

文通过多尺度的方式设置不同膨胀率和卷积核对骨架序列在时间维度上进行多种细粒度的特征提取，从而扩大感受野范围。时域多尺度图卷积模块包含 4 个分支，每个分支均使用一维卷积用于降低通道维度以减少计算量，其中两个分支采用不同大小的膨胀率以提取不同细粒度的时间特征，另一支路采用最大池化用于获取时域中重要特征，最后一个分支保留时间帧上的原始特征，最终将 4 条支路的输出在通道维度上进行拼接，以获得与模块输入特征通道数一致的输出。因篇幅所限，示意图略，读者可扫描本文 OSID 码在“本文开放的科学数据与内容”中查看。

3.3 多层次特征融合模块

尽管空间图卷积和时间图卷积能够有效提取时空特征信息，但在融合不同层次的空间和时间特征时常采用残差连接的方式，可能导致在融合过程中丢失关键的局部特征或过度关注某些区域，从而影响网络模型对全局和局部特征的充分提取。针对上述问题，本文提出多层次特征融合模块，用于取代传统方式中不同语义层次之间的残差连接。模块通过将初始骨架数据和多尺度聚合特征作为模块输入，采用双分支结构将局部通道上下文特征信息与全局通道上下文特征信息融合，从而增强模型对不同语义层次特征的鲁棒性。如图 3 所示，多层次特征融合模块采用的双分支结构，分别是全局通道特征提取与局部通道特征提取，通过使用自适应平均池化的方式实现在通道维度上聚合全局上下文信息，从而弥补不同尺度之间特征不一致的问题。同时，为尽可能保持模型的轻量级，采取逐点卷积的方式仅在通道维度上对骨架空间坐标信息交互，从而有效聚合特征信息并减少计算复杂度。

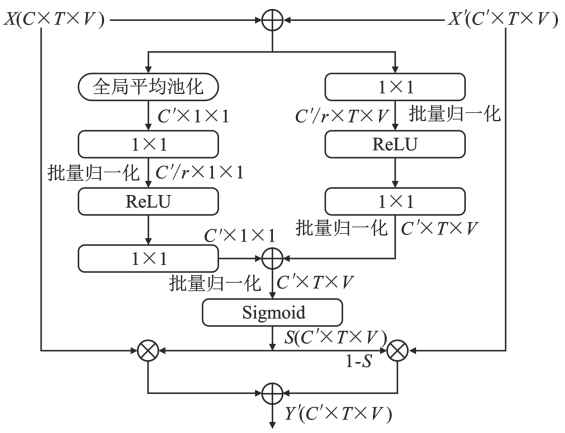


图 3 多层次特征融合模块

第一个分支是全局通道上下文，用于捕获全局通道的上下文信息，其计算方式如下：

$$G(Z)=B(\mathrm{Conv2}(\delta(B(\mathrm{Conv1}(g(Z))))))\quad(4)$$

$$Z=X+X'\quad(5)$$

式中：$G(Z)$ 表示全局通道特征的输出结果；B 表示归一化函数；δ 表示 ReLU 激活函数；$g(Z)$ 表示全局平均池化；Conv1 表示输出通道数为 C'/r 的逐点卷积，r 表示通道缩减率；Conv2 表示输出通道数为 C' 的逐点卷积；$Z\in\mathbb{R}^{C'\times T\times V}$ 表示模块输入；$X\in\mathbb{R}^{C\times T\times V}$ 表示空间图卷积模块的输入特征，即初始特征；$X'\in\mathbb{R}^{C'\times T\times V}$ 表示时间图卷积模块的输出，即多尺度聚合的特征信息；+表示广播加法运算。

第二个分支是局部通道上下文，用于学习多通道局部特征信息之间的关联性来捕获通道上下文信息，其与公式(4)类似，具体公式如下：

$$L(Z)=B(\mathrm{Conv2}(\delta(B(\mathrm{Conv1}(Z)))))\quad(6)$$

式中：$L(Z)$ 表示局部通道特征的输出结果；B 表示归一化函数；δ 表示 ReLU 激活函数；Conv1 表示输出维度为 C'/r 的逐点卷积，r 表示通道缩减率；Conv2 表示输出维度为 C' 的逐点卷积。

