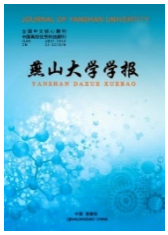

燕山大学学报
Journal of Yanshan University
ISSN 1007-791X,CN 13-1219/N

《燕山大学学报》网络首发论文

题目：基于跨尺度图对比学习的人体骨架动作识别方法
作者：张雪莲，徐增敏，陈家昆，王露露
收稿日期：2022-06-16
网络首发日期：2023-03-27
引用格式：张雪莲，徐增敏，陈家昆，王露露．基于跨尺度图对比学习的人体骨架动作识别方法[J/OL]．燕山大学学报．
https://kns.cnki.net/kcms/detail/13.1219.N.20230324.1620.012.html

燕山大学学报
Journal of Yanshan University

基于跨尺度图对比学习的人体骨架动作识别方法

张雪莲[1]，徐增敏[1,2,*]，陈家昆[1]，王露露[1]

（1. 桂林电子科技大学 数学与计算科学学院，广西 桂林 541004；
2. 桂林电子科技大学 广西高校数据分析与计算重点实验室，广西 桂林 541004；
3. 桂林安维科技有限公司，广西 桂林 541010）

摘　要：传统基于人体骨架的自监督学习方法常用对比学习模块进行表征学习，而现有对比学习模块使用数据增强方法来构建相似的正样本，其余样本皆为负样本，这限制了同类样本的语义信息表达。针对上述问题，提出一种图对比学习与跨尺度一致性知识挖掘的动作识别算法。首先，基于骨架图结构设计了一种新的数据增强方法，对输入的骨架序列进行随机边裁剪，得到两个不同的扩增视图，加强了同一骨架序列不同视图间的语义相关性表达；其次，为缓解同类样本嵌入相似度较低的问题，引入自监督协同训练网络模型，利用同一骨架数据源的不同尺度间的互补信息，从一个骨架尺度获取另一个骨架尺度的正类样本，实现了单尺度内关联及多尺度间语义协同交互；最后，基于线性评估协议对模型效果进行评估，在 NTURGB+D60 与 NTURGB+D120 数据集的实验结果表明，本文所提方法在识别精度上较前沿主流方法平均提升了 2%~3.5%。

关键词：图对比学习；数据增强；跨尺度一致性知识挖掘；协同训练；人体骨架

中图分类号：TP391　　**文献标识码**：A

0　引言

人体动作识别是目前计算机视觉领域非常热门的研究方向，它主要从视频片段中分辨不同类的动作，然后对视频的多帧图像进行处理，并利用全连接层来获得最终的分类结果[1]。作为该领域的研究热点，动作识别在视频监控、人体交互、视频理解等领域[2-3]发挥重要作用。

在过去的工作中，许多基于 RGB 视频的动作识别技术已经取得了显著成果，但在提取 RGB 视频数据时，其易受到遮挡、环境变化与阴影干扰，导致深度图中颜色和纹理特征容易缺失，且处理起来相对耗时。另一种模态数据，人体骨架数据集，利用骨骼关节点的三维坐标来表示人体，实现了一种更加轻量级的表示方法，且骨架数据对于视角变换、人物外貌以及环境变化具有较强的鲁棒性。因此，近年来，基于骨架数据的人体动作识别方法得到了广泛关注，Yan 等人[4]提出一种时空图卷积网络模型（ST-GCN），更好地表述了人体骨骼关节之间的依赖关系；Lei 等人[5]提出了一种双流自适应图卷积网络（2S-AGCN），更加合理地构建了邻接矩阵策略，增强了网络对空间特征的抽取能力；Liu 等人[6]提出了一种多尺度时空聚合方案（MS-G3D），有效地解决有偏加权问题。以上构建的模型虽然取得了较好的识别效果，但都属于全监督学习框架，需要依赖大量人工标注数据，而标注数据是繁琐且昂贵的。

针对以上问题，自监督学习被广泛应用，其无需标注训练样本，可以通过数据增强方法低成本扩充数据集，凭借这一优势，越来越多的研究人员

收稿日期：2022-06-16

基金项目：国家自然科学基金资助项目（61862015）；广西科技基地和人才专项资助项目（AD21220114）；广西重点研发计划资助项目（AB17195025）

作者简介：张雪莲（1997-），女，黑龙江绥化人，硕士研究生，主要研究方向为自监督学习、应用数学；* 通信作者：徐增敏（1981-），男，广西梧州人，博士，副教授，主要研究方向为计算机视觉、人工智能，Email：xzm@guet.edu.cn。

将目光投入到自监督模型构建中。其中,Lin 等人[7]提出一种基于骨架的自监督动作识别方法,可以使编码器学习更多的鉴别特性,解决从单个重建任务中学习骨架表示的过拟合问题;Zheng 等人[8]通过结合编码器,解码器和生成式对抗网络,重新构建了被掩码的 3D 骨架序列;Yang 等人[9]设计了一种骨骼云着色技术,将从未标记的骨架序列中学习到的特征表示用于骨架动作识别的自监督表示方法中。然而,以上基于骨架数据的自监督模型,利用对比学习方法进行建模,没有考虑骨架数据是一种离散数据结构,需要进行图结构学习,且利用数据增强获取正样本的想法过于单一,较少将跨尺度信息联合方法应用到自监督模型中,难以克服单一尺度特征信息不足的缺陷,不利于模型聚类效果。

鉴于此,本文提出基于图对比学习与跨尺度一致性知识挖掘的自监督动作识别方法。所提方法首先结合多种数据增强理论,以获得无标签骨架序列的不同视图,并对不同视图进行编码,建立图对比学习网络;其次将原始骨架序列转化为多尺度骨架图序列,结合跨尺度一致性知识挖掘模块,构建基于骨架的跨尺度图对比学习网络;最后将多尺度骨架图序列输入到所构建的网络模型中,通过个体多尺度映射间的协同关联模式,实现单尺度内关联及多尺度间语义协同交互。

基于以上所述,本文所作贡献可简述如下:

1) 为解决传统方法在扩增骨架数据过程中,存在泛化性不足和传递性不强的问题,融合图数据增强思想,建立图对比自监督动作识别网络。

2) 引入多尺度图来建模三维骨骼特征表示,聚集骨骼关节点的关键相关特征,结合跨尺度一致性知识挖掘方法,实现多尺度信息间的交互。

3) 结合图对比自监督动作识别网络和跨尺度一致性知识挖掘方法,提出一种新的模型框架,并基于线性评估协议对模型效果进行评估。

1 相关工作

1.1 基于骨架的监督动作识别方法

基于骨架的监督动作识别方法旨在从一系列时间连续及有标签的人体骨架序列中识别正在执行的动作[10]。早期人体骨架动作识别算法大多是基于手工特征。近年来,随着机器学习与深度学习的发展,人们将其与骨架序列联系起来,提出许多基于循环神经网络[11-12]的方法,虽然有效地利用了骨架序列的时序信息,但考虑到循环神经网络存在梯度消失等问题,研究者们逐渐将目光转移到卷积神经网络[13-14]上,其可以从不同时间区间内提取到骨架特征的特定局部模式,然而鉴于该网络需要将骨架序列转换成特定的 RGB 图像形式,不利于骨架数据的特征表达,人们又提出了图卷积神经网络[4,15],通过建模骨架数据的自身图结构,进而实现基于骨骼点的动作识别任务。本文受前人启发,采用基于图卷积网络方法,将 ST-GCN[4]作为提取骨架特征的主要组成网络。

1.2 基于自监督对比学习的动作识别方法

对比学习方法着重于学习同类实例之间的共同之处,区分非同类之间的不同之处[16]。最近,研究人员提出许多基于生成实例的自监督对比学习方法[17-20]。其中,MoCo[17]模型建立一个动态字典,用动量对比的学习方法做自监督的表征学习任务,SimCLR[18]模型通过去除存储库(memory bank),简化了 MoCo[17]模型提出的自监督对比学习算法,SimSiam[19]模型通过最大化同一样本不同视图间的相似度,来解决自监督对比学习中出现崩溃解(collapsing solutions)的问题。与本文相似的工作 CoCLR[20]模型是基于 RGB 视频数据与光流数据进行的跨模态自监督行为识别,相对骨架数据,提取 RGB 视频数据与光流数据需要较长的时间,往往导致其复杂度过高。

1.3 基于骨架的自监督动作识别方法

自监督学习是指从大规模未标记数据中学习自身语义信息,为模型及算法提供监督信息。研究人员探索各种模型构建策略,如拼图[21-22]、着色[23]、预测和修复掩码词[24-26]等。相比图像和 RGB 视频,基于骨架数据的用于人体动作识别的自监督学习仍然是一个较新的、值得被关注的问题。其中,MS^2L[7]模型提出一种基于骨架序列表示的多任务自监督学习方法,可以同时解决多个

辅助任务,例如运动预测和骨架拼图等,AS-CAL[27]模型利用未标记骨架序列的不同扩增视图,以自监督对比学习的方式来学习动作表示,AimCLR[28]模型探索极端数据增强带来的不同运动模式,缓解正样本选取的不合理性。以上工作积极探索基于3D骨骼的自监督学习方法,并从无标记骨架数据中学习到有效的动作表示。

1.4　多尺度骨架图

文献[29-30]通过构建不同骨架视图,例如:关节、运动、骨骼等,利用不同视图间的特征相似性,学习丰富的内部监督信息,并将其作为描述身体结构和运动的判别特征。然而,在建立不同骨架视图过程中,往往只从骨架的单一尺度空间中提取这些特征,这将限制从不同身体分区中捕获高层结构信息的能力。例如CrosSCLR[31]模型是基于骨架的跨视图对比学习,DMGNN[32]模型从单一空间尺度和拓扑结构的骨架中提取特征,PoseGait[33]模型将人体关节运动轨迹和预定义的姿态描述符编码为特征向量。本文是在SM-SGE[34]模型的启发下,充分挖掘了身体组成部分中潜在的结构特征,利用一种跨尺度一致性知识挖掘的方法来表达不同层次的骨架结构信息,并结合协同训练,构建跨尺度图对比学习网络模型。

2　跨尺度图对比学习

虽然3D骨架数据在动作识别领域起着至关重要的作用,但在自监督骨架表示方面尚未得到长足发展。数据增强作为对比学习的先决条件,影响着网络模型的最终拟合效果,如何构建适合骨架数据的扩增方法成为本章的研究重点。骨架图是由一系列的骨骼关节点相连组成,通过改变骨架图结构,可以更好地学习骨架的高级语义信息,且包含骨架信息的多尺度图较易获取。因此,本章利用图对比学习方法与多尺度特征间的语义相关性,结合协同训练,构建基于骨架动作表示的跨尺度图对比学习框架。本文主要包括两个关键模块:1) SGCLR:一个用于单尺度自监督学习表示的图对比学习框架(Graph Contrastive Learning for Skeleton-based action Representation, SGCLR);2) CrosScale-SGCLR:该算法将一个尺度的特征信息传递给另一个尺度,通过引入互补的伪标签约束,促进多尺度特征间的信息共享(Cross-Scale Graph Contrastive Learning framework for Skeleton-based action Representation, CrosScale-SGCLR)。

2.1　SGCLR算法

给定一个包含 l 帧连续的3D骨架序列 $\boldsymbol{X}=(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_l)$,其中 $\boldsymbol{X}_i\in\mathbf{R}^{W\times J\times D}$,$W$ 为人的总数,J 为骨骼关节点数,D 为位置向量维度($\boldsymbol{X}_i$ 的位置向量维度为3)。训练集 $\boldsymbol{\Phi}=\{\boldsymbol{X}_i\}_{i=1}^{B}$ 包含了从多个视图和多个人中采集的 B 种不同动作的骨架序列。每个骨架序列 $\boldsymbol{X}_i$ 对应一个标签 y_i,其中 $y_i\in\{a_1,\cdots,a_c\}$,a_i 表示第 i 种动作类别,c 表示动作类别的总个数,每次输入网络中的样本数据批量大小(batch size)为 N。不同于SkeletonCLR[31]模型利用对比学习建模的方法,本节方法虽然同样使用了该方法的基本组成框架来构建网络模型,但在此基础上融合了图对比学习方法,在数据扩增上进行了相应改进,使得同一样本扩增后得到的两个实例具有不同的邻接矩阵。

2.1.1　图对比学习

本文在GraphCL[35]与SimGRACE[36]模型的启发下,为解决传统基于对比学习的动作识别算法在扩增骨架数据过程中,存在泛化性不足和传递性不强的问题,融合图数据增强思想,提出一种基于图对比学习的人体骨架动作识别算法。该算法基于人体骨架数据自身的图结构关系,分别在双路径中处理输入骨架序列,即原路径与图对比路径,两条路径使用相同的自编码图卷积神经网络,将人体骨架数据的不同数据增强得到的实例作为正样本,将存储库中的其他人体骨架序列视为负样本。在每次训练过程中,构成负样本的张量坚持先进先出原则,不断更新存储库中的批量嵌入信息,并利用图对比损失函数训练模型参数,以拉近正样本的距离,远离负样本的距离。SGCLR的总体架构如图1所示。

