·开发研究与工程应用·　　文章编号：1000-3428(2023)10-0264-08　　文献标志码：A　　中图分类号：TP391

基于自适应多尺度图卷积网络的骨架动作识别

刘　宽[1]，奚小冰[2]，周明东[1]

（1.上海交通大学 机械与动力工程学院 上海市复杂薄板结构数字化制造重点实验室，上海 200240；
2.上海交通大学医学院附属瑞金医院，上海 200240）

摘　要：将人体骨架建模为时空拓扑图的图卷积网络在基于人体骨架数据的动作识别任务中得到了广泛应用。但现有图卷积网络存在预定义骨架拓扑图拓扑结构固定、单支路时间图卷积算子提取时空特征粒度单一的问题，极大限制了模型的泛化能力和表达能力。提出基于自适应多尺度图卷积网络的人体骨架动作识别模型，自适应空间图卷积层将骨架的拓扑结构作为参数进行端到端的自适应学习，根据动作生成数据驱动的骨架拓扑图。多尺度时间图卷积层对时间图卷积算子进行多支路扩展，动态融合骨架序列不同时间粒度的时空特征。综合骨架关节点、骨架长度、骨架关节点运动、骨架长度运动4路信息输入模型。实验结果表明，所提模型在NTU RGB+D 60动作识别数据集下的人物划分(CS)模式和视角划分模式实验中分别取得90.5%和96.8%的识别准确率，在NTU RGB+D 120动作识别数据集的CS模式和设置划分模式的实验中分别取得86.0%和88.7%的识别准确率，能有效提取骨架动作的时空特征，提升了人体骨架动作识别的分类性能。

关键词：人体骨架；动作识别；图卷积网络；自适应；多尺度

开放科学（资源服务）标志码（OSID）：

中文引用格式：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.

英文引用格式：LIU K，XI X B，ZHOU M D. Skeleton action recognition based on adaptive multi-scale graph convolutional network[J]. Computer Engineering，2023，49(10)：264-271.

Skeleton Action Recognition Based on Adaptive Multi-scale Graph Convolutional Network

LIU Kuan[1]，XI Xiaobing[2]，ZHOU Mingdong[1]

（1.Shanghai Key Laboratory of Digital Manufacture for Thin-Walled Structures，School of Mechanical Engineering，Shanghai Jiao Tong University，Shanghai 200240，China；
2.Ruijin Hospital，Shanghai Jiao Tong University School of Medicine，Shanghai 200240，China）

【Abstract】The Graph Convolution Network（GCN）is widely used for skeleton-based recognition methods，whereby a predefined spatiotemporal topology graph is used to model human skeletal features. However，the existing GCN has limitations in two aspects：the predefined skeleton topology structure limits the generalization ability of the model，and single-granular spatiotemporal features limit the capacity of the model. To solve the aforementioned problems，an adaptive multi-scale graph convolution network is proposed. The adaptive spatial graph convolution layer regards the topology of the skeleton as parameters to optimize，thereby generating data-driven skeleton topology by samples. The multiscale temporal graph convolution layer uses various temporal graph convolution kernels to dynamically integrate the multi-granular spatiotemporal features. Extensive experiments were conducted by combining joint，bone，joint motion，and bone motion streams as inputs of the proposed model. The experimental results show that the classification accuracy of the proposed model under NTU RGB+D 60 action recognition data set for the Cross-Subject（CS）and Cross-View（CV）subsets was 90.5% and 96.8%，respectively，and the classification accuracy of the NTU RGB+D 120 action recognition dataset for the CS and Cross-Setup（CT）subsets was 86.0% and 88.7%，respectively. The model can effectively extract the spatiotemporal features of skeleton motion，thereby improving the classification performance of human skeleton motion recognition.

【Key words】human skeleton；action recognition；Graph Convolution Network（GCN）；adaption；multi-scale

DOI：10. 19678/j. issn. 1000-3428. 0065882

基金项目：上海交通大学医工交叉重点项目（YG2019ZDA16）。

作者简介：刘　宽（1996—），男，硕士研究生，主研方向为模式识别；奚小冰，主任医师、硕士；周明东（通信作者），副教授、博士。

收稿日期：2022-09-29　　**修回日期：**2022-11-18　**E-mail：**liukuan5625@sjtu.edu.cn

0　概述

人体动作识别广泛应用于人机交互、运动辅助、行为检测等领域。现有研究对表示人体动作的各类数据模态进行了探索，如RGB图像、深度图像、光流、人体骨架等。在这些数据模态中，人体骨架数据只包含骨架关节点的二维或三维空间坐标，高度抽象的动作表达对动态环境和复杂背景具有更强的鲁棒性[1]。同时，由于运动传感器、3D深度相机和人体姿态估计算法的不断发展，基于骨架数据的动作识别方法吸引了愈来愈多的学者进行研究[2-6]。

