

# 基于轻量级语义信息融合的动作识别方法

束 阳,李汪根,高 坤,王志格,葛英奎  
(安徽师范大学 计算机与信息学院,安徽 芜湖 241002)

**摘 要:**针对目前大多数的动作识别方法使用深层网络训练模型导致模型参数量大、验证成本高以及语义信息利用不足等问题,提出一种基于轻量级语义信息融合的动作识别方法(LSIF-GCN),实现了模型的轻量化和对语义信息的充分利用。首先,LSIF-GCN将数据预处理后的关节流、速度流和骨骼流三种不同的输入信息编码至高维空间后,经过一层图卷积操作,以达到特征增强和降低维度的目的,再把三种信息流在通道维度上进行拼接融合。然后,为了充分利用语义信息提取不同关节之间潜在的权重关系,提出一种“瓶颈型”的四层图卷积模块。最后,采用分流网络设计的时间卷积模块,并引入自注意力机制,在减少模型参数量的同时也提高了网络的性能。该模型具有简单的结构和训练过程,便于在低成本的嵌入式设备的实时动作识别系统中部署。在 NTU-RGB+D 60 和 NTU-RGB+D 120 数据集上的大量实验表明,该方法不仅在识别精度和模型复杂度(参数量和 GFLOPs)上优于目前一些主流的轻量级方法,而且与一些近几年的 SOTA 方法相比也具有一定的优势。

**关键词:**语义信息;动作识别;轻量级;自注意力;分流网络

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2023)06-0181-08

doi:10.3969/j.issn.1673-629X.2023.06.027

## Action Recognition Method Based on Lightweight Semantic Information Fusion

SHU Yang, LI Wang-gen, GAO Kun, WANG Zhi-ge, GE Ying-kui  
(School of Computer & Information, Anhui Normal University, Wuhu 241002, China)

**Abstract:** Aiming at the problems that most of the current action recognition methods use deep networks to train models, which leads to large amount of model parameters, high verification cost and insufficient utilization of semantic information, an action recognition method based on lightweight semantic information fusion (LSIF-GCN) is proposed, which realizes the lightweight of the model and the full use of semantic information. First of all, LSIF-GCN encodes three different input information of joint flow, velocity flow and bone flow after data pretreatment into a high-dimensional space, and then goes through a layer of graph convolution operation to achieve the purpose of feature enhancement and dimension reduction. Then, the three information flows are spliced and fused in the channel dimension. Secondly, we propose a "bottleneck" four-layer graph convolution module to make full use of semantic information to extract the potential weight relationship between different joints. Finally, the time convolution module designed by diversion network is adopted, and the self-attention mechanism is introduced, which not only reduces the number of model parameters but also improves the network performance. The model has simple structure and training process, which can be easily deployed in real-time motion recognition system of low cost embedded devices. A large number of experiments on the NTU-RGB+D 60 and NTU-RGB+D 120 dataset show that the proposed method not only outperforms some mainstream lightweight methods in recognition accuracy and model complexity (parameter number and GFLOPs), but also has certain advantages compared with some SOTA methods in recent years.

**Key words:** semantic information; action recognition; lightweight; self-attention; distribution network

## 0 引 言

近些年来,动作识别在计算机视觉领域上扮演着愈来愈重要的角色,充满了挑战性和吸引力。目前,动

作识别在视频监控、人机交互、体育运动<sup>[1-2]</sup>等领域都有着远大的前景。传统的基于 RGB<sup>[3]</sup>图像的动作识别容易受到各种因素的干扰:如光暗程度、摄像机角度

收稿日期:2022-08-24

修回日期:2022-12-28

基金项目:国家自然科学基金项目(61976006);安徽省教育厅高校领军人才引进与培育计划项目(051619)

作者简介:束 阳(1997-),男,硕士研究生,研究方向为深度学习和骨骼识别;通信作者:李汪根(1973-),男,硕士,博士,教授,研究方向为生物计算、智能计算。

和人体自身遮挡或其他遮挡等,往往会造成关键信息的丢失。而当前主流的基于人体骨骼数据的动作识别方法<sup>[4-5]</sup>仅仅关注关节的坐标位置信息,不受这些因素的干扰,因而在动态变化的复杂环境中具有良好的适应性、鲁棒性和稳定性。而且,基于人体骨架数据的动作识别计算量远小于 RGB 图像,这使其应用在灵活的移动设备上成为可能,具有广泛的应用前景。

早期的用于人体骨架的建模,如递归神经网络<sup>[6]</sup>(Recurrent Neural Network, RNN)和卷积神经网络<sup>[7]</sup>(Convolutional Neural Network, CNN)等。这些模型可以同时提取人体关节之间的时空信息,却不能表现非欧式空间下关节间的图形关系,不适用于探索重要关节之间的相关性,导致训练网络时大量动作信息丢失。并且,基于 RNN 的方法需包含大量的计算参数,效率低。