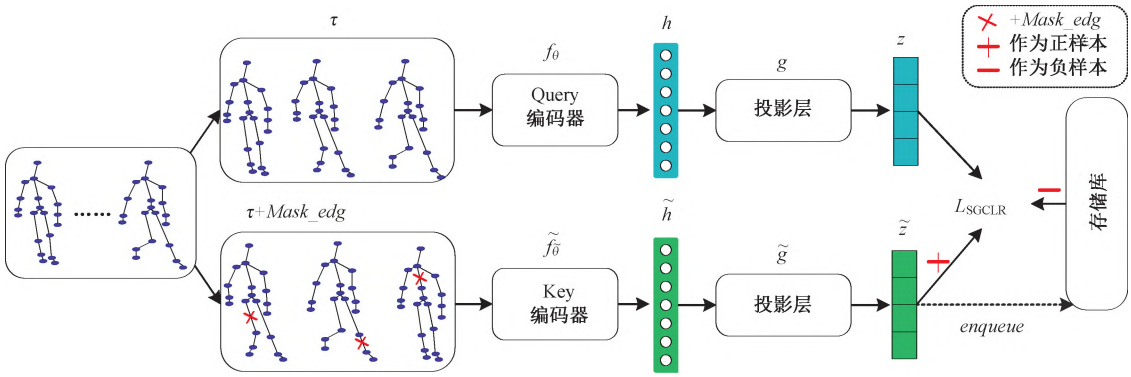

图 1　单尺度 SGCLR 的算法结构图

Fig. 1　Architecture of single-scale SGCLR

2.1.2　算法的实现

2.1.1 节介绍了本节中所提方法的原理和结构,主要步骤如下所示:

1) 获取 3D 骨架序列 $\boldsymbol{X}$,该张量维度为 $[N, D, l, J, W]$,为避免数据冗余和降低计算复杂度,在人体骨架数据集中,统一将骨架序列的连续帧数 l 取为 50,批量大小 $N=128$,位置向量维度 $D=3$,骨骼关节点数 $J=25$,人的总数 $W=2$。

2) 利用数据增强模块 τ 和 $\tau+mask\_edg$ 来获取不同实例 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$,使其作为正样本集,主要步骤如下所示:

① 在步骤 1)获得无标签骨架数据的基础上,分别在原路径与图对比路经上引入剪切(shear)与时序裁剪(temporal crop)的数据增强方法,以得到不同视图 $\boldsymbol{Q}$ 和 $\tilde{\boldsymbol{x}}$,具体方法概述如下。

剪切:剪切变换是通过构建相应的仿射矩阵,使人体关节的三维坐标形状呈任意角度倾斜。仿射矩阵的公式为

$$\boldsymbol{R}_{t,\beta}^{y}=\begin{bmatrix}1 & s_x^y & s_x^z\\ s_y^x & 1 & s_y^z\\ s_z^x & s_z^y & 1\end{bmatrix}, \tag{1}$$

其中,$s_x^y, s_x^z, s_y^x, s_y^z, s_z^x, s_z^y$ 是 6 个错切因子,取值范围在 $-\beta$ 到 β 之间。

时序裁剪是在时间维度上的数据增强,它将一些帧对称地填充到序列中,然后随机地将其裁剪到原始长度。填充长度定义为 l/r,r 为填充比(取值为正整数)。

② 接下来对视图 $\tilde{\boldsymbol{x}}$ 进行随机边裁剪($mask\_edg$),得到不同的图表示向量 $\boldsymbol{K}$。具体思想:利用随机掩码$[0\sim\xi]$,去掉关节点间的连接边,形成新的骨架图结构。

3) 将不同实例 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 分别嵌入到编码器 f_θ 和 $\tilde{f}_{\tilde{\theta}}$ 中,得到编码特征 $\boldsymbol{h}$ 与 $\tilde{\boldsymbol{h}}$,其中 θ 与 $\tilde{\theta}$ 为两个编码器所需参数,$\tilde{\theta}$ 遵循动量更新:$\tilde{\theta}\leftarrow\tilde{\alpha}\theta+(1-\alpha)\theta$,$\alpha$ 为动量系数,$\boldsymbol{h}, \tilde{\boldsymbol{h}}\in\mathbf{R}^{C_h}$,SGCLR 使用 ST-GCN[4] 作为编码器网络。

4) 将得到的编码特征 $\boldsymbol{h}$ 与 $\tilde{\boldsymbol{h}}$ 分别输入到投影层 g 和 $\tilde{g}$ 中,获得较低维空间向量:$\boldsymbol{z}=g(\boldsymbol{h})$,$\tilde{\boldsymbol{z}}=\tilde{g}(\tilde{\boldsymbol{h}})$,其中 $\boldsymbol{z}, \tilde{\boldsymbol{z}}\in\mathbf{R}^{C_z}$。投影层是由一个全连接(FC)层与线性(ReLU)层组成。

5) 存储库 $\boldsymbol{M}=\{\boldsymbol{M}_i\}_{i=1}^{M}$ 中储存大量的负样本,避免了嵌入的冗余计算。它是一个先进先出队列,每次迭代时由 $\tilde{\boldsymbol{z}}$ 更新。具体来看,在每次更新迭代之后,$\tilde{\boldsymbol{z}}$ 将进入队列成为新的负样本,而早期嵌入 $\boldsymbol{M}$ 中的实例将退出队列。

6) 在图对比学习过程中,当一个骨架序列以不同实例输入到两条不同的路径中时,其输出的特征是相似的,本文将 InfoNCE 损失函数作为图对比学习的损失函数,公式如下:

$$L_{SGCLR}=-\log\frac{\exp(\boldsymbol{z}\cdot\tilde{\boldsymbol{z}}/t)}{\exp(\boldsymbol{z}\cdot\tilde{\boldsymbol{z}}/t)+\sum_{i=1}^{M}\exp(\boldsymbol{z}\cdot\boldsymbol{M}_i/t)}, \tag{2}$$

式中,$\boldsymbol{M}_i\in\boldsymbol{M}$ 为存储库中的负样本集,t 是超参数,$\boldsymbol{z}\cdot\tilde{\boldsymbol{z}}$ 表示两个向量的点积,其结果表明两个实例间的相似程度,其中 $\boldsymbol{z}$ 与 $\tilde{\boldsymbol{z}}$ 已被归一化。

在图对比损失 L_{SGCLR} 的约束下,对自监督网络模型进行训练,以区分训练集中的每个样本实例,最后通过线性评估方法验证该模型的有效性。SGCLR 方法的伪代码如算法 1 所示。

