分层残差结构的时空图网络多目标在线康复动作识别

吴冬梅　白　凡　宋婉莹
（西安科技大学通信与信息工程学院　陕西 西安 710054）

摘　要　时空图卷积网络(ST-GCN)可以自动学习骨架数据的空间和时间特征,不受外界复杂环境的干扰。针对原有模型存在的骨架信息特征提取不充分、局部信息建模不强等问题,提出一种分层残差结构的骨架识别模型(Res2-STGCN)。构造分层残差结构的时空图卷积模块结合原模块组成新的网络模型。通过改变模块的尺度来进一步扩大感受野。调整学习率间隔等参数解决过拟合问题。将 Res2-STGCN 与检测、姿态估计与跟踪算法结合实现多目标康复动作识别。在 NTU-RGB + D 和自建数据集上设计实验,对比基准算法 ST-GCN,改进后最优模型的识别准确率在两种不同的数据划分标准下分别提升了 5.61 百分点和 6.03 百分点,在自建数据集上的平均识别准确率为 99.5%,对复杂动作的识别具有较强的鲁棒性。

关键词　时空图卷积　骨架行为识别　分层残差　多尺度特征

中图分类号　TP391　　**文献标志码**　A　　DOI:10.3969/j.issn.1000-386x.2024.11.028

MULTI-TARGET ONLINE REHABILITATION ACTION RECOGNITION BASED ON SPATIO-TEMPORAL GRAPH NETWORK WITH HIERARCHICAL RESIDUAL STRUCTURE

Wu Dongmei　Bai Fan　Song Wanying
(*College of Communication and Information Engineering, Xi'an University of Science and Technology, Xi'an 710054, Shaanxi, China*)

Abstract　Spatio-temporal graph convolutional network (ST-GCN) can automatically learn the spatial and temporal characteristics of skeleton data without interference from the external complex environment. In order to solve the problems of inadequate skeleton information feature extraction and weak local information modeling in the original model, a skeleton recognition model with layered residual structure (Res2-STGCN) is proposed. The spatio-temporal convolution module with layered residual structure was combined with the original module to form a new network model. The receptive field was further expanded by changing the size of the module. Parameters such as learning rate interval were adjusted to solve the overfitting problem. Res2-STGCN was combined with detection, pose estimation and tracking algorithm to realize multi-target rehabilitation action recognition. Experiments were designed on NTU-RGB + D and self-built data sets. Compared with the benchmark algorithm ST-GCN, the recognition accuracy of the improved optimal model is improved by 5.61 and 6.03 percentage points respectively under the two different data partitioning standards. The average recognition accuracy of the optimized model on self-built data sets is 99.5%, showing strong robustness for the recognition of complex actions.

Keywords　Spatio-temporal graph convolution　Skeleton behavior recognition　Stratified residuals　Multi-scale feature

0　引　言

行为识别旨在利用算法模型自动且准确判断出视频或图像中的个体或群体正在发生行为的类别。早期的行为识别主要提取视频的光流、运动轨迹等相关信息分类行为。这类方法对整个视频进行二维卷积操作,具有计算量庞大、特征表现差、训练时间长等问题。

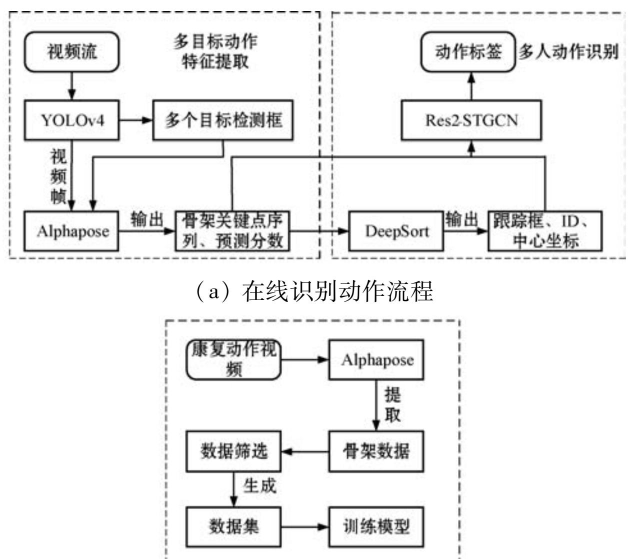
收稿日期:2021－07－19。国家自然科学基金青年科学基金项目(61901358);中国博士后科学基金面上项目(2020M673347)。
吴冬梅,教授,主研领域:智能视频处理。白凡,硕士生。宋婉莹,博士。

