

基于特定帧组与拓扑区的混合骨架序列增强方法

柏永果^{1,2}, 李希^{1,2}, 沈言文^{1,2}, 徐大宏^{1,2}, 刘宏^{1,2}, 郭建平^{2,3}, 张艺瀚^{2,3}

(1. 湖南师范大学信息科学与工程学院; 2. 湖南师范大学体育智能研究重点实验室;
3. 湖南师范大学体育学院, 湖南长沙 410081)

摘要: 骨架数据增强对于基于对比学习的骨架行为识别至关重要, 通常采用全局式增强模式对骨架序列每一帧的所有关节部位进行统一变换处理。然而在实际场景中, 骨架数据的语义信息具有局部特性, 往往集中于局部帧或局部关节部位。仅依赖全局式方式无法充分捕捉关节间的局部关系, 导致难以获得有效反映局部特征的高质量样本。为此, 提出一种基于运动注意力选择帧组与语义感知拓扑区的混合骨架序列数据增强方法SFG-TR, 以获取高质量增强样本。具体来说, 基于运动注意力选择帧组策略帮助模型锁定特定的帧组, 借助语义感知拓扑区策略强化模型对关键关节特征的捕捉能力, 这些关键关节能够更精确地表述人体动作的语义特征。该方法在NTU-RGB+D数据集的关节流(X-Sub)上的识别准确率达到80.1%, 在NTU120-RGB+D数据集的关节流(X-Sub)上的识别准确率达到69.1%, 验证了其有效性。

关键词: 行为识别; 对比学习; 图卷积神经网络; 骨架序列; 混合骨架序列增强; 数据增强

DOI: 10.11907/rjtk.242046

扫描二维码阅读全文:



中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2026)002-0014-07

Hybrid Skeleton Sequence Augmentation Method Based on Specific Frame Groups and Topological Regions

BAI Yongguo^{1,2}, LI Xi^{1,2}, SHEN Yanwen^{1,2}, XU Dahong^{1,2}, LIU Hong^{1,2}, GUO Jianping^{2,3}, ZHANG Yihan^{2,3}

(1. College of Information Science and Engineering, Hunan Normal University; 2. Key Laboratory of Sports Intelligence Research, Hunan Normal University; 3. College of Physical Education, Hunan Normal University, Changsha 410081, China)

Abstract: Skeleton data augmentation is crucial for skeleton behavior recognition based on contrastive learning, and a global augmentation mode is usually used to uniformly transform all joint parts of each frame of the skeleton sequence. However, in practical scenarios, the semantic information of skeleton data has local characteristics and is often concentrated in local frames or joint points. Relying solely on a global approach cannot fully capture the local relationships between joints, making it difficult to obtain high-quality samples that effectively reflect local features. To this end, a data augmentation method SFG-TR based on a hybrid skeleton sequence of motion attention selected frame groups and semantic perception topology regions is proposed to obtain high-quality enhanced samples. Specifically, the frame group selection strategy based on motion attention helps the model lock onto specific frame groups, and the semantic perception topology region strategy enhances the model's ability to capture key joint point features. These key joint points can more accurately express the semantic features of human movements. The recognition accuracy of this method on the joint flow (X-Sub) of the NTU-RGB+D dataset reached 80.1%, and on the joint flow (X-Sub) of the NTU120-RGB+D dataset reached 69.1%, verifying its effectiveness.

Key Words: action recognition; contrastive learning; graph convolutional neural network; skeleton sequence; hybrid skeleton sequence augmentation; data augmentation

收稿日期: 2025-01-13

基金项目: 湖南省普通本科高校教学改革研究项目(202401000057)

作者简介: 柏永果(1996-), 女, CCF会员, 湖南师范大学信息科学与工程学院硕士研究生, 研究方向为计算机视觉、数字图像处理; 李希(1981-), 男, 博士, 湖南师范大学信息科学与工程学院讲师, 研究方向为计算机视觉、体育智能、区块链; 沈言文(1996-), 男, 湖南师范大学信息科学与工程学院硕士研究生, 研究方向为自然语言处理; 徐大宏(1971-), 男, 博士, 湖南师范大学信息科学与工程学院副教授、硕士生导师, 研究方向为机器视觉、大数据等; 刘宏(1963-), 男, 硕士, 湖南师范大学信息科学与工程学院教授、博士生导师, 研究方向为人工智能、智能家居; 郭建平(1970-), 男, 博士, 湖南师范大学体育学院教授, 研究方向为体育智能等; 张艺瀚(1990-), 男, 博士, 湖南师范大学体育学院讲师, 研究方向为体育智能、运动心理学等。本文通讯作者: 李希。

