计算机科学
COMPUTER SCIENCE

基于多尺度特征图卷积网络的教学行为识别及分析

李佳楠, 李锐宜, 赵至夫, 宋娟, 韩嘉泷, 朱桐

引用本文

李佳楠, 李锐宜, 赵至夫, 宋娟, 韩嘉泷, 朱桐. 基于多尺度特征图卷积网络的教学行为识别及分析[J]. 计算机科学, 2024, 51(10): 135-143.

LI Jia'nan, LI Ruiyi, ZHAO Zhifu, SONG Juan, HAN Jialong, ZHU Tong. Recognition and Analysis of Teaching Behavior Based on Multi-scale GCN [J]. Computer Science, 2024, 51(10): 135-143.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于标签传播增强的多通道图卷积网络

Multi-channel Graph Convolutional Networks Enhanced by Label Propagation Algorithm

计算机科学, 2024, 51(8): 304-312. https://doi.org/10.11896/jsjkx.240100139

基于RoBERTa和加权图卷积网络的中文地质实体关系抽取

Chinese Geological Entity Relation Extraction Based on RoBERTa and Weighted Graph Convolutional Networks

计算机科学, 2024, 51(8): 297-303. https://doi.org/10.11896/jsjkx.230600231

基于改进双流视觉Transformer的行为识别模型

Action Recognition Model Based on Improved Two Stream Vision Transformer

计算机科学, 2024, 51(7): 229-235. https://doi.org/10.11896/jsjkx.230500054

基于相似网络融合算法的癌症亚型预测

Cancer Subtype Prediction Based on Similar Network Fusion Algorithm

计算机科学, 2024, 51(6A): 230500006-7. https://doi.org/10.11896/jsjkx.230500006

基于语义扩充和HDGCN的虚假新闻联合检测技术

Unified Fake News Detection Based on Semantic Expansion and HDGCN

计算机科学, 2024, 51(4): 299-306. https://doi.org/10.11896/jsjkx.230700170

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240400089

基于多尺度特征图卷积网络的教学行为识别及分析

李佳楠[1]　李锐宜[1]　赵至夫[2]　宋　娟[1]　韩嘉泷[1]　朱　桐[3]
1 西安电子科技大学计算机科学与技术学院　西安 710071
2 西安电子科技大学人工智能学院　西安 710071
3 西安电子科技大学前沿交叉研究院　西安 710071
(lijianan@xidian.edu.cn)

摘　要　在教育领域,课堂教学评价是提高教学质量的关键环节之一。随着数字化教育的推广,寻求一种智能化的评价方法变得尤为重要。为此,提出了一种基于骨架行为识别和滞后序列分析的新型方法,旨在更准确地对教师的教学行为进行捕获和分析,在减少人力资源消耗的同时,降低教学评价的主观性。首先,提出多尺度特征图卷积网络,并将其用于教师课堂行为分析。该网络在空间维度上使用多尺度语义特征融合模块捕捉骨架点和肢体部位两个尺度的特征;在时间维度上使用多尺度时序特征提取模块,并分别从全局和局部两个角度提取骨架数据的时间特征。然后,构建了教师课堂行为分析数据集,并在该数据集上验证了所提方法的有效性。最后,利用所提的骨架行为识别模型和滞后序列分析法,搭建了一套教学行为识别与分析系统。在进行不同课堂教学行为识别时,所提方法在教室行为识别与分析方面具有显著的优势。

关键词:教师行为分析;骨架序列;数字化教育;图卷积网络;行为识别
中图分类号　TP391

Recognition and Analysis of Teaching Behavior Based on Multi-scale GCN

LI Jia'nan[1],LI Ruiyi[1],ZHAO Zhifu[2],SONG Juan[1],HAN Jialong[1] and ZHU Tong[3]
1 School of Computer and Technology,Xidian University,Xi'an 710071,China
2 School of Artificial Intelligence,Xidian University,Xi'an 710071,China
3 Academy of Advanced Interdisciplinary Research,Xidian University,Xi'an 710071,China

Abstract　In the field of education,classroom teaching evaluation stands as a pivotal element in enhancing teaching quality. With the widespread adoption of digital education,the quest for an intelligent evaluation method becomes increasingly crucial. Therefore,this paper proposes a novel method based on skeleton action recognition and lagged sequence analysis,aiming to more accurately capture and analyze teachers' teaching behaviors while reducing manpower consumption and diminishing the subjectivity of teaching evaluations. Firstly,a multi-scale feature graph convolutional network is proposed and applied to analyze teacher classroom behaviors. This network utilizes a multi-scale semantic feature fusion module to capture features at two scales,skeleton points,and body parts,in the spatial dimension. In the temporal dimension,a multi-scale temporal feature extraction module is employed to extract temporal features of skeleton data from both global and local perspectives. Subsequently,a dataset for analyzing teachers' classroom behaviors is constructed,and the effectiveness of the proposed method is validated on this dataset. Finally,leveraging the proposed skeleton action recognition model and lagged sequence analysis,a system for recognizing and analyzing teaching behaviors is developed. The proposed method demonstrates significant advantages in classroom behavior recognition and analysis when applied to various classroom teaching scenarios.

Keywords　Teaching behavior analysis,Skeleton sequence,Digital education,Graph convolution,Action recognition

到稿日期:2024-04-15　返修日期:2024-07-05
基金项目:西安电子科技大学教育教学改革重点项目(A2304);中央高校基本科研业务费项目(ZYTS24092,QTZX24085);国家自然科学基金青年科学基金(62202356,62302373)
This work was supported by the Key Project of Education Teaching Reform of Xidian University(A2304),Fundamental Research Funds for the Central Universities(ZYTS24092,QTZX24085) and Young Scientists Fund of the National Natural Science Foundation of China(62202356,62302373).
通信作者:赵至夫(zfzhao@xidian.edu.cn)