近年来,基于计算机视觉的骨架行为识别发展较为迅速。与其他类型数据相比,骨骼数据的提取不受复杂环境限制且具有轻量性的优点,因此被应用到康复动作的识别上,Solongontuya 等[1]使用三层长短记忆网络[2](LSTM)依据动作的高使用频率选择特征进行康复动作的分类。Geng 等[3]将获取康复动作的连续肌电信号序列数据转化成视频图像流送入三维卷积神经网络与 LSTM 识别十个动作,效果良好;Tasnim 等[4]将人体的骨骼坐标转换成时空图像利用迁移学习来捕捉时空特征;闫航等[5]利用 openpose 提取人体动作序列,将姿态特征输入到多层 LSTM 网络中进行康复动作识别。上述方法虽然能够有效地识别康复动作,但是本质上都是对动作的时序特征进行提取,没有对关节信息进行建模。Yan 等[6]首次提出基于时空图卷积神经网络(ST-GCN)的行为识别。由于人体骨架关节本质上是拓扑结构的图数据,该算法以骨骼点为图顶点,自然连接的相邻点关节为帧内边,连续帧的相同骨骼点连接为帧间边,在空间和时间维度上构造时空图卷积网络对骨骼信息动态建模实现行为分类。ST-GCN 模型的卷积块按照顺序依次排列图卷积模块(GCN)和时间卷积模块(TCN)提取骨架的时空特征,受网络层数和卷积核大小的限制,模型的感受野较小,因此较难捕获距离关节点周围距离较远的特征。造成对局部信息建模能力不强、特征提取不充分的问题。管珊珊等[7]提出给时空图网络的每个模块加入残差项以提升 ST-GCN 的表征能力,但是对特征的提取仍然不够完全。而通过加深网络深度以及增大卷积核等线性堆积卷积的方式表达的感受野不够灵活且容易造成网络过拟合,或者并行多个尺度的卷积核同时提取特征,却提升了识别精度又增加了计算负载。

为了解决以上问题,本文提出分层残差结构的时空图卷积网络(Res2-STGCN)。将原模型部分模块改造成分层残差结构,实现在类似负载情况下扩展网络层感受野使得模型细粒化、多尺度提取特征。同时结合目标检测跟踪算法、姿态估计 Alphapose[8]算法以及时空图网络结合生成模型(PoseRes-STGCN)实现实时多目标康复动作识别。利用 Alphapose 提取动作视频中的骨架关节点生成康复数据集,将骨架数据归一化处理后送入时空图卷积网络(ST-GCN)提取丰富的时空信息生成康复识别模型。实验分别在 NTU-RGB + D[9]数据集和自建数据集上进行,结果表明,改进后的模型提取骨骼时空特征能力更强、动作识别准确率更高,且对康复动作的识别有较强的鲁棒性。

1 PoseRes2-STGCN 模型

PoseRes2-STGCN 模型识别多人康复动作可分为两个子问题:多目标动作特征提取和多人动作识别,整体框架如图 1 所示。多目标动作特征提取是对多个目标个体进行检测跟踪并提取骨架信息,包括 YOLOv4[10]检测目标、DeepSort[11]跟踪目标、Alphapose 提取骨架节点坐标构建动作特征。多人动作识别是将跟踪的结果及动作特征送入预训练的 Res2-STGCN 模型预测每个人体目标正在发生的动作。另外,图 1(b)表示动作识别模型 Res2-STGCN 的训练流程。由于 Res2-STGCN 是对 2D 的骨架时序信息处理进行康复动作识别,相比于直接对整个视频进行卷积操作的双流网络、3D 卷积神经网络等算法在训练速度、时间等方面具有较强的优势。



(a) 在线识别动作流程

(b) 数据集构建及模型训练

图 1 PoseSR-STGCN 框架

1.1 目标检测 YOLOv4

YOLOv4 具有识别速度快、准确率高、同时检测多个目标的优点,选择其原因有:1) 整个模型必须实时定位动作视频中的每一个人;2) 每个人的检测框信息对于骨架信息的获取及行为识别是至关重要的。在这个过程中,首先将动作视频转换成连续帧,然后将帧的大小从 1 920 ×1 980 调整到 416 ×416,最后调用网络检测个体,结果用边界框的坐标表示,调整帧的大小的目的是保证检测速度和精度。

