基于三维图卷积与注意力增强的行为识别模型

曹 毅*①② 刘 晨①② 盛永健①② 黄子龙①② 邓小龙③

①(江南大学机械工程学院 无锡 214122)

②(江南大学江苏省食品制造装备重点实验室 无锡 214122)

③(江苏信息职业技术学院 无锡 214153)

摘 要：针对当前行为识别方法无法有效提取非欧式3维骨架序列的时空信息与缺乏针对特定关节关注的问题，该文提出了一种基于3维图卷积与注意力增强的行为识别模型。首先，介绍了3维卷积与图卷积的具体工作原理；其次，基于图卷积中可处理变长邻居节点的图卷积核，引入3维卷积的3维采样空间将2维图卷积核改进为具有3维采样空间的3维图卷积核，提出一种3维图卷积方法。针对3维采样空间内的邻居节点，通过3维图卷积核，实现了对骨架序列中时空信息的有效提取；然后，为增强对于特定关节的关注，聚焦重要的动作信息，设计了一种注意力增强结构；再者，结合3维图卷积方法与注意力增强结构，构建了基于3维图卷积与注意力增强的行为识别模型；最后，基于NTU-RGBD和MSR Action 3D骨架动作数据集开展了骨架行为识别的研究。研究结果进一步验证了基于3维图卷积与注意力增强的行为识别模型针对时空信息的有效提取能力及识别准确率。

关键词：行为识别；3维图卷积；注意力增强；时空信息

中图分类号：TN911.73; TP391.41 **文献标识码**：A **文章编号**：1009-5896(2021)07-2071-08

DOI: 10.11999/JEIT200448

Action Recognition Model Based on 3D Graph Convolution and Attention Enhanced

CAO Yi①② LIU Chen①② SHENG Yongjian①②

HUANG Zilong①② DENG Xiaolong③

①(*School of Mechanical Engineering, Jiangnan University, Wuxi* 214122, *China*)

②(*Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Jiangnan University, Wuxi* 214122, *China*)

③(*Jiangsu Information Vocational and Technical College, Wuxi* 214153, *China*)

Abstract: To solve the problems that current behavior recognition methods can not effectively extract the spatial-temporal information in non-European 3D skeleton sequence and lack attention for specific joints, an action recognition model based on 3D graph convolution and attention enhanced is proposed in this paper. Firstly, the specific working principles of the 3D convolution and graph convolution are introduced; Secondly, a 3D graph convolution method is proposed. It is based on the graph convolution kernel that can handle variable-length neighbor nodes in graph and 3D sampling space of 3D convolution is introduced to improve 2D graph convolution kernel to 3D graph convolution kernel with 3D sampling space. For neighbor nodes in 3D sampling space, this method realizes effective extraction of spatial-temporal information with a 3D graph convolution kernel; Thirdly, in order to enhance attention to specific joints and focus important action information, an attention enhanced structure is designed. Besides, through combining 3D graph convolution with attention enhanced structure, action recognition model based on 3D graph convolution and attention enhanced is proposed. Finally, the researches are carried on NTU-RGBD and MSR Action 3D skeleton action dataset.

收稿日期：2020-06-04；改回日期：2021-02-01；网络出版：2021-03-31

*通信作者： 曹毅 caoyi@jiangnan.edu.cn

基金项目：国家自然科学基金(51375209)，江苏省“六大人才高峰”计划项目(ZBZZ-012)，江苏省优秀科技创新团队基金(2019SK07)，高等学校学科创新引智计划(B18027)，江南大学研究生科研与实践创新计划项目(JNSJ19_005, JNKY19_048)

Foundation Items: The National Natural Science Foundation of China (51375209), The Six Talent Peaks Project in Jiangsu Province (ZBZZ-012), The Excellent Technology Innovation Team Fundation Jiangsu Province (2019SK07), The Research and the Innovation Project for College Graduates of Jiangan University (JNSJ19_005, JNKY19_048)

The results further verify the ability to extract spatial-temporal information of this model and its classification accuracy.

Key words: Action recognition; 3D graph convolution; Attention enhanced; Spatial-temporal information

1 引言

骨架行为识别是通过提取骨架序列中的动作特征，进而实现对人体行为的理解与描述的方法。骨架行为识别是机器视觉领域的热点研究方向之一，其可实现计算机准确识别目标对象的动作，进而分析视频中人体的动作，提高了计算机的动态感知能力，因此骨架行为识别技术广泛应用于视频监控[1]、视频理解[2,3]等领域。

针对骨架行为识别，国内外学者分别基于卷积神经网络和图卷积神经网络两类方法开展了大量的理论与实验研究。其中，基于卷积神经网络[3–6]，文献[3]提出将骨架序列的时空信息编码为彩色纹理图像，并使用卷积神经网络学习行为的判别特征；文献[4]将3维卷积引入骨架行为识别，通过3维卷积神经网络学习深度图序列的时空信息，并融合关节特征向量输入的SVM分类结果，实现行为识别；文献[5]将双流结构与3维卷积结合，提出了双流3维卷积网络，并将骨架信息映射到3D坐标空间进行时空信息的编码，实现了时空信息的提取。

