山西大学学报(自然科学版)48(1):55—65,2025
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024141

基于时空增强双分支图卷积网络的骨骼行为识别

施宇航[1],何强[1,2*],王恒友[1,2]
(1.北京建筑大学 理学院,北京 102616;
2.北京建筑大学 大数据建模理论与技术研究所,北京 102616)

摘 要:针对现有基于骨骼行为识别的图卷积的方法存在关节划分固定、重视空间信息而忽视时间信息并且网络参数量较高等问题。首先引入对称关节的信息,增加对称动作的交互特征;其次,加入多尺度金字塔(Multi-scale Pyramid, MSP)时间图卷积模块,形成双分支(Dual-branch, DB)的网络结构,提高网络对时间维度的信息提取能力;最后,本研究利用特征映射和空间聚合(Feature Mapping and Spatial Aggregation,FM-SA),在保留原始拓扑结构信息的前提下,过滤了权重矩阵中的冗余部分,并添加了挤压-激励(Squeeze-and-Excitation,SE)模块,从而有效提升了空间特征的提取能力和特征图的表达能力。实验结果表明,与基准模型相比,网络参数量减少51%,在NTU RGB+D 120数据集上的关节、骨骼流的识别准确率分别提高了0.5%和1.3%,融合准确率提高0.7%,0.5%,在NTU RGB+D、Northwestern-UCLA(NW-UCLA)数据集的识别准确率分别提升0.1%,0.2%,1.5%。本文模型的有效性和可行性得到验证。

关键词:骨骼行为识别;关节分区;时空信息增强;多尺度金字塔;映射聚合
中图分类号:O436 **文献标志码**:A **文章编号**:0253-2395(2025)01-0055-11

Spatiotemporally Enhanced Dual-branch Graph Convolutional Network for Skeleton-based Action Recognition

SHI Yuhang[1], HE Qiang[1,2*], WANG Hengyou[1,2]
(1. *School of Science, Beijing University of Civil Engineering and Architecture, Beijing* 102616*, China;*
2. *Institute of Big Data Modeling Theory and Technology, Beijing University of Civil Engineering and Architecture, Beijing* 102616*, China*)

Abstract: There are issues with existing graph convolution methods for skeleton-based action recognition, such as fixed joint segmentation, an emphasis on spatial information while neglecting temporal information, and a high number of network parameters.To address these issues, firstly, the information of symmetric joint is introduced to increase the interactive features of symmetric action. Secondly, the Multi-scale pyramid (MSP) time graph convolution module is added to form a Dual-branch (DB) network structure to improve the ability of the network to extract time dimension information. Finally, this study employs feature mapping and spatial aggregation (FM-SA) to filter out redundant parts in the weight matrix while preserving the original topological structure information, and incorporate a Squeeze-and-Excitation (SE) module to effectively enhance the extraction of spatial features and the expressive power of the feature maps. The experimental results show that compared with the benchmark model, the number of network parameters is reduced by 51%, the recognition accuracy of joint and bone flow on the NTU RGB+D 120 dataset is increased by 0.5% and

收稿日期:2024-06-20;**接受日期**:2024-10-24
基金项目:国家自然科学基金(62072024;12301581);北京市教育委员会科学研究计划项目(KM202210016002);北京建筑大学硕士研究生创新项目(09081024002)
作者简介:施宇航(1998—),男,河南周口人,硕士研究生,研究方向为计算机视觉、行为识别。E-mail:ansyh0512@126.com
* **通信作者**:何强(HE Qiang),E-mail:heqiang@bucea.edu.cn
引文格式:施宇航,何强,王恒友.基于时空增强双分支图卷积网络的骨骼行为识别[J].山西大学学报(自然科学版),2025,48(1):55-65. DOI:10.13451/j.sxu.ns.2024141.

1.3%, and the fusion accuracy is increased by 0.7% and 0.5%. The recognition accuracy of NTU RGB+D and NW-UCLA datasets is increased by 0.1%, 0.2% and 1.5%, respectively. The validity and feasibility of this model are verified.

Key words: skeleton behavior recognition; joint partitioning; spatiotemporal information enhancement; multi-scale pyramid; mapping aggregation

0 引言

随着互联网的发展,以人为主题的视频数据量愈发庞大,人体行为的智能识别作为计算机视觉领域的热点之一,可以用在例如视频监控、智能人机交互、体育仲裁[1-2]、智能医疗[3]等很多领域,具有重要的理论与应用价值。

早期骨骼行为识别的方法大多数基于循环神经网络(Recurrent Neural Network,RNN)或者结合长短期记忆网络(Long Short-Term Memory,LSTM),例如Du等[4]提出一种端到端的双向RNN模型,将人体的骨架序列划分为五个部分后分别送入不同的子网络,最后对子网络的结果进行融合输出。Lee等[5]提出的时间滑动LSTM网络模型,该网络先将骨骼数据转换到不同的坐标系后输入到网络进行多尺度的时间特征提取。后来出现基于卷积神经网络(Convolutional Neural Network,CNN)的方法,Ke等[6]将骨骼序列划分为片段,转化到图像坐标系中利用CNN分层学习其图像特征,对骨骼序列进行长期的时间建模。

近年来,图卷积神经网络因其能够揭示人体拓扑结构和关节相关性等的优点,被广泛地应用于人体骨骼行为识别。Yan等[7]提出一种时空图卷积网络(Spatial Temporal Graph Convolutional Networks,ST-GCN),用图卷积网络(Graph Convolutional Network,GCN)建模人体关节的相关性并结合时间卷积来提取运动特征。Shi等[8]采用局部网络计算不同关节之间关系生成自适应邻接矩阵,构建双流网络利用骨骼数据的二阶信息(骨骼的长度和方向)实现不同模态的信息互补。Li等[9]提出多流网络模型,增加输入信息,并通过密集连接的方式强化模型的时间提取能力。Shi等[10]将注意力机制融入网络之中,通过加入时空-通道-注意力模块增加模型对重要关节、帧和特征的关注。其他方法[11-12]主要是在网络的结构设计上做出改变,以更好地捕捉多尺度的联合关系。