0 引言

基于骨架的人体行为识别属于计算机视觉领域的重要研究分支,广泛应用于视频监控、体育动作分析和人机交互等智能化领域^[1-3]。早期识别方法均通过监督学习实现,需要人工标注数据。随着数据集规模的扩大,标注成本急剧攀升。为此,研究者们将对比学习框架引入行为识别算法中,并获得了较优性能。对比学习的核心机制在于借助对比损失函数拉近正样本对之间的距离,同时推远负样本对。在该过程中,如何精准获取高质量的正负样本对成为决定对比学习算法性能优劣的关键因素。数据增强技术可用于获取正负样本对,其通常采用全局式增强模式对所有骨架数据进行统一变换处理。然而在实际场景中,骨架数据的语义信息具有局部特性,往往集中在某些特定帧或特定关节部位,仅通过全局式增强方法无法顾及这些局部信息,需要探索新的数据增强方法来更加精细地描述这些局部关系,以生成更高质量的样本对^[4]。

1 相关工作

1.1 基于骨架的行为识别

在早期基于骨架的行为识别研究中,循环神经网络(Recurrent Neural Network, RNN)和卷积神经网络(Convolutional Neural Network, CNN)表现突出。随着图网络技术的不断发展,图卷积神经网络(Graph Convolution Neural Network, GCN)在该领域研究中同样受到高度关注,其中的图和邻接矩阵能够自然地描述关节点与关节点、骨骼与骨骼以及关节点与骨骼之间的关系。而时空图卷积网络(Spatial Temporal Graph Convolutional Network, ST-GCN)的提出进一步为动态描述人体关节间的时空关系提供了新方案^[5]。随后,越来越多基于骨架的行为识别创新实践将ST-GCN作为骨干编码器,如2s-AGCN、AM-GCN等均是在ST-GCN模型的基础上不断改进,并取得了显著成果^[6-7]。

1.2 对比学习

对比学习是一种从未标注数据中获得鲁棒行为表征的方法,正负样本的构建和对比是其核心机制。对比学习设计思想简单且效果显著,受到广泛关注。例如,文献[8]提出的SimCLR通过数据增强生成正样本对,并结合大规模负样本对比和非线性投影头,在无标注图像数据上学习到具有强迁移能力的视觉表征。然而,SimCLR的负样本数量受到样本批次大小的限制,无法提供丰富的负样本对。文献[9]提出的MoCo和文献[10]提出的MoCov2通过引入Memory Bank存储队列增加了负样本对数量,帮助模型改善泛化表现。文献[11]提出的CrosSCLR利用跨视图知识的一致性,引入局部视图挖掘潜在的正样本对,以优

化对比学习性能。文献[12]提出的ISC实现了跨骨架表示的对比学习,旨在学习更高级的语义特征。文献[13]提出的PSTL重构对比学习框架是一种无负样本的三流架构,探索了同一骨架序列中存在的局部时空关系。文献[14]提出的LEWEL证实了局部特征的学习对于提升表征质量的重要作用。文献[15]提出的HiCo不仅关注实例级的特征表示,而且分别从时间和空间的角度逐层捕捉更细粒度的局部特征,以强化对比学习的力度,但其对计算资源的要求较高。

1.3 数据增强

近年来,许多研究通过改进数据增强策略来提升对比学习性能,以达到更好的行为识别效果。例如,文献[16]提出的AS-CAL通过综合运用多种数据增强技术,显著加强了模型对动作本质特征在不同变换模式下的理解力和捕捉力。文献[17]提出的CLSA在对比学习框架中引入更强的增强技术,相较于随机剪切、旋转等弱增强,强增强极大改变了图形的外观和结构。文献[18]提出的AimCLR借鉴文献[17]的思想,将强增强技术引入骨架数据中,创造了新的运动模式,训练模型获得了更具通用性的特征表示。然而,骨架序列通过这种极端增强处理后可能会丢失其原始特征。文献[19]提出的HiCLR提出一种渐增式增强框架,以分层的方式叠加不同的增强策略,从而生成不同增强强度的正样本对,在运动模式变化的同时又确保了行为特征表示的一致性。以上提及的数据增强方式主要侧重以全局式方式对骨架数据的所有帧和关节点进行统一无差别地变换处理,无法充分捕捉关节间的局部关系。文献[20]提出的SkeleMixCLR采用局部式增强模式,通过随机混合不同骨架序列的裁剪和截断部位,生成了新的对比学习样本对。然而,SkeleMixCLR选择时间连续的帧片段和拓扑区关节的方式具有随机性,可能会忽视那些与运动语义特征强相关的帧。因此,需要探索新的数据增强方法来更加精细地描述运动过程中不同关节点以及视频帧之间的局部关系,以生成更高质量的样本对。

