计算机科学
COMPUTER SCIENCE

多流融合的轻量级图卷积行为识别算法

李华, 赵领娣, 陈雨杰, 杨杨, 杜新兆

引用本文

李华, 赵领娣, 陈雨杰, 杨杨, 杜新兆. 多流融合的轻量级图卷积行为识别算法[J]. 计算机科学, 2023, 50(11A): 220800147-6.

LI Hua, ZHAO Lingdi, CHEN Yujie, YANG Yang, DU Xinzhao. Lightweight Graph Convolution Action Recognition Algorithm Based on Multi-streamFusion [J]. Computer Science, 2023, 50(11A): 220800147-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220800147

多流融合的轻量级图卷积行为识别算法

李 华 赵领娣 陈雨杰 杨 杨 杜新兆
长春理工大学计算机科学技术学院 长春 130022
(lihua@cust.edu.cn)

摘 要 传统的基于RGB视频的行为识别容易受到光线强度、观察视角等问题的影响。基于骨骼的行为识别受这些问题的影响较小,成为现在的主流方法之一。但目前基于骨骼信息的行为识别方法参数量较大,运算速度较慢。为了解决这些问题,提出一种多流融合的轻量级图卷积行为识别框架。首先,将融合人体关节、骨骼边、关节速度和骨骼速度的多种信息的数据输入到空间图卷积模块中;其次,在空间图卷积模块中加入了空间注意力机制来更好地提取各个关节之间的关系;最后,在时间卷积模块中使用了深度卷积和逐点卷积减少参数量。提出的网络与基线网络SGN相比,在NTU-RGB+D120数据集中,交叉视角评估下提高了2.3%,交叉设置评估下提高了1.9%,参数量减少了0.12×10^6个,从而验证了提出网络的有效性。

关键词:人体骨骼;行为识别;轻量级;注意力机制;图卷积
中图法分类号 TP391

Lightweight Graph Convolution Action Recognition Algorithm Based on Multi-stream Fusion

LI Hua,ZHAO Lingdi,CHEN Yujie,YANG Yang and DU Xinzhao
College of Computer Science and Technology,Changchun University of Science and Technology,Changchun Jilin 130022,China

Abstract Traditional action recognition based on RGB-based methods is easy to be affected by problems such as light intensity and viewing angle. Skeleton-based action recognition is less affected by these problems and has become one of the mainstream methods. However,the current skeleton-based action recognition methods have a large number of parameters and slow operation speed. In order to solve these problems,a multi-stream fusion lightweight graph convolution action recognition framework is proposed. Firstly,the data fused with various information of joint,bone,joint motion and bone motion are input into the spatial map convolution module. Secondly,the spatial attention mechanism is added to the spatial graph convolution module to better extract the relationship between the joints. Finally,in the time convolution module,depthwise convolution and pointwise convolution are used to reduce the amount of parameters. Compared with the baseline network SGN,in NTU-RGB+D120 dataset,the proposed network increases by 2.3% under cross-subject evaluation,increases by 1.9% under cross-setup evaluation,and the number of parameters reduces by 0.12×10^6. The validity of the proposed network is verified.

Keywords Human skeleton,Action recognition,Lightweight,Attention mechanism,Graph convolution

1 引言

随着计算机视觉的发展,人体行为识别[1]作为其重要领域之一,具有重要的研究意义和广泛的研究背景。其应用范围主要包括视频监控、智能家居和人机交互等。根据输入数据形式的不同,行为识别方法被分为基于RGB的行为识别方法和基于骨架的行为识别方法。基于RGB的行为识别方法以视频帧为输入数据,通过对视频帧的处理实现对行为的识别。这种方法容易受到光线亮度、观察视角、身体遮挡等因素的干扰,所以现在基于骨骼数据的人体行为识别方法受到了广大研究者的青睐。基于骨骼数据的行为识别方法以骨骼序列数据为输入,骨骼序列由多帧骨骼结构数据组成,对于每个人,每帧骨骼结构数据包含固定个数的骨骼点,以此作为人体的高层语义表示。

