中国科学: 信息科学 2023 年 第 53 卷 第 12 期: 2440–2457
SCIENTIA SINICA Informationis
论文

学习成对样本运动显著性的细粒度人体骨架动作识别

李红艳[1,2†], 涂志刚[2†], 谢伟[3], 张嘉旭[2*]

1. 湖北经济学院信息工程学院, 武汉 430205
2. 武汉大学测绘遥感信息工程国家重点实验室, 武汉 430072
3. 华中师范大学计算机学院, 武汉 430079
* 通信作者. E-mail: zjiaxu@whu.edu.cn
† 同等贡献

收稿日期: 2023–03–17; 修回日期: 2023–05–22; 接受日期: 2023–06–27; 网络出版日期: 2023–12–08

湖北省杰出青年学者自然科学基金 (批准号: 2022CFA075) 和国家自然科学基金 (批准号: 62106177) 资助项目

摘要 基于骨架数据的细粒度人体动作识别是一项重要的研究课题, 但未被充分解决. 由于骨架数据缺乏视觉表观信息, 相似类别的人体动作很难被现有的深度网络模型识别. 在这项工作中, 我们提出了一个新型的运动显著性探测器 (motion salience prober, MSP), 并引入了配对学习 (motion salience prober-incorporated pairwise-learning, MSP-PL) 框架, 以实现细粒度的骨架动作识别. 我们的 MSP-PL 框架在构造成对的相似骨架运动样本基础上 (查询样本与探测样本), 利用运动显著性学习机制, 促进编码器学习精细化的运动特征. 其核心模块 MSP 可以在我们设计的探测样本和损失函数的帮助下, 增强查询样本的显著性运动特征, 并消除冗余的噪声. 本文设计了 3 种探测样本构造策略来生成查询 – 探测样本对, 辅助模型识别查询样本的动作, 并测试了它们对模型性能的影响. 在 NTU-RGB+D 120 数据集与 Kinetics-Skeleton 数据集上的大量实验表明, 我们的 MSP-PL 框架是通用的, 大多数骨架特征编码器可以无缝嵌入其中, 并显著提高其准确性. 5 个主流的编码器对精细化动作的平均分类准确率提高了 2.4% 以上. 此外, 我们的 MSP-PL 框架在与最新的编码器相结合时, 在骨架动作识别方面达到了最先进的性能.

关键词 骨架动作识别, 细粒度动作识别, 视觉注意力, 运动显著性学习, 对比学习

1 引言

深度相机 (如微软 Kinect [1]) 和人体姿势估计算法 [2~5] 的最新进展使得快速准确地估计人体的骨架数据成为可能, 且骨架数据的获取成本也愈发低廉. 因此, 基于骨架的人体动作识别, 在人机交

引用格式: 李红艳, 涂志刚, 谢伟, 等. 学习成对样本运动显著性的细粒度人体骨架动作识别. 中国科学: 信息科学, 2023, 53: 2440–2457, doi: 10.1360/SSI-2023-0047
Li H Y, Tu Z G, Xie W, et al. Fine-grained skeleton action recognition with pairwise motion salience learning (in Chinese). Sci Sin Inform, 2023, 53: 2440–2457, doi: 10.1360/SSI-2023-0047

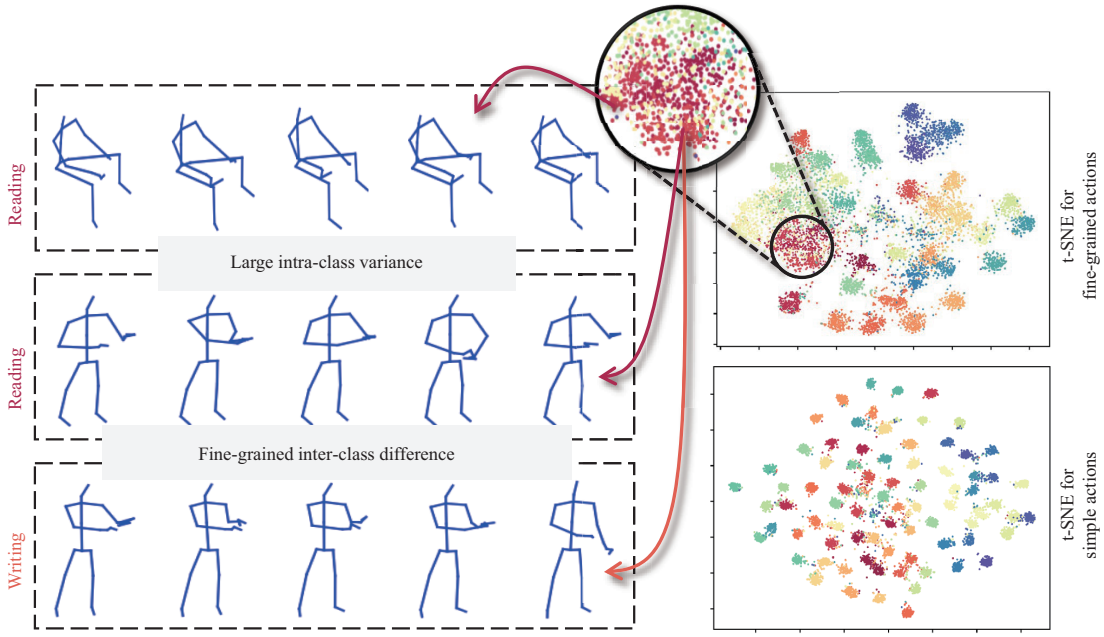

图 1 细粒度骨架动作的例子和本文的研究动机. 左边部分显示了 3 个有代表性的细粒度动作样本的可视化. 右边部分显示了 NTU-120 [16] 数据集中简单动作和细粒度动作的隐式表征的 t-SNE 嵌入

Figure 1 Examples of the fine-grained skeleton actions and our motivation. Left part shows the visualization of three representative fine-grained action samples. Right part shows the t-SNE embeddings of the latent representations for the simple actions and the fine-grained actions in the NTU-120 dataset [16]

互 [6]、自动驾驶 [7]、体育分析 [8,9] 等方面具有广泛的应用前景, 在过去几年中引起了越来越多的关注. 骨架数据作为人类运动的压缩表示, 对背景、灯光和摄像机视角的干扰具有很强的鲁棒性. 最近, 学者们在骨架人体动作识别领域提出了许多先进的方法, 并取得了显著的性能提升 [10~14]. 然而, 由于缺乏表观信息, 仅考虑骨架数据很难识别人体的细粒度动作 [15]. 如图 1 所示, 识别细粒度的骨架动作 (即区分相似的动作类别, 如阅读、书写和打字) 主要面临两个挑战:

(1) 细粒度动作类内差异大. 同样的动作可以由不同的执行者以不同的运动风格来完成. 以 “阅读” 这个动作为例, 有的人站着阅读, 有的人坐着阅读, 这意味着同一个动作, 人体的姿势可能是多种多样的.

(2) 细粒度动作类间差异小. 如果不给出外观背景, 细粒度的骨架动作彼此之间的运动模式非常相似. 以 “阅读” 和 “书写” 这两个动作为例, 它们之间的动作差异仅仅是手部关节的细微频率. 这种微小的差异却是识别这两个动作的关键信息, 但它很难被模型感知和利用.

由于这两个主要的挑战, 在运动隐式特征空间中, 来自相似类别的动作的表征是过度离散且边界模糊的, 彼此无法被模型区分 (如图 1 [16] 的右边). 最近基于骨架数据的动作识别研究侧重于提取骨架序列的时空运动特征和探索骨架拓扑图的物理结构 [17~20]. 然而, 它们忽略了为精细化的运动特征构建一个类内紧凑且类间分散的运动隐式空间, 导致它们无法分辨出细粒度的人体动作. 目前, 只有少数骨架动作识别文献涉及精细化骨架运动特征提取 [15,21], 但它们没有完全解决上述问题. 我们通过实验发现, 在 NTU-120 数据集上评估现有的方法时, 简单动作类别的识别和细粒度的识别之间存在大约 20% 的准确率差距. 因此, 如何通过充分利用显著的运动线索来学习可区分的运动隐式特征空间并构造细粒度动作类别的决策边界是问题的核心所在.

本文尝试设计通用的学习框架, 让深度模型模仿人类的观察对比方式, 以区分细粒度的骨架动作. 具体来说, 通过让深度模型比较、总结和学习成对骨架运动样本的显著运动线索来提高模型识别细粒度骨架动作的能力. 为此, 本文提出了一种新型的模型学习框架: 结合配对学习的运动显著性探测框架 (motion salience prober-incorporated pairwise-learning, MSP-PL). 它可以与大多数现有的骨架特征编码器 (如 ST-GCN [12] 和 MS-G3D [22]) 无缝兼容, 并显著提高其动作识别性能. 本文提出了一种新型的运动显著性探测器 (motion salience prober, MSP), 在自适应构造的探测样本的帮助下学习查询样本的具有辨别性的运动特征. 自适应构造的探测样本与本文提出的两个对比损失函数相结合 (即配对排序损失和门控对比损失), 使网络能够感知每个查询 – 探测样本对之间的微小差异. 在这两个精心设计的损失函数的支持下, MSP 可以从成对的骨架样本中发现差异线索, 增强细粒度动作的显著运动特征, 并通过基于注意力机制的门控特征消除人体的噪声信息. 此外, 本文提出并评估了 3 种不同的探测样本构建策略, 为本文的框架构建有效的查询 – 探测样本对, 这有利于充分研究相似运动样本的特征以及它们对模型性能的影响.

