基于改进时空图卷积网络的人员交互行为识别

雷静思[1,2,3]　刘双广[1,2,3,4]　刘乔寿[1,2,3*]　王祥雪[4]

[1](重庆邮电大学通信与信息工程学院　重庆 400065)

[2](重庆高校市级光通信与网络重点实验室　重庆 400065)

[3](泛在感知与互联重庆市重点实验室　重庆 400065)

[4](高新兴科技集团股份有限公司　广东 广州 510700)

摘　要　针对人员交互行为识别存在的多模态数据融合方法导致的识别准确率与模型性能无法同时满足的问题,提出一种基于改进时空图卷积网络的人员交互行为识别方法。将单模态骨架数据引入级联的密集时空图卷积块网络中获得丰富的时空特征信息,提高特征复用率;设计一种增强时空图卷积网络(EST-GCN)单元提高网络对关节点之间的信息表征能力;引入一种运动特征因子衡量肢体不同关节的重要程度,提高模型识别效果。在 Kinetics 数据集和办案区场景数据集上的实验结果表明,所提出方法在识别效果上具有一定优势,且该方法在模型复杂度及运行效率上具有很好的竞争力。

关键词　交互行为　时空图卷积网络　骨架数据　密集

中图分类号　TP391.41　TP183　　**文献标志码**　A　　DOI:10.3969/j.issn.1000-386x.2024.04.023

HUMAN INTERACTION BEHAVIOR RECOGNITION BASED ON IMPROVED SPATIAL TEMPROAL GRAPH CONVOLUTION NETWORK

Lei Jingsi[1,2,3]　Liu Shuangguang[1,2,3,4]　Liu Qiaoshou[1,2,3*]　Wang Xiangxue[4]

[1](*School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*)

[2](*Chongqing Key Laboratory of Optical Communication and Networks, Chongqing 400065, China*)

[3](*Chongqing Key Laboratory of Ubiquitous Sensing and Networking, Chongqing 400065, China*)

[4](*Gosuncn Technology Group Co., Ltd., Guangzhou 510700, Guangdong, China*)

Abstract　Aimed at the problems that the recognition accuracy and model performance cannot be satisfied by multi-modal data fusion method for human interaction behavior recognition, a human interaction behavior recognition method based on improved spatial temporal graph convolutional network is proposed. The single-modal skeleton data was introduced into the cascaded densely spatial temporal graph convolutional block network to obtain rich spatial-temporal feature information and improve the feature reuse rate. An enhanced spatial temporal convolution network (EST-GCN) unit was designed to improve the information representation ability of the network between joints. A motion characteristic factor was introduced to measure the importance of different joints in the limbs to improve the model recognition effect. The experimental results on the Kinetics dataset and the case-handling area scene dataset show that the proposed method has certain advantages in the recognition effect, and the method is very competitive in model complexity and operating efficiency.

Keywords　Interactive behavior　Spatial temporal graph convolution network　Skeleton data　Densely

---

收稿日期:2021-01-06。国家自然科学基金项目(61901071);重庆市自然科学基金重点项目(cstc2020jcyj-zdxmX0024)。
雷静思,硕士生,主研领域:计算机视觉,深度学习。**刘双广**,高工。**刘乔寿**,副教授。**王祥雪**,博士。

0 引 言

近年来,随着深度学习的发展,人员行为识别逐渐成为计算机视觉领域的研究热点。人类行为识别的研究成果被广泛应用于智能监控、智能家居、视频检索定位、运动属性分析等日常生活领域[1]。目前针对单人的行为分析已取得一定成果,但缺乏对多人的交互行为研究。与单人动作相比,双人交互行为在日常生活中更常见[2]。同时,随着人们对社会安全的要求越来越高,智能监控系统已经广泛应用到现实场景中。特别地,在办案区场景中,由于办案人员有限且在押人员的特殊性质,在押人员间极易发生肢体冲突。因此,针对特定场合的人员交互行为识别的研究具有重要意义。

目前,人员交互行为识别的方法主要分为基于RGB 视频的方法和基于人体骨架序列的方法。RGB 视频包含了丰富的图像信息,但其容易受光照、复杂背景等的影响[3-5]。研究人员提出的基于人体骨架序列[6-7]的行为识别方法,对光照和背景干扰具有更强的鲁棒性,也更能传递出重要信息。Ke 等[8]将骨架序列转换为三个片段,同时提出多任务卷积神经网络,并行处理每个片段的所有帧,以学习骨架序列的时间和空间信息。姬晓飞等[9]利用 RGB 信息和关节点信息的互补性,设计了一种融合 RGB 视频和关节点数据的双流模型,进一步提高了复杂交互行为识别的准确性。