尽管基于图卷积的骨骼行为识别方法较为常用,但仍然存在以下不足之处:(1)空间卷积在人体先验关节连接的基础上提取空间特征,但是先验连接不能体现未连接关节之间的相关性,不能充分地利用人体的结构信息;(2)先验的物理连接拓扑矩阵在学习过程中会出现信息损失的情况;(3)模型大多倾向于空间维度建模或时间维度仅利用单一卷积核处理,不能充分提取时间方面的信息。

针对上述问题,本文提出了一种时空增强的双分支图卷积网络(Spatiotemporally Enhanced Dual-branch Graph Convolutionsl Network,STEDB-GCN)。因为运动中的人体肢体关节内部相关性较强,加入对称关节的划分方式,能够加强内部关节之间的联系,增加有用的信息输入。其次,在学习过程中邻接矩阵和权重矩阵分别会产生信息的损失和冗余[12],根据关节连接的最短距离对于权重矩阵进行映射分组,在保证邻接矩阵信息利用率的前提下也减少了部分冗余信息。最后,由于基础模型通道级拓扑优化图卷积(Channel-wise Topology Refinement Graph Convolution,CTR-GCN)侧重于改善空间特征的提取,所以增加一个包含多种尺度卷积核的金字塔[13]时间图卷积分支,增强了时间特征的提取,更好地聚焦上下文并提取有用信息,使网络得到的时空特征更具判别力的同时也减少了模型的参数量。

总体而言,本文的主要工作包括以下3个方面:

(1)针对现在的大多数方法只在相邻关节点进行卷积的方式,引入了骨骼的对称信息,提取人体动作中的对称行为的特征,对权重矩阵进行分组计算,加入挤压激励模块,提高各通道包含的不同层次语义信息的提取和表达能力;(2)提出双分支的图卷积网络,在普遍使用的时间-空间网络框架基础上增加一个时间分支,强化时间特征;(3)加入金字塔多尺度时空

建模,充分捕捉骨骼序列的长短期时空动作信息,进一步提高网络时空建模能力。

1 基于骨骼行为识别的相关工作

1.1 图卷积网络

图神经网络(Graph Neural Network,GNN)现在已被广泛使用。GNN有两种思想,光谱视角[14]和空间视角[15-16]。本文遵循空间透视GNN的原则。作为经典的GNN模型之一,ST-GCN[7]在每个关节的物理邻接点上应用图卷积,对每帧骨骼数据的空间特征进行编码,还将关节划分不同子集。Shift-GCN[17]采用移位操作收集从所有其他关节到当前关节的信息,将感受野扩大到整个骨架。这些方法使用GCN对骨骼结构进行建模,但是预定义骨骼图忽略了动作识别中样本相关特征,无法学习图的拓扑结构。MotifGCN[18]模型在非物理连接的关节之间构建样本相关的关系进行动作识别。接着,自适应方式应运而生,双流自适应图卷积网络(Two-Stream Adaptive Graph Convolutional Networks,2S-AGCN)[8]学习基于注意力机制的自适应结构,将网络中图的拓扑结构进行单独学习,增加灵活性以适应各种数据样本,并融合骨骼数据一阶和二阶信息进行建模。构建自适应图的信息只来自空间维度,因此很难有效地提取包含空间、时间和通道维度信息的关节之间的复杂连接。CTR-GCN[19]通道拓扑细化图卷积模型,非共享拓扑结构以及动态卷积的方式学习不同通道维度的拓扑特征并聚合。

1.2 最短路径距离

由结点表示的图结构与其邻接矩阵的集合,可以记为$G=(V,A)$,其中G代表完整的图结构,A是图结构的邻接矩阵。图1结点图结构表示为$V=\{v_i|i=1,2,3,4\}$,也可以表示为人体的拓扑结构,节点代表关节,节点间的连接线表示骨骼。邻接矩阵$\boldsymbol{A}$如图2(a),矩阵元素为节点之间的连接关系,1与0分别表示相连和不相连。

最短路径距离计算中会使用距离矩阵$\boldsymbol{D}$如图2(b)所示,表示任意两点之间的距离。路由矩阵$\boldsymbol{R}$如图2(c)所示,记录任意两点之间的路径关系。根据邻接矩阵$\boldsymbol{A}$,如果v_i到v_j是相连的,则距离矩阵$\boldsymbol{D}[i,j]=1$,否则为无穷大,用inf表示。$\boldsymbol{R}[i,j]$表示v_i到v_j需要经过的点,初始化$\boldsymbol{R}[i,j]=j$,即默认v_i到v_j之间是相互连通的,每列数等于该列的数目。把各个节点v_k分别插入图中,比较插入后与原来的距离,若从i经过k到j的距离比直接到达距离短,则更新$\boldsymbol{D}[i,j]=\boldsymbol{D}[i,k]+\boldsymbol{D}[k,j]$,$\boldsymbol{R}[i,j]=k$。$\boldsymbol{R}$中包含两点之间最短距离信息,$\boldsymbol{D}$中包含最短路径的信息。

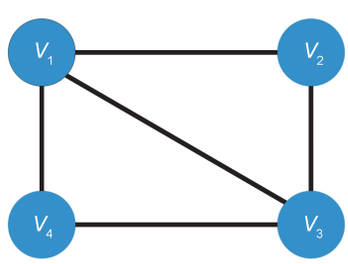


图1 结点的图结构

Fig. 1 Node graph structure

2 基于时空增强双分支图卷积的行为识别网络

2.1 骨骼数据

基于骨骼的行为识别中,人体骨骼可以表示为以关节作为顶点,骨骼作为边的图。这个图表示为$\mathcal{G}=(\mathcal{V},\mathcal{E},\mathcal{X})$,其中$\mathcal{V}=\{v_1,v_2,\cdots,v_N\}$是$N$个关节顶点的集合,$\mathcal{E}$是边的集合,邻接矩

$$\begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{array}\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{array}{cccc} v_1 & v_2 & v_3 & v_4 \end{array}$$

