

基于双流网络与双重注意力机制的人体行为识别

王共杰¹, 李成龙¹, 邢胜南², 高新燕²

(1. 山东建筑大学 计算机学院, 山东 济南 250101;

2. 山东华云三维科技有限公司, 山东 济南 250000)

摘要: 人体行为识别的难点在于目标行为仅占视频场景的一小部分, 易受复杂姿态和背景信息影响, 且现有的人体行为识别方法在处理全局特征提取和长距离依赖关系上有待提升。为此, 提出一种结合双流网络与双重注意力机制, 并融合大选择核卷积的网络模型。首先, 该模型利用双流网络分别对静态的骨骼信息特征与动态的时序特征进行点级特征提取和初步的全局特征提取; 其次, 利用大选择核卷积与双重注意力网络捕获长距离依赖关系, 进行全局特征提取。在 NTU RGB+D、SBU Kinect Interaction 数据集的实验表明, 所提算法通过双重注意力机制与大选择核卷积有效提升了处理长距离依赖关系的能力, 人体行为识别任务的准确率提升较大, 特别在涉及远程联合互动动作时的识别准确率更优。

关键词: 人体行为识别; 双流网络; 大选择核卷积; 双重注意力网络; 骨骼信息

DOI: 10.11907/rjdk.241491

开放科学(资源服务)标识码(OSID):



中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2025)007-0152-09

Human Action Recognition Based on Two-stream Network and Double Attention Mechanism

WANG Gongjie¹, LI Chenglong¹, XING Shengnan², GAO Xinyan²

(1. College of Computer Science and Technology, Shandong Jianzhu University, Ji'nan 250101, China;

2. Shandong Huayun 3D Technology Co., Ltd., Ji'nan 250000, China)

Abstract: Difficulty of human behavior recognition lies in the fact that the target behavior only occupies a small part of the video scene and is easily influenced by complex poses and background information. Moreover, existing human behavior recognition methods need to improve in handling global feature extraction and long-distance dependency relationships. To this end, a network model combining dual stream network and dual attention mechanism, and integrating large selection kernel convolution, is proposed. Firstly, the model utilizes a dual stream network to perform point level feature extraction and preliminary global feature extraction on both static skeletal information features and dynamic temporal features; Secondly, utilizing large selection kernel convolution and dual attention network to capture long-range dependencies for global feature extraction. Experiments on the NTU RGB+D and SBU Kinect Interaction datasets show that the proposed algorithm effectively improves its ability to handle long-distance dependencies through a dual attention mechanism and large selection kernel convolution. The accuracy of human behavior recognition tasks is significantly improved, especially when it comes to remote collaborative interactive actions.

Key Words: human action recognition; two-stream network; large selective kernel convolution; double attention networks; skeletal information

收稿日期: 2024-05-29

扫描二维码阅读全文:



基金项目: 国家自然科学基金项目(62102235); 山东省自然科学基金项目(ZR2020QF029); 山东建筑大学博士基金项目(XNBS1811)

作者简介: 王共杰(1998-), 男, 山东建筑大学计算机学院硕士研究生, 研究方向为计算机视觉、人体行为识别; 李成龙(1988-), 男, 博士, CCF 会员, 山东建筑大学计算机学院副教授、硕士生导师, 研究方向为计算机视觉、增强现实、计算机图形学等; 邢胜南(1986-), 女, 硕士, 山东华云三维科技有限公司高级工程师, 研究方向为计算机辅助设计、计算机视觉等; 高新燕(1987-), 女, 硕士, 山东华云三维科技有限公司高级工程师, 研究方向为计算机辅助设计、计算机视觉等。本文通讯作者: 李成龙。

0 引言

随着深度学习发展,基于骨骼信息的人体行为识别受到越来越多关注。基于骨骼信息的行为识别是一种利用深度学习技术,通过分析人体骨骼信息来识别人体行为的方法。该技术在智能监控、健康管理、虚拟现实等领域得到广泛应用,核心思想是利用深度学习算法从传感器采集到的人体骨骼数据中提取特征,并将这些特征作为输入,通过分类算法区分、识别不同的人体行为。

目前,基于骨骼信息的行为识别方法包括基于循环神经网络(Recurrent Neural Network, RNN)、基于卷积神经网络(Convolutional Neural Network, CNN)和基于图卷积神经网络(Graph Convolutional Neural Network, GCN)等3种主流方向^[1]。面对基于RNN的方法,骨骼序列是关节坐标位置的自然时间序列,可将其看作序列向量,而RNN本身由于其独特的结构,适合处理时间序列数据。长短期记忆网络(Long Short-Term Memory, LSTM)被设计为长期时间依赖性建模,但由于时间建模是在原始输入空间上完成的,因此难以直接从骨骼中学习高级特征^[2]。通过CNN处理基于骨骼的行为识别任务可视为是基于RNN技术的补充,因为基于RNN的方法缺乏空间信息,而CNN架构更注重输入数据中的空间线索,因此CNN成为识别基于骨骼行为的有效方法。

虽然,CNN模型具有出色的高层次信息提取能力,能从骨骼中学习时空特征,但在处理长距离依赖关系上不够完善,特征提取质量较差^[3,4]。为此,本文设计了一种结合双流网络与双重注意力机制,融合大选择核卷积的网络模型,以更好地从所有关节中获取全局响应。首先,通过双流网络从骨骼序列中自动学习分层共现特征;其次,将骨骼序列与骨骼运动序列分别输入双流网络的两个分支,利用较小的卷积核进行点级特征学习,利用较大的卷积核进行初步的全局特征学习;最后,融合双重注意力与大选择核网络,使模型在全局特征学习阶段动态调整特征以提取主干的感受野,更有效地处理、识别广泛变化的背景,更好地捕捉长距离依赖关系,解决行为识别任务中处理长距离依赖关系较差的问题。