早期的人体骨架动作识别方法多采用手工设计特征方法捕捉骨架关节点之间的相对位置关系[7-9]，该方法主要依赖关节点之间的相对平移和旋转提取骨架序列的时空特征，设计复杂的特征提取器限制模型性能[10]。近年来，由于深度学习算法的不断发展，基于深度学习的动作识别方法受到广泛关注。基于深度学习方法的人体骨架动作识别方法主要分为3类：基于循环神经网络的骨架数据动作识别方法[11-17]、基于卷积神经网络（Convolutional Neural Network，CNN）的骨架数据动作识别方法[18-22]以及基于图神经网络（Graph Convolution Network，GCN）的骨架数据动作识别方法[23-34]。基于循环神经网络的骨架数据动作识别方法将骨架序列逐帧编码为向量后输入到循环神经网络单元或长短期记忆网络单元中，学习序列间的时空特征以进行动作识别。基于卷积神经网络的骨架数据动作识别方法首先将骨架数据预处理转化为伪图像，然后采用卷积神经网络对伪图像进行多尺度的特征提取和分类。但是，上述两类动作识别方法将骨架数据转化为向量或伪图像的建模方式忽略了人体骨架的自然拓扑结构，识别效果有限。基于图神经网络的骨架数据动作识别方法的图神经网络将人体骨架的拓扑结构定义为邻接矩阵，使用图结构对骨架序列进行建模以提取人体的时空运动特征。YAN等[23]将图卷积网络（ST-GCN）应用于基于人体骨架数据的动作识别任务中，将自然人体骨架的拓扑结构和时空域中的时间运动依赖性定义为稀疏连接的时空无向图，模型分别使用空间图卷积算子和时间图卷积算子学习骨架序列的空间运动特征和时间运动特征。SHI等[24-25]改进图卷积网络，提出自适应注意力模块学习骨架的拓扑结构图，并将骨架关节点和骨架长度一同作为模型的输入，提升了图卷积网络在动作识别任务中的分类性能。SHI等[26]基于自然人体中关节和骨架之间的运动依赖关系，将骨架数据表示为有向无环图并设计了一种有向图神经网络，提取骨架序列运动的特征信息。LIU等[27]提出一种时间和空间分离的多尺度图卷积算子和时间和空间统一的时空图卷积算子，通过多尺度聚合方法实现有效的动作特征提取。DING等[28]提出语义引导图卷积网络，使用拓扑结构图提取模块、动作图推理模块和注意图迭代模块聚合特征信息并捕获动作的潜在依赖关系。孙琪翔等[29]设计基于图卷积网络的非局部网络模块，有效获取全局特征信息，从而提高网络识别准确率。CHEN等[30]提出通道拓扑细化图卷积模块（CTR-GC）以学习骨架拓扑结构并聚合不同通道的骨架关节点特征，实现基于骨架的动作识别。王小娟等[31]基于空间注意力机制和通道注意力机制，对骨架数据的动作特征进行多粒度卷积和动态融合。ZHANG等[32]提出SATD-GCN网络，基于空间自注意力模块和时间扩展图卷积模块，有效减轻了数据冗余问题并增强模型鲁棒性。TU等[33]采用骨架关节点与骨架长度融合的关系驱动图卷积网络作为特征提取器并采用姿势预测模块实现半监督学习。BIAN等[34]提出人物关系图卷积模块，以自监督的方式学习群体活动识别的骨架动作特征。

上述图卷积网络大多使用预定义的固定拓扑图表达人体不同部位之间的连接关系，但是人体在执行不同动作时，各个部位之间具有不同的关系，比如“脱鞋”动作需要手和脚相互配合才能完成，而上述使用预定义拓扑图的图卷积网络很难捕捉两者之间的关系，因为预定义的拓扑图中手与脚并不存在直接的连接关系。此外，固定的拓扑图无法根据动作调节人体各部位之间连接关系的强弱，例如“脱帽”动作手和头之间应该具有更强的关系，而“踢球”动作则应弱化两者之间的连接关系。上述方法在研究骨架数据的时间维度建模时多采用固定尺寸的单支路时间图卷积算子，但由于各类动作的持续时间、重复次数不同，例如“鼓掌”是重复动作，每次“鼓掌”只持续很短的时间，而“脱鞋”是持续时间很长的单次动作，因此固定尺寸的单支路时间图卷积算子难以捕捉具有不同时空粒度的动作特征。

为了解决上述问题，本文提出基于自适应多尺度图卷积网络的人体骨架动作识别方法，自适应多尺度图卷积网络包括自适应空间图卷积层和多尺度时间图卷积层。自适应空间图卷积层基于自注意力机制，以数据驱动的方式动态构建骨架的拓扑结构，根据动作输入网络自行调节人体各部位的连接关系。多尺度时间图卷积层对单支路时间图卷积核进行多支路扩展，从而捕获动作的多粒度时空特征，使得模型具有多尺度的时空感受野。将骨架关节点、骨架长度、骨架关节点运动、骨架长度运动四路信息作为模型输入，并在NTU RGB+D 60动作识别数据集和NTU RGB+D 120动作识别数据集上开展实验验证所提方法的有效性。

1　图卷积网络

1.1　人体骨架序列

人体骨架序列由每帧中每个人体骨架关节点的二维或三维坐标表示。在具有N个关节点和T帧的骨架序列上构造无向时空图$G=(V,E)$，其中$V=$

$\{v_{ti}|t=1,2,\cdots,N\}$表示骨架序列中的所有骨架关节点，$v_{ti}$代表第$t$帧的第$i$个骨架点。根据骨架结构的自然具有的连接特性，建立同一帧内各个骨架点之间的连接关系。骨架边集E由连接同一帧内人体各个骨架点的骨架边集$E_S=\{v_{ti}v_{tj}|(i,j)\in H\}$和连接前后两帧之间相同骨骼点的骨架边集$E_F=\{v_{ti}v_{(t+1)i}\}$的两个子集组成，$H$表示人体骨架点集。