本文的主要贡献有以下 3 个方面:

• 本文提出了一个通用的 MSP-PL 框架用于细粒度的骨架运动表征挖掘. 在配对排序损失和门控对比损失的驱动下, 本文的 MSP-PL 框架可以对比性地感知查询 – 探测样本对之间的显著运动特征, 并学习具有区分性的时空运动信息.

• 为构建合理且有效的查询 – 探测样本对, 以辅助 MSP-PL 进行运动显著性对比学习, 本文设计并评估了 3 种探测样本构建策略, 以此来促进模型挖掘查询样本的精细化运动特征.

• 广泛的实验表明, 大多数主流的骨架运动特征编码器可以无缝地嵌入到本文的 MSP-PL 框架中, 并有效提高其在细粒度骨架动作识别上的准确性.

2 相关工作

2.1 骨架人体动作识别

基于骨架序列数据的人体动作识别一直是计算机视觉中的一个热门话题. 早期的方法是利用人工设计的骨骼运动特征, 例如骨骼结点的位置、速度以及加速度特征, 对简单动作进行分类 [23~25]. 随着深度学习技术的发展, 深度模型, 如循环神经网络 (recurrent neural network, RNN) 和卷积神经网络 (convolutional neural network, CNN), 显著提升了骨架动作识别的准确率 [26~28]. 基于深度学习的方法不需要人工设计的运动特征, 而是用大量数据训练网络自适应地学习, 因此对更多复杂的人体动作也具有鲁棒性. Yan 等 [12] 首先提出了一种基于图卷积神经网络 (graph convolutional network, GCN) 的方法. 该方法可以挖掘人体骨架图的先验结构, 从而辅助提取骨架的时空运动信息, 实现了较高精度的骨架动作识别. 基于此, 学者们又探索了许多 GCN 方法的变体, 这些变体通常引入一些增量模块, 如注意力机制模块 [17]、环境感知模块 [18, 29]、语义引导模块 [30] 和拓扑细化模型 [31], 以提升神经网络对骨架运动序列的特征学习能力. 一般来说, 由粗到细地提取时空运动信息是基于骨架动作识别的趋势. 然而, 如何从细粒度的骨架动作中进一步提取具有辨识性的运动特征, 仍然是一个亟待解决的问题. Li 等 [15] 提出了一个相似度损失函数来加权惩罚难例数据样本, 用于精细化特征挖掘. 但是, 相似样本对之间细粒度动作的显著运动关联性与差异性没有被充分地利用. 与上述方法相比, 本工作提出的 MSP-PL 框架可以有效地帮助现有的神经网络模型从相似骨架样本对中学习到精细化的显著运动信息, 促进骨架动作识别走向高精度与细粒度.

2.2 细粒度视觉分类

利用局部信息和全局信息来学习具有区分性的显著特征对于细粒度的视觉分类至关重要 [32~35]. Fu 等 [36] 提出了一个循环注意力卷积神经网络, 它可以循环式地学习基于区域的多尺度图像表征, 从而在不同尺度下挖掘细粒度类别图像的精细化特征. Zhuang 等 [37] 提出了一个注意力交互网络来模仿人类的渐进式判别过程, 并通过样本交互学习, 共同利用对比性特征来区分细粒度的图像. 此外. 一些研究还探讨了在视频理解 [38]、点云分类 [39]、实例分割 [40] 和动作检测 [41] 等领域的精细化视觉特征提取. 对于细粒度的骨架动作识别, 不同动作类别显著性的运动特征和相似骨架样本对的细微运动差异也是至关重要的线索, 其在目前的工作中很少被利用. 本文设计了一个具有运动显著性探测器 (MSP) 的配对学习框架来解决上述问题.

2.3 配对学习

类似于人类通过观察样本之间的区别与联系来区分不同事物, 配对学习能够促进深度学习模型对成对样本之间的相对关系进行有效建模 [42, 43]. 最近, 配对学习被引入到细粒度的视觉分类中, 以捕捉样本对的丰富特征. Zhang 等 [44] 设计了一个三元组损失 (triplet loss) 来区分子类别之间的细微差别. Dubey 等 [45] 使用一个成对混淆正则化方法来约束细粒度的类内变化. 另外, 最新的对学习方法, 例如文献 [46~48], 通过构造正负样本对并设计相应的损失函数来促使模型学习到更优的分类边界. 受这些工作的启发, 本文构建了成对的骨架样本并引入了两个额外的损失函数, 即配对排序损失和门控对比损失. 本文使用样本对和损失函数来对比性地促进编码器学习具有区分性的骨架运动特征, 并区分细粒度的运动差异. 与现有对比学习方法不同, 本文提出的面向细粒度骨架动作识别的成对样本运动显著性学习过程, 只需在训练过程中构造时空特征相似的样本对来促进精细化运动特征学习, 而非约束正负样本对在特征空间中的相似性.

3 方法

3.1 概述

细粒度的骨架动作识别是一项具有挑战性的任务, 因为相似类别的动作序列之间的时空运动差异通常十分微小, 难以被深度模型感知. 例如, 通过比较 “阅读” 和 “写作” 这两个动作的骨架序列 (见图 1), 我们可以发现, 在 “阅读” 中, 手的摆动幅度大、频率低; 而在 “写作” 中, 手的摆动幅度小、频率高. 如果没有来自场景和交互物体的外观信息, 我们只能通过比较骨架运动序列之间的这些细微差别来发现细粒度动作的特征. 受此启发, 本文设计了一个配对学习框架, 如图 2 所示. 在训练过程中, 我们将每个训练样本视为查询样本 $\boldsymbol{s}_{\mathrm{q}}$, 并针对查询样本自适应地构建探测样本 $\boldsymbol{s}_{\mathrm{p}}$, 用于精细化表征学习 (详见第 3.3 小节). 之后, $\boldsymbol{s}_{\mathrm{q}}$ 和 $\boldsymbol{s}_{\mathrm{p}}$ 的特征编码都被送入本文提出的运动显著性探测器 (MSP), 以进一步辅助编码器感知它们之间的细微差别. 整个精细化的特征学习过程由我们提出的损失函数 $\mathcal{L}_{\mathrm{fine}}(\cdot)$ 进行监督训练, 其包括了配对排序损失和门控对比损失 (详见 3.2 小节). 有了上面的符号说明, 本文的学习框架可以表述为以下形式:

$$\theta_{\phi}^{*}, \theta_{\mathcal{E}}^{*} = \min_{\theta_{\phi}, \theta_{\mathcal{E}}} \mathcal{L}_{\mathrm{ce}} + \mathcal{L}_{\mathrm{fine}}\Big(\phi\big(\mathcal{E}(s_{\mathrm{q}}),\ \mathcal{E}(s_{\mathrm{p}})\big)\Big), \tag{1}$$

其中 $\mathcal{E}(\cdot)$ 是骨架特征编码器, $\phi(\cdot)$ 是 MSP 模块, $\mathcal{L}_{\mathrm{ce}}(\cdot)$ 是查询样本 $\boldsymbol{s}_{\mathrm{q}}$ 与探测样本 $\boldsymbol{s}_{\mathrm{p}}$ 的交叉熵损失. $\mathcal{L}_{\mathrm{fine}}(\cdot)$ 是用于精细化表征学习的损失函数. θ_{ϕ} 和 $\theta_{\mathcal{E}}$ 分别是 $\phi(\cdot)$ 和 $\mathcal{E}(\cdot)$ 中可学习的网络参数.

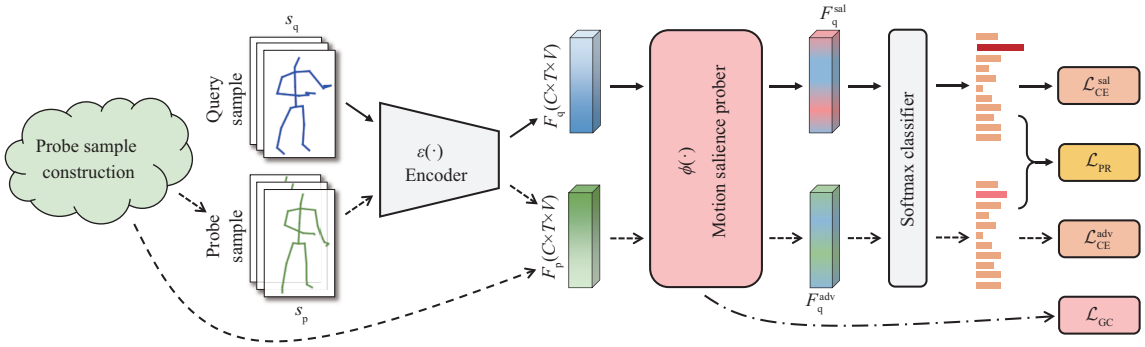

图 2 (网络版彩图) 本文提出的用于细粒度骨架动作识别的 MSP-PL 框架的总体结构图. MSP 是其核心模块, 它可以充分观察对比查询 – 探测样本对, 增强细粒度动作显著的精细化运动线索, 并对比性地消除冗余信息. $\mathcal{L}_{\mathrm{PR}}$ 是配对排序损失, 可以促进编码器和 MSP 从样本对中共同学习具有区分性的运动特征. $\mathcal{L}_{\mathrm{GC}}$ 是门控对比损失, 它可以提高运动特征对相似动作的辨别能力