1 引言

随着科技的不断进步，教育的数字化正成为推动教育领域发展的关键。在数字化教育的进程中，课堂教学评价显得尤为重要，教师课堂上的教学行为是教育最为重要的一环，授课质量直接影响学生的成绩。完整的教育管理需要教学评价来对教师的教学行为进行分析并及时给出建议。传统的教学评价往往是评估者通过线下听课进行打分和记录，或是观看课堂录像。这种方式存在一系列问题，如评分效率低、评分主观性强等，难以在全国范围推广应用。行为识别作为计算机视觉领域的重要研究任务，旨在从一段视频中识别理解主体人的行为，在视频监控、人机交互、医疗看护等领域有着广泛的应用。本文将行为识别引入智慧课堂，以更高效、更准确的方式识别教师的授课行为，减少人力资源的消耗。

根据输入数据的不同，行为识别可以分为基于 RGB 图像和基于骨架的行为识别。而在复杂的课堂场景下，基于骨架的行为识别能够剔除背景和学生的影响，专注于教师的课堂行为分析。当前，基于图卷积的骨架行为识别方法已经取得了显著的效果。但是依旧存在以下问题：1）在提取空间特征时，特征聚合只是在骨架点之间进行的。事实上，人体作为一个层次化的结构，可以看作是不同肢体部位的组成，某些动作也是肢体部位之间的协同合作完成的。2）当前基于图卷积的方法在时间维度建模时，通常采用简单的一维卷积操作，长时时序关系只能通过堆叠卷积操作获取，短时时序关系的提取容易受到卷积核大小的限制，无法同时有效地捕获骨架序列数据的长时特征和短时特征。

为了解决以上问题，本文提出多尺度特征图卷积网络来捕捉骨架数据复杂的时空信息。通过自建教师教学行为数据集验证所提方法对于教师授课行为识别的有效性。此外，搭建教学行为与分析系统，通过行为识别模型实时追踪教师的行为，统计教师授课时各个行为的占比和持续时间。为了更高效、更准确地识别和分析教师的授课行为，系统利用滞后序列分析方法进一步为教学决策提供支持，提高教育评估的科学性，更好地服务学生个性化学习和教育管理。这一创新性的方法有望弥补传统课堂评价的不足，提高评价的客观性和效率，从而进一步推动数字化教育的发展。本文的主要贡献如下：

1）提出了用于骨架行为识别的多尺度特征图卷积网络，提出多尺度语义特征融合模块和多尺度时序特征建模模块，在空间上和时间上提升模型对骨架数据特征的捕获能力；

2）构建了教师课堂行为分析数据集，引入人工智能的方法对教师教学行为进行识别，并在该数据集上验证了所提方法对教师教学行为分析的有效性；

3）搭建了教学行为与分析系统，采用滞后序列分析法（Lag Sequence Analysis）对课堂的教师行为序列进行评估和分析，深入理解教育过程中的因果关系和动态变化，从而为制定教育政策、改进教学方法、优化学生学习过程等提供科学依据。

2 相关工作

2.1 教师行为识别

传统的教学分析方法依赖于人工观察记录教师行为，然后围绕分析软件进行开发研究。

Li 等[1]根据 ITIAS 编码模板，研发了一款用于分析教师教学状态与风格的课堂教学视频分析软件。另一方面，Zhang[2]设计了一种基于达成度分析的实践课程教学过程管理系统，该系统在记录学习者完成教学任务的同时，阶段性地分析其学习能力达成度，并在课程学习结束后计算学习者各项能力达成度指标的最终值。

随着计算机视觉的发展，结合人工智能自动化地识别教师行为的方法，逐渐替代了传统人工观察记录的方式。Chen[3]提取出课堂教学师生运动历史图的 HOG 特征作为人体行为动作的特征，将其送入基于查对表的快速神经网络-支持向量机联合识别分类器进行动作分类。Tan 等[4]利用深度学习中通用的目标检测框架 Faster-RCNN，结合基于 ZFNet 预训练网络模型的迁移学习方法，实现对学生在课堂中的行为特征的提取和分类。Zheng[5]利用 HRNet 深度学习网络获取教师的人体关键点信息，利用关键点信息进一步构建教学评价指标，最后利用模糊综合评价的方法完成对教师的教学行为综合评分。然而，这些方法在评估教师教学行为时，面临着课堂场景错综复杂、教师行为多变的问题，无法有效地完成教学行为识别与分析的任务。

2.2 骨架行为识别

由于背景、光照和视点变化的鲁棒性，基于骨架的行为识别成为计算机视觉中具有吸引力的研究方向。此外，深度相机和姿态估计算法的发展使得获取人体骨骼数据变得更加容易。因此，基于骨架的行为识别近年来一直受到学者们的关注。

早期方法[6-8]通过手工设计特征来研究人类骨骼序列的动态性。例如，文献[6]中使用关节位置随时间的协方差矩阵作为骨架序列的判别描述符。然而，手工设计的特征不够灵活，无法对动作的判别性特征进行建模。相比之下，深度学习模型可以端到端自动学习合适的特征。随着深度学习的发展，最近的方法[9-13]在基于骨架的行为识别方面取得了令人印象深刻的表现。这些方法可以概括为 3 种：基于 CNN 的方法、基于 RNN 的方法和基于 GCN 的方法。

CNN 和 RNN 用于处理网格数据。然而骨架数据是以图的形式存在的，为了利用 RNN 或 CNN 强大的表达能力，早期基于深度学习的方法[9-10,14]将骨架数据重新排列为二维网格数据，然后直接馈送给 RNN 和 CNN。如文献[15]中所述，这些方法不能充分利用骨架数据的结构信息。最近，基于 GCN 的方法[11-13,16-18]以更灵活的方式处理骨骼数据，探索关节之间的关系。Yan 等[11]首次提出将 GCN 用于骨架行为识别。ST-GCN[11]使用空间图卷积和时间卷积来表示运动。为了进一步探索行为识别任务中骨架数据的多样性，2s-AGCN[12]引入了自注意的自适应图和自学习的图邻矩阵掩码。类似地，AS-GCN[16]也使用邻接矩阵进行多尺度建模。而 Zhang 等[17]从骨骼数据中研究人体骨骼，并在其中与骨骼

相对应的图边上进行卷积操作。InfoGCN[19]则提出了一种新的学习框架，将基于信息瓶颈的学习目标与基于注意力的图卷积相结合。此外，Yang 等[20]提出了一种将 GCN 和 CNN 相结合的混合网络，同时利用结构信息对帧间关节之间的复杂依赖关系进行建模。然而，上述方法未能考虑如何探索骨骼序列的时间关系。

