中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

顺序主导和方向驱动下基于点边特征的人体动作识别方法

苏本跃, 郭梦娟, 朱邦国, 盛敏

引用本文:
苏本跃,郭梦娟,朱邦国,盛敏. 顺序主导和方向驱动下基于点边特征的人体动作识别方法[J]. 控制与决策, 2024, 39(9): 3090–3098.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2023.0569

顺序主导和方向驱动下基于点边特征的人体动作识别方法

苏本跃[1,2†], 郭梦娟[1,2], 朱邦国[1,2], 盛 敏[3]

(1. 安庆师范大学 计算机与信息学院, 安徽 安庆 246133；2. 铜陵学院 数学与计算机学院，安徽 铜陵 244061；
3. 安庆师范大学 数理学院，安徽 安庆 246133)

摘 要: 人体运动是肢体运动方向、关节活动顺序以及动作幅度相互协调的过程. 然而,现有方法往往直接对原始3D骨骼关节点信息进行建模,容易忽略肢体关节活动的顺序关系、运动方向性以及动作幅度变化影响. 因此,提出一种顺序主导和方向驱动下基于点边特征的骨骼卷积神经网络,通过刻画人体关节点运动顺序、帧间距离和骨骼边方向向量等特征对人体动作分类识别. 该网络包含顺序主导单元和方向驱动单元. 顺序主导单元对骨骼边末端关节点进行建模,利用关节点的排列方式、帧间距离信息对关节活动顺序和肢体变化幅度进行表征. 方向驱动单元利用骨骼边方向向量信息表征肢体运动的方向性. 最后,将顺序主导单元与方向驱动单元进行特征融合,对人体日常行为动作进行分类识别. 实验结果表明,在两个大型数据集NTU-RGB+D60和NTU-RGB+D120上的实验结果分别较基准方法提升了2.6%、3.5%和5.9%、6.1%. 因此,所提出方法能有效利用多特征之间的协同互补性对人类日常行为运动进行深层次刻画,提高人体动作识别的精度.

关键词: 人体动作识别；骨骼数据；骨骼边方向向量；有序关节点；帧间距离；卷积神经网络

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2023.0569

引用格式: 苏本跃,郭梦娟,朱邦国,等. 顺序主导和方向驱动下基于点边特征的人体动作识别方法[J]. 控制与决策, 2024, 39(9): 3090-3098.

Sequence-driven and direction-driven human action recognition method based on joint-bone features

SU Ben-yue[1,2†], *GUO Meng-juan*[1,2], *ZHU Bang-guo*[1,2], *SHENG Min*[3]

(1. School of Computer and Information, Anqing Normal University，Anqing 246133，China；2. School of Mathematics and Computer，Tongling University，Tongling 244061，China；3. School of Mathematics and Physics，Anqing Normal University，Anqing 246133，China)

Abstract: Human action is the process of coordinating the direction of limb movement, the sequence of joint activity and the amplitude of motion. However, existing methods tend to directly model the original 3D skeletal joint information, which easily ignores the sequential relationship between limb joint activities, motion directionality and movement amplitude variation. Therefore, this paper proposes a skeletal convolutional neural network based on point-bone features in a sequence-driven and direction-driven manner to recognize human actions by characterizing the sequence of human joint point movements, inter-frame distances and skeletal bone direction vectors. The network consists of a sequence-driven unit and a direction-driven unit. The sequence-driven unit models the joint points at the end of the skeletal bone, and characterizes the sequence of joint movements and the magnitude of limb changes by using the joint arrangement and inter-frame distance information. The direction-driven unit uses the direction vector information of the skeletal bone to characterize the directionality of the limb movement. Finally, the sequence-driven unit is fused with the direction-driven unit features maps to classify and recognize human daily behavioral actions. The experimental results show that the results on two large datasets, NTU-RGB+D60 and NTU-RGB+D120, improve 2.6%, 3.5% and 5.9%, 6.1%, respectively, compared with the benchmark method. The proposed method can effectively utilize the synergistic complementarity between multiple features to deeply characterize human daily behavioral movements and effectively improve the accuracy of human action recognition.

Keywords: human action recognition；skeleton data；skeletal bone direction vector；sequential joints；distance between frames；convolutional neural network

收稿日期: 2023-04-27；**录用日期:** 2023-09-07.

基金项目: 安徽省领军人才团队项目(皖教秘人[2019]16号)；安庆师范大学与铜陵学院联合培养研究生科研创新基金项目(22tlaqsflhy2).

†通讯作者. E-mail: subenyue@sohu.com.

0 引　言

人体动作识别是计算机视觉、计算机图形学、虚拟现实等领域的重要研究课题,其目的是通过与计算机进行交互来实现对人体行为的预测[1],在人机交互、智能视频监控、运动分析和机器人等领域具有广阔的应用前景[2-4]. 在深度学习模型中,可利用多种数据对人体骨架进行动作特征提取,如: RGB视频数据、深度数据、骨骼数据等[5-6]. 与RGB视频数据和深度数据相比,骨骼数据具有简洁性、易于存储和受光照影响较小等优点[2,6-7],在人体动作识别领域得到了快速发展.

近年来,常见的处理关节点骨骼数据的深度学习方法有循环神经网络(recurrent neural network, RNN)、图卷积神经网络(graph convolutional network, GCN)和卷积神经网络(convolutional neural network, CNN)[6-8]. 其中RNN一般将骨骼数据处理为长向量的形式,主要着重于在时间维度上对人体动作建模[8]. GCN将人体骨架转化为以关节为顶点,骨骼为边的拓扑结构对人体建模[9-10]. 但基于GCN的拓扑建模通常使用固定大小的卷积核提取骨骼特征,泛化能力较弱,同时GCN模型结构较为复杂,时间成本较高[7,11]. 而CNN模型易构建,具有提取数据高级特征的能力,可以利用强大的局部卷积特征和自注意力机制更有效提取动作的时空特征[12].

