

DOI:10.16136/j.joel.2022.02.0384

基于时空图卷积网络的学生在线课堂行为识别

胡锦涛林, 齐永锋*, 王佳颖

(西北师范大学 计算机科学与工程学院, 甘肃 兰州 730070)

摘要: 为了有效地识别学生在线课堂行为, 提出了一种融合全局注意力机制和时空图卷积网络的人体骨架行为识别模型。首先在时空图卷积网络的空间图卷积网络和时间卷积网络之间加入全局注意力模块, 空间图卷积网络输出的空间特征图作为注意力模块的输入。其次引入按时间维度的平均池化和最大池化操作, 以增加模型学习全局特征信息的能力。最后用三个加入注意力机制的时空图卷积神经网络和类激活图(class activation map, CAM), 构造对遮挡数据识别能力更强的丰富激活图卷积网络(RA-GCNv2-A)模型, 并通过迁移学习实现学生在线课堂行为识别功能。在 NTU-RGB+D 和 NTU-RGB+D120 数据集上进行实验验证, 与 RA-GCNv2 模型相比, 在 NTU-RGB+D 和 NTU-RGB+D120 数据集上的识别准确率分别提高了(cross-subject, CS)1.3%、(cross-view, CV)1.2%和(cross-subject, CSub)1.6%、(cross-setup, CSet)1.4%。实验结果表明, 提出的方法是一种有效的学生在线课堂行为识别方法。

关键词: 人体骨架; 行为识别; 注意力机制; 时空图卷积神经网络; 迁移学习

中图分类号: TP391.4 **文献标识码:** A **文章编号:** 1005-0086(2022)02-149-08

Recognition of students' online classroom action based on spatiotemporal graph convolutional network

HU Jinlin, QI Yongfeng*, WANG Jiaying

(College of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu 730070, China)

Abstract: In order to effectively identify students' online classroom action, a human skeleton action recognition model integrating global attention mechanism and spatiotemporal convolution network is proposed. Firstly, a global attention module is added between the spatial graph convolutional network and the temporal convolutional network of the Spatiotemporal graph convolutional neural network, and the spatial feature map output by the spatial graph convolutional network is used as the input of the attention module; Secondly, average pooling and maximum pooling operations according to the time dimension are introduced to increase the ability of the model to learn global feature information. Finally, three spatiotemporal graph convolutional neural networks and class activation map (CAM) added to the attention mechanism are used to construct a rich activation map convolutional network with stronger ability to recognize occlusion data (RA-GCNv2-A) model, and realize student online classroom action recognition function through transfer learning. Experimental verification is performed on the NTU-RGB+D and NTU-RGB+D120 two datasets. Compared with the RA-GCNv2 model, the recognition accuracy on the NTU-RGB+D dataset is increased by 1.3% (cross-subject, CS), 1.2% (cross-view, CV), the recognition accuracy on the NTU-RGB+D120 dataset is increased by 1.6% (cross-subject, CSub), 1.4% (cross-setup, CSet) respectively. The experimental results show that the proposed method is an effective way to recognize students' online classroom action.

Key words: human skeleton; action recognition; attention mechanism; spatiotemporal graph convolutional neural network; transfer learning

* E-mail: qiyf@nwnu.edu.cn

收稿日期: 2021-06-03 修订日期: 2021-07-03

基金项目: 甘肃省科技计划项目(18JR3RA097)资助项目

1 引 言

随着互联网的快速发展,人类行为识别在现实生活的许多场景中都得到了广泛的应用,比如人机交互、视频监控、视频理解等^[1]。人类行为识别的核心是如何提取具有区别性的丰富特征,以充分描述人体动作的空间和时间信息。目前,与 RGB 视频行为识别方法比较,由于基于骨架的行为识别方法对动态环境和复杂背景都具有很强的适应性,所以受到越来越多的关注^[2,3]。基于骨架的行为识别方法主要分为^[4]:基于递归神经网络(recurrent neural network, RNN)、卷积神经网络(convolutional neural network, CNN)和图卷积神经网络(graph convolutional network, GCN)。在最新的工作中,通过使用图卷积神经网络构建的模型处理骨架序列表现出了优越的性能,其中,最著名的时空图卷积神经网络(spatial temporal graph convolutional network, ST-GCN)模型^[5]通过构造一个时空图来编码骨架序列,堆叠一组时空图,采用卷积提取特征并进行预测识别。基于图卷积神经网络的人体骨架行为识别中,SI 等^[6]结合长短期记忆(long short-term memory, LSTM)网络和 GCN 提出了注意力增强图卷积 LSTM 网络(attention enhanced graph convolutional LSTM network, AGC-LSTM),该网络模型能够有效地捕获骨架图上具有鉴别性的时空特征信息,提高了学习高层次语义时空特征的能力。SHI 等^[7]提出了双流自适应图卷积网络(two-stream adaptive graph convolutional network, 2s-AGCN)结构,该网络结构通过自适应的方式学习得到骨架图的拓扑结构,改变了传统的手动设置,提高了模型的灵活性。LIU 等^[8]提出了一种多尺度图卷积和统一的时空图卷积算子方法,该方法能够提取多尺度结构特征和进行长期上下文依赖关系建模。SONG 等^[9]构建了丰富激活的图卷积网络(richly activated graph convolutional network, RA-GCNv2)结构,丰富激活的图卷积网络实质上是多流的时空图卷积神经网络模型,每一流是一个 ST-GCN 网络模型,类激活图(class activation map, CAM)^[10]定位激活的关节点。RA-GCNv2 模型在由 NTU-RGB+D120^[11]数据集合成的帧遮挡、部分遮挡、块遮挡和随机遮挡骨架图上识别性能良好,其中,块遮挡骨架图是把腰部以下关节点遮挡,与学生在线课堂行为骨架图非常相似。