1 相关工作

目前,基于骨骼信息的行为识别方法主要包括RNN、CNN、GCN三大类^[5,6]。为了提升人体行为识别任务中的特征提取质量、处理长距离依赖关系的能力,在模型中融入注意力机制及一些其他的卷积网络。

视频中提取出的骨骼信息包括时序信息,而RNN和LSTM具有时序记忆功能,在处理三维骨架关节等具有长期依赖性的时序数据时具有较大优势^[7]。最初,首选循环

神经网络进行识别,例如Wang等^[8]借鉴双流网络思想,在时间域上通过时间RNN模块学习行为中的运动特征,在空间域上采用基于邻域关系的遍历方法,根据依赖关系的强弱将关节依次输入空间RNN模块中来学习空间特征,但仅能处理时序信息,难以从骨骼序列中学习高级信息。Li等^[9]提出一个自适应的RNN树的模型,有效解决了单一网络难以处理细粒度行为类别的难题。Wei等^[10]通过一个门控循环单元(Gate Recurrent Unit, GRU)表示一个关节,提出高阶联合相对运动特征表示法描述关节与其子节点间的相对运动关系,但都是从原始的人体骨骼中学习运动表示,只通过连接每个帧来建模时域信息而忽略了空间信息。

近些年,基于CNN学习骨骼特征的方法层出不穷,文献[11,12]直接将骨骼序列中的坐标转换为经过变换的伪图像,以生成形状为 $K \times T$ 的二维输入(K 为关节数, T 为时间长度),但该输入无法利用卷积网络的局部性,因此在流行的基准测试中效果较差。如果只考虑卷积核内的相邻关节来学习共现特征,虽然在后续的卷积层中感受野覆盖了骨骼所有关节,但很难有效地从所有关节中挖掘共现现象。考虑到动作完成依赖于身体各个部位,宋震等^[13]将骨骼关节进行分区建模后进行全局特征融合,较好地学习了人体骨骼的局部特征,但忽略了某些关节之间的潜在信息。Yang等^[14]通过深度优先树的遍历顺序重新设计了骨架表示,增强了骨架图像语义,更好地保留了结构信息。

然而,上述模型都根据空间位置特征进行隐式学习,而对时间进行显式建模能得到更丰富的动作描述,以平衡空间与时间的信息表达。为此,Simonyan等^[15]提出双流CNN,通过提取一个静态图像流与另一个光流来融合物体外观和短期的运动信息。Li等^[16]提出一个能全局共现的分层特征学习网络,借助双流网络学习所有关节点进行全局共现特征,并将该网络拓展到更多人数,但对长距离依赖关系及特征提取质量的处理仍不够完善。

随着深度学习发展,注意力机制被引入深度学习模型,以解决输入序列长度不同、上下文关系复杂等问题。Luong等^[17,18]提出Global Attention与Local Attention两种注意力机制,丰富了注意力机制的应用场景。在计算机视觉领域,注意力机制也被引入图像分类、目标检测、图像生成等任务中,并取得了很好的效果。Si等^[19]提出一种注意力增强图卷积LSTM网络(AGC-LSTM),从骨骼数据中识别人类动作,不仅能捕获空间构型和时间动态的判别特征,还能探索空间与时间域之间的共现关系。Chen等^[20]提出一种具有注意力机制的时空可变形ConvNet模块,考虑了时间域与空间域之间的相互关联,以有效捕捉视频动作中的长距离依赖关系。该方法将基于注意力的可变形模块作为3D ConvNets的通用模块,以自适应地学习更准确的时空偏移来建模动作的不规则性。Peng等^[21]利用双重注意力块从输入图像或视频的整个时空空间中聚合、传播信

息中丰富的全局特征,使后续的卷积层能有效访问整个空间特征。

近年来,大核卷积神经网络得到了广泛的研究关注。Peng等^[21]提出一种新的全局卷积网络(global convolutional network, GCN)解决语义分割任务中的分类与定位问题。GCN采用大核卷积来增强特征图和每个像素分类器之间的稠密连接,从而提升模型处理不同图像相互变换的能力,还介绍了一种基于残差结构的边界细化模块以进一步提升目标边界的定位精度。同时,Vision Transformer (ViT)、Swin Transformer、PVT等方法在图像识别任务中也表现出良好的性能,在计算机视觉领域逐步得到广泛适用^[22-31]。Yan等^[32-34]研究发现,大的感受野是技术表现优异的关键,具有大感受野的卷积网络可与基于transformer的模型竞争。例如,ConvNeXt在其骨干中使用7×7深度卷积显著提升了下游任务的性能;RepLKNet甚至通过重新参数化,使用31×31卷积核实现了优异的性能^[35,36]。随后,SLaK通过内核分解与稀疏组技术,将内核大小进一步扩展到51×51^[37]。VAN在卷积注意力机制中引入了大核分解^[38]。Liu等^[39]提出一种大选择核卷积网络(LSKNet),可动态调整特征提取主干的感受野,提升了模型在不同场景中检测对象的效果。

2 本文方法

本文提出了一种继承了HCN优势的网络架构——A²TLS-HCN。该模型融合了大选择核卷积网络与双重注意力,有效提升了模型捕捉长距离依赖关系的能力和人体行为识别准确率。

2.1 分层共现网络HCN

HCN是人体行为识别任务中一个经典的网络架构,是一个端到端卷积共现的特征学习框架,采用分层方法逐渐聚集不同层次的上下文信息^[16]。首先,对每个关节的点级信息进行独立编码;其次,将编码组合成空间与时间域的语义表示。在HCN模型中,将联合维放入CNN的输入通道中,以在全局范围内聚合共现特征,并引入骨骼运动的表示将其显式地输入网络中。坐标系 t 中的人体骨骼可表述为 $S^t = \{J_1^t, J_2^t, \dots, J_N^t\}$, N 为关节个数, $J = (x, y, z)$ 为三维关节坐标。为此,本文将骨骼运动定义为连续两帧之间每个关节的时间差:

$$M^t = S^{t+1} - S^t = \{J_1^{t+1} - J_1^t, J_2^{t+1} - J_2^t, \dots, J_N^{t+1} - J_N^t\} \quad (1)$$

