doi:10.3969/j.issn.1672-1292.2024.04.006

骨骼双流注意力增强图卷积人体姿态识别

陈　斌,樊飞燕,陆天易

(南京师范大学信息化建设管理处,江苏 南京 210023)

[**摘要**]　为解决骨骼关键点分类算法中运动时间线中运动关联信息的价值分析缺乏,以及骨骼节点关联性和依赖关系信息含义丢失问题,提出了一种骨骼双流注意力增强图卷积人体姿态识别模型. 以提取骨骼特征节点为基础,构建骨骼关节点之间空域连接矩阵和运动时间线时域信息矩阵,在此基础上进行双流骨骼节点信息处理. 利用通道注意力机制对上下文处理的优势,构解关键节点间依赖关系以及全局骨骼运动含义,构建邻域节点加权的双域骨骼拓扑. 在 Kinetics 和 NTU RGB+D 两个数据集上的对比验证显示,该模型在不同数据集上均有较好的执行效果. 与领域内较具代表性的主流方法的横向比对显示,该模型在选定的 9 种行为姿态的识别精度上均优于其他模型. 该方法在人体姿态识别上体现了较优的识别率及稳定性,并佐证了时空双域骨骼特征信息的挖掘价值.

[**关键词**]　姿态识别,时空双域,注意力机制,图卷积,骨骼特征,运动信息表示
[**中图分类号**]TP39;TH691.9　[**文献标志码**]A　[**文章编号**]1672-1292(2024)04-0057-11

Bone Dual-Stream Attention Enhancement Graph Convolving Human Posture Recognition

Chen Bin,Fan Feiyan,Lu Tianyi

(Information Construction Management Division,Nanjing Normal University,Nanjing 210023,China)

Abstract:In order to solve the lack of value analysis of motion correlation information in the loss of meaning of skeletal nodes and dependency information, the paper proposes a model of bone dual-stream attention enhancement graph convolving human posture recognition. The airspace connection matrix and time domain information matrix between bone joints are constructed on the basis of extracting bone feature nodes. With this basis, dual-flow bone node information processing is performed. Taking advantage of the channel attention mechanism for context processing, decturing key node dependencies and global bone motion implications, a two-domain bone topology weighted by neighborhood nodes is constructed. The comparative validation on two datasets Kinetics and NTU RGB+D shows that the model performs better on different datasets. Horizontal comparison with the more representative mainstream methods in the field is shown, the model outperforms the other models in the recognition accuracy of the nine selected behavioral poses. This method reflects the better recognition rate and stability in human posture recognition, and proves the mining value of spatial-temporal dual-domain bone feature information.

Key words:posture recognition,time and space double domain,attention mechanism,figure convolution,skeletal features, movement information representation

人体姿态识别的原理是将所捕获视频中的姿态进行分类,以达到对姿态行为做出判断的目的. 人体姿态识别是人体动作识别的基础,而姿态本身也是动作序列的一部分. 人体姿态识别通常以行为特征分析为基础,主要行为特征对象有光流特征、图像深度特征、RGB 视频特征和骨骼特征等. 在多种人体姿态识别典型方法中,基于骨骼特征的行为识别研究近年来较为热门. 与传统基于视频的行为分类研究相比,骨骼特征数据是对人体骨架关键点的一种近似拓扑信息呈现,其更具稳定性、抗干扰性,受姿态运动变化

收稿日期:2024-06-20.
基金项目:江苏省现代教育技术研究 2023 年度智慧校园专项课题(2023-R-107311).
通讯作者:陈斌,博士,高级工程师,研究方向:模式识别、机器学习、大数据分析方面的研究. E-mail:60167@njnu.edu.cn

及角度变化的影响较小,且由于特征维度相对较低,消耗的计算资源较低,计算效率相对较高.

在人体姿态识别研究的初期,使用较为广泛的方法主要是基于人体几何特征法[1]. 由于人体姿态往往有着较大的变化,而人体几何特征的提取受这种形变的影响非常明显,于是发展出以运动特征为基础的研究方法. 运动特征提取主要聚焦于随时间线变化的人体运动过程的变化[2],较有代表性的方法有基于光流场信息的多方向矢量图方法[3],其通过对横轴纵轴的双向光流场信息的坐标系矢量分析,对分析结果以归一化方式提取运动特征信息. 随着神经网络和机器学习在智能图像领域的深入研究,针对人体姿态识别领域的研究也在逐渐拓展. 基于神经网络的姿态识别研究主要有基于卷积神经网络、循环神经网络、Transformer 网络及图卷积网络等. 其中,图卷积网络由于引入了图的拓扑结构中信息之间的关联性,促使识别性能有了较大的提高[4],较为经典的有图卷积理论及骨架序列在动态时空交互变化基础上与人体拓扑节点邻接矩阵构成的基于时空图的卷积神经网络[5]、双流自适应图卷积网络[6]和有向图神经网络等[7].

注意力机制以其有利于改善卷积神经网络的上下文关联性的优点,近年来在图像分析及人体骨架提取领域已有较成功的研究基础,具有代表性的研究有时空注意力神经网络在行为识别研究中的应用[8]、全局注意力网络在图像像素类内关系连接上的研究[9]、在人体行为识别中的双重注意力网络的研究[10]、分别以关节点数据及成对骨架数据为输入的双流注意力循环网络[11]、以注意力机制为基础的视频帧逐帧分级关注骨骼动态信息增强获取机制[12]、用于骨架行为识别的多维特征嵌合注意力机制所提出的通过增强特征表现能力对图卷积网络识别性能的提升以及对时空双域动态信息和通道维度中蕴含的上下文依赖关系的解析能力[13].