通过上述的全局通道特征 $G(Z)$ 与局部通道特征 $L(Z)$，进一步聚合两条支路的权重并将权重大小映射为 0~1 之间，具体公式如下：

$$S=\mathrm{Sigmoid}(G(Z)+L(Z))\quad(7)$$

式中：$S\in\mathbb{R}^{C'\times T\times V}$ 表示两条支路的融合权重输出；Sigmoid 表示激活函数；+表示广播加法运算。

结合上述公式，再将两分支输出分别与模块输入 X 和 X' 加权融合，从而有效融合低维特征和高维特征。多层次特征融合模块的计算公式如下：

$$Y'=S\otimes X\oplus(1-S)\otimes X'\quad(8)$$

式中：Y' 表示多层次特征融合后的特征；S 表示两条支路的融合权重输出；X 和 X' 表示模块的输入；$\otimes$ 表示逐元素乘法；$\oplus$ 表示逐元素相加。

3.4 肢体注意力模块

由于不同动作对身体各个肢体部位的依赖程度各不相同，因此本文提出了肢体注意力模块，其作用于最后 4 个网络层，通过引入额外的编码机制来强调位于特定身体部位的动作特征，可以更加精准地学习不同动作对相应肢体部位的关注程度。模块包含肢体分割、通道注意力和空间注意力，通过对身体各部位分配不同的权重系数来聚焦动作幅度较大的关节部位。如图 4 所示，首先根据人体先验知识，将

人体分为 5 个部分,分别是脊椎、左臂、右臂、左腿和右腿,对每个肢体部位分别计算其通道维度上的注意力,使得特征图具有更高的细粒度,并允许不同通道的卷积核学习更丰富的信息,使模型在多种姿态等不确定因素时表现出更好的鲁棒性和识别性能。

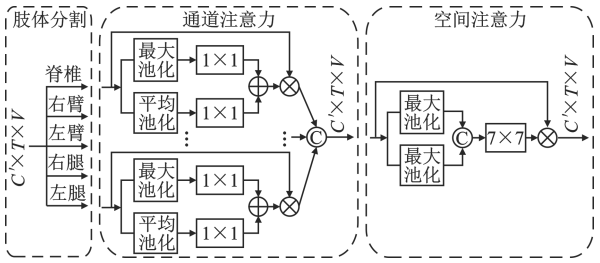


图 4　肢体注意力模块

模块的输入特征为 $Y' \in \mathbb{R}^{C' \times T \times V}$,首先需要对输入特征在空间维度上按照肢体位置分为 5 个部分,记为 $Y'_i \in \mathbb{R}^{C' \times T \times V_i}$,$i \in [1,5]$,$V_i$ 表示每个身体部位中的关节集合,其遵循对称原则对身体关节集进行分割。为有效计算通道注意力,本文将分割后的肢体关节集合分别进行基于时间和空间的全局最大池化和全局平均池化,分别获取不同肢体部位在时间和空间维度的突出特征和全局特征。为了增加网络的非线性,在卷积层中对特征图通道数进行升维和降维操作,同时为获取不同身体部位在通道维度上的重要特征,将每个肢体部位在不同池化层和一维卷积层生成的特征图在通道维度上相加,以获得不同肢体部位在通道维度的权重,并与肢体部位特征图 Y'_i 相乘,进而获得各部位的通道注意力特征图 $M_i \in \mathbb{R}^{C' \times T \times V}$,$i \in [1,5]$,其中每个通道代表不同肢体部位下不同运动模式的特征和其重要程度。最后将各肢体部位特征图在关节维度进行拼接以获得易于网络理解的细化特征输出。上述实现过程可以总结为下式:

$$M_i = \sigma(\mathrm{Conv}(\mathrm{AvgPool}(Y'_i)) + \mathrm{Conv}(\mathrm{MaxPool}(Y'_i)))\otimes Y'_i \tag{9}$$

$$M = \mathrm{Cat}(M_1 + M_2 + \cdots + M_5) \tag{10}$$

式中:σ 表示 Sigmoid 激活函数;Conv 表示包含两个用于升维和降维的一维卷积;AvgPool 表示在空间和时间维度上的平均池化;MaxPool 表示在空间和时间维度上的最大池化;$\otimes$ 表示逐元素乘法;Cat 表示特征图在关节维度上的拼接;$M \in \mathbb{R}^{C' \times T \times V}$ 表示人体骨架在通道注意力后的特征图。