在基于CNN方法的骨骼动作识别中,人体动作识别的关键在于同步挖掘骨骼序列在空间和时间域的特征信息[5,13]. 人体是关节与骨骼形成的铰链系统[2,5,8],关节点和骨骼边等特征在人体动作识别起着关键作用. 然而,现有方法往往基于原始关节点、骨骼边等特征对动作直接建模[8-9],未充分考虑人体运动的协调规律性.

人体运动是肢体运动方向、关节活动顺序和动作幅度相互协调的过程. 一方面,人体在执行每类动作时,各肢体关节具有先后次序性且运动幅度存在差异性. 例如“挥手”动作,先是肘部关节运动,而肘部关节又带动手部关节运动;同时,手部关节较肘部关节细节变化更为显著,运动幅度较大. 另一方面,人体相同关节运动方向不同会产生不同动作类别,例如“穿外套”(方向由外向内)、“脱外套”(方向由内向外),关节点序列相同,而运动方向相反. 基于此,本文考虑通过有序关节点、帧间距离、骨骼边方向向量等特征实现人体运动的深层次描述.

此外,人体在运动过程中,动作大多由身体上半身、下半身或全身关节支配完成. 例如,“剪纸”“写作”等动作仅用上半身关节信息即可识别;而“走路”“踢”等动作仅用下半身关节信息识别. 因此,在对人体结构合理划分的基础上,基于关节点、骨骼边等特征对涉及人体局部关节的动作细粒度建模,更准确、丰富地表征人体运动.

综上所述,本文提出一种顺序主导和方向驱动下基于点边特征的骨骼卷积神经网络(sequential & directional driven skeletal convolutional neural networks, SDD-CNN). 该网络包含顺序主导单元和方向驱动单元. 顺序主导单元通过骨骼边末端关节点的排列顺序、帧间距离信息在空间域建模,同时引入关节点的速度信息在时间域对动作建模,对人体关节活动顺序、肢体运动幅度进行刻画;方向驱动单元基于分区策略考虑人体上半身和下半身的骨骼边向量信息,表征肢体运动的方向性.

1 方　法

本节首先介绍动作识别的时空特征构造方式,然后分别对顺序主导单元和方向驱动单元详细描述,最后构建SDD-CNN网络并提出算法流程. SDD-CNN网络如图1所示.

1.1 基于CNN的骨骼数据伪图像表示

传统基于CNN的方法将骨骼数据表示为伪图像形式,其中伪图像的宽度、高度和通道分别表示骨骼序列的关节、时间和坐标维度. 为了更好地利用人体骨骼数据,防止噪声扰动,使得关节点、骨骼边等信息更稳定地聚集在伪图像中,本文参考以往文献[5]在卷积操作前构建了矩阵A,具体流程如图2所示.

具体地, Input $=$ $R^{C\times T\times V\times M}$, C代表通道,即关节点的xyz坐标维度;T为时间帧;V为关节维度;M为人数,在单人动作中为1,双人动作为2. 原始关节坐标信息为$v_{\text{in}}^t = [v_1^t, v_2^t, \ldots, v_N^t], t \in \{1, 2, \ldots, T\}$, N表示人体骨骼关节点个数,T表示最终帧数. 其中

$$A_{N\times N} = I_{N\times N} + \theta_{N\times N}, \tag{1}$$

$I_{N\times N}$为单位矩阵,θ为一个随机扰动矩阵,以防止元素为零.

$$v_{\text{out}}^t = v_{\text{in}}^t * A_{N\times N}. \tag{2}$$

$v_{\text{out}}^t = [\beta_1^t, \beta_1^t, \ldots, \beta_N^t]$为变换后骨骼坐标信息;$*$表示矩阵乘法.