1.2 目标跟踪 DeepSort

DeepSort 定义包含了 YOLOv4 得到的边界框中心坐标及高度的 8 维向量$(x,y,\gamma,h,x',y',\gamma',h')$,$\gamma$ 是长宽比,其他四个变量表示对应的在坐标系中的速度

信息。对于每单个追踪目标，假设记 a_k 为当前时间与其之前的最后一次成功匹配的帧数，如果该值大于设定阈值 A_{max}，则当前追踪结束；如果目标多次发生无法与已存在追踪关联的情况，则可认为新目标出现。跟踪算法 DeepSort 引入了运动和外观信息以提高较长遮挡目标追踪的精确度。其中，利用运动目标的已有的 Track 卡尔曼预测结果与新的 dection 测量结果的平方马氏距离关联运动信息，如式(1)所示。其中：d_j 表示第 j 个 detection，y_i 表示第 i 个 track，S_i^{-1} 表示协方差计算。

$$d^{(1)}(i,j)=(d_j-y_i)^{\mathrm{T}}S_i^{-1}(d_j-y_i) \tag{1}$$

利用最小余弦距离关联外观信息，达到更加精确预测个体 ID 的目的。如式(2)所示。其中：$\boldsymbol{r}_j^{\mathrm{T}}\boldsymbol{r}_k^{(i)}$ 表示余弦相似度，$\boldsymbol{r}_j$ 代表每个检测框的外观描述符。

$$d^{(2)}(i,j)=\min\{1-\boldsymbol{r}_j^{\mathrm{T}} \mid \boldsymbol{r}_k^{(i)} \in R_i\} \tag{2}$$

将上述两个距离度量值加权得到综合匹配值，如式(3)所示。

$$c_{i,j}=\lambda d^{(1)}(i,j)+(1-\lambda)d^{(2)}(i,j) \tag{3}$$

式中：λ 表示加权系数。

DeepSort 是在线实时跟踪器，用于执行多人跟踪，本文基于 YOLOv4 的检测结果为每个目标个体创建专属 ID 号，每个人通过其边界框与 ID 相联系。如果成功跟踪目标，显示跟踪框，返回目标的 ID 和中心点位置。

1.3　姿态估计 Alphapose

本文采用具有极高准确度的 Alphapose 算法检测并提取康复视频的骨架关节点，表 1 为主流姿态估计模型在标准数据集 COCO 上的精度对比，AP@ 表示 IOU 设定为某一值时的准确率。可明显看出 Alphapose 算法各项精度指标最优。

表 1　主流姿态估计模型精度对比(%)

模型	AP@0.5:0.95	AP@0.5	AP@0.95
Openpose[12]	61.8	84.9	67.5
Mask R-CNN[13]	63.1	87.3	68.7
G-RMI[14]	68.8	87.1	75.5
Alphapose[8]	73.3	89.2	79.1

Alphapose 的核心在于检测到视频中的目标后分别针对各个个体做关键点检测，以减少漏检以及关节连接偏离，利用对称空间变换网络(SSTN)和姿态非极大值抑制器(PNMS)解决了以往算法检测框定位易错误以及冗余的问题，提升检测精度。

时空图卷积网络精确识别并分类动作是整个模型结构改进的核心部分。

2　Res2-STGCN 网络

为了使模型能够获取更广泛、精细的特征，Gao 等[15]在残差网络(Resnet)的基础上提出了分层残差网络(Res2net)，在 Resnet 网络内部嵌入小的残差块来增大每一个网络层的感受野，降低忽略重要信息的概率，提升网络性能。因此本文提出分层残差结构的时空图卷积模型，并命名为 Res2-STGCN。