基于图卷积神经网络，文献[7]融合图卷积神经网络与时间卷积网络，提出了一种时空图卷积模型，以提取骨架序列的空间信息与时间信息；文献[8]结合图卷积与长短时记忆网络(LSTM)，提出了一种图卷积LSTM网络，通过图卷积与LSTM网络，分别提取骨架序列中的空间信息与时间信息；为捕获关节间更丰富的依赖关系，文献[9]引入了一种编码器-解码器结构以捕获动作的潜在依赖关系，并通过图卷积与时间卷积分别学习空间与时间信息，实现了时空信息的提取。

基于上述，针对骨架行为识别国内外诸多学者尽管开展了大量研究并取得了一定的研究成果[3–9]，但不难发现：(1) 3维卷积无法直接针对具有非欧式空间数据的3维骨架序列进行时空信息的提取；(2) 图卷积仅能提取空间信息，时空信息依赖图卷积与LSTM(或时间卷积)分别进行提取，且未考虑空间与时间信息间的关联性；(3) 缺少对于特定关节的关注，无法聚焦重要的动作信息。

针对上述问题，本文提出了一种基于3维图卷积与注意力增强的行为识别模型。本文首先介绍了3维卷积与图卷积的具体工作原理；其次基于图卷积中可处理变长邻居节点的图卷积核，引入3维卷积的3维采样空间将2维图卷积核改进为具有3维采样空间的图卷积核，提出了一种3维图卷积方法；然后，为增强对于特定关节的关注，聚焦重要的动作信息，设计了一种注意力增强结构；再者，结合3维图卷积与注意力增强结构，构建了基于3维图卷积与注意力增强的行为识别模型；最后，基于NTU-RGBD和MSR Action 3D骨架动作数据集开展了骨架行为识别的研究。研究结果进一步验证了本文提出的行为识别模型的时空信息的有效提取能力及优秀的识别准确率。

2 3维卷积与图卷积

2.1 3维卷积

3维卷积的3维采样空间由多个连续帧中相同位置的采样区域构成，其包含时间与空间2个维度。通过3维卷积核将多个连续帧中采样区域的数据进行堆叠求和生成多维数据，从而实现了对3维采样空间的卷积操作[10,11]，如图1所示。设3维卷积核的卷积核尺寸为$[P_i, Q_i, R_i]$，则第i层网络中第j张特征图的(x, y, z)位置响应可表示为

$$v_{ij}^{xyz}=\sigma\left(b_{ij}+\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1}w_{ij}^{pqr}v_{(i-1)j}^{(x+p)(y+q)(z+r)}\right) \tag{1}$$

其中，P_i, Q_i为3维卷积核的两个空间维度尺寸，R_i为3维卷积核的时间维度尺寸，w_{ij}^{pqr}表示3维卷积核中的采样权重，b_{ij}表示偏置值；$\sigma(\cdot)$函数包含批量标准化与激活函数等操作。

3维采样通过将前一层输出中多个连续帧进行加权叠加，其不仅能采集空间信息，且能构建当前特征图与前一层输出中多个连续帧的连接，实现了多帧范围内时间信息的捕捉。因此，3维卷积不仅能同时实现空间与时间信息的采集，且能保留两者的关联性，故3维卷积可适用于连续动作视频帧序列等欧式空间内3维序列型数据的时空特征采集。

2.2 图卷积

图卷积是学习图结构数据的一种通用有效的方式。图卷积通过可处理变长邻居节点的图卷积核，将邻居节点的隐藏状态进行加权求和，以此来聚合邻居节点的信息，实现了图结构数据的卷积操作，提取了图上信息[12]。因此，图卷积能处理具有广义拓扑结构的图结构数据，故其广泛运用于骨架行为识别[2]和姿态估计[13]等领域。

设第l层网络的输出图中有m个节点，从第1个

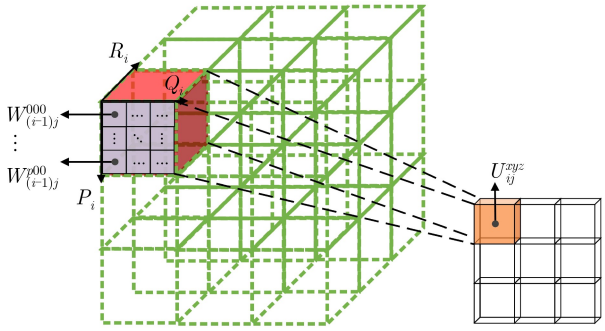


图 1 3维卷积的卷积操作

节点到第m个节点的n维隐藏状态表示为$\boldsymbol{h}^l{}_1$, $\boldsymbol{h}^l{}_2,\cdots,\boldsymbol{h}^l{}_m$，如图2所示。图节点状态记为$\boldsymbol{H}^l[\boldsymbol{h}^l{}_1, \boldsymbol{h}^l{}_2,\cdots,\boldsymbol{h}^l{}_m]\in\mathbb{R}^{m\times n}$，连接关系可通过邻接矩阵$\boldsymbol{A}\in\mathbb{R}^{m\times m}$表述，则$l$+1层输出的第1个节点响应为