随着教育信息化和新冠肺炎疫情的影响,在线教学^[12]已经成为一种普遍的模式。然而,现有的学生课堂行为识别工作都是基于线下的模

式^[13],线上课堂学生行为识别未能得到解决。虽然 RA-GCNv2 模型对存在遮挡的骨架数据具备良好的识别能力,但其只通过可学习的重要性加权去增加局部关节点之间的重要程度,未从全局角度考虑和充分利用注意力机制。本文通过增加全局注意力机制,得到识别能力更优的 RA-GCNv2-A 模型,并迁移在 NTU-RGB+D120 数据集上训练得到的预训练模型,解决学生在线课堂行为识别问题。

2 基于人体骨架的在线课堂行为识别

2.1 人体骨架信息

人体骨架信息可通过硬件方法(Kinect 等深度摄像机)和软件方法(OpenPose^[14]等人体姿态估计算法)获取。一帧中的骨架信息通常由一系列向量表示,每个人体关节的二维或三维坐标由相应的向量表示。人体骨架信息图如图 1 所示。图 1(a)是使用 OpenPose 算法进行姿态估计得到 18 个人体关节点的人体骨架图示例,关节之间通过基于人体的自然连通性连接。针对学生在线课堂行为视频数据,通过姿态估计算法最多能得到 12 个关节点,如图 1(a)中虚线框所示。

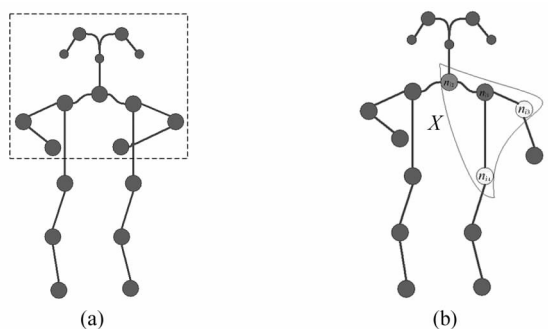


图 1 人体骨架信息图:(a) 人体骨架示例图;
(b) 空间配置分区图

Fig. 1 Human skeleton infographic:
(a) Example of human skeleton;
(b) Space allocation zoning graph

2.2 时空图卷积神经网络

ST-GCN 在得到由坐标构成的骨架序列数据后,沿空间和时间维度对这些关节之间的结构化信息进行建模^[5],空间维度是指同一帧中的关节所在的维度,时间维度是指某一行为所有帧的相同关节所在的维度。时空图卷积神经网络是由 9 个 ST-GCN 单元构成,每个 ST-GCN 单元又由 GCN 和时间卷积网络(temporal convolutional network, TCN)构成,ST-GCN 单元之间加入残差机制(residual),GCN 与 TCN 之间 Dropout 层的丢失率为 0.5,避免

过拟合。TCN 模块由批标准化层、Relu 层、Dropout 层、一维卷积层和批标准化层构成^[5]。

人体骨架图中的每个节点 V_i 的空间图卷积计算如下:

$$f_{\text{out}}(V_i) = \sum_{V_j \in P_i} \frac{1}{Z_{ij}} f_{\text{in}}(V_j) \cdot \omega(l_i(V_j)), \quad (1)$$

式中, f_{in} 表示输入特征图, f_{out} 表示输出特征图, V 表示骨架图中的关节点。顶点 V_i 图卷积时的采样区域定义为 P_i (图 1(b) 中被曲线框起来的部分), 它表示的是所有与节点 V_i 的距离为 1 的相邻节点的集合。由于采样区域 P_i 中的节点个数是变化的, 所以, l_i 作为映射函数 (时空图卷积神经网络中有三种映射函数, RA-GCNv2 选取的是基于距离的映射函数, 本文选择了最优的空间配置分区函数) 把所有相邻节点映射为一个固定数目的子集, 每个子集对应一个唯一的权重向量 (权重向量由权重函数 ω 提供)。 Z_{ij} 是归一化操作, 目的是平衡不同子集对输出的贡献。