空间注意力是将通道注意力的输出 M 分别在通道维度上进行池化,以捕获全局上下文信息和重点特征信息,并使用二维卷积获取不同肢体部位在空间维度上的权重。最终将空间注意力的输出与通道注意力的输出融合,捕获各通道下重要关节的特征信息,从而提高模型性能。具体实现过程可总结为下式:

$$L = \sigma(\mathrm{Conv2d}(\mathrm{AvgPool}(M) + \mathrm{MaxPool}(M)))\otimes M \tag{11}$$

式中:σ 表示 Sigmoid 激活函数;Conv2d 表示二维卷积;AvgPool 表示在通道维度上的平均池化;MaxPool 表示在通道维度的最大池化;$\otimes$ 表示逐元素乘法;L 表示肢体注意力模块的输出特征。

4　实验与结果分析

4.1　数据集

NTU-RGB+D 数据集[21]是公开的三维人体动作识别的数据集,包含 60 个动作类别,56 880 个骨架序列样本。该数据集按照不同的划分标准分为 Cross-Subject(CS)和 Cross-View(CV)两个子集,其中,CS 是根据志愿者 ID 来划分训练集和测试集,分别包含 40 320 个样本和 16 560 个样本;CV 是根据相机 ID 来划分训练集和测试集,分别包含 37 920 个样本和 18 960 个样本。

NTU RGB+D 120 数据集[22]是目前用于动作识别最大的数据集,包含 120 个动作类别,113 945 个骨架序列样本。数据集按照 Cross-Subject(X-Sub)和 Cross-Setup(X-Set)两种评价指标进行划分,其中,X-Sub 是根据志愿者 ID 来划分训练集和测试集,分别包含 63 026 个样本和 50 919 个样本;X-Set 是根据摄像设置 ID 来划分训练集和测试集,分别包含 54 468 个样本和 49 477 个样本。

4.2　实验设置

本文在一台单卡 GPU 为 V100-SXM2 的设备,Pytorch 版本为 1.1.0,Python 版本为 3.7,Cuda 版本为 10.0,操作系统为 ubuntu18.04 上进行实验。模型的损失函数采用交叉熵函数,权重衰减参数设置为 0.000 4,样本批次的大小设置为 64,学习率初始值为 0.1,在第 35 与第 55 个 epoch 分别下调为原来的 1/10。训练 epoch 设置为 65,为了使训练更稳定,前 5 个 epoch 中使用了预热策略[23]。

4.3　消融实验

针对现有研究方法在特征提取时存在特征丢失

和不同肢体动作下重要关节部位关注程度不足的问题,本文提出了多层次特征融合模块和肢体注意力机制,用于有效利用不同层次的特征信息,增强模型的泛化能力,以及深入挖掘不同肢体部位之间的依赖关系,从而有效提高模型识别相似动作的能力。为了验证本文所提方法的有效性,以 CTR-GCN 为基线,在 NTU-RGB+D 120 数据集 X-Sub 评估模式下进行消融实验,通过实验验证了不同模块对模型识别精度的影响。其中,4s、MFF 和 LA 分别表示本文所使用的全局和局部预处理下获取的四流输入数据、多层次特征融合模块和肢体注意力模块,w/o 表示在本文模型的基础上将后缀模块去除进行消融实验,分别验证其在模型中的有效性。实验结果如表 1 所示,可以观察到,当模型不使用 LA 模块时,模型的识别准确率下降了 0. 6%;当模型再去除 MFF 模块后,识别准确率下降了 1. 0%;在上述的基础上去除本文使用的全局和局部模式下的预处理方式,采用关节流、骨骼流、关节运动流和骨骼运动流作为模型输入,识别准确率相较于本文提出的模型识别准确率下降了 1. 3%。上述实验结果验证了本文提出的方法对模型整体性能有一定的提升。