随着图卷积网络(Graph Convolution Network, GCN)<sup>[8]</sup>的出现,研究人员发现,基于 GCN 的模型比其他模型具有更好的性能。Yan 等人<sup>[9]</sup>首先提出时空图卷积神经网络(Spatial Temporal Graph Convolutional Networks, ST-GCN),骨架数据表示为图形数据,自然骨架连接用于构建每个骨架图的邻接矩阵。Shi 等人<sup>[10]</sup>认为不同的关节之间也具有相关性,固定的图形拓扑结构会限制模型的性能,因此提出双流自适应图卷积神经网络(Two-Stream Adaptive Graph Convolutional Networks, 2s-AGCN),其中引入了自我注意系数和自我学习的图形残差掩码来捕捉不同关节之间的关系。同时,添加骨骼流以提高 2s-AGCN 的性能。在此基础上,Sun 等人<sup>[11]</sup>提出基于骨架动作识别的多流快慢图卷积网络(Multi-stream slowFast graph convolutional networks for skeleton-based action recognition, MSSF-GCN),其中除了关节的坐标信息外,还引入了五个高阶序列,包括骨骼边、关节和骨骼边的空间差异和时间差异,以增强人类行为的表示。Fang 等人<sup>[12]</sup>认为平等的对待骨架的每一帧需要一个大规模 GCN 模型来建模,这会造成大量的冗余信息,因此提出基于骨架动作识别的时空慢快速图卷积网络

(STSF-GCN), STSF-GCN 包含快速路径和慢速路径,可以有效地捕获长程和短程时空联合关系,以更低的计算成本实现更先进的性能。多流网络虽然动作分析的准确度较高,但由于参数量过高和计算量过大,导致其不易在移动设备上应用。

早期,在机器翻译和图像识别等领域, Ashish 等人<sup>[13]</sup>和 Zheng 等人<sup>[14]</sup>已经利用了与语义相关的显式探索。他们分别对序列中标记的位置进行编码和将组索引编码为卷积信道表示,借此来保留任务中的时序信息。最近, Zhang 等人<sup>[15]</sup>提出了语义引导的神经网络(semantics-guided neural networks for efficient skeleton-based human action recognition, SGN),在 GCN 模型中显性地添加了骨骼关节和帧索引的语义信息,以保留空间身体结构的重要信息和动作在时间上的连贯性,对模型性能的提升起到很大的作用。Jing 等人<sup>[16]</sup>提出了一种轻量级多信息图卷积神经网络(LMI-GCN)。引入由语义信息(关节类型和帧索引)拼接成的自适应图,来聚合重要关节特征,并且提出一种分流设计的时间卷积块来降低模型参数量和计算量,然而却没有充分地利用语义信息。网络参数量较高,导致不适合应用于实际场景。

针对以上问题,该文以 LMI-GCN 为基线,通过多信息输入模块提取骨架重要数据信息,并提出能充分利用语义信息来提取骨骼关节特征的图卷积模块。而且,网络模型中的自注意力时间模块在提升模型性能的同时减少了参数量,总体来说,LSIF-GCN 比文献[16]模型复杂度更小,识别精度更高,模型效率更强。

主要贡献如下:(1)提出一种高效的基于轻量级语义信息融合的人类骨骼动作识别方法(LSIF-GCN);(2)提出的“瓶颈型”的图卷积模块,充分利用了语义信息以提高模型性能;(3)针对基线中的分流时间卷积块,引入了压缩和激励(SE)模块<sup>[17]</sup>并且使用再分组的方法,使模型更小,识别精度更高;(4)在目前流行的 NTU-RGB+D60<sup>[18]</sup>和 NTU-RGB+D120<sup>[19]</sup>大规模数据集上进行多次实验证明 LSIF-GCN 是一种高效的轻量级方法。

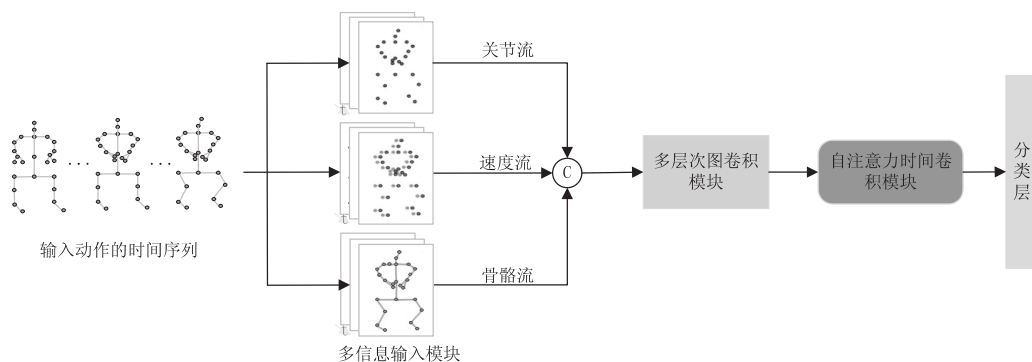


图 1 LSIF-GCN 整体模型框架

## 1 模型结构

该文提出一种多层次高级语义融合的骨骼动作识别模型,网络的整体结构如图1所示。网络模型整体结构由三部分组成,分别是多信息输入模块、图卷积模块和自注意力时间卷积模块。