早期的行为识别方法主要通过手工设计特征[2]对视频时空判别性特征进行建模,这些手工设计特征包括局部和全局的特征。Su等[3]设计了一个根据动作级别选择不同的特征,进而选择特定的分类器的框架,使用了支持向量机和隐马尔可夫模型;Li等[4]使用姿态估计技术提取出骨骼点,使用基于帧窗口矩阵的特征描述方法来进行多人行为识别。虽然这些手工设计的特征取得了不俗的效果,但为了获取到这些特征,需要消耗大量的人力,并且需要具备专业领域知识。此外,手工设计的特征在大型数据集上具有泛化能力较弱的缺点。近几年来,随着深度学习的发展,基于深度学习的行为识别方法得到了广泛的关注并取得了惊人的效果。Jiang等[5]提出基于多维特征嵌合注意力机制的图卷积识别方法,利用时空建模与通道之间的相关性提取出更为丰富的动作信息,得出了更加准确的分类结果;Lee等[6]使用GCN网络,使用一种新颖的

基金项目:国家自然科学基金(U19A2063);吉林省科技厅自然科学基金项目(20210101412JC)
This work was supported by the National Natural Science Foundation of China(U19A2063) and Natural Science Fund Project of Science and Technology Department of Jilin Province(20210101412JC)
通信作者:赵领娣(1025205283@qq.com)

层次分解图网络生成具有语义信息的邻接矩阵，提出的网络结构在 NTU-RGB+D 等数据集上都优于最先进的方法。

以上方法都取得了不错的效果，但是忽略了参数量以及计算速度等。本文提出了融合多种信息的轻量级行为识别算法，其网络结构如图 1 所示。本文主要贡献包括：1)在数据输入阶段，融合关节信息、关节运动信息、骨骼边信息和骨骼边运动信息，使输入的信息更加多样化，提高了模型的性能；2)在空间图卷积中加入空间注意力机制，来进行骨骼特征提取，增强了各个关节之间的权重区分度，以此来更好地提取各个关节之间的权重关系；3)在时间卷积使用深度卷积和逐点卷积，保证准确率的同时，大幅度减少运算参数，提高了模型计算速度。