图 1(b) 表示空间配置分区函数的具体分区策略, 图中的 X 是骨架重心 (所有关节的平均坐标), 采样区域 P_i 中的 n_{i1} 节点是根节点, n_{i2} 节点表示向心群, 向心群被定义为比根节点更靠近骨架重心的关节点的集合。 n_{i3} 和 n_{i4} 节点是离心组 (采样区域中除去根节点和向心群的其他节点), 权重函数根据空间配置分区函数的结果保证同一子集中的关节点共享可学习的权重函数。

通过用邻接矩阵 $\mathbf{A}$ 表示人体骨架中关节之间的关系, 式 (1) 等价式为:

$$f_{\text{out}} = \sum_{l \in \{0,1,2\}} \omega_l f_{\text{in}} (\mathbf{A}_l^{-\frac{1}{2}} \mathbf{A} \mathbf{A}_l^{-\frac{1}{2}} \otimes \mathbf{M}_l), \quad (2)$$

式中, $l \in \{0,1,2\}$ 代表空间配置分区函数划分的 3 类子集, l 取值为 0 代表根节点子集, l 为 1 代表向心群子集, l 为 2 代表离群组子集。 ω_l 是子集 l 的权重函数, $\mathbf{A}_l$ 是邻接矩阵, $\mathbf{A}_l^{\#} = \sum_j (\mathbf{A}_l^{ij}) + \vartheta$ 是归一化对角矩阵, ϑ 取值为 0.001, 目的是避免计算时归一化对角

矩阵中出现空行, $\mathbf{M}_l$ 是一个可学习的参数矩阵, $\otimes$ 指两个矩阵的按元素相乘操作。

2.3 融合全局注意力机制的 ST-GCN-A 模型和 RA-GCNv2-A 模型

1) 融合全局注意力的时空图卷积神经网络 (ST-GCN-A) 模型

ST-GCN-A 结构图, 如图 2 所示。图 2(a) 是由 9 个 ST-GCN-A 单元构成的 ST-GCN-A 模型, ST-GCN-A 单元之间加入残差机制 (residual)。图 2(b) 是单个 ST-GCN-A 单元的结构图, ST-GCN-A 单元是在 ST-GCN 模型的 ST-GCN 单元基础上增加全局注意力 (attention) 模块, attention 模块得到的权重矩阵与空间图卷积网络得到的输出特征图融合作为 TCN 模块的输入, attention 模块的注意力权重计算式为:

$$\mathbf{M} = \delta(g(\text{AvgPools}(f_{\text{in}}) \oplus \text{MaxPools}(f_{\text{in}}) \mathbf{w})), \quad (3)$$

式中, $\mathbf{M}$ 为 Attention 模块的注意力权重矩阵, f_{in} 是经过空间图卷积网络处理得到的输出特征图, 作为 Attention 模块的输入信息, AvgPools 和 MaxPools 分别指沿时间维度的平均池化操作和最大池化操作, $\oplus$ 表示两个张量的拼接操作, g 为一维卷积操作, $\mathbf{w}$ 是可学习的参数矩阵, δ 为 Relu 激活函数。

2) 融合全局注意力的丰富激活图卷积网络 (RA-GCNv2-A) 模型

图 3 是 RA-GCNv2-A 模型图, 用 3 个 ST-GCN-A 模型替换原来的 3 个 ST-GCN 模型, 再加上两个 CAM^[10] 实现定位激活关节的功能。 $\mathbf{X}$ 是对原始输入的骨架序列数据进行预处理后的数据, 它是由原始骨架数据、相对坐标和时间位移经过级联操作后得到, 使得输入数据更具有信息性。相对坐标是每一帧中其他关节坐标与中心关节坐标的差值, 时间位移通过计算相邻两帧上相同关节的坐标之差得到。mask 是掩码矩阵, $\oplus$ 表示聚合操作, 把三个分支流

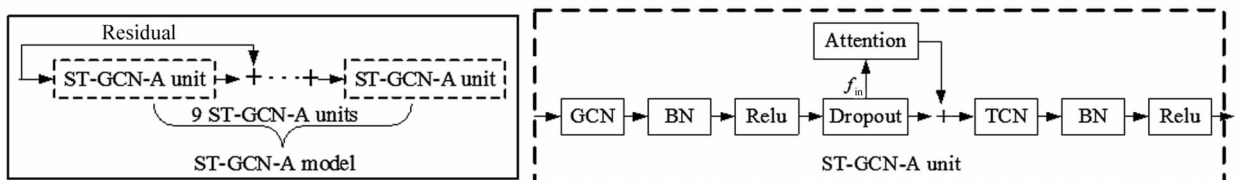


图 2 ST-GCN-A 结构图: (a) ST-GCN-A 模型图; (b) ST-GCN-A 单元图

Fig. 2 ST-GCN-A Structure diagram: (a) ST-GCN-A Model diagram; (b) ST-GCN-A unit diagram