为更有效地表达关节点以及视频帧之间的局部关系,本文提出一种基于运动注意力选择帧组(Motion-attention Frame Group Selection, MFGS)与语义感知拓扑区(Semantic-aware Topological Regions, STR)的混合骨架序列数据增强方法SFG-TR(Specific Frame Groups and Topological Regions)。主要贡献为:①提出MFGS策略来选择携带符合要求信息的帧组,在整个上下文信息融合交换过程中,可以在保持表征关键信息不缺失的同时保证语义信息的相对连续性;②利用STR策略获取运动骨架中特定语义聚集的关节拓扑区域,进一步提升动作表征的精确度;③融合MFGS策略与STR策略,使两个不同的骨架序列在关键帧组中融合交换特定拓扑区域的骨架部分,从而产生更高质量的正负样本对,增强对比学习的效果;④在两个骨架数据集上进行的大量实验证实了所提方法的有效性和优

越性。

2 SFG-TR方法构建

本文在 SkeleMixCLR 的基础上设计了一种基于 MFGS 与 STR 的混合骨架序列数据增强方法 SFG-TR, 包括全局对比学习和局部对比学习两个分支, 结构如图 1 所示。全局对比学习分支基于 SkeletonCLR 框架设计, 用于提取人体行为的全局运动特征; 局部对比学习分支则通过运动注意力值判断并选择关键帧组, 并利用语义感知模块确定重要的拓扑关节区域。通过整合不同骨架序列中特定帧组和关节区域的信息, 融合多样化上下文语义, 可生成更具挑战性的混合样本。该方法旨在为对比学习添加更高质量的正负样本对, 为自监督学习提供有益补充。通过这种方式, 模型能够更深入地捕捉和学习骨架序列中的关键运动特征和语义信息。

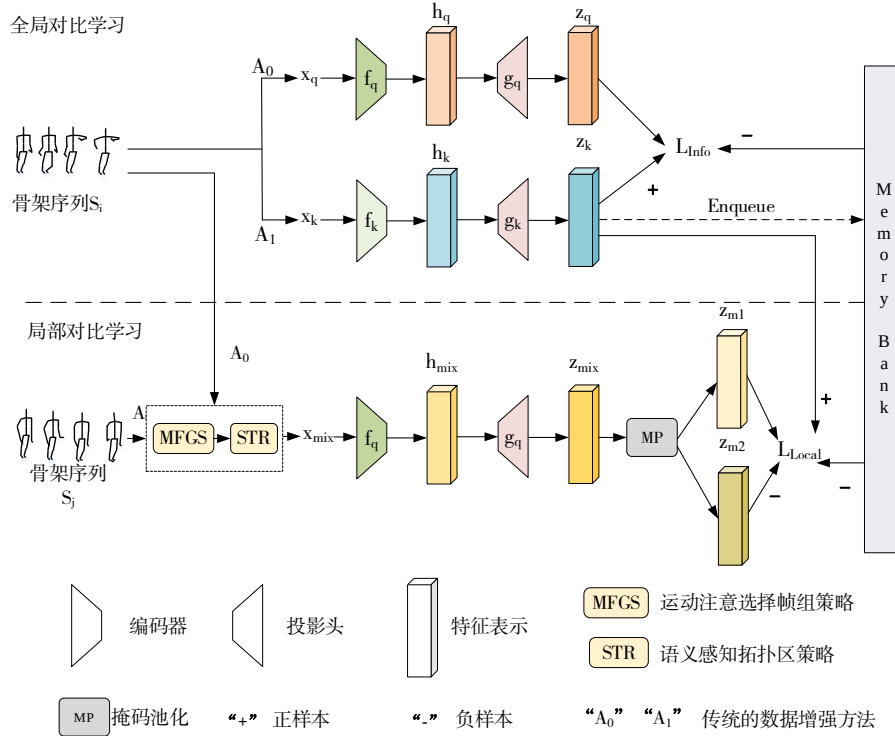


Fig. 1 Structure of SFG-TR

图1 SFG-TR结构

2.2 STR策略

根据骨架的拓扑结构, 将全身关节划分为5个部分, 其中左、右手各6个关节, 左、右腿各4个关节, 躯干5个关节。如图3所示, “梳头”这个动作由局部的一只手控制, 该区域特征在整体动作的语义特征中起着决定性作用, 其他关节区的影响甚微。本文提出的STR策略便是为了掌握对行为识别至关重要的部分特征。