$$\begin{aligned}\boldsymbol{h}_1^{l+1} &= \sigma(b+\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\boldsymbol{H}^l\boldsymbol{W})\\ &= \sigma(b+[a_{1,1}\boldsymbol{h}_1^l,\ a_{1,2}\boldsymbol{h}_2^l,\ a_{1,3}\boldsymbol{h}_3^l,\ \ldots,\ a_{1,m}\boldsymbol{h}_m^l]\boldsymbol{W})\end{aligned} \tag{2}$$

其中，$\boldsymbol{D}$表示$\boldsymbol{A}$的度矩阵，a为$\boldsymbol{A}$的元素用以判断节点是否为存在连接的邻居节点，$\boldsymbol{W}$表示图卷积的权重矩阵，b表示偏置值，$\sigma(\cdot)$表示非线性变化的激活函数。

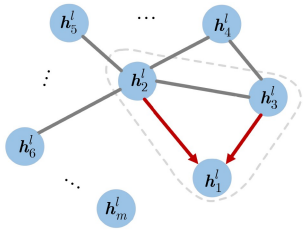


图 2 图卷积的卷积操作

3　基于3维图卷积与注意力增强的行为识别模型

3.1　3维图卷积

3.1.1　3维图卷积原理

骨架序列的空间结构特征与时间特征能够表述骨架序列中动作的完整信息，且两者之间存在关联不可独立分析。因此，为实现骨架序列中时空信息的有效提取，开展3维图卷积方法的研究是非常有必要的。

值得指出的是，3维卷积中3维采样空间为栅格化采样，其仅适用于欧式空间内3维序列型数据的特征采集，对于非欧式空间3维数据的采样存在采样空间中邻居节点数量不固定的问题。因此，(1) 3维卷积无法针对具有非欧式空间3维数据的骨架序列进行时空信息的提取；(2) 图卷积通过可处理变长邻居节点的图卷积核，其仅能实现图上空间信息的提取。为提取骨架序列的时空信息，基于图卷积中可处理变长邻居节点的图卷积核，以3维卷积中的3维采样空间为改进思想，将2维图卷积核改进为具有3维采样空间的图卷积核，本文提出了一种3维图卷积方法，其能有效提取非欧式空间内3维骨架序列的时空信息。

3维图卷积针对骨架序列的采样操作中，3维采样空间的邻居节点既包含当前帧内与节点存在连接的邻居节点也包含多个连续帧内相同位置节点的邻居节点。基于3维图卷积核，通过3维采样空间内邻居节点数据的加权堆叠求和来生成多维数据，从而实现了骨架序列的3维图卷积，有效提取了骨架序列的时空信息。如图3所示，设3维采样空间中有L张连续骨架帧，从第1帧到第L帧记作$\boldsymbol{G}^0,\boldsymbol{G}^1,\cdots,\boldsymbol{G}^{L\text{-}1}$，则3维图卷积的输出结果可表示为

$$\boldsymbol{x}' = \sigma\left(b+\sum_{t=0}^{L-1}\sum_{c=0}^{C-1}\sum_{k=0}^{K-1}\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}G_{c,\,k}^{\,t}\boldsymbol{W}_{c,\,k}^{\,t}\right) \tag{3}$$

其中，$\boldsymbol{A}$表示连接关系的邻接矩阵，$\boldsymbol{D}$表示$\boldsymbol{A}$的度矩阵，$G_{c,\,k}^{\,t}$表示3维采样空间中第t帧的第k个邻居节点的第c通道特征值，$\boldsymbol{W}_{c,\,k}^{\,t}$表示3维图卷积的权重矩阵，$b$表示偏置值；$\sigma(\cdot)$函数包含批量标准化与激活函数等操作。