2.1　时空骨架图的构建

基于时空图卷积的骨架行为识别通常在时间-空间维度构建动态骨架图，对于输入的连续帧骨架序列，分别以帧内和帧间两步构建时空图。

第一步，依据人体骨骼点之间的自然连接构造单帧帧内空间图。假设第 t 帧有骨骼点数，令对应的第 i 节点矩阵为：

$$\boldsymbol{V}=(v_{ti} \mid t=1,2,\cdots,,i=1,2,\cdots,N) \tag{4}$$

连接任意两个相邻节点的边集可表示为：

$$E_s=\{v_{ti}v_{tj} \mid (i,j)\in H\} \tag{5}$$

式中：H 是人体骨骼点连接的集合。

第二步，将相邻帧之间同一骨骼点连接起来形成时空骨架图，如图 1 的第三部分所示。连续不同帧之间的连接关系为：

$$E_F=\{v_{ti}v_{(t+1)i} \mid (i,j)\in H\} \tag{6}$$

时空骨架图如图 2 所示，将构建好的骨架序列送入分层残差结构的时空图卷积网络训练。

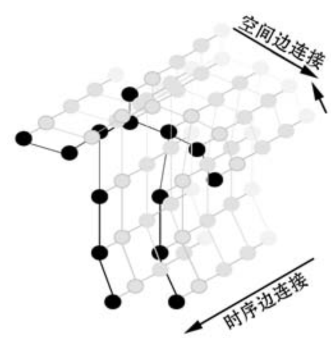


图 2　时空骨架图

2.2　模型结构

本文提出的网络模型的单元模块主要由两部分构成：基础的由 GCN 和 TCN 组成的 stgcn 模块和重构的分层残差结构的 res2-stgcn 模块。网络整体框架如图 3 所示。网络整体包括 1 个批量归一化层(BN)、10 个时空图卷积层、1 个全局池化层及 1 个全连接层。前 4 层输出为 64 通道，中间三层输出为 128 通道，最后两层为 256 通道。每一模块中的 2D 卷积核为 1×1，TCN 的 1D 卷积核为 9，步长取 1。

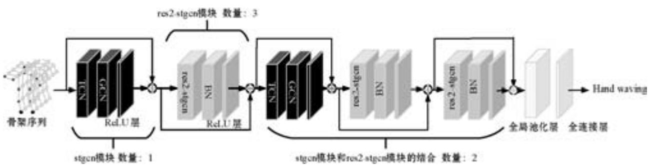


图 3　分层残差的时空图卷积网络整体架构

res2-stgcn 模块对变换后的矩阵进行时空维度归一化，即对一个关节点下的不同帧随机分布的位置信息例如坐标、置信度、帧的数量、节点的边的数量等规范统一化，以达到利于收敛的目的。通过池化层和全连接层对获取到的高级特征进行分类得到动作类别。

2.3　分层残差结构模块

res2-stgc 模块结构如图 4 所示。相比于普通的时空图模块，分层残差结构的每一个子块都可以学习上一个子块的特征，因此可获取更多通道的特征信息，通过使用分层、层叠的特征组增加块内的感受野提高模型对于较远距离关节点之间信息的“抓取”能力。

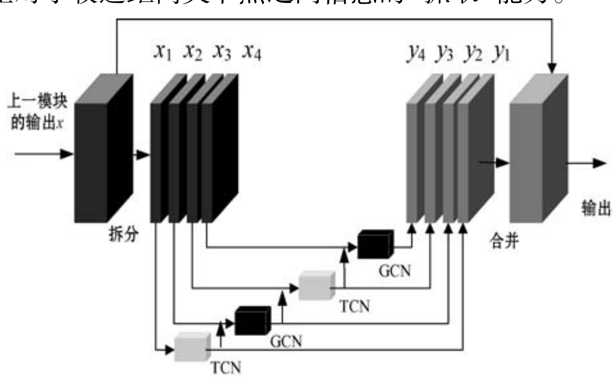


图 4　分层残差结构模块

如图 4 所示，前一层模块的输出经过 1 ×1 卷积被划分为 n 部分，每一部分用 $\boldsymbol{x}_i$表示，其中 $i \in (1,2,\cdots,n)$。所有的 $\boldsymbol{x}_i$具有相同的空间尺寸并且通道数皆为原始的 $1/n$，记 $\boldsymbol{x}_i$对应的输出为 $\boldsymbol{y}_i$，假设每一层的时间卷积或图卷积输出为 $\boldsymbol{K}_i$，$\boldsymbol{y}_i$的计算式如下：

$$\boldsymbol{y}_i = \begin{cases} \boldsymbol{K}_i \boldsymbol{x}_i & i = 1 \\ \boldsymbol{K}_i(\boldsymbol{x}_i + \boldsymbol{y}_{i-1}) & 1 < i \leqslant n \end{cases} \tag{7}$$