### 1.1 多信息输入模块

根据先前的研究<sup>[20-21]</sup>表示,骨骼边信息对基于人体骨架的动作识别至关重要。该文在数据预处理阶段提取了六种输入序列信息,分别是关节位置信息、关节相对位置信息、关节速度信息、骨骼边信息、骨骼边速度信息和骨骼边加速度信息。

假设  $G_t(V, E)$  表示第  $t$  帧的人体骨架图,其中  $V$  表示所有关节节点的集合,  $E$  表示所有骨骼边的集合。可以从原始骨架坐标获得关节数据,表示为:

$$p_{i,t} = (x_{i,t}, y_{i,t}, z_{i,t}) \in \mathbb{R}^3 \quad (1)$$

式中,  $i = 1, 2, \dots, N, t = 1, 2, \dots, T$ 。  $p_{i,t}$  表示第  $t$  帧中关节  $i$  的3D坐标位置,  $N$  是关节节点总数量,  $T$  是骨骼序列帧的总数。

关节相对位置可以通过从其他关节的坐标减去中心关节的坐标得出。对于不同的应用场景和不同任务,中心关节可以进行调整。如果更多关注集中在手上的动作细节,如手势识别,可以将中心关节设定为指关节或腕关节。考虑到该文关注的是对人体各种动作的识别,因此,将动作序列的脊椎上部和下部,以及手掌腕部三个关节节点确定为中心关节,捕捉相似动作之间的细微差别,如式(2)所示。

$$p'_{i,t} = p_{i,t} - p_{c,t} \quad (2)$$

其中,  $p_{c,t}$  表示的是中心关节。骨骼边指的是人体骨架各个关节节点之间的自然连接拓扑图,是有大小和方向的向量,在人体骨骼动作识别中起着重要的作用。把远离中心关节的关节节点称为目标关节节点,把靠近中心关节的关节节点称为源关节节点。可以表示为:

$$e_{i,t} = p_{i,t} - p_{s,t} \quad (3)$$

其中,  $e_{i,t}$  表示骨骼边信息,  $p_{s,t}$  表示源关节节点。为了更好地了解关节和骨骼的变化信息,提升模型的性能,引入了速度信息和加速度信息。速度信息指相邻帧之间的关节坐标位置和骨骼边位置的变化,加速度信息表示骨骼速度在相邻帧之间的差值,如式(4)所示。相邻帧之间时间差值为1,由于时间是线性的,所以帧差的数量会少一个,为此,该文在速度和加速度信息的第一帧中添加一个值为0的向量。

$$\begin{aligned} v &= p_{i,t} - p_{i,t-1} \\ v' &= e_{i,t} - e_{i,t-1} \\ a &= v'_{i,t} - v'_{i,t-1} \end{aligned} \quad (4)$$

式中,  $v$  表示关节速度信息,  $v'$  表示骨骼边的速度信

息。 $a$  表示骨骼边的加速度信息。

文献[16]对输入信息的处理是将四种信息均先编码到高维空间后再进行相加融合,一定程度上减少了模型的参数量和计算成本,但是相加融合会使一部分信息丢失。因此,该文先在低维空间中对输入信息进行通道维度的拼接融合,再编码到高维空间,通过文献[21]提出的 stgcnn 模块对输入数据进行特征提取和降维。最后,将三种输入数据进行拼接融合,作为图卷积模块的输入。经过数据预处理后的输入特征主要分三类:(1)关节流;(2)速度流;(3)骨骼流。

$$\begin{aligned} \tilde{p} &= \text{stgcnn}(\text{embed}(\text{cat}[p, p'])) \\ \tilde{v} &= \text{stgcnn}(\text{embed}(v)) \\ \tilde{e} &= \text{stgcnn}(\text{embed}(\text{cat}[e, v', a])) \\ f_{in} &= \text{cat}[\tilde{p}, \tilde{v}, \tilde{e}] \end{aligned} \quad (5)$$

式中,  $\tilde{p}$  表示关节流,  $\tilde{v}$  表示速度流,  $\tilde{e}$  表示骨骼流。 $f_{in}$  表示下一层的输入, cat 表示拼接操作, embed 是编码操作,由两个  $1 \times 1$  的卷积组成,达到升维的目的。stgcnn 是图卷积操作,如式(6)所示。

$$f'_{out} = \sigma \left( \sum_{d=0}^D W_d (f'_{in} (\Lambda^{-\frac{1}{2}} A_d \Lambda^{-\frac{1}{2}} \odot M_d) + W_{in} f_{in}) \right) \quad (6)$$

其中,  $D$  是人体关节图预定义关节节点之间的最大距离,  $f'_{in}$  和  $f'_{out}$  表示输入和输出的特征图,  $\odot$  表示逐点卷积,  $A_d$  表示图形距离为  $d$  的关节对的邻接矩阵,  $\Lambda$  用来规范化  $A_d$ 。 $\sigma$  是 Relu 激活函数,  $W_d$ 、 $W_{in}$  和  $M_d$  都是可学习的参数。

