基于多维动态拓扑学习图卷积的骨架动作识别

罗会兰,曹立京
(江西理工大学信息工程学院,江西赣州 341000)

摘 要: 图卷积由于其对图数据的强大表示能力被广泛应用于基于骨架的动作识别任务中. 但是现有的图卷积方法在所有帧或通道上都使用共享的图拓扑进行特征聚合,这极大限制了图卷积网络的表示能力. 为了解决这些问题,本文提出多维动态拓扑学习图卷积用于动态建模具有时序与通道特异性的拓扑结构. 多维动态拓扑学习图卷积主要包含三个组成部分:纯粹节点拓扑学习图卷积(pure Joint topology learning Graph Convolution, J-GC)、动态时序特异性拓扑学习图卷积(Dynamic Temporal-Wise topology learning Graph Convolution, DTW-GC)和通道特异性拓扑学习图卷积(Channel-Wise topology learning Graph Convolution, CW-GC). 特别地,在DTW-GC中使用了动态骨架拓扑建模方法(Dynamic Skeleton Topology Learning, DSTL),以高效地建模富含全局时空拓扑特征的动态骨架拓扑. 将多维动态拓扑学习图卷积与多尺度时间卷积(Multi-Scale Temporal Convolution, MS-TC)相结合,本文构建了具有强大建模能力的图卷积网络. 此外,为了对骨架数据的空间信息进行补充,本文额外引入了相对节点数据和相对骨骼数据进行多流网络的融合. 本文所提出的方法在NTU-RGB+D与NTU-RGB+D 120数据集上分别取得了92.64%和89.29%的准确率,超过了当前最先进方法.

关键词: 动作识别;深度学习;图卷积;动态骨架拓扑;数据融合

基金项目: 国家自然科学基金(No.61862031);江西省主要学科技术带头人领军人才计划资助项目(No.20213BCJ22004);江西省学位与研究生教育教学改革研究重点项目(No.JXYJG-2020-120)

中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2024)03-0991-11

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221106

Multi-Dimensional Dynamic Topology Learning Graph Convolution for Skeleton-Based Action Recognition

LUO Hui-lan, CAO Li-jing
(*School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou , Jiangxi* 341000, *China*)

Abstract: Graph convolution is widely used in skeleton-based action recognition because of its effectiveness of processing graph data. However, the existing graph convolution methods use the shared graph topology for feature aggregation on all frames or channels, which greatly limits the representation ability of graph convolution network. In order to solve these problems, a multi-dimensional dynamic topology learning graph convolution is proposed in this paper to dynamically model the topology with temporal and channel specificity. The multi-dimensional dynamic topology learning graph convolution mainly includes three parts: pure joint topology learning graph convolution (J-GC), dynamic temporal-wise topology learning graph convolution (DTW-GC) and channel-wise topology learning graph convolution (CW-GC). In particular, in DTW-GC, a dynamic skeleton topology modeling method (DSTL) is designed to efficiently model the dynamic skeleton topology with rich global spatio-temporal topological features. Finally, by combining multi-dimensional dynamic topology learning graph convolution with multi-scale temporal convolution (Muti-Scale TCN), a graph convolution network with powerful modeling capability is constructed in this paper. In addition, in order to supplement the spatial information of skeleton data, the relative joint data and relative bone data are introduced for multi-stream network fusion. Our method achieves 92.64% and 89.29% accuracy on NTU-RGB+D and NTU-RGB+D 120 datasets, respectively, which is superior to the current state-of-the-art methods.

收稿日期:2022-09-30;修回日期:2023-03-15;责任编辑:梅志强

Key words: action recognition; deep learning; graph convolution; dynamic skeleton topology; data fusion
Foundation Item(s): National Natural Science Foundation of China (No.61862031); The Project Supported by the Leading Talents Plan for the Technical Leaders of Major Disciplines in Jiangxi Province (No.20213BCJ22004); Jiangxi Province Degree and Postgraduate Education and Teaching Reform Research Key Project (No.JXYJG-2020-120)

1 引言

动作识别是计算机视觉领域的研究热点,其在人机交互、公共安全监控、影视制作与医疗康复等领域有着广泛的应用. 由于深度传感器和实时人体姿态估计技术[1]的发展,骨架数据变得愈加广泛和廉价. 与原始RGB、RGB+D数据相比,骨架数据信息密度高,并提供了高语义层次的结构信息,因此使用骨架数据进行动作识别可以提高计算效率和识别性能,特别是在复杂场景情况下,它的鲁棒性更强.

基于骨架序列的动作识别方法可分为三大类:基于递归神经网络(Recurrent Neural Network, RNN)的方法,基于卷积神经网络(Convolutional Neural Network, CNN)的方法和基于图卷积网络(Graph Convolutional Network, GCN)的方法. 其中RNN具有建模时序关系的固有优势,但RNN对于空间结构信息的提取能力存在较大缺陷. 基于CNN的骨架动作识别方法对于空间序列信息的提取能力更加优异,但是由于骨架数据不同于传统视频图像,使用传统的CNN网络不能表达节点间的复杂拓扑结构信息. 而GCN能够保留骨架结构及节点间潜在的空间关系,使得其处理骨架数据拥有天然的优势,因此GCN成为基于骨架数据的动作识别任务的主流方法.

Sijie 和 Yuanjun[2] 首次提出的 ST-GCN (Spatial-Temporal Graph Convolutional Network),开创了GCN在骨架动作识别任务应用的先河. 在此基础上,MSAGC-SRU(Multi-Stream spatial Attention Graph Convolutional SRU Network)[3]将循环卷积网络与图卷积结合,在简单循环单元中嵌入图卷积进行骨架动作识别,得到了一定的性能提升. 由于ST-GCN采用了固定的拓扑结构,所以无法建模非自然连接节点之间的关系. 针对此局限性,2s-AGCN(Two-stream Adaptive Graph Convolutional Network)[4]、MS-AAGCN(Multi-Stream Adaptive Graph Convolutional Network)[5]、Dynamic GCN(Dynamic Graph Convolutional Network)[6]和SMotif-GCN(Sparse Motif-based Graph Convolutional Network)[7]等方法通过自注意力机制、卷积聚合和节点间距离来引入自适应拓扑结构;AS-GCN(Actional-Structural Graph Convolutional Network)[8]通过学习高阶邻接拓扑和动作相关连接拓扑来学习灵活多变的节点间的拓扑关系;MS-G3D(Disentangling and unifying Graph convolutions)[9]通过时间窗口来获取多尺度图拓扑结构,实现窗口内的联合时空关系建模;InfoGCN(Info Graph Convolutional Network)[10]使用基于注意力的图卷积来捕获人体动作中上下文相关的内在拓扑结构. 虽然上述方法可以建模非自然连接关系,但是它们在进行图卷积时,所有的时序帧和通道都使用相同的拓扑结构进行特征聚合,这限制了图卷积网络的学习能力. 由于不同的时序帧代表动作执行的不同时刻,而不同的通道上存放着不同的运动模式特征,在运动执行的不同时刻和不同的运动模式下,关节点之间的关系并不总是相同,因此在不同时序和通道维度使用相同的拓扑结构进行图卷积限制了网络的学习能力.