表 1 模块的消融实验结果

方法	参数量/10^6	识别准确率/%
本文 w/o 4s,MFF,LA	5. 84	88. 5
本文 w/o MFF,LA	5. 84	88. 8
本文 w/o LA	10. 04	89. 2
本文	11. 44	89. 8

4.4 多流融合模型

目前,许多方法采用了多流融合策略来对模型的识别精度进行提升,采用的数据流多为关节、骨骼、关节运动、骨骼运动以及骨骼角度等状态的数据,但某些状态的数据流对于模型多流融合后识别效果贡献有限,因此本文采用了局部关节流和局部骨骼流两种状态的数据流取代关节运动流和骨骼运动流,进一步验证本文采用的预处理方式的优越性,实验结果如表 2 所示,其中“J^S”“B^S”“V^J”“V^B”“J^F”和“B^F”分别表示全局关节流、全局骨骼流、关节运动流、骨骼运动流、局部关节流和局部骨骼流。6 种输入数据在 NTU-RGB+D 的 CS 评估模式下进行实验,在表中可以看到局部数据流作为模型输入的识别精度均比运动流的识别精度高,且四流网络“$J^S+B^S+J^F+B^F$”的融合准确率相比于传统的四流网络“$J^S+B^S+V^J+V^B$”也具有更高的准确率,同时六流网络“$J^S+B^S+J^F+B^F+V^J+V^B$”相对四流网络“$J^S+B^S+J^F+B^F$”在增加大量计算量情况下准确率仅提升 0. 1%,因此本文采用四流融合网络进行实验。

表 2 多流骨架特征输入的有效性对比

流	识别精度/%
J^S	90. 1
B^S	91. 2
V^J	88. 0
V^B	87. 5
J^F	89. 2
B^F	90. 1
J^S+B^S	92. 4
$J^S+B^S+V^J+V^B$	92. 7
$J^S+B^S+J^F+B^F$	93. 0
$J^S+B^S+J^F+B^F+V^J+V^B$	93. 1

为了进一步验证本文模型采用的局部关节流与局部骨骼流作为输入的有效性,本文在 NTU-RGB+D 和 NTU-RGB+D 120 数据集的不同评估模式下分别进行单流和多流融合实验,实验结果如表 3 所示。从表中可以看到在两个大规模数据集上,多流融合的方法相较于单流的识别精度有明显提升;而且在加入局部关节流和局部骨骼流后,相比于仅有“J^S+B^S”的双流融合方法来说模型性能得到一定改善,在 NTU-RGB+D 的 CS 和 CV 上准确率分别提升了 0. 6% 和 0. 7%,在 NTU-RGB+D 120 的 X-Sub 和 X-Set 上准确率提升了 0. 8% 和 0. 5%。

表 3 多流融合在不同评估模式的性能对比

流	识别准确率/%			
	NTU-RGB+D		NTU-RGB+D 120	
	CS	CV	X-Sub	X-Set
J^S	90. 1	94. 9	84. 6	86. 9
B^S	91. 2	94. 9	86. 9	88. 2
J^F	89. 2	94. 8	83. 0	84. 6
B^F	90. 1	95. 0	85. 7	86. 7
J^S+B^S	92. 4	96. 2	89. 0	90. 6
$J^S+B^S+J^F+B^F$	93. 0	96. 9	89. 8	91. 1