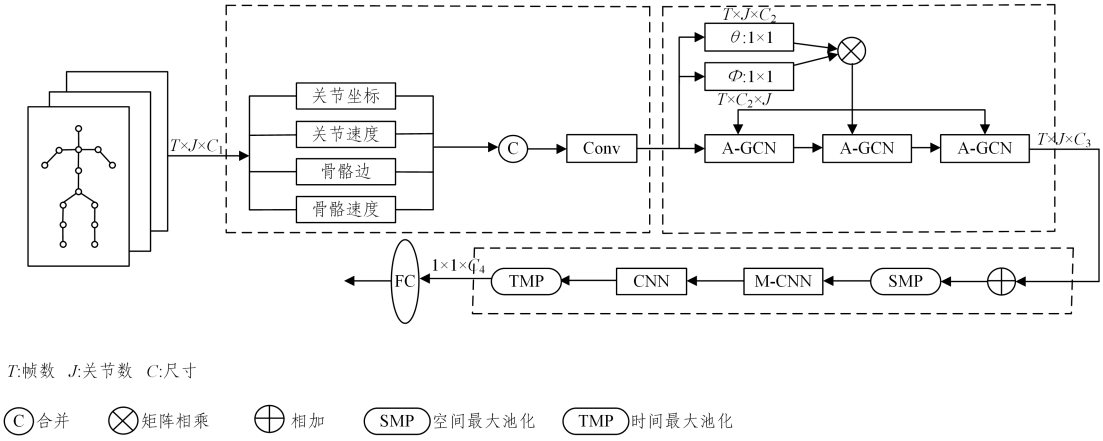

图 1 网络结构

Fig. 1 Network structure

2 相关工作

基于深度学习的方法主要包括三大类：基于卷积神经网络(Convolutional Neural Network，CNN)[7]、基于循环神经网络(Recurrent Neural Network，RNNs)[8]和基于图卷积网络(Graph Convolutional Network，GCN)[9]的方法。Li 等[10]提出了一种端到端的共现特征学习框架，使用了 CNN 来自动从骨架序列中学习分层的共现特征，显著提高了行为识别的性能。Du 等[11]用 LSTM 对人体骨骼的时空特征直接建模，提出了一种级联地组合人体骨骼各个部分运动的方法，关节揭示了人体的拓扑结构，是人体骨骼的重要信息。然而，基于 CNN 和 RNN 的方法却忽略了关节之间的相关性。Yan 等[12]首次将图卷积应用于人体行为识别，首先利用图来建模人体关节之间的关联，并应用图卷积和时间卷积来提取运动特征。之后，基于图卷积的研究越来越多。骨骼中有更多的细节和信息，例如骨骼长度、方向、角度等。为了融合骨骼的更多信息，Shi 等[13]提出了使用关节点信息和骨骼信息的双流网络，并且提出了自适应的图卷积，其可以自动学习出针对不同样本的不同拓扑结构，提高了行为识别的准确率。Qin 等[14]第一次提出将角度特征纳入时空图卷积，角度特征可以很容易地融合到现有的动作识别架构中，以进一步提高性能，角度特征是关节和骨骼的补充信息。Cheng 等[15]在图卷积的基础上用 Shift 卷积算子取代传统卷积算子而诞生出来的，可以用更少的参数量和计算量达到更好的模型性能。Liu 等[16]利用复杂的跨时空关节关系，提出了一种统一时空图卷积算子，该算子有助于跨时空的直接信息流，以实现有效的特征学习。但是基于图卷积的骨骼行为识别方法在健壮性和可伸缩性等方面受到了限制。Duan 等[17]提出了基于 3D-CNN 的 PoseConv3D 行为识别方法，在提取时空特征方面更加有效，对于姿态估计噪声更具有鲁棒性。PoseConv3D 由骨骼关节的热图堆栈表示，而不是由骨骼图上操作的坐标表示，在跨数据集环境中更具有通用性。由于其他网络参数量大，训练时间长，本文提出一种轻量化网络，引入了多种信息，空间模块和时间模块分别使用图卷积和卷积神经网络进行提取特征，解决了以往参数量过大的问题，并且达到了较高的识别精度。

3 多流融合的轻量级图卷积模型设计

3.1 多流融合

对于基于骨架的动作识别任务，关节坐标和骨骼边的方向和长度及其运动信息都值得研究。在这项工作中，我们对关节坐标、骨骼边坐标、关节运动和骨骼边运动进行建模。

首先，对于一个给定的骨架序列，其关节点的定义如式(1)所示：

$$s=\{v_{i,t}\mid i=1,2,\cdots,N;t=1,2,\cdots,T\} \tag{1}$$

其中，N 为关节总数，T 为序列中的总帧数，$v_{i,t}$ 表示第 t 帧的关节 i。本文定义距离骨架重心较近的关节为源关节，距离重心较远的关节为目标关节。每个骨骼边表示从其源关节指向其目标关节的向量，该向量不仅包含长度信息，还包含方向信息。给定第 t 帧中的骨骼边，其源关节的定义为 $v_{i,t}=(x_{i,t},y_{i,t},z_{i,t})$，其目标关节的定义为 $v_{j,t}=(x_{j,t},y_{j,t},z_{j,t})$，骨骼边的向量计算为：

$$e_{i,t}=(x_{j,t}-x_{i,t},y_{j,t}-x_{i,t},z_{j,t}-z_{i,t}) \tag{2}$$

由于骨骼边数据图中没有循环，因此可以为每个骨骼边指定唯一的目标关节。根关节未指定给任何骨骼边，所以关节数比骨骼边数多一个。为了简化网络的设计，将一个值为 0 的空骨骼边分配给根关节。因此，骨骼边的图和网络与关节的图和网络设计相同，每个骨骼边都可以与唯一的关节绑定。