为了进一步增强图卷积的特征提取能力,DC-GCN+ADG(Decoupling GCN with DropGraph module)[11]与CTR-GCN(Channel-wise Topology Refinement Graph Convolution Network)[12]通过建模通道分组拓扑和通道特异性拓扑进行特征聚合,然而它们对于不同时序帧仍然使用共享的拓扑结构. 而SGN(Semantics-Guided neural Network)[13]虽然建模了具有时序特异性的拓扑结构,但是却在所有通道上使用了同一拓扑结构进行特征聚合. 为了同时建模通道和时序特异拓扑结构,本文提出了一种新颖的多维动态拓扑学习图卷积方法(Multi-Dimensional Dynamic Topology Learning Graph Convolution, MD^{2}TL-GC),同时对节点维、时序维和通道维进行拓扑结构建模,使得每一帧与每一个通道都拥有独立、动态的特征聚合参数,从而使模型拥有更大的灵活程度和更强的学习能力. 此外,为了减少获得全局语义信息所需的网络深度,本文提出了动态骨架拓扑学习(Dynamic Skeleton Topology Learning, DSTL),将动态骨架序列表达成单张动态骨架图,在此基础上进行图卷积,从全局时序的角度进行特征的聚集,使网络只需要较浅的深度就能学习到高级时空特征. MD^{2}TL-GC利用全局时空信息在各个数据维度上自适应建模拓扑结构,用它构造的网络仅需要5层就可以超越当前主流方法. 另外,本文提出了相对节点和相对骨骼的骨架信息表达方法,为原始骨架序列数据提供互补信息,与经典的节点运动流和骨骼运动流相比,获得了更高的性能提升.

本文的贡献总结如下:

(1) 提出的MD^{2}TL-GC通过分级的方式动态地建模各个维度的拓扑结构,从而实现了灵活有效的拓扑关系建模和高效的特征提取.

(2) 提出的动态骨架拓扑建模方法,学习到骨架序列数据的全局时空表达,基于此全局动态拓扑结构进行图卷积,可以增强多维拓扑学习中的全局特征,实现应用较浅的网络达到更好的性能.

(3) 引入了全新的相对节点和相对骨骼模态数据,进一步提高了动作识别的性能. 在两个骨架动作识别大型数据集 NTU-RGB+D[14] 与 NTU-RGB+D 120[15] 上进行了大量实验,结果表明利用 MD^{2}TL-GC 所构造的网络 MD^{2}TL-GCN(MD^{2}TL-GC Network)获得了优越的性能,与 ST-GCN 相比,在 NTU-RGB+D 数据集的 CS(Cross-Subject)和 CV(Cross-View)基准上分别获得了+11.14%和+8.06%的显著提升.

2　方法

本节首先介绍了经典图卷积的过程与动态图像的计算方法;然后对本文所提出的多维动态拓扑学习图卷积(MD^{2}TL-GC)进行了详细的论述;最后阐述了用于骨架动作识别的多维动态拓扑图卷积网络(MD^{2}TL-GCN).

2.1　图卷积网络

有 N 个节点和 T 帧的骨架序列时空图定义为 $G=\{V,E\}$,其中:

$$\begin{cases} V=\left\{v_{ti}\middle| t=1,\cdots,T;\ i=1,\cdots,N\right\} \\ E_{\mathrm{S}}=\left\{v_{ti}v_{tj}\middle|(i,j)\in H\right\} \\ E_{\mathrm{T}}=\left\{v_{ti}v_{(t+1)i}\right\} \\ E=\left\{E_{\mathrm{S}},E_{\mathrm{T}}\right\} \end{cases} \tag{1}$$

在上述时空图中边集合包括 E_{S} 和 E_{T},分别指空间边(spatial edges)集合和时序边(temporal edges)集合;H 表示自然连接的人体关节点对的集合.

空间边集合 E_{S} 的连接情况使用空间域邻接矩阵 $\boldsymbol{A}_{\mathrm{S}}\in\mathbf{R}^{N\times N}$ 表示:如果节点 i 和 j 在空间上直接相连,则 $A_{\mathrm{S}ij}=1$,其余为 0. 自 ST-GCN[2] 以来,空间图卷积核的大小一般设置为:$K_{\mathrm{S}}=3$,即将邻接节点集划分为 3 个子集,相应的邻接矩阵 $\boldsymbol{A}_{\mathrm{S}}$ 分为 $\boldsymbol{A}_{\mathrm{S}}^1$、$\boldsymbol{A}_{\mathrm{S}}^2$ 和 $\boldsymbol{A}_{\mathrm{S}}^3$,分别代表根节点、向心运动节点和离心运动节点的连接情况. 空间图卷积分别基于这三个邻接子集进行卷积,然后将所得结果相加.

图卷积操作通过不断聚集邻域信息来更新当前节点的特征. 其中,空间维度上的图卷积过程可表示为:

$$\begin{cases} \boldsymbol{f}_{\mathrm{S}}=\sum\limits_{k=1}^{K_{\mathrm{S}}}\boldsymbol{W}_{\mathrm{S}}^{(k)}\boldsymbol{f}_{\mathrm{in}}\widetilde{\boldsymbol{A}_{\mathrm{S}}^{(k)}}\circledast\boldsymbol{M}^{(k)} \\ \widetilde{\boldsymbol{A}_{\mathrm{S}}^{(k)}}=\boldsymbol{D}_{\mathrm{S}}^{(k)^{-\frac{1}{2}}}\boldsymbol{A}_{\mathrm{S}}^{(k)}\boldsymbol{D}_{\mathrm{S}}^{(k)^{-\frac{1}{2}}} \\ \boldsymbol{D}_{\mathrm{S}}^{(k)}=\sum\limits_{j}\boldsymbol{A}_{\mathrm{S}ij}^{(k)}+\varepsilon \end{cases} \tag{2}$$

其中输入特征图 $\boldsymbol{f}_{\mathrm{in}}\in\mathbf{R}^{C_{\mathrm{in}}\times T\times N}$,$C_{\mathrm{in}}$,$T$,$N$ 分别表示输入通道数、帧数和关节点数量;$\boldsymbol{W}_{\mathrm{S}}^{(k)}\in\mathbf{R}^{C_{\mathrm{out}}\times C_{\mathrm{in}}}$ 表示 1×1 空域图卷积运算的可训练权重向量;$\boldsymbol{M}^{(k)}\in\mathbf{R}^{N\times N}$ 是一个简单的注意力掩膜矩阵,代表每个关节的重要性;$\circledast$ 为点乘操作;$\boldsymbol{A}_{\mathrm{S}}^{(k)}$ 使用对角矩阵 $\boldsymbol{D}_{\mathrm{S}}^{(k)}$ 规范化后得到 $\widetilde{\boldsymbol{A}_{\mathrm{S}}^{(k)}}$,以保证各划分间的平衡和卷积前后的幅值稳定;$\varepsilon$ 用于避免分母为 0,一般设置为 0.001.

对于时间维度上的图卷积,利用普通的卷积就可以实现:

$$\boldsymbol{f}_T=\mathrm{Conv\,1D}_{K_t}\left(\boldsymbol{f}_{\mathrm{S}}\right) \tag{3}$$

其中 $\mathrm{Conv\,1D}_{K_t}$ 表示时间维度上卷积核大小为 K_t 的一维卷积操作.

将输入骨架序列先进行空域图卷积,再进行时间维卷积,完成一次图卷积,重复这个过程便可以构建出深度图卷积网络[2].

与上述采用固定的节点邻接矩阵表示节点间的空域拓扑关系不同,本文提出的 MD^{2}TL-GC 在节点维、时序维和通道维学习节点间的拓扑关系,在此基础上进行更加灵活的空域图卷积,使得模型具有更强的表达能力. 另外,MD^{2}TL-GCN 在时间维上采用了多尺度的卷积方法,使得模型在时序上具有多尺度的感受野.