4.5 先进方法的对比

本文在 NTU-RGB+D 和 NTU-RGB+D 120 两个数据集的不同评估模式下与其他前沿方法的识别性能进行了对比。本文使用的基线 CTR-GCN 是对传

统的图卷积方法在拓扑结构上进行优化,将静态拓扑结构改进为动态拓扑结构,并对拓扑结构进行通道级细化,以提高模型的表达能力;而本文在基线 CTR-GCN 的基础上构建了多层次特征融合模块以充分利用不同层次的特征信息,同时提出一种肢体注意力机制用于分割肢体部位并分别进行权重训练,以提高模型对相似动作的识别能力。比较结果如表 4 所示,其中,基线 CTR-GCN 的准确率为官方代码复现结果,其他方法的准确率均来自于论文中的结果。从表 4 可以看出,本文提出的算法超越了大多数基于 GCN 的先进方法,在 NTU-RGB+D 数据集的 CS、CV 评估模式下分别取得了 93.0% 和 96.9% 的准确率,相比于 CTR-GCN 分别提升了 1.1% 和 0.4%;在 NTU-RGB+D 120 数据集的 X-Sub、X-Set 评估模式下分别取得了 89.8% 和 91.1% 的准确率,相比于 CTR-GCN 分别提升了 1.3% 和 1.0%。

表 4　NTU 数据集上与其他方法的比较

方法	识别准确率/%			
	NTU-RGB+D		NTU-RGB+D 120	
	CS	CV	X-Sub	X-Set
ST-GCN[6]	81.5	88.3	70.7	73.2
2s-AGCN[7]	88.5	95.1	82.5	84.2
SGN[27]	89.0	94.5	79.2	81.5
ST-TR[14]	90.3	96.3	85.1	87.1
Shift-GCN[25]	90.7	96.5	85.9	87.6
IST-GCN[16]	90.8	96.2	87.0	88.1
Dynamic GCN[24]	91.5	96.0	87.3	88.6
MS-G3D[8]	91.5	96.2	86.9	88.4
MST-GCN[26]	91.5	96.6	87.5	88.8
CTR-GCN[9]	91.9	96.5	88.5	90.1
SAGGAN[11]	92.1	96.7	88.1	89.5
STF[28]	92.5	96.9	88.9	89.9
STTGC-Net[29]	92.8	96.5	89.6	91.1
FR Head[30]	92.8	96.8	89.5	90.9
本文	93.0	96.9	89.8	91.1

上述对比结果表明,本文提出的网络模型能够将不同层次特征信息进行有机融合,从而增强模型对不同语义层次特征的鲁棒性,并且能够有效地关注每个肢体部位的重要关节,使得针对不同动作具有更好的识别效果,有效提升了模型的识别性能。

4.6　数据可视化

4.6.1　不同动作的可视化分析

本文所提出的算法通过构建肢体注意力机制,提高了特定动作对相关肢体部位的关联强度。为了验证该机制对模型性能的影响,本文在图 5 中可视化了不同动作下每个关节的响应程度,其中每个关节的受关注程度以圆的大小直观展示,圆的尺寸越大代表关注度越高。在“扔”的动作中,手臂部位的运动幅度最大,因此在动作识别过程中手臂关节的响应程度更高,其他身体部位的关节受关注程度相对较低。在“踢东西”的动作中,由于在整个运动序列中以左腿的肢体幅度最大,因此注意力聚焦在左腿关节部位,在动作识别中起到引领作用。

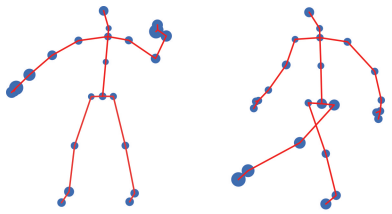


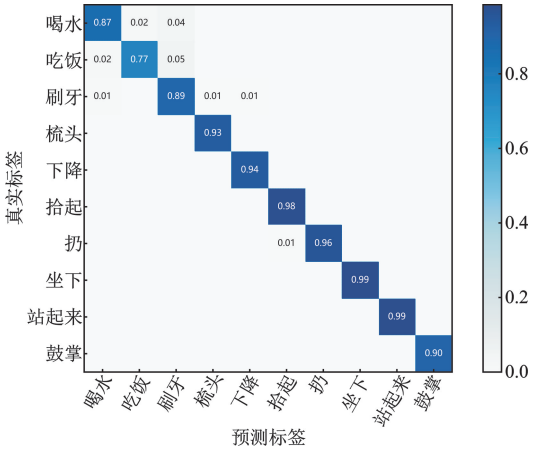
图 5　不同动作的响应程度

本文通过将人体骨架分割成 5 个部分,针对不同动作分别训练不同肢体部位实现细化特征图的目的,使模型能够根据不同动作重点关注不同身体部位关节在全局的权重比例,减少与动作不相关或关联度较低的关节对模型性能的影响。另一方面,多层次特征融合模块通过融合不同层次的特征信息,增强了模型对关节相关性的建模能力,以此提高了模型对不同行为动作的识别能力。