1.2 空间图卷积

图卷积网络是针对具有图结构数据设计的基于神经网络的特征提取算法，被广泛应用于有图性质的推荐系统、社交网络、交通预测等任务中[35-37]。图卷积网络通过式(1)计算：

$$\boldsymbol{f}_{\text{out}}=\boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{f}_{\text{in}}\otimes W \tag{1}$$

空间图卷积算子对骨架序列的第τ帧进行图卷积操作，如图1黑色框线所示，该帧包含有N个骨架点的骨架点集V_τ和连接该帧内各个骨架点的骨架边集$E_S(\tau)=\{v_{ti}v_{tj}|\tau=t,(i,j)\in H\}$，空间图卷积算子表达式为：

$$\boldsymbol{f}_{\text{out}}=\boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{f}_{\text{in}}\otimes \boldsymbol{W}_S \tag{2}$$

其中：$\boldsymbol{f}_{\text{in}}$表示维度为$C_{\text{in}}\times T\times N$的输入骨架序列；$\boldsymbol{f}_{\text{out}}$表示维度$C_{\text{out}}\times T\times N$的输出骨架序列；$\otimes$为卷积操作；$\boldsymbol{W}_S$表示维度为$C_{\text{in}}\times C_{\text{out}}\times 1\times 1$的空间图卷积核；$\boldsymbol{A}$是维度$N\times N$的邻接矩阵，表达式如式(3)所示：

$$\boldsymbol{A}[i,j]=\begin{cases}0, v_{ti}v_{tj}\notin E_S(\tau)\\ 1, v_{ti}v_{tj}\in E_S(\tau)\end{cases} \tag{3}$$

式(2)中的$\boldsymbol{\Lambda}$为防止图卷积网络在反向传播时出现梯度消失或梯度爆炸问题的度矩阵，表达式为：

$$d(v_i)=\sum_j \boldsymbol{A}[i,j] \tag{4}$$

$$\boldsymbol{\Lambda}=\text{diag}(d(v_1),d(v_2),\cdots,d(v_N)) \tag{5}$$

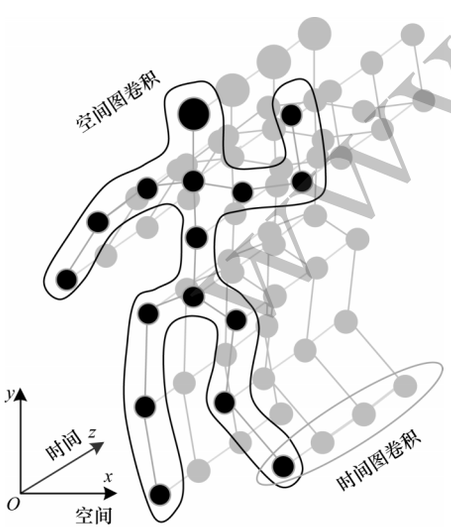


图1 空间图卷积算子和时间图卷积算子

Fig.1 Spatial graph convolution operator and temporal graph convolution operator

1.3 时间图卷积

时间图卷积算子对骨架序列的第ι个骨架点进行图卷积操作，如图1浅色框线所示，包含T帧中所有该骨架点组成的点集V_ι和前后连接该骨架点的骨架边集$E_F=\{v_{ti}v_{(t+1)i}|i=\iota\}$。时间图卷积算子和空间图卷积算子类似，时间图卷积算子表达式为：

$$\boldsymbol{f}_{\text{out}}=\boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{f}_{\text{in}}\otimes \boldsymbol{W}_F \tag{6}$$

其中：$\boldsymbol{W}_F$表示维度为$C_{\text{in}}\times C_{\text{out}}\times T\times 1$的时间图卷积核，是时间图卷积核的训练参数。

2 自适应多尺度图卷积网络

2.1 自适应空间图卷积层A-GCN

本文基于自注意力机制[24-25,38]提出自适应空间图卷积层(A-GCN)，根据骨架序列输入样本对骨架拓扑结构进行数据驱动的自适应优化，将骨架的拓扑结构作为参数与空间图卷积内核一起作为参数进行学习，流程如图2所示。