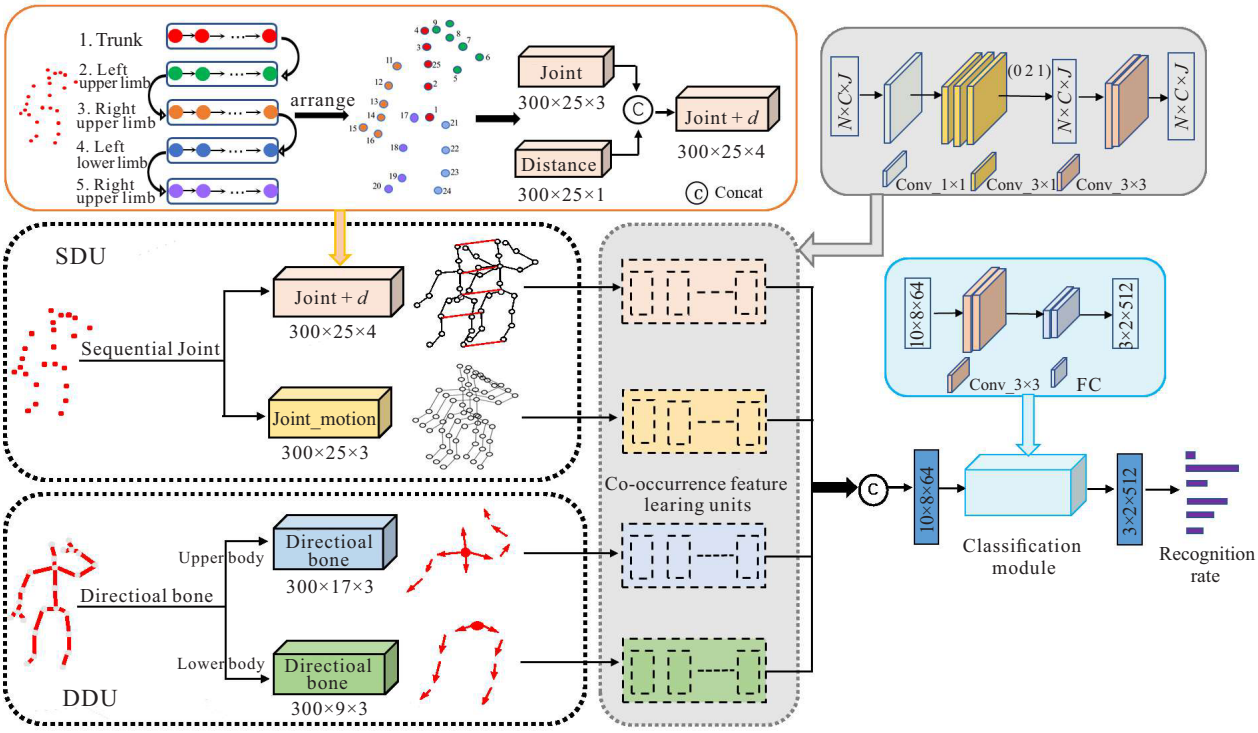


图 1 本文方法整体网络架构

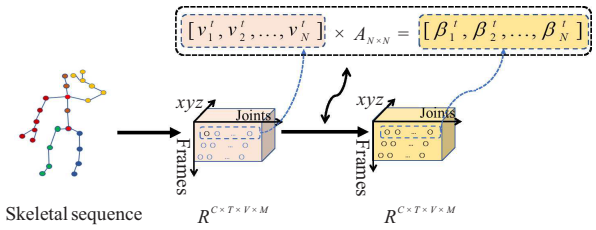
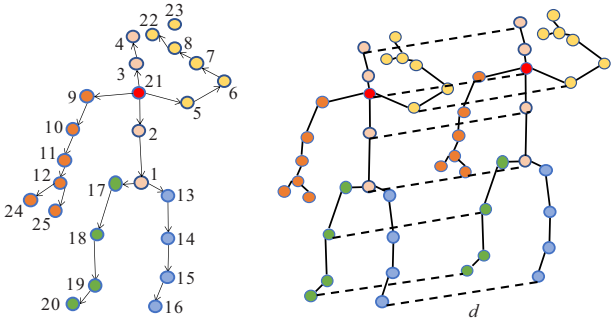


图 2 CNN伪图像示意图

1.2 时空特征构建方式

1.2.1 顺序主导单元特征构造

顺序主导单元主要利用有序关节点、帧间距离和速度等特征对人体动作进行识别分类. 通常认为, 人体在执行每类动作时, 各关节运动具有先后次序性, 而骨骼边末端关节点的排列方式可对人体关节活动顺序进行表征. 如图3所示, 以人体较为稳定的脊柱关节点21为中心点, 在中心点向外的趋势上构建骨骼边向量, 其中被指向的关节点为骨骼末端关节点, 以手肘和手腕连接形成的骨骼边向量为例, 手腕为骨骼末端关节点. 因此, 本文将人体按照躯干、左上肢、右上肢、左下肢和右下肢的排列方式依次对骨骼边末端关节点排序, 对人体关节活动顺序进行表征. 如图3(a)所示, 不同的部位在图中有不同的颜色. 排列顺序见表1. 为进一步刻画人体运动幅度变化, 如图3(b)所示, 基于有序关节点构建距离 d_i^t (d_i^t 表示相邻两帧关节点之间的距离), 对肢体运动幅度大小进行刻画; D 表示所有关节点距离信息的集合. 具体表示为



(a) 人体骨架 (b) 关节点帧间距离

图 3 人体骨骼示意图

表 1 关节点顺序构造方式

部分	骨骼边顺序	关节点顺序
躯干	(2 1) (21 2) (21 3) (3 4)	1→2→3→4
左上肢	(21 5) (5 6) (6 7) (7 8) (8 22) (8 23)	5→6→7→8→22→23
右上肢	(21 9) (9 10) (10 11) (11 12) (12 24) (12 25)	9→10→11→12→24→25
左下肢	(1 13) (13 14) (14 15) (15 16)	13→14→15→16
右下肢	(1 17) (17 18) (18 19) (19 20) (21 21)	17→18→19→20→21

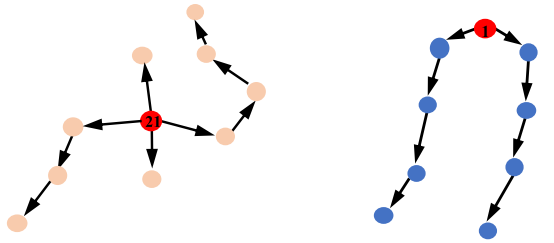
$$d_i^t=\sqrt{(x_i^{t+1}-x_i^t)^2+(y_i^{t+1}-y_i^t)^2+(z_i^{t+1}-z_i^t)^2},$$
$$i\in\{1,2,\dots,N\},t\in\{1,2,\dots,T\};\tag{3}$$
$$D=\{d_1^t,d_2^t,\dots,d_N^t\},t\in\{1,2,\dots T\}.\tag{4}$$