(a) 邻接矩阵

$$\boldsymbol{D}=\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & \inf \\ 1 & 1 & 0 & 1 \\ 1 & \inf & 1 & 0 \end{bmatrix}$$

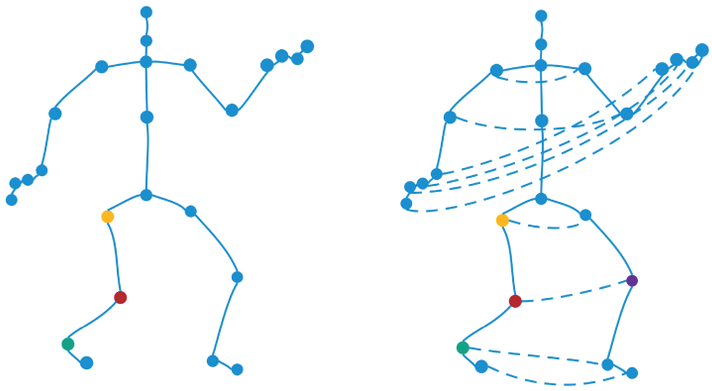
(b) 距离矩阵

$$\boldsymbol{R}=\begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

(c) 路由矩阵

图2 矩阵表示图

Fig. 2 Matrix representation diagram



(a) ST-GCN关节划分图 (b) 本文关节划分图

注:红色为根节点,黄色为离心点、绿色为近心点、紫色是对称;实线是自然的连接方式,虚线是对称关节的连接。

图3 关节点划分图

Fig. 3 Joint point division diagram

阵表示为$\boldsymbol{A}\in\boldsymbol{R}^{N\times N}$,$\boldsymbol{A}$的元素$a_{ij}$表示关节$v_i$和$v_j$之间是否相连,$v_i$的邻域表示为$\mathcal{N}(v_i)=\{v_j|a_{ij}\neq 0\}$,$\mathcal{X}$是$N$个顶点的特征的集合表示为矩阵$\boldsymbol{X}\in\boldsymbol{R}^{N\times C}$,$v_i$的特征表示为$x_i\in\boldsymbol{R}^C$。

$$H^{(l)}=\sigma(\boldsymbol{A}^{(l)}H^{(l-1)}\boldsymbol{W}^{(l)})。\tag{1}$$

其中$\boldsymbol{A}\in\boldsymbol{R}^{V\times V}$是空间聚合的邻接矩阵,$H\in\boldsymbol{R}^{V\times T\times M}$是隐藏表示,$\boldsymbol{W}\in\boldsymbol{R}^{M\times M}$是特征投影的权重矩阵。$M$表示隐藏特征的数量、$l$表示层数、$C$是通道数、$V$是关节数目、$T$表示输入的序列的长度。

2.2 关节分区策略

图卷积有优异的时空建模能力,ST-GCN首次将图卷积用于行为识别,提出将人体相邻关节点划分为3个子集的策略,如图3(a)所示。本文在此基础上加入对称关节(Symmetrical Joint, SJ)子集,如图3(b)所示。结点归属方式如式(2)。

$$l(v_j)=\begin{cases}0 & \text{if} \quad r_i=r_j\\ 1 & \text{if} \quad r_i<r_j\\ 2 & \text{if} \quad r_i>r_j\\ 3 & \text{if} \quad v_jv_i\in V_s\end{cases}。\tag{2}$$

其中r_i是第i个关节点到中心关节点的距离,V_s是所有对称点对的集合。

结合上述关节点的划分方式和公式(2),在空间维度上,顶点v_i的图卷积的公式[7]表示为:

$$f_{\text{out}}(v_i)=\sum_{v_j\in B_i}\frac{1}{Z_{ij}}f_{\text{in}}(v_j)\cdot w(l_i(v_j))。\tag{3}$$

其中$f_{\text{out}}(v_i)$是第i个关节的输出特征,B_i是通过划分方法划分的邻域集合,Z_{ij}等于相应子集中的基数,$w\left(l_i(v_j)\right)$是索引张量的加权函数。本文结合空间结构划分方式进行骨骼序列分割,将$l_i(v_j)$分为四个不同的值。

2.3 空间映射聚合(FM-SA)

基于图卷积的方法中,拓扑结构的选择至关重要。早期方法使用骨骼的固定拓扑结构描述骨骼信息,有效但是也存在局限性。目前的方法使用可以学习的邻接矩阵来捕获关节之间的关系,但是在训练学习过程中,包含大量运动信息的邻接矩阵$\boldsymbol{A}$中的拓扑无法完好的保留,造成结构的损坏。所以本文选择利用最短路径距离来描述骨骼关节之间的连接。

$$B_{ij}=e_{d_{i,j}},\tag{4}$$

$$d_{i,j}=\min_{p\in\text{Path}(\mathcal{G})}\left\{|p|,P_1=v_i,P_{|p|}=v_j\right\},\tag{5}$$

$$H^{(l)}=\sigma\left(\begin{bmatrix}(A_1+B_1)\\ \vdots\\ (A_k+B_k)\\ \vdots\end{bmatrix}H^{(l-1)}\begin{bmatrix}\boldsymbol{W}_1^{(l)} & & & \\ & \ddots & & \\ & & \boldsymbol{W}_k^{(l)} & \\ & & & \ddots\end{bmatrix}\right)。\tag{6}$$

其中权重参数B_{ij}从训练参数$E=\{e_{\text{index}}\}$中得到,根据关节对的最短路径距离$d_{i,j}$将B_{ij}分配给关节对,P_1和$P_{|p|}$分别是路径上的第一个和最后一个顶点。$\boldsymbol{W}_k\in R^{D/K\times D/K}(k=1,\cdots,K)$是权重矩阵,$H_k\in R^{V\times T\times D/K}$是隐藏层特征。