3 教师教学行为识别及分析

3.1 整体流程

为了更有效地对教师行为进行识别与分析，为教学评价提供更科学的依据，搭建了基于多尺度特征建模图卷积网络的教师行为识别及分析系统。利用行为识别模型和滞后序列分析法，减少传统教学评价的人力资源消耗，促进人工智能与教育领域的结合。所提系统能够对一段课堂视频中的教师行为进行识别与分析，整体流程如图 1 所示。对于要分析的课堂视频，首先，提取视频中的教师骨架序列数据，采用公开的 Openpose 算法，人体姿态模型选择“COCO”，每一帧会提取 18 个关键点的坐标；然后，将所提取到的骨架数据送入训练好的多尺度特征建模图卷积网络，以获得教师教学行为的识别结果；最后，采用滞后序列分析法对课堂的教师行为序列进行评估和分析，深入理解教育过程中的因果关系和动态变化。

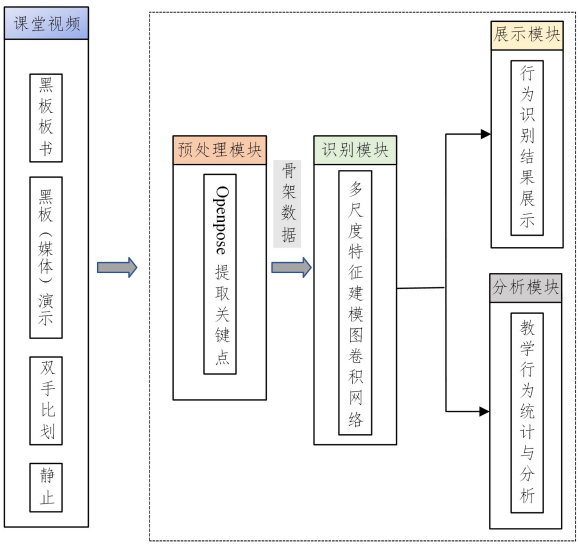

图 1 流程图

Fig. 1 Flow chart

选取黑板板书、黑板（多媒体）演示、双手比划、静止这 4 种出现频率高，对教学评价有意义，容易分辨的行为作为此次教师教学的行为类别。行为类别的具体描述如表 1 所列。

表 1 教师行为

Table 1 Teacher actions

编号	行为类别	行为表现	行为描述
A	黑板板书	用笔在黑板上写字	教师将需要强调的重要内容在黑板上记录下来
B	黑板（多媒体）演示	用手指黑板或者多媒体屏幕	教师对黑板或者多媒体屏幕的内容进行讲解
C	双手比划	用双手在空中比划进行讲解	教师站在讲台上用双手向学生进行教学内容讲解
D	静止	静止	教师在讲台上静止不动，等待学生反馈

相较于人工观察视频的方法，根据行为识别模型统计结果，能够准确、高效地获得一段视频中教师的行为序列，然后将其送入滞后序列分析软件 GSEQ 5.0，得到各行为占比与转换频次，以及调整后的残差值表，最后根据残差值表生成行为转换图。通过对比由各个教师课堂视频得到的数据，总结课堂行为特征，为教学评估提供科学的依据。

3.2 多尺度特征图卷积网络

本节提出了基于多尺度特征建模图卷积网络的骨架行为识别模型，以弥补当前骨架行为识别方法中时空信息提取的不足。具体来说，提出多尺度语义特征融合模块，将其嵌入网络的低层中，在获得骨架点特征图的基础上，重塑特征图，以获得更高级的语义信息。同时提出了多尺度时序特征建模模块，从全局和局部两个角度分别提取骨架序列数据的长时特征和短时特征，利用多尺度的方法进一步扩大时序感受野，提升其捕获长期信息的能力。

3.2.1 网络模型

多尺度特征建模图卷积网络的结构如图 2 所示，包含 7 个时空图卷积模块、3 个多尺度特征建模模块，以及一层全局平均池化层（Global Average Pooling，GAP）和一层全连接层（Fully Connected，FC）。时空图卷积模块由基准模型的图卷积模块（Graph Convolutional Networks，GCN）和时间卷积模块（Temporal Convolutional Networks，TCN）构成。分别利用 GCN 和 TCN 提取骨架数据的空间特征和时间特征。GCN 的计算如下：

$$\boldsymbol{H}^{(l+1)}=f(\boldsymbol{H}^{l},\boldsymbol{A})=\sigma(\boldsymbol{D}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}) \tag{1}$$

其中，$\boldsymbol{H}^{(l+1)}$ 为 l 层的输出特征；$\boldsymbol{H}^{(l)}$ 为第 l 层的输入特征；$\boldsymbol{W}^{(l)}$ 是第 l 层的参数矩阵；$\sigma(\cdot)$ 是神经网络的激活函数，如 Sigmoid，ReLU 函数；$\boldsymbol{A}$ 为邻接矩阵，$\mathbf{I}_N$ 为单位矩阵，$\boldsymbol{D}$ 为度矩阵，$\widetilde{\boldsymbol{A}}=\boldsymbol{A}+\mathbf{I}_N$。在每一层对节点及其邻居节点进行加权运算。多阶邻居节点的信息则是通过堆叠多层图卷积神经网络，扩大卷积核的感受野。

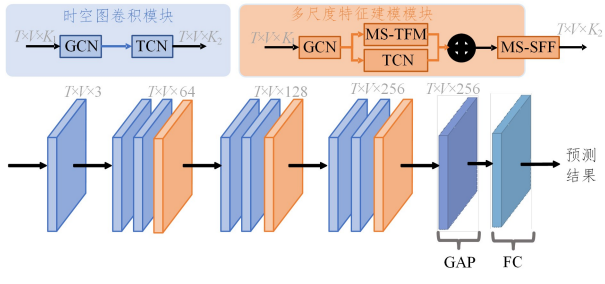

图 2 网络模型

Fig. 2 Network model

多尺度特征建模模块在时空图卷积模块的基础上，将多尺度时序特征建模模块（Multi-Scale Temporal Feature Modeling，MS-TFM）嵌入其中，与 TCN 平行地处理 GCN 的输出；同时，为了提取骨架数据的不同尺度的语义特征，将多尺度语义特征融合模块（Multi-Scale Semantic Feature Fusion，MS-SFF）嵌入在模块的最后。其中每个模块的输出通道数为 64，64，64，64，64，128，128，128，256，256，256。第 10 个时空图卷积模块的输出被送入全局平均池化层后，经过全连接层