网络的输出合并。FC(fully connected layer)是全连接层,后接 Softmax 分类器,输出识别结果。

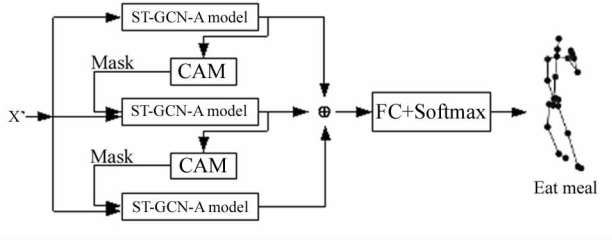


图 3 RA-GCNv2-A 模型图

Fig. 3 RA-GCNv2-A Model diagram

整个 RA-GCNv2-A 模型的工作流程是接收到预处理后的数据输送给第一个分支流,第一流网络处理接收到的数据,然后通过 CAM 确定哪些关节被激活,进而改变掩码矩阵中某关节对应的值,更新后的掩码矩阵传到第二流。第二流网络的输入是预处理后的数据 X' 与掩码矩阵进行逐元素相乘后的结果。同样,第三流网络的输入是 X' 与从第二流得到的掩码矩阵进行逐元素相乘后的结果,最后通过三个支流网络激活关节并把结果聚合,如此保证了第二、三流的输入仅由先前流未激活的关节组成,使得 RA-GCNv2-A 模型能够探索区分所有关节的更多特征信息。

损失函数在交叉熵损失函数的基础上改进,达到监督多流网络的功能,定义计算式为:

$$L = -y \log \hat{y} - \sum_{s=1}^S y \log \hat{y}_s, \quad (4)$$

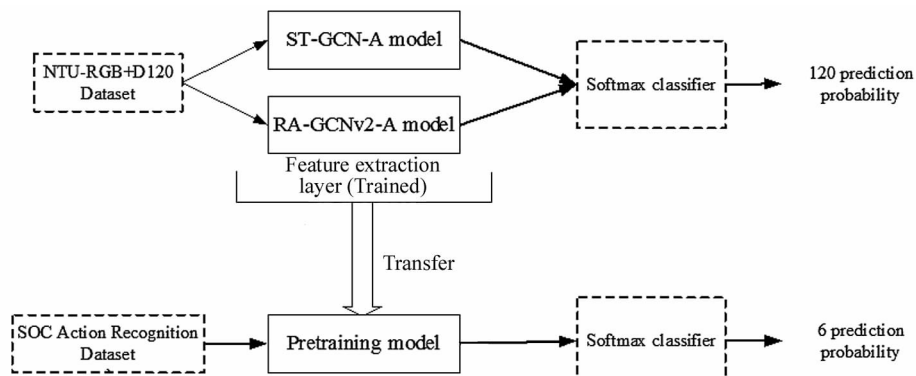


图 4 迁移预训练模型流程图

Fig. 4 Migration pre-training model flowchart

3 实 验

3.1 数据集

1) 学生在线课堂行为识别数据集

针对在线课堂教学环境,主要针对学生上课经

式中, $\hat{y}_s$ 为第 s 个流的预测输出, $\hat{y}$ 是整个模型的预测输出, y 是标注好的真实的值。

2.4 基于迁移学习的学生在线课堂行为识别

目前暂时未有公开的学生在线课堂行为识别数据集,针对自己采集的学生在线课堂行为识别数据集的数据量小及存在过拟合等问题,通过迁移学习,将从相似领域训练得到的模型作为训练目标数据的预训练模型,从而有效地解决相关问题并提高模型的学习效率。

观察发现,RA-GCNv2-A 模型在大型的骨架行为识别数据集 NTU-RGB+D 和 NTU-RGB+D120 上性能良好,并且学生在线课堂行为数据集中的 4 种行为(饮食、玩手机、阅读和写作)都出现在 NTU-RGB+D120 数据集中。以 NTU-RGB+D120 数据集作为源域,学生在线课堂行为数据集(SOC Action Recognition Dataset)作为目标域,发现二者具有有效的相似性。学生在线课堂行为识别数据集是在一个摄像头下采集的一位同学做出的动作数据,是以人为对象的,所以通过迁移 NTU-RGB+D120 数据集上 cross-subject(CSub)划分方式下训练得到的预训练模型。然后,在采集的学生在线课堂行为识别数据集上进行训练精调,使得修改后的网络模型表现出良好的性能。迁移预训练模型去处理学生在线课堂行为数据集,其流程图如图 4 所示。

常出现的饮食(diet)、交头接耳(talk_others)、玩手机(play_phone)、阅读(read)、睡觉(sleep)和写字(write)6 种行为进行数据采集。考虑到学生在线课堂环境复杂多变,受到电脑的摆放位置、学生的坐姿等影响,为此采集数据时加入这些影响因素,使得学