尽管在人体骨架提取及姿态识别领域已有一些前沿性的研究,与早期方法相比在识别准确性、效率方面也有了较大提升和改进,但仍普遍存在一些问题. 首先,大部分相关研究聚焦在人体骨骼关节的提取输入方面,而更多的输入仅以坐标参数作为关键点信息[14],分类算法以此信息进行分类,忽视了动作时间线中运动关联信息的保留和分析. 其次,即便有些研究考虑到了动态骨架点运动的关联性情况,但对骨骼关键点之间的依赖关系,以及这些依赖关系在运动过程中体现出的信息含义缺乏深入关联分析.

为解决以上问题,本文聚焦于提出一种骨骼双流注意力增强图卷积人体姿态识别模型. 以时域空域双流通道融合交互提取人体骨骼关键点,抽取对人体姿态识别有重要影响的关节点,并以关节连接及非连接时间线相关性建立关节点运动信息矩阵,以此为基础构建局部运动信息关联模型. 通过引入注意力增强机制对时空双流通道的采集信息进行上下文关联性分析,并对关键点之间的依赖关系及全局骨骼运动信息进行定义和构造. 在此基础上,通过图卷积机制对双流骨骼运动信息及注意力增强运动含义信息进行关联分析,按照邻域节点对动作影响重要程度划分相应权值,建立双域骨骼拓扑,构建运动含义模型.

1 基本原理及相关工作

1.1 人体姿态识别原理及方法

人体姿态估计是一种以单一目标导向为基础的估计任务,目的在于通过简单的图像输入对其中人体关键点位置的坐标进行快速判断,从而达到人体姿态估计的目的[15]. 人体姿态估计往往着眼于有限局部范围的姿态位置估计,是一种相对简单低层级的分类,也是人体姿态识别的基础[16]. 人体姿态跟踪介于人体姿态估计与人体姿态识别之间,是一种递进的人体姿态估计,主要针对三维图像或视频流,但处理方式上还是以人体关键点独立的空间位置关系为基础[17]. 人体姿态识别本质上是在人体姿态估计基础上建立的更高级的人体结构化表示及处理,是对人体姿态行为的更精准更深层的语义构建及语义理解,因此属于更高级别的分类实现[18]. 常用的人体姿态识别方法往往都会与人体姿态估计相结合,以二维或三维人体骨骼关键点的特征提取和预测估计作为基础,再与其他预测网络、深度学习网络、高层分类算法等结合,实现高级别的分类.

1.2 人体姿态识别代表性数据集

作为以数据分析为驱动力的人工智能技术,科学、精准、有效的数据集样本库是关键基础资源. 人体姿态识别研究虽然发展时间较短,但仍有一些数据集有着较好的使用效果.

NTU RGB+D 数据集由微软 Kinect 建立,包含 NTU RGB+D60 及 NTU RGB+D120 两个版本,分别包含了 56 880 个数据样本和 114 480 个数据样本. 样本类型包含 RGB 视频、深度图序列、3D 骨骼数据和红外

视频[19],分辨率为 1 920×1 080(RGB)和 512×424(其他),视频帧关节点数量为 25.

KTH 数据集包含了行走、慢跑、奔跑、拳击、挥手和鼓掌 6 类行为数据,每类数据都由 25 位受试者在无尺度变化室外、有尺度变化室外、无尺度变化室内、有尺度变化室内 4 种不同的场景采集制作,共有 2 391 个数据样本,帧率为 25 fps,分辨率为 160×120,平均时长为 4 s[20].

Hollywood 数据集包含 Hollywood、Hollywood2 及 Hollywood extended 3 个版本[21],分别包含了 8 类行为共 233 个视频样本、12 类行为共 3 669 个视频样本、16 类行为共 937 个视频样本.

UCF 数据集包含 UCF Sports Action Dataset、UCF YouTube、UCF50 及 UCF101 等多个版本[22],数据样本来源于广播电视频道或互联网视频网站. 其中,UCF Sports Action Dataset 侧重于专业体育运动类姿态;UCF YouTube 侧重于通用运动姿态;UCF50 及 UCF101 区别在于数据集数量不同,主要针对人物互动、单纯身体动作、多人交互动作、器乐演奏及运动五大分类.

HMDB51 数据集内容主要采集于电影片段,另有一部分来源于公共网络视频平台,包含 6 849 个视频,含 51 个行为分类,每个分类至少包含百个以上视频[23]. 这些分类总体上分为五个大类,包括一般类面部动作(微笑、说话、眨眼等)、控制类相关面部动作(抽烟、喝水、吃饭等)、一般类身体运动(静坐、跳跃、跑步等)、人与物交互类身体动作(踢球、骑车、射击等)以及人与人交互类身体动作(握手、拥抱、搏斗等).

HiEve 数据集是一个以人为中心的复杂事件数据集[24],包含了公共用餐、地铁上下车、地震逃生等不同挑战性场景. 该数据集的提出主要是针对密集人群或反常的个人或集体行为的复杂事件,旨在理解各种现实事件中,尤其是聚集性人群及复杂事件中的人体动作和姿态. 该数据集包含 9 种不同场景下的大量姿态数(>1 M),最大数量的复杂事件动作标签数(>56 k),以及最大数量的长期持续轨迹(平均轨迹长度>480 s)[25],在多目标追踪、姿态估计与追踪、动作识别等领域发挥着独特的科研价值.

AVA 数据集由 Google 发布,是一个用于理解人类动作的精细标记视频数据集,样本来源于 YouTube 中公开的视频,由 80 种时空局部化原子动作标记而成. 数据集由 5.76 万个视频片段、9.6 万个标记动作行为人及 21 万个动作标签组成[26],其特点是动作标签直接与行为人相关而不与视频剪辑分类相关,原子动作标签限定在很小尺度.