2.2　动态图计算

在文献[16]中首次将动态图(dynamic image)应用于视频动作识别任务中. 具体来说,对于输入视频 V:$\{\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_T\}$,它的动态图为 $\boldsymbol{d}^*$,计算过程如式(4)所示:

$$\begin{cases} \boldsymbol{d}^*=\underset{\boldsymbol{d}}{\operatorname{argmin}}\,E(\boldsymbol{d}) \\ E(\boldsymbol{d})=\dfrac{\lambda}{2}\|\boldsymbol{d}\|^2+\dfrac{2}{T(T-1)} \\ \qquad\times\sum\limits_{q>t}\max\left\{0,1-S(q|\boldsymbol{d})+S(t|\boldsymbol{d})\right\} \end{cases} \tag{4}$$

其中 $q,t\in\{1,2,\cdots,T\}$,$S(t|\boldsymbol{d})=\left\langle\boldsymbol{d},\boldsymbol{V}_t\right\rangle$,$\boldsymbol{V}_t=\dfrac{1}{t}\sum\limits_{\tau=1}^{t}\varphi(\boldsymbol{I}_\tau)$ 表示前 t 帧的平均特征,$\varphi(\boldsymbol{I}_\tau)\in\mathbf{R}^D$ 表示第 τ 帧 $\boldsymbol{I}_\tau$ 的特征向量. 可以使用 RankSVM 方法优化式(4),学习获得视频 V 的动态图 $\boldsymbol{d}^*\in\mathbf{R}^D$ 来聚合视频帧序列的全局时空特征.

但是,精确的优化学习方法效率较低,故在文献[16]中,作者提出动态图的近似计算方法. 具体来说,通过对式(4)进行一次手动梯度计算,得到近似动态图的计算公式:

$$\boldsymbol{d}^*\cong\sum_{t=1}^{T}r_t\varphi(\boldsymbol{I}_t) \tag{5}$$

其中 $r_t=2t-T-1$. 本文也利用了这种近似计算方法来设计 DSTL,以学习到节点间的全局动态拓扑关系.

2.3　MD^{2}TL-GCN 概述

MD^{2}TL-GCN 的整体框架如图 1 所示,共包含 5 层网

络结构(L1~L5),每一层网络结构(在图1中表示为MD²TL block)包含多维动态拓扑图卷积(MD²TL-GC)和多尺度时间卷积(MS-TC). 沿用文献[2]中先空间维卷积再时间维卷积的方法,MD²TL-GC和MS-TC通过串联的方式组合成一个MD²TL block,以有效提取时空特征. 通过叠加多层MD²TL block,便可构造出用于骨架动作识别任务的端到端层级网络MD²TL-GCN.

如图1所示,其中MD²TL-GC主要包括三部分:纯粹节点拓扑学习图卷积(pure Joint topology learning Graph Convolution, J-GC)、动态时序特异性拓扑学习图卷积(Dynamic Temporal-Wise topology learning Graph Convolution, DTW-GC)和通道特异性拓扑学习图卷积(Channel-Wise topology learning Graph Convolution, CW-GC). 为了减少计算复杂度,在同样的通道数中表达更加多样的运动模式特征,MD²TL-GC先将通道数平均划分为两部分,一部分用于学习J-GC,另一部分用于学习DTW-GC. 然后,将学习到的节点特异性拓扑结构和时序特异性拓扑结构沿通道维度串接融合,再进行CW-GC学习,实现对每一个通道中所代表的运动模式特征的细化与调整.

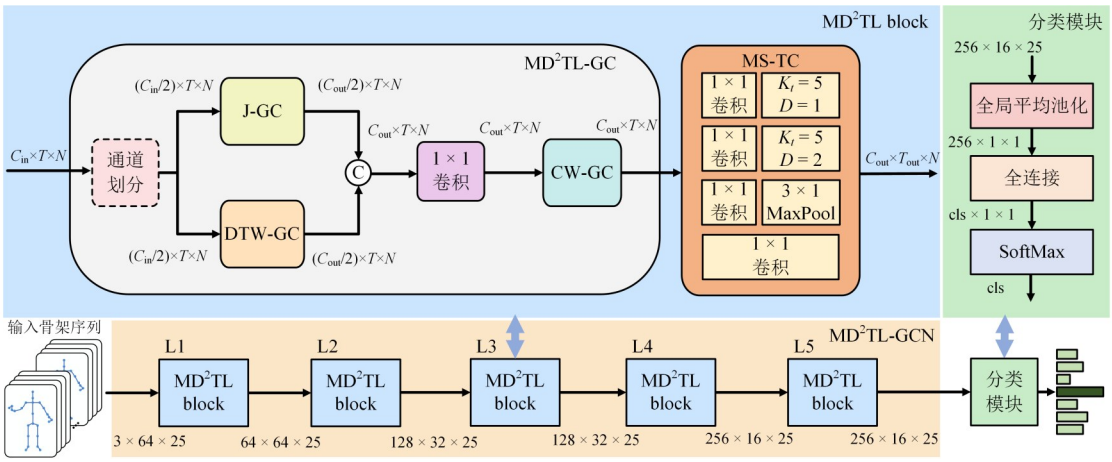


图1 MD²TL-GCN网络结构

MD²TL-GC的总体计算过程如式(6)所示.

$$\begin{cases} \boldsymbol{f}_{\text{in1}}, \boldsymbol{f}_{\text{in2}} = \text{split}(\boldsymbol{f}_{\text{in}}) \\ \boldsymbol{f}_{\text{out}} = G_{\text{C}}(\text{Conv}([G_{\text{J}}(\boldsymbol{f}_{\text{in1}}), G_{\text{DT}}(\boldsymbol{f}_{\text{in2}})])) \end{cases} \tag{6}$$

其中split(*)表示通道划分操作,即对于$\boldsymbol{f}_{\text{in}} \in \mathbf{R}^{C_{\text{in}} \times T \times N}$,通道划分后$\boldsymbol{f}_{\text{in1}}, \boldsymbol{f}_{\text{in2}} \in \mathbf{R}^{(C_{\text{in}}/2) \times T \times N}$(值得注意的是,对于第一个MD²TL block,由于输入通道数只有3,不进行通道划分操作,此时$\boldsymbol{f}_{\text{in1}}, \boldsymbol{f}_{\text{in2}} \in \mathbf{R}^{C_{\text{in}} \times T \times N}$);$G_{\text{J}}$、$G_{\text{DT}}$和$G_{\text{C}}$分别表示J-GC、DTW-GC和CW-GC;Conv表示1×1卷积.

2.4 纯粹节点拓扑结构学习图卷积

针对不同的动作,不同的数据样本,为了获取到节点之间的灵活拓扑结构,我们使用了文献[5]中的自适应图卷积方法学习纯粹节点间的拓扑结构.

如图2所示,J-GC学习两种类型的拓扑结构:动作特定的纯粹节点拓扑结构$\boldsymbol{G}_p^{(k)} \in \mathbf{R}^{N \times N}$与数据特定的纯粹节点拓扑结构$\boldsymbol{G}_j^{(k)} \in \mathbf{R}^{N \times N}$. 其中$\boldsymbol{G}_p^{(k)}$与输入特征没有关系,直接通过动作分类损失进行训练. 而$\boldsymbol{G}_j^{(k)}$的计算过程如式(7)所示:

$$\begin{cases} \boldsymbol{f}_\theta^{(k)} = \text{Re shape}(\boldsymbol{W}_{\theta^{(k)}} \boldsymbol{f}_{\text{in}}) \\ \boldsymbol{f}_\varphi^{(k)} = \text{Re shape}(\boldsymbol{W}_{\varphi^{(k)}} \boldsymbol{f}_{\text{in}}) \\ \boldsymbol{G}_j^{(k)} = \text{Re shape}(\text{Tanh}(\boldsymbol{f}_\theta^{(k)} \boldsymbol{f}_\varphi^{(k)})) \end{cases} \tag{7}$$

其中$\boldsymbol{f}_\theta^{(k)} \in \mathbf{R}^{N \times C_{\text{e}} T}$,$\boldsymbol{f}_\varphi^{(k)} \in \mathbf{R}^{C_{\text{e}} T \times N}$,故它们经过矩阵相乘后得到数据特定的纯粹节点拓扑结构$\boldsymbol{G}_j^{(k)} \in \mathbf{R}^{N \times N}$,$\boldsymbol{W}_{\theta^{(k)}}, \boldsymbol{W}_{\varphi^{(k)}} \in \mathbf{R}^{C_{\text{e}} \times (C_{\text{in}}/2)}$表示1×1卷积的可学习参数,用于特征学习和通道维度的调整,在我们的实验中$C_{\text{e}} = C_{\text{in}}/8$;Tanh表示激活函数.