上述方法通常将骨架信息转化为灰度图或仅利用单个骨骼点信息,忽略了人体关键点之间的自然连接结构。最近,图卷积网络(Graph Convolutional Network,GCN)突破了传统卷积网络处理非欧氏空间数据的局限性[10],将卷积操作拓展到了图数据上。Yan 等[11]设计了一种时空图卷积网络(Spatial Temporal Graph Convolution Network,ST-GCN)通过在人体骨架上构建时空拓扑图和卷积操作实现行为识别,成为第一个将图卷积网络应用到行为识别领域的研究。Shi 等[12]构建多流图卷积网络,自适应地学习关节间的连接关系。Li 等[13]通过探索关节间的依赖关系得到更多全局联合信息,进而实现行为识别和预测。Li 等[14]在 GCN 框架中设计了空间图路由器和时间图路由器,以得到关节间的联通关系及结构信息。Peng 等[15]使用骨架邻接图的高阶表征以及动态图建模机制寻找隐式的关节联系。管珊珊等[16]引入残差项的动态骨架模型提高了人体行为特征的表征能力,增强模型泛化性能。

上述基于图卷积的方法中,大多利用了多模态数据提高交互模型识别效果,即将人体骨架数据采用不同的描述方法形成新的数据。这种方法虽然提高了精度,但由于需要多次训练或增加额外的网络学习特征,导致其在模型复杂度和运行效率上的表现不够优秀。人体骨架数据虽然表现形式单一,但关节间存在着丰富的运动信息。人体运动过程中,关节点之间存在着强相关性,并且肢体上的不同关节往往具有不同的运动幅度。例如,挥手动作时带动腕、肘、肩三个关节的运动,而腕关节比肩关节运动范围更大,具有更明显的判别特征。

针对以上问题,本文基于文献[11]的架构,提出一种改进时空图卷积网络的人员交互行为识别方法。具体地,将单模态的人体骨架数据输入到级联的密集[17]时空图卷积块网络中,增加深层网络特征复用率,获得大量空间域和时间域中的特征信息;同时设计一种增强时空图卷积网络(EST-GCN)单元弥补 ST-GCN 基本单元对人体关节间信息表征能力上的不足,并大幅降低模型参数量;引入一种运动特征因子衡量肢体上不同关节的重要程度,提高模型识别效果。在 Kinetics 公共数据集和办案区场景数据集上的实验结果表明,本文方法能够在极低的模型复杂度和较高的识别速度下实现人员交互行为识别,且识别精度分别达到 32.94% 和 98.19%。

1 相关工作

1.1 时空图卷积网络

图卷积网络将传统卷积网络扩展到图结构数据上,通过定义图上的卷积操作实现相应的任务。GCN 的构造方法主要分为两种:频谱域[18]和空间域[19]。频谱域上,卷积主要利用图傅里叶变换实现,利用图的拉普拉斯矩阵导出其频域上的拉普拉斯算子,再类比到欧氏空间中的卷积以导出图卷积公式。空间域上,其核心在于聚合邻居节点的信息,一种简单的无参数卷积方式是将所有直连的邻居节点的隐藏状态加和来更新当前节点的隐藏状态。

针对基于视频的行为识别任务,时空图卷积网络 ST-GCN 从空域角度上,分别设计了空间维度和时间维度的卷积。ST-GCN 通过定义图上的采样函数和权重函数来聚合邻居关节点的信息。空域上的图卷积关键是在图上找到合适的采样函数,ST-GCN 将节点的输出定义为:

$$f_{\text{out}}(v_{ti}) = \sum_{v_{tj}\in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{\text{in}}(p(v_{ti},v_{tj})) \cdot w(v_{ti},v_{tj}) \quad (1)$$

式中:$B(v_{ti}) = \{v_{tj} \mid d(v_{ti},v_{tj}) \leqslant D\}$ 表示节点 v_{ti} 的邻居集集合,$d(v_{ti},v_{tj})$ 表示从 v_{tj} 到 v_{ti} 的任何路径的最小长度;归一化项 $|Z_{ti}(v_{tj}) = \{v_{tk} \mid l_{ti}(v_{tk}) = l_{ti}(v_{tj})\}|$ 等于相应子集的基数,$l_{ti}(v_{tj})$ 表示当前帧在 v_{ti} 的标签映射;$p(v_{ti},v_{tj})$ 为采样函数;$w(v_{ti},v_{tj})$ 为权重函数。