但是一些动作，例如“站起来”和“坐下来”，很难从空间信息中识别。本文提取了关节运动信息和骨骼边运动信息，以帮助识别。骨架数据表示为关节的坐标，关节的运动很容易计算为沿时间维度的坐标差。第 t 帧中的关节 $v_{j,t}=(x_{j,t},y_{j,t},z_{j,t})$，第 $t+1$ 帧中的相同关节 $v_{j,t+1}=(x_{j,t+1},y_{j,t+1},z_{j,t+1})$，$v_{j,t}$ 和 $v_{j,t+1}$ 之间的运动信息表示：

$$m_{i,t}=(x_{j,t+1}-x_{j,t},y_{j,t+1}-y_{j,t},z_{j,t+1}-z_{j,t}) \tag{3}$$

第 t 帧中的骨骼边为 $e_{i,j,t}$，第 $t+1$ 帧中的相同骨骼边为 $e_{i,j,t+1}$，则 $e_{i,j,t}$ 和 $e_{i,j,t+1}$ 之间的运动信息表示为：

$$n_{i,t}=e_{i,j,t+1}-e_{i,j,t} \tag{4}$$

本文使用了 4 个流：第一个流使用原始骨架坐标作为输入，称为“关节流”；第二个流使用空间坐标的差分作为输入，称为“骨骼边流”；第三和第四个流使用时间维度上的差分作为输入，分别称为“关节运动流”和“骨骼边运动流”。

以关节编码为例，使用两个全连接层（FC）对关节 $v_{i,t}$ 进行编码，如式（5）所示：

$$\tilde{v}_{i,t}=\sigma(W_2(\sigma(W_1 v_{i,t}+b_1))+b_2) \tag{5}$$

其中，$W_1\in\mathbb{R}^{c_1\times 3}$，$W_1\in\mathbb{R}^{c_1\times c_1}$ 是权重矩阵，b_1 和 b_2 是偏差向量，σ 表示 ReLU 激活函数。以同样的方式可以得到骨骼边编码 $\tilde{e}_{i,t}$、关节运动编码 $\tilde{m}_{i,t}$ 和骨骼边运动边编码 $\tilde{n}_{i,t}$。

得到编码之后，将关节编码、骨骼边编码、关节运动编码和骨骼边运动编码分别嵌入到相同的高维空间，即 $\tilde{v}_{i,t}$，$\tilde{e}_{i,t}$，$\tilde{m}_{i,t}$ 和 $\tilde{n}_{i,t}$，通过求和将它们融合在一起，如式（6）所示。融合之后输入二维卷积层变换为合适的通道，以便能够更好地输入空间图卷积模块。

$$z_{i,t}=\tilde{v}_{i,t}+\tilde{e}_{i,t}+\tilde{m}_{i,t}+\tilde{n}_{i,t}\in\mathbb{R}^{c_1} \tag{6}$$

其中，c_1 是关节的尺寸。

3.2　空间图卷积 A-GCN

本文在 SGN[18] 的基础上，重新设计了空间模块和时间模块。空间模块主要采用图卷积来探索关节数据之间的相关性，其网络结构如图 2 所示。

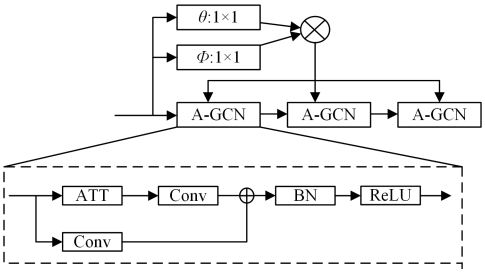

图 2　空间模块

Fig. 2　Space module

本文采用图卷积网络（GCN）来探索结构骨架数据的相关性。以前一些基于 GCN 的方法将关节作为节点，并基于先验知识预先定义图连接（边）[12] 或学习内容自适应图[14]；本文也学习了内容自适应图，但不同的是，结合了将关节类型连接到 GCN 层，以实现更有效的学习。本文将关节类型的语义合并到 GCN 层，以实现更有效的学习。文中从 4 个方面充分利用语义信息来增强 GCN 层的能力。此外，还使用关节类型、骨骼边类型和动力学的语义信息来学习不同关节之间的图连接，利用空间语义信息来增强 GCN 层的能力。关节信息不仅有助于学习合适的邻接矩阵，还参与了 GCN 层的消息传递。