J-GC利用动作特定的纯粹节点拓扑结构$\boldsymbol{G}_p^{(k)}$与数据特定的纯粹节点拓扑结构$\boldsymbol{G}_j^{(k)}$进行卷积的过程描述

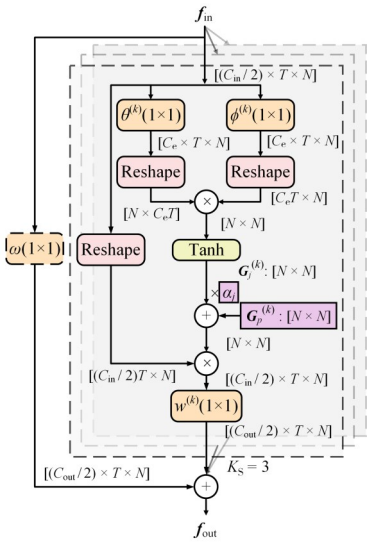


图2 J-GC的计算流程

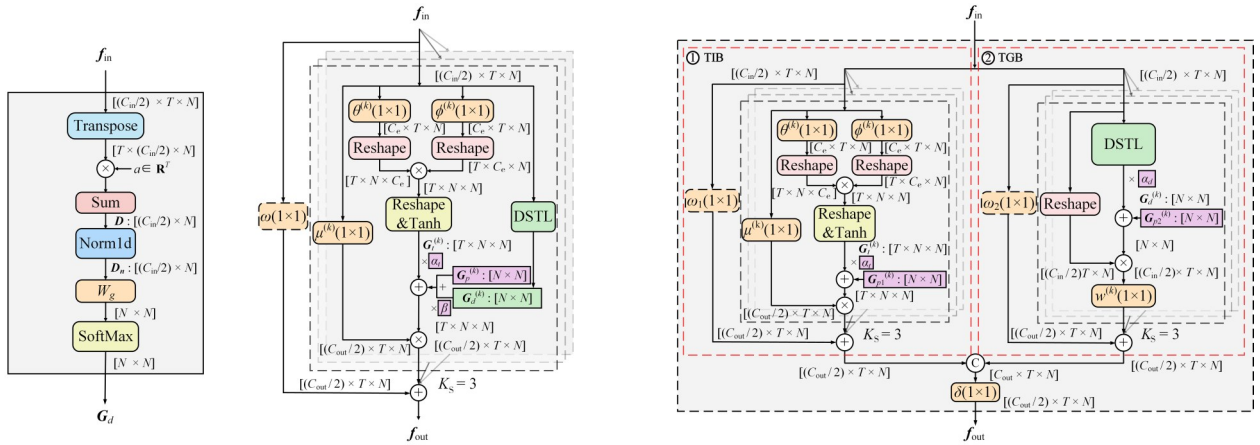
如式(8)所示:

$$\begin{cases} \boldsymbol{f}_{\mu}^{(k)} = \text{Re shape}(\boldsymbol{f}_{\text{in}}) \\ \boldsymbol{G}^{(k)} = \boldsymbol{G}_p^{(k)} + \alpha_j \boldsymbol{G}_j^{(k)} \\ \boldsymbol{f} = \begin{cases} \boldsymbol{W}_{\omega} \boldsymbol{f}_{\text{in}}, C_{\text{in}} \neq C_{\text{out}} \\ \boldsymbol{f}_{\text{in}}, C_{\text{in}} = C_{\text{out}} \end{cases} \\ \boldsymbol{f}_{\text{out}} = \sum_{k=1}^{K_{\text{S}}} \boldsymbol{W}_{w^{(k)}} \boldsymbol{f}_{\mu}^{(k)} \boldsymbol{G}^{(k)} + \boldsymbol{f} \end{cases} \tag{8}$$

其中α_j为可学习参数,用于调节两拓扑结构之间的相对重要性;$\boldsymbol{W}_{w^{(k)}}$是图卷积参数;$\boldsymbol{W}_{\omega}$为1×1卷积参数,用于调节通道数.

2.5　动态时序特异性拓扑学习图卷积

在骨架序列中,在不同的时序帧上,节点间的拓扑关系是不同的,为了学习时序维上的特异性拓扑结构,受文献[5]和文献[16]的启发,本文提出了动态时序特异性拓扑学习图卷积(DTW-GC). 如图3(*c*)所示,DTW-GC包含两个分支:时序特异性分支(Temporal Individual Branch,TIB)和时序全局性分支(Temporal Global Branch,TGB),用于分别学习两类拓扑结构:时序特异性拓扑$\boldsymbol{G}_t$和时序全局动态拓扑$\boldsymbol{G}_d$. 为了训练的稳定性,在MD^{2}TL-GCN的L1,$\boldsymbol{G}_t$和$\boldsymbol{G}_d$的融合采用了直接拓扑结构的融合(如图3(*b*)所示);而在L2~L5,则对基于$\boldsymbol{G}_t$和$\boldsymbol{G}_d$的图卷积结果进行通道维的串接融合(如图3(*c*)所示). 时序特异性拓扑$\boldsymbol{G}_t$使得网络在不同的时序帧上可以使用不同的特征聚合方式,而时序全局动态拓扑$\boldsymbol{G}_d$给网络提供全局动态拓扑结构信息,让图卷积在时间维上具有全局感受野.



(*a*) DSTL　　(*b*) DTW-GC(L1)　　(*c*) DTW-GC(L2~L5)

图3　DSTL与DTW-GC的计算流程

2.5.1　时序特异性拓扑分支

如图3(*c*)所示,TIB包含时序特异性拓扑$\boldsymbol{G}_t^{(k)} \in \mathbf{R}^{T\times N\times N}$和用于调整时序拓扑的动作特定拓扑结构$\boldsymbol{G}_{p1}^{(k)} \in \mathbf{R}^{N\times N}$. 与J-GC类似,$\boldsymbol{G}_{p1}^{(k)}$是可训练的参数,它与输入特征无关,通过动作识别损失直接进行训练,学习到与动作相关的特定拓扑结构,用于对各时序特定拓扑结构$\boldsymbol{G}_t^{(k)}$进行调整和补充. 而$\boldsymbol{G}_t^{(k)}$是与输入特征相关的拓扑结构,它的计算过程如式(9)所示:

$$\begin{cases} \boldsymbol{f}_{\theta}^{(k)} = \text{Re shape}(\boldsymbol{W}_{\theta^{(k)}} \boldsymbol{f}_{\text{in}}) \\ \boldsymbol{f}_{\varphi}^{(k)} = \text{Re shape}(\boldsymbol{W}_{\varphi^{(k)}} \boldsymbol{f}_{\text{in}}) \\ \boldsymbol{G}_t^{(k)} = \text{Re shape}(\text{Tanh}(\boldsymbol{f}_{\theta}^{(k)} \boldsymbol{f}_{\varphi}^{(k)})) \end{cases} \tag{9}$$

其中$\boldsymbol{f}_{\theta}^{(k)} \in \mathbf{R}^{T\times N\times C_{\text{e}}}$,$\boldsymbol{f}_{\varphi}^{(k)} \in \mathbf{R}^{T\times C_{\text{e}}\times N}$,故它们经过矩阵相乘后得到时序特异性拓扑$\boldsymbol{G}_t^{(k)} \in \mathbf{R}^{T\times N\times N}$,每个时序帧上对应着不同的拓扑结构.