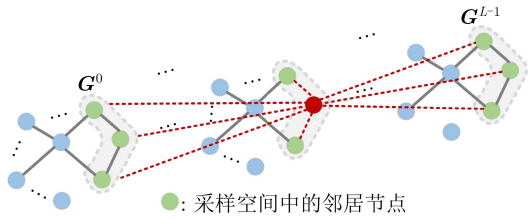


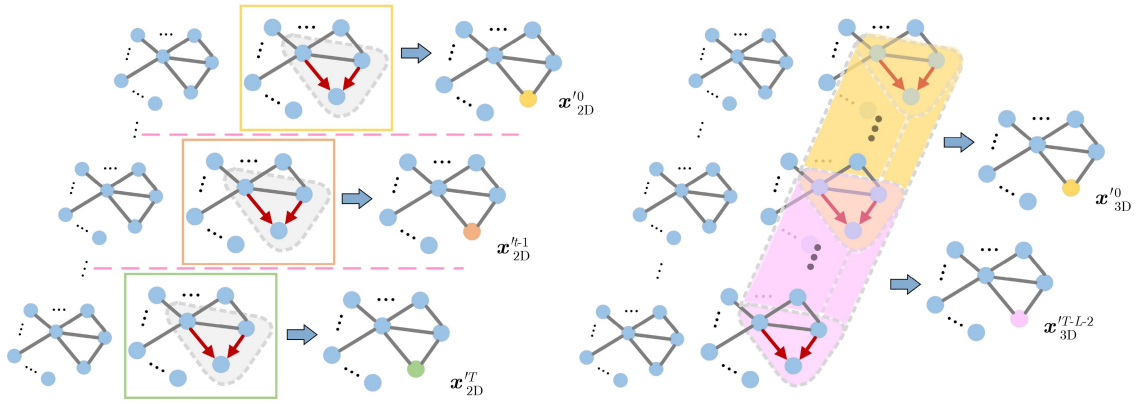
图 3 骨架序列中的3维图卷积

值得注意的是，3维图卷积在时间维度上采样骨架序列中的连续L帧，在未进行填充操作的情况下，每一次3维图卷积操作将使骨架序列减少L–1帧的序列长度。且基于padding填充操作，通过设置时间维度的采样步长，3维图卷积可实现倍率减少序列长度。

3.1.2　3维图卷积的有效性

骨架序列中3维采样空间是3维图卷积的核心，为证明3维图卷积的有效性，开展了3维图卷积与2维图卷积提取骨架序列特征的差异性研究。

如图4(a)所示，应用于骨架序列的2维图卷积仅输出对应当前第T帧的单帧图(式(4))，故每次图卷积运算仅处理当前第T帧内的空间信息，未对时间信息进行提取[7]，且2维图卷积切断了骨架帧间的时间关系，无法提取时空信息。



(a) 骨架序列的2维图卷积 (b) 骨架序列的3维图卷积

图 4 骨架序列中2维图卷积与3维图卷积的差异性

$$\begin{aligned}\boldsymbol{x}'^{T}_{2\mathrm{D}} &= \boldsymbol{\sigma}(b+\boldsymbol{x}^{T}_{2D})\\ &= \sigma\left(b+\sum_{c=0}^{C-1}\sum_{k=0}^{K-1}\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\boldsymbol{G}_{c,k}\boldsymbol{W}_{c,k}\right)\end{aligned} \tag{4}$$

对比2维图卷积，3维图卷积(图4(b))则通过采集当前第T帧的空间信息与第T帧周围L–1帧的时间信息(式(5))，保留了骨架帧间的时间关系，实现了时空信息的提取。通过将卷积层输出特征图与多个相邻帧相连，既提取了空间信息又捕获了时间信息[12]。其次，通过同时联合空间信息与时间信息进行提取，3维图卷积解决了2维图卷积与时间卷积网络融合带来的空间信息与时间信息关联性被破坏的问题，保留了两者的关联性。本文将进一步开展实验，以验证3维图卷积对比2维图卷积的有效性。

$$\begin{aligned}\boldsymbol{x}'^{T}_{3\mathrm{D}} &= \sigma\left(b+\sum_{t=0}^{L-1}\boldsymbol{x}^{t}_{2D}\right)\\ &= \sigma\left(b+\sum_{t=0}^{L-1}\sum_{c=0}^{C-1}\sum_{k=0}^{K-1}\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\boldsymbol{G}^{t}_{c,k}\boldsymbol{W}^{t}_{c,k}\right)\end{aligned} \tag{5}$$

由图4(a)、图4(b)与式(4)、式(5)的对比可知：(1) 2维图卷积仅对单张骨架帧的空间信息进行处理，由于其切断了骨架帧间的时间关系，故无法提取时间信息；(2) 基于具有时间与空间两个采样维度的3维采样空间，3维图卷积通过添加聚合时间维度上相关的邻居节点信息，既提取了骨架序列间的时间信息，又提取了空间信息，实现了时空信息的有效提取，且保留了空间与时间信息的相关性。

综上所述，针对具有非欧式空间3维数据的骨架序列，基于3维采样空间，3维图卷积通过聚合空间与时间维度上的邻居节点信息，实现了骨架序列中时空信息的有效提取。

3.2 注意力增强结构

骨架行为识别中动作的大部分动作信息可由少数关节表示，如挥手的大部分动作信息可由肩、肘、腕3个关节表示，故聚焦特定关节能一定程度提升骨架行为的识别准确率。

3维图卷积的输入特征中各个关节的权重均一致，针对特定动作其存在缺乏对于特定关节关注的问题。注意力机制通过注意力矩阵表示骨架序列中各关节对应的注意力权重并加权输入模型，实现了针对特定关节的关注[14]。基于上述，为解决3维图卷积缺乏对于特定关节关注的问题，本文设计了一种注意力增强结构。其不仅能增强对于特定关节的关注，且不削弱非关注关节点的信息，如图5所示。