通过不同关节的空间亲和力来构建边缘权重，如式（7）所示：

$$S_t(i,j)=\theta(z_{i,t})^{\mathrm{T}}\varphi(z_{j,t}) \tag{7}$$

其中，θ 和 φ 表示两个变换函数，变换函数由全连接层实现，即 $\theta(x)=W_3x+b_3$，$\varphi(x)=W_4x+b_4$。通过式（7）计算同一帧中所有关节对的亲和力，得到邻接矩阵。在邻接矩阵的每一行上使用 SoftMax 进行归一化，以便连接到目标节点的所有边缘值的总和为 1。采用残差图卷积层实现消息在节点间的传递，如式（8）所示：

$$\begin{cases}Y_t=G_tZ_tW_y\\ Z_t'=Y_t+Z_tW_z\end{cases} \tag{8}$$

其中，G_t 表示归一化邻接矩阵，W_y 和 W_z 是变换矩阵。不同时间帧的权重矩阵是共享的。Z_t' 是输出。图卷积层使用了残差图卷积和空间注意力机制（ATT），来更好地提取各个关节之间的权重关系，堆叠 3 个图卷积层可以实现相同邻接矩阵关节点之间进一步的消息传递。

3.3　时间卷积 M-CNN

本网络采用空间最大池化层（SMP）来合并帧中的跨关节信息。因此，序列特征的维数为 $T\times 1\times C_3$。首先借鉴 MobileNet[19-21] 的核心思想，设计一个时间卷积层 M-CNN，用于建模不同帧之间的相关性。时间卷积层的设计如图 3 所示。

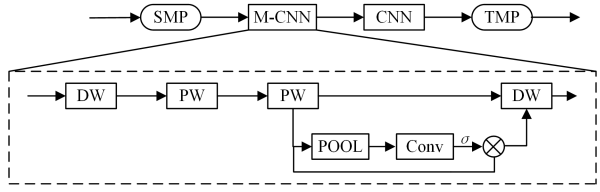

图 3　时间模块

Fig. 3　Time module

时间卷积层使用深度卷积（DW）和逐点卷积（PW）。对于深度卷积，卷积滤波器仅应用于一个对应的信道，而点卷积使用 1×1 卷积层来组合深度卷积的输出并调整输出信道的数量。使用深度卷积和逐点卷积，可以大大减少参数量且不会降低准确率。由于深度卷积和逐点卷积的感受野不大，为了融合更多尺度的信息，在最后一层的逐点卷积和深度卷积之间加了通道注意力机制 Efficient Channel Attention（ECA）[22]。ECA 避免了维度的减少，使用一维卷积高效实现了局部跨信道交互，提取通道之间的交互关系。激活函数使用了 h-swish 函数，h-swish 相较于 sigmoid 函数的优点在于它是一种非单调的函数，处处连续且可导，更容易训练。

经过时间卷积层之后，再经过一个 CNN 层，通过将其映射到卷积核大小为 1 的高维空间，来增强学习特征的表示能力。在两个卷积层之后，应用时间最大池化层（TMP）来聚合所有帧的信息，并获得 C_4 维的序列级特征表示。最后，用 Softmax 完全连接层来执行分类。

4　实验结果及分析

4.1　数据集

1）NTU-RGB+D60[23] 数据集包含来自 40 个不同主题的 56 000 多个视频样本，一共 400 万帧。其中包含 60 个不同的动作类，分为三大类：40 个日常动作，如饮酒、进食、阅读等；9 种与健康相关的动作，如打喷嚏、踉跄、摔倒等；11 种相互动作，如拳打脚踢、拥抱等。一共邀请了 40 名不同的受试者进行数据收集，受试者的年龄在 10 到 35 岁之间。该数据集有两种评估方式。（1）交叉主题评估（CS）：将 40 名受试者分为训练组和测试组。每组由 20 名受试者组成。评估中训练对象的 ID 为：1，2，4，5，8，9，13，14，15，16，17，18，19，25，27，28，31，34，35，38；剩下的受试者留作测试。对于该评估，训练集和测试集分别有 40 320 和 16 560 个样本。（2）交叉