1.3　人体姿态识别关键问题及解决思路

作为计算机视觉的一个重要分支,人体姿态识别经历了单人姿态识别、多人姿态识别、人体姿态跟踪及三维人体姿态识别等多个发展阶段,取得了很多实质性进展[27],但仍存在很多共性问题需要解决. 其中,对骨骼关键点的运动时间线中运动关联信息的保留和分析,对骨骼关键点之间依赖关系在运动过程中体现出的信息含义的深入剖析,这些问题对姿态识别效果及效率有重要影响. 本文借助一种骨骼双流注意力增强图卷积人体姿态识别模型,通过时域空域双流通道融合交互提取人体骨骼关键点,提取对人体姿态识别有重要影响的关节点,以关节点时间线相关性为基础建立运动信息矩阵,并在此基础上构建局部运动信息关联模型,以此解决局部关节运动信息表征问题. 同时,借助注意力增强机制关联分析双流通道的上下文信息,定义及构造关键点依赖关系和全局骨骼运动信息矩阵,以此解决全局运动信息定义问题. 进一步,通过图卷积机制关联分析双流骨骼运动信息及注意力增强运动含义信息,以邻域节点对动作的影响重要程度划分相应权值,建立双域骨骼拓扑,构建运动含义模型,以此完成人体姿态识别.

2　理论原型及模型框架

2.1　骨骼节点机理、特征提取及局部运动信息表示

人体行为动作发生时,人体关节会以局部区域为活动单元从而完成动作[28]. 骨骼数据有多种体系结构,与一般二维或三维图像的网络结构不同,其具有不规则性及复杂性的特点,是通过动作序列中每一关节的二维或三维坐标对人体骨架进行标注,对人体关节在时域及空域中的动态变化建立映射关系[29]. 通过单帧图像中相邻关节点之间的邻接关系,以及连续帧图像中相同位置关节点之间的时空域信息关联关系,可以构建骨骼节点时空依赖关系[30]. 在研究骨骼机理过程中,通常需要先定义骨骼结构图[31]. 本文在典型的骨骼结构图 Kinetics 数据集上利用 OpenPose 获取到 25 个关键节点,选择其中的 18 个作为核心关键节点,以完成骨骼构图模型的定义,如图 1 所示. 其中,定义了 25 个骨骼关节点索引关系,并对其中 18 个核心节点进行了 * 标注,如表 1 所示.

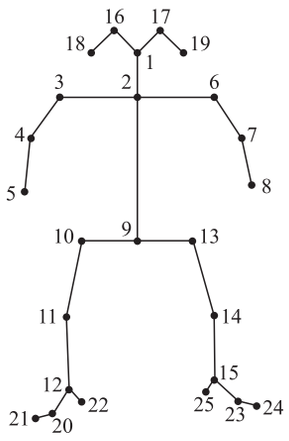


图 1　人体骨骼关键点结构图

Fig. 1　Key point structure diagram of human bone

表 1　骨骼关节点索引表

Table 1　Bone joint point index table

索引号	关节点	索引号	关节点
1*	鼻	14*	左膝
2*	脖	15*	左踝
3*	右肩	16*	右眼
4*	右肘	17*	左眼
5*	右手腕	18*	右耳
6*	左肩	19*	左耳
7*	左肘	20	右脚内
8*	左手腕	21	右脚外
9	中胯	22	右脚跟
10*	右胯	23	左脚内
11*	右膝	24	左脚外
12*	右踝	25	左脚跟
13*	左胯		

骨骼特征提取为人体姿态特征分类提供了基础数据[32]. 本文通过骨骼节点流和骨骼架构流双通道融合机制,对节点连续流和架构连续流分别进行交互分析[33],提取对姿态识别有重要影响的骨骼关键点,定义骨骼节点序列的空间运动信息矩阵,以此建立骨骼关键点之间的时间线相关性,并进一步构建局部运动信息关联模型,来对关节运动信息进行表示.

在人体姿态识别过程中,骨骼关键节点起着非常重要的作用. 对于骨骼关键节点的运动信息的表征是建立局部运动信息关联模型、进行相关卷积分类等工作的基础. 对骨骼关键点的空间结构特征进行表示,通过定义关键节点彼此对应关系及骨骼尺寸的向量关系,构建关键节点连接相关性关系. 定义关键节点坐标为(X,Y),设置可信率为Z,输入视频帧流图关键节点特征元组为$I(X,Y,Z)$,则关键节点连接相关性关系定义如表 2 所示.

在对人体行为进行描述时,关键节点序列中蕴含了非常重要的运动信息. 由于 18 个关节点中颈关节相对其他关节点更接近人体中轴线,在人体运动过程中位置变化幅度较小,特征更为明显,所以将其设定为基准位 0,定义其他关节位与基准位的向量关系为局部运动信息表示. 对整个总长为R帧的运动视频的第T帧,定义其中第m个关节点的局部运动信息为:

$$I_{Tm}^{R}=(x_{Tm},y_{Tm})-(x_{T0},y_{T0}).$$

表 2　关键节点连接相关性关系(部分)

Table 2　Part of the key nodes connection correlation relationship

关节索引连接	关节连接说明	关节索引连接	关节连接说明	关节索引连接	关节连接说明
(1,2)	鼻-脖	(2,2)	脖-脖	(3,4)	右肩-右肘
(6,7)	左肩-左肘	(9,10)	中胯-右胯	(9,13)	中胯-左胯
(10,11)	右胯-右膝	(13,14)	左胯-左膝	(11,12)	右膝-右踝
(14,15)	左膝-左踝	(1,16)	鼻-右眼	(1,17)	鼻-左眼
(16,18)	右眼-右耳	(17,19)	左眼-左耳	(2,9)	脖-中胯
(4,5)	右肘-右手腕	(7,8)	左肘-左手腕	(12,20)	右踝-右脚内
(15,23)	左踝-左脚内	(12,22)	右踝-右脚跟	(15,25)	左踝-左脚跟