具体来说, 在每次迭代中, 首先计算样本整体的平均运动值, 并将其与当前骨架序列一并作为STR模块的输入; 其次, 该模块分析并为每个关节分配重要性评分, 根

2.1 MFGS策略

SkeleMixCLR 从骨架序列中随机选择一段在时间上连续的帧片段, 因此可能忽视了那些与运动语义特征强相关的帧^[20]。为此, 本文提出 MFGS 策略, 可以根据运动注意力值的大小灵活选择符合要求的帧组。如图2所示, 将运动变化率值视为每帧的注意力权重, 因此, 第 t 时间步的运动注意力值 $m = S_{:,t+1,:} - S_{:,t,:}$, 根据值的大小对所有帧进行排序, 选出符合要求的帧。从0到 T 帧, 每4帧为一组, 64帧的骨架序列被有序地均分为16组。随后, 根据 m 大小对包含整个运动持续变化的所有帧进行排序, 并在排序完成后进行遍历。当遍历到第 t 帧时, 某组中出现的帧数越多, 该组的排序越靠前; 对于帧数相同的组, 出现得越早, 其排名越靠前。排名靠前的帧组将优先进入信息融合交换帧的范畴。由于帧组内的帧是连续的, 而帧组之间可能不连续, 这在一定程度上既保证了关键信息的完整性, 又保持了语义信息的相对连续性。

据评分计算5个拓扑集的均值并进行优先级排序, 最终选择部分拓扑区域作为信息融合交换的子集。该过程亦可以被抽象为一个选择机制, 其根据拓扑关节区域的重要性来决定后续用作融合交换的骨架部分。

2.3 损失函数

在全局特征提取框架中, 输入为骨架序列 $S \in R^{C \times T \times V}$, 包含 C 个通道、 T 个帧、 V 个关节。骨架数据在经过裁剪与错切等全局式数据增强后, 得到两个在视觉上略有不同但语义信息高度一致的视图 x_q 和 x_k 。随后, 分别将 x_q 和 x_k 作为查询编码器 (Query) 和键编码器 (Key) 的输