模型采用分割-融合的策略，分层的残差连接使得一个块内的感受野具有多个尺度，为了融合不同尺度的特征信息 $\boldsymbol{y}_i$，拆分后的 $\boldsymbol{x}_i$特征经过分层的时空卷积操作以后再进行 1 ×1 卷积整合，将整合后的结果送入下一层网络。

除了深度、宽度、通道数等参数外，改进的模型新增加了一个新的维度，即尺度，它代表了时空块的数量。尺度越大，卷积次数越多，网络层的感受野越大，对高层的复杂特征学习的能力越强。不同的感受野组合有利于全局信息的提取。由于卷积块间以级联的方式连接，因此尺度的增加对内存的影响可忽略不计。

3　实验与结果分析

本文根据专业医生的指导自建了一组康复动作数据集，同时为了保证评价改进算法的客观性与真实性，选取了公共数据集 NTU-RGB + D[9] 对比。

3.1　数据集及评价标准

自建数据集采集 10 位实验者不同环境下以正前方为基准左右偏离 60 度的各个角度的 6 种康复动作视频。康复动作每个动作包含 400 段时长为 10 s 的视频。帧率 30 帧/s。自建数据集的骨架数据是通过 Alphapose 提取视频中个体的骨架数据保存为. json 文件后按照 9∶1的比例划分为训练集和测试集，每类动作的数量占比均衡。数据集示例如图 5 所示。康复动作说明如表 2 所示。

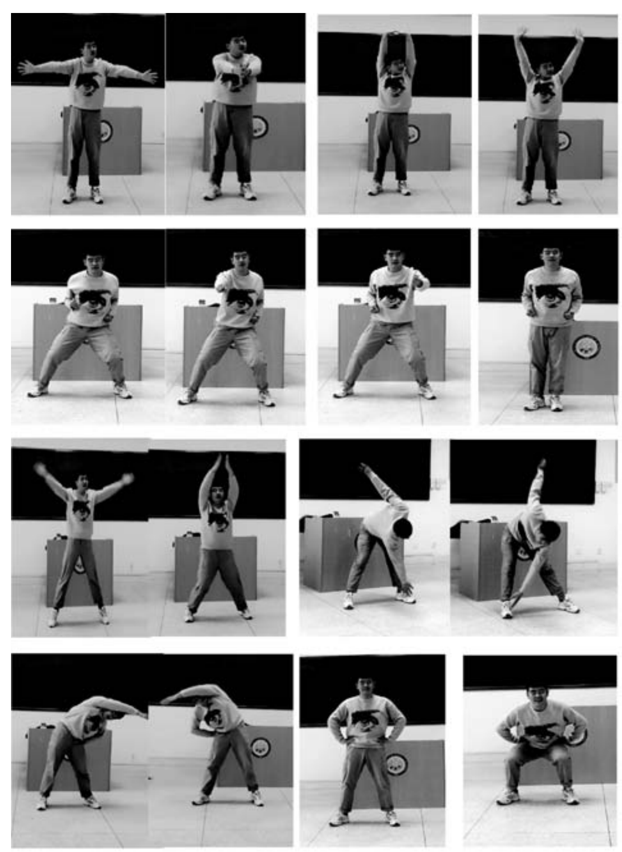


图 5　康复训练数据样本（从左到右依次为占位扣手上举、开合跳、左右深触脚尖、下蹲出拳、弯腰单手上摆、相扑深蹲）

表 2　康复动作说明

动作类型	动作说明	目的
占位扣手上举	两脚平行开立，与肩同宽。两臂分别自左右身侧向上高举过头，十指交叉，翻转掌心极力向上托	提拉胸腹
相扑深蹲	双腿分开，缓慢下蹲，身体微微倾斜	调节骨骼肌肉

续表 2

动作类型	动作说明	目的
左右深触脚尖	双手交替触摸脚尖,双腿伸直不可弯曲	伸展
弯腰单手上摆	占位叉腰,单手掌心向上摇摆两次	协调上下肢
下蹲出拳	身体平直下蹲,左右手依次出拳收拳	改善躯体稳定性
开合跳	身体直立,用力向上跳起,跳到最高处时双手前后交叉,膝盖微屈回到地面	增强运动耐力