为防止学习过程中邻接矩阵发生变化造成关节位置信息的丢失,在训练过程中只对嵌入权重进行优化不对邻接矩阵处理,确保由关节距离表示的骨骼的连通信息得到保留。原始特

征信息划分为K组隐藏特征,将权重矩阵$\boldsymbol{W}_k$排列为对角矩阵的形式,相较于每个结点参与计算,只有矩阵对角块参与计算减少参数量并且投影的特征组之间是彼此独立。此外,在空间特征提取网络中加入挤压-激励(Squeeze-and-Excitation, SE)模块,它将全局的特征进行压缩后进行通道特征学习,得到各个通道的权重,再将其与原始特征图逐通道相乘,加强通道的特征提取能力,提高了模型的性能,如图4所示。

2.4 多尺度金字塔时间建模(MSP)

为了更加全面高效地提取人体行为的时间特征,采用多尺度时间建模方式,利用五个尺度的时间卷积核,能覆盖大部分尺度的感受野并且提取全粒度的时间特征信息。但是单纯的串联增加时间卷积核的数目会增加网络的深度,提高网络的运算负荷。

通过对金字塔形式的探索[20],大尺寸卷积核的深度较浅,小尺寸卷积核深度较深。所以使用多种尺度的时候,将卷积核的深度也对应进行变化,卷积核{3,5,7,9}对应的深度变化为{1,1/4,1,1/4}。

$$f_{\text{out}}=f(f_{\text{in}},f_1,f_{35},f_{79})。\tag{7}$$

其中f表示1×1的卷积、批量归一化和激活函数组成的复合函数,f_{35}和f_{79}是两个金字塔图卷积的输出特征图。如图5所示,输入特征f_{in}经过1×1卷积后分别并行经过多个卷积核通道数进行变换,经过每阶段的残差连接拼接后利用1×1卷积重塑通道之后得到输出特征f_{out}。

网络深度减少,提高了GPU并行计算的能力,通过分组卷积的方式,降低网络的参数量,增强了时间特征的提取能力,可以更加全面地捕捉到人体的短期动作细节和长期运动趋势。

2.5 双分支网络(DB)

本文网络的基本模块STEDB(Spatiotemporally Enhanced Dual-branch)主要分为三个部分,时间卷积层、金字塔时间卷积层和空间图卷积层,如图6(b)所示。

特征f_{in}首先输入到空间图卷积层(Feature Mapping and Spatial Aggregation, FM-SA),空间图卷积层中使用SE模块利用全局通道信息学习得到通道细化后的特征图,再结合自适应学习权重和分组后的权重得到输出$f_{S\text{-}GC}$,此过程强化空间特征提取能力。接着将$f_{S\text{-}GC}$分别输入到并行的时间卷积层和金字塔时间图卷积层(Multi-scale Pyramid, MSP)分支得到$f_{T\text{-}GC}$和$T(f_{S\text{-}GC})$,加强时间特征信息的提取能力。基本模块的输出是经过增强的空间信息和时间信息

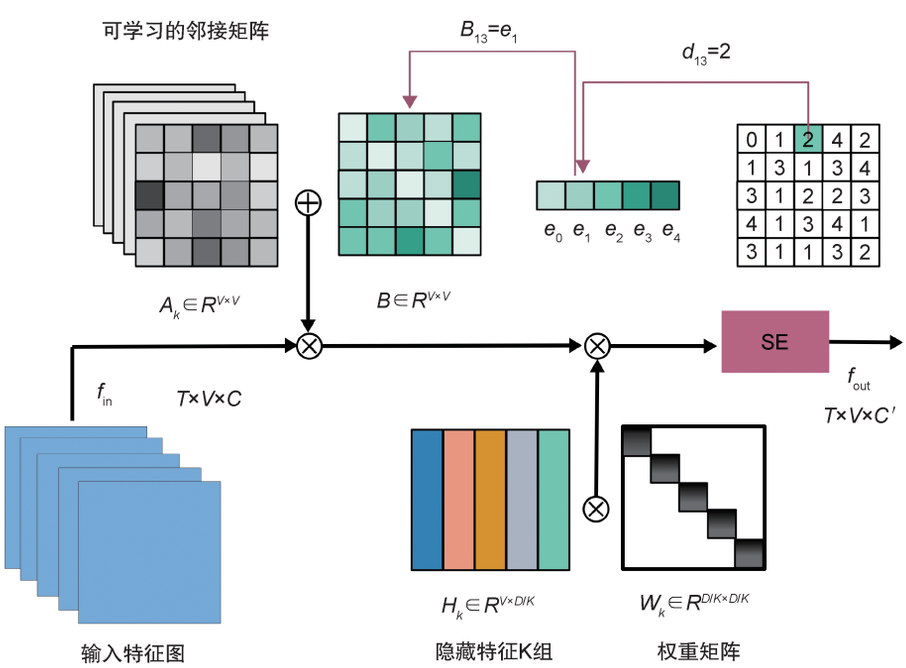
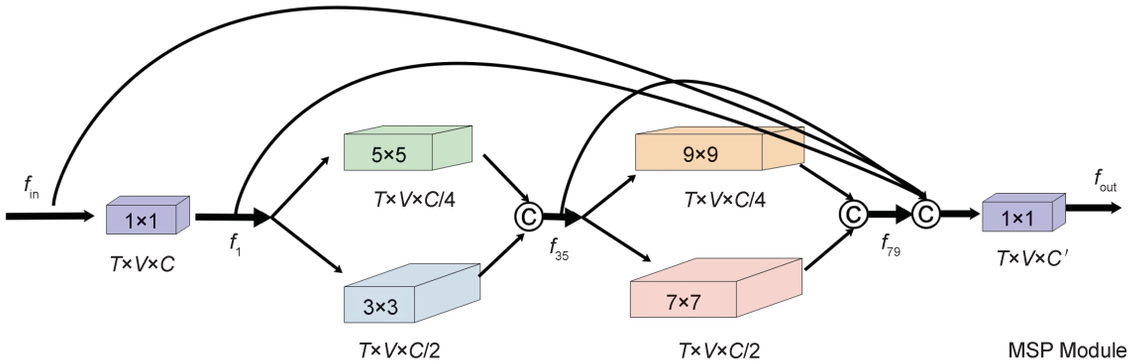