Figure 2 (Color online) Overview of the proposed MSP-PL framework for fine-grained skeleton action recognition. The MSP is a core module which can fully utilize the sample pairs to attentively enhance the salient motion clues and contrastively eliminate the redundant information. $\mathcal{L}_{\mathrm{PR}}$ is a pair ranking loss which can facilitate the encoder and MSP to jointly learn discriminative features from the sample pairs. $\mathcal{L}_{\mathrm{GC}}$ is a gate contrast loss which can increase the discrimination of the motion features with respect to similar actions

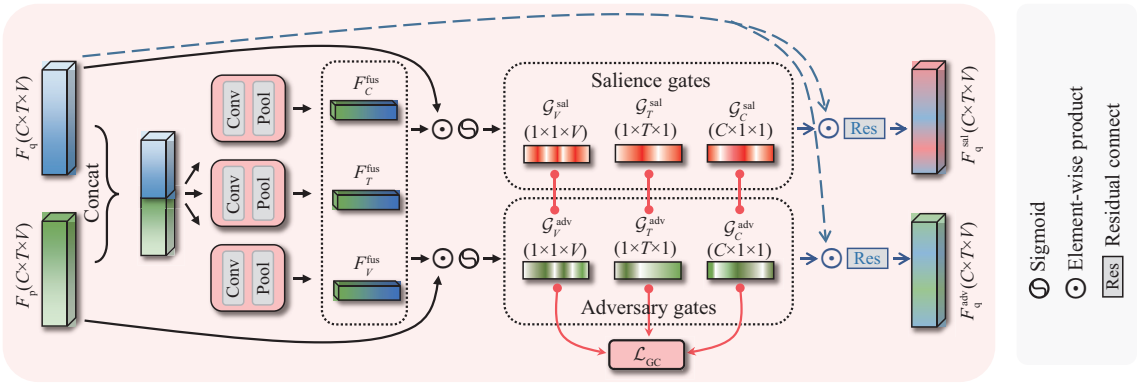

图 3 (网络版彩图) 本文提出的运动显著性探测器 (MSP) 的流程图. 它可以对比性地利用成对的查询 - 探测样本来生成特征门控向量, 并促进编码器学习具有区分能力的运动特征

Figure 3 (Color online) The flow chart of the proposed motion salience prober (MSP). It can contrastively utilize the query-probe sample pair to generate the gate vectors and promote the encoder to learn discriminative features

3.2 运动显著性探测器 (MSP)

令 $F_{\mathrm{q}} \in \mathbb{R}^{C\times T\times V}$ 表示编码器输出的查询样本的特征, 即 $F_{\mathrm{q}} = \mathcal{E}(\boldsymbol{s}_{\mathrm{q}})$, $F_{\mathrm{p}} \in \mathbb{R}^{C\times T\times V}$ 表示探测样本的特征, 即 $F_{\mathrm{p}} = \mathcal{E}(\boldsymbol{s}_{\mathrm{p}})$, 其中 C 是通道维度, T 是时间维度, V 是骨骼结点维度. 本文利用 MSP 在 F_{p} 的辅助下从 3 个维度上对 F_{q} 进行特征增强. MSP 模块输出两个特征向量, 其中一个是带有运动显著性增强的显著性特征 $F_{\mathrm{q}}^{\mathrm{sal}} \in \mathbb{R}^{C\times T\times V}$. 另一个是对抗性特征 $F_{\mathrm{q}}^{\mathrm{adv}} \in \mathbb{R}^{C\times T\times V}$, 其带有噪声信息. 接下来, 我们将它们送入 FC-Softmax 分类器, 得到两个类别预测分数. 最后, 在运动类别标签的监督下, 可以计算出 4 个损失. 交叉熵损失 $\mathcal{L}_{\mathrm{CE}}^{\mathrm{sal}}$ 和 $\mathcal{L}_{\mathrm{CE}}^{\mathrm{adv}}$ 旨在使模型预测出查询样本的正确标签. 损失 $\mathcal{L}_{\mathrm{PR}}$ 和 $\mathcal{L}_{\mathrm{GC}}$ 是精细化特征学习的损失. MSP 的作用是通过比较查询样本和探测样本的异同来增强查询样本具有显著区分性的运动线索, 并促进编码器提取精细化的运动特征. MSP 的流程图如图 3. 首先, MSP 可以从 F_{q} 和 F_{p} 中学习 3 种融合特征, 分别是结点、时间和通道维度的特征. 这个过程可

以表示为

$$F_d^{\text{fus}} = \text{pool}_d(\text{conv}_d([F_{\text{q}}, F_{\text{p}}])), \tag{2}$$

其中 F_d^{fus} 是融合特征, $d \in \{\text{joint}, \text{time}, \text{channel}\}$. $\text{pool}_d(\cdot)$ 是一个池化操作, $\text{conv}_d(\cdot)$ 是一个具有 1×1 内核的卷积层. $[\cdot]$ 表示在通道维度上进行特征拼接. F_d^{fus} 融合了查询样本和探测样本的信息, 这代表了该样本对中的时空运动对比线索.

在学习了包含丰富对比线索的融合特征 F_d^{fus} 后, MSP 将其与 F_{q} 和 F_{p} 进行比较, 从而分别生成两种门控向量. 分别是, 显著性门控 $\mathcal{G}^{\text{sal}}$, 其由 F_{q} 生成, 用于增强查询样本的显著性运动线索. 对抗门控 $\mathcal{G}^{\text{adv}}$, 其由 F_{p} 生成, 用于突出查询样本的冗余与易混淆特征. MSP 将为结点、时间和通道 3 个维度分别生成相应的门控向量. 具体来说, MSP 将融合特征 F_d^{fus} 与查询样本特征 F_{q} (或探测样本特征 F_{p}) 进行元素乘积, 然后用一个 sigmoid 激活函数来探测查询样本的显著性运动线索 (或噪声特征). 门控向量的生成可以表示为

$$\begin{aligned} \mathcal{G}_d^{\text{sal}} &= \sigma(F_d^{\text{fus}} \odot F_{\text{q}}), \\ \mathcal{G}_d^{\text{adv}} &= \sigma(F_d^{\text{fus}} \odot F_{\text{p}}), \end{aligned} \tag{3}$$

其中, $\odot$ 指向量对应元素的乘积. 接下来, MSP 利用门控向量对查询样本的特征向量 F_{q} 进行残差注意力操作, 输出增强运动显著特征 $F_{\text{q}}^{\text{sal}}$ 和对抗信息特征 $F_{\text{q}}^{\text{adv}}$. 这个过程可以表述为

$$\begin{aligned} F_{\text{q}}^{\text{sal}} &= F_{\text{q}} + (F_{\text{q}} \odot \mathcal{G}_V^{\text{sal}} \odot \mathcal{G}_T^{\text{sal}} \odot \mathcal{G}_C^{\text{sal}}), \\ F_{\text{q}}^{\text{adv}} &= F_{\text{q}} + (F_{\text{q}} \odot \mathcal{G}_V^{\text{adv}} \odot \mathcal{G}_T^{\text{adv}} \odot \mathcal{G}_C^{\text{adv}}). \end{aligned} \tag{4}$$

经过 MSP 的处理, 查询样本的特征会产生两个具有不同侧重点的特征向量, 即 $F_{\text{q}}^{\text{sal}}$ 和 $F_{\text{q}}^{\text{adv}}$. $F_{\text{q}}^{\text{sal}}$ 是由显著性门控增强的, 而 $F_{\text{q}}^{\text{adv}}$ 是由对抗性门控增强的. 直观地说, $F_{\text{q}}^{\text{sal}}$ 是一个更具辨别力的运动特征, 其中包含了查询样本的精细化运动线索, 而 $F_{\text{q}}^{\text{adv}}$ 是一个表达能力较弱的运动特征, 其中包含更多的噪声和非关键信息. 最后, 式 (1) 中的损失 $\mathcal{L}_{\text{ce}}(\cdot)$ 和 $\mathcal{L}_{\text{fine}}(\cdot)$ 可以表述为

$$\begin{aligned} \mathcal{L}_{\text{ce}} &= \alpha \mathcal{L}_{\text{CE}}^{\text{sal}} + \beta \mathcal{L}_{\text{CE}}^{\text{adv}}, \\ \mathcal{L}_{\text{fine}} &= \gamma \mathcal{L}_{\text{GC}} + \mathcal{L}_{\text{PR}}, \end{aligned} \tag{5}$$

其中 α, β 和 γ 是超参数. $\mathcal{L}_{\text{CE}}^{\text{sal}}$ 和 $\mathcal{L}_{\text{CE}}^{\text{adv}}$ 是交叉熵损失:

$$\mathcal{L}_{\text{CE}} = -y_{\text{q}}^{\text{T}} \log(p_{\text{q}}), \tag{6}$$

其中 y_{q} 是查询样本的标签. T 表示矩阵转置. p_{q} 是查询样本的 softmax 类别预测得分.