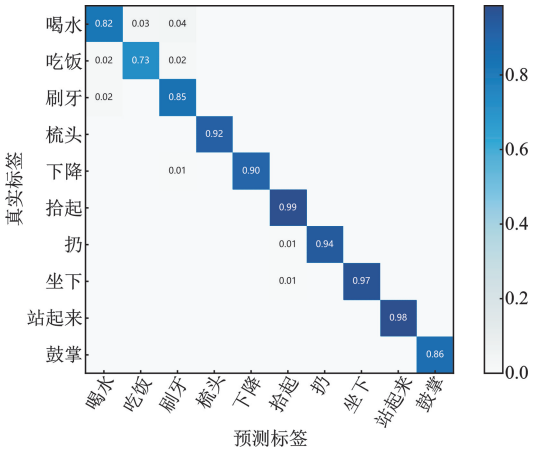
4.6.2　混淆矩阵可视化对比

为更好地验证本文方法的性能,在 NTU RGB+D 数据集的 X-Sub 评估模式下分别可视化了本文提出的网络模型和 CTR-GCN 网络模型。如图 6 所示,为直观表达模型效果,仅在数据集的前 10 个动作类别上以热力图的形式绘制混淆矩阵,其中,对角线上的元素值表示预测正确的比例,非对角线上元素值表示预测错误的比例,热力图色标的颜色深度与混淆矩阵中每个元素的数值大小成正比。通过对比可看出在识别相似动作“喝水”“吃饭”和“刷牙”3 个动作时,本文提出的算法的识别精度相较于 CTR-GCN 的识别精度均具有一定提升,表明本文提出的肢体注意力机制能够对相似动作做出更准确的识别效果,在处理复杂时空关系的动作时能够更精细化

关注重要关节,相比于 CTR-GCN 具有更好的捕捉不同运动状态下独有特征的能力。



(a)本文



(b)CTR-GCN

图 6 混淆矩阵可视化对比

5 结束语

本文提出了基于时空特征融合与注意力机制的图卷积动作识别方法。该方法采用双分支结构对不同语义层次的特征信息进行特征提取和相关性融合,同时使用肢体注意力机制对人体骨架进行肢体划分,并对其分别进行权重训练以聚焦运动肢体的重要关节,使模型对不同动作具有更高的识别精度。本文模型分别在两个大型数据集 NTU RGB+D 和 NTU RGB+D 120 上进行大量实验,并与多个主流模型准确率对比验证了其先进性。其中,对于 NTU RGB+D 数据集,本文模型在 X-Sub、X-Set 评估模式下的准确率分别为 93. 0% 和 96. 9%,较基线模型 CTR-GCN 分别提升了 1. 1% 和 0. 4%;对于 NTU-RGB+D 120 数据集,本文模型在 X-Sub、X-Set 评估模式下的准确率分别为 89. 8% 和 91. 1%,较基线模型 CTR-GCN 分别提升了 1. 3% 和 1. 0%。实验结果表明本文所提方法能够提高动作识别性能。

考虑到模型训练参数量较大的情况,下一步聚焦于如何在不牺牲模型识别性能的基础上有效降低参数量的研究。

参考文献:

[1] 刘宽,奚小冰,周明东. 基于自适应多尺度图卷积网络的骨架动作识别[J]. 计算机工程,2023,49(10):264-271.

[2] 杨思佳,辛山,刘悦,等. 基于 3D ResNet-LSTM 的多视角人体动作识别方法[J]. 电讯技术,2023,63(6):903-910.

[3] KE Q H, BENNAMOUN M, AN S J, et al. A new representation of skeleton sequences for 3D action recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu:IEEE,2017:4570-4579.

[4] LIU M Y, YUAN J S. Recognizing human actions as the evolution of pose estimation maps[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE,2018:1159-1168.