NTU-RGB + D 数据集是骨架动作识别领域经典的 3D 数据集之一,包含由 60 类动作提取的 56 880 个骨架序列。所有的动作采集由 40 位不同年龄阶段的人完成,同一动作由摄像机于相同垂直高度的三个水平角度(分别是 −45°、0°、45°)拍摄采集。动作类型主要分三类:第一类是日常行为活动,穿衣服、穿鞋等;第二类是异常动作,例如摔倒等;第三类是交互活动,例如推搡他人、脚踢他人等。本文通过两种数据集评价标准验证算法:1) 交叉对象(X-Sub):该标准下训练集和测试集依据拍摄对象区分,分别来自不同的参与者;2) 交叉视角(X-View):该标准下训练集和测试集依据摄影机的拍摄角度区分,分别来自不同的摄影设备,其中训练集数据选自 2、3 号设备,而测试集选自 1 号设备。

3.2　实验配置及训练策略

实验环境:实验在具有 NVIDIA GTX-2080Ti 显卡、显存 8 GB 的 Ubuntu 18.04 环境下运行,深度学习框架为 PyTorch 1.7。

实验策略:实验过程中,单次输入 180 帧骨架序列,如果某一单个样本不足 300 帧,重复直至规定帧数,训练周期设为 80,优化算法采用随机梯度下降算法,前向传播与验证数据批处理大小均为 32,基准学习率为 0.08,60 次迭代后降低为 0.008,70 次迭代后降低为 0.000 8,dropout 设置为 0.5,采用 Softmax 分类器。

3.3　实验结果分析

3.3.1　分层残差结构时空图网络对识别准确率的影响分析

依据上述的评价标准,分别以 Top-1 以及 Top-5 识别准确率评估动作分类识别性能。本节设计三组对比验证多尺度特征的时空图卷积网络在骨架动作识别的有效性。

第一组对比:不同尺度的分层残差结构模型对比。尺度代表了 GCN 和 TCN 的数量总和,尺度的变化精度提升有一定的影响。实验比较了 2、4、8、16 四种尺度下网络在 NTU + RGB + D 数据集识别精度,如表 3 所示。

表 3　不同尺度下 Res2-STGCN 模型实验结果(%)

尺度	X-SUb	X-View
2	85.47	92.56
4	86.12	93.15
8	86.49	93.32
16	86.51	93.33

可以看出,尺度增大有利于精度的提升,但是网络到一定深度后提升幅度变得不够明显,这是因为网络层数过多,容易产生过拟合,同时计算量变大,训练时间变长。因此综合考虑选取尺度为 8 的模型继续优化。

第二组对比:不同参数下 Res2-STGCN 模型的精度对比。在训练 Res2-STGCN 的过程中发现,模型在训练后期会发生训练损失远小于验证损失的情况,网络发生过拟合,因此通过调整学习率的间隔等参数来保证网络收敛的同时提升准确率。本组实验比较了尺度为 8 时不同迭代次数以及间隔数下 Res2-STGCN 模型的识别准确率。

如表 4 和表 5 的结果所示,Res2-STGCN 模型在 step = [60,70],epoch = 80 时在数据集的两种划分方式下均达到准确率的最优值。

表 4　不同参数下 Res2-STGCN 模型在 X-Sub 的实验结果(%)

参数	Top-1	Top-5
step = [10,50],epoch = 70	86.34	96.59
step = [30,50],epoch = 70	86.38	97.23
step = [40,60],epoch = 70	86.67	97.26
step = [50,60],epoch = 70	86.55	97.48
step = [60,70],epoch = 80	**87.11**	**97.45**
step = [70,80],epoch = 90	86.30	97.32

表 5　不同参数下 Res2-STGCN 模型在 X-View 的实验结果(%)

参数	Top-1	Top-5
step = [10,50],epoch = 70	93.32	98.78
step = [30,50],epoch = 70	93.89	99.12
step = [40,60],epoch = 70	93.24	99.18
step = [50,60],epoch = 70	93.78	99.17
step = [60,70],epoch = 80	**94.33**	**99.17**
step = [70,80],epoch = 90	94.02	99.20