2.2　双流注意力增强及全局运动信息表示

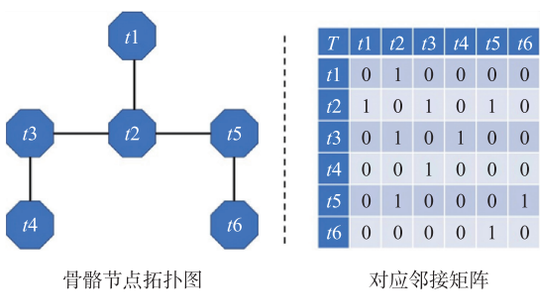
在人体姿态识别领域,也有研究将骨架流和关节流作为双流网络组合完成自适应卷积训练,卷积模块由空域卷积层、时域卷积层、标准化处理、线性修正单元激活函数以及残差块处理组成. 空域卷积层以人体所有关节点之间拓扑结构的邻接矩阵完成卷积过程,对骨骼结构空域特征进行提取. 时域卷积层对骨骼行为信息进行时域特征提取. 关节流及骨架流作为空域及时域双流网络不同的处理对象,关节流利用 OpenPose 的 25 个关键点中的 18 个作为输入单人体对象的对标节点,由此形成运动数据流的坐标信息表示;骨架流由关节流数据计算得到,即以节点间连接边缘所构成的边缘集合.

由于骨骼关节点信息与原始视频帧信息不同,骨骼关节点是由算法工具计算出来的节点坐标序列,若直接将双流注意力应用于人体骨骼姿态特征提取,对这种生成节点序列过度池化处理,会使得骨骼节点信

息噪声被成倍放大,卷积信息大幅衰减,反而大大降低了卷积处理的效果. 为改善该问题的影响,对空域及时域注意力信息进行增强处理,在处理上下文之间依赖关系的基础上,同时增强处理双域输入特征,对不同维度注意力模块进行多维特征堆叠和嵌合,以支持对骨骼姿态识别任务中多维度之间的依赖性分析,从而提升卷积效果. 增强方式包括对图像执行动态变形、动态翻转、动态拉伸、动态旋转等方式的数量扩充及同类数据训练样本多样化处理,以此训练并增强模型多元化处理能力及持续稳定性.

骨骼节点拓扑图及对应邻接矩阵示意图如图 2 表示,图中的节点表示骨骼的关节点,节点之间的连接边表示骨骼本身. 按照人体骨骼拓扑关节点数量为 J 表示(本文中 J 的实际取值为 18),骨骼拓扑图可以定义为 $T=\{t_i \mid i=1,2,\cdots,J\}$. 骨骼拓扑全局运动信息可用邻接矩阵 $\boldsymbol{M}_m$ 表示,i 和 j 分别为骨骼拓扑中任意两个关节点,若两关节点有骨骼连通则表示为 1,否则为 0:

$$\boldsymbol{M}_m[i][j]=\begin{cases}1, & i \text{ 与 } j \text{ 连通};\\ 0, & i \text{ 与 } j \text{ 不连通}.\end{cases}$$



T	$t1$	$t2$	$t3$	$t4$	$t5$	$t6$
$t1$	0	1	0	0	0	0
$t2$	1	0	1	0	1	0
$t3$	0	1	0	1	0	0
$t4$	0	0	1	0	0	0
$t5$	0	1	0	0	0	1
$t6$	0	0	0	0	1	0

图 2　骨骼节点拓扑图及对应邻接矩阵示意图

Fig. 2　Topology diagram of the bone nodes and corresponding adjacency matrix

在双流网络的基础上辅以注意力增强机制可以更好地增进卷积效果,利用邻接矩阵为基础对关节流及骨架流进行双流全局运动信息表示.

2.3　骨骼图卷积及运动含义构解

对骨骼图序列而言,某种行为是由一系列视频帧连贯而成的,任何一帧都由一组关节点坐标序列构成. 对于一个由 J 个关节点组成的 R 帧的骨骼序列时空图 $\boldsymbol{M}(P,Q)$,关节点的集合定义为 $P=\{p_{ki} \mid k=1,2,\cdots,R;i=1,2,\cdots,J\}$,帧 k 的关节点特征向量可表示为 $\boldsymbol{F}(p_{ki})$. 按照人体骨骼节点关联性,在一帧中可以将相邻关节点互相连接,在 R 帧视频构建的坐标系中,一个关节点的运动轨迹坐标构成一个坐标集,同一帧内骨骼连接集合可以定位为 $Q_a=\{p_{ki}p_{kj} \mid (i,j)\in G\}$,$G$ 为骨骼关节集合,连续帧间同一关节连接可以定位为 $Q_b=\{p_{ki}p_{t+1,i}\}$.

图卷积聚集邻域节点的权重由当前节点连接的重要程度决定. 对于图卷积模型 P_k,在第 k 帧,关节点数量为 J,帧内骨骼连接集合为 $Q_a=\{p_{ki}p_{kj} \mid (i,j)\in G\}$. 对图像进行降维并卷积,设步长为 1,卷积核为 $H\times H$,定义输入图像为 F_i,通道数为 x,则可将空域位置中的 l 点位的基于通道的图像卷积表达为:

$$F_o(l)=\sum_{m=1}^{H}\sum_{n=1}^{H}F_i(S(l,m,n)\times n(m,n)),$$

式中,$S(l,m,n)$用于获取样本,代表采样点 l 及其邻域节点之间的空域关系;$n(m,n)$用于定义权重,代表通道维度为 x 的输入向量在进行卷积时赋予的权重.