生在线课堂行为识别研究更具有现实意义。

本次实验选取 50 名高校同学作为研究对象。通过模拟在线课堂环境,采集 50 名同学做出上述 6 种行为的数据。每位同学的一种行为需做两组,这样可保证至少有一组是考虑坐姿等影响因素下采集的数据。每种行为的数据都是视频文件,总共采集到 600 组视频文件。为了保证不影响每种行为的识别效果,对视频数据文件进行裁剪,保证每份视频中只有一位同学,通过 OpenPose 提取到的人体骨架只有一份,其中未能检测到的人体骨架关节点的二维坐标都为 0。

使用 OpenPose 进行姿态估计,得到每帧中 18 个关节点的二维坐标(X, Y)和对应的每个关节点置信度 S 。通过用(X, Y, S)元组表示每个关节,一帧骨架信息由 18 个元组构成的数组表示。在实验过程中,对学生在线课堂行为骨架数据集进行打乱,随机抽取其中的 75% 作为训练集,剩余的 25% 作为测试集。

2) NTU-RGB+D 和 NTU-RGB+D120 数据集

NTU-RGB+D^[15]数据集包含 60 种动作,来自 40 名动作对象,共得到 56 880 个动作样本。NTU-RGB+D120^[11]数据集是 NTU-RGB+D 数据集的扩展,包含 120 种动作,来自 106 名动作对象,共 114 480 个动作样本。两个数据集中的骨架数据是人体主要的 25 个关节的 3 维坐标,通过 3 个摄像机实时捕获。NTU-RGB+D 数据集作者给定 cross-subject (CS) 和 cross-view (CV) 两种数据集划分方式,其中 CS 指跨动作对象划分数据集为测试和训练样本, CV

指跨视角划分数据集为测试和训练样本。NTU-RGB+D120 数据集作者推荐了 cross-subject (CSub) 和 cross-setup (CSet) 两种划分方式, CSub 是跨动作对象划分数据集为测试和训练样本, CSet 是所有样本设定编号,偶数编号样本为训练集,奇数编号样本为测试集。

3.2 实验环境与配置

实验硬件环境的配置为:两张 Tesla T4 (GPU), 32 G 显存, Intel(R) Xeon(R) Gold 5218 (CPU), 128 G 内存(恒源智享云算力平台);实验所用的编程语言为 Python,实验使用 Pytorch 框架和 OpenPose(pytorch 版本)开源库等。

综合实验硬件及实验效果考虑,学生在线课堂行为识别数据集训练过程的样本大小为 32、训练轮数设为 65, NTU-RGB+D 和 NTU-RGB+D120 数据集的训练样本大小为 16、训练轮数设为 50, 初始学习率设置为 0.1, 在第 10 轮和第 30 轮迭代时,学习率减小 10 倍,优化策略采用 (SGD) 随机梯度下降算法(其 momentum 设置为 0.9), 权重衰减率为 10^{-4} 。

3.3 实验结果与分析

1) NTU-RGB+D 数据集实验结果及分析

NTU-RGB+D 数据集训练结果图,如图 5 所示。图 5(a)为该数据集的训练准确率变化曲线,图 5(b)为损失率变化曲线。在图 5(b)中 ST-GCN-A 和 RA-GCNv2-A 的损失率值在前 30 轮相差很大,是因为二者选取的损失函数计算方法不一样, ST-GCN-A 采用的交叉熵损失函数,而 RA-GCNv2-A 采用的损失函数计算式(4)。