图4 空间聚合和特征映射

Fig. 4 Spatial aggregation and feature mapping



注:1×1、3×3、5×5、7×7、9×9表示卷积核大小,$T\times V\times C$是提取特征尺寸,每阶段的曲线为残差连接,©表示拼接操作。

图5 多尺度金字塔卷积模块

Fig. 5 Multi-scale pyramid convolution module

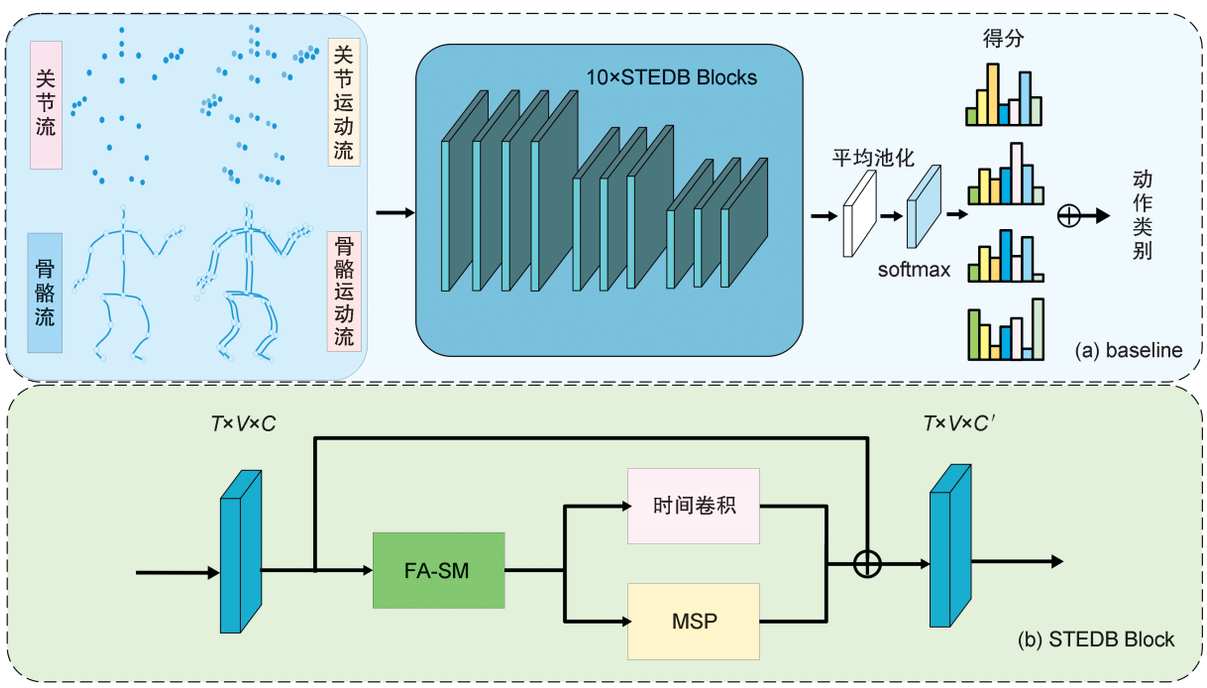


图6 时空增强双分支图卷积网络模型

Fig. 6 Spatiotemporal enhanced dual-branch graph convolution network model

融合后的结果。

$$f_{S\text{-}GC}=BN(\boldsymbol{W}f_{SE}f_{in}G^{ad})+L_{res}(f_{in}), \tag{8}$$

$$f_{T\text{-}GC}=BN(\boldsymbol{W}f_{S\text{-}GC}), \tag{9}$$

$$f_{out}=f_{T\text{-}GC}+T(f_{S\text{-}GC})。 \tag{10}$$

其中$f_{S\text{-}GC}$,$f_{T\text{-}GC}$分别表示空间图卷积层和时间图卷积层的特征输出,L_{res}表示残差连接,$\boldsymbol{W}$是分组后的权重矩阵,G^{ad}为空间卷积层中学习到的自适应图,T是基于多尺度金字塔的时间图卷积操作。

2.6 模型结构

网络模型的总体流程如图6(a)所示,首先将骨骼、关节和对应的运动流分别输入网络中,然后经过10个STEDB模块、平均池化和softmax操作,再将4个流的得分融合得到最终的分类结果。每个STEDB模块结构如图6(b),输入信息经过特征提取得到维度为$T\times V\times C$的特征,接着将其输入FA-SM强化对空间信息提取,后两个并行分支分别为普通时间图卷积和MSP模块增强时间维度特征提取能力,将增强后的特征与原始特征融合后输出。

3 实验结果分析

3.1 数据集

Northwestern-UCLA数据集包含10个动作

类别,每个动作类别由10个不同的对象执行,用来训练的数据是来自前两个Kinetic摄像头,剩余一个Kinetic摄像头数据用来进行测试。

NTU RGB+D数据集中包含60个动作类别,共计有56 880个样本。它使用两个标准来划分训练集和测试集,一是跨主题(X-sub),二是交叉视图(X-view)。

NTU RGB+D 120数据集是现有最大的人体运动3D数据集。该集合在NTU RGB+D数据集的基础上添加了60个动作类别进行扩展,且使用两个评估标准对训练集和测试集进行划分,第一跨主题(X-sub),第二交叉设置(X-set)。