TIB的图卷积计算过程如式(10)所示. 其中$\boldsymbol{f}_1$为调整通道维度后的残差连接,α_t为可学习参数,用于调节两拓扑结构之间的相对重要性. 通过该时序特异性拓扑学习支流,DTW-GC可以捕捉到不同时刻节点间的拓扑结构.

$$\begin{cases} \boldsymbol{f}_{\mu 1}^{(k)} = \boldsymbol{W}_{\mu^{(k)}} \boldsymbol{f}_{\text{in}} \\ \boldsymbol{G}_1^{(k)} = \boldsymbol{G}_{p1}^{(k)} + \alpha_t \boldsymbol{G}_t^{(k)} \\ \boldsymbol{f}_1 = \begin{cases} \boldsymbol{W}_{\omega 1} \boldsymbol{f}_{\text{in}}, C_{\text{in}} \neq C_{\text{out}} \\ \boldsymbol{f}_{\text{in}}, C_{\text{in}} = C_{\text{out}} \end{cases} \\ \boldsymbol{f}_{\text{out1}} = \sum_{k=1}^{K_{\text{S}}} \boldsymbol{f}_{\mu 1}^{(k)} \boldsymbol{G}_1^{(k)} + \boldsymbol{f}_1 \end{cases} \tag{10}$$

2.5.2　动态骨架拓扑分支

在TGB中,为了学习到全局动态骨架拓扑,在文献[16]的启发下,本文设计了DSTL模块学习动态骨架拓扑结构.

DSTL计算流程如图3(*a*)所示,对于骨架序列输入特征$\{\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_T\}$,首先计算近似排序池化系数:

$$a_t = 2t - T - 1 \tag{11}$$

随后使用该系数进行近似排序池化得到动态骨架,即:

$$\boldsymbol{D}=\sum_{t=1}^{T}a_t\boldsymbol{f}_t \tag{12}$$

随后对该动态骨架进行卷积映射,用以获取到包含全局时空特征的拓扑结构信息,其计算过程为:

$$\begin{cases}\boldsymbol{D}_n=\mathrm{Norm1d}(\boldsymbol{D})\\ \boldsymbol{G}_d=\mathrm{soft\,max}(\boldsymbol{W}_g\boldsymbol{D}_n)\end{cases} \tag{13}$$

其中Norm1d为归一化操作,在输入通道上进行归一化得到$\boldsymbol{D}_n$;而$\boldsymbol{W}_g\in\mathbf{R}^{N\times(C_{\mathrm{in}}/2)}$表示使用一维卷积进行特征映射,将维度为$(C_{\mathrm{in}}/2)\times N$的$\boldsymbol{D}_n$映射为维度为$N\times N$的邻接矩阵$\boldsymbol{G}_d$;最后使用soft max函数激活便可以得到全局视角下的动态骨架拓扑结构$\boldsymbol{G}_d\in\mathbf{R}^{N\times N}$. DSTL学习到的动态骨架拓扑$\boldsymbol{G}_d$具有全局时序维感受野,在此基础上进行图卷积,可以高效学习到高级语义特征.

2.5.3 拓扑融合

考虑到网络浅层缺少高级语义特征,直接在其上进行时序特异性图卷积会引入较大的方差,故此,对于MD^{2}TL-GCN的第一层(L1),如图3(*b*)所示,以相加的方式将含有全局时序信息的动态骨架拓扑$\boldsymbol{G}_d$与时序维特异性拓扑$\boldsymbol{G}_t$进行融合,在此基础上再进行图卷积. L1层的融合拓扑结构计算如式(14)所示:

$$\boldsymbol{G}^{(k)}=\boldsymbol{G}_p^{(k)}+\alpha_t\boldsymbol{G}_t^{(k)}+\beta\boldsymbol{G}_d^{(k)} \tag{14}$$

其中$\boldsymbol{G}_p^{(k)}$表示可学习的动作特定拓扑结构,α_t和β都是可学习的参数,用于调整各拓扑结构间的比重.

而对于MD^{2}TL-GCN的其他层(L2~L5),如图3(*c*)所示,是将TIB和TGB的图卷积结果进行通道维串接融合. 首先,TGB分支基于DSTL学习到的动态骨架拓扑$\boldsymbol{G}_d$进行图卷积,过程如式(15)所示.

$$\begin{cases}\boldsymbol{f}_{\mu 2}^{(k)}=\mathrm{Reshape}(\boldsymbol{f}_{\mathrm{in}})\\ \boldsymbol{G}_2^{(k)}=\boldsymbol{G}_{p2}^{(k)}+\alpha_d\boldsymbol{G}_d^{(k)}\\ \boldsymbol{f}_2=\begin{cases}\boldsymbol{W}_{\omega 2}\boldsymbol{f}_{\mathrm{in}},C_{\mathrm{in}}\neq C_{\mathrm{out}}\\ \boldsymbol{f}_{\mathrm{in}},C_{\mathrm{in}}=C_{\mathrm{out}}\end{cases}\\ \boldsymbol{f}_{\mathrm{out2}}=\sum_{k=1}^{K_{\mathrm{S}}}\boldsymbol{W}_{w^{(k)}}\boldsymbol{f}_{\mu 2}^{(k)}\boldsymbol{G}_2^{(k)}+\boldsymbol{f}_2\end{cases} \tag{15}$$

其中$\boldsymbol{G}_{p2}^{(k)}$表示动作特定的可学习拓扑结构,$\boldsymbol{f}_2$为调整通道维度后的残差连接,$\alpha_d$为可学习参数,用于调节两拓扑结构之间的相对重要性.

然后,将两个分支的图卷积结果进行通道维拼接后,输入一个1×1卷积进行特征融合的进一步学习和通道维调整.

2.6 通道特异性拓扑学习图卷积

在将J-GC与DTW-GC的卷积结果拼接后,所获取到的特征图中通道维度包含丰富的时空动态特征与运动模式特征. 因此为了进一步细化特征,我们提出了通道特异性拓扑学习图卷积(CW-GC),通过学习通道维上的特异性拓扑结构来进一步增强网络的学习能力.

如图4所示,通道特异性拓扑结构$\boldsymbol{G}_{\mathrm{c}}^{(k)}$的学习过程如式(16)所示:

$$\begin{cases}\boldsymbol{f}_\theta^{(k)}=\mathrm{Re\,shape}(\boldsymbol{W}_{\theta^{(k)}}\boldsymbol{f}_{\mathrm{in}})\\ \boldsymbol{f}_\varphi^{(k)}=\mathrm{Re\,shape}(\boldsymbol{W}_{\varphi^{(k)}}\boldsymbol{f}_{\mathrm{in}})\\ \boldsymbol{G}_{\mathrm{ce}}^{(k)}=\mathrm{Re\,shape}(\boldsymbol{f}_\theta^{(k)}\boldsymbol{f}_\varphi^{(k)})\\ \boldsymbol{G}_{\mathrm{c}}^{(k)}=\boldsymbol{W}_{w^{(k)}}(\mathrm{Mean}(\mathrm{Tanh}(\boldsymbol{G}_{\mathrm{ce}}^{(k)})))\end{cases} \tag{16}$$

其中$\boldsymbol{W}_{\theta^{(k)}},\boldsymbol{W}_{\varphi^{(k)}}\in\mathbf{R}^{C_{\mathrm{e}}\times C_{\mathrm{out}}}$表示1×1卷积的可学习参数,用于特征学习和通道维度的调整,为了构建特征丰富的通道特异性拓扑结构,设置$C_{\mathrm{e}}=C_{\mathrm{out}}/4$;$\boldsymbol{f}_\theta^{(k)}\in\mathbf{R}^{C_{\mathrm{e}}N\times T}$,$\boldsymbol{f}_\varphi^{(k)}\in\mathbf{R}^{T\times C_{\mathrm{e}}N}$,故它们经过矩阵相乘和维度转换后得到$\boldsymbol{G}_{\mathrm{ce}}^{(k)}\in\mathbf{R}^{C_{\mathrm{e}}\times N\times C_{\mathrm{e}}\times N}$. 然后,沿$\boldsymbol{G}_{\mathrm{ce}}^{(k)}$倒数第二维进行平均池化操作,得到维度为$C_{\mathrm{e}}\times N\times N$的具有通道特异性的拓扑结构,将通道维度的自相关性融合于邻接矩阵中. 之后,$\boldsymbol{G}_{\mathrm{ce}}^{(k)}$使用$\boldsymbol{W}_{w^{(k)}}$进行维度调整得到通道特异性拓扑结构$\boldsymbol{G}_{\mathrm{c}}^{(k)}\in\mathbf{R}^{C_{\mathrm{out}}\times N\times N}$.