算法 1　SGCLR 方法的伪代码

输入:批量输入骨骼点的坐标数据矩阵 $\boldsymbol{X}[N, D, l, J, W]$ 数据增强参数 β,r 和 ξ,动量更新参数 α

超参数 t,f_q $=f_\theta$,f_k $=\tilde{f}_{\tilde{\theta}}$,X_q $=\boldsymbol{Q}$,X_k $=\boldsymbol{K}$

输出:N 维损失值向量

编码器函数参数更新

```
f_k.params = f_q.params #初始化编码器函数
for X in loader: #加载带有 N 个样本的 minibatch X
    X_q = aug(X) #使用随机数据增强 τ,根据 2.1.2 节 2)①
    X_k = aug(X) #使用随机数据增强 τ,根据 2.1.2 节 2)①
    X_k = mask_edg(X_k)
    #随机边裁剪 mask_edg,根据 2.1.2 2)②
q = f_q.forward(X_q) #得到编码特征 h: N×C
q = F.normalize(q, dim=1) #正则化编码特征 h: N×C
k = f_k.forward(X_k) #得到编码特征 h̃: N×C
k = F.normalize(k, dim=1) #正则化编码特征 h̃: N×C
k = k.detach( ) #编码特征 h̃ 不进行梯度计算
l_pos = einsum('nc,nc->n', [q, k]).unsqueeze(-1)
#正样本 logits: N×1
l_neg = einsum('nc,ck->nk', [q, self.queue.clone( ).detach( )])
#负样本 logits: N×K
# queue:储存负样本的队列,采用逐步更新方法(N×K)
# logits:N×(1+K)
logits = cat ([l_pos, l_neg], dim=1)
labels = zeros(logits.shape[0], dtype=torch.long)
#生成伪标签
loss = CrossEntropyLoss(logits/t, labels) #公式(2) ▷ L_SGCLR
#求解图对比损失,t = t 是超参数取值为 0.07
loss.backward( ) #根据 loss 来计算网络参数的梯度
update(f_q.params) # f_q 采用 SGD 更新参数
f_k.params =m * f_k.params+(1-m) * f_q.params
# f_k 采用动量更新参数,m =α 是动量系数取值为 0.999
self.dequeue_and_enqueue(k) #更新字典
```

2.2　CrosScale-SGCLR 算法

鉴于人体的运动主要是通过骨骼围绕各个关节进行旋转而实现,可以根据骨骼关节点的分布,将人体分割成粗细粒度不同的功能部件[34]。本文将人体关节点作为基本构件,将空间上相邻的关节点进行组合,形成不同尺度的骨架图,并基于各尺度间具有语义信息互补的特性,提出跨尺度一致性知识挖掘方法,利用一个尺度图中特征信息的相似性,来促进另一个尺度图中相似特征进行有效聚类。相比于 CrosSCLR[31] 方法,本文在不使用骨架视图(motion, bone)的情况下,通过构建多尺度骨架图来实现不同尺度间的信息融合,也可以很好地学习到不同图结构丰富的内部监督信息。

2.2.1　构建多尺度图

如图 2 所示,首先,给定一个包含 l 帧的骨架序列 $\boldsymbol{X}$,将其称为关节点尺度(即身体关节作为节点),记作 $\boldsymbol{\Theta}^1$。其次,构建粗粒度比例图,将运动者的骨架结构分为 10 个部分(包括躯干、头、右臂上、右臂下、左臂上、左臂下、右腿上、右腿下、左腿上和左腿下)和 5 个部分(包括躯干、右上肢、左上肢、右下肢和左下肢),并将每部分所包含的骨骼关节点进行位置坐标平均,合并为新的骨骼关节点,将其命名为粗关节点尺度(即身体部分作为关节节点),记作 $\boldsymbol{\Theta}^2$ 和 $\boldsymbol{\Theta}^3$。最后,基于以上操作,得到不同尺度的骨架图 $\boldsymbol{\Theta}^m(\boldsymbol{V}^m,\boldsymbol{\varepsilon}^m)$($m\in\{1,2,3\}$),其中 $\boldsymbol{V}^m=\{\boldsymbol{\nu}_1^m,\boldsymbol{\nu}_2^m,\cdots,\boldsymbol{\nu}_{n_m}^m\}$($\boldsymbol{\nu}_i^m\in\mathbf{R}^D$,$i\in\{1,2,\cdots,n_m\}$)表示不同骨架尺度图对应的关节点集合,$\boldsymbol{\varepsilon}^m=\{\boldsymbol{e}_{i,j}^m|\boldsymbol{\nu}_i^m,\boldsymbol{\nu}_j^m\in\boldsymbol{V}^m\}$($\boldsymbol{e}_{i,j}^m\in\mathbf{R}$)表示不同骨架尺度图边结构关系集合,$n_m$ 是第 m 个尺度图 $\boldsymbol{\Theta}^m$ 的关节点数。鉴于骨架的多尺度数据是由合并关节点所组成,导致图结构发生改变,不能直接输入到上文所建立的 SGCLR 模块中,为此,本文将针对不同尺度骨架图构建相应图结构,并将其命名为 multi-scale SGCLR(SGCLR(25), SGCLR(10), SGCLR(5))。在之后的实验中,将选取关节点数为 25 和 10 的多尺度骨架图 $\boldsymbol{\Theta}^1$ 与 $\boldsymbol{\Theta}^2$ 作为主要研究对象。

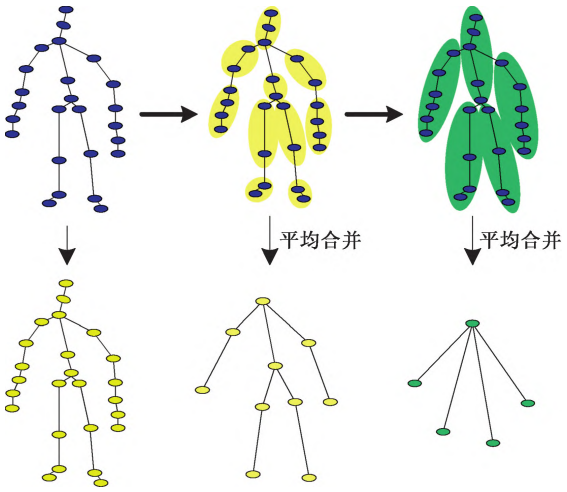

图 2　多尺度骨架结构图

Fig. 2　Multi-scale skeleton structure

2.2.2 跨尺度图对比学习网络

本节为了可以从骨架数据的不同尺度图中获取语义互补信息,协助网络从相似的负样本中挖掘出更多的正样本,拟结合多尺度骨架图数据,提出跨尺度图对比学习网络,该网络模型不仅可以从互补尺度图中挖掘出高置信度的正样本,而且使嵌入的上下文在多个尺度图中保持一致。具体模型框架如图 3 所示。