将空间图卷积网络扩展到时空域,使邻域信息包含关节点的时间信息,获得的时空模型如式(2)、式(3)所示。

$$B(v_{ti}) = \{v_{qj} \mid d(v_{tj}, v_{ti}) \leqslant K, |q-t| \leqslant \lfloor \Gamma/2 \rfloor\} \quad (2)$$

$$l_{ST}(v_{qj}) = l_{ij}(v_{ij}) + (q - t + \lfloor \Gamma/2 \rfloor) \times K \quad (3)$$

式中:q 表示时间维度帧,$q-t$ 定义了时间邻居范围;K 为 v_{tj} 到 v_{ti} 的最短路径,取 $K=1$;Γ 控制邻域的时间范围;l_{ST} 为时空域上的标签映射。

1.2　构建人体骨架拓扑图

人体姿态估计一般分为两种方法:(1) 自顶向下,该方法将人体检测和关键点检测分离,该方法准确率较高但速度较慢;(2) 自底向上,该方法先检测图像中人体部件,然后将图像中多个人体的部件分别组合成人体,这类方法速度更快但准确率稍低。综合考虑,OpenPose[20] 算法在速度和性能上都具有一定优势。

OpenPose 在只能进行单人识别的 CPM[21] 算法上改进获得一个多人姿态估计算法框架,可以提取视频中的多人关键点信息。生成的关键点信息由 (X,Y,C) 组成,其中:(X,Y) 表示关键的位置信息;C 表示该关键点置信度。单个人体由 18 个关键点表示,人体关键点序号及连接方式如图 1 所示。

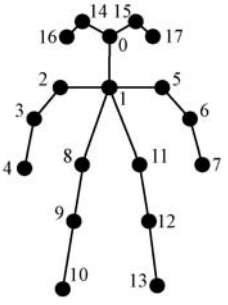

图 1　人体骨架连接

对 N 个关节点,T 帧图像,根据人体自然连接关系,人体骨骼序列图可表示为 $G=(V,E)$。节点集 $V=\{v_{ti} \mid t=1,2,\cdots,T, i=1,2,\cdots,N\}$ 包含了图上所有关节点。边集 E 由两个子集组成,子集 $E_S=\{v_{ti}v_{tj} \mid (i,j) \in H\}$ 表示每帧内关节的连接关系,其中 H 为人体自然连接的关节点集合;子集 $E_F=\{v_{ti}v_{(t+1)i}\}$ 表示连续帧中相同关节点的连接关系。人体骨架序列连接关系如图 2 所示。

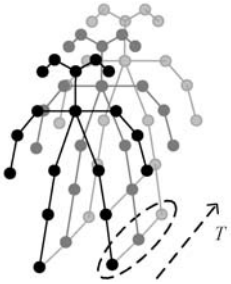

图 2　人体骨架序列

2　改进时空图卷积的人员交互行为识别方法

2.1　改进时空图卷积网络模型设计

改进的时空图卷积网络算法主要由两部分组成:关键点提取和改进的特征提取网络。首先,采用 OpenPose 算法提取视频人体关键点,并构建人体骨架拓扑图;然后,将骨架拓扑图输入到改进的时空图卷积网络中进行特征提取并通过池化层和全连接层输出分类结果。其中,特征提取网络由多个 EST-GCN 单元和密集时空图卷积块级联而成,每个密集块之间使用一个 EST-GCN 单元作为过渡层,主要完成下采样和通道数保持任务。模型整体结构如图 3 所示。

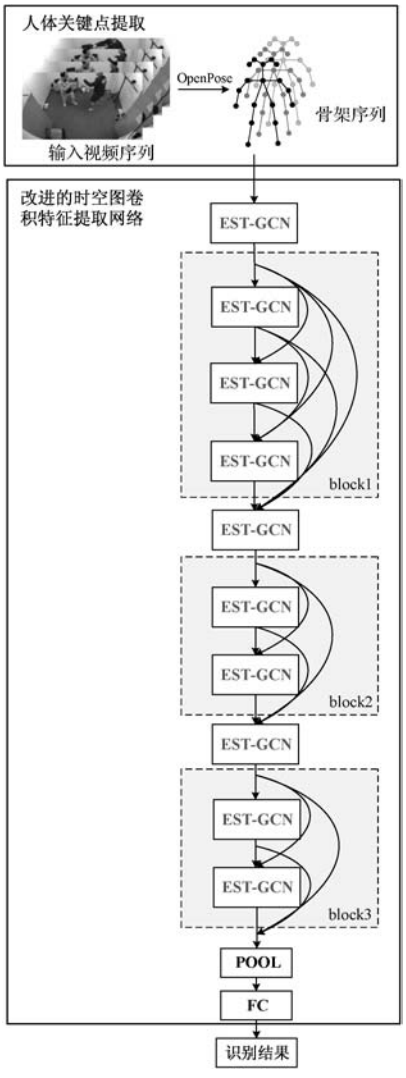

图 3　基于改进时空图卷积网络的人员交互行为识别模型

2.2　密集时空图卷积块

ST-GCN 由 10 层时空图卷积单元级联而成,然而随着模型深度的增加,网络获得的语义信息更加高级,却忽略了低层网络具有的丰富特征。密集连接恰好弥

补了这一缺点,它通过将低层网络特征传递到高层网络中,使得高层网络包含丰富的语义信息,提高网络的特征学习效率。不同于残差连接的加法操作,密集连接通过拼接的方式将低层特征连接到当前层特征的后面,保证每个特征被重新复用。密集连接基本方法如式(4)所示。