视角评估(CV):训练集包括动作的前视图和两侧视图,而测试集包括动作性能的左视图和右 45 度视图。对于该评估,训练集和测试集分别有 37920 和 18960 个样本。

2) NTU-RGB+D120[24] 是对 NTU-RGB+D60 的补充。该数据集来自 106 个不同的主题,包含超过 11.4 万个视频样本,一共 800 万帧。该数据集包含 120 个不同的动作类,包括日常、相互和健康相关活动。这些受试者来自 15 个不同的国家。他们的年龄在 10 到 57 岁之间,身高在 1.3m 到 1.9m 之间。NTU-RGB+D120 采用了交叉视角评估(CS)和交叉设置评估(SS)。交叉视角评估和 NTU-RGB+D60 大致相似,交叉设置评估指的是选择所有具有偶数采集设置 ID 的样本进行训练,并选择具有奇数设置 ID 的样本进行测试,即 16 个设置用于训练,其他 16 个设置用于测试。

4.2 实验设置

本文设置模型框架中的通道维数 C_1, C_2, C_3, C_4 分别为 64,128,256 和 512,骨架序列选取帧数 T 为 20。本文使用的优化器是 Adam 优化器,初始学习率为 1×10^{-3},权重衰减为 1×10^{-4}。使用 MultiStepLR 学习率调整策略,学习率在第 60,90,110 个 epoch 分别衰减为原来的 1/10,训练在第 120 个 epoch 结束。NTU-RGB+D60 和 NTU-RGN+D120 数据集的批量大小分别设置为 64,64。所有实验均使用标签平滑分类交叉熵损失用于训练网络,其中将平滑因子设置为 0.1。实验硬件环境如表 1 所列。

表 1 实验设置

Table 1 Experimental setup

环境名称	具体配置
内存/GB	128
操作系统	Ubuntu18.0.4
GPU	NVIDIA GeForce RTX 3080
CPU	Intel © Xeon © Gold 6226R
Python/Pytorch	3.7/1.8.1
CUDA	11.1

4.3 实验结果及分析

为了验证多流融合信息的有效性,引入不同的信息进行消融实验,进一步对比多种信息的不同融合方式对模型的影响。表 2 列出不同信息的融合对模型性能的影响。其中,J 表示关节流,JV 表示关节运动流,B 表示骨骼边流,JV 表示骨骼边运动流,+表示将不同信息先在通道维度上合并后通过卷积变换为合适的通道输入到空间卷积模块中。将只有关节流的实验作为基础实验,然后在此基础上融合其余 3 种信息。实验结果表明,同时将关节流、关节运动流、骨骼边流和骨骼边运动流进行融合的方式准确率最高,同时参数量也不会增加,相较于单流输入的准确率提高了,以此可以证明多流融合输入的有效性。

表 2 多流融合实验

Table 2 Multi-stream fusion experiment

方法	参数量/($\times10^6$)	CS/%	CV/%
AM-GCN(J)	0.57	84.9	92.2
AM-GCN(J+B)	0.56	85.9	92.3
AM-GCN(J+JV)	0.56	87.5	93.5
AM-GCN(B+BV)	0.56	86.2	91.6
AM-GCN(J+JV+B)	0.57	88.5	94.0
AM-GCN(J+B+BV)	0.57	87.7	93.7
AM-GCN(J+JV+B+BV)	0.56	89.3	94.7

为了验证时间卷积模块的有效性,进行了对比实验,不同的深度卷积和逐点卷积的分配会产生不同的准确率。表 3 列出了不同的深度卷积和逐点卷积的组合。其中,DW 表示深度卷积,PW 表示逐点卷积。使用不同的 DW 和 PW 会有不同的参数量和准确率,实验证明,DW+PW+PW+DW 是最好的搭配组合,在保证参数量少的前提下提高了准确率。