$\mathcal{L}_{\text{GC}}$ 是对两种门控施加门控对比损失, 可表述为

$$\mathcal{L}_{\text{GC}} = \begin{cases} \log(1 + \text{e}^{\psi(\mathcal{G}^{\text{sal}}, \mathcal{G}^{\text{adv}})}), & \text{if } y_{\text{q}} = y_{\text{p}}, \\ \log(1 + \text{e}^{-\psi(\mathcal{G}^{\text{sal}}, \mathcal{G}^{\text{adv}})}), & \text{if } y_{\text{q}} \neq y_{\text{p}}, \end{cases} \tag{7}$$

其中 $\psi(\mathcal{G}^{\text{sal}}, \mathcal{G}^{\text{adv}})$ 是显著性门控和对抗性门控的均方误差, y_{q} 和 y_{p} 分别是查询样本和探测样本的标签. 根据式 (7), $\mathcal{L}_{\text{GC}}$ 不仅可以使相同动作的门控特征相互接近, 也可以使不同动作的门控特征相互远离, 从而提高细粒度运动特征的辨别能力.

$\mathcal{L}_{\text{PR}}$ 是配对排序损失, 可以表述为

$$\mathcal{L}_{\text{PR}} = \max(0, p_{\text{q}}^{\text{sal}}(c_i) - p_{\text{q}}^{\text{adv}}(c_i) + \epsilon), \tag{8}$$

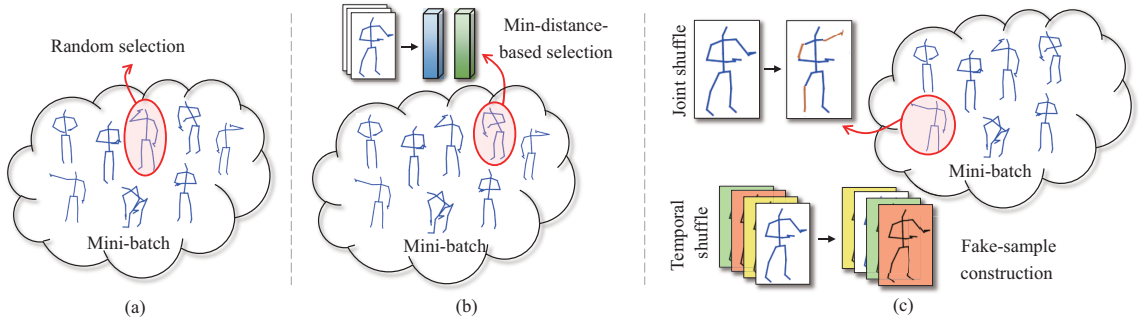

图 4　(网络版彩图) 探测样本的 3 种构造策略

Figure 4　(Color online) Illustration of the probe sample construction strategies. (a) Random selection; (b) min-distance-based selection; (c) fake-sample construction

其中 $p_{\mathrm{q}}^{\mathrm{sal}}(c_i)$ 和 $p_{\mathrm{q}}^{\mathrm{adv}}(c_i)$ 分别是由向量 $p_{\mathrm{q}}^{\mathrm{sal}}$ 和 $p_{\mathrm{q}}^{\mathrm{adv}}$ 得到的分类得分, c_i 是与查询样本的真实标签有关的类别索引. ϵ 是一个可以人工指定的边界阈值系数. $\mathcal{L}_{\mathrm{PR}}$ 可以促进特征 $F_{\mathrm{q}}^{\mathrm{sal}}$ 获得比 $F_{\mathrm{q}}^{\mathrm{adv}}$ 更好的分类结果, 从而优化显著性门控向量, 以样本对比的方式学习丰富的精细化信息. 需要注意的是, 特征 $F_{\mathrm{q}}^{\mathrm{adv}}$ 的分类结果也受到了查询样本标签的监督 (见式 (6)). 通过这种方式, 整个网络可以从查询 – 探测样本对中共同学习精细化的动作特征, 并有效地减少相似骨架动作的混淆.

3.3　探测样本的构造

在本文的 MSP-PL 框架中, 探测样本可以辅助网络学习查询样本的精细化特征. 因此, 为查询样本构建一个合理的探测样本是至关重要的. 本小节介绍了 3 种探测样本的构建策略来讨论这个问题.

(1) 随机选择策略. 如图 4(a) 所示, 最简单的构造方法是在小批量的训练数据中随机选择一个探测样本. 该随机选择的过程, 不会增加任何额外的计算, 是一种快速的构造策略. 在具体实践中, 我们随机打乱训练批次内的查询样本的顺序, 以此构造探测样本.

(2) 基于最小距离的选择策略. 虽然随机选择是最为快速的, 但它往往不是最优的构造方式. 如图 4(b) 所示, 我们引入了基于特征最小距离的选择策略, 为每个查询样本选择更相似的探测样本. 具体来说, 通过计算小批次查询样本的特征之间的距离, 将每个查询样本与最接近的样本作为探测样本进行匹配. 计算特征距离的方式可以采用矢量内积、余弦相似度、L1 距离或 L2 距离等. 不同距离计算方法的影响将在第 4.3 小节讨论.

(3) 虚假样本构造策略. 构造探测样本的核心思想是制定与查询样本相似但在运动细节信息上有微小差异的样本. 因此, 这样的样本对可以促使网络捕捉细粒度骨架动作的细微的相似性和差异性. 按照这一思路, 本文提出了一种虚假样本构造方法. 如图 4(c) 所示, 首先用其他样本的相应骨骼结点替换查询样本的骨骼结点 (结点替换), 然后对时间维度进行分组并打乱分组顺序 (时间乱序). 结点替换操作可以通过 3 种方式进行: 替换随机结点, 替换上半身结点, 或替换下半身结点. 不同的方式和替换结点的数量将在第 4.3 小节中讨论.

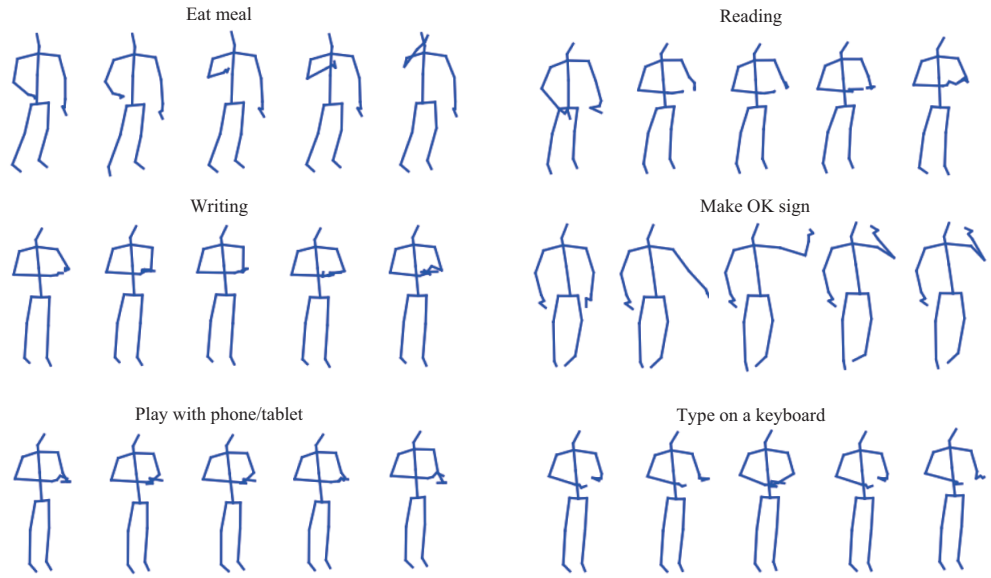

图 5 (网络版彩图) NTU-FG 数据集中一些细粒度骨架动作样本的可视化

Figure 5 (Color online) Visualization of some fine-grained action samples in the NTU-FG dataset

4 实验

4.1 数据集

NTU-RGB+D 120: NTU-RGB+D 120 [16] (NTU-120) 包含 113945 个 3D 骨架动作序列. 该数据集利用 32 个不同视角的摄像机捕捉 106 位动作执行者的 120 个类别的动作. 在骨架动作序列中, 每个人体用 25 个骨骼结点表示. 该数据集包括两种划分方式. (1) 跨主体 (X-Sub) 基准: 106 个动作执行者被分成训练组和测试组. 每组包含 53 个执行者. (2) 跨视角 (X-Set) 基准: 训练数据来自具有偶数相机 ID 的样本, 而测试数据来自具有奇数相机 ID 的样本.

NTU-FG: 为验证并突出我们的方法在细粒度骨架动作识别方面的有效性. 本文基于 NTU-120, 筛选出 42 个分类精度较低的细粒度的动作类别并构建了一个子数据集 NTU-FG. 该 42 个细粒度动作的分类精度均低于 NTU-120 数据集的 120 个动作类别的平均分类精度, 因此识别难度较高. 这些细粒度的骨架动作类别包括 “1. 吃东西”, “2. 阅读”, “3. 书写”, “4. 脱鞋”, “5. 玩手机/平板电脑”, “6. 用键盘打字”, “7. 指东西”, “8. 揉搓双手”, “9. 打喷嚏/咳嗽”, “10. 头痛”, “11. 胸部疼痛”, “12. 颈部疼痛”, “13. 嘘”, “14. 梳头发”, “15. 拇指向上”, “16. 拇指向下”, “17. 做 OK 手势”, “18. 做胜利手势”, “19. 订书”, “20. 数钱”, “21. 剪指甲”, “22. 剪纸 (用剪刀)”, “23. 打响指”, “24. 开瓶”, “25. 闻 (嗅)”, “26. 抛硬币”, “27. 折纸”, “28. 把纸团起来”, “29. 玩魔方”, “30. 在脸上涂油”, “31. 在手背上涂油”, “32. 把东西放进袋子”, “33. 从包里取出东西”, “34. 打开箱子”, “35. 握拳”, “36. 摘掉帽子”, “37. 打哈欠”, “38. 擤鼻子”, “39. 用东西打别人”, “40. 向别人挥刀”, “41. 抢别人的东西”, “42. 用枪向别人射击”. 图 5 显示了一些行动类别在 NTU-FG 上的可视化例子. NTU-FG 总共包含 40008 个骨架动作序列. 与 NTU-120 一样, 其包括两种划分方式, 即 X-Sub 和 X-Set. 如图 5 所示, 该数据集着重关注运动差异较小的细粒度骨架动作, 因此对于动作识别模型具有挑战性.