原始骨骼坐标 S 与骨骼运动 M 以双流模式独立输入网络,为了融合来自两个来源的信息,在网络后续层中跨通道连接特征映射。同时,HCN还引入了一种全局空间聚合方案,能学习到相较于局部聚合更优越的联合共现特征,虽然在一定程度上关注了全局特征,但提取的特征质量较差,对长距离依赖关系的处理也较差。

2.2 大选择核卷积网络与双重注意力机制

2.2.1 大选择核卷积网络

为了解决HCN模型存在的问题,本文在分层特征学习网络中引入大选择核卷积,如图1所示。为了提升网络识别目标中最相关的空间上下文区域,使用空间选择机制从不同尺度的大型卷积核中选择特征图。首先,将不同核函数得到的不同感受野范围的特征进行拼接;其次,拼接特征通过基于通道的平均和最大池化层来有效提取空间关系;再次,对每个空间注意力图使用Sigmoid激活函数获得每个分解的大内核单独空间选择掩码;最后,利用空间选择掩码加权大内核序列的输入特征,并通过卷积层融合得到注意力特征,最终输出结果为输入特征与注意力特征的逐元素乘积。

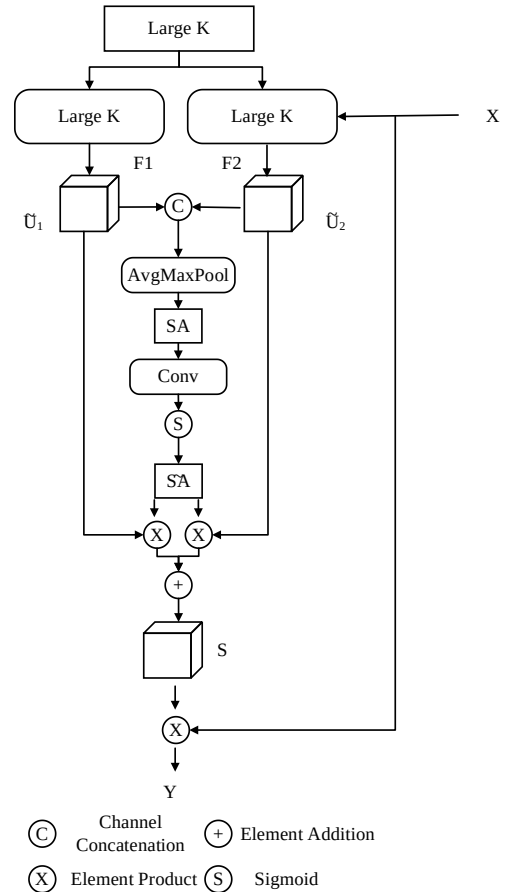


Fig. 1 Structure of large selective kernel convolutional network model

图1 大选择核卷积网络模型结构

2.2.2 双重注意力网络

首先,模型通过空间选择机制对一系列大深度核处理的特征进行加权;其次,在空间上将其合并。核的权重根据输入动态确定,允许模型自适应地使用不同的核,并根据需要调整空间中每个目标的感受野。其中,核的大小和增加扩张率保证了感受野足够快的扩张,并对膨胀率设置了一个上限,以保证膨胀卷积不会在特征图之间引入间隙。该设计的优点在于:①显式地产生具有各种大感受野的多个特征,使后续核的选择更容易;②顺序分解相较于

简单地应用单个较大的内核更有效,在相同的理论感受野下,顺序分解大卷积核减少了参数量;③引入大选择核卷积网络能使模型在人体行为识别任务中更好地处理长距离依赖关系。

传统神经网络结构在处理长距离依赖关系时表现不佳,现有 CNN 模型通过增加深度来建模这种依赖关系,但非常低效。研究发现,引入注意力机制可帮助模型更好地捕捉长距离依赖关系,例如鼓掌、穿鞋等动作,解决了 HCN 对长距离依赖关系处理不足的问题。本文方法引入了双重注意力块,从输入信息的整个时空空间中聚合与传播信息丰富的全局特征,使后续卷积层能有效访问整个空间的特征。双重注意力机制操作为:①通过双线性池化得到两个输入特征图 A 、 B 内的所有特征向量对 (a_i, b_i) 的外积二阶特征集合,双线性池化相较于传统的平均池化和最大池化而言只计算一阶统计量,可更好地捕捉和保持复杂的关系;②引入双线性池化的输出,并将第二个特征 B 重写为 $B = [\bar{b}_1; \dots; \bar{b}_n]$, $\bar{b}_i$ 为一个 dhw 维的行向量, d 为时间维度, h, w 为输入帧的空间维度,当 $\bar{b}_i$ 在所有位置上都密集参与时捕获全局特征,当 $\bar{b}_i$ 在特定区域上稀疏参与时捕获特定语义;③对注意力图进行 softmax 操作,将不同位置的视觉基元收集到一个全局描述集合中;④通过另一个注意力自适应地选择、分配特征到每个位置,根据每个位置特征的需要来自适应的分布视觉基元集合,每个位置可选择与当前特征互补的特征,使模型训练更容易并有助于捕获更复杂的关系。

特征选择步骤与生成注意力图的方式类似,注意力权重向量集也由卷积层与 softmax 归一化生成,计算流程图 2 所示。首先,对输入特征阵列 X 进行 3 个不同卷积的操作生成特征矩阵 A 、 B 和 V ,并进行 softmax 归一化;其次,通过两个矩阵乘法及必要的转置操作得出输出结果 Z ;最后,