注意力增强结构首先通过计算关节相似性度权重系数，求解生成中间特征，然后利用两层感知机，实现骨架序列中关节权重分布的提取，最后，结合结构输入特征，实现对于特定关节的注意力增强，注意力增强算法流程如下所示：

输入. 具有n维m个关节的骨架序列特征；

输出. 由输入骨架序列与关节加权的骨架序列求和生成的骨架序列；

步骤 1 基于相似度计算函数Score求解各关节间的相似度，并利用softmax函数进行相似度归一化，实现关节相似性权重系数α的生成；

步骤 2 基于权重系数α进行关节信息的加权求和并与原始特征拼接，实现中间特征$\boldsymbol{H}'$的生成；

步骤 3 通过两层感知机(s,u)结合tanh与sigmoid非线性化操作，实现关节权重矩阵$\boldsymbol{V}$的计算；

步骤 4 基于关节权重v_i针对骨架序列中关节进行加权，并通过求和结构输入特征$\boldsymbol{h}_i$得到结构输出$\hat{\boldsymbol{h}}_i$。

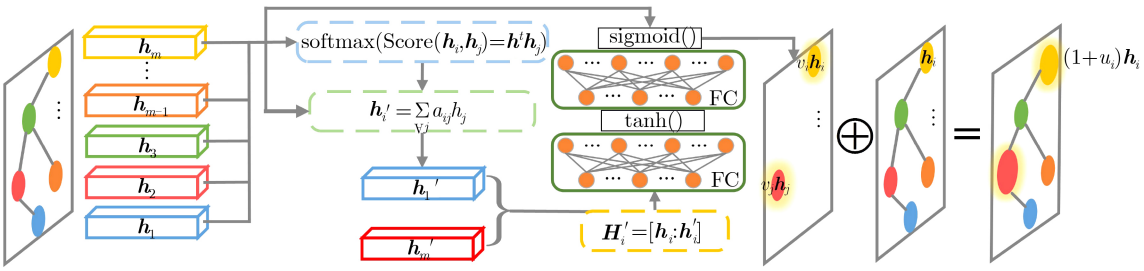


图 5 注意力增强结构示意图

$$\left.\begin{aligned}&\alpha_{ij}=\mathrm{softmax}(\mathrm{Score}(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{h}_j))\\&\boldsymbol{h}_i{}'=\sum_{\forall j}\alpha_{ij}\boldsymbol{h}_j,\boldsymbol{H}'_i=[\boldsymbol{h}_i:\boldsymbol{h}_i{}']\\&\boldsymbol{V}=\mathrm{sigmoid}(\boldsymbol{W}_u\tanh(\boldsymbol{W}_s\boldsymbol{H}'+b_s)+b_u)\\&\hat{\boldsymbol{h}}_i=(1+v_i)\boldsymbol{h}_i\end{aligned}\right\}\tag{6}$$

基于上述研究，注意力增强结构通过计算骨架序列中关节的权重并结合结构输入特征构建增强特征，既实现了对于特定关节点的注意力增强，又不削弱非关注关节点的信息，更有利于模型学习重要特征。

综上所述，基于3维图卷积与注意力增强的行为识别模型具有以下特点：(1) 3维图卷积将2维图卷积的2维采样区域扩展到3维采样空间，包含空间与时间2个维度，实现了骨架序列中时空信息的有效提取；(2) 基于注意力增强结构，增强了对于特定关节点的关注，更有利于模型学习重要特征。

4　实验设计与结果分析

4.1　实验数据集及评价指标

NTU-RGBD[15]：该数据集为最为广泛应用的行为识别大型实验数据集之一，其包含56880个动作样本。动作样本可划分为60个动作类别，每一个动作类别均通过3个视角的Kinect相机采集40个志愿者的25个关节点动作来构建。数据集具有基于视角(X-View)与基于运动对象(X-Sub)两种划分方式。为验证基于3维图卷积与注意力增强的行为识别模型的性能，评价指标采用Top-1识别准确率和Top-5识别准确率，针对模型在两种数据集划分方式下的性能进行综合评价。

MSR Action 3D[16]：该数据集包含16个动作类别的320个动作样本，每一个动作样本均有Kinect相机采集人体的20个关节点来构成。数据集可划分为3类子集(AS1, AS2, AS3)，各类子集均包含8类动作，其中AS1和AS2子集均为简单的相似动作，AS3子集为复杂动作[17]，评价标准采用Top-1识别准确率评价模型。

4.2　网络结构与参数配置

网络结构：单特征输入由于限制了网络模型从多种特征中学习各种信息，故影响了网络模型的识别准确率。为学习多种特征进一步提高网络模型的识别准确率，基于3维图卷积与注意力增强结构并以双特征作为输入，构建了基于3维图卷积与注意力增强的行为识别模型。双特征分别为表示静态特性的骨架特征与表示运动特性的骨架序列帧差特征。

该模型中的单流网络由3维图卷积与注意力增强结构构成，每一层3维图卷积前均设置注意力增强结构；利用3维图卷积的串联，构建了单流网络；通过叠加平均双流网络的预测分数，实现了双流融合并预测动作标签。若采用NTU数据集中25个关节表示的300帧骨架序列的动作样本，则模型具体结构如表1所示。