Kinetics-Skeleton: Kinetics [49] 数据集包括 300000 个视频片段和 400 个动作类别. Kinetics 数据

集中的视频来自 YouTube 网站, 其被世界各地的用户自行上传, 因此类型丰富多样也更符合实际的应用需求. 但此数据集中的视频不包括骨架标注的原始视频. Yan 等 [12] 使用 OpenPose 工具箱 [2] 估计了每个视频帧上人体的 18 个关节的位置, 并相应地发布了 Kinetics-Skeleton 数据集. 在此数据集中, 所有视频首先转换为 30 fps 的帧速率, 分辨率为 340×256. 然后根据 OpenPose 工具箱, 使用这些标准化的视频生成每个人体 18 个关节的 2D 坐标和置信度分数, 以此描述人体骨骼结点的运动. 对于包含有多个人体的视频, 该数据集只保留人体关键点置信度最高的两个个体. 相比于 NTU-120 数据集中的准确 3D 骨架表达, 该数据集中的骨架均为 2D 表示, 且包含有更多噪声.

4.2 实现细节

我们的 MSP 模块中的 3 个卷积层结构是 Conv(2C $\rightarrow$ 128)、Conv(2C $\rightarrow$ 128) 和 Conv(2C $\rightarrow$ C), 分别用于处理结点、时间和通道维度. 其中 $\rightarrow$ 表示特征通道数目的映射, 通道 C 取决于编码器的输出通道数目. 损失函数中的超参数 α, β 和 γ 被设定为 0.1. 在配对排序损失中, 边界阈值系数 ϵ 被设置为 0.05. 本文基于 PyTorch 深度学习框架 [50] 实现我们的模型. 本文应用随机梯度下降 (SGD) 算法, 设置 Nesterov 动量因子为 0.9, 作为优化器. 优化器的权重正则化被设置为 0.0005. 本文使用 4 个 Nvidia GTX 2080Ti GPU 进行模型训练, 并将批次大小设置为 32. 对于所有的数据集, 学习率被设置为 0.1, 训练周期数被设置为 70. 在训练过程中, 学习率衰减在第 30 个周期、第 50 个周期和第 60 个周期执行, 系数被设置为 0.1. 在训练过程中, 本文的 MSP-PL 框架可以辅助模型学习精细化的运动特征. 因此, 在测试过程中, 无需用训练时复杂的方式构造探测样本, 仅需要将测试样本的复制当作其探测样本即可. MSP-PL 框架的相关代码将被开源: https://github.com/FineAct/FineSkeleton.

4.3 消融实验

为了研究本文提出的 MSP-PL 框架的有效性, 本文在 NTU-FG 数据集上对它的效果进行了评估. 值得注意的是, 通过实验, 我们发现如果在 NTU-FG 数据集上训练现有的主流骨架动作识别模型, 它们对细粒度类别的分类精度略低于在 NTU-120 完整数据集上的训练结果. 以 MS-G3D [22] 为例, 在 NTU-FG 和完整 NTU-120 上训练的细粒度动作的分类准确率分别为 71.79% 和 72.02%. 其原因是, 完整的 NTU-120 数据集包含更多的动作样本, 这些多样的样本能够帮助模型学习到更鲁棒的运动特征. 这一现象间接证明了我们之前的阐述, 即在样本类型有限的情况下, 现有的模型不能有效地提取精细化骨架动作的显著特征, 因此本文构造相似的探测样本来辅助模型的精细特征学习.

MSP-PL 框架的有效性. 我们将多种目前主流的编码器嵌入到本文的 MSP-PL 框架中, 包括基于 CNN 的编码器 [51]、基于 GCN 的编码器 [12, 19, 22], 以及基于 Transformer 的编码器 [52]. 在 NTU-FG 数据集上的实验结果展示在表 1 中. 正如预期的那样, 本文的框架可以显著提高各种编码器的精细化特征提取性能. 具体来说, 它为目前使用最广泛的骨架动态特征编码器 ST-GCN [12] 在 X-Sub 基准和 X-Set 基准上分别带来了 4.08% 和 4.13% 的 top-1 精度提升. 图 6 显示了将 ST-GCN 编码器嵌入我们的 MSP-PL 框架前后, 42 个细粒度动作类别精度在 X-Set 基准上的准确性比较. 可以看出, 本文的方法几乎提高了所有精细化动作类别的识别准确性. 其中, 对于 “把纸团起来”、“ 擤鼻子”、“做胜利的手势”、“玩魔方” 和 “拇指向下” 等动作, 其准确率绝对提高了 10% 以上. 这些结果表明, 我们的配对学习框架 MSP-PL 和 MSP 可以与各种编码器无缝集成, 并能有效地帮助编码器学习骨架动作的精细化运动表征, 提升识别准确性.

探测样本的构建策略. 我们测试了本文提出的 3 种不同的策略来构造探测样本. 从表 2 所示的结果中, 我们得到了一些有意义的结论. 表 2 中, “随机替换” 后的数字内容指代替换的结点数目. 首先,

表 1 将不同编码器嵌入 MSP-PL 框架, 在 NTU-FG 数据集上动作识别的 top-1 和 top-5 准确率的比较

Table 1 Comparison of the top-1 and top-5 accuracy on the NTU-FG dataset with different model configurations[a)]

Encoder	X-Sub		X-Set	
	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
HCN [51]	63.75	89.12	66.68	90.89
HCN (MSP-PL)	**66.51**$^{\uparrow 2.76}$	**90.65**$^{\uparrow 1.53}$	**69.92**$^{\uparrow 3.24}$	**92.33**$^{\uparrow 1.44}$
ST-GCN [12]	59.84	87.76	63.32	88.91
ST-GCN (MSP-PL)	**63.92**$^{\uparrow 4.08}$	**90.53**$^{\uparrow 2.77}$	**67.45**$^{\uparrow 4.13}$	**91.01**$^{\uparrow 2.10}$
AGCN [19]	65.53	89.79	68.96	91.52
AGCN (MSP-PL)	**67.85**$^{\uparrow 2.32}$	**91.26**$^{\uparrow 1.47}$	**71.74**$^{\uparrow 2.78}$	**93.03**$^{\uparrow 1.51}$
ST-TR [52]	68.16	92.45	71.21	92.87
ST-TR (MSP-PL)	**69.78** $^{\uparrow 1.72}$	**93.14**$^{\uparrow 0.69}$	**73.45**$^{\uparrow 2.24}$	**93.29**$^{\uparrow 0.42}$
MS-G3D [22]	68.82	92.97	71.79	93.26
MS-G3D (MSP-PL)	**70.22** $^{\uparrow 1.40}$	**93.17**$^{\uparrow 0.20}$	**73.79**$^{\uparrow 2.00}$	**93.76**$^{\uparrow 0.50}$

a) "(MSP-PL)" means with our MSP-based pairwise-learning framework.

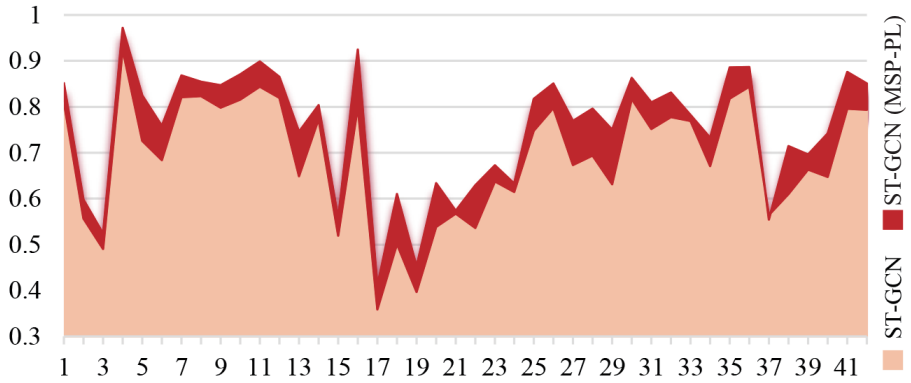

图 6 (网络版彩图) ST-GCN 模型嵌入 MSP-PL 框架前后, 在 NTU-FG X-Set 基准上对 42 个细粒度的动作类别的分类准确性进行比较. 动作 ID 可参考第 4.1 小节

Figure 6 (Color online) Comparison of the classification accuracy of 42 fine-grained action categories before and after equipping STGCN with MSP-PL on the NTU-FG X-Set benchmark. The action ID can be referred to Subsection 4.1

无论采用何种构建策略, 我们的 MSP-PL 框架都能提高基准模型的准确性. 在最简单的随机选择策略中, top-1 的准确性提高了 1.04% (72.83% vs. 71.79%), 这进一步验证了我们方法的有效性和稳健性. 其次, 基于最短距离的选择策略比其他两种策略取得了更好的性能. 对于最短距离的计算方法, 内积法获得了最好的 top-1 准确率 (73.79%), 而余弦相似度法获得了最好的 top-5 准确率 (93.88%). 这些结果反映了我们的 MSP-PL 框架能够从具有相似运动特征的样本对中捕捉到显著性的运动信息. 因此, 在网络训练过程中, 构造与查询样本具有相似运动特征的探测样本是最佳选择. 最后, 虚假样本构造策略的准确率相对其他两种构造策略较低, 这是由于构造出的虚假样本并不包含在训练集中, 且类别标签存在噪声, 不能很好地帮助网络区分数据集特定特征空间中的细粒度样本.