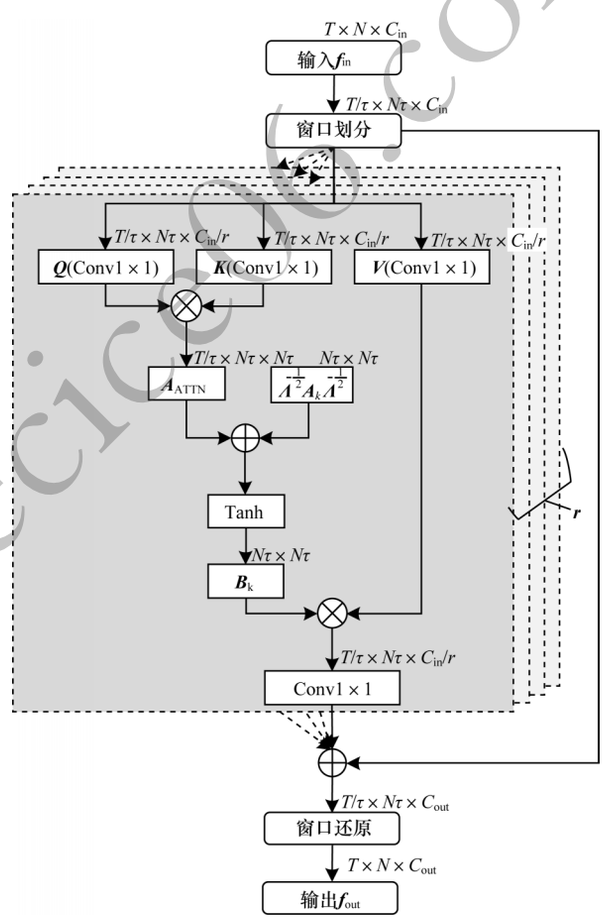


图2 自适应空间图卷积层的流程

Fig.2 Procedure of adaptive spatial graph convolution network

首先，按每τ帧长度将输入骨架序列$\boldsymbol{f}_{\text{in}}$划分为$T/\tau$个不重叠的维度为$N\times\tau\times C_{\text{in}}$的时空窗口，自适应空间图卷积层分别在每个时空窗口内计算τ帧内所有骨架关节点之间的连接关系。

维度为 $C_{in}\times C_{in}/r\times 1\times 1$ 的空间图卷积核 $\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$ 将窗口划分后的输入 $\boldsymbol{f}_{in}$ 分别投射为 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 3个特征矩阵，提取骨架动作特征的同时将骨架序列的通道维度降维为 C_{in}/r 以提升计算效率。

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{f}_{in}\otimes\boldsymbol{W}_Q\\ \boldsymbol{K}=\boldsymbol{f}_{in}\otimes\boldsymbol{W}_K\\ \boldsymbol{V}=\boldsymbol{f}_{in}\otimes\boldsymbol{W}_V\end{cases}\tag{7}$$

在骨架序列的每个时空窗口内分别将 $\boldsymbol{Q}$、$\boldsymbol{K}^{\mathrm{T}}$ 矩阵相乘，得到维度为 $T/\tau\times N\times\tau\times N\times\tau$ 的自适应骨架拓扑结构邻接矩阵 $\boldsymbol{A}_{ATTN}$，将自适应骨架拓扑矩阵 $\boldsymbol{A}_{ATTN}$ 与归一化后的预定义骨架拓扑邻接矩阵 $\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}_k\boldsymbol{\Lambda}^{-\frac{1}{2}}$ 相加后使用Tanh激活函数将矩阵元素归一化到 $(-1,1)$ 之间，得到最终的骨架拓扑结构矩阵 $\boldsymbol{B}_k$，矩阵元素 $\boldsymbol{B}_k[i,j]$ 代表第 i 个骨架点和第 j 个骨架点之间连接关系的强弱，$\boldsymbol{A}_k$ 定义如图3所示。每个骨架样本在不同时空窗口内其拓扑结构都是唯一的、由输入样本生成的自适应拓扑结构矩阵，数据驱动的骨架拓扑动态调整人体各部位的连接关系。$\boldsymbol{A}_{ATTN}$、$\boldsymbol{B}_k$、Tanh函数的表达式分别如下所示：

$$\boldsymbol{A}_{ATTN}=\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{C_{in}/r}}\tag{8}$$

$$\boldsymbol{B}_k=\mathrm{Tanh}\left(\boldsymbol{A}_{ATTN}+\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}_k\boldsymbol{\Lambda}^{-\frac{1}{2}}\right)\tag{9}$$

$$\mathrm{Tanh}(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}\tag{10}$$

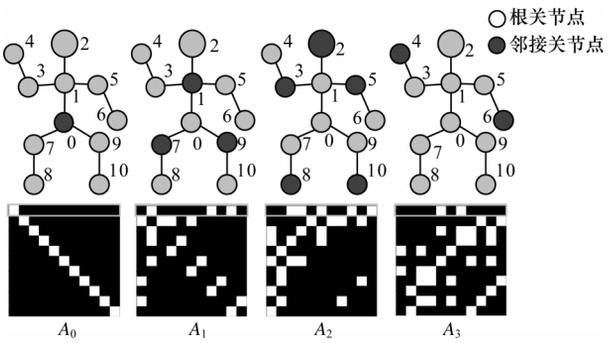


图3　k 阶邻接矩阵

Fig.3　k-hop adjacency matrix

基于多头注意力机制[38]，对骨架序列输入重复 r 次上述操作。多头注意力网络能够感知并提取骨架不同邻域范围的空间信息，加强了人体各关节点之间的联系。

骨架拓扑结构矩阵 $\boldsymbol{B}_k$ 与特征矩阵 $\boldsymbol{V}$ 相乘并经过维度为 $C_{in}/r\times C_{out}\times 1\times 1$ 的空间图卷积核 $\boldsymbol{W}$，将骨架序列的通道维度升维为 C_{out}，最后合并 r 层骨架动作特征并加入残差连接[39]，窗口还原后得到自适应空间图卷积层输出 $\boldsymbol{f}_{out}$。

$$\boldsymbol{f}_{out}=\sum_{k=0}^{r}\boldsymbol{B}_kV\otimes W+\boldsymbol{f}_{in}\tag{11}$$

2.2　多尺度时间图卷积层MS-TCN

自适应空间图卷积层的各个时空窗口之间相对闭合，信息无法进行有效交互，因此需要对骨架序列的时间维度进行特征建模。本文采用多尺度时间图卷积层(MS-TCN)将单支路时间图卷积核进行多支路扩展[27]，以捕获人体动作特征不同粒度的时空模式。多尺度时间图卷积层的算法流程如图4所示。