骨骼图卷积中对人体运动含义的构解本质上是对图卷积生成特征进行 Softmax 分类. 人体不同关节部分蕴含的运动信息量不同,提取同一帧不同关节点之间邻域关系及节点之间连接长度,并补充骨骼序列的时序动态特征变化关系,再计算邻域节点的权重. 构解流程步骤如下:通过骨骼预处理机制提取视频帧中全部关节特征点坐标向量,构建关键节点及骨骼长度空间结构特征矩阵,通过多帧间的关节位移形成时序动态特征矩阵;通过关节点特征及节点序号形成基础骨骼信息表,多张骨骼信息表集合成为骨骼时空域特征矩阵;将骨骼时空域特征矩阵通过平均池化至通道,经全连接层 FCN 及 Softmax 分类器,输出动作分类评分;根据分类评分加权计算,输出最佳分类结果. 骨骼图卷积及运动含义构解结构图如图 3 所示.

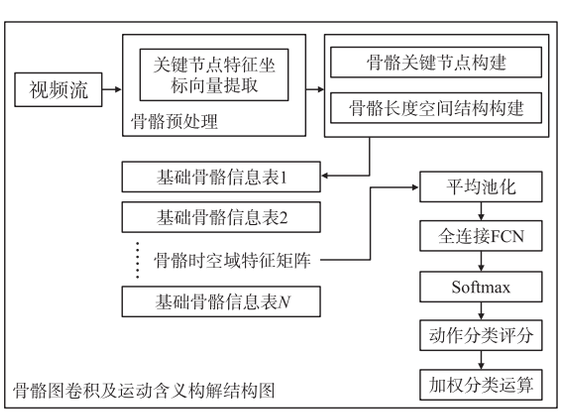


图 3　骨骼图卷积及运动含义构解结构图

Fig. 3　Bone diagram convolution and motion meaning structure diagram

2.4　基于骨骼特征点双流注意力增强图卷积人体姿态识别模型

针对人体的连续性动作,各关节点及其相邻关节点组成一系列局部运动群组,根据关节点或群组的变化特征信息来对运动进行分类,判断运动含义. 本文将人体运动视频连续帧中的骨骼特征数据作为骨骼特征点双流注意力增强图卷积人体姿态识别模型(skeleton double-flow attention enhance graph convolution, SDFAEGC)的输入信息,在模型中提取一系列的关节点特征坐标,构建骨骼关节点坐标向量及特征点距离集合空间特征矩阵、骨骼关节点坐标移动向量组和距离变化集合时序特征矩阵,从而形成原始全局骨骼信息矩阵. 在此基础上,通过双流注意力增强网络进行适应卷积训练、标准化处理及线性修正激活等处理,并将双流信息进行最大池化处理和上下文相关性关联. 进而,构建时空双域图卷积策略,在图卷积过程中构造动态骨骼时空单元卷积网络,由此对人体骨骼全局特征进行图卷积,生成更高层级的特征. 再经全连接处理和分类器预测,得到人体骨骼运动姿态行为评估结果. 基于骨骼特征点双流注意力增强图卷积人体姿态识别模型框架如图 4 所示.

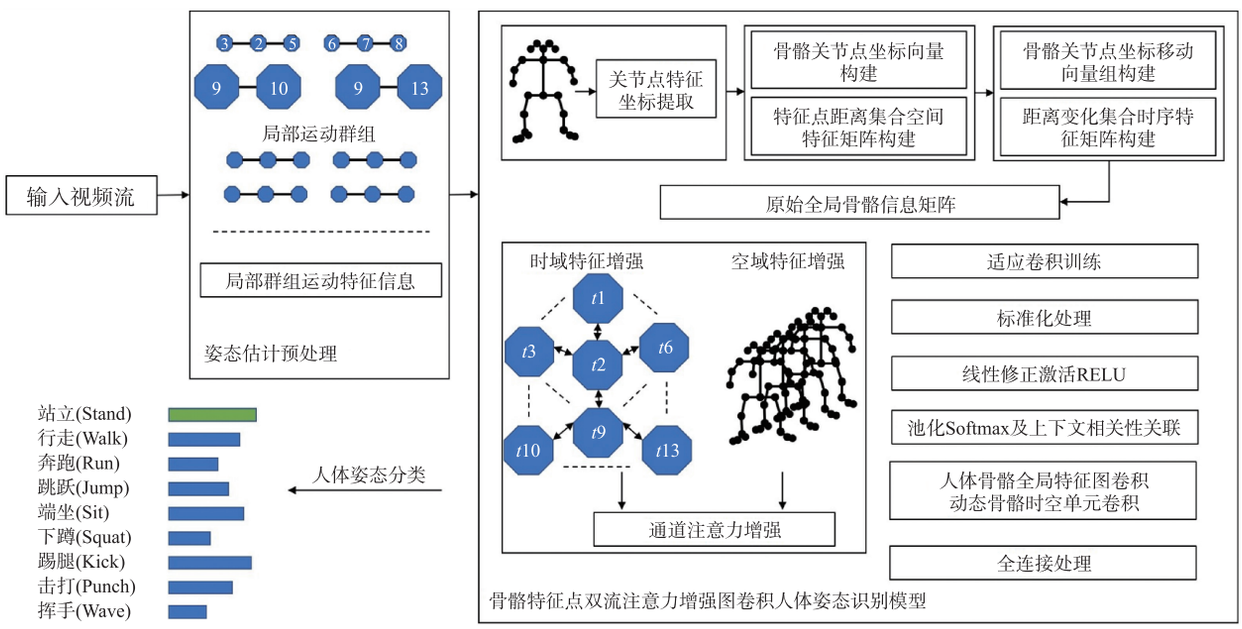


图 4　基于骨骼特征点双流注意力增强图卷积人体姿态识别模型框架

Fig. 4　Model framework of bone dual-stream attention enhancement graph convolving human posture recognition

3　实验及分析

3.1　实验环境

本项目实验环境为:操作系统为 64 位 Linux,2 * Intel Xeon Gold 5318Y CPU,128GB DDR4 SDRAM 内存,2 * NVIDIA GeForce A10 GPU 加速卡. 数据运行平台采用 Anaconda3 2022.10(64-bit)集成工具包,包括 Conda 和 Python 等 190 多个科学包及其依赖项. 代码集成开发环境采用 PyCharm2022.3. 机器学习算法运行平台采用 TensorFlow2.10. 人体姿态识别基础架构使用开源的 OpenPose 工具,在该开源工具基础上进行开发修改和定制. 因人体姿态识别易受周围环境噪声的影响,本实验采用白墙体背景模式,以达到更理想的实验效果,减少背景噪声引发的干扰.