得到所有动作类别最终的特征向量，最后通过 Softmax 层获得每个动作类别的预测结果。

3.2.2 多尺度语义特征融合模块

在行为识别任务中，某些动作是由各个肢体一同完成的，而图卷积提取空间特征的方法仅仅聚合了各个骨架点的特征。只有在深层的网络当中，才能获取到更高级别的语义特征。多尺度语义特征融合模块的提出，是为了在获得骨架点级别的特征的基础上，进一步获取其人体肢体部位的特征，然后将二者进行融合，同时保留关节点和肢体部位两个不同尺度的语义特征，进一步增强模型的特征表示能力，提高识别准确率。

如图 3 所示，假定输入的骨架数据特征图为 $\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{C\times T\times V}$，其中 C 表示通道数，T 表示帧数，V 表示节点数。根据肢体部位划分的规则将骨架点分为 N 个部分，N 为划分的肢体部位数，每个部位的特征为其所包含的骨架点的特征相加后的表示，此时的特征图 $\boldsymbol{X}\in\mathbb{R}^{C\times T\times N}$。由于每个样本只包含一个动作，因此不考虑每个肢体部位在时间上的变化，统一提取骨架序列中肢体部位的特征。首先进行平均池化操作，保留骨架数据特征图的特征通道维度 C 和空间维度 N，此时的特征图为 $\boldsymbol{X}\in\mathbb{R}^{(C\times N)\times 1\times 1}$。然后通过两个卷积层进行降维和升维的操作，第一次将通道数降至 $C\times N/4$，第二次将特征通道数升至 $C\times N$，其中 N 表示人体肢体部位的个数。接着将特征图转换为 $(C,N,1)$ 维的向量，使用 Softmax 函数对向量进行归一化，得到的新向量是 N 个肢体部位在通道数上的得分，表示每个肢体部位对此样本动作的重要性。由于之前存在池化操作，此时的肢体部位权重图不包含时间维度 T，需要进行扩展和复制。具体操作为：根据每个肢体部位所包含的骨架点，首先进行空间上的扩展，将每个肢体部位的得分根据其包含的骨架点一一对应，得到新的 $(C,1,V)$ 维向量，然后在空间上复制，得到最终的肢体部位权重图 $\boldsymbol{W}_p\in\mathbb{R}^{C\times T\times V}$。

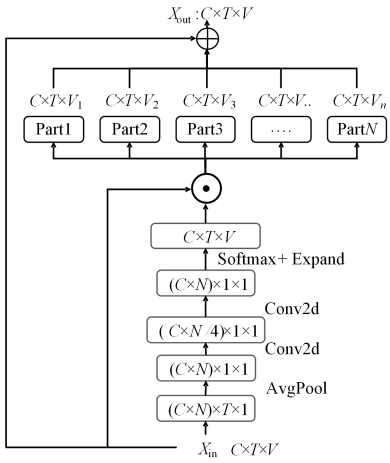

图 3 多尺度语义特征融合模块

Fig. 3 Multi scale semantic feature fusion module

在得到权重图之后，将对原始输入的骨架数据特征图进行重塑。重塑操作可以表示为：

$$f_{\text{out}}=\delta(f_{\text{in}}\odot\boldsymbol{W}_p+f_{\text{in}}) \tag{2}$$

其中，f_{in} 表示原始的输入特征，f_{out} 表示最终输出的融合特征，“$\odot$”表示逐元素相乘操作，“δ”表示激活函数 ReLu。经过重塑的特征图已具有了肢体部位这个级别的特征表示。同时，为了保留原来骨架点级别的特征，将原始的输入特征连接一同作为多尺度语义特征融合模块的输出。

3.2.3 多尺度时序特征建模模块

在骨架行为识别当中，现有的基于 GCN 的方法通常采用固定的一维卷积核在时间维度上建模，长时时序特征都是通过堆叠局部时序卷积间接获取。Liu 等[21]通过平行的不同膨胀率的卷积核扩大时序感受野，增强时序建模能力。在面对长期时序信息的提取时，Chen 等[22]提出多尺度时序图卷积模块。Liu 等[23]使用注意力机制来增强时序关系的特征表示，能够帮助一维卷积更有效地提取时序信息。

然而，这些方法无法有效提取复杂的时间动态信息。复杂的时间动态信息主要反映在长期时序信息和短期时序信息的联合建模上，这对于骨架行为识别是至关重要的任务。为了有效提取骨架数据复杂的时间特征，本文提出了多尺度时序特征建模模块，设计了一个适应全局信息的可变形时间卷积核(Deformable Temporal Kernel，DTK)。除此之外，为了进一步增强捕捉长期时序和短期时序的能力，设计了一个双重多尺度网络。在局部分支中，通过局部卷积获得对时间位置敏感的权重图，以此对短期时序建模。在全局分支中，通过可变形卷积核来聚合复杂的长期时序信息。相比简单的堆叠卷积操作，多尺度时序特征建模模块利用多尺度的方法扩大时序感受野，高效地提取骨架序列数据的时序特征，增强模型的性能。本节将会对可变形时间卷积核和双重多尺度策略进行详细的介绍。

1)可变形时间卷积核

可变形时间卷积核的提出是为了在面对不同的动作类别时，能够结合其全局信息，生成自适应的具有全局视野的卷积核，来提取骨架序列的长期时序特征。可变形时间卷积核的产生过程如图 4 所示。在骨架行为识别当中，给定输入的特征图，其中 C 表示通道数，T 表示帧数，V 表示节点数。因为可变形时间卷积核只关注时序特征，所以首先通过全局平均池化操作，将特征图在空间维度上压缩，可以表示为：

$$\hat{\boldsymbol{X}}=\phi(\boldsymbol{X})=\frac{1}{v}\sum_{i}^{v}\boldsymbol{X}_{c,t,i} \tag{3}$$