表 3 时间模块实验

Table 3 Time module experiment

方法	参数量/($\times10^6$)	CS/%	CV/%
DW+PW	0.59	89.2	94.1
DW+PW+PW	0.56	89.0	94.3
DW+PW+DW	0.59	89.0	94.5
DW+PW+PW+DW	0.56	89.3	94.7
DW+PW+PW+PW+DW	0.57	89.0	94.0

为了验证算法的有效性,在 NTU-RGB+D60 和 NTU-RGB+D120 数据集上进行实验,实验如表 4 所列。同时为了验证该网络在较低参数量的情况下的具体表现,选择近两年内提出的方法作为参考比较的对象。实验结果表明,相较于基线网络 SGN[18],网络结构参数量减小并且在两个数据集上的识别准确率都有所上升,甚至在 NTU-RGB+D120 数据集上的 SS 评估设置下上升了 1.9%。相比较于近两年其他网络,对于 ST-GDN 网络,在 NTU-RGB+D60 数据集上准确率不如 ST-GDN,但在 NTU-RGB+D120 数据集上却比 NTU-GDN 高出 1%左右。对于 2s-AGCN 网络,网络参数量仅为 2s-AGCN 网络参数量的 1/10 左右,在 NTU-RGB+D60 数据集上的识别准确率也要比 2s-AGCN 网络的识别准确率要高出 0.8%。

表 4 NTU-RGB+D120 数据集上的识别准确率

Table 4 Recognition accuracy on NTU-RGB+D120 dataset

方法	年份	参数量/($\times10^6$)	NTU-RGB+D60		NTU-RGB+D120	
			CS/%	CV/%	CS/%	SS/%
ST-GCN[12]	2018	3.10	81.5	88.3	70.7	73.2
SR-TSL[25]	2018	19.07	84.8	92.4	74.1	79.9
RA-GCNv1[26]	2019	6.21	85.9	93.5	74.4	79.4
AS-GCN[27]	2019	9.5	86.8	94.2	77.9	78.5
2s-AGCN[13]	2019	6.94	88.5	95.1	**82.5**	**84.2**
SGN[18]	2020	0.69	89.0	94.5	79.5	81.5
RA-GCNv2[28]	2021	6.21	87.3	93.6	81.1	82.7
ST-GDN[29]	2021	—	**89.7**	**95.9**	80.8	82.3
Ours	—	0.57	89.3	94.7	81.8	83.4

综上所述,所提算法网络参数量小,训练速度更快,在 NTU-RGB+D 两个数据集上也能表现出不错的准确率。为了更直观地展现,还设计了如图 4 所示的散点图。散点图表示的是各个方法在 NTU-RGB+D60 数据集上 CS 评估设置下的准确率,其中 ST-GDN 没有对应的参数量,因此未在图中显示。

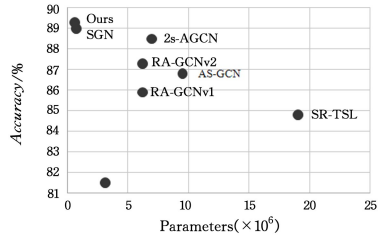

图 4 NTU-RGB+D60 数据集的识别准确率

Fig. 4 Recognition accuracy on NTU-RGB+D60 dataset

结束语　本文提出了一种多流融合的轻量级网络，将关节流、关节运动流、骨骼流和骨骼运动流4种信息进行融合作为网络输入，能够有效提高训练过程的准确率；在轻量化网络改进方面，在时间卷积模块借鉴了MobileNet[19-21]网络，采用了深度卷积和逐点卷积。实验证明，所提方法与SGN[18]相比，参数量减少了0.12×10^6，并且识别准确率有所上升，该研究对实时性要求较高的场景更加友好。但是正是由于参数量比较小，网络层数少，所以该方法无法提取出更深层的特征，相较于复杂度较高的网络结构识别准确率较低。在以后的工作中可以进一步研究如何在保证参数量的同时进一步提高准确率，例如在输入骨骼信息的基础之上融合RGB视频、光流等数据集，在时间语义信息上进行更深层的提取等。

参 考 文 献

[1] DENG M L，GAO Z D，LI L，et al. Overview of Human Behavior Recognition Based on Deep Learning[J]. Computer Engineering and Applications，2022，58(13)：14-26.