CW-GC也使用了动作特定的可学习邻接矩阵$\boldsymbol{G}_p^{(k)}$,用于自适应调整$\boldsymbol{G}_{\mathrm{c}}^{(k)}$. 其图卷积过程如式(17)所示:

$$\begin{cases}\boldsymbol{f}_\mu^{(k)}=\boldsymbol{W}_{\mu^{(k)}}\boldsymbol{f}_{\mathrm{in}}\\ \boldsymbol{G}^{(k)}=\boldsymbol{G}_p^{(k)}+\alpha_{\mathrm{c}}\boldsymbol{G}_{\mathrm{c}}^{(k)}\\ \boldsymbol{f}_{\mathrm{out}}=\sum_{k=1}^{K_{\mathrm{S}}}\boldsymbol{f}_\mu^{(k)}\boldsymbol{G}^{(k)}+\boldsymbol{f}_{\mathrm{in}}\end{cases} \tag{17}$$

其中$\boldsymbol{W}_{\mu^{(k)}}\in\mathbf{R}^{C_{\mathrm{out}}\times C_{\mathrm{out}}}$是图卷积参数;$\alpha_{\mathrm{c}}$为可学习参数,用于调节两拓扑结构之间的相对重要性. CW-GC网络可以对具有丰富时空信息的特征图进行通道特异性拓扑

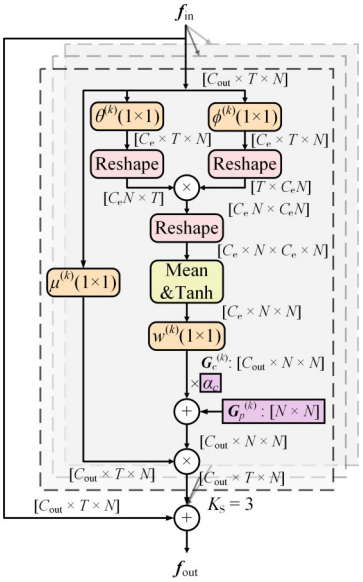


图4 CW-GC的计算流程

结构的学习,使得每一个通道都拥有独立的特征聚合参数,实现特征细化.

2.7　多尺度时间维卷积模块 MS-TC

考虑到动作的快慢和持续时间的不同,在时序建模中,使用了 CTR-GCN[12]中的多尺度时间卷积. MS-TC 的结构如图 1 所示,它包含 4 个分支,其中前 3 个分支先使用 1×1 卷积将通道数降为输入通道数的 1/4,然后分别使用空洞率 D 为 1 与卷积核大小 $K_t=5$ 的空洞卷积、空洞率 D 为 2 与 $K_t=5$ 的空洞卷积和时序维最大池化操作,以获得不同时序感受野的特征;而第 4 个分支作为残差连接只使用了 1×1 卷积调整通道数为输入通道数的 1/4. 最后,将所有分支的结果进行通道拼接获得最终的输出. 在 MD^{2}TL-GCN 的 L2 和 L4 中,MS-TC 在时序维上采用了步长为 2 的卷积,实现了时序维的减半,而其他层维持时序维的大小不变.

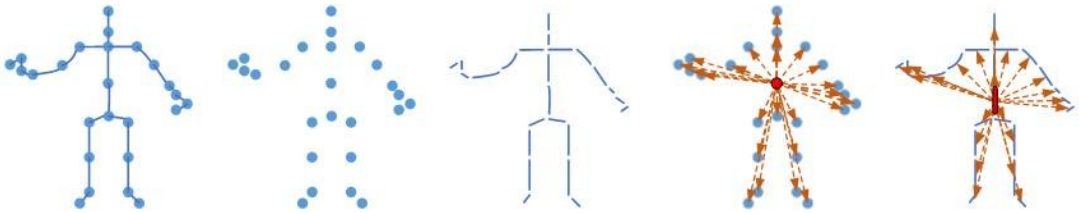
2.8　多流 MD^{2}TL-GCN

在骨架动作识别任务中,除了使用初始的节点数据以外,骨骼数据和运动信息也是十分重要的,通过使用节点流、骨骼流与它们的运动信息所构成的多流架构,可以大大增强网络识别能力. 图 5 中示例了常用的节点数据,骨骼数据,以及本文提出的相对节点数据和相对骨骼数据.

人体关节点自然连接得到的每条边作为一个骨骼,用它关联的两相邻节点的 3 维坐标的差值来表示. 而对于节点运动信息的计算,可通过关节点在相邻时间帧上的 3 维坐标差得到.

除此之外,节点相对于人体重心的位置和骨骼相对于人体重心的位置对于动作的识别也非常重要,所以我们引入了相对节点数据和相对骨骼数据. 如图 5(*d*)和图 5(*e*)所示,它们分别通过计算每个节点与中心节点之间的相对坐标向量和每个骨骼与中心骨骼之间的相对坐标向量所获得.

六流网络架构 6s-MD^{2}TL-GCN 如图 6 所示,分别包含节点数据流,相对节点数据流,节点运动数据流,骨骼数据流,相对骨骼数据流和骨骼运动数据流,它们可以并行训练和测试,最后将六个流的 softmax 得分进行相加融合,作为最终的分类结果.



(*a*) 骨架结构　(*b*) 节点数据　(*c*) 骨骼数据　(*d*) 相对节点数据　(*e*) 相对骨骼数据

图 5　各种骨架输入数据说明

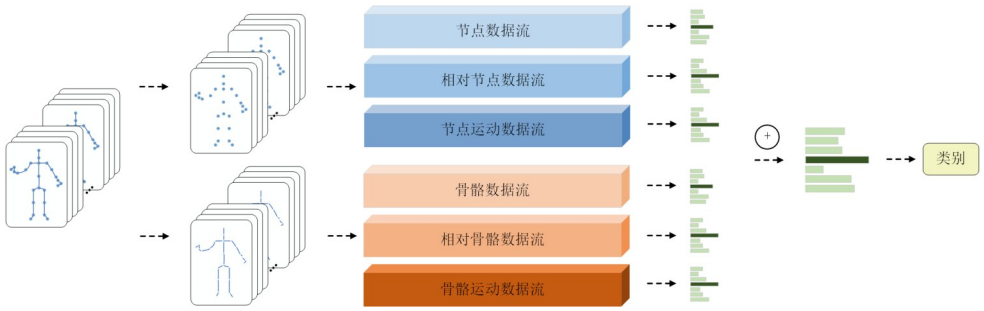


图 6　多流网络架构 6s-MD^{2}TL-GCN

3　实验

为了验证本文所提出的 MD^{2}TL-GCN 方法的优势与有效性,我们在 2 个大规模骨架动作识别数据集上进行实验. 大量消融实验的结果验证了各关键模块的有效性.

3.1　数据集与实验设置

(1)　NTU-RGB+D 数据集

NTU-RGB+D 数据集[14]包含 60 个动作类中的 56 880 个视频样本. 该数据提供了 3D 骨骼数据,其中包含每人 25 个身体关节的 3D 坐标,且每个视频样本中至多包含 2 人的身体骨架.

该数据集拥有两个评价指标,对应于两种训练集与测试集划分方式:Cross-Subject(CS)和 Cross-View(CV). CS 按照志愿者 ID 来划分训练集和测试集,训练集 40 320 个样本,测试集 16 560 个样本. CV 按相机 ID 来划分训练集和测试集,相机 1 采集的样本作为测试集,相机 2 和 3 的样本作为训练集,样本数分别为 18 960

和37 920. 三个相机的垂直高度相同,但水平角度分别为−45°、0°和45°.