### 1.2 图卷积模块

在之前的工作中<sup>[15,22-23]</sup>,已经证实将高级骨骼关节的语义信息引入到图卷积中,对模型性能的提高起到显著的作用。早期工作<sup>[15,23]</sup>是简单的把语义信息与输入信息进行融合,然而随着图卷积层的深入,某些重要的语义信息会丢失,模型的效率也会下降。

针对这一问题,文献[16]提出了一种简单的自适应图  $G$ 。图  $G$  是网络训练数据过程中自动学习而来,可以聚合不同通道权重的关节特征。然而,图  $G$  聚合的通道有所单一,语义信息优势不能完全表现出来。

因此,该文提出一种“瓶颈型”结构的四层图卷积模块,前两层图卷积提取特征并降维,后两层图卷积在提取特征的同时进行升维。“瓶颈型”的四层图卷积结构不仅比基线方法<sup>[16]</sup>提出的三层图卷积的参数量低,模型性能也好。

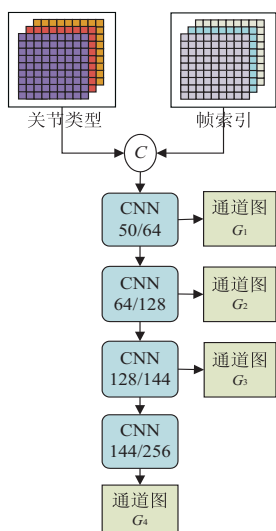
如图2所示,为了匹配四层图卷积,充分利用语义信息,该文提出一种新的多通道自适应图  $G$ 。图  $G$  使用了四层卷积进行学习,利用卷积核参数的自动学习机制,可以根据模型训练过程中输入数据的变化不断



自适应调整,如式(7)所示:

$$\begin{aligned} G_0 &= \text{cat}[s_p, s_t] \\ G_j &= \sigma(\text{conv}(G_{j-1})) \end{aligned} \quad (7)$$

其中,  $s_p$  和  $s_t$  分别指的是关节类型和帧索引语义信息, cat 指的是通道拼接操作。  $G_j$  是训练不同通道的学习图,  $j = 1, 2, 3, 4$ 。 conv 是  $1 \times 1$  的卷积。通过多层卷积操作得到图  $G_j$  后,再将其输入到图卷积中。



注: 关节类型和帧索引语义信息大小均为  $N \times N \times N$

图2 多通道自适应图  $G$

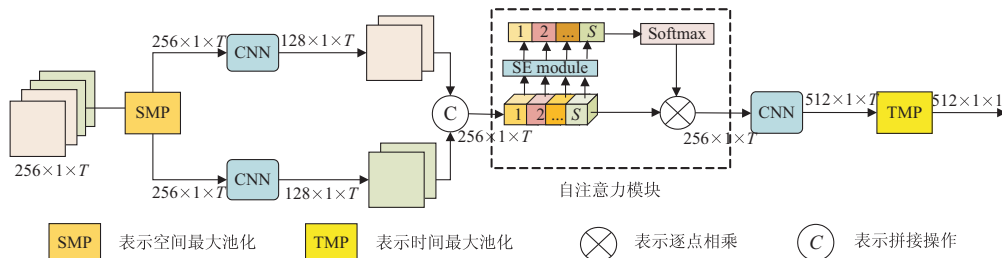


图3 自注意力时间卷积模块

## 2 实验

### 2.1 数据集

NTU RGB+D 60 数据集<sup>[18]</sup>: NTU 60 是当前主流的基于骨架动作识别的数据集之一, 包含从 40 个不同的主题和 3 个 Microsoft Kinect V2 深度摄像头同时捕获的 60 个动作类别的 56 880 个骨架序列。每个骨架序列包含 25 个关节的三维空间坐标。该数据集提供了两个评估基准: 交叉主题 (Cross-Subject, CS) 和交叉视角 (Cross-View, CV)。CS 由 40 名受试者完成动作, 其中一半受试者用于培训, 其余用于测试。CV 选择摄像机 2 和 3 捕获的样本进行训练, 其余用于测试。

NTU-RGB+D 120 数据集<sup>[19]</sup>: NTU 120 是一种大型室内捕捉的 3D 人体动作识别的公共数据集。该数据集是 NTU RGB+D 60 数据集在动作类别和演员人数上的拓展, 包含 106 名演员参与的 114 480 个动作视

$$f_{\text{out}} = \sigma(\text{conv}(f_{\text{in}} \otimes G_j) + \text{conv}(f_{\text{in}})) \quad (8)$$

如式(8)所示,  $f_{\text{in}}$  和  $f_{\text{out}}$  分别为图卷积的输入和输出,  $\otimes$  表示矩阵相乘,  $\sigma$  为 Relu 激活函数, conv 为  $1 \times 1$  的卷积, 具有不同的训练权重。