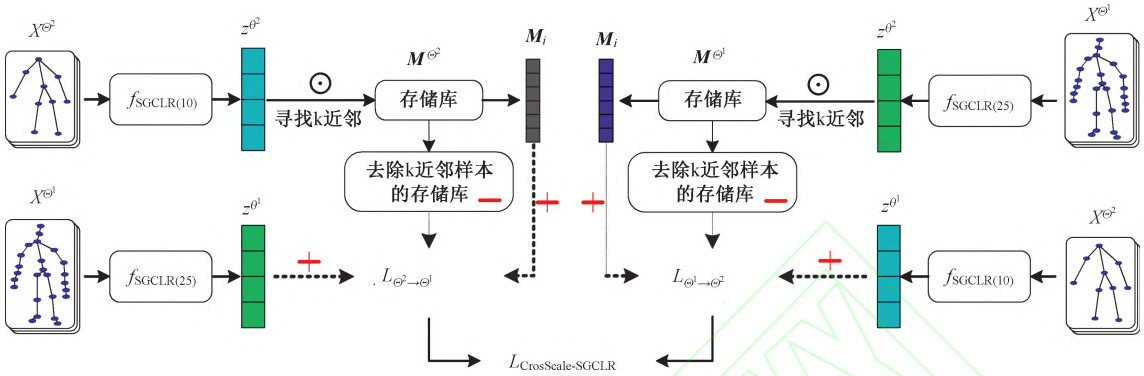

图 3 跨尺度 CrosScale-SGCLR 的算法结构图

Fig. 3 Architecture of cross-scale CrosScale-SGCLR

作为前一种方法的扩展,在图对比学习网络训练结束后,再进行跨尺度图对比学习,以获得更强的学习表征能力,避免网络从头开始训练时的错误分类。具体来看,给定一个骨架序列 $\boldsymbol{X}$,需要得到两种不同的尺度图 $\boldsymbol{X}^{\Theta^m}=(\boldsymbol{X}^{\Theta^1},\boldsymbol{X}^{\Theta^2})$,在本文中 $\boldsymbol{\Theta}^1$ 和 $\boldsymbol{\Theta}^2$ 分别是由 25 个和 10 个骨骼关节点组成的不同尺度骨架图,$\boldsymbol{X}^{\Theta^1}$ 和 $\boldsymbol{X}^{\Theta^2}$ 分别表示对应尺度下的骨架序列,CrosScale-SGCLR 表示学习的目的是学习具有较好泛化性的 $f_{\theta^1}(\cdot)$ 和 $f_{\theta^2}(\cdot)$,其中 $z^{\theta^1}=f_{\theta^1}(\boldsymbol{X}^{\theta^1})$,$z^{\theta^2}=f_{\theta^2}(\boldsymbol{X}^{\theta^2})$ 是 $\boldsymbol{\Theta}^1$ 和 $\boldsymbol{\Theta}^2$ 的特征表示,可以有效地执行各种下游任务。其主要思想与 SGCLR 方法不同之处在于需要重新构建 $\boldsymbol{X}^{\Theta^m}$ 的正样本集和负样本集,即在 $\boldsymbol{\Theta}^1$ 尺度中很难发现的正样本,可以在 $\boldsymbol{\Theta}^2$ 中发现。将多尺度数据 $\boldsymbol{\Theta}^m$ $(\boldsymbol{V}^m,\boldsymbol{\varepsilon}^m)$ $(m\in\{1,2\})$ 输入到 multi-scale SGCLR 网络中,通过两个不同图结构的 SGCLR 模块获得图编码特征 z^{Θ^m},以及相应的存储库 $\boldsymbol{M}^{\Theta^m}$,随着训练的进行,逐渐增强模型的表示学习能力。

最后利用对比损失函数进行参数更新,主要介绍尺度 $\boldsymbol{\Theta}^1(\boldsymbol{V}^1,\boldsymbol{\varepsilon}^1)$ 与 $\boldsymbol{\Theta}^2(\boldsymbol{V}^2,\boldsymbol{\varepsilon}^2)$ 之间的损失函数公式,具体如下:

$$L_{\Theta^2\to\Theta^1}=-\log\frac{\exp(z\cdot\tilde{z}/t)+\sum\limits_{i\in N_+^{\Theta^2\to\Theta^1}}\exp(z\cdot M_i\Theta^1/t)}{\exp(z\cdot\tilde{z}/t)+\sum\limits_{i=1}^{N}\exp(z\cdot M_i\Theta^1/t)},\quad(3)$$

其中,t 是超参数,$M_i\Theta^1\in\boldsymbol{M}^{\Theta^1}$ 为存储库中的负样本集,分子包含 $1+k$ 个正样本,分母包含 $1+k$ 个正样本和 $N-k$ 个负样本在内的共 $N+1$ 个样本,k 是相似样本特征嵌入的索引,由 $topk(\cdot)$ 函数进行选取,实验中 k 值取为 1。

同样地,在 $\boldsymbol{\Theta}^1$ 尺度特征空间中相似的实例也可以作为伪标签,帮助 $\boldsymbol{\Theta}^2$ 尺度下的网络进行更好地表征学习。其损失函数如下:

$$L_{\Theta^1\to\Theta^2}=-\log\frac{\exp(z\cdot\tilde{z}/t)+\sum\limits_{i\in N_+^{\Theta^1\to\Theta^2}}\exp(z\cdot M_i\Theta^2/t)}{\exp(z\cdot\tilde{z}/t)+\sum\limits_{i=1}^{N}\exp(z\cdot M_i\Theta^2/t)},\quad(4)$$

其参数意义与公式(3)相同,两个网络互相为对方采样正样本,以增强网络模型性能并获得更好聚类效果。

将公式(3)与公式(4)联立求和并取平均,即得到 CrosScale-SGCLR 方法的总损失函数,具体操作如下:

$$L_{\text{CrosScale-SGCLR}}=\frac{L_{\Theta^2\to\Theta^1}+L_{\Theta^1\to\Theta^2}}{2}。\quad(5)$$

多尺度损失函数 $L_{\text{CrosScale-SGCLR}}$ 与单尺度损失函数 L_{SGCLR} 相比,拉近了更多的高置信度正样本,使特征空间中同类样本特征更加容易聚合。

3 实验

3.1 实验数据集

NTU RGB+D 60[37] 由 56 880 个动作序列组

成,是目前基于骨架动作识别研究中应用最广泛的数据集。该数据集由 3 个 Microsoft Kinect v2 摄像头从不同的视角捕获,动作样本由 40 名演员执行,包含 60 种动作分类,其中 40 类为日常行为动作,9 类为与健康相关的动作,11 类为双人交互动作。本文采用该数据集的两种评价基准:1) Cross-Subject(xsub)基准,即训练数据来自 20 名演员,测试数据来自其他 20 名演员;2) Cross-View(xview)基准,其中训练数据来自摄像机视图 2 和 3,测试数据来自摄像机视图 1。