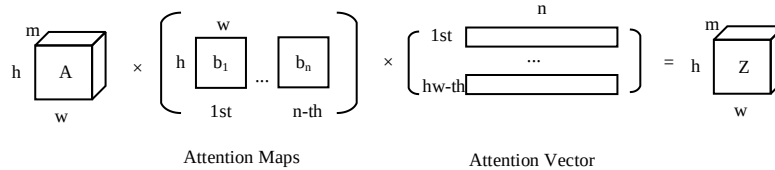


Fig. 3 Double attention operation process

图 3 双重注意操作流程

通过张量 $T \times N \times V$ 表示输入的骨骼序列, T 为序列中的帧数, N 为批次大小, V 为坐标维数(例如三维骨骼为 3)。将其作为两种输入流直接输入网络,两个网络分支共享相同的体系结构,具体参数并不单独共享和学习,特征映射在 conv4 后通过沿通道拼接进行融合。

阶段一:通过 1×1 (Conv1) 和 3×1 (Conv2) 卷积层对点级特征进行编码。

$$X_{x_i} = \text{Conv1}(x_i) \quad (2)$$

$$X'_{x_i} = \text{Conv2}(X_{x_i}) \quad (3)$$

式中: x_i 代表输入的特征。

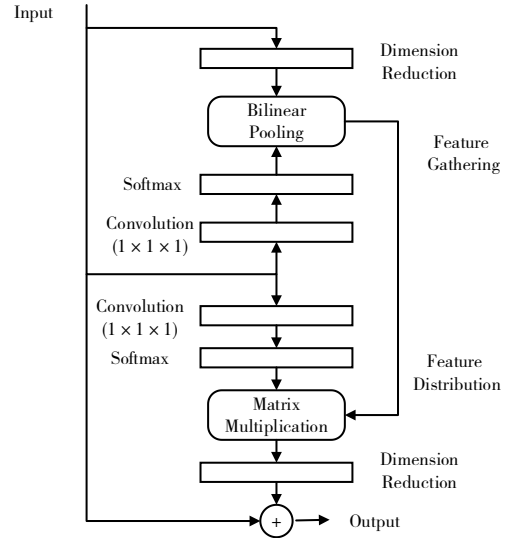


Fig. 2 Calculation process of dual attention block

图 2 双重注意力块的计算流程

在末尾添加一个额外的卷积层以扩展 Z 的通道数,从而通过逐元素加法将其编码回输至 X 。图 3 为双重注意力操作,先对三维输入数组 A 产生一组全局特征,再生成新的局部特征 Z_i 。

2.3 A²TLS-HCN

图 4 为 A²TLS-HCN 模型结构,具有参数量少、轻量化的特点,无需进行预训练且拥有较快的训练速度。模型按照特征提取过程可分为 4 个阶段:阶段一、阶段二使用 HCN 的双流网络结构进行点级特征提取和初步的全局特征提取;阶段三、阶段四使用大选择核卷积 LSKNet 和双重注意力网络 A²-Nets 进行全局特征提取。其中,大选择核卷积 LSKNet 的大内核可增大感受野,自适应的感受野能在一定程度减少参数量;双重注意力网络 A²-Nets 可使后续的卷积层访问整个空间的特征更有效。

由于沿关节维的卷积核大小为 1,使模型从每个关节的三维坐标中独立地学习点级表示。后续,本文以参数 $(0, 2, 1)$ 对特征映射进行转置得到 X''_{x_i} ,以便将联合维度移动到张量通道中。

阶段二:通过两个 3×3 (Conv3 和 Conv4) 卷积层分别对骨骼信息和运动信息进行全局学习。

$$X'''_{x_i} = \text{Conv3}(X''_{x_i}) \quad (4)$$

$$S_{x_i} = \text{Conv4}(X'''_{x_i}) \quad (5)$$

式中: X'''_{x_i} 表示 conv3 的输出; S_{x_i} 表示骨骼分支在阶

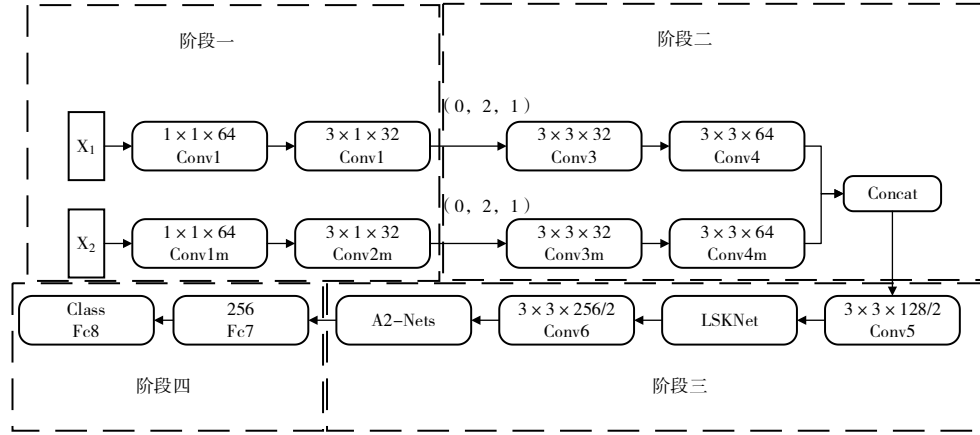


Fig. 4 A²TLS-HCN model structure

图4 A²TLS-HCN模型结构

阶段二的结果;时间分支同理可得到第二阶段的结果 M_{x_i} 。

阶段三:将阶段二的两个分支结果进行拼接,再经过一个 3×3 (conv5)卷积层。

$$F_{x_i} = \text{Concat}(S_{x_i}, M_{x_i}) \quad (6)$$

$$F'_{x_i} = \text{Conv5}(F_{x_i}) \quad (7)$$

模型在经过 LSKNet 后增大了网络感受野,即每个神经元对输入数据的感知范围增加,使网络能更好地理解输入数据的结构和关联性。同时,不同通道之间的交互更稀疏,有助于减少不同通道之间的冗余信息传递,提升特征的独立性和多样性,使特征学习更高效。此外,自适应感受野既提升了特征提取效果,又减少了参数量。

$$F''_{x_i} = \text{LSKNet}(F'_{x_i}) \quad (8)$$

经过一个 3×3 (conv6)卷积层后可得:

$$F'''_{x_i} = \text{Conv6}(F''_{x_i}) \quad (9)$$

随后,通过双重注意力层可使模型学习到不同行为中的关键点,从而提升行为识别的准确性,帮助模型动态调整不同时刻的关注度,使其关注与当前行为更相关的骨骼节点,有助于提升模型对于长距离依赖关系的建模能力。