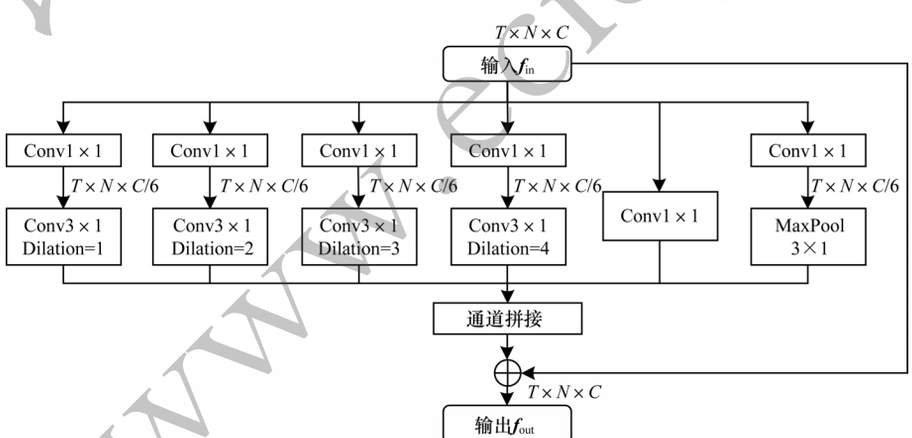


图4　多尺度时间图卷积层的算法流程

Fig.4　Algorithm procedure of multi-scale temporal graph convolution network

多尺度时间图卷积层采用瓶颈设计对骨架数据的时间维度进行建模，在6条支路分别使用维度为 $C_{in}\times C_{in}/6\times 1\times 1$ 的时间图卷积核将骨架序列的通道维度降维为 $C_{in}/6$ 以提升计算效率。其中4条支路分别使用膨胀率为1、2、3、4的维度为 $C_{in}\times C_{out}\times 3\times 1$ 的时间图卷积核，分别以3帧、6帧、9帧、12帧的采样率对骨架序列进行多种时间粒度的特征提取。为进一步扩张模型的时空感受野，加入维度为 $C_{in}\times C_{in}/6\times 1\times 1$ 时间图卷积支路和最大池化支路。最后在通道维度上对6条支路的输出进行拼接将骨架序列的通道维度升维为 C_{out}，并在输入和输出之间加入残差连接[39]得到多尺度时间图卷积层的最终输出。

2.3　时空图卷积网络总体架构

时空图卷积网络模型的整体架构如图5所示。

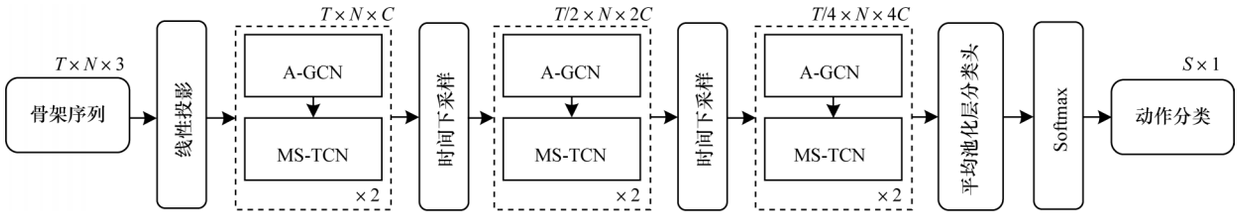


图5 时空图卷积网络模型的整体架构

Fig.5 Overall architecture of spatiotemporal graph convolutional network model

模型输入数据维度为$T\times N\times 3$的骨架序列，线性投影层将骨架序列的通道维度升维为C并将帧数T减半，使用6层串联连接的自适应空间图卷积层（A-GCN）和多尺度时间图卷积层（MS-TCN）提取骨架时空特征，在第2层尾部和第4层尾部对骨架特征的时间维度进行下采样，通道数C翻倍的同时帧数T减半。

原始骨架序列通过6层A-GCN+MS-TCN模块建立维度为$T/8\times N\times 8C$的多层次骨架动作特征表示，采用Softmax函数对经过平均池化层和分类头的骨架特征进行动作预测，最终模型输出维度为$S\times 1$的向量，该向量即代表模型最终得出S类动作的概率分布。

3 实验结果与分析

3.1 数据集

本文使用NTU RGB+D 60数据集和NTU RGB+D 120数据集。

NTU RGB+D 60数据集[41]是使用3台动作捕捉相机拍摄的动作识别任务的数据集，包含由40名10～35岁志愿者完成的共60类动作以及56 880个数据样本。每个骨架样本数据的维度为$T\times N\times 3$，其中T表示该骨架序列的总帧数，N表示每帧中的总骨架关节点数，3表示每个骨架关节点都有(x,y,z)3个坐标值。NTU RGB+D 60数据集使用人物划分（Cross-Subject，CS）和视角划分（Cross-View，CV）两种模式划分训练集和测试集，其中CS划分模式按照志愿者划分训练集和测试集，训练集包含40 320个样本，测试集包含16 560个样本；CV划分模式按照相机划分训练集和测试集，将相机2和相机3采集的37 920个样本作为训练集，相机1采集的19 960个样本作为测试集。