NTU RGB+D 120[38] 为 NTU RGB+D 60 数据集的扩展,该数据集包含来自 106 个演员执行的 120 种动作,相机的摆放位置由 17 个增加到 32 个,动作骨架序列总数扩充到 114 480。同样,本文采用该数据集的两种评价基准:Cross-Subject(xsub)和 Cross-Setup(xset)。在 xsub 基准中,身份标识为 1、2、4、5、8、9、13、14、15、16、17、18、19、25、27、28、31、34、35、38 等演员所做出的动作用作训练,其余的用于测试。在 xset 基准中,训练数据和验证数据分别由身份标识数字的奇偶进行确定。

3.2　实验设置

本文实验所用的硬件平台包括运行内存 128 GB 的 4 块 TITAN XP 显卡,软件平台包括 Python 3.6 和 PyTorch 1.2.0 框架。使用的参数配置与文献[31]保持一致,编码器 f_θ 和 $\tilde{f}_{\tilde{\theta}}$ 主要使用 ST-GCN[4] 网络,隐藏层维度为 256,特征维度为 128,f_θ 采用随机梯度下降法更新参数,$\tilde{f}_{\tilde{\theta}}$ 采用动量更新,动量系数 α 取值为 0.999,剪切常数 β 取值为 0.5,填充率 r 取值为 6,超参数 τ 取值为 0.07,ξ 取值为 2,随机边裁剪的个数范围在[0, 2]之间,训练过程中,将批量大小设为 128,存储库中负样本个数 M = 32 768,迭代次数设置为 250,权重系数为 0.000 1,每个模型均运行 300 epochs,其学习率初值为 0.1,在训练了 250 epochs 之后变为 0.01,线性评估均运行 100 epochs,其学习率初值为 0.3,在评估了 80 epochs 之后变为 0.03。

3.3　实验结果分析

本文的图对比学习网络基于 SkeletonCLR[31] 模型,在该网络模型的对比路径上加入了图增强方法,并使用 ST-GCN[4] 模型作为主干网络,在每个编码器后附加一个投影层以产生固定大小为 128 维的特征向量。在计算对比损失之前,对嵌入图进行归一化处理。由于随机边裁剪的范围在[0, 2]之间,当选到 0 条边裁剪时,其精度将会与原模型保持一致。

3.3.1　定量结果分析

如表 1 所示,将本文方法与其他基于骨架数据的自监督学习方面的工作进行了比较,主要对比了 LongT GAN[8]、MS^{2}L[7]、P&C[39]、AS-CAL[27] 与 SkeletonCLR[31],SGCLR 在 NTU RGB+D 60 数据集上的 xsub 与 xview 两个评价基准上的精度分别是 71.5%和 76.5%。相比于 SkeletonCLR[31],分别提升了 3.2%与 0.1%的精度,且 CrosView-SGCLR 在跨视图(joint+motion)上取得了 70.4%与 77.9%的精度。除此之外,本文基于图对比学习网络和不同尺度特征间的互补性,构建了跨尺度(joint25+joint10)协同训练网络模型,即 CrosScale-SGCLR,其精度分别达到了 70.3%和 75.2%。

表 1　NTU RGB+D 60 数据集上的实验精度对比

Tab.1　Comparison of accuracy on NTU RGB+D 60 dataset　%

方法	年份	xsub	xview
LongT GAN	2018	39.1	48.1
MS^{2}L	2020	52.6	—
P&C	2020	50.7	76.3
AS-CAL	2021	58.5	64.8
SkeletonCLR	2021	68.3	76.4
SGCLR	**2022**	**71.5**	**76.5**
CrosView-SGCLR	**2022**	**70.4**	**77.9**
CrosScale-SGCLR	**2022**	**70.3**	**75.2**

为了更好证明本文所提方法的有效性,同样在 NTU RGB+D 120 数据集上也做了相应的比较。如表 2 所示,SGCLR 网络模型在 NTU RGB+D 120 数据集的 xsub 和 xset 上分别达到了 57.6%和 54.6%的精度。CrosView-SGCLR 在跨视图(joint+motion)上取得了 60.1%与 62.2%的精度,在跨尺度图(joint25+joint10)上取得 59.0%与 63.6%的精度。结果表明随机边裁剪的图增强方法对人体骨架的图结构搭建起到促进作用,该方法在 xset 评价基准上性能改善较为明显;跨尺度图对比学习网络,在不使用骨架视图(motion, bone)情况下,利用多尺度间的协同训练方法,也可以达到较

好的识别效果。

表 2 NTU RGB+D 120 数据集上的实验精度对比

Tab. 2 Comparison of accuracy on NTU RGB+D 120 dataset %

方法	年份	xsub	xset
LongT GAN	2018	35.6	39.7
MS^{2}L	2020	—	—
P&C	2020	42.7	41.7
SkeletonCLR	2021	56.8	55.9
SGCLR	**2022**	**57.6**	**54.6**
CrosView-SGCLR	**2022**	**60.1**	**62.2**
CrosScale-SGCLR	**2022**	**59.0**	**63.6**

3.3.2 定性结果分析

本文利用 t-SNE[40] 降维算法，可视化预训练了 300 epochs 后的 SkeletonCLR[31]、SGCLR、CrosView-SGCLR 和 CrosScale-SGCLR 的嵌入分布。如图 4 所示，从 NTU RGB+D 60 的 xsub 数据集中选取 10 类样本进行嵌入比较，可以得出与表 1 相似的结论。与 SkeletonCLR[31] 相比，本文所提方法可以更加紧凑的聚合相同类别的嵌入特征，分离不同类别的嵌入特征，具有更好的识别能力。

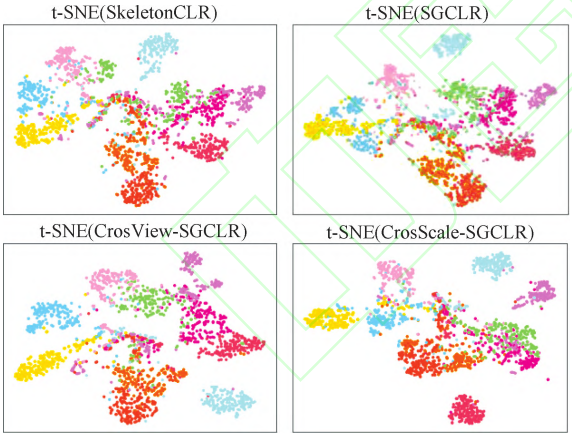

图 4 t-SNE 可视化不同模型嵌入特征的结果对比图

Fig. 4 Comparison of the results of the t-SNE visualization of different model embedding features