3.2 参数配置

本文使用PyTorch框架在NVIDIA RTX(NVIDIA Ray Tracing Texel eXtreme)A6000 GPU上进行所有实验,部分超参数沿用基准模型CTR-GCN进行设置。模型使用随机梯度下降(Stochastic Gradient Descent,SGD)优化器训练了75个迭代周期,动量设置为0.9,并且在三个数据集训练模型时,我们在前5个迭代周期中采用了热身策略。在NTU RGB+D和NTU RGB+D 120数据集的权重衰减为0.000 4,批量大小设置为64,在Northwestern-UCLA数据集的衰减权重是0.000 1,批量大小设置为16。本文实验中采用交叉熵损失,学习率设置为0.1,在35、55、70个迭代周期阶段把学习分别率缩减10倍。

3.3 模型对比实验

为了验证本文提出模型的有效性并使结果能够相对公平,我们利用关节、骨骼、关节运动和骨骼运动四个流进行了融合实验,在NTU RGB+D、NTU RGB+D 120和NW-UCLA数据集上和主流方法进行比较。

由表1可知,在NTU RGB+D数据集X-sub和X-view划分标准中,我们实验的准确率相比基准模型CTR-GCN分别提升了0.1%和0.2%,比最新的方法分别提升0.2%,0.1%,验证了本文模型具有较好的竞争优势。在NTU RGB+D 120数据集X-Sub和X-Set的划分标准下的准确率比基准模型分别提升0.7%和0.5%,比最新的方法分别提高了0.4%和0.2%,进一步验证了本文模型的有效性。

由表2可知,在NW-UCLA数据集中,本文方法与参考模型CTR-GCN相比识别精度上升1.5%,比最优型提升0.4%,提升效果较为显著。

本文实验结果在准确率最优的同时,参数量并没有显著增加。相较于CTR-GCN的方法,本文模型参数量减少2.3 M,与最新的方法相比高了0.6 M,综合三个数据集上的实验表现,可知本文模型能够在不同数据集上表现出良好的性能。

3.4 消融实验

模型主要是基于骨骼、关节、骨骼运动和关节运动四个流进行融合得到最终的效果,我们

表1 NTU RGB+D和NTU RGB+D 120骨架数据集上的最新方法的识别准确率比较

Table 1 Comparison of recognition accuracy against state-of-the-art methods on NTU RGB+D and NTU RGB+D 120 Skeleton dataset

方法	NTU RGB+D		NTU RGB+D 120		参数量/M
	X-sub/%	X-view/%	X-sub/%	X-set/%	
ST-LSTM[21]	69.2	77.7	55.7	57.9	—
SGN[22]	—	—	79.2	81.5	0.69
ST-GCN[7]	81.5	88.3	—	—	3.1
2s-AGCN[8]	88.5	95.1	82.9	84.9	6.9
Shift-GCN[17]	90.7	96.5	85.9	87.6	2.8
Dynamic GCN[23]	91.5	96.2	87.3	88.6	14.4
CTR-GCN[19]	92.4	96.4	88.9	90.3	4.5
EfficientGCN-B4[24]	91.7	95.7	88.3	89.1	2.0
FF-TMN[25]	91.2	96.4	86.8	88.2	2.5
GSTLN[26]	91.9	96.6	88.1	89.3	1.5
InfoGCN[27]	92.3	96.5	89.2	90.6	1.6
STEDB-GCN(Ours)	**92.5**	**96.6**	**89.6**	**90.8**	**2.2**

注:—表示该方法的此项数据未知。

采用NTU RGB+D 120数据集的X-sub划分标准进行各个流的测试。

表2 在Northwestern-UCLA数据集上识别准确率方法比较

Table 2 Comparison of recognition accuracy methods on the Northwestern-UCLA dataset

方法	Northwestern-UCLA/%	参数量/M
Lie Group[28]	74.2	—
AGC-LSTM[29]	93.3	—
Shift-GCN[17]	94.6	2.8
CTR-GCN[19]	95.1	4.5
Ta-GCN[30]	96.1	—
GSTLN[26]	94.8	1.5
infoGCN[27]	96.1	1.6
STEDB-GCN(Ours)	96.6	2.2

注:—表示该方法的此项数据未知。

表3 本文模型在不同流上的准确率(%)比较

Table 3 Accuracy (%) comparison of our models on different streams

流	CTR-GCN/%	STEDB-GCN (Ours)/%
Bone	85.7	87.0
Joint	84.9	85.4
Bone motion	81.2	82.5
Joint motion	81.4	82.4

由表3可知,本文模型在四个流的准确率相比于CTR-GCN分别提高了1.3%、0.6%、1.3%和1%,因为本文方法不仅在原始模型CTR-GCN的基础上增加了拓扑连接方式,也在时间图卷积层和空间图卷积层进行了改进,能够更好地提取骨骼的时空信息。实验结果也验证了本文模型对于每个流的有效性和适用性。

为了验证所提出的不同模块的有效性,以CTR-GCN作为基准,分别对几个模块进行验证,实验结果如表4所示。本文提出的各改进模块与CTR-GCN比较,在参数量和识别精度上均有改善,特别是在对称关节(SJ)、特征聚合映射(FM-SA)、多尺度金字塔(MSP)三个模块上有显著的提升,进一步验证了所提模型的有效性和正确性。

表4 NTU RGB+D和NTU RGB+D 120数据集X-sub下消融实验的单流识别准确率(%)

Table 4 Accuracy (%) of single flow recognition in ablation experiments under X-sub with NTU RGB+D and NTU RGB+D 120 data sets

方法	参数量/M	NTU RGB+D 120 (Bone)/%	NTU RGB+D (joint-motion)/%
CTR-GCN(baseline)	4.5	85.69	87.88
Ours w/o SJ	2.2	86.96	88.45
Ours w/o SE	2.2	85.84	88.11
Ours w/o FM-SA	3.0	86.97	88.39
Ours w/o MSP	1.9	86.07	88.29
Ours w/o DB	2.3	85.95	87.98

为了验证模型的稳定性,对NTU RGB+D、NTU RGB+D 120数据集X-sub和NW-UCLA数据集下不同的模态分别进行3次实验,并计算出平均精确度和对应方差。方差都小于0.05,可知网络模型是稳定的,本文选取精确度最高的作为最终结果,如表5所示。