(2) NTU-RGB+D 120数据集

NTU-RGB+D 120[15]数据集是目前用于骨架动作识别最大的数据集,共有114 480个骨架序列样本,包含120个动作类别,由106名志愿者执行,使用三个不同视角的摄像头捕获. 此数据集同样包含两种评价指标:X-Sub(Cross-Subject)和X-Set(Cross- Setup). X-Sub按照志愿者ID来划分训练集和测试集,训练数据来自53个志愿者的动作样本,测试数据来自其他53个志愿者的动作样本. X-Set按摄像设置ID来划分训练集和测试集,训练数据来自摄像设置ID为偶数的样本,测试数据来自摄像设置ID为奇数的样本.

(3) 实验设置

所有实验都是基于4块Tesla P100 GPU设备进行,采用交叉熵损失函数和Nesterov动量为0.9的随机梯度下降(SGD)策略作为网络优化策略. 在训练中,训练批次大小为64;权重衰减设置为0.000 1;并且为了使训练更稳定,在训练前五代使用warmup策略[17];在NTU-RGB+D数据集与NTU-RGB+D 120数据集中,初始学习率都设置为0.1,在第35代与第55代分别进行系数为0.1的学习率衰减,一共训练75代.

3.2 消融实验

本小节的消融实验在NTU-RGB+D和NTU-RGB+D 120数据集上分别使用CS和X-Sub评价指标进行比较,而且所有参与比较的方法只使用关节点坐标序列作为输入.

3.2.1 MD²TL-GC的有效性

为了验证MD²TL-GC模块的有效性,我们复现5层的ST-GCN[2]网络作为比较的基线网络(表示为ST-GCN_L5),实验结果如表1所示. 在表1中,"ST-GCN_L5+MS-TC"表示在ST-GCN_L5基础上增加残差连接,并将它的时间卷积替换为MS-TC;"ST-GCN_L5+MS-TC+*i*MD²"表示在"ST-GCN_L5+MS-TC"基础上,将前*i*层图卷积层替换成MD²TL-GC. 同时,我们还复现了通道特异性图卷积网络CTR-GCN[12],比较了只有5层图卷积的CTR-GCN(表示为"CTR-GCN_L5")和10层图卷积的CTR-GCN,以验证本文提出的多维动态拓扑图卷积MD²TL-GC的优越性.

从表1可以看出,在CS和X-Sub评价指标上,当ST-GCN_L5的图卷积层逐步替换成MD²TL-GC后,网络的识别准确率在逐步提升,当全部替换之后,MD²TL-GCN的识别准确率分别比ST-GCN_L5+MS-TC高了3.23%和5%. 与同是5层图卷积的CTR-GCN_L5相比,MD²TL-GCN的识别准确率分别高出0.56%和0.36%,比10层图卷积的CTR-GCN分别高0.21%和0.02%. 这些实验结果充分验证了我们所提出的MD²TL-GC的有效性.

表1 MD²TL-GC的有效性分析

方法	CS/%	X-Sub/%
ST-GCN_L5	86.00	79.17
ST-GCN_L5+MS-TC	86.37	79.32
ST-GCN_L5+MS-TC+1MD²	88.26	82.66
ST-GCN_L5+MS-TC+2MD²	88.82	83.85
ST-GCN_L5+MS-TC+3MD²	89.05	83.80
ST-GCN_L5+MS-TC+4MD²	89.42	84.22
MD²TL-GCN	**89.60**	**84.32**
CTR-GCN_L5	89.14	83.96
CTR-GCN	89.39	84.30

3.2.2 各个模块的有效性

为了验证J-GC、DTW-GC、CW-GC与MS-TC对模型性能的影响,我们进行了大量的消融实验,实验结果如表2所示. 表2中的TCN模块为ST-GCN中的时序建模方法,它只是在时间维上进行了卷积核大小为9的一维卷积. 实验结果总结如下:

(1)比较表2中A到F行可以发现,相比于基准模型ST-GCN_L5,在分别添加J-GC、DTW-GC、CW-GC后,模型识别准确率都得到了显著提升,在NTU-RGB+D数据集上的CS准确率分别提高了1.84%、2.39%和2.75%;在NTU-RGB+D 120数据集上的X-Sub准确率分别提高了2.48%、3.08%和3.15%. 这充分验证了我们所提出的J-GC、DTW-GC与CW-GC的有效性. 当用MS-TC替换TCN后,在两个数据集上的性能分别提升了0. 37%和0. 15%.

(2)比较表2中B到D行可以发现,在NTU-RGB+D数据集和NTU-RGB+D 120数据集上,DTW-GC中单独使用TIB比基线ST-GCN_L5的识别准确率分别提高了1.8%和2.77%,DTW-GC中单独使用TGB比基线的识别准确率分别提高了2.02%和2.34%,这充分验证了我们提出的DSTL和TIB的有效性. 此外,在融合TIB和TGB后,识别准确率进一步分别提升了0.59%和0.31%,这证明了DSTL和TIB能学习到互补特征,从而提高模型的性能.

(3)比较表2中G-I行和MD²TL-GCN可以发现,MD²TL-GCN的识别准确率比任意两个维度拓扑建模的组合都更高,这验证了多维拓扑建模的有效性.

3.3 多流融合模型

为了分析多流数据输入对MD²TL-GCN性能的影响,特别是本文提出的相对节点流数据输入和相对骨骼流数据输入的有效性,我们在NTU-RGB+D和NTU-RGB+D 120数据集上比较了MD²TL-GCN在不同数据流组合下的识别准确度. 实验结果如表3和表4所示,其中"Js"、"Bs"、"RJs"、"RBs"、"JMs"和"BMs"分别表示

表2　提出的各模块对性能的影响

方法	空间建模				时序建模		NTU-RGB+D	NTU-RGB+D 120
	J-GC	DTW-GC		CW-GC	TCN	MS-TC	CS/%	X-Sub/%
		TIB	TGB					
ST-GCN_L5					√		86.00	79.17
A	√				√		87.84	81.65
B		√			√		87.80	81.94
C			√		√		88.02	81.51
D		√	√		√		88.39	82.25
E				√	√		88.75	82.32
F						√	86.37	79.32
G	√	√	√			√	89.23	83.61
H	√			√		√	89.13	84.00
I		√	√	√		√	89.40	83.87
MD^{2}TL-GCN	√	√	√	√		√	**89.60**	**84.32**

节点流数据输入、骨骼流数据输入、相对节点流数据输入、相对骨骼流数据输入、节点运动流数据输入和骨骼运动流输入;“2s”指“Js”与“Bs”两流融合;“b4s”指“Js”、“Bs”、“JMs”和“BMs”四流融合;“4s”指“Js”、“Bs”、“RJs”和“RBs”的四流融合. 可以看到在两个大规模数据集上,使用多流融合的方法会大幅好于基于单流的方法. 而且基于相对节点流数据输入和相对骨骼流数据输入的四流网络“4s-MD^{2}TL-GCN”相比于以往的四流数据融合方法,即“b4s-MD^{2}TL-GCN”具有明显的优势,在NTU-RGB+D 120上的X-Sub准确率提升了0.53%,X-Set准确率提升了0.46%.