[2] CAI Q，DENG Y B，LI H S，et al. Survey on Human Action Recognition Based on Deep Learning[J]. Computer Science，2020，47(4)：85-93.

[3] SU B Y，WU H，SHENG M，et al. Accurate Hierarchical Human Actions Recognition From Kinect Skeleton Data[J]. IEEE Access，2019，7.

[4] LI M H，XU H J，SHI L X，et al. Multi-person Activity Recognition Based on Bone Keypoints Detection[J]. Computer Science，2021，48(4)：138-143.

[5] JIANG Q Y，WU X J，XU T Y. M2FA：multi-dimensional feature fusion attention mechanism for skeleton-based action recognition[J]. Journal of Image and Graphics，2022，27(8)：2391-2403.

[6] LEE J，LEE M，LEE D，et al. Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition[J]. arXiv：2208.10741，2022.

[7] DUAN H，ZHAO Y，XIONG Y，et al. Omni-sourced webly-supervised learning for video recognition[C]// European Conference on Computer Vision. Cham：Springer，2020：670-688.

[8] ATEFE A，ALI N，EBRAHIMI M M. Sparse Deep LSTMs with Convolutional Attention for Human Action Recognition[J]. SN Computer Science，2021，2(3).

[9] CHEN Y，ZHANG Z，YUAN C，et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：13359-13368.

[10] LI C，ZHONG Q，XIE D，et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[J]. arXiv：1804.06055，2018.

[11] 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. 2015：1110-1118.

[12] YAN S，XIONG Y，LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// Thirty-second AAAI Conference on Artificial Intelligence. 2018.

[13] 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. 2019：12026-12035.

[14] QIN Z Y，LIU Y，JI P，et al. Fusing Higher-Order Features in Graph Neural Networks for Skeleton-Based Action Recognition[J]. arXiv：2015.01563，2022.

[15] CHENG K，ZHANG Y，HE X，et al. Skeleton-based action recognition with shift graph convolutional network[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2020.

[16] 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. 2020：143-152.

[17] DUAN H，ZHAO Y，CHEN K，et al. Revisiting skeleton-based action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022：2969-2978.

[18] ZHANG P，LAN C，ZENG W，et al. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2020.

[19] HOWARD A G，ZHU M，CHEN B，et al. Mobilenets：Efficient convolutional neural networks for mobile vision applications[J]. arXiv：1704.04861，2017.

[20] SANDLER M，HOWARD A，ZHU M，et al. Mobilenetv2：Inverted residuals and linear bottlenecks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：4510-4520.

[21] HOWARD A，SANDLER M，CHU G，et al. Searching for MobileNetV3[J]. arXiv：1905.02244，2019.

[22] WANG Q，WU B，ZHU P，et al. ECA-Net：Efficient channel attention for deep convolutional neural networks[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2020.

[23] SHAHROUDY A，LIU J，NG T T，et al. NTU RGB+D：A Large Scale Dataset for 3D Human Activity Analysis[J]. arXiv：1604.02808，2016.

[24] LIU J，AMIR A，LISBOA P 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).

[25] CHEN Y S，YA J，WEI W，et al. Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network[J]. Pattern Recognition，2020，107.

[26] SONG Y F，ZHANG Z，WANG L. Richly Activated Graph Convolutional Network for Action Recognition with Incomplete

Skeletons[J]. arXiv:1905.06774,2019.

[27] LI M S,CHEN S H,CHEN X,et al. Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition [J]. arXiv:1904.12659,2019.

[28] SONG Y F,ZHANG Z,SHAN C,et al. Richly Activated Graph Convolutional Network for Robust Skeleton-Based Action Recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology,2021,31(5).

[29] PENG W,SHI J,ZHAO G. Spatial temporal graph deconvolutional network for skeleton-based human action recognition[J]. IEEE Signal Processing Letters,2021,28:244-248.

LI Hua, born in 1977, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include computer vision and virtual reality technology.

ZHAO Lingdi, born in 1997, master. Her main research interest is virtual reality.