### 1.3 自注意力时间卷积模块

众所周知, 分组卷积可以减少卷积操作的参数量, 但缺陷却是不同组的特征图之间不能进行通信, 这会降低网络的特征提取能力。受文献[24]的启发, 本文提出了一个自注意力时间卷积模块, 如图3所示。

在时间卷积模块中, 首先对输入信息最大池化以聚合关节空间信息, 再将信息在通道维度上分成多组, 每一组使用的卷积核依次增大, 如卷积核大小为  $1 \times 1$ ,  $1 \times 3$ , 以获取不同尺度的感受野, 提取不相邻帧之间的特征信息。考虑到使用大卷积核所产生的参数量和计算量也会随之增大, 因此, 对每一组再次进行分组卷积, 如卷积核为  $1 \times 3$  时, 分组数量为 2, 在经过不同大小的卷积后, 在通道上进行拼接。然后, 通过 SE 注意力模块对每组的通道的权重值进行提取, 对每组的权重值进行 Softmax 归一化并加权, 再将这些组进行拼接。自注意力模块的结构如图3虚线框所示。

最后, 在经过一个  $1 \times 1$  卷积进行特征提取和通道升维后, 通过时间的最大池化聚合全局帧特征。

频。该数据集共有 120 个动作类别, 包括 82 种日常生活, 12 种医疗条件和 26 种两人互动情况下的动作。该数据集有两个评估基准: 交叉主题 (Cross-s-Setup, SS)。CS 根据视频中的不同演员将此数据集划分为训练集 (63 026 个视频) 和验证集 (50 919 个视频)。SS 是根据视频编号奇偶来划分数据集。54 468 个偶数视频作为训练集, 59 477 个奇数视频作为测试集。

### 2.2 实验细节

在多信息输入阶段, 将三类信息流在通道维度上编码到 64 维, 通过公式(5)降到 48 维后进行拼接, 拼接得到的通道维数为 144。在图卷积模块中, 每层图卷积的输入维数分别为 144、128、64 和 128, 最终输出维度为 256。对于空间和时间语义信息, 本文对它们拼接后形成的自适应图  $G$  进行升维, 维度大小与每层图卷积的输入维度相同。值得注意的是, 最终输出数据与图  $G$  通过相加融合后输入自注意力时间卷积模

块。关于时间卷积的过程,已在1.3节中详细阐述。最后,使用全连接层与Softmax进行输出,输出维度是对应的数据集类别数。

将模型中epoch的数量设置为120,将一个epoch的批量大小设置为64,初始学习率设置为0.001,并在迭代过程中不断减小。当迭代次数为80、100时,学习率下降十倍。同时,也使用Adam对模型进行优化,其中权重衰减为0.000 1。交叉熵损失用于训练网络。该文帧索引取值为25,实验取3次测试的最佳结果,以对比其他模型的最佳结果。

实验环境配置:处理器为Intel i5-10400F,内存为32G,显卡为单块NVIDIA RTX 3060ti,操作系统为Ubuntu 20.04,编程语言为Python 3.8,开发环境是Cuda 11.3和Pytorch 1.7。

### 2.3 实验结果与分析

目前大多数主流方法使用的是多流多信息融合的输入方式,这导致模型参数量成倍增长,模型变得更为复杂。因此,在引入多信息的同时,该文尝试通过在通道维度上进行拼接来最小化参数量。

表1显示的是不同的输入信息流对模型性能的影响,其中joint表示关节流,velocity表示速度流,bone表示骨骼流。通过表1的实验结果验证得到以下两点结论:一是使用三流融合输入在CS和SS上的识别精度高于双流融合输入和单流的信息输入;二是对比多流网络<sup>[10,21]</sup>参数量成倍增长,所提信息融合方法在仅仅增加了约1/7的参数量和不足1/2的GFLOPs情况下,大幅度提升了模型的性能。GFLOPs表示推理一个样本所需的浮点运算量,单位为十亿次。

表1 LSIF-GCN在NTU 120两种评估方式上的对比

Input	Param./M	GFLOPs	CS/%	SS/%
Only joint	0.46	0.50	83.4	84.9
Only bone	0.46	0.50	84.9	85.9
Only velocity	0.46	0.50	75.8	77.5
joint and bone	0.49	0.64	85.6	86.9
joint and velocity	0.49	0.64	85.6	87.3
bone and velocity	0.49	0.64	86.0	87.6
ALL	0.53	0.78	86.6	87.9

注:Param.表示模型参数量,单位为百万(M)

在表2中,w/o stgc表示在多信息输入模块中没有使用stgc模块,w/o G表示在多层次卷积模块中没有使用多通道的自适应图。表2实验结果表明,如果不使用stgc,模型在CS和SS基准上的准确度会分别下降0.5个百分点和0.4个百分点。如果不使用多通道的自适应图G,模型准确度会分别下降2.4个百分点和2.3个百分点。原因是分别消去stgc和G后,原来的图