3.2　实验数据集

本实验在 Kinetics 和 NTU RGB+D 两个有代表性的大规模数据集上对比验证了所提出的模型的有效性,并与其他相关算法进行了比较.

针对 Kinetics 数据集,考虑到要获取骨骼所有关节点位置信息,调整了视频分辨率及单帧速率,分辨率设置为 352×288ppi,单帧速率调整为 50FPS,训练周期设置为 200 帧,初始学习率为 10%,按照 10 个周期一轮迭代,每轮迭代学习率衰减 10 倍递归迭代. 针对 NTU RGB+D 数据集,采用原始分辨率和单帧速率,训练周期设

置为 300 帧,初始学习率设置为 10%,按照 30 个周期一轮迭代,每轮迭代学习率衰减 10 倍递归迭代. 为了防止过拟合现象的发生,在实验中加入优化器对每个时空单元的 Dropout 衰变值进行调参.

Kinetics 数据集包括超过 65 万个动作视频流,动作类别超过 700 个,且每一类动作至少有超过 600 个视频流,每一个视频流长度都超过 10 s. 但 Kinetics 数据集仅提供 RGB 视频流,不包含人体骨骼关节信息,需通过 OpenPose 对视频流进行批量处理,获取所有关键点,并提取实验所需的特征关键点,以键值对(JavaScript object notation,JSON)格式保存. 本实验采用 Top-1 精度指标作为评判依据,按照全部测试视频分类准确为基数得出算法分类精度.

NTU RGB+D 数据集包含超过 5 万个人体动作视频流,包括 60 个动作行为类别,每一个视频流中都含有 25 个骨骼关节点的三维坐标数据. 基准数据子集有两类,一个是 CS 基准子集,包含 40 320 个视频流的训练集及 16 560 个视频流的测试集;另一个是 CV 基准子集,包含 37 920 个视频流的训练集及 18 960 个视频流的测试集.

3.3　实验结果及分析

本实验在近距离低噪声条件下进行,为达到更理想的实验效果,进行了充足的采光及背景单一化处理. 摄像机与人体呈垂直设置,排除了角度差异带来的识别差异. 因为人体姿态连续识别分类对计算资源有较高要求,实验采用 GPU 加速和高性能算法相结合,为计算的实时性连续性提供保障. 针对本模型,实验中以下列 9 类动作作为分类标签测试模型的性能情况:站立(Stand),行走(Walk),奔跑(Run),跳跃(Jump),端坐(Sit),下蹲(Squat),踢腿(Kick),击打(Punch),挥手(Wave). 实验分别采用未加入双流注意力增强的骨骼图卷积模型及加入双流注意力增强后的骨骼图卷积模型以不同姿态的识别百分率作为测试目标,并分别在 Kinetics 数据集及 NTU RGB+D 数据集上进行测试. 为增加实验的鲁棒性及跨数据集性能测试的准确度,每种数据集的每一类动作姿态都分别进行 20 轮(25 s/轮)的实验测试,从每组测试结果中挑选出最优结果作为该数据集此类姿态的最优实验结果.

实验数据结果如表 3 所示. 根据实验结果可以得出如下结论:加入双流注意力增强的骨骼图卷积模型比未加入双流注意力增强机制前的骨骼图卷积模型(skeleton graph convolution,SGC)识别率有较大提升. 从 Kinetics 及 NTU RGB+D 两个数据集上的增强效果来看,在 NTU RGB+D 数据集上的增强效果更为明显. Kinetics 数据集不论从规模上还是从分类的数量上都比 NTU RGB+D 更胜一筹,但 Kinetics 数据集只有原始视频流,没有骨骼关键点信息,需要在实验过程中提取特征点键值;NTU RGB+D 则对骨骼三维坐标数据进行了标注,所以对识别算法的输入特征有着更好的支撑. 从本文识别模型针对 9 种不同行为姿态的识别效果来看,在 NTU RGB+D 数据集上各种姿态最高识别率均达到了 100%;在 Kinetics 上也有 3 种姿态的识别率达到了 100%,其余 6 种姿态识别率也都超过了 90%;可见本文模型在不同数据集上均有较好的执行效果. 实验效果图例和现场测试环境展示如图 5 所示.

表 3　双流注意力增强机制前后骨骼特征点图卷积人体姿态识别模型在 Kinetics 和 NTU RGB+D 数据集上的实验结果

Table 3　Experimental results of skeletal feature dot map convolved human pose recognition model before and after the dual-flow attention enhancement on datasets Kinetics and NTU RGB+D　%

姿态类别	Kinetics		NTU RGB+D	
	SGC	SDFAEGC	SGC	SDFAEGC
站立(Stand)	91.50	100.00	96.51	100.00
行走(Walk)	76.28	93.10	90.08	100.00
奔跑(Run)	89.74	100.00	90.97	100.00
跳跃(Jump)	83.47	90.19	92.77	100.00
端坐(Sit)	86.05	94.45	85.66	100.00
下蹲(Squat)	72.96	92.85	81.39	100.00
踢腿(Kick)	81.33	90.93	88.46	100.00
击打(Punch)	84.00	100.00	95.79	100.00
挥手(Wave)	89.46	96.08	91.53	100.00

图 5　实验效果图例和现场测试环境展示

Fig. 5　Experimental effect legend and test environment display

3.4　实验结果与主流方法的比较

在 NTU RGB+D 数据集基础上,将本文所提模型算法与 6 种主流方法进行比较,结果如表 4 所示.