其中，输出的 $\hat{\boldsymbol{X}}\in\mathbb{R}^{C\times T}$ 表示时序特征图。

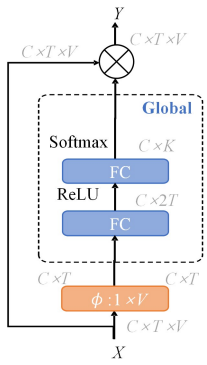

图 4 可变形时间核

Fig. 4 Deformable temporal kernel

全局平均池化操作并不会改变输入特征图的通道维度大小，学习到的可变形时间卷积核将以逐通道的方式对特征图进行卷积操作。如图 4 所示，在产生可变形卷积核时，不会将空间维度纳入其中。为了充分利用全局时序信息，使用全连接层来生成可变形卷积核。第一层全连接层将时间维度升维至 $2T$，并通过激活函数 ReLU 将其送入第二层全连接层。在第二层全连接层，将时间维度变化为指定的卷积核大小，最后通过 Softmax 函数对其进行归一化，生成最终的可变形时间卷积核。由此生成的带有全局感受野的卷积核，在聚合时序特征时能够关注到长期时序特征。具体来说，对于第 c 个通道，可变形时间卷积核可以表示为：

$$\boldsymbol{\theta}_c = G(x)_c = \mathrm{softmax}(F(\boldsymbol{W}_2, \delta(F(\boldsymbol{W}_1, \phi(x)_c)))) \tag{4}$$

其中，$\boldsymbol{\theta}_c \in \mathbb{R}^K$ 代表学习到的卷积核；K 代表卷积核的大小，是一个可设置超参数；δ 代表激活函数 ReLU；$F(\cdot)$ 代表全连接层；$\phi(x)_c$ 代表经过空间池化层的输出；$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 分别代表第一层和第二层全连接层的参数矩阵。

2）双重多尺度策略

如图 5 所示，双重指从全局和局部两个分支的视角来提取骨架数据序列的长期和短期时序特征；多尺度策略指在全局生成可变形时间卷积核时，通过设置不同大小的并行的卷积核扩大其时序感受野，进一步增强模型对长期时序捕捉的能力。

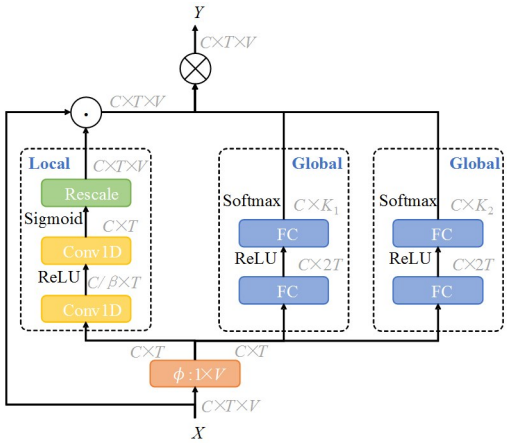

图 5 双重多尺度卷积

Fig. 5 Dual multi-scale convolution

（1）局部分支

在基于骨架的行为识别中，由于不同动作的时间、速度等存在差异，骨架序列中会存在一些关键帧，这些关键帧对于动作的识别具有判别性特征。局部分支的提出，正是为了增强模型捕获判别性特征的能力，生成对时间位置敏感的时序敏感度图，关注短期的局部细节。

如图 5 中 Local 分支所示，局部分支同样只关注特征图的时序维度特征，将经过空间全局平均池化操作后的特征图作为输入。首先经过第一层时间卷积层，将通道数从 C 降维至 C/β，经过 ReLU 激活函数后送入第二层时间卷积层。在第二层时间卷积层，将通道数从 C/β 升维至 C，然后经过 Sigmoid 函数产生时序敏感图 C。因为局部分支主要是为了捕获骨架序列的短期信息，所以将两层时间卷积层的卷积核大小均设置为 3，基于局部时间窗口生成能够提取判别性特征的时序敏感度图。此时产生的时序敏感度图仅包含时间维度，需要将其在空间维度上扩展，可以表示为：

$$\hat{\boldsymbol{V}} = F_{\mathrm{rescale}}(\boldsymbol{V}) \tag{5}$$

其中，$\hat{\boldsymbol{V}} \in \mathbb{R}^{C\times T\times V}$，$F_{\mathrm{rescale}}(\cdot)$ 表示在空间维度上复制。最终，局部分支的输出可以表示为：

$$\boldsymbol{Z} = L(\boldsymbol{X}) = \hat{\boldsymbol{V}} \odot \boldsymbol{X} \tag{6}$$

其中，“$\odot$”表示逐元素相乘，$\boldsymbol{X} \in \mathbb{R}^{C\times T\times V}$ 表示输入特征图，$\boldsymbol{Z} \in \mathbb{R}^{C\times T\times V}$ 表示局部分支的输出。从式(6)可以得出，局部分支并不会改变输入特征图的大小。

（2）多尺度全局分支

为了更有效地捕捉骨架序列的长期时序特征，在全局分支，使用可变形的时间卷积核对局部分支产生的特征图进行卷积操作。单尺度的聚合操作可以表示为：

$$\boldsymbol{Y} = \boldsymbol{G}_K(\boldsymbol{X}) \otimes \boldsymbol{L}(\boldsymbol{X}) \tag{7}$$

其中，$\boldsymbol{G}_K(\boldsymbol{X})$ 代表由全局分支产生的可变形时间卷积核，大小为 K；$\boldsymbol{L}(\boldsymbol{X})$ 代表局部分支中经过时序敏感度图重塑后的特征图；“$\otimes$”代表卷积操作。因为在产生可变形时间卷积核时利用了骨架数据的全局信息，而局部分支产生的时序敏感度图关注了骨架数据的判别性短期信息，所以经过特征聚合操作之后，能够达到同时捕捉长期时序和短期时序特征的目的。

进一步地，尽管生成的可变形时间卷积核具有了全局视野，但是由于卷积操作仍然是在局部窗口进行的，单一固定的卷积核无法突破此限制，对于跨度大的时间信息提取较为困难，长期时序信息只能通过简单的堆叠局部卷积获取。为了对动作的时序信息完整建模，既需要短期的细节，也需要长时间范围内的动态演化。在全局分支的基础上，使用了多尺度的方法，提取多个时间尺度内的特征。