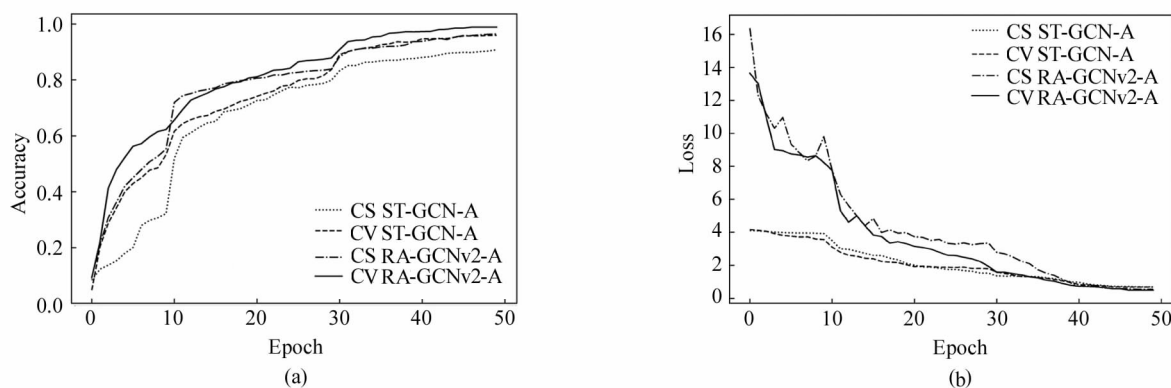


图5 NTU-RGB+D数据集训练结果图:(a)准确率;(b)损失率

Fig. 5 NTU-RGB+D dataset training result graph: (a) Accuracy; (b) Loss

NTU-RGB+D数据集的性能比较如表1所示,分析可知,ST-GCN-A模型在NTU-RGB+D数据集上的测试准确率分别为83.4%(CS)和89.6%

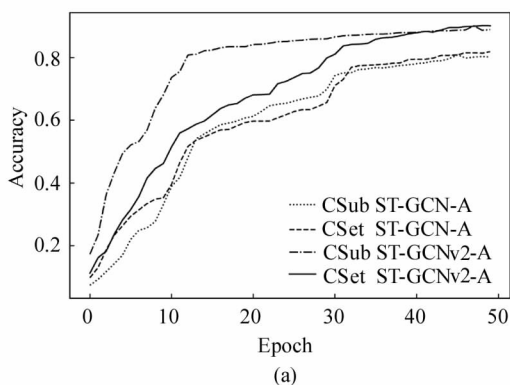
(CV)。RA-GCNv2-A模型在NTU-RGB+D数据集上的测试准确率分别为88.6%(CS)和94.9%(CV)。对比发现在RA-GCNv2模型基础上改变映

射函数和增加全局注意力模块,得到 RA-GCNv2-A 模型在 NTU-RGB+D 数据集上的识别准确率分别提高 1.3%(CS)和 1.2%(CV)。

表 1 NTU-RGB+D 数据集的性能比较

Tab. 1 Performance comparison of NTU-RGB+D dataset

Method	Year	CS Accuracy/%	CV Accuracy/%
ST-GCN ^[5]	2018	81.5	88.3
AS-GCN ^[16]	2019	86.8	94.2
AGC-LSTM ^[6]	2019	87.5	93.5
GR-GCN ^[17]	2019	87.5	94.3
2s-AGCN ^[7]	2019	88.5	95.1
RA-GCNv2 ^[9]	2020	87.3	93.6
ST-GCN-A(ours)	—	83.4	89.6
RA-GCNv2-A(ours)	—	88.6	94.8



2) NTU-RGB+D120 数据集实验结果及分析

NTU-RGB+D120 数据集训练结果图,如图 6 所示。图 6(a)为该数据集的训练准确率变化曲线,图 6(b)为损失率变化曲线。在损失率图中由于两种模型选取的损失函数计算方法不一样,导致两种模型得到的损失率值有差异。通过观察训练过程,发现 RA-GCNv2-A 模型对 NTU-RGB+D120(CSub)数据集具有很好的适应性和更快的收敛速度,原因是融合全局注意力机制后,模型能够学习到更重要的特征信息,对于关节点坐标变化不大的不同动作,具有很强的区分能力。

NTU-RGB+D120 数据集的性能比较如表 2 所示,分析可知,ST-GCN-A 模型的测试准确率分别为 73.5%(CSub)和 75.8%(CSet)。RA-GCNv2-A 模型的测试准确率分别为 82.7%(CSub)和 84.1%(CSet)。对比发现在 RA-GCNv2 模型基础上改变映射函数和增加全局注意力模块,在 NTU-RGB+D120 数据集上的识别准确率分别提高 1.6%(CSub)和 1.4%(CSet)。

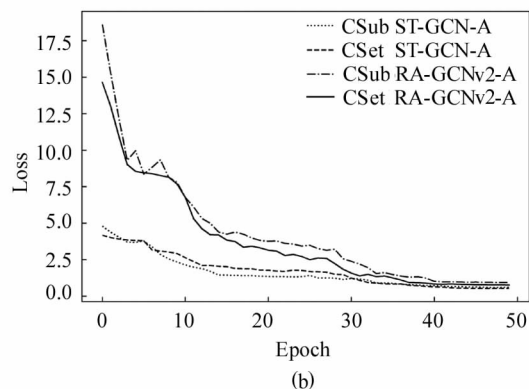


图 6 NTU-RGB+D120 数据集训练结果图:(a) 准确率;(b) 损失率

Fig. 6 NTU-RGB+D120 dataset training result graph:(a) Accuracy;(b) Loss

表 2 NTU-RGB+D120 数据集的性能比较

Tab. 2 Performance comparison of NTU-RGB+D 120dataset

Method	Year	CSub accuracy / %	CSet accuracy / %
ST-GCN ^[5]	2018	70.7	73.2
AS-GCN ^[16]	2019	77.7	78.9
2s-AGCN ^[7]	2019	82.5	84.2
RA-GCNv2 ^[9]	2020	81.1	82.7
ST-GCN-A(ours)	—	73.5	75.8
RA-GCNv2-A(ours)	—	82.7	84.1

3) 学生在线课堂行为识别数据集实验结果及

分析

为了验证模型的效果,对视频数据采用 DeepSort 目标跟踪算法^[18],以此来跟踪标识视频中的学生,然后调用训练好的学生在线课堂行为识别模型实现对数据集的实验,在线课堂行为识别结果可视化图,如图 7 所示。