值得注意的是, 从假样本构建的结果中, 我们可以得出两个有价值的结论来指导未来的研究. (1) 选择数据集中包含的, 与查询样本的运动相似的样本作为探测样本, 可以提高网络对细粒度样本的区

表 2 在 NTU-FG X-Set 基准测试中, 采用不同的探测样本构建策略, 比较 top-1 和 top-5 的准确性. 使用 MS-G3D 作为编码器

Table 2 Comparison of the top-1 and top-5 accuracy on the NTU-FG X-Set benchmark with different probe sample construction strategies. Using MS-G3D as the encoder[a)]

Strategy		X-Set	
		Top-1 (%)	Top-5 (%)
Random selection		72.83	93.39
Min-dis selection	Inner product	**73.79**	93.76
	Cosine similarity	73.51	**93.88**
	L1 distance	73.04	93.62
	L2 distance	73.17	93.72
Fake-sample construction	Random (5)	72.40	93.28
	Random (10)	72.74	93.41
	Random (15)	72.91	93.55
	Upper body	<u>73.14</u>	<u>93.61</u>
	Lower body	72.59	93.34

a) Bold and underline indicate the optimal and suboptimal results, respectively.

表 3 在 NTU-FG X-Set 基准测试中, 使用不同的门控向量生成策略, 比较 top-1 和 top-5 的准确性. 使用 MS-G3D 作为编码器

Table 3 Comparison of the top-1 and top-5 accuracy on the NTU-FG X-Set benchmark with different configurations to generate the gate vectors. Using MS-G3D as the encoder[a)]

Operations	X-Set	
	Top-1 (%)	Top-5 (%)
Sum-Conv-Pool + Sigmoid	72.85	93.61
Concat-Conv-Pool + Sigmoid	**73.79**	**93.76**
Concat-Conv-Pool + Softmax	73.23	93.56

a) Bold indicates the optimal results.

表 4 在 NTU-FG X-Set 基准测试中, 使用不同的门控向量, 比较 top-1 和 top-5 的准确性. 使用 MS-G3D 作为编码器

Table 4 Comparison of the top-1 and top-5 accuracy on the NTU-FG X-Set benchmark with different gate vectors. Using MS-G3D as the encoder[a)]

Joint	Time	Channel	X-Set	
			Top-1 (%)	Top-5 (%)
–	–	–	71.79	93.26
✓	–	–	72.86	93.63
–	✓	–	72.55	93.48
–	–	✓	72.44	93.74
✓	✓	✓	**73.79**	**93.76**

a) Bold indicates the optimal results.

分能力. (2) 替换人体上半身结点比替换下半身结点更有效 (73.14% vs. 72.59% top-1 准确率). 因为细粒度的动作主要集中在人体上肢, 而下半身的动作一般是冗余信息. 因此, 探测样本应该与查询样本

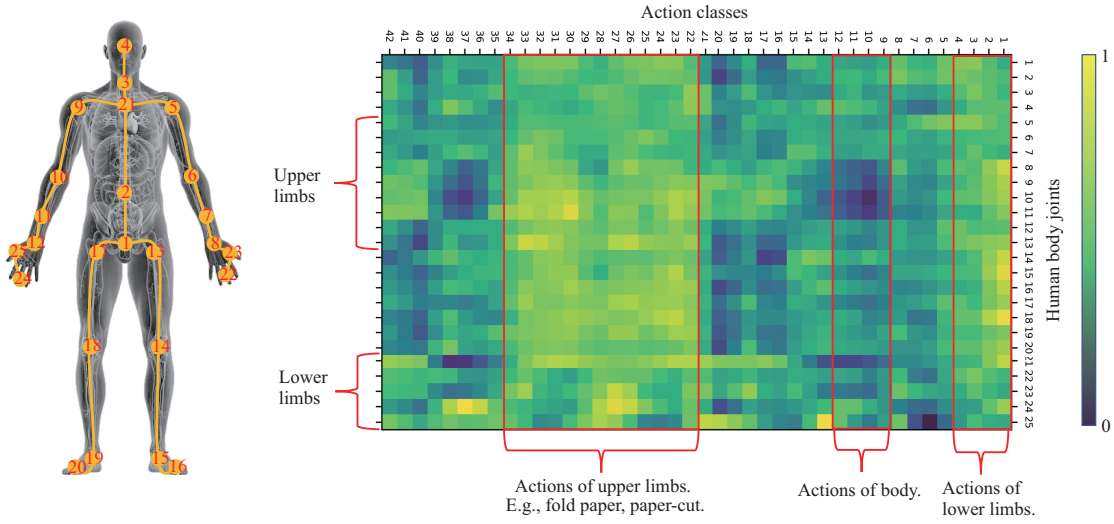

图 7 细粒度动作的结点门控可视化

Figure 7 Visualization of the joint gate of fine-grained actions

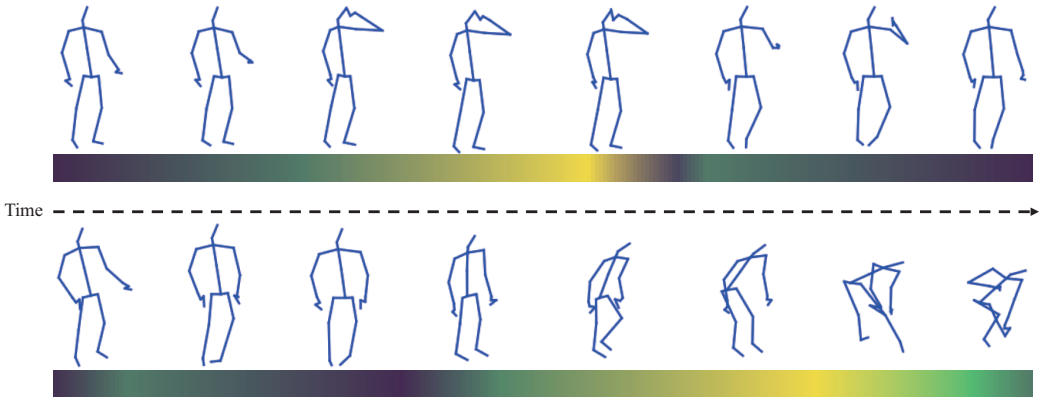

图 8 细粒度动作的时间门控可视化

Figure 8 Visualization of the time gate in fine-grained actions

具有不同的显著运动信息, 但噪声信息和冗余信息应该相似. 此外, 这些结果还显示, 我们的 MSP-PL 框架可以从样本对中学习关键的运动区域, 增强显著的运动信息, 并消除噪声信息.

门控向量的生成策略. 对于门控向量的生成策略, 表 3 中的实验结果表明, 特征拼接 (Concat) 是一种比元素相加 (Sum) 更合适的融合方法, 而 Sigmoid 函数是一种最佳激活函数. 因此, 在 MSP 中, 利用 Concat-Conv-Pool 操作和 Sigmoid 函数来生成门控向量 (见图 3)

三种门控向量对性能的影响. 我们在实验中测试了结点门控、时间门控和通道门控对模型性能的影响. 表 4 中的结果显示, 各种门控向量都对模型的细粒度动作识别有帮助. 其中结点门控向量所带来的提升最显著, 模型 top-1 的准确率提高了 1.07% (72.86% vs. 71.79%). 骨架运动序列的结点维度包含丰富的精细化信息, 使网络聚焦于人体的显著运动结点, 同时增强显著的结点特征是区分细粒度骨架动作的关键. 在 MSP 中, 同时使用三种门控来充分增强查询样本的运动特征.

细粒度动作的结点与时间门控可视化. 图 7 展示了 MSP 模块的结点门控在 NTU-FG 数据集中

表 5 目前主流的骨架动作识别模型在 NTU-RGB+D 120 X-Set 基准测试中, 简单动作类别和细粒度动作类别的 top-1 准确率比较

Table 5 Comparison of the top-1 accuracy on the NTU-RGB+D 120 X-Set benchmark for the easy samples and the fine-grained samples. Only the accuracy of the joint stream is reported

Method	Easy samples (%)	Fine-grained samples (%)	Gaps (%)
ST-GCN [12]	79.98	53.04	↓26.94
AGCN [19]	89.64	68.37	↓21.27
MS-G3D [22]	91.05	72.02	↓19.03

表 6 在不同的边界阈值系数 ϵ 条件下, NTU-FG X-Set 基准的 top-1 和 top-5 准确率比较

Table 6 Comparison of the top-1 and top-5 accuracy on the NTU-FG X-Set benchmark with different ϵ a)

The value of ϵ	Top-1 (%)	Top-5 (%)
0.1	73.18	93.59
0.05	**73.79**	**93.76**
0.01	72.38	93.15

a) Bold indicates the optimal results.