[5] NIEPERT M, AHMED M, KUTZKOV K. Learning convolutional neural networks for graphs[C]//The 33rd International Conference on International Conference on Machine Learning. New York:ACM,2016:2014-2023.

[6] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//The 32nd AAAI Conference on Artificial Intelligence. New Orleans:AAAI Press,2018:7444-7452.

[7] 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. Long Beach:IEEE,2019:12018-12027.

[8] 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. Seattle:IEEE,2020:140-149.

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

[10] 赵登阁,智敏. 用于人体动作识别的多尺度时空图卷积算法[J]. 计算机科学与探索,2023,17(3):719-732.

[11] PANG C, GAO X Y, CHEN Z Y, et al. Self-adaptive graph

with nonlocal attention network for skeleton-based action recognition[J]. IEEE Transactions on Neural Networks and Learning Systems,2024,35(12):17057-17069.

[12] 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,2024,127:1-10.

[13] ZHANG H P, LIU X, YU D J, et al. Skeleton-based action recognition with multi-stream, multi-scale dilated spatial-temporal graph convolution network[J]. Applied Intelligence,2023,53(14):17629-17643.

[14] PLIZZARI C, CANNICI M, MATTEUCCI M. Spatial temporal Transformer network for skeleton-based action recognition[C]//The 25th International Conference on Pattern Recognition. Cham:Springer,2021:694-701.

[15] YANG H G, REN Z L, YUAN H Q, et al. Multi-scale and attention enhanced graph convolution network for skeleton-based violence action recognition[J]. Frontiers in Neurorobotics,2022,16:1-14.

[16] XING Y L, ZHU J, LI Y, et al. An improved spatial temporal graph convolutional network for robust skeleton-based action recognition[J]. Applied Intelligence,2023,53(4):4592-4608.

[17] 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. Paris: IEEE, 2023: 10410-10419.

[18] QIU H L, HOU B, REN B, et al. Spatio-temporal segments attention for skeleton-based action recognition [J]. Neurocomputing,2023,518:30-38.

[19] WANG H J,BAI B Q,LI J H,et al. Action recognition method based on multi-stream attention-enhanced recursive graph convolution[J]. Applied Intelligence, 2024,54(20):10133-10147.

[20] HOU R J,WANG Z H,REN R M,et al. Multi-channel network: constructing efficient GCN baselines for skeleton-based action recognition [J]. Computers & Graphics,2023,110:111-117.

[21] 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. Las Vegas:IEEE,2016:1010-1019.

[22] 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.

[23] HE K M,ZHANG X Y,REN S Q,et al. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas:IEEE,2016:770-778.

[24] YE F F,PU S L,ZHONG Q Y,et al. Dynamic GCN: context-enriched topology learning for skeleton-based action recognition [C]//The 28th ACM International Conference on Multimedia. Seattle:ACM,2020:55-63.

[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. Seattle:IEEE,2020:180-189.

[26] CHEN Z, LI S C, YANG B, et al. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition[J]. Computer Science,2021,35(2): 1113-1122.

[27] 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. Seattle:IEEE,2020:1109-1118.

[28] KE L P,PENG K C,LYU S W. Towards To-a-T spatio-temporal focus for skeleton-based action recognition [EB/OL]. [2024-07-15]. https://doi. org/10. 1609/aaai. v36i1. 19998.

[29] YAGNESHWAR T, MUKHERJEE S. STTGC-net: spatial-temporal transformer with graph convolution for skeleton-based action recognition[C]//The Fourteenth Indian Conference on Computer Vision, Graphics and Image Processing. Rupnagar:ACM,2024:1-10.

[30] ZHOU H Y, LIU Q J, WANG Y H. Learning discriminative representations for skeleton based action recognition [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: ACM,2023:10608-10617.

作者简介:

王晓路　男,1977 年生于四川广安,2010 年获工学博士学位,现为副教授,主要研究方向为物联网、人工智能。

谭永辉　男,1999 年生于河南周口,2022 年获工学学士学位,现为硕士研究生,主要研究方向为人工智能。

李晓婷　女,2000 年生于陕西蒲城,2022 年获工学学士学位,现为硕士研究生,主要研究方向为人工智能。