图 7 为 6 种常见的学生在线课堂行为识别视频结果,每张图左下角的 Human 代表检测到当前视频同学个数,Time 指的是运行时长,Frame 代表当前处理的帧数。图中间(比如 ID-1 write)代表对学生的行为识别结果(ID-1 代表检测到的同学编号)。

为了进一步对比 ST-GCN-A 和 RA-GCNv2-A 预训练模型对学生在线课堂行为数据的识别能力,输出在测试集上得到的混淆矩阵如图 8 所示。纵坐标轴表示真实标签,横坐标轴表示预测标签,对角线上的值为对应识别正确的个数。迁移 ST-GCN-A 预

训练模型对学生在线课堂行为识别数据集上的准确率为 93.1%,RA-GCNv2-A 模型的分析准确率为 96.1%。对比迁移的两个预训练模型,RA-GCNv2-A 预训练模型的识别效果更佳,对一些难以区分的动作更具有识别能力。

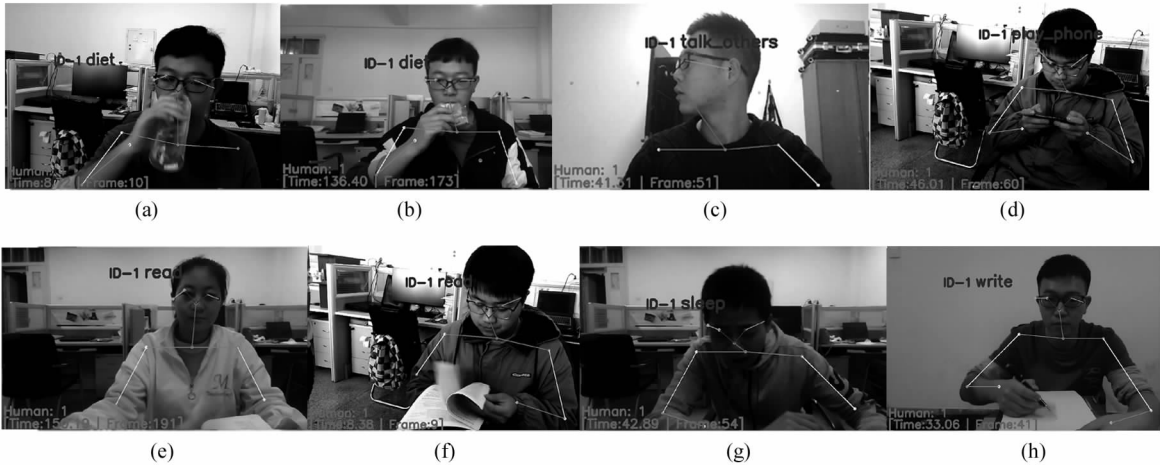


图 7 在线课堂行为识别结果可视化图:(a)(b) 饮食;(c) 交头接耳;(d) 玩手机;
(e)(f) 阅读;(g) 睡觉;(h) 写字

Fig. 7 Visualization of online classroom action recognition results:(a) (b) Diet;
(c) Talk_others; (d) Play_phone; (e) (f) Read; (g) Sleep; (h) Write