人体 25 个结点对于 42 个细粒度动作类别的激活图, 该激活图对结点维度进行归一化 (即行归一化). 从图中可以看到, 人体的上肢动作, 例如折纸、剪纸等对上肢结点的激活值较高. 相反, 例如头痛、胸部疼痛、颈部疼痛等动作, 模型更加关注其躯干的区域来区分这些细粒度运动. 该结果证明, SMP-PL 框架中的结点门控能够辅助模型捕捉到该运动显著的区域, 从而提升模型对于这些易混淆动作的识别能力.

图 8 展示了 MSP 模块在 NTU-FG 数据集中, 两个动作样本的时间门控激活图. 从图中可以看到, 时间门控可以辅助模型捕捉到区分度较强的人体运动的关键时间段, 弱化其余区分度较弱的冗余时间段对细粒度动作分类所造成的影响. 该实验证明本文提出的 SMP-PL 框架对于时序信息的敏感性高, 提取能力强, 能够有效辅助模型捕捉到运动的显著时间段.

简单动作类别与细粒度动作类别之间的精度差距. 表 5 报告了现有主流骨架动作识别模型在简单动作类别和细粒度动作类别之间的精度差距. 这些方法在简单动作样本和困难 (细粒度) 动作样本之间有巨大的准确性差距 (约 20%). 因此, 细粒度骨架动作的低分类精度是目前骨架动作识别领域的一个突出的问题. 针对此问题, 上述实验证明, 本文提出的 MSP-PL 框架可以有效提升现有模型在细粒度骨架动作识别上的准确性.

边界阈值系数 ϵ 的影响. 对于 Loss_{RK} 中的边界阈值系数 ϵ 的取值对模型精度的影响, 我们进行了实验分析, 表 6 的实验结果显示, 当 ϵ 在 0.05 左右时, 它对最终识别精度的影响变得稳定, 且达到最优.

混淆矩阵. 图 9 展示了 MSP-PL (MS-G3D) 在 NTU-FG 数据集上各类别识别精度的混淆矩阵. 可以看到, MS-G3D 编码器与 MSP-PL 框架结合后取得了较高的分类准确性, 但个别细粒度动作类别依旧较难区分, 例如 "15. 拇指向上" 与 "16. 拇指向下". 其原因是目前的人体骨架动作数据集并未精确捕捉手指运动, 这也是骨架动作识别的一个值得研究的问题.

4.4 对比实验

我们将本文提出的方法与目前最先进的骨架动作识别方法在 NTU-120 [16] 的完整数据集上进行比

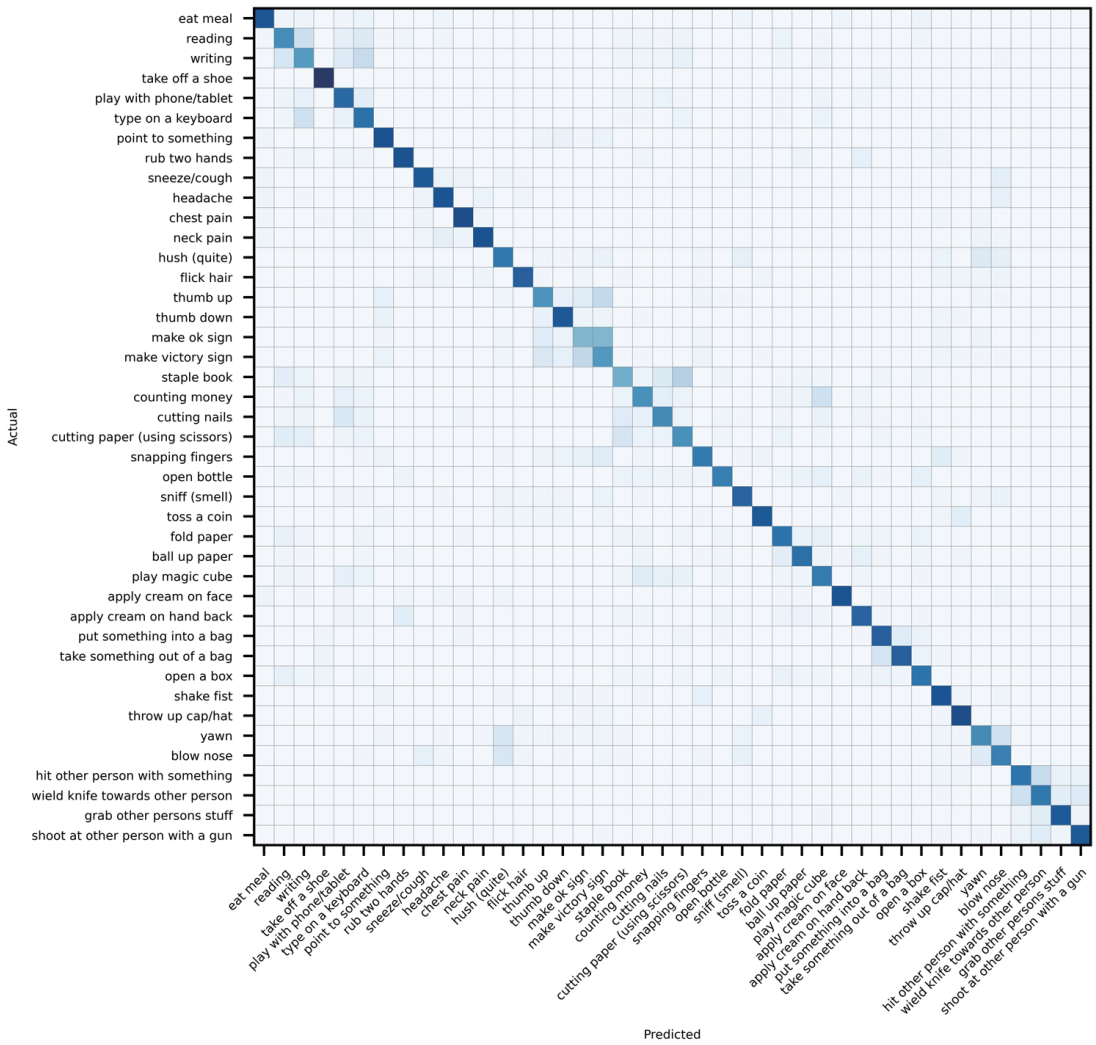

图 9 MSP-PL (MS-G3D) 在 NTU-FG X-Set 数据集上各类别的混淆矩阵

Figure 9 Confusion matrix of our MSP-PL with the MS-G3D encoder on the NTU-FG X-Set benchmark

较. 为与现有方法公平对比, 表 7 中报告的是结点 – 骨骼双流融合模型的结果. 当与最优的编码器 MS-G3D 相结合时, 本文的 MSP-PL 框架在这两个基准上都取得了最佳性能. 如表 7 所示, 本文的 MSP-PL 框架在 X-Sub 基准和 X-Set 基准上将 MS-G3D 的准确性分别提高了 1.1% 和 0.9%. 与之前的精细化特征提取器 STATT-MS 相比, 本文的 MSP-PL 框架结合 MS-G3D 编码器在 X-Sub 基准和 X-Set 基准上的准确率分别高出了 3.9% 和 4.0%. 值得注意的是, 原始 MS-G3D 的准确性低于 DualHead-Net, 但在将 MS-G3D 嵌入本文的 MSP-PL 框架后, 其准确度超过了 DualHead-Net (在 X-Set 基准上为 89.3% 对 89.1%). 另外, 这些精细化的特征提取器 (例如 STATT-MS 与 DualHead-Net) 也可以被嵌入到本文的 MSP-PL 框架中, 以进一步学习细粒度的运动特征.

在 Kinetics-Skeleton 数据集上, 如表 8 所示, 当与最优的编码器 MS-G3D 相结合时, 我们的 MSP-PL 框架也取得了领先的精度. 相比于基准模型 ST-GCN, 本文的 top-1 精度与 top-5 精度分别提升了 7.7% 和 8.8%. 这些结果表明, 提取精细化的运动特征有助于识别骨架动作, 而本文的 MSP-PL 框

表 7　在 NTU-120 数据集上, 我们的模型与现有方法的 top-1 的准确度对比

Table 7　Comparison of the top-1 accuracy with state-of-the-arts on the NTU-120 dataset[a)]

Method	X-Sub (%)	X-Set (%)
ST-GCN [12]	70.7	73.2
2s-AGCN [19]	82.9	84.9
MS-G3D [22]	86.9	88.4
ST-TR [52]	85.1	87.1
CTR-GCN (Paper) [31]	88.9	90.6
CTR-GCN (Code) [31]	87.1	88.6
*STATT-MS [15]	84.1	85.3
*DualHead-Net [21]	87.9	89.1
*MSP-PL (AGCN)	84.3	86.6
*MSP-PL (MS-G3D)	**88.0**	**89.3**

a) * means this method is especially designed for fine-grained motion feature extraction. Bold indicates the optimal results.

表 8　在 Kinetics-Skeleton 数据集上, 我们的模型与现有方法的 top-1 和 top-5 的准确度对比

Table 8　Comparison of the top-1 and top-5 accuracy with state-of-the-arts on the Kinetics-Skeleton dataset[a)]

Method	Top-1 (%)	Top-5 (%)
Deep LSTM (2016) [16]	16.4	35.3
ST-GCN (2018) [12]	30.7	52.8
2s-AGCN (2019) [19]	35.9	58.6
ST-TR [52]	37.0	59.7
CTR-GCN [31]	38.0	60.9
MSP-PL (MS-G3D)	**38.4**	**61.6**

a) Bold indicates the optimal results.