$$x_l = H_l([x_1, x_2, \cdots, x_{l-1}]) \tag{4}$$

式中:$[x_1, x_2, \cdots, x_{l-1}]$为前面所有层的输出特征连接而成的张量,经过 $H_l(\cdot)$后输出当前层特征图 x_l。

随着网络层数的增加,密集的连接将使得当前层接受来自上层的全部信息,从而导致网络维度的大幅上升。具体地,当每层输出通道数为 k 时,第 l 层的通道数将增加到 $k_0 + k(l-1)$,其中:k_0为当前层输出特征图数量,k 定义为增长率。

为了解决上述问题,本文通过跳跃式的密集连接将 10 层 EST-GCN 划分为 3 个密集时空图卷积块,仅在每个密集块的内部进行密集连接。每个密集时空图卷积块之间使用一个 EST-GCN 单元避免通道数增加以及完成下采样操作。虽然密集连接只作用于密集时空图卷积块内部,但每个密集块仍会包含前面所有层的信息。

具体结构如图 3 特征提取网络部分所示:第 1 层为输入层,第 2 层至第 4 层为第一个密集时空图卷积块;第 6 层和第 7 层组成第二个密集时空图卷积块;第 9 层和第 10 层组成第三个密集时空图卷积块;第 5 层和第 8 层作为过渡层,同时完成下采样和通道数控制。

2.3 增强时空图卷积单元

ST-GCN 中的基本网络单元由 GCN 块和 TCN 块同时完成时空上的特征采样。为了获得关节点变化信息,GCN 单元利用逐点卷积对单个节点进行采样。这种方式只关注了节点本身的信息,而忽略了与其他关节的内部联系。

因此,本文设计 EST-GCN 单元替换 ST-GCN 基本单元。如图 4 所示,人体关节点信息的空域采样由 EGCN 单元完成,时域上的采样单元与 ST-GCN 使用相同的 TCN 时间卷积单元实现,并使用残差连接 EGCN 和 TCN。

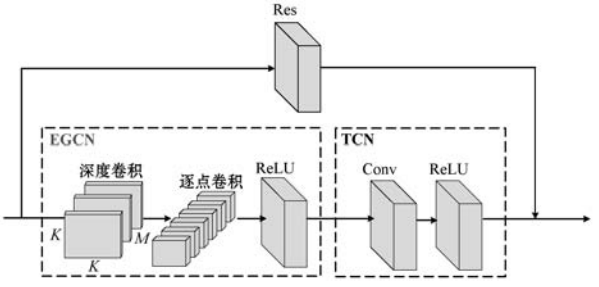

图 4 增强时空图卷积 EST-GCN 单元

为了建立单个关节点与其他节点的联系,EGCN 使用 3×3 卷积核提取邻域内关节点信息,获得肢节的运动特征。然而,与 GCN 块的逐点卷积相比,大尺度的卷积核同时给网络带来了更大的计算量。

文献[22]中提到,普通卷积操作是为了获取通道间相关性和空间相关性的联合映射结果,而卷积层通道间的相关性和空间相关性是可以退耦合的,将它们分开映射可以获得更好的效果,同时可以减少训练参数和模型复杂度。因此,对 EGCN 单元引入深度可分离卷积代替普通卷积可以大幅减少大尺度卷积核带来的网络参数量。深度可分离卷积由深度卷积和逐点卷积组成,深度卷积完成每个通道上的独立卷积,逐点卷积利用空间位置上的特征信息生成新的特征图。

2.4 运动特征因子

时空图卷积网络中,为了实现式(3)中的标签映射,设计三种关节点划分策略来实现图的标签映射:唯一划分、基于距离划分、空间构型划分。

具体地,唯一划分将节点本身及其邻域划为一组;基于距离的划分将节点本身及其邻域分成两个子集,即节点本身为一个子集,相邻节点为一个子集;空间构型划分将节点本身及其邻域分成三个子集:根节点本身,向心节点,离心节点。由于人体结构特性,人体运动时肢体以人体重心为中点,基于空间构型的划分策略更具有合理性。空间构型划分可以用式(5),其中骨骼重心由每一帧图像骨骼上所有关节点的平均坐标计算得出的。

$$l_{ti}(v_{tj}) = \begin{cases} 0 & r_j = r_i \\ 1 & r_j < r_i \\ 2 & r_j > r_i \end{cases} \tag{5}$$