3.3.3 结果分析

通过上述实验，可以得出以下结论：利用骨架数据间存在的紧密关联，融合图对比学习思想，改变原有模型的数据增强方法，不仅可以增强同一骨架序列不同视图间的语义相关性表达，而且可以学习到骨架图的高级语义信息；利用跨尺度一致性知识挖掘方法，从未标记骨架序列中获得人员动作识别的有效表示，为不同数据尺度中传递特定行为语义，捕捉更丰富的尺度信息，弥补单一尺度特征难以准确表达复杂的人体动作。两个模型各有优势，同时也具有极强的互补关系，这也再次证明了本文所构建模型的有效性。

3.4 消融实验结果分析

本节所有实验均在 NTU RGB+D 数据集上进行，该数据集可以分为以下四个基准，分别是 xsub60、xview60、xsub120 和 xset120，测试实验遵循线性评估协议，在消融实验参数选取上参考模型 AimCLR[28] 与 SkeletonCLR[31]。

3.4.1 验证原数据增强 τ 有效性

数据增强可以学习到相同实例不同的表示状态，对图对比学习方法起着至关重要的作用。本文利用剪切与时序裁剪的方法对骨架数据进行扩增，通过改变参数 β 与 r 的值，在模型 SGCLR 上进行相应实验，结果如表 3 所示，在四种基准数据集中，当 $\beta=0.5$，$r=8$ 时，测试结果的平均精度达到了最大值，由此说明数据增强 τ 的有效性以及不同数据增强强度对图对比学习方法的影响程度。由于本文模型构建基于 SkeletonCLR[31]，因此在参数选取上应与其保持一致，选取 $\beta=0.5$，$r=6$。

表 3 不同参数 β 与 r 对网络性能的影响

Tab. 3 The impact of different parameters β and r on network performance %

β	r	xsub60	xview60	xsub120	xset120
0.2	0	62.1	65.8	51.2	53.9
0.5	0	64.2	66.8	49.4	51.6
1	0	63.4	63.6	51.0	48.2
0.5	4	69.5	73.4	54.1	**58.4**
0.5	6	**71.5**	76.5	**57.6**	54.6
0.5	**8**	69.2	**76.8**	56.6	58.2

3.4.2 验证所提数据增强 *mask_edg* 有效性

由于超参数 ξ 决定了边裁剪的数量范围，影响着骨架图结构表示学习。为了验证边裁剪数据增强方法的有效性，本文做了相应的消融实验。如表 4 所示，将 ξ 分别取值为 2、3、4、5、10 和 15。可以看出，当 $\xi=2$ 时，模型的平均精度达到了最大值。然而，较大的 ξ 会降低所构建模型的性能，因为裁剪较多的边会导致邻接矩阵变成稀疏矩阵，不利于图结构的特征表达，致使自监督网络模型

的错误分类。

表 4　不同参数 ξ 对网络性能的影响

Tab. 4　The impact of different parameters ξ on network performance　%

ξ	xsub60	xview60	xsub120	xset120
2	**71.5**	**76.5**	**57.6**	54.6
3	52.4	57.1	35.7	41.1
4	52.4	57.0	46.1	38.7
5	58.5	64.0	35.6	38.8
10	59.4	57.8	47.6	43.1
15	70.5	70.4	54.7	**55.7**

3.4.3　验证所提图对比学习有效性

为了验证自监督模型在加入图对比学习(GCL)方法时,识别效果会有不同程度的提升。本文选用 SGCL(单尺度)以及 CrosScale-SGCLR(跨尺度)为基准模型,对比了加入 GCL 和去除 GCL 的网络模型识别效果,将其记作 SGCLR† 与 CrosScale-SGCLR†。如表 5 所示,分别在 NTU RGB+D 60 与 NTU RGB+D 120 两个数据集上进行验证,得到相应结论:加入图数据增强方法之后,可以学习到骨架序列潜在的高级语义信息,增强同一样本不同视图的语义相关性表达,所以在不同基准下测试该模型,总体上加入了图对比学习要优于没加图对比学习的识别效果。

表 5　模型 SGCLR 和 CrosScale-SGCLR 消融实验测试结果

Tab. 5　Ablation experimental test results on the model SGCLR and CrosScale-SGCLR　%

方法		NTU-60		NTU-120	
网络模型	GCL	xsub	xview	xsub	xset
SGCLR†	×	68.3	76.4	56.8	**55.9**
SGCLR	√	**71.5**	**76.5**	**57.6**	54.6
CrosScale-SGCLR†	×	70.3	74.5	57.9	62.8
CrosScale-SGCLR	√	**70.3**	**75.2**	**59**	**63.6**

4　结论

本文提出一种用于人体骨架动作识别的跨尺度图对比学习网络模型,通过聚合相关骨骼关节点,结合跨尺度感知一致性,构建出多个尺度的骨架图,摆脱了传统方法在扩增骨架数据过程中,存在泛化性不足和传递性不强的问题,增强了高级语义信息的表达,提高了最近邻挖掘策略,使学习过程更加合理。为了验证该方法的有效性,分别在公开的人体骨架动作识别数据集 NTU RGB+D 60 和 NTU RGB+D 120 上进行了大量实验,结果表明,本文所提方法的识别效果均比文中所涉及的其他方法有一定的提高。在以后的工作中,将继续研究基于多尺度骨架图序列的自监督动作识别方法,以实现利用更加轻量的网络模型达到更高的识别精度这一目标。

参考文献

[1] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231.

[2] 胡正平, 张敏姣, 李淑芳, 等. 智能视频监控系统中行人再识别技术研究综述[J]. 燕山大学学报, 2019, 43(5): 377-393.
HU Z P, ZHANG M J, LI S F. Review of person re-identification technology in intelligent video surveillance system[J]. Journal of Yanshan University, 2019, 43(5): 377-393.

[3] 聂栋栋, 贺悦悦, 马勤勇. 基于 PCA_LDA 和协同表示分类的人脸识别算法[J]. 燕山大学学报, 2019, 43(2): 176-181.
NIE D D, HE Y Y, MA Q Y. Face recognition based on PCA_LDA and collaborative representation[J]. Journal of Yanshan University, 2019, 43(2): 176-181.

[4] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//The AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018: 1113-1122.

[5] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 12026-12035.

[6] LIU Z, ZHANG H, CHEN Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//The IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 140-149.

[7] LIN L, SONG S, YANG W, et al. MS^{2}L: multi-task self-supervised learning for skeleton based action recognition[C]//The ACM International Conference on Multimedia, Dublin, Ireland, 2020: 2490-2498.