架不仅能有效地识别 3D 表示的细粒度的骨架动作, 而且还对 3D 以及 2D 表示的, 包含有大量噪声数据的骨架动作识别具有普适性. 因此, 本文的 MSP-PL 框架具有较强的实用性.

5　结论

本文提出了一种新型的配对学习框架 MSP-PL, 用于细粒度的骨架动作识别, 其具有较高的实际应用价值且在以前的相关工作中并未被充分研究. 运动显著性探测器 (MSP) 是本文 MSP-PL 框架的核心模块, 该模块可以从人体关键运动结点、重点运动时段以及局部特征通道准确地增强细粒度骨架动作的显著性运动线索, 通过相似样本对比消除冗余信息, 并有效地促进编码器提取具有区分能力的骨架运动特征. 在门控对比损失和配对排序损失的帮助下, 本文的 MSP-PL 框架可以驱动编码器充分探索查询 – 探测样本对之间的相似性与差异性, 并学习丰富的精细化运动特征. 实验证明, 大多数流行的骨架特征编码器可以直接嵌入到本文的 MSP-PL 框架中, 以显著提高细粒度动作分类的性能. 此外, 本文研究并讨论了多种探测样本构造策略, 其中基于最小距离的选择策略最为有效. 最后, 当 MSP-PL 框架与先进的编码器相结合时, 其在 NTU-RGB+D 120 数据集与 Kinetics-Skeleton 数据集上进行的骨架动作识别达到了目前最佳的性能.

参考文献

1 Zhang Z. Microsoft Kinect sensor and its effect. IEEE Multimedia, 2012, 19: 4–10
2 Cao Z, Hidalgo G, Simon T, et al. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 172–186
3 Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7103–7112
4 Zheng C, Zhu S, Mendieta M, et al. 3D human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 11656–11665
5 Zhang J, Tu Z, Yang J, et al. MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 13232–13242
6 Carfi A, Mastrogiovanni F. Gesture-based human-machine interaction: taxonomy, problem definition, and analysis. IEEE Trans Cybern, 2023, 53: 497–513
7 Xu F, Xu F, Xie J, et al. Action recognition framework in traffic scene for autonomous driving system. IEEE Trans Intell Transp Syst, 2022, 23: 22301–22311
8 Gupta P, Thatipelli A, Aggarwal A, et al. Quo Vadis, skeleton action recognition? Int J Comput Vis, 2021, 129: 2097–2112
9 Zhang J, Jia Y, Xie W, et al. Zoom transformer for skeleton-based group activity recognition. IEEE Trans Circuits Syst Video Technol, 2022, 32: 8646–8659
10 Song S, Lan C, Xing J, et al. Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. IEEE Trans Image Process, 2018, 27: 3459–3471
11 Cao C, Lan C, Zhang Y, et al. Skeleton-based action recognition with gated convolutional neural networks. IEEE Trans Circuits Syst Video Technol, 2018, 29: 3247–3257
12 Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the 32nd AAAI Conference On Artificial Intelligence, 2018. 7444–7452
13 Si C, Chen W, Wang W, et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In: Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition, 2019. 1227–1236
14 Chen S, Xu K, Mi Z, et al. Dual-domain graph convolutional networks for skeleton-based action recognition. Mach Learn, 2022, 111: 2381–2406
15 Li X, Liu S, Li Y, et al. Spatial-temporal attention network with multi-similarity loss for fine-grained skeleton-based action recognition. In: Proceedings of International Conference on Neural Information Processing, 2021. 620–631
16 Liu J, Shahroudy A, Perez M, et al. NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding. IEEE Trans Pattern Anal Mach Intell, 2019, 42: 2684–2701
17 Zhang J, Ye G, Tu Z, et al. A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. CAAI Trans Intel Tech, 2020, 7: 46–55
18 Zhang X, Xu C, Tao D. Context aware graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 14333–14342
19 Shi L, Zhang Y, Cheng J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 12026–12035
20 Peng W, Hong X, Zhao G. Tripool: graph triplet pooling for 3D skeleton-based action recognition. Pattern Recognition, 2021, 115: 107921
21 Chen T, Zhou D, Wang J, et al. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021. 4334–4342
22 Liu Z, Zhang H, Chen Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 143–152
23 Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3D skeletons as points in a lie group. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. 588–595
24 Vemulapalli R, Chellapa R. Rolling rotations for recognizing human actions from 3D skeletal data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 4471–4479
25 Shan Y, Zhang Z, Huang K. Learning skeleton stream patterns with slow feature analysis for action recognition.

In: Proceedings of European Conference on Computer Vision, 2014. 111–121
26 Zhang S, Yang Y, Xiao J, et al. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks. IEEE Trans Multimedia, 2018, 20: 2330–2343
27 Wang H, Wang L. Beyond joints: learning representations from primitive geometries for skeleton-based action recognition and detection. IEEE Trans Image Process, 2018, 27: 4382–4394
28 Zhu K, Wang R, Zhao Q, et al. A cuboid CNN model with an attention mechanism for skeleton-based action recognition. IEEE Trans Multimedia, 2019, 22: 2977–2989
29 Zhang J, Xie W, Wang C, et al. Graph-aware transformer for skeleton-based action recognition. Vis Comput, 2023, 39: 4501–4512
30 Zhang P, Lan C, Zeng W, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 1112–1121
31 Chen Y, Zhang Z, Yuan C, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 13359–13368
32 Ji R, Wen L, Zhang L, et al. Attention convolutional binary neural tree for fine-grained visual categorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 10468–10477
33 Sun M, Yuan Y, Zhou F, et al. Multi-attention multi-class constraint for fine-grained image recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 805–821
34 Peng Y, He X, Zhao J. Object-part attention model for fine-grained image classification. IEEE Trans Image Process, 2017, 27: 1487–1500
35 Cai S, Zuo W, Zhang L. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 511–520
36 Fu J, Zheng H, Mei T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4438–4446
37 Zhuang P, Wang Y, Qiao Y. Learning attentive pairwise interaction for fine-grained classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 13130–13137
38 Zhang C, Gupta A, Zisserman A. Temporal query networks for fine-grained video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 4486–4496
39 Qiu S, Anwar S, Barnes N. Geometric back-projection network for point cloud classification. IEEE Trans Multimedia, 2022, 24: 1943–1955
40 Yu F, Liu K, Zhang Y, et al. PartNet: a recursive part decomposition network for fine-grained and hierarchical shape segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 9491–9500
41 Singh B, Marks T, Jones M, et al. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1961–1970
42 Ying Y, Wen L, Lyu S. Stochastic online AUC maximization. In: Proceedings of Advances in Neural Information Processing Systems, 2016
43 Sohn K. Improved deep metric learning with multi-class N-pair loss objective. In: Proceedings of Advances in Neural Information Processing Systems, 2016
44 Zhang X, Zhou F, Lin Y, et al. Embedding label structures for fine-grained feature representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1114–1123
45 Dubey A, Gupta O, Guo P, et al. Pairwise confusion for fine-grained visual classification. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 70–86
46 Yeh C, Hong C, Hsu Y, et al. Decoupled contrastive learning. In: Proceedings of the 17th European Conference Computer Vision, Tel Aviv, 2022. 668–684
47 Wang H, Guo X, Deng Z, et al. Rethinking minimal sufficient representation in contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 16041–16050
48 Yang J, Li C, Zhang P, et al. Unified contrastive learning in image-text-label space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 19163–19173

49 Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset. 2017. ArXiv:1705.06950
50 Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of Advances in Neural Information Processing Systems, 2019
51 Li C, Zhong Q, Xie D, et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018. 786–792
52 Plizzari C, Cannici M, Matteucci M. Skeleton-based action recognition via spatial and temporal transformer networks. Comput Vision Image Understanding, 2021, 208–209: 103219

Fine-grained skeleton action recognition with pairwise motion salience learning

Hongyan LI[1,2†], Zhigang TU[2†], Wei XIE[3] & Jiaxu ZHANG[2*]

1. *School of Information Engineering, Hubei University of Economics, Wuhan* 430205*, China*;
2. *State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan* 430072*, China*;
3. *School of Computer Science, Central China Normal University, Wuhan* 430079*, China*
* Corresponding author. E-mail: zjiaxu@whu.edu.cn
† Equal contribution

Abstract Fine-grained skeleton-based action recognition is a crucial but less explored topic. Due to the lack of appearance context, skeleton-based actions from similar categories are not easily recognized. In this work, we propose a novel motion salience prober-incorporated pairwise-learning (MSP-PL) framework to cope with fine-grained skeleton action recognition. Our MSP-PL framework applies motion salience learning on each similar skeleton sample pair to force the encoders to learn fine-grained motion features. The core module MSP enhances the salient motion clues and eliminates the noise of the query samples with the help of the designed probe samples and losses. We exploit three types of probe sample construction strategies to generate the sample pairs and test their impact on the model performance. Extensive experiments on the NTU-RGB+D 120 dataset demonstrate that our MSP-PL framework is general so that most skeleton feature encoders can be seamlessly embedded into it for a considerable boost in accuracy. The average accuracy improvement on the fine-grained action classification of five popular encoders is over 2.4%. In addition, our MSP-PL framework achieves state-of-the-art performance on skeleton action recognition when combined with the advanced encoder.

Keywords skeleton action recognition, fine-grained action recognition, visual attention, motion salience learning, contrastive learning