阶段三的特征输出结果 T_{x_i} 为:

$$T_{x_i} = \text{A2Nets}(F'''_{x_i}) \quad (10)$$

阶段四:特征图被平铺成一个向量,通过两个全连接层进行最终分类。对于识别任务,本文使用 softmax 函数对网络输出进行归一化,序列 X 属于第 i 类的概率为:

$$P(CilX) = \frac{e^{o_i}}{\sum_{j=1}^C e^{o_j}}, i = 1, 2, \dots, C \quad (11)$$

式中: $\mathbf{o} = (o_1, o_2, \dots, o_c)^T$ 为网络输出; C 为类的数量。

3 实验结果与分析

3.1 实验环境

本文模型共进行 30 万次迭代,批次大小为 32,优化器为 Adam,学习率初始化为 0.001,每 1 000 步以指数衰减,

速率为 0.99,如表 1 所示。在训练过程中从整个序列中随机截取一个子序列,裁剪比取自 $[0.5, 1]$ 之间的均匀分布。在训练过程中,以 0.9 的比例对子序列进行中心裁剪。由于不同动作的持续时间不同,输入序列被归一化为固定长度(本文取 32),并沿着帧维进行双线性插值。为了缓解过拟合现象,在 Conv4、Conv5、Conv6 和 fc7 后附加了比率为 0.5 的 Dropout,并对 fc7 的权重施加 0.001 的衰减率。

Table 1 Experimental environment

表 1 实验环境

类目	参数
操作系统	Windows 11 专业版
处理器	Intel(R) Xeon(R) Gold 5218R CPU @ 2.10 GHz
GPU	NVIDIA RTX 3060Ti
深度学习框架	Pytorch 2.0.1 版本
编程语言	Python 3.8

3.2 数据集

本文研究使用经典的行为识别数据集 SBU Kinect Interaction 数据集和 NTU RGB+D 60 数据集。其中,SBU 数据集是一个复杂的人类活动数据集,描述了一个人的互动情况,包括同步视频、深度和运动捕捉数据;NTU RGB+D 60 数据集是基于骨骼的人类动作识别的经典数据集,包含了 60 个种类的动作,共 56 880 个样本,有 40 类为日常行为动作,9 类为与健康相关的动作,11 类为双人互动动作。

图 5 为 NTU RGB+D 数据集动作可视化结果。图 5(a)、图 5(b)为单人日常行为动作,分别为脱鞋和戴帽子;图 5(c)、图 5(d)为双人交互动作,分别为推拉和拥抱。NTU RGB+D 数据集有跨主体评估(Cross-Subject Evaluation, CSE)和跨视角评估(Cross-View Evaluation, CVE)两种推荐的评估方案。

CSE 将 20 个人作为训练集并编码为 1、2、4、5、8、9、13、14、15、16、17、18、19、25、27、28、31、34、35、38,剩余数据为测试集。在此评估方案中,训练集与测试集主体不重叠,意味着训练模型的数据来自于某些主体,而测试模型的数据来自于其他主体,原因是为了测试模型在未见过的

主体上的泛化能力,以更好地模拟实际应用中模型对新主体的适应能力。

CVE按相机来划分训练集与测试集,将相机1采集的样本作为测试集,相机2、相机3采集的数据作为训练集,样本数分别为18 960和37 920,相机垂直高度保持相同,水平角度设置不同,分别为 -45° 、 0° 和 45° 。在此评估方案中,训练集、测试集的视角不重叠,训练模型的数据来自于某些视角(摄像机角度),而测试模型的数据来自于其他视角,以测试模型在不同视角下的鲁棒性和泛化能力。

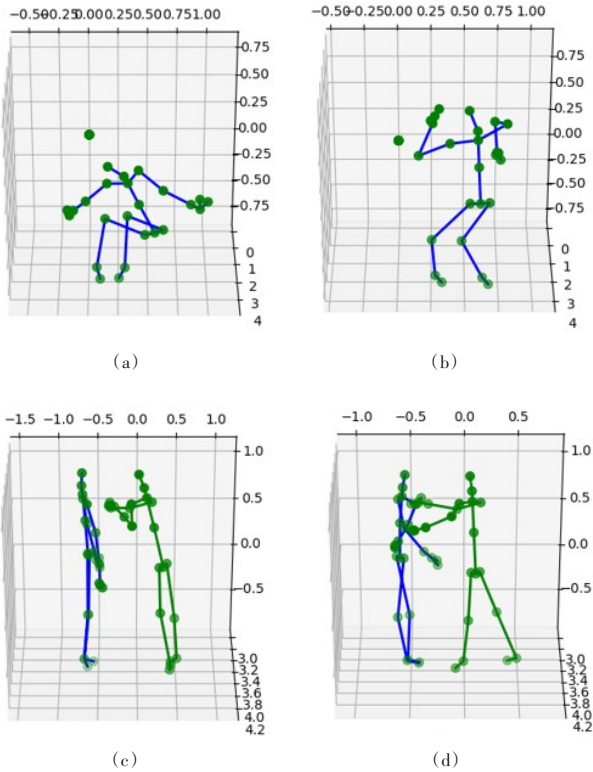


Fig. 5 Visualization results of actions in NTU RGB+D dataset

图5 NTU RGB+D数据集中的动作可视化结果

3.3 评价指标

行为识别是一个分类任务,准确率(accuracy, A)必然是关注的重点。准确率即正确分类个数在全部分类中的占比情况,一般任务中准确率数值越高分类模型的性能越好。本文采用五折交叉验证算法,在NTU RGB+D数据集与SBU Kinect Interaction数据集上得到模型的评价指标,并取评价结果的平均值。