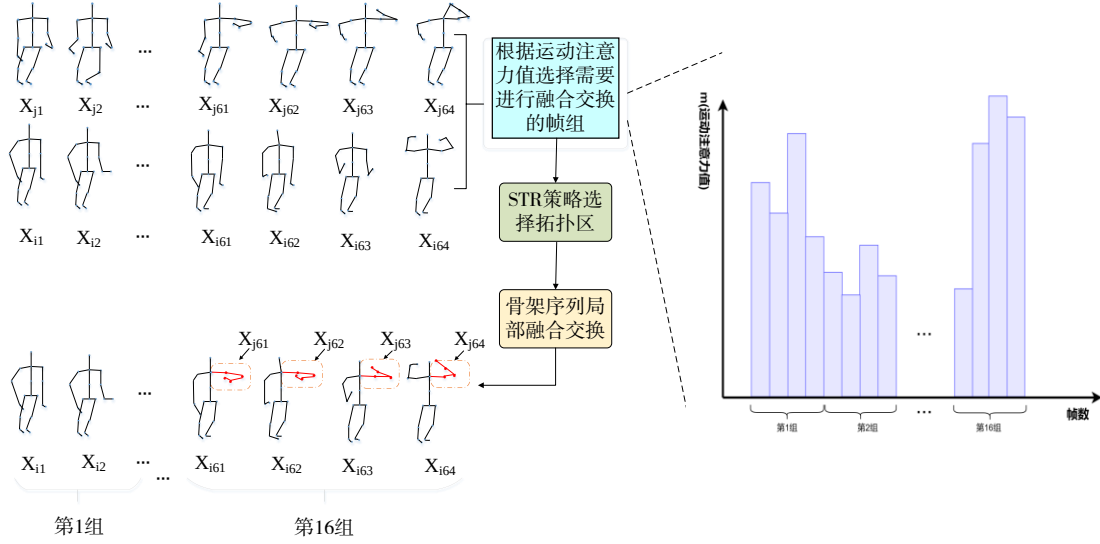


Fig. 2 Motion-attention frame group selection strategy

图 2 MFGS 策略

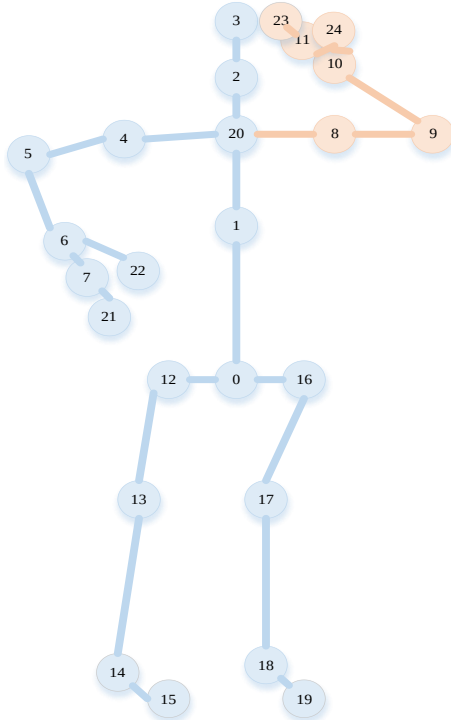


Fig. 3 Skeleton topology diagram of the "combing hair" action

图 3 “梳头”动作的骨架拓扑图

入, 编码器的输出继而作为投影头的输入, 最终得到特征表示 z_q 和 z_k 。在同一轮迭代中, z_q 和 z_k 互为正样本对, 负样

本 m_i 来自 Memory Bank 存储队列中的第 i 个样本。在对比学习过程中, 先前迭代得到的特征表示 z_k 被存储在 Memory Bank 队列中, 用于构建大量负样本。按照存储顺序对 Memory Bank 进行出队操作, 即越早存储的特征越早出队, 出队后成为负样本 m_i , 与当前迭代计算得到的 z_q 配对, 形成负样本对。这些样本对随后被用于对比学习, 参与 InfoNCE 损失的计算过程。公式为:

$$L_{Info} = -\log \left(\frac{\exp(z_q \cdot z_k / \tau)}{\exp(z_q \cdot z_k / \tau) + \sum_{i=1}^n \exp(z_q \cdot m_i / \tau)} \right) \quad (1)$$

式中: $\cdot$ 表示点积, 用于计算两个样本之间的相似度; τ 为温度超参数, 用于控制相似度的平滑度。

融合 MFGS 策略与 STR 策略后, 两个不同的骨架序列有序地合成了一个新的混合骨架序列增强样本 $x_{mix} \in R^{C \times T \times V}$ 。新样本在经过编码器时, 随时间下采样生成特征 $h_{mix} \in R^{C' \times T' \times V}$, 其特征维度发生了变化。在池化层中, 通过掩码池化模块 (Masking Pooling, MP) 将不同部位的特征分开处理, 保持各自的独立性, 即从混合特征 h_{mix} 中提取含有不同上下文信息的独立特征 z_{m1} 和 z_{m2} , 其中 z_{m1} 携带信息的骨架部分来自骨架序列 S_i , z_{m2} 携带信息的骨架部分来自骨架序列 S_j 。在局部特征损失计算中, z_{m1} 与 z_{m2} 互为困难负样本对, z_{m1} 与 m_i 互为负样本对, z_{m1} 与 z_{k1} 互为正样本对。计算关于 z_{m1} 的对比损失 L_1 。公式为:

$$L_1 = -\log \left(\frac{\exp(z_{m1} \cdot z_{k1} / \tau)}{\exp(z_{m1} \cdot z_{k1} / \tau) + \exp(z_{m1} \cdot z_{m2} / \tau) + \sum_{i=1}^n \exp(z_{m1} \cdot m_i / \tau)} \right) \quad (2)$$

类似地, 计算 z_{m2} 的对比损失 L_2 。公式为:

$$L_2 = -\log \left(\frac{\exp(z_{m2} \cdot z_{k2} / \tau)}{\exp(z_{m2} \cdot z_{k2} / \tau) + \exp(z_{m1} \cdot z_{m2} / \tau) + \sum_{i=1}^n \exp(z_{m2} \cdot m_i / \tau)} \right) \quad (3)$$

在全局对比学习和局部对比学习的双重作用下, 模型得到了更好的优化, 鲁棒性有所提升。总损失 L 的计算公

式为:

$$L = L_{Info} + h_1 * L_1 + h_2 * L_2 \quad (4)$$

式中:超参数 $h_1, h_2 \in (0, 1)$ 。

3 实验方法与结果分析

3.1 实验环境与参数设置

本文实验采用 Python 编程语言,基于 PyTorch 深度学习框架构建并训练模型,所有计算过程均在 GPU 上执行。在数据准备阶段,将每个骨架序列样本的帧数统一至 64 帧。模型共进行 300 轮训练, Batch Size 为 128, 优化器选用 SGD, 初始学习率为 0.1。为确保模型稳定收敛,权重衰减率设置为 0.000 1。实验主干网络采用 ST-GCN, 温度超参数设为 0.2。在测试阶段采用线性评估方案评估模型性能,即以 Top-1 识别准确率作为唯一评价指标。该指标反映了模型在预测样本时概率值最高的类别与其真实标签之间的匹配程度。本文提及的准确率皆指 Top-1 识别准确率。测试训练周期为 150 轮,初始学习率为 3,从第 101 轮起降为前 100 轮的 1/10。