表 4　本文方法与主流方法效果的比较分析

Table 4　Comparative analysis of the effect between mainstream methods and the method in this paper　%

方法	站立	行走	奔跑	跳跃	端坐	下蹲	踢腿	击打	挥手
方法 1	93.34	98.50	92.33	93.79	93.90	92.13	94.66	93.30	98.62
方法 2	93.49	93.50	92.69	93.30	92.65	94.45	92.19	98.33	98.82
方法 3	94.73	99.50	94.35	92.14	95.21	93.74	94.51	95.76	91.56
方法 4	95.14	94.33	95.54	94.05	96.57	95.23	95.88	98.26	98.50
方法 5	96.07	96.25	97.50	95.44	97.70	96.22	96.07	98.75	98.33
方法 6	97.11	100.00	96.25	96.86	98.02	97.93	97.99	100.00	100.00
本文方法	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

基于光流特征及密集加速鲁棒特征算法运动检测模型(方法 1)将动作和行为轨迹以及相关的视觉描述相融合,形成基于兴趣点轨迹的人体姿态识别策略[34]. 动作相关性关键姿态特征可迁移有限动作识别

模型(方法 2)通过从 RGB 视频流中获取运动特征,并以此作为 DCNN 的输入源完成识别过程[35],在这种模式下,从大规模数据集训练中得到的 CNN 表示具备了向数据集有限动作识别任务中迁移的能力. 混合 SVM 及 K 近邻分类器动作识别模型(方法 3)是由构建于预处理机制上的 DCNN 模型的特征提取及特征表示,再通过混合 SVM 及 K 近邻分类器完成动作识别[36]. 基于注意力机制的全局上下文感知骨骼关键点识别模型(方法 4)基于全局上下文的由注意力引导的长短期记忆神经网络,借助全局上下文记忆单元对视频流中每一帧当中的人体关节点进行有选择性、有侧重性的关注,在循环注意力机制的协同下对网络性能进行改善,并以此完成基于骨骼关键点的动作识别[37]. 基于目标检测及定位的融合长短期记忆网络行为识别模型(方法 5)基于 YOLO(You Only Look Once)方法进行对特定目标检测及定位,并在此基础上与长短期记忆神经网络相结合,利用神经网络进行行为识别[38]. 基于姿态估计的人体异常行为识别模型(方法 6)利用基于深度学习的人体姿态估计算法提取人体的骨骼关键点坐标,组成包含空间信息和时间序列信息的时空图模型,同时对时空图进行多阶段的时空图卷积操作以提取高级特征,最后用 Softmax 分类器进行行为分类,得到行为结果并判断是否为异常行为[39].

从实验数据可知,本文在选定的 9 种行为姿态识别精度上均优于其他模型. 究其原因,以上方法或通过光流特征行为轨迹,或通过直接从 RGB 视频流获取运动特征,或通过预处理干预并利用 SVM 及 K 近邻分类特征,这些方法与人对姿态识别过程有较大区别,只能在特定或局部情况下起到一定作用. 也有方法利用注意力机制感知全局骨骼关键点,或通过记忆网络模型进行检测定位,或利用深度学习时空双域的方式提取关键点坐标并提取高级特征进而分类,这些方法改善了识别策略,更多从全局视角及模拟人的识别过程出发,取得了更优效果. 本文方法将局部与全局方法相结合,既发挥了双流通道机制对局部运动信息的细微处理能力,又发挥了注意力机制对全局运动信息的宏观处理能力,从而获得了较好的效果.

4　结论

针对骨骼关键点分类算法中过多聚焦于输入参数而忽视运动时间线中运动关联信息的分析价值,以及骨骼节点关联性和依赖关系信息含义丢失问题,本文提出了一种骨骼双流注意力增强图卷积人体姿态识别模型. 通过以提取骨骼特征节点为基础,构建骨骼关节点之间空域连接矩阵和运动时间线时域信息矩阵,继而进行双流骨骼节点信息处理. 利用通道注意力机制对上下文处理的优势,对关键节点间依赖关系以及全局骨骼运动含义进行构解,构建邻域节点加权的双域骨骼拓扑. 将本文模型在 Kinetics 和 NTU RGB+D 两个数据集上进行验证,并与领域内具有代表性的主流方法进行对比,可以看出本模型具有较高识别精度和较强鲁棒性.

由于实验是在去噪的实验场景中完成的,在背景噪声较大或多人交互的复杂现实场景衰减较大,后续工作将以去噪和多人交互为继续研究的重点.

[参考文献](References)

[1] YANG X D, TIAN Y L. Effctive 3D action recognition using eigenjoints[J]. Journal of Visual Communication and Image Representation, 2014, 25(1): 2-11.

[2] 石跃祥,朱茂清. 基于骨架动作识别的协作卷积 Transformer 网络[J]. 电子与信息学报,2023,45(4):1485-1493.

[3] CHAUDHRY R, RAVICHANDRAN A, HAGER G, et al. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009.

[4] 姜权晏,吴小俊,徐天阳. 用于骨架行为识别的多维特征嵌合注意力机制[J]. 中国图象图形学报,2022,27(8):2391-2403.

[5] 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.