具体来说，在对局部分支的输出特征图进行卷积操作时，生成不同尺度可变形时间卷积核进行特征聚合操作，然后将每个尺度的输出相加作为多尺度全局分支的输出，可以表示为：

$$\begin{aligned}\boldsymbol{Y} = {} & \boldsymbol{G}_{K_1}(\boldsymbol{X}) \otimes \boldsymbol{L}(\boldsymbol{X}) + \boldsymbol{G}_{K_2}(\boldsymbol{X}) \otimes \boldsymbol{L}(\boldsymbol{X}) + \cdots + \\ & \boldsymbol{G}_{K_n}(X) \otimes \boldsymbol{L}(\boldsymbol{X})\end{aligned} \tag{8}$$

其中，$K_1, K_2, \cdots, K_n$ 表示可变形时间卷积核的大小。

3.3 教学行为滞后序列分析法

滞后序列分析法是一种用于时间序列数据的统计分析方法，主要用于探索变量之间的因果关系和动态变化。它是在经典的时间序列分析方法的基础上发展而来，通常用于观察变量之间的滞后效应。

通过软件 GSEQ5.0 对输入的行为序列进行数据统计与分析，将序列按照格式要求输入，能够获得各个动作类别转换的频次表以及残差值表。频次转换表体现了某个行为之后发生另一个行为的频率，残差值表体现了两个行为之间转换的显著性。残差值(z-score)的计算如下：

$$z\text{-}score = \frac{\sigma - \bar{\sigma}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\sigma_i - \bar{\sigma})}} \tag{9}$$

其中，σ 表示特定数据点的值，$\bar{\sigma}$ 表示总体均值，N 表示样本大小，σ_i 表示每个数据点的值。通过将一个数据点的值与总体均值进行比较，并将其标准化为标准差的单位来计算残差值 z-score。依据滞后序列分析原理，当残差值大于 1.96 时，认为该行为序列存在统计学意义上的显著性。在教育领域，滞后序列分析法可以用于研究学生学习过程中的动态变化、学习成果等。本文通过教学行为识别模型能够获得课堂视频中教师的教学行为序列，使用滞后序列分析来评估不同教学行为序列的效果。在实际应用场景中，可以比较不同教师的课堂行为模式，例如将获得优秀教师称号的教师课堂行为序列与普通教师的课堂行为序列进行对比，通过观察不同教学方法实施后学生学习成绩的变化情况，可以确定哪种方法对学生学习效果更为有效。

4 实验结果与分析

4.1 实验数据集

4.1.1 NTU-RGB+D 60 数据集

NTU-RGB+D 60[24]是目前骨架行为识别任务中最广泛使用的大型数据集之一，包含 56 880 个骨架序列数据。该数据集涵盖了 60 种动作类别，由 40 名志愿者完成。每个样本仅包含一个动作，由两名志愿者完成，分别通过 3 台 Microsoft Kinect v2 相机在两个视角同时拍摄。作者提供了两个用来评估模型性能的基准。1）跨视角（Cross View，CV）：数据集根据相机的拍摄视角划分。选取相机视角 2 和视角 3 的 37 290 个样本来构建训练集，由视角 1 拍摄的 18 960 个样本用于构建测试集。2）跨主体（Cross Subject，CS）：数据集按照拍摄的志愿者划分。训练集和测试集都由 20 名志愿者构成，分别有 40 320 和 16 560 个样本。

4.1.2 教师教学行为数据集

目前暂未有公开的教师教学行为骨架数据集，因此采用自建数据集的方法。根据 3.1 节中所划分的教学行为，从网上下载视频进行分类和裁减，视频全部来自慕课网上的公开教学视频。在裁减视频时，只关注教师本身的行为，学生行为暂不考虑，因此保证视频内教师占主体部分并且只会出现教师一人，方便人体姿态算法进行骨架点数据的提取；同时，保证教师与黑板或多媒体屏幕的交互在视频中完整体现。在画面内容裁减完成后，接着对视频的长度进行裁减，每个视频片段的长度在 5～8 s 之间，每个片段只包含一种动作，保证教学行为动作清晰，能够被正确分辨，不存在不同类别的动作模糊相似。

视频裁减完成后，进行数据增强操作，具体为将每个视频的画面进行镜像翻转。视频片段构建完成之后，将每个视频片段分别放在对应动作类别的文件夹内，将每个文件下的视频分辨率改为 340×256，帧率为 30fps。选取 Openpose 人体姿态估计算法，开始进行关键点的提取，以构建骨架数据集。人体姿态模型选择“COCO”，每一帧会提取 18 个关键点的坐标。最终每个视频将对应一个骨架序列，然后生成对应的标签。根据数据集的划分原则，选取 80% 的骨架数据作为训练集，20% 的数据作为测试集，具体的统计情况如表 2 所列。

表 2 教师行为数据集统计

Table 2 Statistics of teacher action dataset

行为类别	样本数量
黑板板书	188
黑板（多媒体）演示	200
双手比划	144
静止	102

4.2 实验设置

所有的实验使用 PyTorch 框架完成，在单张 RTX 3080 GPU 上进行。使用 Stochastic Gradient Descent（SGD）优化算法。训练参数设置分别为：动量因子（momentum）为 0.9，训练轮数为 80，每次训练的批量大小为 16，初始学习率为 0.05，权重衰减（Weight Decay）率设置为 0.0004，学习率在第 10 轮和第 50 轮分别降低为初始学习率的 10%。骨架序列数据处理与 ST-GCN 采用相同的方式，通过扩充或者裁减，每个样本固定为 300 帧。

4.3 实验分析

4.3.1 NTU-RGB+D 60 消融实验

本节通过实验验证多尺度语义特征融合模块和多尺度时序特征建模模块的有效性，使用的数据集为 NTU-RGB+D 60 的跨主体（Cross Subject，CS），骨架数据的模态使用原始骨架序列，也就是关节流。选择 Top1 准确率作为评价标准。

1）多尺度语义特征融合模块

将人体骨架点划分为 5 个部分，分别是躯干、左臂、右臂、左腿、右腿，如图 6 所示。多尺度语义特征融合模块嵌入在第 4 时空卷积层之后，使模型在前期的学习过程中就能够提取骨架数据的不同尺度的语义特征。