$$A = \frac{N_{\text{corr}}}{N_{\text{total}}} \quad (12)$$

式中: N_{corr} 表示分类正确的样本数量; N_{total} 表示总样本数量。

3.4 实验结果

本文在训练过程中以0.9的比例对子序列进行中心裁剪,输入序列被归一化固定为32,训练轮次设置为400。由图6可知,识别准确率在400轮时最高,训练轮次小于400轮时最佳轮次约等于训练轮次;训练轮次大于400轮

时最佳训练轮次也趋近于400轮,但准确率并未得到提升。

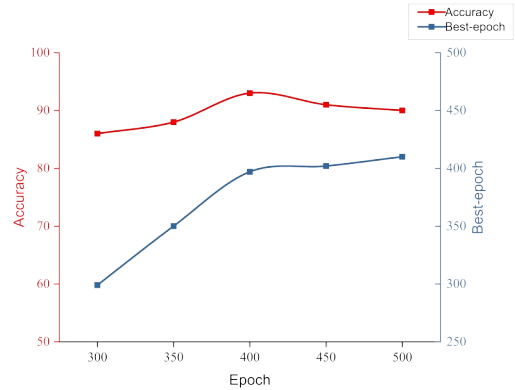


Fig. 6 Relationship between model training epochs and accuracy

图6 模型训练轮次与准确率之间的关系

图7(彩图扫OSID码可见,下同)为模型实验效果,在前100轮中提升较明显,模型整体性能稳定。其中,top1错误率指预测输出的概率最高类别与人工标注类别相符的准确率;top5指预测输出的概率最高的前5个类别,只要出现了正确概率(与人工标注类别一致)即为预测正确,否则预测错误。蓝色线表示训练集top1错误率,绿色线表示测试集top1错误率,模型在前300轮中训练集错误率高于测试集,在400轮时训练集与测试集的错误率几乎相同,模型也达到了最优效果。

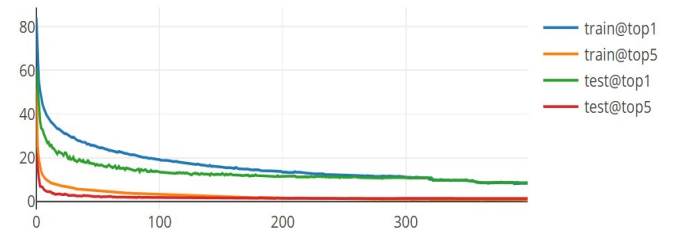


Fig. 7 Error rates of training and testing sets

图7 训练集与测试集错误率

图8为第200轮与400轮训练后的混淆矩阵。由第400轮训练后的混淆矩阵可知,所提模型对一些局部动作相似、动作幅度不大且没有明显交互的动作识别准确率欠佳,但相较于以往方法在涉及长距离依赖关系的动作识别效果方面提升明显。由第200轮训练后的混淆矩阵可知,模型在200轮时就能良好识别大部分动作,相较于400轮训练后的混淆矩阵而言,在识别远程联合互动动作(鼓掌)时的效果更优。

图9为在NTU RGB+D数据集上本文方法相较于HCN方法在部分动作上的准确率差异,仅显示了变化大于2%的类别。由此可见,穿鞋这一动作类别识别准确率提升超过4%,鼓掌、脱鞋识别准确率提升超过3%,接打电话、刷牙、摘下眼镜、跳跃、擦脸、摸头这6个动作类别识别准确率提升超过2%,识别准确率提升较大的动作类别大多涉及长距离依赖关系。综上,本文模型能有效提升远程联合动作类别的识别准确率,明显提升了长距离依赖关系的处

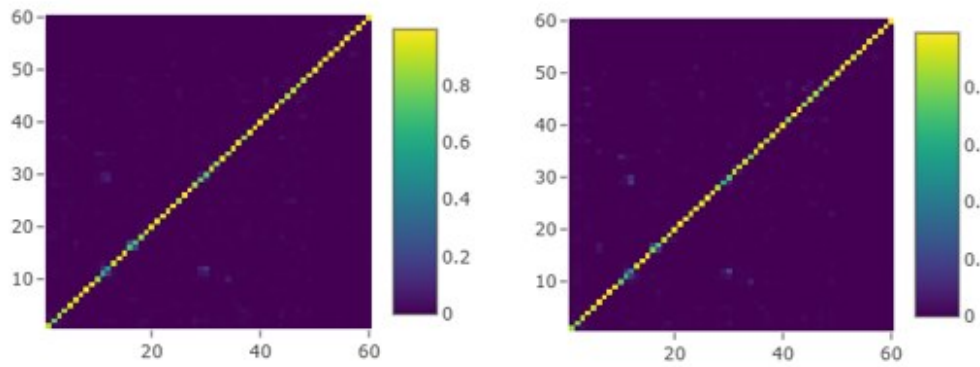


Fig. 8 Confusion matrix after the 200th and 400th rounds of training

图8 第200轮与第400轮训练后的混淆矩阵

理能力。

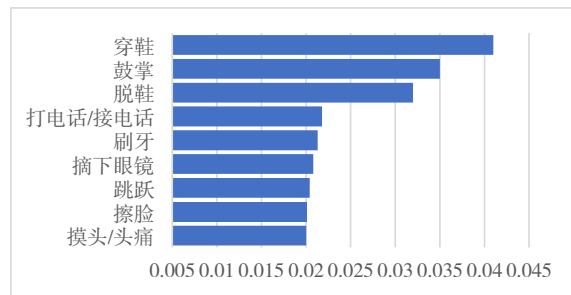


Fig. 9 Difference in accuracy of some actions between the proposed method and HCN method on the NTU RGB+D dataset