3.2 数据集

本文实验选用骨架数据集 NTU-RGB+D (NTU-60) 和 NTU120-RGB+D (NTU-120) 进行模型训练与测试。其中, NTU-60 数据集包含 60 个动作类别,共 56 880 个样本,由 40 个受试者完成拍摄。算法评估设置包括跨受试者 (X-Sub) 和跨视角 (X-View) 两种方式,其在训练集和测试集的划分上有所不同:在 X-Sub 中,训练集和测试集分别由 20 个不同的受试者组成,但训练集的样本数比测试集多出 23 760 个;在 X-View 中,由 2 和 3 号摄像机拍摄的 3 7920 个样本作为训练集,剩余样本由 1 号相机拍摄,作为测试集^[21]。

在 NTU-60 的基础上, NTU-120 扩增至 120 类别,共采集 114 480 个样本。该数据集包含 106 个受试者完成的 82 个日常行为(如梳头、擦脸等)、12 个健康相关行为(如打哈欠、腰酸背痛等)以及 26 个双人互动行为(如拥抱、推等)。评估设置方式分为跨受试者评估设置 (X-Sub) 和跨场景评估设置 (X-Set) 两种^[22]。

3.3 实验结果分析

3.3.1 超参数对实验结果的影响

(1) 训练周期对实验结果的影响。训练周期的长短直接影响到模型学习的效果,如果训练时间短,模型可能因为没有充沛时间提取到复杂特征而导致学习能力不够;相反,训练时间太长,容易导致模型过拟合,进而影响其在测试集上的线性评估结果。由表 1 可知,模型在 NTU-60 数据集 (X-View) 上训练 100 个 epoch 时,本文方法的准确率为 82.8%,随着训练周期数的不断增加,识别准确率也在随之增加,当训练到 300 个 epoch 时,本文方法的识别准确率达到 85.1%。因此将 300 作为本次实验训练周期的默认值。

(2) 动量系数 m 对实验结果的影响。本文方法遵循

Table 1 Comparison of recognition accuracy across different training epochs

表 1 不同训练周期的识别准确率比较 (%)		
Epoch	X-View	X-Sub
100	82.8	76.0
200	84.5	77.6
300	85.1	80.1

MoCo 框架, Key 编码器的参数需要依靠 Query 编码器参数和动量系数 m 来更新,以保证特征表示的稳定性和连续性。 m 的取值范围在 0~1 之间, m 值设置得越大意味着 Key 编码器的参数变化越缓慢,更多地依赖于其自身的历史参数,而较少地受到 Query 编码器最新参数的影响。表 2 为本文方法在 NTU-60 数据集两种协议下关节流数据上取不同 m 值时的识别准确率。可以看出,动量系数 m 的取值对模型的识别准确率影响较大。当 m 取值 0.999 时,模型的识别准确率会比 m 取值 0 和 0.9 高很多,故而将 0.999 作为本文动量系数 m 的默认值。

Table 2 Comparison of recognition accuracy for different values of momentum coefficient m

表 2 不同动量系数 m 下的识别准确率 (%)		
动量系数 m	X-View	X-Sub
0	10.9	12.0
0.9	76.6	48.4
0.999	85.1	80.1

3.3.2 与 SkeleMixCLR 方法的比较

SFG-TR 方法与 SkeleMixCLR 方法的比较结果见表 3, 分别展示了两种方法在关节流 (J)、运动流 (M)、骨骼流 (B) 以及三流集成 (J+M+B) 数据上的实验结果,其中三流集成指模型在训练过程中通过融合关节、骨骼和运动 3 个数据流的特征实现联合优化。可以看出,除了 NTU-60 数据集的 X-View, SFG-TR 在关节流、运动流、骨骼流以及三流集成数据上的识别准确率均高于 SkeleMixCLR。