参数配置：动作帧数规整化(NTU：300帧；MSR：100帧)；设置批量处理尺寸(batch_size)为32；采用SGD为模型优化器；设置初始学习率为0.1；循环轮数(epoch)为80，并在第50轮进行学习率衰减。

4.3　模型对比实验

4.3.1　模型深度实验

模型深度在一定程度上影响模型的识别准确率，浅层模型识别准确率不高，深层模型存在过拟合。为探究最优的模型深度，以结合注意力增强结构的3维图卷积层数为变量，分别构建了5层至11层的网络结构，并基于以X-View划分的NTU数据集开展骨架行为识别的实验研究，实验结果如表2所示。

表 1　基于3维图卷积与注意力增强的行为识别模型的网络结构

结构层	输入	[注意力增强结构 3维图卷积]	…	[注意力增强结构 3维图卷积]	…	[注意力增强结构 3维图卷积]	…	Flatting	FC	Fusion
特征1	[3,300,25]	[16,300,25]	…	[32,150,25]	…	[64,75,25]	…	[120000]	[64]	[60]
特征2	[3,300,25]	[16,300,25]	…	[32,150,25]	…	[64,75,25]	…	[120000]	[64]	

表 2 不同模型深度的识别准确率对比(%)

模型深度	5层3DGCN	6层3DGCN	7层3DGCN	8层3DGCN	9层3DGCN	**10层3DGCN**	11层3DGCN
Top-1	92.18	92.59	92.76	92.93	93.04	**93.30**	93.01
Top-5	99.05	99.07	99.07	99.10	99.07	**99.49**	99.17

由表2可知：当模型层数为10层时，基于Top-1与Top-5评价指标，模型均取得最高识别准确率，分别为93.30%与99.49%，故模型最优模型深度为10层。

4.3.2 时间维度的邻居采样范围实验

3维图卷积通过扩展多个连续帧内相同位置节点的邻居节点，实现了时间信息的采样。时间维度上的邻居采样范围会影响模型时间信息的采样能力，长采样范围无法关注短时重要信息，短采样范围则无法提取上下文信息。为探究最优的采样范围，本文分别设置了5类采样范围并基于以X-View划分的NTU数据集开展骨架行为识别的实验研究，实验结果如表3所示。

由表3可知：当邻居采样范围为9帧时，基于Top-1与Top-5评价指标，模型取得最高的识别准确率，分别为93.30%与99.49%，故模型最优的采样范围选用9帧。

4.3.3 注意力对比实验

为验证注意力增强结构相较于其他注意力机制在3维图卷积模型上的优势，基于注意力增强结构与3种注意力机制分别开展了对比实验。实验均基于以X-View划分的NTU数据集开展，实验结果如表4所示。

表 3 不同邻居采样范围的识别准确率对比(%)

采样范围	3帧采样范围	5帧采样范围	7帧采样范围	**9帧采样范围**	11帧采样范围
Top-1	92.55	92.73	92.90	**93.30**	93.08
Top-5	99.11	99.40	99.00	**99.49**	99.10

表 4 注意力增强结构与多种注意力机制的识别准确率对比(%)

模型	3DGCN	3DGCN+Hard Attention	3DGCN+Soft Attention	3DGCN+Self Attention	**3DGCN+注意力增强结构**
Top-1	92.90	92.87	93.04	92.98	**93.30**
Top-5	**99.14**	99.02	99.04	99.12	99.09

由表4可知：(1) 相较于未使用注意力增强结构的3维图卷积模型，结合注意力增强结构的3维图卷积模型虽在Top-5评价指标下识别准确率下降了0.05%，但在Top-1评价指标下识别准确率提升了0.4%，达到最高93.30%的识别准确率；(2) 对比其他3种注意力机制，通过结合注意力增强结构，3维图卷积模型在Top-1评价指标下的识别准确率得到最显著提升，其也充分论证了注意力增强结构对于增强关节关注的有效性。

因此，注意力增强结构不仅能够增强3维图卷积针对特定关节的关注，且进一步提高了识别准确率。

4.4 识别准确率对比实验

为评估基于3维图卷积与注意力增强的行为识别模型的性能，基于NTU-RGBD和MSR Action 3D数据集，开展了骨架行为的识别准确率对比实验。

4.4.1 NTU-RGBD

为验证基于3维图卷积与注意力增强的行为识别模型相较于基于3维卷积与图卷积行为识别模型的优秀性能，分别基于X-View与X-Sub划分的NTU数据集开展了骨架行为识别的实验研究，并采用Top-1准确率为评价指标，实验结果如表5所示。

由表5可知：

(1) 在以X-View与X-Sub两种方式划分的NTU数据集上，基于3维图卷积与注意力增强的行为识别模型均取得最高的识别准确率，分别为93.30%与89.43%；

表 5 NTU数据集上不同模型的识别准确率对比(%)