图9 在NTU RGB+D数据集上本文方法相较于HCN方法部分动作上的准确率差异

表2为本文方法在NTU RGB+D数据集上与其他基于CNN与LSTM方法的比较结果。由此可知,在NTU RGB+D这一大规模数据集上,本文方法相较于之前的方法在识别准确率上存在明显提升。相较于基于文献[40]的方法,CS设置的准确率提升10.4%,CV设置的准确率提升5.5%。相较于文献[8]的方法,CS设置的准确率提升6.4%,CV设置的准确率提升3.9%。相较于文献[16]的方法,CS设置的准确率提升3.1%,CV设置的准确率提升2.1%。此外,本文方法相较于流行的复杂大模型的不同之处在于:网络仅包含大约90万个参数,属于一个轻量级模型,极小的模型能轻松地从头开始训练网络,无需进行预训练。

表3为本文方法在SBU Kinect Interaction数据集上与

Table 2 Experimental results of different models on NTU RGB+D dataset

表2 不同模型在NTU RGB+D数据集上的实验结果

模型	准确率/%	
	CS	CV
STA-LSTM ^[39]	73.4	81.2
Clips+CNN+MTLN ^[13]	79.6	84.8
VA-LSTM ^[40]	79.2	87.7
Two-stream CNN ^[8]	83.2	89.3
HCN ^[16]	86.5	91.1
A ² TLS-HCN	89.6	93.2

其他方法的比较结果。由此可见,本文方法在小型SBU数据集上也优于其他方法。相较于文献[41]的方法精度提升8.5%。然而,在SBU Kinect Interaction数据集上的识别任务相较于在NTU RGB+D数据集上更容易,分层共现网络的识别准确率已达到98.6%,但A²TLS-HCN方法将识别准确率提升到98.9%,相较于HCN文献[16]的方法提升0.3%。

Table 3 Experimental results of different models on the SBU Kinect Interaction dataset

表3 不同模型在SBU Kinect Interaction数据集上的实验结果

模型	准确率/%
Raw skeleton ^[41]	79.4
Joint feature ^[41]	86.9
ST-LSTM ^[42]	88.6
Co-occurrence RNN ^[41]	90.4
STA-LSTM ^[43]	91.5
ST-LSTM+Trust Gate ^[42]	93.3
VA-LSTM ^[36]	97.6
HCN	98.6
A ² TLS-HCN	98.9

为了进一步评估双重注意力与大选择核卷积对模型的影响,进行了以下消融实验:分别去掉A²TLS-HCN模型中的双重注意力与大选择核卷积模块,在NTU RGB+D数据集和SBU Kinect Inter-action数据集上进行实验,结果如表4、表5所示。由此可知,A²TLK-HCN在NTU RGB+D数据集和SBU Kinect Interaction数据集上的表现始终优于其他模型,证实了双重注意力机制与大选择核卷积对模型学习全局特征具有积极作用,特别是在处理长距离依赖关系时的效果更优,且跨主体评估模型的效果更明显,表明强化全局特征的学习可更好地解决不同主体的差异。

Table 4 Ablation experiments on the NTU RGB+D dataset

表4 在NTU RGB+D数据集上的消融实验

方法	准确率/%	
	CS	CV
HCN	86.5	91.1
HCN+LSKNet	87.0	91.8
HCN+A ² -Nets	87.5	92.7
A ² TLS-HCN	89.6	93.2

Table 5 Ablation experiments on the SBU Kinect Interaction dataset**表 5 在 SBU Kinect Interaction 数据集上的消融实验**

方法	准确率/%
HCN	98.6
HCN+LSKNet	98.8
HCN+A ² -Nets	98.7
A ² TLS-HCN	98.9

4 结语

本文提出一种融合双流网络、双重注意力网络及大选择核卷积网络的 A²TLS-HCN 模型。该模型从所有关节中获取全局响应,既提升了模型对全局特征的提取能力,又强化了特征提取的质量,还有效提升了模型捕捉长距离依赖关系的能力。

在 NTU RGB+D 数据集、SBU Kinect Interaction 数据集上的实验表明,所提算法在行为识别任务中的准确率始终优于其他模型,证实了双重注意力机制与大选择核卷积对模型学习全局特征具有积极作用,特别是在处理长距离依赖关系时的效果更优,且跨主体评估模型的效果更明显,表明强化全局特征的学习可更好地解决不同主体之间的差异。后续,将进一步对模型各个部分进行优化,以提升模型的识别性能和鲁棒性。

参考文献:

- [1] LU J, LI X F, ZHAO B, et al. A review of skeleton-based human action recognition[J]. Journal of Image and Graphics, 2023, 28(12): 3651-3669.
卢健, 李莹峰, 赵博, 等. 骨骼信息的人体行为识别综述[J]. 中国图象图形学报, 2023, 28(12): 3651-3669.
- [2] SAINATH T N, VINIYALS O, SENIOR A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]// 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, 2015: 4580-4584.
- [3] DU Y, FU Y, WANG L. Skeleton based action recognition with convolutional neural network[C]// 2015 3rd IAPR Asian Conference on Pattern Recognition, 2015: 579-583.
- [4] KE Q, BENNAMOUN M, AN S, et al. A new representation of skeleton sequences for 3D action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3288-3297.
- [5] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7291-7299.
- [6] SHOTTON J, FITZGIBBON A, COOK M, et al. Real-time human pose recognition in parts from single depth images[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011: 1297-1304.
- [7] NIELSEN M A. Neural networks and deep learning[EB/OL]. <http://neuralnetworksanddeeplearning.com/>.
- [8] WANG H, WANG L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 499-508.
- [9] LI W, WEN L, CHANG M C, et al. Adaptive RNN tree for large-scale human action recognition[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 1444-1452.
- [10] WEI S, SONG Y, ZHANG Y. Human skeleton tree recurrent neural network with joint relative motion feature for skeleton based action recognition[C]// 2017 IEEE International Conference on Image Processing, 2017: 91-95.
- [11] CAETANO C, SENA J, BRÉMOND F, et al. Skelemotion: a new representation of skeleton joint sequences based on motion information for 3D action recognition[C]// 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2019: 1-8.
- [12] JOZE H R V, SHABAN A, IUZZOLINO M L, et al. MMTM: multimodal transfer module for CNN fusion[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 13289-13299.
- [13] SONG Z, ZHOU Y F, JIA J G, et al. Local feature fusion temporal convolutional network for human action recognition[J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(3): 418-424.
宋震, 周元峰, 贾金公, 等. 面向人体动作识别的局部特征融合时间卷积网络[J]. 计算机辅助设计与图形学学报, 2020, 32(3): 418-424.
- [14] YANG Z, LI Y, YANG J, et al. Action recognition with visual attention on skeleton images[C]// 2018 24th International Conference on Pattern Recognition, 2018: 3309-3314.
- [15] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[DB/OL]. <https://arxiv.org/abs/1406.2199>.
- [16] LI C, ZHONG Q, XIE D, et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[DB/OL]. <https://arxiv.org/abs/1804.06055>.
- [17] LUONG M T, LE Q V, SUTSKEVER I, et al. Multi-task sequence to sequence learning[DB/OL]. <https://arxiv.org/abs/1511.06114>.
- [18] SI C, CHEN W, WANG W, et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 1227-1236.
- [19] LI J, LIU X, ZHANG M, et al. Spatio-temporal deformable 3D convnets with attention for action recognition[J]. Pattern Recognition, 2020, 98: 107037.
- [20] CHEN Y, KALANTIDIS Y, LI J, et al. A²-nets: double attention networks[DB/OL]. <https://arxiv.org/abs/1810.11579>.
- [21] PENG C, ZHANG X, YU G, et al. Large kernel matters — improve semantic segmentation by global convolutional network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4353-4361.
- [22] LI P, XIE J, WANG Q, et al. Is second-order information helpful for large-scale visual recognition?[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 2070-2078.
- [23] WANG D, ZHANG Q, XU Y, et al. Advancing plain vision transformer toward remote sensing foundation model[DB/OL]. <https://arxiv.org/abs/>

- 2208.03987.
- [24] BAHDANAU D, CHO K, BENGIO Y J A P A. Neural machine translation by jointly learning to align and translate[DB/OL]. <https://arxiv.org/abs/1409.0473>.
- [25] BAZI Y, BASHMAL L, RAHHAL M M A, et al. Vision transformers for remote sensing image classification[J]. *Remote Sensing*, 2021, 13(3): 516.
- [26] DENG P, XU K, HUANG H J I G, et al. When CNNs meet vision transformer: a joint framework for remote sensing scene classification[J]. *IEEE Geoscience and Remote Sensing Letters*, 2021, 19: 1–5.
- [27] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// *Proceedings of the IEEE International Conference on Computer Vision*, 2021: 10012–10022.
- [28] HAO S, WU B, ZHAO K, et al. Two-stream swin transformer with differentiable sobel operator for remote sensing image classification[J]. *Remote Sensing*, 2022, 14(6): 1507.
- [29] XU X, FENG Z, CAO C, et al. An improved swin transformer-based model for remote sensing object detection and instance segmentation[J]. *Remote Sensing*, 2021, 13(23): 4779.
- [30] WANG W, XIE E, LI X, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]// *Proceedings of the IEEE International Conference on Computer Vision*, 2021: 568–578.
- [31] RANFTL R, BOCHKOVSKIY A, KOLTUN V. Vision transformers for dense prediction[C]// *Proceedings of the IEEE International Conference on Computer Vision*, 2021: 12179–12188.
- [32] YAN H, LI Z, LI W, et al. Contnet: why not use convolution and transformer at the same time?[DB/OL]. <https://arxiv.org/abs/2104.13497>.
- [33] ZHENG S, LU J, ZHAO H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021: 6881–6890.
- [34] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022: 11976–11986.
- [35] DING X, ZHANG X, HAN J, et al. Scaling up your kernels to 31×31: revisiting large kernel design in CNNs[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022: 11963–11975.
- [36] LIU S, CHEN T, CHEN X, et al. More convnets in the 2020s: scaling up kernels beyond 51×51 using sparsity[DB/OL]. <https://arxiv.org/abs/2207.03620>.
- [37] ZHANG P F, LAN C L, XING J L, et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data[C]// *Proceedings of the IEEE International Conference on Computer Vision*, 2017: 2117–2126.
- [38] SONG S J, LAN C L, XING J L, et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data[DB/OL]. <https://arxiv.org/abs/1611.06067>.
- [39] LIU J, SHAHROUDY A, XU D, et al. Spatio-temporal lstm with trust gates for 3D human action recognition[DB/OL]. <https://arxiv.org/abs/1607.07043>.
- [40] GUO M H, LU C Z, LIU Z N, et al. Visual attention network[DB/OL]. <https://arxiv.org/abs/2202.09741>.
- [41] ZHU W, LAN C, XING J, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks[DB/OL]. <https://arxiv.org/abs/1603.07772>.
- [42] LI Y, HOU Q, ZHENG Z, et al. Large selective kernel network for remote sensing object detection[DB/OL]. <https://arxiv.org/abs/2303.09030>.
- [43] JI Y, YE G, CHENG H. Interactive body part contrast mining for human interaction recognition[C]// *2014 IEEE International Conference on Multimedia and Expo Workshops*, 2014: 1–6.

(责任编辑:刘嘉文)