卷积就变成了普通的卷积,模型聚合不同关节特征之间的能力就会变弱,从而导致模型在NTU 120数据集两种基准上的准确度也随之下降。

表2 NTU 120上stgc模块和多通道自适应图的有效性对比

Methods	Param./M	CS/%	SS/%
w/o stgc	0.51	86.1	87.5
w/o G	0.40	84.2	85.6
ALL	0.53	86.6	87.9

表3显示在“瓶颈型”图卷积模块中,探究不同层数的图卷积操作对模型性能的影响。其中,LSIF-GCN(three)表示LSIF-GCN使用三层图卷积,每层的输出通道数分别是128、64、256。LSIF-GCN(five)表示模型使用五层图卷积,每层的输出通道数分别是128、64、128、144和256。该文采用的是四层图卷积结构,每层的输出通道分别为128、64、128和256。实验结果表明,使用四层图卷积时,模型能更充分利用语义信息,效果更好。

表3 在NTU 120上不同图卷积层数对模型性能的影响

Methods	Param./M	CS/%	SS/%
LSIF-GCN(three)	0.48	85.8	87.2
LSIF-GCN(four)	0.53	86.6	87.9
LSIF-GCN(five)	0.57	86.1	87.7

如表4所示,对比于基线模型<sup>[16]</sup>的时间卷积模块,所提模型在CS和SS基准上分别高出0.1个百分点和0.2个百分点,模型参数量也低于基线模型。其原因有以下两点:一是将原来分组后的通道再次分组可以降低模型参数量,而在参数量较少的情况下,分组卷积相当于正则化操作,从而防止模型出现过拟合;二是注意力模块能自动学习不同通道特征的重要程度,增强特征提取能力,从而提高模型性能。其中,Old TCN表示文献[16]使用的时间卷积网络,Old TCN w/ SE表示在该方法的基础上引入了SE注意力模块。New TCN表示该文使用的时间卷积网络,New TCN w/o SE表示在时间卷积上只使用再分组的方法,而不引用SE注意力模块。

表4 在NTU 120上时间卷积模块和注意力模块有效性对比

Methods	Param./M	CS/%	SS/%
Old TCN	0.58	86.5	87.7
Old TCN w/ SE	0.58	86.5	87.8
New TCN w/o SE	0.53	86.2	87.7
LSIF-GCN(New TCN)	0.53	86.6	87.9

图 4 显示 LSIF-GCN 和 SGN<sup>[15]</sup> 以及 LMI-GCN<sup>[16]</sup> 两种模型在 NTU 120 数据集 CS 基准上的收敛情况对比。为了公平对比,三种模型的超参数设置和数据预处理方法均保持一致。从图中可以得出结论,所提模型的收敛速度和收敛程度均优于另外两种模型。

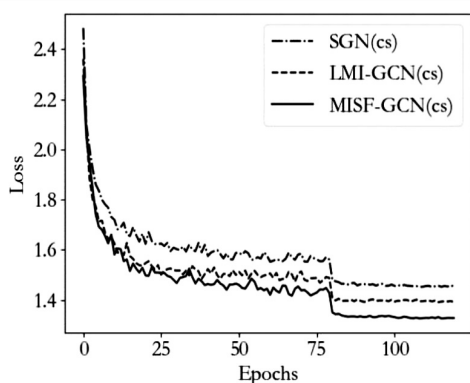


图 4 在 NTU 120 数据集 CS 评估中收敛性的对比

表 5 在 NTU 60 上与一些主流方法的比较

Methods	Year	Param. /M	CS/%	CV/%
ST-GCN	2018	3.10 *	81.5 *	88.3 *
2s-AGCN	2019	6.94 *	88.5 *	95.1 *
SGN	2020	0.69	89.0	94.5
4s-Shift-GCN	2020	2.76 *	90.7 *	96.5 *
MS-G3D	2020	6.4 *	91.5 *	96.2 *
FLAGCN	2021	0.83 <sup>+</sup>	89.4 <sup>+</sup>	94.8 <sup>+</sup>
NLB-ACSE	2021	1.21 <sup>+</sup>	91.0 <sup>+</sup>	96.1 <sup>+</sup>
4s-Shift-GCN++	2021	1.80 <sup>+</sup>	90.5 <sup>+</sup>	96.3 <sup>+</sup>
LMI-GCN	2021	0.58	89.9	94.9
PTF-GCN	2022	0.79 <sup>+</sup>	89.7 <sup>+</sup>	95.2 <sup>+</sup>
EfficientB0	2022	0.29 *	90.2 *	94.9 *
EfficientB2	2022	0.51 *	91.4 *	95.7 *
1s-ST-GCN++	2022	1.39 <sup>+</sup>	90.1 <sup>+</sup>	95.5 <sup>+</sup>
2s-ST-GCN++	2022	1.39 <sup>+</sup>	91.4 <sup>+</sup>	96.7 <sup>+</sup>
1s-LSIF-GCN (ours)	-	0.5	91.0	95.7
2s-LSIF-GCN (ours)	-	1.0	91.3	96.0