有序关节点和距离特征的构建充分反映了人体关节活动顺序、肢体运动幅度的变化,主要在空间维度对人体动作表征. 与此同时,关节点随时间的变化也蕴含着丰富的信息,可在时间维度推断不同的动作类别. 因此,本文基于有序关节点引入了连续两帧之间的速度信息m_i^t来表征人体运动在时间维度上的变化;其中M表示所有速度信息的集合,具体表示为

$$m_i^t=v_i^{t+1}-v_i^t,\ i\in\{1,2,\dots,N\},\ t\in\{1,2,\dots,T\};\tag{5}$$
$$M=\{m_1^t,m_2^t,\dots,m_N^t\},\ t\in\{1,2,\dots,T\}.\tag{6}$$

1.2.2　方向驱动单元特征构造

方向驱动单元主要基于分区策略通过人体上半身和下半身骨骼边方向向量对人体运动方向进行表征. 如图4所示,本文基于分区策略分别对人体上半身和下半身骨骼边方向向量进行建模. 具体来说,上半身、下半身分别以人体较为稳定的脊椎关节点21、脊柱关节点1为中心点,在中心点向外的趋势上构建有向骨骼边. 其中上半身包括人体躯干和所有手指、手臂关节,下半身包括两条腿的所有关节.



(a)　上半身骨骼边　　(b)　下半身骨骼边

图 4　人体上半身、下半身骨骼边示意图

具体地,假设v_i^t为第i个关节点第t帧的关节坐标信息,T为最终帧数,则骨骼边向量b_i^t可表示为

$$b_i^t=v_k^t-v_j^t,\ k,j\in\{1,2,\dots,N\};\tag{7}$$
$$B_U=\{b_1^t,b_2^t,\dots,b_{17}^t\};\tag{8}$$
$$B_L=\{b_1^t,b_2^t,\dots,b_9^t\},\ t\in\{1,2,\dots,T\}.\tag{9}$$

其中:b_i^t表示相邻关节点形成的骨骼边向量,v_k^t和v_j^t表示相邻关节点关节坐标信息;B_U和B_L分别表示身体上半身、下半身骨骼边向量集合.

1.3　SDD-CNN 网络

本节介绍了点边特征融合的骨骼卷积神经网络结构框架(SDD-CNN). 如图1所示,该网络主要包括顺序主导单元(sequence-driven unit, SDU),方向驱动单元(direction-driven unit, DDU),共现特征学习单元(co-occurrence feature learning unit, CFLU).

1.3.1　顺序主导单元

顺序主导单元(SDU),主要利用人体骨架的点级信息构建,利用有序关节点和关节点的帧间距离特征在人体动作空间域进行学习;同时,将关节点速度信息在时间域进行学习. 如图1(SDU 单元)所示,将关节点和距离信息以25 × 300 × 4输入作为一个独立通道输送到共现特征单元,对关节运动顺序和幅度进行刻画,从而提取出丰富的空间语义信息;将速度信息以25 × 300 × 3输入到共现特征单元,获得人体时间维度关节的依赖关系. 然后,将时空域特征通过通道维度上的特征映射进行拼接融合. 最后,采用多层卷积层和最大池化层相结合的方法,学习骨骼关节点信息的高级运动模式,对人体运动规律进行研究.

1.3.2　方向驱动单元

方向驱动单元(DDU),主要由人体骨骼边级信息构成. 首先,分别将人体上半身和下半身骨骼边方向向量作为两个单独通道输入到网络模型中. 其中上半身输入信息为17 × 300 × 3,下半身输入信息为9×300×3;然后,将上半身、下半身的骨骼边方向向量通过两个并行分支输入共现特征学习单元;最后,使用多个并行卷积块来建模骨架之间的结构关系. 再依次生成64维、256维、256维和512维特征,并将人体上半身、下半身特征进行拼接操作,实现人体行为动作的细粒度刻画.

1.3.3　共现特征单元

共现特征学习单元(CFLU),旨在基于人体骨架的点级、边级信息中学习共现特征,这有利于建模具有远程关节交互的动作,例如“穿鞋”. 如图5所示,共现特征学习单元包含6个卷积层. 其中,时间维度的卷积核大小分别为1和3,用来获取不同时间尺度的关节坐标特征,之后将第4卷积层的输出张量进行转置,将关节维数作为通道,从而在卷积运算过程中通过跨通道的逐个元素求和实现关节信息的全局聚合. 为了充分挖掘多特征之间的协同互补性,对动作实现精确识别,并将SDU和DDU单元的输出信息沿

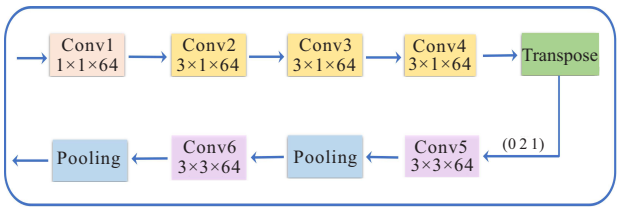


图 5　多特征卷积大小流程图

通道维度拼接起来. 然后将拼接的特征馈送到神经元大小为512线性层中,对每个任务进行监督学习.