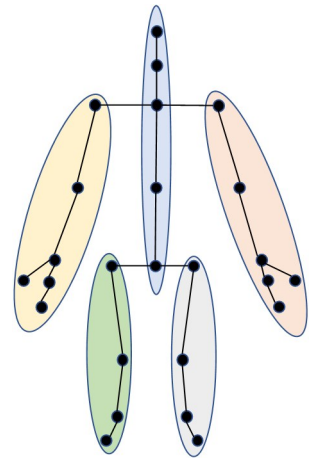

图 6 骨架点划分

Fig. 6 Skeleton point division

在 NTU-RGB+D 60 数据集上的实验结果如表 3 所列。从表中可以看出，在不同的基准网络上，嵌入了多尺度语义特征融合模块之后，模型的识别准确率均有了提升。

表 3 多尺度语义特征融合模块的准确率

Table 3 Accuracy of MS-SFF module

（%）

基准网络	未嵌入模块	嵌入模块
ST-GCN[25]	81.5	84.7
2s-AGCN[26]	85.8	87.6
CTR-GCN[22]	89.6	89.8

2）多尺度时序特征建模模块

为了增强模型提取长期时序特征的能力以面对不同类型的动作，3.2.3 节中提出了多尺度时序特征建模模块。为了验证其有效性，在 CTR-GCN 上进行消融实验，将多尺度时序特征建模模块嵌入到基准网络的第 4，7，10 层，其余的结构与原模型保持一致，实验结果如表 4 所列，括号里的数字代表选取的多尺度卷积核组合。当可变形时间卷积核为 3，5，7 时，相比基准网络，模型的识别准确率均有了提升，设置卷积核为 5 和 7 时，模型准确率达到了 90.2%，相比基准网络提升了 0.6%。毫无疑问，得益于局部分支和全局分支对时序特征的同时提取，可变形时间卷积核能够更有效地对骨架序列的时间维度建模。另外，从表中可以看出，使用多尺度的可变形时间卷积核比使用单尺度的可变形时间卷积核的效果更好。即使是多尺度策略下的最低性能 90.4%，仍然比单尺度最高的识别准确率高 0.2%。当设置卷积核大小为 3 和 7 时，识别准确率达到了最高的 90.6%，比基准网络提升了 1.0%。之所以多尺度可变形时间卷积核组合为 3 和 7 时模型的准确率最高，可能是因为基准网络 CTR-GCN 的 TCN 模块所使用的卷积核大小为 5，而多尺度可变形时间卷积核则完成了另外两个时间尺度下的时序特征提取，3 个尺度下的信息互补达到了最优性能。

表 4 可变形时间卷积核的准确率

Table 4 Accuracy of DTK

（%）

方法		准确率
基准网络	CTR-GCN[22]	89.6
单尺度可变形时间卷积核	+DTK(3)	90.1
	+DTK(5)	90.2
	+DTK(7)	90.2
多尺度可变形时间卷积核	+MS-DTK(3,5)	90.4
	+MS-DTK(3,7)	90.6
	+MS-DTK(5,7)	90.3
	+MS-DTK(3,5,7)	90.3

除此之外，在实验中还将基准网络换成了其他的骨架行为识别模型，选取了 ST-GCN 和 2s-AGCN，同样将多尺度时序特征建模模块嵌入到网络模型的第 4，7，10 层。3 个基准网络的模型结构基本一致，GCN 模块和 TCN 模块均保留了原网络的处理方法。值得注意的是，在 ST-GCN 和 2s-AGCN 中，多尺度可变形时间卷积核组合设置为 3，5，7，这是因为 ST-GCN 和 2s-AGCN 的时间卷积模块所使用的卷积核大小为 9。实验结果如表 5 所列，从表中可以看出，在加入多尺度时序特征建模模块之后，基准网络的识别准确率均有了提升，这意味着多尺度时序特征建模模块可作为即插即用的模块嵌入到其他骨架行为识别模型当中，从而提升其捕捉时序特征的能力。

表 5 不同基准网络实验的准确率

Table 5 Accuracy of different benchmark network experiments

（%）

基准网络	准确率	
	未嵌入 MS-DTK	嵌入 MS-DTK
ST-GCN[25]	81.5	84.7
2s-AGCN[26]	85.8	87.3

4.3.2 与先进方法对比

为了校验多尺度特征图卷积网络在 NTU-RGB+D 60 数据集上的性能，本节使用多流融合的策略，将多尺度特征图卷积网络与当前的先进方法进行对比。实验结果如表 6 所列。从表中可以看出，本文所提出的方法在 NTU-RGB+D 60 数据集上的表现基本达到了目前先进的水平。与 HD-GCN 相比，由于本文采用的是四流融合策略，准确率高于 2-ensemble 方法；相比 6-ensemble 方法，本文方法在较少参数量的情况下，可以达到与之接近的准确率。

表 6 在 NTU-RGB+D 60 上与先进方法的准确率对比

Table 6 Accuracy comparison with advanced methods on NTU-RGB+D 60

（%）

方法	年份	CS	CV
MST-GCN[27]	2021	91.6	96.6
CTR-GCN[22]	2021	92.4	96.8
EfficientGCN-B4[28]	2022	91.7	95.7
HD-GCN(2-ensemble)[29]	2023	92.4	96.6
HD-GCN(6-ensemble)[29]	2023	93.4	97.2
Ours	—	92.8	96.8

4.3.3 教学行为数据集上的实验结果

在自建教学行为数据集上的实验结果如表 7 所列，可以看出，本文提出的行为识别模型在小型的教学行为数据集上已经达到了极高的准确率。同时，表 8 列出了每个动作类别的识别准确率，可以看出在面对黑板板书、黑板（多媒体）演示、双手比划、静止时，行为识别模型能够在很大程度上替代人工观察的方法，对课堂视频中的教师教学行为进行识别与统计。