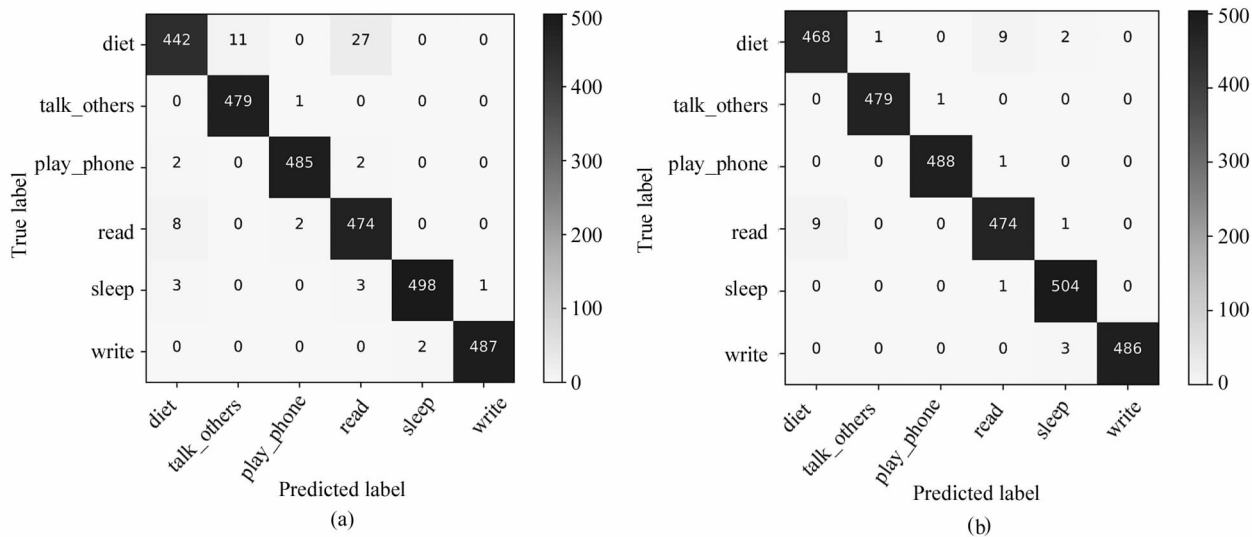


图 8 混淆矩阵:(a) ST-GCN-A; (b) RA-GCNv2-A

Fig. 8 Confusion matrix:(a) ST-GCN-A; (b) RA-GCNv2-A

4 结 论

为了保证在基于人体骨架的行为识别过程中,全局特征信息能得到有效的利用并解决在线课堂行为识别问题,本文提出了一种融合全局注意力机制和时空图卷积网络的人体骨架行为识别方法。首先

把空间图卷积网络处理得到的输出特征图作为全局注意力模块的输入数据,然后把空间图卷积网络的输出特征图与注意力模块得到的权重矩阵融合作为时间卷积网络的输入数据,提取更丰富的特征信息。在两个公开数据集 NTU-RGB+D 和 NTU-RGB+D120 上验证了融合全局注意力机制和时空图卷积网

络方法的有效性,并在学生在线课堂行为识别数据集上得到了良好的识别效果。虽然融合全局注意力的丰富激活图卷积网络(RA-GCNv2-A)模型对学生在线课堂行为识别数据集提供了良好的识别能力,但网络参数量大,识别行为需要花费一定的时间。下一步工作将寻求既能保证高准确率且网络模型更轻巧的学生在线课堂行为识别方法,以便更具有应用价值。

参考文献:

- [1] LI S, YI J, FARHA Y A, et al. Pose refinement graph convolutional network for skeleton-based action recognition[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 1028-1035.
- [2] KAWAMURA K, MATSUBARA T, UEHARA K. Deep state-space model for noise tolerant skeleton-based action recognition[J]. IEICE Transactions on Information and Systems, 2020, 103(6): 1217-1225.
- [3] CUI R, ZHU A, WU J, et al. Skeleton-based attention-aware spatial-temporal model for action detection and recognition[J]. IET Computer Vision, 2020, 14(5): 177-184.
- [4] HUANG Q, ZHOU F, QIN R. View transform graph attention recurrent networks for skeleton-based action recognition[J]. Signal, Image and Video Processing, 2021, 15(3): 599-606.
- [5] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//AAAI Conference on Artificial Intelligence, Feb. 2-7, 2018, Seattle, USA. New York: ACM, 2018: 7444-7452.
- [6] SI C, CHEN W, WANG W, et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 16-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 1227-1236.
- [7] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 16-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 12026-12035.
- [8] LIU Z, ZHANG H, CHEN Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 13-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 143-152.
- [9] 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, 2020, 31(5): 1915-1925.
- [10] PIERSON E, CUTLER D M, LESKOVEC J, et al. An algorithmic approach to reducing unexplained pain disparities in underserved populations[J]. Nature Medicine, 2021, 27(1): 136-140.
- [11] 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.
- [12] DICK G. Teaching online: creating student engagement[J]. Communications of the Association for Information Systems, 2021, 48(1): 65-72.
- [13] WEI Y T, QIN D Y, HU J M, et al. Recognition of students' classroom action based on deep learning[J]. Modern Educational Technology, 2019, 29(7): 88-92.
魏艳涛, 秦道影, 胡佳敏, 等. 基于深度学习的学生课堂行为识别[J]. 现代教育技术, 2019, 29(7): 88-92.
- [14] STRINGER C, WANG T, MICHAELLOS M, et al. Cellpose: a generalist algorithm for cellular segmentation[J]. Nature Methods, 2021, 18(1): 100-106.
- [15] SHAHROUDY A, LIU J, NG T T, et al. Ntu rgb+ d: A large scale dataset for 3d human activity analysis[C]//IEEE Conference on Computer Vision and Pattern Recognition, Jun. 27-30, 2016, Las Vegas, USA. New York: IEEE, 2016: 1010-1019.
- [16] LI M, CHEN S, CHEN X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 16-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 3595-3603.
- [17] GAO X, HU W, TANG J, et al. Optimized skeleton-based action recognition via sparsified graph regression[C]//27th ACM International Conference on Multimedia, Oct. 21-25, 2019, Nice, France. New York: ACM, 2019: 601-610.
- [18] CHEN C, LIU B, WAN S, et al. An edge traffic flow detection scheme based on deep learning in an intelligent transportation system[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(3): 1840-1852.

作者简介:

齐永锋 (1972—), 男, 博士, 教授, 硕士生导师, 主要从事图像处理与模式识别方面的研究。