3.5 可视化结果

NTU RGB+D数据集的原始骨骼序列有25个关节点,邻接矩阵可视化如图7所示,图中颜色越深表示两节点之间关系越密切,印证了对称关节信息对于动作的表示相关程度的重要性。

对模型进行评价时,不仅要考虑精度也需要考虑收敛的速度,所以将四个流的准确率和损失函数利用折现图直观地体现出来。本文设置的35,55,70分别为学习率变化的epoch节点,如图8所示,可以清晰地看出模型迭代过40次之后便趋向于收敛趋势。

4 结论

本文提出了用于骨骼行为识别的双分支时空增强网络。该网络包含更为合理关节点区域的划分,加入对称关节点的分区,提取的信息

表5 模型精度(%)的稳定性验证

Table 5 Verification of stability on the accuracy (%) of the models

数据集	模态	1/%	2/%	3/%	平均精确度/%	方差
NTU RGB+D 120	X-sub(joint)	84.9	85.2	85.4	85.17	0.042
NTU RGB+D	X-sub(bone-motion)	87.4	87.6	87.7	87.57	0.016
NW-UCLA	Bone	92.1	92.2	92.5	92.27	0.029

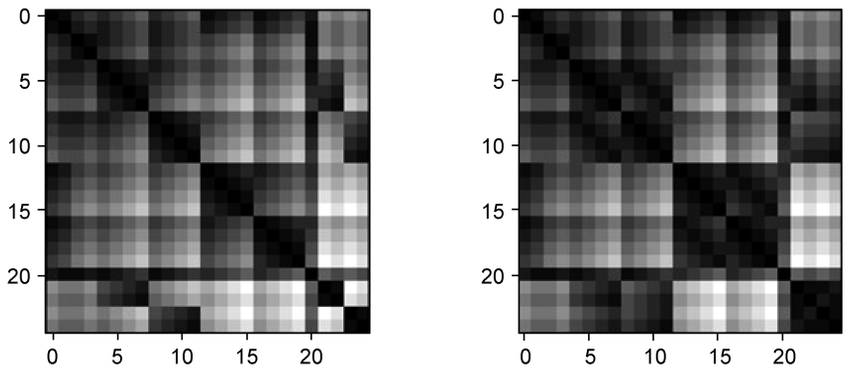


图7 邻接矩阵对比图

(横纵作坐标分别对应25个关节点的编号。)

Fig. 7 Comparison chart of adjacency matrix

(The horizontal and vertical coordinates correspond to the numbers of the 25 nodes.)

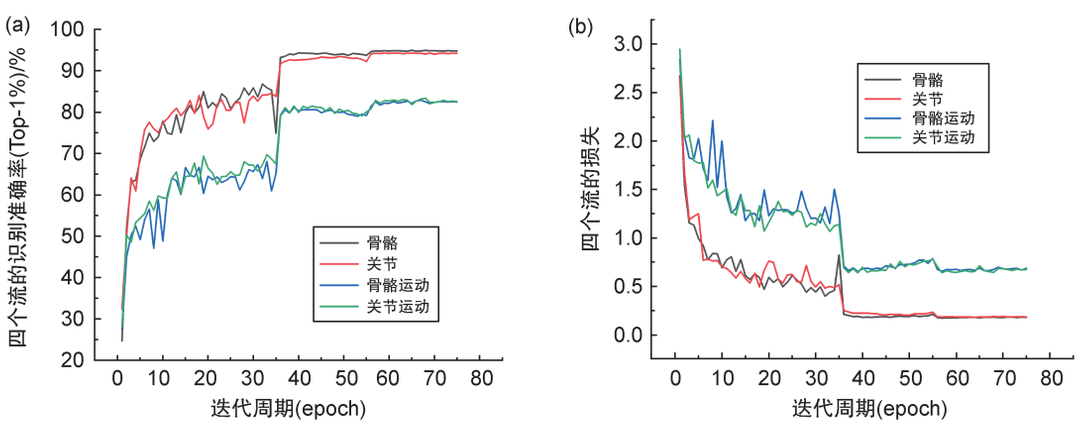


图8 NTU RGB+D 数据集X-view下的迭代准确率(Top-1%) (a)和迭代损失图(b)

Fig. 8 Iterative accuracy (Top-1%) (a) and iterative loss graph (b) under NTU RGB+D dataset X-view

更加细致。加入最短距离和权重矩阵分组,能够减少特征信息模型学习过程中的损失和冗余,挤压激励模块能够更准确地刻画通道之间的关系,提升模型的表达能力。采用多尺度金字塔模型,多个尺度有利于提取特征的上下文信息,能捕获不同尺度的关键特征进而为网络提供更加丰富的特征表示。在原始的网络中加入一个时间建模分支,进一步增强对骨骼序列时间特征的提取能力。本文工作主要体现在时空增强和参数量减少,并且在三个数据集上的表现都较优异。未来的主要工作可以进一步减少网络参数量,提高识别的准确率。

参考文献:

[1] WEINLAND D, RONFARD R, BOYER E. A Survey of Vision-based Methods for Action Representation, Segmentation and Recognition[J]. *Comput Vis Image Underst*, 2011, **115**(2): 224−241. DOI: 10.1016/j.cviu.2010.10.002.

[2] POPPE R. A Survey on Vision-based Human Action Recognition[J]. *Image Vis Comput*, 2010, **28**(6): 976−990. DOI: 10.1016/j.imavis.2009.11.014.

[3] MOCCIA S, MIGLIORELLI L, CARNIELLI V, *et al*. Preterm Infants' Pose Estimation with Spatio-temporal Features[J]. *IEEE Trans Biomed Eng*, 2020, **67**(8): 2370−2380. DOI: 10.1109/TBME.2019.2961448.

[4] DU Y, WANG W, WANG L. Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2015: 1110−1118. DOI: 10.1109/CVPR.2015.7298714.