NTU RGB+D 120数据集[40]是对NTU RGB+D 60数据集的扩展数据集，另外增加包含60类新动作的57 367个数据样本，数据集共包含120类动作以及113 945个数据样本。NTU RGB+D 120数据集使用CS和设置划分（Cross-Setup，CT）两种模式划分训练集和测试集，其中CS划分模式按照志愿者划分训练集和测试集，训练集包含63 026个样本，测试集包含50 919个样本；CT划分模式按照相机设置方案划分训练集和测试集，训练集包含54 468个样本，测试集包含49 477个样本。

3.2 数据预处理

为将数据样本的各维度坐标归一化到同一区间，并消除特征间的相关性以便于模型学习，使用如下步骤对原始骨架序列进行预处理[27]：

1）将骨架序列的第1帧坐标值线性投影到$[-1,1]$的区间内，其余帧采取相同线性变换与第1帧进行对齐；

2）对骨架序列进行三维坐标旋转，使骨架的左右肩线平行于x轴，脊柱平行于z轴；通过重新放映骨架动作，将所有的骨架序列帧数统一填充为288帧。预处理前后的骨架序列样本对比如图6所示。

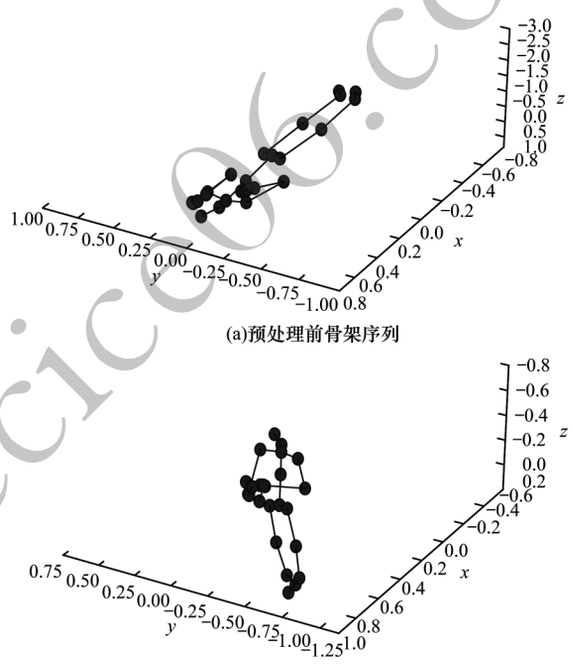


图6 原始骨架序列和预处理后骨架序列对比

Fig.6 Comparison between original skeleton sample and preprocessed skeleton sample

3.3 网络设置与实验数据预处理

使用Python语言PyTorch深度学习框架建立网络，NTU RGB+D数据集预处理方式如第3.2节所述，整体网络模型如第2.3节所述。通过大量实验最终确定使用6层A-GCN+MS-TCN搭建自适应多尺度时空图卷积网络模型，每层的通道数C分别定义为96、96、192、192、384、384；注意力头数量r定义为4；时空窗口长度τ定义为1。训练批数据大小为32，优化算法使用带动量的随机梯度下降算法，动量设

置为0.9。初始学习率为0.05，模型共训练100轮，在训练到35轮和70轮时学习率乘1/10。损失函数使用交叉熵函数，权重衰减参数设置为0.000 4。为防止模型出现过拟合，在每个模块的残差连接部分加入drop path层，drop path层随机丢弃概率设置为0.2。使用标签平滑方法加速训练同时防止网络过拟合，标签平滑参数设置为0.1。

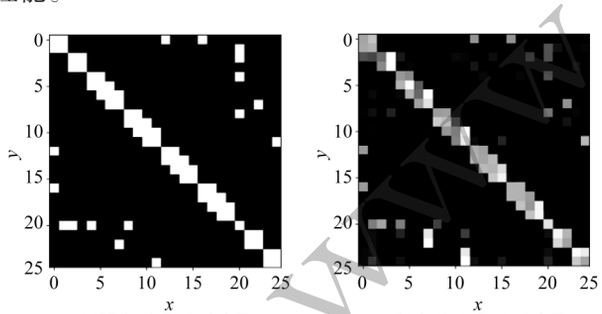
为验证本文所提的自适应空间图卷积层和多尺度时间图卷积层的有效性，在NTU RGB+D 60数据集CV划分模式下进行消融实验，网络仅输入骨架关节点坐标，不使用数据增强。实验结果如表1所示，采用自适应的空间图卷积层A-GCN和固定尺寸的单支路时间图卷积层TCN，相比于ST-GCN基线模型，模型的识别准确率提升了5.1个百分点；同时采用自适应的空间图卷积层A-GCN和多尺度时间图卷积层MS-TCN时模型识别准确率进一步提升至95.6%。

表1　NTU RGB+D 60数据集下的消融实验结果

Table 1　Ablation experiment results under the NTU RGB+D 60 data set　%

模型	输入数据类型	识别准确率
GCN+TCN(ST-GCN)	骨架关节点	88.3
A-GCN+TCN	骨架关节点	93.4
A-GCN+MS-TCN	骨架关节点	95.6

图7为预定义的骨架拓扑图和自适应的骨架拓扑图实例对比，矩阵中元素的灰度表示两骨架关节点之间连接关系的强弱。图7(a)是预定义的骨架拓扑图，图7(b)是A-GCN+MS-TCN模型在NTU RGB+D 60数据集的CV划分模式下学习到的骨架拓扑图实例。自适应的空间图卷积算子能够根据输入动作自适应调整骨架关节点之间联系的强弱，显著增强模型的表达能力，有效提升模型的分类性能。