1.3.4 算法流程

具体算法流程如下(其中θ为较小值矩阵; J, M, D, B_U, B_L分别表示关节点,关节点速度,距离信息,上半身、下半身骨骼边向量的集合; $S \in \{J, D\}$表示人体关节点、距离信息集合):

输入: 人体骨架特征信息S, M, B_U, B_L和矩阵θ;

输出: 动作识别率acc.

step 1: 初始化输入集合temp_set.

step 2: 数据预处理操作

for i in temp_set do;

$A = I + \theta$

$P = i + A$

temp_set.append(P)

end for

step 3: 对4种特征进行遍历

for epoch = 0 to 800

do in parallel

for $i = 0$ to 3 do

for $l = 0$ to 3 do

$J_1 = \text{pool}(f_l(\text{temp\_set}[i]))$

J_1.append(J_1)

end for

end for

step 4: 维度转换操作

$J_1^{N\times V\times T\times C} \Leftarrow J_1^{N\times C\times T\times V}$

End parallel

step 5: 对特征进行拼接操作

$J_2 = \text{concat}(J_1[0], J_1[1], J_1[2], J_1[3])$.

step 6: 对特征进行卷积、池化、全连接操作

$J_3 = \text{fully\_connected}(\text{pool}(\text{conv}(J_2, \text{filters}), \text{pool\_size}), \text{weights})$.

step 7: 根据loss反向传播,模型参数更新

loss = cross_entropy(real_lable, predict_lable)

end epoch

step 8: 得到预测标签

predict_lable = softmax(J_4).

step 9: 得到识别率

count = 0.

for i in to len(predict_lable)

if real_lable[i] == predict_lable[i]

count + = 1

end if

end for

acc = count/len(predict_lable).

step 10: 输出识别率

return acc

2 实验与结果

2.1 数据集及实验细节

2.1.1 数 据 集

NTU-RGB+D60[14]是一个大规模的骨骼动作识别数据集,由40个不同的受试者执行60类不同的动作. 这些动作涵盖了人体的上半身、下半身和全身运动,包含了40种日常行为动作、9种与健康相关的动作以及11种双人交互动作. 该数据集包含25个人体骨骼关节点,由微软Kinect V2摄像头采集得到,共56 880个样本. 该数据集在划分训练集和测试集时采用了跨对象(cs, cross-subject)和跨视角(cv, cross-view)两种评估标准. 其中cs评估中20个对象的40 320个视频组成训练集,其余20个对象的16 560个视频作为测试集. cv评估中通过摄像机ID2和ID3采集到的37 920个视频为训练集,通过摄像机ID1捕获的18 960个视频作为测试集. 值得注意的是,实验中删除了302个具有缺失或者不完整骨骼数据样本.

NTU-RGB+D120[15]是NTU-RGB+D60数据集的扩展版,增加了60个动作类别,一共11 3945个动作样本. 这些动作共由106名受试者完成,其中82类为日常行为动作,12类为与健康相关的动作,26类为双人相互动作. 与NTU-RGB+D60中的动作相比,NTU-RGB+D120数据集中添加的动作涉及更多难以区分的细粒度动作. 此数据集有两种评估标准,一种是交叉对象评估(cs, cross-subject),即53名受试者的样本用于训练,其余53名受测者用于测试. 另一种是交叉设置评估(c-set, cross-set),即以偶数ID的样本组成训练集,奇数ID的样本组成测试集. 同样实验中删除了532个具有缺失或不完整骨骼数据样本.

2.1.2 实验细节

所有的实验采用PyTorch框架来设计动作识别深度学习模型. 通过不同的参数设置来验证所提出的深度学习模型的有效性和鲁棒性. 本文以0.001的学习率训练所提出模型的800个epoch. Batch size为64,采用随机梯度下降法(SGD)对模型进行优化. 每100个epoch后,学习率降低10%. 在数据处理中,对于不同长度的序列,在时间帧的维度上进行双线性插值[16]. 与此同时,为了更好地实现人体动作的分类与识别,本文在基准方法[9]的基础上增加了两层卷积操作与矩阵A. 最后,使用交叉熵损失函数对所提出的

模型进行训练.

2.2　实验结果与分析

2.2.1　消融实验与分析

本文在NTU RGB+D60、NTU-RGB+D120两个数据集上对动作进行识别,结果显示在表2中. 由表2可以看出,与基准方法[9]所用特征(关节点和速度信息)相一致的情况下,本文在NTU-RGB+D60(评价指标为cs、cv)和NTU-RGB+D120(评价指标为cs、c-set)数据集下,识别率分别达到了87.5%、93.5%和77.1%、78.2%,与基准方法相比,分别提升了1.0%、2.4%和0.6%、1.6%,验证了本文模型改进的有效性. 同时,为验证本文其他特征的有效性,进一步做了消融实验分析. 其中: B_U、B_L 分别表示人体上半身、下半身的骨骼边信息; J_0、J、D、M 分别表示无序关节点、有序关节点、距离、速度信息.

表2　本文所选特征的有效性验证分析　%

Method	NTU-RGB+D60		NTU-RGB+D120	
	cs	cv	cs	c-set
J_0+M	86.7	92.6	76.5	77.7
$J+M$	87.5	93.5	77.1	78.2
$J+M+D$	88.7	94.2	79.2	81.0
$J+M+D+B_U$	89.1	94.4	81.4	82.4
$J+M+D+B_L$	88.8	94.2	80.1	82.1
$J+M+D+B_U+B_L$ (SDD-CNN)	**89.1**	**94.6**	**82.4**	**82.7**
HCN (base)[9]	86.5	91.1	76.5	76.6

为了验证本文顺序主导单元模型的有效性,对无序的关节点和距离信息进行了相应的对比实验. 由表2可知,有序关节点较无序关节点识别率在NTU-RGB+D60(评价指标为cs和cv)、NTU-RGB+D120(评价指标为cs和c-set)数据集下的识别率分别提升0.8%、0.9%和0.6%、0.5%,说明了本文所考虑的关节点顺序的合理性. 当加入距离 D 时,在NTU-RGB+D60(评价指标为cs、cv)、NTU-RGB+D120(评价指标为cs、c-set)数据集下的识别率分别提升了1.2%、0.7%和2.1%、2.8%. 这是因为距离特征信息可以有效地刻画人体运动,对人体运动幅度变化进行表征. 由此可见,顺序主导单元所选取特征对动作识别分类有显著帮助.