式中:r_j是骨骼重心到所有关节 i 的平均距离;l_{ti}表示单一帧在 v_{ti}的标签映射。

然而,空间构型的划分策略只关注了它们与重心的距离关系,而未注意到不同子集的重要程度。物理学中,质点在做圆周运动时完成相同周期的运动,半径越大的点需要更大的线速度。按照空间构型的划分策略,离心点具有比根节点和向心点更明显的运动状态,而向心点的运动状态最弱。基于此,引入运动特征因子$\{\alpha_i\}_{i=0}^{N}$。对于不同子集,根据其运动幅度给予不同的关注能帮助网络获得更有利的信息。标签映射公式表示如下:

$$\omega = \omega'(\alpha_i l_{ti}(v_{tj})) \tag{6}$$

式中:ω 表示权重函数;α_i表示运动特征因子,通过控制各子集的缩放系数探究运动特征因子对结果的影响。

3　实验数据及预处理

3.1　公共数据集 Kinetics

公共数据集 Kinetics 包括 Kinetics-600 和 Kinetics-400 两个子集,具体为演奏乐器等人机以及握手和拥抱等交互行为,约 65 万个视频剪辑,涵盖 700 个人类动作类,每个动作类至少有 600 个视频剪辑片段,每个剪辑片段由一个动作类注释,持续时间约为 10 s。本文所用的数据集为 Kinetics-400,其中用于模型训练的数据量为 24 万个视频片段,用于评估模型精度的数据量为 2 万个视频片段。

3.2　办案区场景数据集

本文构建了办案区场景的人员交互行为数据集,数据集由真实数据和模拟数据组成。真实数据由公安部门提供,模拟数据分为真实环境下的模拟数据和模拟环境下的模拟数据,模拟数据均由多名志愿者按照真实行为类型进行模拟。该数据集由 3 036 段视频组成,分割为若干段 10 s 的剪辑。原始视频尺寸为 1 920 × 1 080,视频内人员数量不等。

模拟环境旨在模拟办案区内部结构,本文最终选取室内房间作为模拟环境。为了数据采集的多样性,在房间内多角度安装摄像头,人员在摄像区域内随机分布。办案区场景内主要防范在押人员间由于冲突发生的肢体交互。根据研究,本文对肢体冲突进行具体划分。具体地,交互类行为包含双人挥拳、脚踢、锁喉、推搡、抱摔,其他行为包含单人坐立、行走、摆臂。数据集行为展示如图 5 所示。

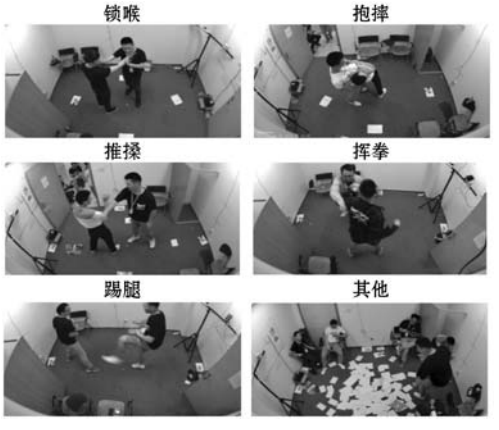

图 5　办案区模拟数据展示

3.3　数据预处理

为了获得人体关键点数据,本文首先将原始视频调整到 340 × 256 像素,然后使用 OpenPose 提取视频中人体关节的 2D 坐标。OpenPose 定位的每个人体均由 18 个关节表示,单个关节采用 (X,Y,C) 的形式表示。其中:(X,Y) 表示关节在图像中的 2D 坐标;C 表示关节置信度。为了减小坐标分布范围过大的影响,对每个 2D 坐标相对视频尺寸执行归一化操作。

4　实　验

4.1　实验平台

本文实验中,训练及测试采用 8 GB 内存、2 张 NVIDIA Geforce GTX 1080 的 GPU,Intel Xeon E5-2620 型号 CPU 的硬件平台和 Ubuntu 16.04 LTS 操作系统。深度学习框架为 PyTorch 0.4.0,采用 Python 脚本语言实现本文算法。