模型	使用方法	X-View	X-Sub
文献[18]	3维卷积+双流	72.58	66.85
文献[6]	图卷积+TCN	88.30	81.50
文献[19]	图卷积	89.60	82.60
本文模型	**3维图卷积+注意力增强**	**93.30**	**89.43**

(2) 相较于同样采用双流结构的3维卷积方法，基于注意力增强的3维图卷积模型，在X-View与X-Sub下分别提高了20.72%与22.58%；

(3) 相较于使用2维图卷积的文献[6]，基于3维图卷积与注意力增强的模型识别准确率在X-View与X-Sub下分别提高了5.0%与7.93%，实验结果进一步验证了3维图卷积对比2维图卷积的有效性。

4.4.2 MSR Action 3D

上述单一数据集上的对比实验仅反映模型在单一数据集上的性能表现，为进一步验证模型在不同数据集上的性能，综合表现模型的泛化性能，需在全新数据集上开展模型性能的测试。为验证基于3维图卷积与注意力增强的行为识别模型的泛化性能，基于MSR Action 3D骨架行为识别数据集的3类子集(AS1, AS2, AS3)，开展了识别准确率对比实验。对比实验采用Top-1准确率为评价指标，实验结果如表6所示。

由表6可知：

(1) 基于注意力增强的3维图卷积模型，在AS1, AS2, AS3 3种训练条件下均取得了高于3维卷积与图卷积的识别准确率，进一步验证了模型时空信息提取的有效性；

表 6　MSR Action 3D数据集上3种训练条件下的识别准确率对比(%)

模型	使用方法	AS1	AS2	AS3	平均
文献[4]	3维卷积+SVM	92.03	88.59	95.54	92.05
文献[20]	3维卷积+SPMF	96.73	97.35	98.77	97.62
文献[21]	图卷积	93.70	95.80	96.60	95.20
本文模型	**3维图卷积+注意力增强**	**96.78**	**98.56**	**99.02**	**98.12**

(2) 本文提出的基于3维图卷积与注意力增强的行为识别模型，在NTU-RGBD与MSR Action 3D数据集上均取得了优秀的识别准确率，进一步验证了模型具有良好的泛化性能。

综上实验结果表明：基于3维图卷积与注意力增强的行为识别模型相较于基于3维卷积与图卷积的行为识别方法，既实现了骨架序列中时空信息的有效提取与对特定关节的注意力增强，又具有优秀的识别准确率与泛化性能。

5　结束语

为有效提取非欧式空间中3维骨架序列的时空信息，并实现针对特定关节的关注，本文提出了一种基于3维图卷积与注意力增强的行为识别模型。(1) 3维图卷积基于具有时间与空间两个采样维度的3维采样空间，聚合前帧内邻居节点的空间信息与时间维度上相关的邻居节点时间信息，实现了时空信息的有效提取，且保留了空间与时间信息的相关性。(2) 对比传统注意力机制，注意力增强结构不仅能增强对于特定关节的关注，且不削弱非关注关节点的信息。实验结果表明：首先，对比传统注意力机制，注意力增强结构能更有效增强对于特定关节的关注，且能进一步提高识别准确率。其次，基于3维图卷积与注意力增强的行为识别模型具有优秀的识别准确率与泛化性能。

值得指出的是，人体行为在未剪辑视频中仅占据小段时间，本文所提出的行为识别模型仅是基于人工剪辑视频，未剪辑视频中准确定位动作的问题在研究中未予以考虑。针对如何在未剪辑视频中准确定位动作与识别动作的问题，其在后续的研究中拟进一步展开。

参 考 文 献

[1] 周风余, 尹建芹, 杨阳, 等. 基于时序深度置信网络的在线人体动作识别[J]. 自动化学报, 2016, 42(7): 1030–1039. doi: 10.16383/j.aas.2016.c150629.
ZHOU Fengyu, YIN Jianqin, YANG Yang, *et al*. Online recognition of human actions based on temporal deep belief neural network[J]. *Acta Automatica Sinica*, 2016, 42(7): 1030–1039. doi: 10.16383/j.aas.2016.c150629.

[2] 刘天亮, 谯庆伟, 万俊伟, 等. 融合空间-时间双网络流和视觉注意的人体行为识别[J]. 电子与信息学报, 2018, 40(10): 2395–2401. doi: 10.11999/JEIT171116.
LIU Tianliang, QIAO Qingwei, WAN Junwei, *et al*. Human action recognition via spatio-temporal dual network flow and visual attention fusion[J]. *Journal of Electronics & Information Technology*, 2018, 40(10): 2395–2401. doi: 10.11999/JEIT171116.

[3] 吴培良, 杨霄, 毛秉毅, 等. 一种视角无关的时空关联深度视频行为识别方法[J]. 电子与信息学报, 2019, 41(4): 904–910. doi: 10.11999/JEIT180477.
WU Peiliang, YANG Xiao, MAO Bingyi, *et al*. A perspective-independent method for behavior recognition in depth video via temporal-spatial correlating[J]. *Journal of Electronics & Information Technology*, 2019, 41(4): 904–910. doi: 10.11999/JEIT180477.