Table 3 Comparison results of SFG-TR and SketchMixCLR

表3 SFG-TR与SkeleMixCLR比较结果									(%)
模型	数据流	NTU-60				NTU-120			
		X-Sub		X-View		X-Sub		X-Set	
		准确 率	Δ	准确 率	Δ	准确 率	Δ	准确 率	Δ
SkeleMixCLR	J	77.6		84.5		67.6		67.6	
	M	73.6		78.3		48.8		53.3	
	B	76.9		82.0		66.1		56.2	
3s-SkeleMixCLR	J+M+B	81.3		86.4		70.2		68.1	
SFG-TR	J	80.1	+2.5	85.1	+0.6	69.1	+1.5	68.3	+0.7
	M	74.1	+0.5	77.0	-1.3	51.2	+2.4	55.1	+1.8
	B	77.7	+0.8	82.0	0.0	66.2	+0.1	62.5	+6.3
3s-SFG-TR	J+M+B	82.6	+1.3	87.2	+0.8	71.0	+0.8	70.8	+2.7

3.3.3 与先进方法的比较

为验证本文方法的优越性,分别在 NTU-60 和 NTU-120 数据集上将其与目前的一些先进方法进行比较,结果见表 4 和表 5。其中,单流表示仅使用一种数据流训练,此处为关节流;三流表示融合关节、运动和骨骼 3 种数据流

进行训练。由表 4 可以看出,在 NTU-60 数据集的 X-Sub 和 X-View 两种设置下,本文方法在单流数据上的识别准确率分别达到 80.1% 和 85.1%,综合来看优于 ISC^[12]、PSTL^[13]、AS-CAL^[16]、AimCLR^[18]等先进方法。本文方法在多流数据上的识别准确率分别达到 82.6% 和 87.2%,相较于 3s-SkeletonCLR^[11]、3s-CrosSCLR^[11]、3s-AimCLR^[18]和 3s-HiCLR^[19]等先进方法具有一定竞争力。

Table 4 Comparison of recognition accuracy between the proposed method and advanced methods on the NTU-60 dataset

表 4 本文方法与先进方法在 NTU-60 数据集上的识别准确率比较 (%)

数据流	方法	NTU-60	
		X-Sub	X-View
单流	ISC	76.3	85.2
	AS-CAL	58.5	64.8
	AimCLR	74.3	79.7
	PSTL	77.3	81.8
	SFG-TR	80.1	85.1
	3s-AimCLR	78.9	83.8
三流	3s-SkeletonCLR	75.0	79.8
	3s-CrosSCLR	77.8	83.4
	3s-HiCLR	80.4	85.5
	3s-SFG-TR	82.6	87.2

由表 5 可以看出,在 NTU-120 数据集的 X-Sub 和 X-Set 两种设置下,本文方法在单流数据上的识别准确率分别达到 69.1% 和 68.3%,在多流数据上的识别准确率分别达到 71.0% 和 70.8%,与在 NTU-60 数据集上的表现一致,均优于对照方法。以上结果证实了本文方法的有效性以及在不同数据集上的泛化能力。

Table 5 Comparison of recognition accuracy between the proposed method and advanced methods on the NTU-120 dataset

表 5 本文方法与先进方法在 NTU-120 数据集上的识别准确率比较 (%)

数据流	方法	NTU-120	
		X-Sub	X-Set
单流	SkeletonCLR	56.8	55.9
	ISC	67.1	67.9
	AimCLR	63.4	63.4
	PSTL	66.2	67.7
	SFG-TR	69.1	68.3
	3s-AimCLR	68.2	68.8
三流	3s-SkeletonCLR	60.7	62.6
	3s-CrosSCLR	67.9	66.7
	3s-HiCLR	70.0	70.4
	3s-SFG-TR	71.0	70.8

3.3.4 消融实验

为证明 MFSG 策略和 STR 策略的有效性,在 NTU-60 数据集的关节流数据上进行消融实验,结果见表 6。可以看出,在 NTU-60 数据集的 X-Sub 上, SkeleMixCLR 的识别准确率为 77.6%,单独添加 MFSG 策略后,模型的识别准确率为 78.7%,提高了 1.1%,证明了该策略的有效性。单独添加 STR 策略后,模型的准确率为 79.1%,提高了 1.5%,证

明了该策略的有效性。同时添加两种策略后,模型的准确率为 80.1%,提高了 2.5%,证明了采用两种策略共同训练模型比单独采用其中一种策略效果更好。

Table 6 Results of the ablation study on the NTU-60 dataset

表 6 NTU-60 数据集上的消融实验结果 (%)				
方法	MFSG 策略	STR 策略	X-Sub	X-View
SkeleMixCLR			77.6	84.5
SkeleMixCLR+ MFSG	√		78.7	83.3
SkeleMixCLR+ STR		√	79.1	84.8
SkeleMixCLR+ MFSG+ STR	√	√	80.1	85.1