4.2　模型训练与评价指标

本文模型训练参数设置如表 1 所示。其中,在 30 轮训练后,学习率每隔 20 个训练轮次以 0.1 的速率下降,最终获取精度最高的模型。

表 1　模型训练参数设置

训练参数	量值
初始学习率	0.01
epoch	100
动量	0.9
mini-batch	16
梯度下降	SGD
权重衰减系数	1×10^{-4}

同时,本文采用以下 4 个评价指标对本文方法进行评估:

(1) 识别准确率(accuracy,Acc)。识别准确率主要用来评估模型正确识别样本的能力。计算方式如式(7)所示。

$$A_{cc}=(\text{分类正确样本数}/\text{总样本数})\times100\% \quad (7)$$

(2) 模型参数量(parameters,Params)。对于模型空间复杂度的衡量通常引入模型参数量计算,其主要来源于卷积层和全连接层,计算方式如式(8)、式(9)所示。

$$P_{\text{arams}_{\text{conv}}}=k_w\times k_h\times c_{in}\times c_{out}+c_{out} \quad (8)$$

$$P_{\text{arams}_{\text{dw}}}=k_w\times k_h\times c_{in}+c_{in}\times c_{out} \quad (9)$$

$$P_{\text{arams}_{\text{fc}}}=n_{in}\times n_{out}+n_{out} \quad (10)$$

式中:$P_{\text{arams}_{\text{conv}}}$、$P_{\text{arams}_{\text{dw}}}$、$P_{\text{arams}_{\text{fc}}}$ 分别代表普通卷积层、深度可分离卷积层及全连接层的参数个数;n_{in}、n_{out} 分别为当前层输入通道数和输出通道数;k_w、k_h 为卷积核的宽和高。可以看出,卷积层的参数量与卷积核的大小和输入及输出的通道数有关,而全连接层参数量只与通道数有关。

(3) 模型计算量(Floating Point Operations,FLOPs)。FLOPs 浮点运算次数指标从时间复杂度上评估模型性能,直接决定模型前向推理时间。对于卷积层,其 FLOPs 为参数量与当前层特征图大小的乘积。而对于全连接层,其计算量与参数量相等。其中,1 GFLOPs = 10^9 FLOPs。

(4) 单帧识别速度。单帧识别速度用于评价模型完成一次识别的平均每帧处理时间,以 ms 为单位。模型测试平台与实验平台相同。

4.3 办案区场景数据集的实验结果与分析

为了验证本文方法对交互行为识别的效果,本文通过控制变量法分别探讨了密集连接、信息增强单元中卷积核大小及深度可分离卷积作用于模型时对识别准确率、模型复杂度等的影响。实验结果如表 2 所示。

表 2 不同策略下的模型准确率及性能

策略	Params/MB	FLOPs/GFLOPs	速度/ms	Acc/%
—	3.06	5.80	6.89	97.40
$k=16$	0.07	0.22	2.85	96.75
$k=32$	0.19	0.70	3.47	97.53
$k=64$	0.78	2.80	6.12	96.80
Conv3	8.27	17.03	7.56	97.58
Conv3 + Dw	3.08	5.93	7.25	97.86
$k=32$ + Conv3	1.07	3.77	4.03	97.70
$k=32$ + Conv3 + Dw	0.22	0.65	3.75	98.03

由于密集连接增长率 k 将决定网络参数量递增速度,本文设置了不同的 k 值以获得最优的模型,其中 $k=\{16,32,64\}$。由表 2 可知,k 为 16 和 64 时,获得的网络模型准确率分别为 96.75% 和 96.80%,相比 k 为 32 时未获得更好的提升。原因为:(1) 过小的增长率导致网络学习不到足够的信息;(2) 网络中的密集连接对模型性能的提升和增长率之间并不成正比关系。因此,当 k 为 32 时模型获得最高的准确率 97.53%,相比其他两种 k 值准确率分别提升了 0.78 百分点和 0.73 百分点,单帧识别速度较 k 为 64 时提升了 2.65 ms,较 k 为 16 下降了 0.62 ms。由此可知,本文方法的密集连接增长率选择 k 为 32 的条件下进行。

表 2 中,使用 Conv3 方法表示将 EGCN 块的卷积核大小设置为 3×3,并使用 padding = 1 填充以保持输入和输出特征图的大小一致性。其模型准确率提升到 97.58%,但随之带来了模型性能的下降。联合 Conv3 和深度可分离卷积方法的结果显示,在模型复杂度和运行速度表现相当时,其在准确率上获得了 97.86%。

当同时使用密集连接和 EGCN 模块时,网络模型准确率提升到 98.03%,模型参数量和计算量都具有较好的结果,单帧识别速度仅为 3.75 ms,获得了最好的效果。