(a)预定义骨架拓扑图　(b)自适应骨架拓扑图

图7　预定义骨架拓扑图和自适应骨架拓扑图对比

Fig.7　Comparison between predefined skeleton topology graph and adaptive skeleton topology graph

为提高算法模型的动作识别准确率，将骨架关节点、骨架长度、以及骨架关节点运动和骨架长度运动共4路信息输入模型，在NTU RGB+D 60数据集CV划分模式下进行消融实验，结果如表2所示，由表2可知，当仅使用骨架关节点作为模型输入时，模型的top-1准确率为95.6%，将骨架长度和骨架关节点两路信息一同输入模型后准确率提高至96.5%，将骨架关节点、长度、关节点运动和长度运动完整4路信息输入时模型准确率进一步提升至96.8%，与仅输入骨架关节点相比，网络的识别准确率提高了1.2个百分点。

表2　多流融合的NTU RGB+D 60数据集下的消融实验结果

Table 2　Ablation experiment results of multi stream fusion on the NTU RGB+D 60 data set　%

模型	输入数据类型	识别准确率
2s-AGCN[24]	骨架关节点	93.7
2s-AGCN[24]	骨架长度	93.2
2s-AGCN[24]	骨架关节点+长度	95.1
A-GCN+MS-TCN	骨架关节点	95.6
A-GCN+MS-TCN	骨架长度	95.5
A-GCN+MS-TCN	骨架关节点+骨架长度	96.5
A-GCN+MS-TCN	骨架关节点运动	93.3
A-GCN+MS-TCN	骨架长度运动	93.4
A-GCN+MS-TCN	骨架关节点+骨架长度+骨架关节点运动+骨架长度运动	96.8

将本文所提的A-GCN+MS-TCN完整模型与在NTU RGB+D 60数据集和NTU RGB+D 120数据集上识别效果较好的其他模型进行对比，如表3、表4所示，表中加粗数字表示该组数据最大值。由表3可知，A-GCN+MS-TCN模型在NTU RGB+D 60数据集的CS、CV划分模式下分别取得90.5%和96.8%的动作识别准确率，相比于ST-GCN基线模型，在CS、CV划分模式下的识别准确率分别提升了9.0、8.5个百分点。由表4可知，A-GCN+MS-TCN模型在NTU RGB+D 120数据集的CS、CT划分模式下分别取得86.0%和88.7%的动作识别准确率。

表3　不同模型在NTU RGB+D 60数据集下的结果对比

Table 3　Results comparison of different models under the NTU RGB+D 60 data set　%

模型	识别准确率	
	CS模式	CV模式
P-LSTM[11]	62.9	70.3
IndRNN[12]	81.8	88.0
Clips+CNN+MTLN[18]	79.6	84.8
ST-GCN[23]	81.5	88.3
2s-AGCN[24]	88.5	95.1
DGNN[26]	89.9	96.1
MS-G3D[27]	91.5	96.2
Sem-GCN[28]	86.2	94.2
CTR-GCN[30]	92.4	96.8
SATD-GCN[32]	89.3	95.5
A-GCN+MS-TCN	**90.5**	**96.8**

表4 不同模型在NTU RGB+D 120数据集下的结果对比

Table 4 Results comparison of different models under the NTU RGB+D 120 data set %

模型	识别准确率	
	CS模式	CT模式
ST-LSTM[16]	55.7	57.9
GCA-LSTM[17]	61.2	63.3
Body Pose Evolution Map[22]	64.6	66.9
ST-GCN[23]	70.7	73.2
2s-AGCN[24]	82.9	84.9
MS-G3D[27]	86.9	88.4
CTR-GCN[30]	88.9	90.6
A-GCN+MS-TCN	**86.0**	**88.7**

以上结果表明,自适应多尺度图卷积网络能够根据输入动作动态调节骨架关节点间的连接关系提取骨架动作的多粒度的时空特征,具有良好的泛化性能和表达能力,提升了图卷积网络的识别性能。

4 结束语

本文提出基于自适应多尺度图卷积网络的人体骨架动作识别方法。针对预定义骨架拓扑图连接关系固定导致模型泛化能力不足的问题,提出自适应空间图卷积层对骨架拓扑结构进行自适应调整,使用数据驱动的骨架拓扑图提升模型的泛化能力。针对单支路图卷积核提取特征导致模型表达能力不足的问题,提出多尺度时间图卷积层对时间图卷积算子进行多支路扩展,采用动态融合的骨架序列的时空特征提升模型的表达能力。在NTU RGB+D 60和NTU RGB+D 120动作识别数据集上验证本文模型的有效性,实验结果表明本文模型相比现有图卷积网络动作识别模型显著提升了分类准确率,具有良好的泛化性和鲁棒性。但本文模型仅针对人体骨架进行动作识别,由于忽略了环境、场景等辅助信息,导致模型对某些特定动作的识别能力较差,因此下一步将对多模态数据融合的动作识别模型进行研究,以提升模型的动作识别准确率。

参考文献

[1] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]// Proceedings of IEEE Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 7794-7803.

[2] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 4724-4733.

[3] WANG L, HUYNH D Q, KONIUSZ P. A comparative review of recent kinect-based action recognition algorithms[J]. IEEE Transactions on Image Processing, 2020, 29: 15-28.