表3　NTU-RGB+D数据集上多流融合性能对比

方法	NTU-RGB+D	
	CS/%	CV/%
Js MD^{2}TL-GCN	89.60	94.49
Bs MD^{2}TL-GCN	90.38	94.69
RJs MD^{2}TL-GCN	89.58	94.88
RBs MD^{2}TL-GCN	90.19	94.47
JMs MD^{2}TL-GCN	87.69	93.21
BMs MD^{2}TL-GCN	87.15	91.92
2s MD^{2}TL-GCN	91.90	96.01
b4s MD^{2}TL-GCN	92.31	96.27
4s MD^{2}TL-GCN	92.42	**96.47**
6s MD^{2}TL-GCN	**92.64**	96.36

3.4　与其他先进骨架动作识别算法的比较

MD^{2}TL-GCN在两个大规模骨架动作识别数据集NTU-RGB+D和NTU-RGB+D 120上与其他先进方法的比较结果如表5和表6所示,其他方法的准确度均来自于他们论文中报导的结果. 从表5和表6中的结果可以看出,MD^{2}TL-GCN几乎在所有指标上优于当前所有主流方法,特别是在NTU-RGB+D 120数据集的X-Sub评价指标和NTU-RGB+D数据集的CS评价指标上,与当前最先进的方法CTR-GCN[12]相比,识别准确率分别提升了0.39%和0.24%.

表4　NTU-RGB+D 120数据集上多流融合性能对比

方法	NTU-RGB+D 120	
	X-Sub/%	X-Set/%
Js MD^{2}TL-GCN	84.32	85.79
Bs MD^{2}TL-GCN	86.10	87.20
RJs MD^{2}TL-GCN	84.44	86.26
RBs MD^{2}TL-GCN	85.87	87.10
JMs MD^{2}TL-GCN	80.78	83.01
BMs MD^{2}TL-GCN	81.22	82.72
2s MD^{2}TL-GCN	88.52	89.51
b4s MD^{2}TL-GCN	88.64	89.91
4s MD^{2}TL-GCN	89.17	90.37
6s MD^{2}TL-GCN	**89.29**	**90.49**

我们还进一步比较了MD^{2}TL-GCN与ST-GCN[2]、AAGCN[5]和CTR-GCN[12]算法在NTU-RGB+D数据集中的“书写”、“阅读”、“吃饭”、“玩手机或平板”、“触头或头痛”、“打喷嚏或咳嗽”、“喝水”和“恶心或呕吐”类上仅使用节点流作为输入的CS识别准确度. 这些类是文献报导中性能较低的动作类,是公认的比较难识别的动作. 实验比较结果如图7所示,MD^{2}TL-GCN在所有困难类上的性能都优于其他三种模型. 特别是在“玩手机或平板”,MD^{2}TL-GCN相比于ST-GCN[2]、AAGCN[5]和CTR-GCN[12]算法,识别准确率分别提升了20%、8.73%和3.28%. 这说明与其他图卷积算法相比,我们所提出的MD^{2}TL-GCN可以有效地提取到长时动作中多种微妙的运动模式特征,并对其进行更加准确的分类.

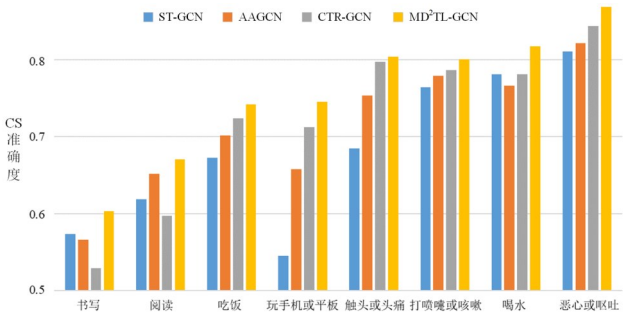


图7 MD²TL-GCN与其他先进算法在困难类中识别准确率

表5 NTU-RGB+D数据集的比较

方法	年份	NTU-RGB+D	
		CS/%	CV/%
Deep LSTM[14]	2016	60.7	67.3
ST-LSTM[18]	2016	69.2	77.7
HCN[19]	2018	86.5	91.1
ST-GCN[2]	2018	81.5	88.3
ASGCN[8]	2019	86.8	94.2
MSAGC-SRU[3]	2022	87.3	92.7
2s-AGCN[4]	2019	88.5	95.1
SGN[13]	2020	89.0	94.5
2s-AAGCN[5]	2020	89.4	96.0
MS-AAGCN[5]	2020	90.0	96.2
MS-G3D[9]	2020	91.5	96.2
SMotif-GCN[7]	2022	91.7	96.7
CTR-GCN(4s)[12]	2021	92.4	**96.8**
2s MD²TL-GCN	—	91.90	96.01
4s MD²TL-GCN	—	92.42	96.47
6s MD²TL-GCN	—	**92.64**	96.36

表6 NTU-RGB+D 120数据集的比较

方法	年份	NTU-RGB+D 120	
		X-Sub/%	X-Set/%
ST-LSTM[18]	2016	55.7	57.9
SGN[13]	2020	79.2	81.5
MS-G3D[9]	2020	86.9	88.4
DynamicGCN[6]	2020	87.3	88.6
SMotif-GCN[7]	2022	88.4	88.9
InfoGCN(2s)[10]	2022	88.5	89.7
CTR-GCN(2s)[12]	2021	88.7	90.1
CTR-GCN(4s)[12]	2021	88.9	**90.6**
2s MD²TL-GCN	—	88.52	89.51
4s MD²TL-GCN	—	89.17	90.37
6s MD²TL-GCN	—	**89.29**	90.49

4 结论

本文提出了一种新颖的基于多维动态拓扑学习图卷积的骨架动作识别方法，它可以动态地建模同时具有时序特异性与通道特异性的人体骨架拓扑结构. 在时序维特异性拓扑建模中，通过学习动态骨架拓扑图，增强全局时空动态信息. 另外，我们还提出了相对节点数据流和相对骨骼数据流用于补充骨架信息. 在两个大规模的公共骨架动作识别数据集NTU-RGB+D和NTU-RGB+D 120上的实验结果证明了动态骨架拓扑和多维特异性拓扑能学习到更细微的动作差别，提出的网络模型MD²TL-GCN在只堆叠5层时就可以超越当前所有的主流方法.

参考文献

[1] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 1302-1310.

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

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

[4] 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). Piscataway: IEEE, 2020: 12018-12027.

[5] 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: a Publication of the IEEE Signal Processing Society, 2020, 29: 9532-9545.

[6] 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. New York: ACM, 2020: 55-63.

[7] WEN Y H, GAO L, FU H B, et al. Motif-GCNs with local and non-local temporal blocks for skeleton-based action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2009-2023.

[8] LI M S, CHEN S H, CHEN X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 3590-3598.

[9] 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 (CVPR). Piscataway: IEEE, 2020: 140-149.

[10] CHI H G, HA M H, CHI S, et al. InfoGCN: Representation learning for human skeleton-based action recognition [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 20154-20164.

[11] CHENG K, ZHANG Y F, CAO C Q, et al. Decoupling GCN with DropGraph module for skeleton-based action recognition[C]//Computer Vision ECCV 2020. Cham: Springer International Publishing, 2020: 536-553.

[12] 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). Piscataway: IEEE, 2022: 13339-13348.

[13] 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). Piscataway: IEEE, 2020: 1109-1118.

[14] 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 (CVPR). Piscataway: IEEE, 2016: 1010-1019.

[15] 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.

[16] BILEN H, FERNANDO B, GAVVES E, et al. Dynamic image networks for action recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 3034-3042.

[17] 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 (CVPR). Piscataway: IEEE, 2016: 770-778.

[18] LIU J, SHAHROUDY A, XU D, et al. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 3007-3021.

[19] LI C, ZHONG Q Y, XIE D, et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm: International Joint Conferences on Artificial Intelligence Organization, 2018: 786-792.

作者简介

罗会兰　女,1974年9月生于江西上高.现为江西理工大学图像处理实验室教授、硕士生导师.主要从事机器学习、模式识别等方面的研究.

E-mail: luohuilan@sina.com

曹立京　男,1997年10月生于江西赣州.现为江西理工大学信息工程学院硕士研究生,研究方向为骨架动作识别.

E-mail: 2870256076@qq.com