同样,在两大数据集上分别对人体上半身、下半身的骨骼边信息进行了消融实验分析. 可以发现,当SDU单元加入 B_U 时,识别率高于加入 B_L 特征识别率,这是因为人体的动作大部分都是由上半身肢体完成,而只有小部分动作类别是由下半身肢体关节控制. 可见,当上半身、下半身骨骼边方向向量都加入时,取得了较好的结果,分别在NTU-RGB+D60(评价指标为cs、cv)、NTU-RGB+D120(评价指标为cs、c-set)数据集下识别率提升2.6%、3.5%和5.9%、6.1%. 主要原因是因为有些动作只用局部相关关节特征就可识别,减少了无效关节对动作的干扰. 由此说明,方向驱动单元特征对人体动作识别的有效性. 因此,顺序主导单元和方向驱动单元的特征信息对模型的性能提升有显著帮助,同时在使用两单元全部信息时,模型的效果达到最佳. 由此可见,本文提出的特征信息对模型的性能提升很有帮助.

图6、图7描述了在NTU-RGB+D60数据集下,顺序主导单元添加上半身、下半身有向骨骼边信息时,动作类别精度的比较结果. 由图6实验结果可以看出,当加入上半身有向骨骼边信息后,大多数含有上半身动作或者涉及手部动作的分类识别率都有所提升. 同时,从图7可以看出,像站立、摔等下半身动作识别率与幅度变化单元特征相比识别率也略有提升,进一步验证了本文所选用特征的合理性.

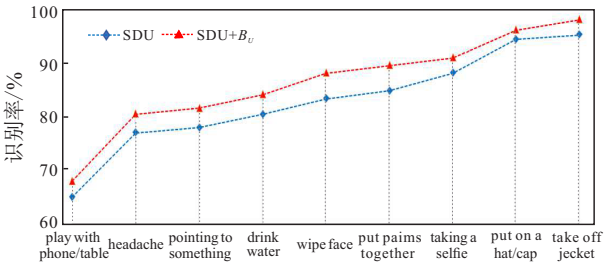


图 6　SDU与SDU+B_U 部分动作识别率对比曲线

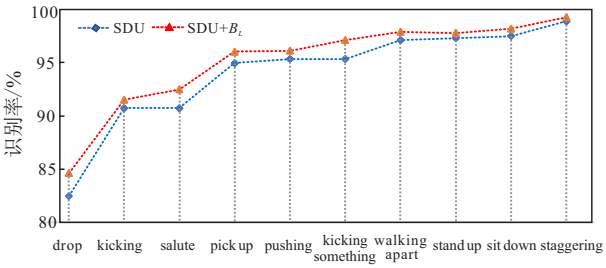


图 7　SDU与SDU+B_L 部分动作识别率对比曲线

2.2.2　模型适用性分析

为进一步地验证本文方法在不同动作类别上的适用性,本文将NTU-RGB+D60和NTU-RGB+D120数据集按照动作活动范围的不同特点划分为4类动作子集: 上半身动作子集、下半身动作子集、全身动作子集和交互动作子集. 对这些子集进行了算法性能测试,实验结果如表3所示. 由表3可以看出,下半身动作子集、全身动作子集和交互动作子集在NTU-RGB+D60、NTU-RGB+D120数据集上识别率有显著提升. 而上半身动作子集较全部动作识别率分别在NTU-RGB+D60(评价指标为cs、cv)、NTU-RGB+D120(评价指标为cs、c-set)数据集下识别率降低了

2.6%、3.5%和5.9%、6.1%. 主要原因是上半身动作特征与下半身、全身和交互动作相比动作相似性较高,例如“梳头”与“抚摸头发”等动作在手部运动和姿势上可能没有明显的差异,这增加了识别上的难度. 同时,全身动作和双人交互动作涉及到不同身体部位之间的相对位置关系,且这些动作幅度较大. 为了捕捉整体身体运动的一致性和综合特征,本文将距离特征整体度量,但同时也会失去一些上半身动作的细节信息. 因此,对于上半身等过于细粒度的动作较难区分,从而导致识别率较低.

表3 不同动作子集类别的实验结果 %

子集类别	NTU-RGB+D60		NTU-RGB+D120	
	cs	cv	cs	c-set
上半身动作	83.7	88.6	75.9	76.2
下半身动作	97.0	98.5	96.5	94.6
全身动作	92.2	95.8	89.2	88.7
交互动作	95.1	97.3	88.6	88.9

此外,全身动作、交互动作通常由连续的动作组成,每个动作的顺序具有特定的逻辑性. 通过捕捉顺序和方向,可以更好地理解动作序列的时序关系,并从中提取出更准确和丰富的运动信息. 例如,击球运动中,先挥动球拍再击打球的顺序是固定的,识别算法可以通过捕捉这种时序关系来提高准确性. 而与全身动作和交互动作相比,上半身动作(如“打哈欠”“嗅闻”“握拳”等)在空间上没有明显的顺序、方向上的变化,因此识别率较低.

2.3 与其他方法对比

在本节中,为了验证本文模型的鲁棒性和优越性,在NTU-RGB+D60、 NTU-RGB+D120数据集上与其他方法进行了比较. 本文的对比方法包括基于RNN的方法、 基于GCN的方法和基于CNN的方法. 在这3类方法中,本文模型属于基于CNN的方法.