进一步,为了探究不同权重因子对模型的影响,本文在原始权重(1,1,1)基础上按照比例设置了三组不同的权值参数对关键点集合进行缩放,并通过实验测试了其对精度的影响。每组实验中的三个参数分别表示近心点、根节点和离心点的权值。实验结果如表 3 所示。

表 3 不同特征因子权重值下的模型准确率

算法	权重值	Acc/%
本文方法	(1,1,1)	98.03
	(1,2,4)	98.19
	(−1,1,4)	98.03
	(−1,2,4)	97.53

设置子集权值为(1,2,4)时,准确率为最优,较原始权重(1,1,1)提高了 0.16 百分点。当权重值设置为(−1,2,4)时,模型准确率下降 0.5 百分点。综上,本文方法中不同子集的缩放程度会对模型产生不同的影响,如果只弱化近心节点集,而放大根节点和离心点的作用,将使模型不能较好地学习到行为特征,从而导致网络模型准确率下降。

为了进一步验证本文算法的效果,本文在办案区数据集上对 2s-AGCN、ST-GCN、本文算法分别在 GPU 上进行了测试,实验结果如表 4 所示。

表 4 办案区数据集不同算法实验结果对比

算法	Params/MB	FLOPs/GFLOPs	速度/ms	Acc/%
2s-AGCN[24]	6.84	13.40	15.29	98.87
ST-GCN[11]	3.06	5.80	6.89	97.40
本文方法	0.22	0.65	3.75	98.19

结果显示,在模型空间复杂度上,本文算法模型参数量为 0.22 MB,较 2s-AGCN 和 ST-GCN 明显下降;在模型时间复杂度上,本文算法计算量仅为 0.65 GFLOPs,约为 ST-GCN 的十分之一;在算法准确率度量上,本文算法较 ST-GCN 提高了 0.79 百分点,而低于 2s-AGCN 算法 0.68 百分点;在模型处理速度上,本文算法耗时 3.75 ms,低于 ST-GCN 算法的 6.89 ms 和 2s-AGCN 的 15.29 ms。

4.4 公共数据集 Kinetics 的实验结果与分析

为了客观证明本文算法的效果,本文在公共数据集 Kinetics 上与其他五种算法进行比较,实验结果如

表 5 所示。

表 5　Kinetics 数据集上对比实验结果

算法	Params/MB	FLOPs/GFLOPs	速度/ms	Acc/%
TCN[23]	—	—	—	20.30
ST-GCN[11]	3.16	5.80	6.89	30.70
STGR-GCN[14]	>3.16	>5.80	>6.89	33.60
2s-AGCN[24]	6.98	13.40	15.29	36.10
MS-AAGCN[12]	>6.98	>13.40	>15.29	37.80
本文方法	0.26	0.65	3.75	32.94

表 5 中 TCN 表示传统卷积方法上实现的时域卷积。ST-GCN 是第一个提出将图卷积用于人体行为识别的方法,也是本文的基线方法。STGR-GCN 在保留 ST-GCN 骨干网络的同时,额外增加了空间路由和时间路由获取骨架更丰富的信息,虽然精度得到了一定提升,但额外的网络模块也使得模型整体复杂提高。2s-AGCN 利用自适应模块学习人体骨架的高阶信息,融合关节流和骨骼流模型获得精度的提升。MS-AAGCN 在 2s-AGCN 双流结构基础上增加网络分支,利用多流网络的融合效果作为最终结果。与 STGR-GCN 相似,MS-AAGCN 同样以模型性能换取了准确率的提升。

根据表 5 的结果,与 TCN 相比,本文方法在准确率上提升了 12.64 百分点,说明图卷积方法在基于人体骨架的交互行为识别任务中的突出表现;与基线方法相比,本文方法不仅在准确率上提升了 2.24 百分点,模型复杂度和识别速度也表现得更好;与其他模型(STGR-GCN,2s-AGCN,MS-AAGCN)相比,本文方法虽然准确率稍有落后,但在模型性能上具有最好的结果。

综上所述,本文算法虽然在 Kinetics 数据集上没有获得最高的准确率,但在模型速度和复杂度上具有明显优势,这在工程应用中至关重要,可以节省计算成本和硬件成本。

5　结　语

本文提出一种改进时空图卷积网络的人员交互行为识别方法。本文方法通过级联的密集时空图卷积块网络获得了大量单模态人体骨架数据的时空特征,提高特征复用率;通过 EST-GCN 单元获得人体关节间的相关性,提高网络的特征提取能力的同时减小模型复杂度;根据关节点运动程度的差异,引入一种运动特征因子缩放其重要性,帮助模型更好地识别运动特征。实验结果表明,本文算法在获得准确率提升的同时,在模型复杂度和识别速度上均优于其他算法,证明了本文方法在人员交互识别任务上具有良好的效果。由于基于图卷积网络的行为识别任务依赖人体关节点数据的准确度,因此如何提高人体关节点数据的有效性,进而提高人员交互行为识别的准确率是未来工作的研究方向。

参　考　文　献

[1] 胡琼,秦磊,黄庆明. 基于视觉的人体动作识别综述[J]. 计算机学报,2013,36(12):2512 - 2524.