[4] 胡建芳,王熊辉,郑伟诗,等. RGB-D行为识别研究进展及展望[J]. 自动化学报,2019,45(5):829-840.
HU J F, WANG X H, ZHENG W S, et al. RGB-D action recognition: recent advances and future perspectives[J]. Acta Automatica Sinica, 2019, 45(5): 829-840. (in Chinese)

[5] 赫磊,邵展鹏,张剑华,等. 基于深度学习的行为识别算法综述[J]. 计算机科学,2020,47(S01):139-147.
HE L, SHAO Z P, ZHANG J H, et al. Review of deep learning-based action recognition algorithms[J]. Computer Science, 2020, 47(S01): 139-147. (in Chinese)

[6] ZHANG Z Y. Microsoft kinect sensor and its effect[J]. IEEE MultiMedia, 2012, 19(2): 4-10.

[7] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human action recognition by representing 3D skeletons as points in a lie group[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 588-595.

[8] KONIUSZ P, CHERIAN A, PORIKLI F. Tensor representations via kernel linearization for action recognition from 3D skeletons[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 37-53.

[9] KONIUS Z P, WANG L, CHERIAN A. Tensor representations for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 648-665.

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

[11] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB D: a large scale dataset for 3D human activity analysis[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 1010-1019.

[12] LI S, LI W Q, COOK C, et al. Independently recurrent neural network: building a longer and deeper RNN[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 5457-5466.

[13] SI C Y, JING Y, WANG W, et al. Skeleton-based action recognition with spatial reasoning and temporal stack learning[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 106-121.

[14] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 1110-1118.

[15] LI W B, WEN L Y, CHANG M C, et al. Adaptive RNN tree for large-scale human action recognition[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 1453-1461.

[16] LIU J, SHAHROUDY A, XU D, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 816-833.

[17] LIU J, WANG G, DUAN L Y, et al. Skeleton-based human

action recognition with global context-aware attention LSTM networks [J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2018, 27(4): 1586-1599.

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

[19] KIM T S, REITER A. Interpretable 3D human action analysis with temporal convolutional networks [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington D. C., USA: IEEE Press, 2017: 1623-1631.

[20] LIU H, TU J, LIU M. Two-stream 3D convolutional neural network for skeleton-based action recognition[EB/OL]. [2022-08-10]. https://arxiv. org/abs/1705. 08106.

[21] LIU M, LIU H, CHEN C. Enhanced skeleton visualization for view invariant human action recognition[J]. Pattern Recognition, 2017, 68: 346-362.

[22] LIU M Y, YUAN J S. Recognizing human actions as the evolution of pose estimation maps[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 1159-1168.

[23] 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.

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

[25] SHI L, ZHANG Y F, CHENG J, et al. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[J]. IEEE Transactions on Image Processing, 2020, 29: 9532-9545.

[26] SHI L, ZHANG Y F, CHENG J, et al. Skeleton-based action recognition with directed graph neural networks [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 7904-7913.

[27] LIU Z Y, ZHANG H W, CHEN Z H, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 140-149.

[28] DING X L, YANG K, CHEN W. A semantics-guided graph convolutional network for skeleton-based action recognition [C]//Proceedings of the 4th International Conference on Innovation in Artificial Intelligence. New York, USA: ACM Press, 2020: 130-136.

[29] 孙琪翔，何宁，张聪聪，等. 基于轻量级图卷积的人体骨架动作识别方法[J]. 计算机工程，2022，48(5)：306-313.
SUN Q X, HE N, ZHANG C C, et al. Human skeleton action recognition method based on lightweight graph convolution[J]. Computer Engineering, 2022, 48(5): 306-313. (in Chinese)

[30] CHEN Y X, ZHANG Z Q, YUAN C F, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition [C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2022: 13339-13348.

[31] 王小娟，钟云，金磊，等. 基于骨架的自适应尺度图卷积动作识别[J]. 天津大学学报(自然科学与工程技术版)，2022，55(3)：306-312.
WANG X J, ZHONG Y, JIN L, et al. Scale adaptive graph convolutional network for skeleton-based action recognition[J]. Journal of Tianjin University (Science and Technology), 2022, 55(3): 306-312. (in Chinese)

[32] ZHANG J X, YE G X, TU Z G, et al. A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition [J]. CAAI Transactions on Intelligence Technology, 2022, 7(1): 46-55.

[33] TU Z G, ZHANG J X, LI H Y, et al. Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition[J]. IEEE Transactions on Multimedia, 2023, 25: 1819-1831.

[34] BIAN C L, FENG W, WANG S. Self-supervised representation learning for skeleton-based group activity recognition[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York, USA: ACM Press, 2022: 5990-5998.

[35] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[EB/OL]. [2022-08-10]. https://arxiv. org/abs/1312. 6203.

[36] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. [2022-08-10]. https://arxiv. org/abs/1609. 02907.

[37] NIEPERT M, AHMED M, KUTZKOV K. Learning convolutional neural networks for graphs[C]// Proceedings of the 33rd International Conference on Machine Learning. New York, USA: ACM Press, 2016: 2014-2023.

[38] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of Conference on Advances in Neural Information. Washington D. C., USA: IEEE Press, 2017: 76-85.

[39] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition . Washing D. C., USA: IEEE Press, 2016: 770-778.

[40] 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.

编辑　赖玉玲