为了验证本文方法的性能优势,在NTU-RGB+D60数据集上与当前主流的方法加以对比,结果如表4所示. 与基准方法[9]相比,本文模型分别在cs和cv评估设置上高出了2.6和3.5个百分点. 这主要是因为相较基准方法,本文选用的不同特征能够获得不同信息之间的协同互补性,从而能够对人体动作深层次表征. 与2023年基于CNN的DG-2sCNN[6]模型相比,本文模型分别优于其1.9%(cs)、3.4%(cv),这是因为本文对关节点排列方式和距离信息进行了刻画,对人体关节运动顺序和幅度信息进行了表征. 同时,也表明了本文模型改进的鲁棒性和不同特征信息选用的有效性. 通常情况下,基于GCN的方法性能优于基于CNN的方法,基于CNN的方法优于基于RNN的方法. 即便如此,本文模型仍旧取得了比大多数主流基于GCN的方法更好的性能,在cv指标下略逊于ST-TR[23]方法,ST-TR中的Transformer注意力机制具有非常出色的全局特性和模态融合能力,但其缺点是计算效率低下,成本消耗较大.

表4 本文方法与其他方法在NTU-RGB+D60数据集上的结果比较

Category	years	Methods	cs/%	cv/%
RNN-based	Two-stream RNN [17]	2017	71.3	79.5
	TCN[18]	2017	74.3	83.1
	ARRN-LSTM [19]	2019	80.7	88.8
	BGC-LSTM [20]	2019	81.8	89.0
GCN-based	ST-GCN[8]	2018	74.9	86.3
	AS-GCN [21]	2019	86.8	94.2
	PGCN-TCA [22]	2020	88.0	93.6
	ST-TR (joint) [23]	2021	88.7	95.6
	MST-GCN [24]	2022	89.0	95.1
CNN-based	Frame interpolation [25]	2020	88.9	94.6
	View-domain [26]	2021	88.0	93.3
	AFE-CNN [7]	2022	86.2	92.2
	DG-2sCNN [6]	2023	87.1	91.2
	HCN (base) [9]	2018	86.5	91.1
	本文方法(SDD-CNN)		**89.1**	**94.6**

综上所述,本文方法在NTU-RGB+D60数据集上有较好的实验效果,表明本文所定义的不同运动特征能够更深层次地理解人体动作,同时也表明了本文模型在基于CNN方法中的优越性.

表5所示为本文方法在NTU-RGB+D120数据集上与主流方法的对比, Two-Stream Attention LSTM[28]是RNN方法中比较经典的算法,本文方法在评价指标cs、c-set下分别提升了21.5、19.4个百分点;在GCN

体系结构下,本文方法相较于经典模型ST-GCN[8]在两种评估设置上分别高出11.7和9.5个百分点. 同时与ST-TR[23]模型相比,本文在cs评估标准上提升了0.5个百分点,但在c-set指标下略有下降,这说明本文模型对不同对象之间的识别更有优势. 与MST-GCN[24]相比,虽然本文方法准确度略低于该模型,但GCN模型受限于人体关节图结构,而CNN模型不受此结构影响. 对于基于CNN的方法,本文方法较基准方法[9]在评价指标cs和c-set下,精度分别提升了5.9%和6.1%. 与最新的方法DG-2SCNN[6]相比,本文方法性能也取得了显著的优势,这是因为本文所考虑不同特征的协同互补性可对人体运动的协调性规律进行深层次刻画. 由此可见,与最新的方法相比,本文模型在性能方面有显著优势.

表5　本文方法与其他方法在NTU-RGB+D120数据集上的结果比较

Category	Methods	years	cs/%	cv/%
RNN-based	Spatial-Temporal LSTM [27]	2016	55.7	57.9
	Two-Stream Attention LSTM [28]	2017	61.2	63.3
	GCA-LSTM[29]	2017	58.3	59.2
GCN-based	ST-GCN[8]	2018	70.7	73.2
	AS-GCN [22]	2019	77.7	78.9
	ST-TR (joint) [23]	2021	81.9	84.1
	MST-GCN [24]	2022	82.8	84.5
CNN-based	FSNet[30]	2021	59.9	62.4
	AFE-CNN [7]	2022	80.4	81.6
	DG-2SCNN [6]	2023	78.0	81.0
	HCN (base) [9]	2018	76.5	76.6
	本文方法(SDD-CNN)		**82.4**	**82.7**

3　结　论

本文提出了一种方向驱动和顺序主导下基于点边特征的骨骼卷积神经网络人体动作识别方法. 通过有序关节点、骨骼边向量、帧间距离等特征对人类日常行为动作进行分类识别. 实验结果表明,本文提出的方法能够挖掘多特征之间的协同互补性,有效提高人类日常动作识别的精度. 在未来工作中,将继续深入研究不同的动作类别,例如残疾人如何进食和书写等动作,并通过合理的特征描述进行分类识别. 同时,还将进一步考虑不同的距离度量方式,并考虑如何度量不同身体部位之间的距离等特征,深层次地表征人体行为动作.

参考文献(References)

[1] 南静, 宁传峰, 建中华, 等. 基于随机配置网络的轻量级人体行为识别模型[J]. 控制与决策, 2023, 38(6): 1541-1550.
(Nan J, Ning C F, Jian Z H, et al. A lightweight model for human activity recognition using stochastic configuration networks[J]. Control and Decision, 2023, 38(6): 1541-1550.)

[2] Yan G L, Hua M, Zhong Z C. Multi-derivative physical and geometric convolutional embedding networks for skeleton-based action recognition[J]. Computer Aided Geometric Design, 2021, 86: 101964.

[3] 苏本跃, 倪钰, 盛敏, 等. 基于改进卷积神经网络的动力下肢假肢运动意图识别[J]. 控制与决策, 2021, 36(12): 3031-3038.
(Su B Y, Ni Y, Sheng M, et al. Intent recognition of power lower-limb prosthesis based on improved convolutional neural network[J]. Control and Decision, 2021, 36(12): 3031-3038.)