第三组对比:模型改进前后的准确率对比。表 6

展示了分层残差结构的时空图网络(Res2-STGCN)与基准网络(ST-GCN)在两种数据集下的对比。可以看出,分层残差结构网络对比 ST-GCN 提升明显。

表 6　分层残差结构网络在两种数据集上识别准确率对比(%)

网络	X-Sub	X-View	康复数据集
ST-GCN	81.50	88.30	92.87
Res2-STGCN	87.11	94.33	99.50

3.3.2　分层残差结构时空图网络对实时性的影响分析

为了模型更好地应用于实际并保证其实时性,模型参数和计算量要减少,对高性能设备的依赖也要降低。网络结构的改变会带来计算量和内存占用量的变化,具体参数如表 7 所示。

表 7　计算量和内存参数

网络	参数总量	内存
ST-GCN	3 098 832	5 208
Res2-STGCN	1 155 472	4 796

可以看出,分层残差结构网络所需的参数量远小于基准网络,且因卷积块间以级联的方式连接,尺度的增加对内存产生影响不大,并且模型在自建数据集上可达到 20 帧/s,满足实时性的要求。

3.3.3　本文算法与骨架动作识别领域其他算法的精度对比

将 Res2-STGCN 模型在 NTU-RGB + D[9] 数据集上与近年来的方法进行了比较,比较结果如表 8 所示。

表 8　与其他算法比较的结果(%)

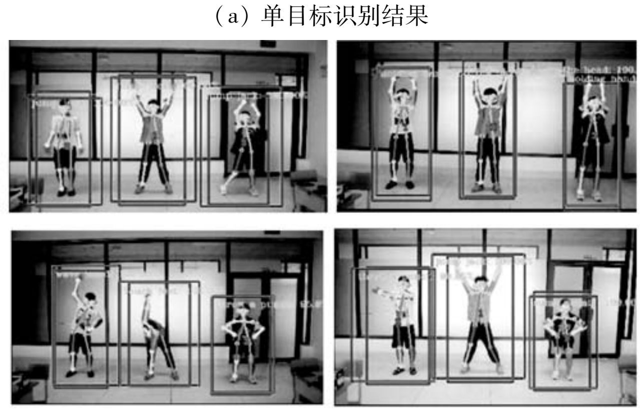
算法	X-SUb	X-View
Spatio-Temporal-LSTM + TS[16]	69.20	77.70
Res-TCN[17]	74.30	83.10
CNN + MTLN[18]	79.60	84.80
ST-GCN[6]	81.50	88.30
Res-STGCN[7]	83.30	89.23
CNN + Motion + Trans[19]	83.20	89.30
3scale ResNet152[20]	85.00	92.30
3S RA-GCN[21]	85.90	93.50
ST-AGCN[22]	86.40	92.10
JRIN-SGCN[23]	86.20	91.90
AS-GCN[24]	86.80	94.20
Res2-STGCN	**87.11**	**94.33**

如表 8 所示,在 NTU-RGB + D[9] 数据集上共比较了 11 种骨架动作识别算法。本文模型在数据集的两种评价标准下准确率有较大提升。对比传统的 LSTM、TCN 及 CNN 精度分别提升了 17.91 百分点和 16.63 百分点、12.81 百分点和 11.23 百分点、7.51 百分点和 9.53 百分点,因此可以得出结论,图卷积更适用于骨架的动作识别。对比其他时空图卷积的方法可以看出本文算法的准确度也均有不同程度的提升,可以证明,分层残差网络有助于时空图卷积网络深度挖掘骨架特征提升识别准确率。

3.3.4　连续康复动作识别的可视化

本文通过结合人体目标检测和跟踪技术来识别单人及多人的连续康复动作。图 6 为模型在单目标和多目标识别场景中对康复动作识别的可视化结果。在实验中,测试者依据自身执行不同的动作。测试单目标时,对于每个动作都有 15 ~ 20 个实例,平均识别准确率可以达到 99.5%。结合测试和跟踪后,模型同样能够准确识别各个目标人体正在发生的动作,且处理速度可达到每秒 20 帧,表明本文方法可以用于实时应用。