[4] HOU Yonghong, LI Zhaoyang, WANG Pichao, *et al*. Skeleton optical spectra-based action recognition using convolutional neural networks[J]. *IEEE Transactions on*

Circuits and Systems for Video Technology, 2018, 28(3): 807–811. doi: 10.1109/TCSVT.2016.2628339.

[5] LIU Zhi, ZHANG Chenyang, and TIAN Yingli. 3D-based deep convolutional neural network for action recognition with depth sequences[J]. *Image and Vision Computing*, 2016, 55: 93–100. doi: 10.1016/j.imavis.2016.04.004.

[6] TU Juanhui, LIU Mengyuan, and LIU Hong. Skeleton-based human action recognition using spatial temporal 3D convolutional neural networks[C]. Proceedings of 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, USA, 2018: 1–6. doi: 10.1109/ICME.2018.8486566.

[7] YAN Sijie, XIONG Yuanjun, and LIN Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. Proceedings of the 32nd AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, USA, 2018: 7444–7452.

[8] SI Chenyang, CHEN Wentao, WANG Wei, *et al*. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019: 1227–1236. doi: 10.1109/CVPR.2019.00132.

[9] LI Maosen, CHEN Siheng, CHEN Xu, *et al*. Actional-structural graph convolutional networks for skeleton-based action recognition[C]. Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019: 3590–3598.

[10] KIM T S and REITER A. Interpretable 3D human action analysis with temporal convolutional networks[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, USA, 2017: 1623–1631. doi: 10.1109/CVPRW.2017.207.

[11] JI Shuiwang, XU Wei, YANG Ming, *et al*. 3D convolutional neural networks for human action recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(1): 221–231. doi: 10.1109/TPAMI.2012.59.

[12] 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. doi: 10.11897/SP.J.1016.2020.00755.

XU Bingbing, CEN Keting, HUANG Junjie, *et al*. A survey on graph convolutional neural network[J]. *Chinese Journal of Computers*, 2020, 43(5): 755–780. doi: 10.11897/SP.J.1016.2020.00755.

[13] CAI Yujun, GE Liuhao, LIU Jun, *et al*. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks[C]. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019: 2272–2281. doi: 10.1109/ICCV.2019.00236.

[14] CHO S, MAQBOOL M H, LIU Fei, *et al*. Self-attention network for skeleton-based human action recognition[C]. Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, USA, 2020: 624–633. doi: 10.1109/WACV45572.2020.9093639.

[15] SHAHROUDY A, LIU Jun, NG T T, *et al*. NTU RGB+D: A large scale dataset for 3D human activity analysis[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 1010–1019. doi: 10.1109/CVPR.2016.115.

[16] LI Wanqing, ZHANG Zhengyou, and LIU Zicheng. Action recognition based on a bag of 3D points[C]. Proceedings of 2010 Computer Vision and Pattern Recognition-Workshops, San Francisco, USA, 2010: 9–14. doi: 10.1109/CVPRW.2010.5543273.

[17] 冉宪宇, 刘凯, 李光, 等. 自适应骨骼中心的人体行为识别算法[J]. 中国图象图形学报, 2018, 23(4): 519–525. doi: 10.11834/jig.170420.

RAN Xianyu, LIU Kai, LI Guang, *et al*. Human action recognition algorithm based on adaptive skeleton center[J]. *Journal of Image and Graphics*, 2018, 23(4): 519–525. doi: 10.11834/jig.170420.

[18] LIU Hong, TU Juanhui, and LIU Mengyuan. Two-stream 3D convolutional neural network for skeleton-based action recognition[EB/OL]. https://arxiv.org/abs/1705.08106, 2017.

[19] GAO Xuesong, LI Keqiu, ZHANG Yu, *et al*. 3D skeleton-based video action recognition by graph convolution network[C]. Proceedings of 2019 IEEE International Conference on Smart Internet of Things (SmartIoT), Tianjin, China, 2019: 500–501. doi: 10.1109/SmartIoT.2019.00093.

[20] PHAM H H, KHOUDOUR L, CROUZIL A, *et al*. Skeletal movement to color map: A novel representation for 3D action recognition with inception residual networks[C]. Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 2018: 3483–3487. doi: 10.1109/ICIP.2018.8451404.

[21] BATTISTONE F and PETROSINO A. TGLSTM: A time based graph deep learning approach to gait recognition[J]. *Pattern Recognition Letters*, 2019, 126: 132–138. doi: 10.1016/j.patrec.2018.05.004.

曹 毅：男，1974年生，教授，博士，研究方向为机器人机构学、机器人控制系统、机器学习.

刘 晨：男，1995年生，硕士生，研究方向为图像处理、计算机视觉.

盛永健：男，1996年生，硕士生，研究方向为故障诊断、深度学习.

黄子龙：男，1996年生，硕士生，研究方向为深度学习、音频分类.

邓小龙：男，1972年生，教授，研究方向为信号处理、机器学习.

责任编辑：陈 倩