[4] 盛敏, 刘双庆, 王婕, 等. 基于改进模板匹配的智能下肢假肢运动意图实时识别[J]. 控制与决策, 2020, 35(9): 2153-2161.
(Sheng M, Liu S Q, Wang J, et al. Real-time motion intent recognition of intelligent lower limb prosthesis based on improved template matching technique[J]. Control and Decision, 2020, 35(9): 2153-2161.)

[5] Huang H E, Su H, Chang Z G, et al. Convolutional neural network with adaptive inferential framework for skeleton-based action recognition[J]. Journal of Visual Communication and Image Representation, 2020, 73: 102925.

[6] Su B Y, Zhang P, Sun M Z, et al. Direction-guided two-stream convolutional neural networks for skeleton-based action recognition[J]. Soft Computing, 2023, 27(16): 11833-11842.

[7] Guan S N, Lu H Y, Zhu L C, et al. AFE-CNN: 3D skeleton-based action recognition with action feature enhancement[J]. Neurocomputing, 2022, 514: 256-267.

[8] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 7444-7452.

[9] Li C, Zhong Q Y, Xie D, et al. Co-occurrence feature learning from skeleton data for action recognition and

detection with hierarchical aggregation[C]. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, 2018: 786-792.

[10] 张冰冰, 葛疏雨, 王旗龙, 等. 基于多阶信息融合的行为识别方法研究[J]. 自动化学报, 2021, 47(3): 609-619.
(Zhang B B, Ge S Y, Wang Q L, et al. Multi-order information fusion method for human action recognition[J]. Acta Automatica Sinica, 2021, 47(3): 609-619.)

[11] 苏本跃, 孙满贞, 马庆, 等. 单视角下基于投影子空间视图的动作识别方法[J]. 系统仿真学报, 2023, 35(5): 1098-1108.
(Su B Y, Sun M Z, Ma Q, et al. Action recognition method based on projection subspace views under single viewing angle[J]. Journal of System Simulation, 2023, 35(5): 1098-1108.)

[12] Liu S H, Bai X Y, Fang M, et al. Mixed graph convolution and residual transformation network for skeleton-based action recognition[J]. Applied Intelligence, 2022, 52(2): 1544-1555.

[13] 盛敏, 夏安琦, 王可林, 等. 基于几何与物理特征融合的智能下肢假肢运动意图识别[J]. 控制与决策, 2022, 37(4): 953-961.
(Sheng M, Xia A Q, Wang K L, et al. Movement intention recognition of intelligent lower limb prosthesis based on the fusion of geometric and physical features[J]. Control and Decision, 2022, 37(4): 953-961.)

[14] Shahroudy A, Liu J, Ng T T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis[J/OL]. 2016, arXiv: 1604.02808.

[15] 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.

[16] 陆昱翔, 徐冠华, 唐波. 基于视觉Transformer时空自注意力的工人行为识别[J]. 浙江大学学报: 工学版, 2023(3): 446-454.
(Lu Y X, Xu G H, Tang B. Worker behavior recognition based on temporal and spatial self-attention of vision Transformer[J]. Journal of Zhejiang University: Engineering Science, 2023(3): 446-454.)

[17] Wang H S, Wang L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, 2017: 499-508.

[18] Kim T S, Reiter A. Interpretable 3D human action analysis with temporal convolutional networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, 2017: 20-28.

[19] Zheng W, Li L, Zhang Z X, et al. Relational network for skeleton-based action recognition[C]. 2019 IEEE International Conference on Multimedia and Expo (ICME). Shanghai, 2019: 826-831.

[20] Zhao R, Wang K, Su H, et al. Bayesian graph convolution LSTM for skeleton based action recognition[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 6882-6892.

[21] Li M, Chen S, Chen X, et al. Actional-structural graphconvolutional networks for skeleton-based actionrecognition[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 3595-3603.

[22] 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.

[23] Plizzari C, Cannici M, Matteucci M. Skeleton-based action recognition via spatial and temporal transformer networks[J]. Computer Vision and Image Understanding, 2021, 208/209: 103219.

[24] Chen Z, Li S C, Yang B, et al. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1113-1122.

[25] Jiang Y G, Xu J, Zhang T. View-independent representation with frame interpolation method for skeleton-based human action recognition[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(12): 2625-2636.

[26] Gedamu K, Ji Y L, Yang Y, et al. Arbitrary-view human action recognition via novel-view action generation[J]. Pattern Recognition, 2021, 118: 108043.

[27] 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, 2018, 27(4): 1586-1599.

[28] Liu J, Wang G, Hu P, et al. Global context-aware attention LSTM networks for 3D action recognition[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, 2017: 1647-1656.

[29] 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. Cham: Springer, 2016: 816-833.

[30] Chen H, Jiang Y F, Ko H. Action recognition with domain invariant features of skeleton image[C]. 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Washington DC, 2021: 1-7.

作者简介

苏本跃(1971—), 男, 教授, 博士, 从事模式识别与机器学习、图形图像处理等研究, E-mail: subenyue@sohu.com;

郭梦娟(1997—), 女, 硕士生, 从事模式识别、深度学习等研究, E-mail: 1914407015@qq.com;

朱邦国(1997—), 男, 硕士生, 从事模式识别、深度学习等研究, E-mail: 2461611988@qq.com;

盛敏(1975—), 女, 教授, 博士, 从事模式识别、图像及视频处理等研究, E-mail: msheng0125@aliyun.com.