[5] LEE I, KIM D, KANG S, *et al*. Ensemble Deep Learning for Skeleton-based Action Recognition Using Temporal Sliding LSTM Networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2017: 1012−1020. DOI: 10.1109/ICCV.2017.115.

[6] KE Q H, BENNAMOUN M, AN S J, *et al*. A New Rep-

resentation of Skeleton Sequences for 3D Action Recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 4570−4579. DOI: 10.1109/CVPR.2017.486.

[7] YAN S J, XIONG Y J, LIN D H. Spatial Temporal Graph Convolutional Networks for Skeleton-based Action Recognition[J]. *Proc AAAI Conf Artif Intell*, 2018, **32**(1): 7444−7452. DOI: 10.1609/aaai.v32i1.12328.

[8] 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 (CVPR). New York: IEEE, 2019: 12018−12027. DOI: 10.1109/CVPR.2019.01230.

[9] LI F J, ZHU A C, XU Y G, *et al*. Multi-stream and Enhanced Spatial-temporal Graph Convolution Network for Skeleton-based Action Recognition[J]. *IEEE Access*, 2020, **8**: 97757−97770. DOI: 10.1109/ACCESS.2020.2996779.

[10] SHI L, ZHANG Y F, CHENG J, *et al*. Skeleton-based Action Recognition with Multi-stream Adaptive Graph Convolutional Networks[J]. *IEEE Trans Image Process*, 2020, **29**: 9532-9545. DOI:10.1109/TIP.2020.3028207.

[11] BAI R W, LI M, MENG B, *et al*. Hierarchical Graph Convolutional Skeleton Transformer for Action Recognition[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2022: 1−6. DOI: 10.1109/ICME52920.2022.9859781.

[12] CHENG K, ZHANG Y F, CAO C Q, *et al*. Decoupling GCN with Drop Graph Module for Skeleton-based Action Recognition[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 536−553. DOI: 10.1007/978-3-030-58586-0_32.

[13] DUTA I C, LIU L, ZHU F, *et al*. Pyramidal Convolution: Rethinking Convolutional Neural Networks for Visual Recognition[EB/OL]. (2020−6−20) [2024−02−24]. http://arxiv.org/abs/2006.11538.

[14] JIA N, TIAN X L, ZHANG Y, *et al*. Semi-supervised Node Classification with Discriminable Squeeze Excitation Graph Convolutional Networks[J]. *IEEE Access*, 2020, **8**: 148226−148236. DOI: 10.1109/ACCESS.2020.3015838.

[15] MONTI F, BOSCAINI D, MASCI J, *et al*. Geometric Deep Learning on Graphs and Manifolds Using Mixture Model CNNS[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 5425−5434. DOI: 10.1109/CVPR.2017.576.

[16] ALSARHAN T, ALI U, LU H T. Enhanced Discriminative Graph Convolutional Network with Adaptive Temporal Modelling for Skeleton-based Action Recognition [J]. *Comput Vis Image Underst*, 2022, **216**: 103348. DOI: 10.1016/j.cviu.2021.103348.

[17] 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 (CVPR). New York: IEEE, 2020: 180−189. DOI: 10.1109/CVPR42600.2020.00026.

[18] WEN Y H, GAO L, FU H B, *et al*. Graph CNNS with Motif and Variable Temporal Block for Skeleton-based Action Recognition[J]. *Proc AAAI Conf Artif Intell*, 2019, **33**(1): 8989−8996. DOI: 10.1609/aaai.v33i01.33018989.

[19] 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 (ICCV). New York: IEEE, 2021: 13339−13348. DOI: 10.1109/ICCV48922.2021.01311.

[20] LI F J, ZHU A C, LIU Z Y, *et al*. Pyramidal Graph Convolutional Network for Skeleton-based Human Action Recognition[J]. *IEEE Sens J*, 2021, **21**(14): 16183−16191. DOI: 10.1109/JSEN.2021.3075722.

[21] LIU J, SHAHROUDY A, XU D, *et al*. Spatio-temporal LSTM with Trust Gates for 3D Human Action Recognition[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 816−833. DOI: 10.1007/978-3-319-46487-9_50.

[22] 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 (CVPR). New York: IEEE, 2020: 1109−1118. DOI: 10.1109/CVPR42600.2020.00119.

[23] 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. ACM. New York: Association for Computing Machinery, 2020: 55−63. DOI: 10.1145/3394171.3413941.

[24] SONG Y F, ZHANG Z, SHAN C F, *et al*. Constructing Stronger and Faster Baselines for Skeleton-based Action Recognition[J]. *IEEE Trans Pattern Anal Mach Intell*, 2023, **45**(2): 1474−1488. DOI: 10.1109/TPAMI.2022.3157033.

[25] LI F J, ZHU A C, LI J J, *et al*. Frequency-driven Channel Attention-augmented Full-scale Temporal Modeling Network for Skeleton-based Action Recognition[J]. *Knowl Based Syst*, 2022, **256**: 109854. DOI: 10.1016/j.knosys.2022.109854.

[26] DAI M, SUN Z H, WANG T Y, *et al*. Global Spatio-temporal Synergistic Topology Learning for Skeleton-based Action Recognition[J]. *Pattern Recognit*, 2023,

140: 109540. DOI: 10.1016/j.patcog.2023.109540.

[27] HUANG X, ZHOU H, WANG J, *et al*. Graph Contrastive Learning for Skeleton-Based Action Recognition[EB/OL]. (2023−6−10)[2024−04−9]. http://arxiv.org/abs/2301.10900.

[28] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2014: 588−595. DOI: 10.1109/CVPR.2014.82.

[29] SI C Y, CHEN W T, WANG W, *et al*. An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-based Action Recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 1227−1236. DOI: 10.1109/CVPR.2019.00132.

[30] XU K L, YE F F, ZHONG Q Y, *et al*. Topology-aware Convolutional Neural Network for Efficient Skeleton-based Action Recognition[J]. *Proc AAAI Conf Artif Intell*, 2022, **36**(3): 2866−2874. DOI: 10.1609/aaai.v36i3.20191.