\* 表示结果来自于文献[21];+表示结果由原文作者提供。

为了验证所提模型的性能,在 NTU-RGB+D 120 数据集的 CS 和 SS 基准上比较了 LSIF-GCN 与最近几年的 SOTA 方法在准确性、模型参数数量和 GFLOPs 上的效率。表 6 实验结果显示,第一个基于骨架动作识别的图卷积模型 ST-GCN 的模型参数数量和 GFLOPs 分别是单流网络 1s-LSIF-GCN 的 5.84 倍和 20.92 倍。在模型参数数量方面,1s-LSIF-GCN 仅低于 EfficientB0,但在 SS 识别精度上却高出 2.9 百分点。而对比于所有模型的 GFLOPs,除了 SGN<sup>[15]</sup> 与 1s-LSIF-

在 NTU RGB+D 60 数据集的 CS 和 CV 基准上,与最近一些主流的方法<sup>[9-10,15-16,21-29]</sup>进行了比较。从表 5 可以看出,LSIF-GCN 在两种基准上的最佳性能分别是 91.0% 和 95.7%。表中有三种典型的方法值得注意。第一,对比于目前最为流行的基于骨骼的动作识别基线模型 ST-GCN<sup>[9]</sup>,单流网络 1s-LSIF-GCN 在 CS 和 CV 基准上分别高出 9.5 百分点和 7.4 百分点。第二,2020 年的 SOTA 方法 MS-G3D<sup>[25]</sup>,在 CS 和 CV 评估基准上的识别准确度比 1s-LSIF-GCN 均高出 0.5 百分点,而参数量却是文中的 13 倍。第三,EfficientB0<sup>[21]</sup> 和 EfficientB2<sup>[21]</sup> 是目前使用图卷积的轻量级 SOTA 方法。单流网络 1s-LSIF-GCN 参数量比 EfficientB0 高,但在两种基准上的准确度均高出 EfficientB0 模型 0.8 百分点。对比于 EfficientB2 方法,双流网络 2s-LSIF-GCN 在 CS 基准上低 0.1 百分点,而 CV 的识别精度高出 0.3 百分点。

GCN 相当外,其余模型均远高于文中方法,而在 CS 和 SS 评估基准上,该文比 SGN 分别高出 7.4 百分点和 6.4 百分点。在多流网络(2s-AGCN<sup>[10]</sup>、4s-Shift-GCN<sup>[26]</sup>、4s-Shift-GCN++<sup>[27]</sup>、2s-ST-GCN++<sup>[28]</sup>)中,双流网络 2s-LSIF-GCN 的模型参数数量和 GFLOPs 也达到了最低,SS 识别精度也仅比 2s-ST-GCN++ 低 0.6 百分点。总体来说,文中方法在识别精度和模型复杂度(参数量和 GFLOPs)上,达到了先进的水平,相比于大多数的 SOTA 方法,更适用于资源有限的移动

设备和实际应用场景。

表6 在 NTU-RGB+D 120 数据集上与 SOTA 方法的准确度、参数量和 GFLOPs 的比较

Methods	Param. /M	Ratio	GFLOPs	Ratio	CS/%	SS/%
ST-GCN	3.10 *	5.84x	16.32 *	20.92	70.7 *	73.2 *
2s-AGCN	6.94 *	13.09x	37.32 *	47.85x	82.5 *	84.2 *
SGN	0.72	1.35x	0.8	1.14x	79.2	81.5
4s-Shift-GCN	2.76 *	5.21x	10 *	12.82x	85.9 *	87.6 *
MS-G3D	6.4 *	12.07x	48.88 *	62.67x	86.9 *	88.4 *
NLB-ACSE	1.21 <sup>+</sup>	2.28x	—	—	86.2 <sup>+</sup>	88.1 <sup>+</sup>
4s-Shift-GCN++	1.80 <sup>+</sup>	3.40x	1.7 <sup>+</sup>	2.18x	85.6 <sup>+</sup>	87.2 <sup>+</sup>
EfficientB0	0.32 *	0.60x	2.73 *	3.5x	86.6 *	85.0 *
EfficientB2	0.54 *	1.02x	4.05 *	5.19x	88.0 *	87.8 *
1s-ST-GCN++	1.39 <sup>+</sup>	2.62x	2.8 <sup>+</sup>	3.59x	85.6 <sup>+</sup>	87.5 <sup>+</sup>
2s-ST-GCN++	1.39 <sup>+</sup>	2.62x	2.8 <sup>+</sup>	3.59x	87.0 <sup>+</sup>	89.1 <sup>+</sup>
1s-LSIF-GCN(ours)	0.53	1x	0.78	1x	86.6	87.9
2s-LSIF-GCN(ours)	1.06	2x	1.56	2x	87.3	88.5