4 结语

骨架数据增强方法通常以全局式增强模式对骨架数据进行统一变换处理,忽视了骨架数据中的语义信息往往集中在局部视频帧和局部拓扑区的骨架部分这一问题。为此,本文提出基于 MFSG 与 STR 的混合骨架序列增强的方法,更加精细地描述运动过程中不同关节点之间以及视频帧之间的局部关系。其中, MFSG 策略通过注意力值引导模型专注在特定帧组,从而提高复杂动作的识别准确率; STR 策略选择对运动起重要作用的关节点拓扑区,进一步提升了动作表征的精确度。此外,两种策略的融合还为对比学习提供了高质量且有效的对比对,进一步提升了模型的准确性和鲁棒性。在 NTU-60 和 NTU-120 两个大型数据集上的实验结果证实了所提方法的有效性。然而,该方法在泛化能力和计算效率方面依存在提升空间,未来可尝试将在关键区学到的深层语义知识传递给参与全局特征学习的模型,以增强其对复杂场景的理解与识别能力。

参考文献:

[1] YAN X Y, KUANG Y X, BAIG R, et al. Student classroom behavior recognition method based on deep learning [J]. Computer Engineering, 2023, 49(7): 251-258.
闫兴亚, 匡娅茜, 白光睿, 等. 基于深度学习的学生课堂行为识别方法[J]. 计算机工程, 2023, 49(7): 251-258.

[2] SUN M Z, ZHANG P, SU B Y. Survey of human action recognition methods based on skeleton data features [J]. Software Guide, 2022, 21(4): 233-239.
孙满贞, 张鹏, 苏本跃. 基于骨骼数据特征的人体行为识别方法综述[J]. 软件导刊, 2022, 21(4): 233-239.

[3] WU F, WANG Q Z, BIAN J, et al. A survey on video action recognition in sports: Datasets, methods and applications [J]. IEEE Transactions on Multimedia, 2022, 25: 7943-7966.

[4] LIANG C W, YANG J, HU W, et al. Temporal dynamic frame selection and spatio-temporal graph convolution for interpretable skeleton-based action recognition [J]. Journal of Graphics, 2024, 45(4): 791-803.
梁成武, 杨杰, 胡伟, 等. 基于时间动态帧选择与时空图卷积的可解释骨架行为识别[J]. 图学报, 2024, 45(4): 791-803.

[5] YAN S J, XIONG Y J, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition [DB/OL]. <https://arxiv.org/>

- abs/1801.07455.
- [6] SHI L, ZHANG Y F, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12026–12035.
 - [7] WANG X, ZHU M Q, BO D Y, et al. AM-GCN: adaptive multi-channel graph convolutional networks[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020: 1243–1253.
 - [8] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//International Conference on Machine Learning, 2020: 1597–1607.
 - [9] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 9729–9738.
 - [10] CHEN X L, FAN H Q, GIRSHICK R, et al. Improved baselines with momentum contrastive learning [DB/OL]. <https://arxiv.org/abs/2003.04297>.
 - [11] LI L G, WANG M S, NI B B, et al. 3d human action representation learning via cross-view consistency pursuit[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 4741–4750.
 - [12] THOKER F M, DOUGHTY H, SNOEK C G M. Skeleton-contrastive 3D action representation learning[C]//Proceedings of the 29th ACM International Conference on Multimedia, 2021: 1655–1663.
 - [13] ZHOU Y J, DUAN H D, RAO A, et al. Self-supervised action representation learning from partial spatio-temporal skeleton sequences[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 3825–3833.
 - [14] HUANG L, YOU S, ZHENG M K, et al. Learning where to learn in cross-view self-supervised learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 14451–14460.
 - [15] DONG J F, SUN S K, LIU Z L, et al. Hierarchical contrast for unsupervised skeleton-based action representation learning [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 525–533.
 - [16] RAO H C, XU S H, HU X P, et al. Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition[J]. Information Sciences, 2021, 569: 90–109.
 - [17] WANG X, QI G J. Contrastive learning with stronger augmentations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(5): 5549–5560.
 - [18] GUO T Y, LIU H, CHEN Z, et al. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 762–770.
 - [19] ZHANG J H, LIN L L, LIU J Y. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 3427–3435.
 - [20] CHEN Z, LIU H, GUO T Y, et al. Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition[DB/OL]. <https://arxiv.org/abs/2207.03065>.
 - [21] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+D: a large scale dataset for 3D human activity analysis[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1010–1019.
 - [22] 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.

(责任编辑:尹晨茹)