[8] ZHENG N, WEN J, LIU R, et al. Unsupervised representation learning with long-term dynamics for skeleton based action recognition[C]//The AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018: 2644-2651.

[9] YANG S, LIU J, LU S, et al. Skeleton cloud colorization for unsupervised 3d action representation learning[C]//The IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 13423-13433.

[10] 李龙. 融合注意力机制的人体骨骼点动作识别方法研究[D].

成都: 成都理工大学, 2019.
LI L. Research on human bone point action recognition method integrating attention mechanism [D]. Chengdu: Chengdu University of Technology, 2019.

[11] BANERJEE A, SINGH P K, SARKAR R. Fuzzy integral-based cnn classifier fusion for 3d skeleton action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(6): 2206-2216.

[12] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition [C]//The IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1110-1118.

[13] LI C, ZHONG Q, XIE D, et al. Skeleton-based action recognition with convolutional neural networks [C]//The IEEE International Conference on Multimedia & Expo Workshops, Hong Kong, China, 2017: 597-600.

[14] ZHANG X, XU C, TIAN X, et al. Graph edge convolutional neural networks for skeleton-based action recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(8): 3047-3060.

[15] ZHANG P, LAN C, XING J, et al. View adaptive neural networks for high performance skeleton-based human action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1963-1978.

[16] PAN T, SONG Y, YANG T, et al. Videomoco: contrastive video representation learning with temporally adversarial examples [C]//The IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 11205-11214.

[17] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning [C]//The IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 2575-7075.

[18] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]// The International Conference on Machine Learning, Vienna, Australia, 2020: 2640-3498.

[19] CHEN X, HE K. Exploring simple siamese representation learning[C]//The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 15750-15758.

[20] HAN T, XIE W, ZISSERMAN A. Self-supervised co-training for video representation learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 5679-5690.

[21] NOROOZI M, FAVARO P. Unsupervised learning of visual representations by solving jigsaw puzzles [C]//The European Conference on Computer Vision, Amsterdam, Holland, 2016: 69-84.

[22] NOROOZI M, VINJIMOOR A, FAVARO P, et al. Boosting self-supervised learning via knowledge transfer [C]//The IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 9359-9367.

[23] ZHANG R, ISOLA P, EFROS A A. Colorful image colorization [C]//The International European Conference on Computer Vision, Amsterdam, Holland, 2016: 649-666.

[24] ZHANG J, ZHAO Y, SALEH M, et al. Pegasus: pre-training with extracted gap-sentences for abstractive summarization[C]// The International Conference on Machine Learning, Shangri-La, China, 2020: 11328-11339.

[25] FENG Z, XU C, TAO D. Self-supervised representation learning by rotation feature decoupling[C]//The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 10364-10374.

[26] ZHAI X, OLIVER A, KOLESNIKOV A, et al. S4l: self-supervised semi-supervised learning [C]//The IEEE/CVF International Conference on Computer Vision, Long beach, USA, 2019: 1476-1485.

[27] RAO H, XU S, HU X, et al. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition [J]. Information Sciences, 2021, 569: 90-109.

[28] GUO T, LIU H, CHEN Z, et al. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition [C]// The AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2022: 762-770.

[29] MUNARO M, FOSSATI A, BASSO A, et al. One-shot person re-identification with a consumer depth camera[M]//GONG S, CRISTANI M, YAN S, et al. Person Re-Identification: Advances in Computer Vision and Pattern Recognition. London: Springer, 2014: 161-181.

[30] PALA P, SEIDENARI L, BERRETTI S, et al. Enhanced skeleton and face 3d data for person re-identification from depth cameras[J]. Computers & Graphics, 2019, 79: 69-80.

[31] LI L, WANG M, NI B, et al. 3D human action representation learning via cross-view consistency pursuit[C]//The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 4741-4750.

[32] LI M, CHEN S, ZHAO Y, et al. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction [C]//The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 214-223.

[33] LIAO R, YU S, AN W, et al. A model-based gait recognition method with body pose and human prior knowledge[J]. Pattern Recognition, 2020, 98: 107069.

[34] RAO H, HU X, CHENG J, et al. SM-SGE: a self-supervised multi-scale skeleton graph encoding framework for person re-identification[C]//The 29th ACM International Conference on Multimedia, Chengdu, China, 2021: 1812-1820.

[35] YOU Y, CHEN T, SUI Y, et al. Graph contrastive learning with augmentations[J]. Advances in Neural Information Processing

Systems, 2020, 33: 5812-5823.

[36] XIA J, WU L, CHEN J, et al. Simgrace: A simple framework for graph contrastive learning without data augmentation[C]//The ACM Web Conference, Lyon, France, 2022: 1070-1079.

[37] SHAHROUDY A, LIU J, NG T T, et al. Ntu rgb+ d: a large scale dataset for 3d human activity analysis [C]//The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1010-1019.

[38] 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.

[39] SU K, LIU X, SHLIZERMAN E. Predict & cluster: unsupervised skeleton based action recognition[C]//The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 9631-9640.

[40] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.

Human skeleton for action recognition based on cross-scale graph contrastive learning

ZHANG Xuelian[1], XU Zengmin[1,2], CHEN Jiakun[1], WANG Lulu[1]

(1. School of Mathematics and Computing Science, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China;
2. Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China;
3. Anview. ai, Guilin Anview Technology Co. Ltd., Guilin, Guangxi 541010, China)

Abstract: Traditional self-supervised learning models based on the human skeleton usually use contrastive learning modules for representation learning, while existing contrastive learning modules use data augmentation methods to construct similar positive samples, and the rest of the samples are all negative samples, which limits the expression of semantic information for similar samples. To solve these issues, an action recognition algorithm with graph contrastive learning and cross-scale consistent knowledge mining is proposed. First, a new data augmentation method is designed based on the skeleton graph structure, which performs random edge cuttings on the input skeleton sequence to obtain two different views, thus enhancing semantic correlation expression between different views of the same skeleton sequence. Second, to alleviate the problem of low embedding similarity to similar samples, a self-supervised co-training network model is introduced to obtain positive class samples from one skeleton scale and another skeleton scale by using complementary information between different scales of the same skeleton data source, to realize the association within a single scale and semantic collaborative interaction between multi-scales. Finally, the effectiveness of the model is evaluated based on the linear evaluation protocol, and the experimental results on NTU RGB+D 60 and NTU RGB+D 120 datasets show that the recognition accuracy of the proposed method is improved by 2% ~ 3. 5% on average compared with the cutting-edge mainstream methods.

Keywords: graph contrastive learning; data augmentation; cross-scale consistent knowledge mining; co-training; human skeleton