[2] 杨文璐,于孟孟,谢宏. 基于关键姿势的双人交互行为识别[J]. 计算机应用,2020,40(8):2231 - 2235.

[3] Liu K, Liu W, Gan C, et al. T-C3D: Temporal convolutional 3D network for real-time action recognition[C]//32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence,2018:7138 - 7145.

[4] Ke Q, Bennamoum M, An S, et al. Leveraging structural context models and ranking score fusion for human interaction prediction[J]. IEEE Transactions on Multimedia,2018,20(7):1712 - 1723.

[5] 李洪均,丁宇鹏,李超波,等. 基于特征融合时序分割网络的行为识别研究[J]. 计算机研究与发展,2020,57(1):145 - 158.

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

[7] Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition,2015:1110 - 1118.

[8] Ke Q, Bennamoun M, An S, et al. Learning clip representations for skeleton-based 3D action recognition[J]. IEEE Transactions on Image Processing, 2018, 27(6): 2842 - 2855.

[9] 姬晓飞,秦琳琳,王扬扬. 基于 RGB 和关节点数据融合模型的双人交互行为识别[J]. 计算机应用,2019,39(11):3349 - 3354.

[10] 孔玮,刘云,李辉,等. 基于图卷积网络的行为识别方法综述[J]. 控制与决策,2021,36(7):1537 - 1546.

[11] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//American Association for Artificial Intelligence,2018:7444 - 7452.

[12] Shi L, Zhang Y, Cheng J, et al. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[J]. IEEE Transactions on Image Processing,2020,29:9532 - 9545.

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

[14] Li B, Li X, Zhang Z, et al. Spatio-Temporal graph routing for skeleton-based action recognition[C]//33rd AAAI Conference on Artificial Intelligence and 31st Innovative Applications of Artificial Intelligence Conference and 9th AAAI Symposium on Educational Advances in Artificial Intelligence,2019:8561 - 8568.

[15] Peng W, Hong X, Chen H, et al. Learning graph convolutional network for skeleton-based human action recognition by neural searching[C]//34th AAAI Conference on Artificial Intelligence,2020:2669 - 2676.

[16] 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.

[17] Huang G, Liu Z, Weinberger K, et al. Densely connected convolutional networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2261 - 2269.

[18] Levie R, Monti F, Bresson X, et al. Cayleynets: Graph convolutional neural networks with complex rational spectral filters[J]. IEEE Transactions on Signal Processing,2019,67(1):97 - 109.

[19] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[EB]. arXiv:1710.10903,2018.

[20] Cao Z, Hidalgo G, Simon T, et al. Openpose: Realtime multi-person 2D pose estimation using part affinity fields [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2021,43(1):172 - 186.

[21] Wei S, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition,2016:4724 - 4732.

[22] Chollet F. Xception: Deep learning with Depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition,2017:1800 - 1807.

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

[24] Shi L, Zhang Y, Cheng J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:12018 - 12027.

(上接第 134 页)

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition,2014:580 - 587.

[5] Girshick R. Fast R-CNN[C]//IEEE International Conference on Computer Vision,2015:1440 - 1448.

[6] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137 - 1149.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition,2016: 779 - 788.

[8] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]//European Conference on Computer Vision,2016:21 - 37.

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

[10] Sandler M, Howard A, Zhu M L, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510 - 4520.

[11] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection [C]//IEEE International Conference on Computer Vision,2017:2980 - 2988.

[12] Redmon J, Farhadi A. Yolov3: An incremental improvement [EB]. arXiv:1804.02767,2018.

[13] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(9):1904 - 1916.

[14] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition,2018:8759 - 8768.

[15] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]//15th European Conference on Computer Vision,2018:3 - 19.

[16] Chen Y, Bai Y L, Zhang W, et al. Destruction and construction learning for fine-grained image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition,2019:5152 - 5161.

[17] Hou Q B, Cheng M, Hu X W, et al. Deeply supervised salient object detection with short connections[C]//IEEE Conference on Computer Vision and Pattern Recognition,2017: 3203 - 3212.

[18] Everingham M, Gool L, Williams C, et al. The pascal visual object classes(VOC) challenge[J]. International Journal of Computer Vision,2010,88(2):303 - 338.