(a) 单目标识别结果



(b) 多目标识别结果

图 6　对康复动作识别的可视化结果

4　结　语

本文提出分层残差结构的骨架动作识别模型 Res2-STGCN,对比基准算法,改进后的模型在精度上有明显的提升。模型利用关节点之间的信息构造骨架

序列的时空图卷积捕获人体行为变化的运动特征，引入的分层残差网络可以多尺度、细粒化地提取特征，提高行为识别的准确率。在 NTU-RGB + D 数据集上进行了多次实验，实验证明，分层残差结构的时空图卷积模型优于其他行为识别的模型，具有良好的识别效果。将 Res2-STGCN 与目标检测跟踪等技术结合生成康复动作识别模型（PoseRes2-STGCN），能够实现在线准确识别康复动作。Res2net 的扩展性也为其他模型与 Res2net 相结合提供了方向。人体行为识别未来的研究工作还有很多，例如研究更合适变化的动作的骨骼点的划分规则、更加充分提取特征的神经网络等。

参 考 文 献

[1] Solongontuya B, Cheoi K J, Kim M H. Novel side pose classification model of stretching gestures using three-layer LSTM[J]. The Journal of Supercomputing, 2021, 77(5): 10424 - 10440.

[2] Graves A. Long short-term memory[M]//Supervised Sequence Labeling with Recurrent Neural Networks. Springer, 2012.

[3] Geng T J, Jia X Q, Guo Y L. Lower limb joint nursing and rehabilitation system based on intelligent medical treatment[J]. Journal of Healthcare Engineering, 2021, 2021(4): 1 - 12.

[4] Tasnim N, Islam M K, Baek J H. Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints[J]. Applied Sciences, 2021, 11(6): 2675.

[5] 闫航，陈刚，佟瑶，等. 基于姿态估计与 GRU 网络的人体康复动作识别[J]. 计算机工程，2021，47(1)：12 - 20.

[6] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//32nd AAAI Conference on Artificial Intelligence, 2018: 7444 - 7452.

[7] 管珊珊，张益农. 基于残差时空图卷积网络的 3D 人体行为识别[J]. 计算机应用与软件，2020，37(3)：198 - 201，250.

[8] Fang H S, Xie S Q, Tai Y W, et al. RMPE: Regional multi-person pose estimation[C]//IEEE International Conference on Computer Vision, 2017: 2353 - 2362.

[9] Shahroudy A, Liu J, Ng T, et al. NTU RGB + D: A large scale dataset for 3D human activity analysis[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1010 - 1019.

[10] Bochkovskiy A, Wang C Y, Liao H Y. YOLOv4: Optimal speed and accuracy of object detection[EB]. arXiv: 2004.10934, 2020.

[11] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE International Conference on Image Processing, 2017: 3645 - 3649.

[12] Zhe C, Simon T, Wei S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[13] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//IEEE International Conference on Computer Vision, 2017: 2980 - 2988.

[14] Papandreou G, Zhu T, Kanazawa N, et al. Towards accurate multi-person pose estimation in the wild[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3711 - 3791.

[15] Gao S H, Cheng M, Zhao K, et al. Res2Net: A new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(2): 652 - 662.

[16] Liu J, Shahroudy A, Dong X, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[C]//European Conference on Computer Vision, 2016: 816 - 833.

[17] Kim T S, Reiter A. Interpretable 3D human action analysis with temporal convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017: 1623 - 1631.

[18] Ke Q H, Bennamoun M, An S J, et al. A new representation of skeleton sequences for 3D action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4570 - 4579.

[19] Li C, Zhong Q Y, Xie D, et al. Skeleton-based action recognition with convolutional neural networks[C]//IEEE International Conference on Multimedia & Expo Workshops, 2017: 597 - 600.

[20] Bo L, Dai Y C, Cheng X L, et al. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN[C]//IEEE International Conference on Multimedia & Expo Workshops, 2019: 601 - 604.

[21] Zhang P F, Lan C L, Zeng W J, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2020: 1109 - 1118.

[22] 曹毅，刘晨，黄子龙，等. 时空自适应图卷积神经网络的骨架行为识别[J]. 华中科技大学学报(自然科学版)，2020，48(11)：10 - 15.

[23] Ye F, Tang H M, Wang X W, et al. Joints relation inference network for skeleton-based action recognition[C]//IEEE International Conference on Image Processing, 2019: 16 - 20.

[24] Li M S, Chen S H, Chen X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2019: 3590 - 3598.