表 7 教学行为数据集上的实验结果

Table 7 Experiment result on teaching action dataset

（%）

方法	准确率
多尺度特征建模图卷积网络	99.2

表 8 教学行为识别的准确率

Table 8 Accuracy of teaching action recognition

（%）

动作类别	准确率
黑板板书	97.2
黑板（多媒体）演示	100.0
双手比划	100.0
静止	100.0

4.3.4 滞后序列分析法实验结果

根据 3.1 节所选定的教学行为，通过多尺度特征建模图卷积网络对课堂视频中的教师行为进行编码，固定时间间隔采样，能够获得一段教师教学行为序列。将其输入交互行为分析软件 GSEQ5.0，得到转换频次表（见表 9）和残差值表（见表 10）。根据残差值表，能够绘制图 7 所示的行为模式转换图。从图中可以看出，黑板板书是一个连续的行为，因为 A 行为之后大概率会继续 A 行为（$A\rightarrow A$），在黑板板书之后通常会接着对黑板进行演示（$A\rightarrow B$），双手比划动作通常会持续进行（$C\rightarrow C$）。

表 9 行为转换频次

Table 9 Action conversion frequency

	A	B	C	D	Total
A	7	5	1	0	13
B	1	1	4	1	7
C	2	1	9	3	15
D	3	0	2	2	7
Total	13	7	16	6	42

表 10 残差值

Table 10 Residual values

	A	B	C	D
A	2.15	2.54	−2.72	−1.77
B	−1.04	−0.19	1.14	0.00
C	−1.84	−1.30	2.18	0.79
D	0.75	−1.30	−0.57	1.18

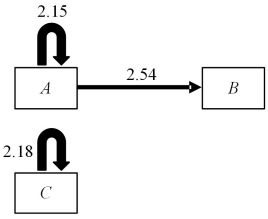

图 7 行为模式转换图

Fig. 7 Action pattern transition diagram

通过以上实际场景的教学行为统计与分析，不仅可以让教师更好地理解在不同教学情境中可能出现的教学行为问题，进而推动他们进行深入的教学反思，还能够清晰地展示各种教学行为之间的内在联系，有助于教师识别并判断可能遭遇的难题。

结束语 本文提出了基于多尺度特征建模图卷积网络的行为识别模型，在空间维度上提高了模型的多尺度语义特征提取能力，在时间维度上增强了模型的长时特征和短时特征提取能力。通过大量实验证明了所提出的多尺度语义融合模块和多尺度时序特征建模模块的有效性。同时自建教师教学行为数据集，引入人工智能的方法对课堂视频的教学行为进行统计，采用滞后序列分析法对统计结果进行进一步的分析，为课堂评价提供科学的依据，有利于促进数字化教育的发展。需要注意的是，自建教师教学行为数据集较小，针对的课堂场景有限，在未来的研究中理应从数据集的构建出发，进一步发挥行为识别模型的潜能。

参 考 文 献

[1] LI B H, ZHANG X Y. Design and Implementation of Classroom Teaching Video Analysis Software Based on ITIAS[J]. Software, 2019, 40(1): 46-50.

[2] ZHANG N L. Research on the Design of a Practical Course Teaching Process Management System Based on Achievement Analysis[J]. Education and Teaching Forum, 2021(41): 4.

[3] CHEN J T. Research on Intelligent Image Recognition and Analysis of Classroom Behavior [D]. Hangzhou: Zhejiang University, 2019.

[4] TAN B, YANG S. Research on Student Classroom Behavior Detection Algorithm Based on Faster R-CNN [J]. Modern Computer, 2018(22): 45-47.

[5] ZHENG Y. A Teacher Teaching Behavior Evaluation Method Based on Posture Recognition [J]. Software Engineering, 2021, 24(4): 6-9.

[6] HUSSEIN M E, TORKI M, GOWAYYED M A, et al. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations[C]// International Joint Conference on Artificial Intelligence(IJCAI). 2013: 2466-2472.

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

[8] VEERIAH V, ZHUANG N, QI G J. Differential recurrent neural networks for action recognition[C]// IEEE International Conference on Computer Vision(ICCV). 2015: 4041-4049.

[9] ZHANG P, LAN C, XING J, et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data[C]// IEEE International Conference on Computer Vision(ICCV). 2017: 2136-2145.

[10] LIU J, WANG G, DUAN L Y, et al. Skeleton-based human action recognition with global context-aware attention LSTM networks[J]. IEEE Transactions on Image Processing, 2017, 27(4): 1586-1599.

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

[12] 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.

[13] LI B, LI X, ZHANG Z, et al. Spatio-temporal graph routing for skeleton-based action recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 8561-8568.

[14] LI S, LI W, COOK C, et al. Independently recurrent neural network(indrnn): Building a longer and deeper rnn[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5457-5466.

[15] MONTI F, BOSCAINI D, MASCI J, et al. Geometric deep learning on graphs and manifolds using mixture model cnns[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5115-5124.

[16] LI M, CHEN S, CHEN X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3595-3603.

[17] ZHANG X, XU C, TIAN X, et al. Graph edge convolutional neural networks for skeleton-based action recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(8): 3047-3060.

[18] XU K, YE F, ZHONG Q, et al. Topology-aware convolutional neural network for efficient skeleton-based action recognition [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 2866-2874.

[19] CHI H,HA M H,CHI S,et al. Infogcn:Representation learning for human skeleton-based action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:20186-20196.

[20] YANG W,ZHANG J,CAI J,et al. HybridNet:Integrating GCN and CNN for skeleton-based action recognition[J]. Applied Intelligence,2023,53(1):574-585.

[21] 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.

[22] 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.

[23] LIU Y,ZHANG H,XU D,et al. Graph transformer network with temporal kernel attention for skeleton-based action recognition[J]. Knowledge-Based Systems,2022,240:108146.

[24] 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. 2016:1010-1019.

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

[26] 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.

[27] CHEN Z,LI S,YANG B,et al. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition[C]// Proceedings of the AAAI Conference on Artiffcial Intelligence. 2021:1113-1122.

[28] SONG Y F,ZHANG Z,SHAN C,et al. Constructing stronger and faster baselines for skeleton-based action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,45(2):1474-1488.

[29] LEE J,LEE M,LEE D,et al. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023:10444-10453.

LI Jia'nan, born in 1991, Ph. D, is a member of CCF (No. K8171M). Her main research interests include video understanding and action recognition.

ZHAO Zhifu, born in 1990, Ph. D, is a member of CCF (No. A0143M). His main research interests include deep learning, video understanding, and compressive sensing.

(责任编辑:柯颖)