\*表示结果来自于文献[21];+表示结果由原文作者提供。

### 3 结束语

针对传统的动作识别模型参数量较高和计算成本较大的问题,通过通道维度上的拼接操作将多种输入信息进行融合,避免了各种特征之间的相互干扰,提高了特征的利用率,同时提出了一种能够充分利用语义信息的“瓶颈型”的四层图卷积模块,并且将激励和压缩(SE)模块融入到时间卷积模块中,通过再分组的方法,提高模型效率的同时降低了模型的参数量。最后,在目前主流的动作识别数据集的实验结果表明,该模型在计算成本、模型参数量和识别精度上,与其他的主流方法对比,具有一定的优越性。

在未来工作中,笔者将针对特定场景更加注重细微动作间的区分,如手势识别,将该方法应用于特定场景中。

#### 参考文献:

- [1] D'SA A G, PRASAD B G. A survey on vision based activity recognition, its applications and challenges [C]//2019 second international conference on advanced computational and communication paradigms (ICACCP). Gangtok: IEEE, 2019:1-8.
- [2] GUPTA A, GUPTA K, GUPTA K, et al. A survey on human activity recognition and classification [C]//2020 international conference on communication and signal processing. Chennai: IEEE, 2020:915-919.
- [3] 梁 绪, 李文新, 张航宇. 人体行为识别方法研究综述[J]. 计算机应用研究, 2022, 39(3):651-660.
- [4] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human action recognition by representing 3d skeletons as points in a lie group [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Columbus: IEEE, 2014: 588-595.
- [5] LIU M, LIU H, CHEN C. Enhanced skeleton visualization for view invariant human action recognition [J]. Pattern Recognition, 2017, 68:346-362.
- [6] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston: IEEE, 2015:1110-1118.
- [7] DU Y, FU Y, WANG L. Skeleton based action recognition with convolutional neural network [C]//2015 3rd IAPR Asian conference on pattern recognition (ACPR). Kuala Lumpur: IEEE, 2015:579-583.
- [8] 徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述 [J]. 计算机学报, 2020, 43(5):755-780.
- [9] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition [C]//Thirty-second AAAI conference on artificial intelligence. Menlo Park: AAAI, 2018:7444-7452.
- [10] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE, 2019:12026-12035.
- [11] SUN N, LENG L, LIU J, et al. Multi-stream slowFast graph convolutional networks for skeleton-based action recognition [J]. Image and Vision Computing, 2021, 109:104141.
- [12] FANG Z, ZHANG X, CAO T, et al. Spatial-temporal slow-fast graph convolutional network for skeleton - based action



- recognition[J]. IET Computer Vision, 2022, 16(3): 205–217.
- [13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 2017(30): 5998–6008.
- [14] ZHENG H, FU J, ZHA Z J, et al. Learning deep bilinear transformation for fine-grained image representation[J]. Advances in Neural Information Processing Systems, 2019, 2019(32): 4279–4288.
- [15] ZHANG P, LAN C, ZENG W, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: IEEE, 2020: 1112–1121.
- [16] 井 望, 李汪根, 沈公仆, 等. 轻量级多信息图卷积神经网络动作识别方法[J]. 计算机应用研究, 2022, 39(4): 1247–1252.
- [17] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 7132–7141.
- [18] SHAHROUDY A, LIU J, NG T T, et al. Ntu rgb+ d: a large scale dataset for 3d human activity analysis [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 1010–1019.
- [19] LIU J, SHAHROUDY A, PEREZ M, et al. Ntu rgb+ d 120: a large-scale benchmark for 3d human activity understanding [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10): 2684–2701.
- [20] TU Z, ZHANG J, LI H, et al. Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition[J]. arXiv: 2202.04075, 2022.
- [21] SONG Y F, ZHANG Z, SHAN C, et al. Constructing stronger and faster baselines for skeleton-based action recognition [J]. arXiv: 2106.15125, 2022.
- [22] CHEN H, LI M, JING L, et al. Lightweight long and short-range spatial-temporal graph convolutional network for skeleton-based action recognition [J]. IEEE Access, 2021, 9: 161374–161382.
- [23] JIANG Y, YANG X, LIU J, et al. A lightweight hierarchical model with frame-level joints adaptive graph convolution for skeleton-based action recognition[J]. Security and Communication Networks, 2021, 2021(11): 1–13.
- [24] ZHANG H, ZU K, LU J, et al. Epsanet: an efficient pyramid split attention block on convolutional neural network [J]. arXiv: 2105.14447, 2021.
- [25] LIU Z, ZHANG H, CHEN Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: IEEE, 2020: 143–152.
- [26] CHENG K, ZHANG Y, HE X, et al. Skeleton-based action recognition with shift graph convolutional network [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: IEEE, 2020: 183–192.
- [27] CHENG K, ZHANG Y, HE X, et al. Extremely lightweight skeleton-based action recognition with shiftgcn++ [J]. IEEE Transactions on Image Processing, 2021, 30: 7333–7348.
- [28] DUAN H, WANG J, CHEN K, et al. PYSKL: towards good practices for skeleton action recognition [J]. arXiv: 2205.09443, 2022.
- [29] 曾胜强, 李 琳. 基于姿态校正与姿态融合的2D/3D骨架动作识别方法[J]. 计算机应用研究, 2022, 39(3): 900–905.