[6] PENG W, HONG X P, CHEN H Y, et al. Learning graph convolutional network for skeleton-based human action recognition by neural searching[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020.

[7] 马钰锡,谭励,董旭,等. 面向智能监控的行为识别[J]. 中国图象图形学报,2019,24(2):282-290.

[8] DU W B, WANG Y L, QIAO Y. Recurrent spatial-temporal attention network for action recognition in videos[J]. IEEE

Transactions on Image Processing,2018,27(3):1347-1360.
[9] XIA R J,LI Y S,LUO W H. LAGA-Net:Local-and-global attention network for skeleton based action recognition[J]. IEEE Transactions on Multimedia,2022,24:2648-2661.
[10] JIANG X H,XU K,SUN T F. Action recognition scheme based on skeleton representation with DS-LSTM network[J]. IEEE Transactions on Circuits and Systems for Video Technology,2020,30(7):2129-2140.
[11] SI C Y,CHEN W T,WANG W,et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,USA:IEEE,2019.
[12] 李扬志,袁家政,刘宏哲. 基于时空注意力图卷积网络模型的人体骨架动作识别算法[J]. 计算机应用,2021,41(7):1915-1921.
[13] 田志强,邓春华,张俊雯. 基于骨骼时序散度特征的人体行为识别算法[J]. 计算机应用,2021,41(5):1450-1457.
[14] SHI L,ZHANG Y F,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.
[15] CAO B,XIA H,LIU Z. A video abnormal behavior recognition algorithm based on deep learning[C]//Proceedings of the 2021 IEEE 4th Advanced Information Management,Communicates,Electronic and Automation Control Conference(IMCEC). Chongqing,China:IEEE,2021.
[16] 薛盼盼,刘云,李辉,等. 基于时域扩张残差网络和双分支结构的人体行为识别[J]. 控制与决策,2022,37(11):2993-3002.
[17] NAVEENKUMAR M,DOMNIC S. Spatio temporal joint distance maps for skeleton-based action recognition using convolutional neural networks[J]. International Journal of Image and Graphics,2021,21(5):s0219467821400015.
[18] 钱银中,沈一帆. 姿态特征与深度特征在图像动作识别中的混合应用[J]. 自动化学报,2019,45(3):626-636.
[19] YANG H Y,GU Y Z,ZHU J C,et al. PGCN-TCA:Pseudo graph convolutional network with temporal and channel-wise attention for skeleton-based action recognition[J]. IEEE Access,2020,8:10040-10047.
[20] SULTANI W,CHEN C,SHAH M,et al. Real-world anomaly detection in surveillance videos[C]//Proceedings of the 2018 IEEE/AVF Conference on Computer Vision and Pattern Recognition. Salt Lake City,USA:IEEE,2018.
[21] PENG W,SHI J G,ZHAO G Y. Spatial temporal graph deconvolutional network for skeleton-based human action recognition[J]. IEEE Signal Processing Letters,2021,28:244-248.
[22] YU W,YANG K Y,YAO H X,et al. Exploiting the complementary strengths of multi-layer CNN features for image retrieval[J]. Neurocomputing,2017,237:235-241.
[23] LIU J,SHAHROUDY A,WANG G,et al. Skeleton-based online action prediction using scale selection network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(6):1453-1467.
[24] PENG W,SHI J,VARANKA T,et al. Rethinking the ST-GCNs for 3D skeleton-based human action recognition[J]. Neurocomputing,2021,454:45-53.
[25] MIRZA A,SIDDIQI I. Recognition of cursive video text using a deep learning framework[J]. IET Image Processing,2020,14(14):3444-3455.
[26] ZHANG S Y,YANG Y,XIAO J,et al. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks[J]. IEEE Transactions on Multimedia,2018,20(9):2230-2343.
[27] CHEN C,LIU B,WAN S H,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.
[28] 张全贵,蔡丰,李志强. 基于耦合多隐马尔可夫模型和深度图像数据的人体动作识别[J]. 计算机应用,2018,38(2):454-457.
[29] LI C,XIE C Y,ZHANG B C,et al. Memory attention networks for skeleton-based action recognition[J]. IEEE Transactions on Neural Networks and Learning Systems,2022,33(9):4800-4814.
[30] 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,2020,42(10):2684-2701.
[31] 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.
[32] HUANG L J,HUANG Y,OUYANG W L,et al. Part-level graph convolutional network for skeleton-based action recognition[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York,USA:AAAI,2020.

[33] HATIRNZA E,SAH M,DIREKOGLU C A. Novel framework and concept-based semantic search Interface for abnormal crowd behaviour analysis in surveillance videos[J]. Multimedia Tools & Applications,2020,79(25/26):17579-17617.

[34] 王佳铖,鲍劲松,刘天元,等. 基于工件注意力的车间作业行为在线识别方法[J]. 计算机集成制造系统,2021,27(4):1099-1107.

[35] 苏江毅,宋晓宁,吴小俊,等. 多模态轻量级图卷积人体骨架行为识别方法[J]. 计算机科学与探索,2021,15(4):733-742.

[36] JI S W,XU W,YANG M,et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(1):221-231.

[37] 黄海新,王瑞鹏,刘孝阳. 基于 3D 卷积的人体行为识别技术综述[J]. 计算机科学,2020,47(增刊 2):139-144.

[38] ZHANG B W,WANG Y D,HOU W X,et al. Flexmatch:Boosting semi-supervised learning with curriculum pseudo labeling [C]//Proceedings of the 35th Conference on Neural Information Processing System (NeurIPS 2021). Online,Canada:NeurIPS,2021.

[39] CHEN P,GAO Y,MA A J. Multi-level attentive adversarial learning with temporal dilation for unsupervised video domain adaptation[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Waikoloa,USA:IEEE,2022.

[责任编辑:严海琳]