计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):261−269 [doi: 10.15888/j.cnki.csa.008729]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于融合几何特征时空图卷积网络的动作识别①

邹浩立

(华南师范大学 计算机学院, 广州 510631)
通信作者: 邹浩立, E-mail: haolizou@m.scnu.edu.cn

摘　要: 最近, 基于骨架的动作识别研究受到了广泛关注. 因为图卷积网络可以更好地建模非规则数据的内部依赖, ST-GCN (spatial temporal graph convolutional network) 已经成为该领域的首选网络框架. 针对目前大多数基于 ST-GCN 的改进方法忽视了骨架序列所蕴含的几何特征. 本文利用骨架关节几何特征, 作为 ST-GCN 框架的特征补充, 其具有视觉不变性和无需添加额外参数学习即可获取的优势, 进一步地, 利用时空图卷积网络建模骨架关节几何特征和早期特征融合方法, 构成了融合几何特征的时空图卷积网络框架. 最后, 实验结果表明, 与 ST-GCN、2s-AGCN 和 SGN 等动作识别模型相比, 我们提出的框架在 NTU-RGB+D 数据集和 NTU-RGB+D 120 数据集上都取得了更高准确率的效果.

关键词: 几何特征; 特征融合; 骨架; 时空图卷积网络; 动作识别; 深度学习

引用格式: 邹浩立.基于融合几何特征时空图卷积网络的动作识别.计算机系统应用,2022,31(10):261–269. http://www.c-s-a.org.cn/1003-3254/8729.html

Spatio-temporal GCN with Geometric Features Fusion for Action Recognition

ZOU Hao-Li

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: Recently, the research on skeleton-based action recognition has attracted a lot of attention. As the graph convolutional networks can better model the internal dependencies of non-regular data, the spatio-temporal graph convolutional network (ST-GCN) has become the preferred network framework in this field. However, most of the current improvement methods based on the ST-GCN framework ignore the geometric features contained in the skeleton sequences. In this study, we exploit the geometric features of the skeleton joint as the feature enhancement of the ST-GCN framework, which has the advantage of visual invariance without additional parameters. Further, we integrate the geometric feature of the skeleton joint with earlier features to develop ST-GCN with geometric features. Finally, the experimental results show that the proposed framework achieves higher accuracy on both NTU-RGB+D dataset and NTU-RGB+D 120 dataset than other action recognition models such as ST-GCN, 2s-AGCN, and SGN.

Key words: geometric features; feature fusion; skeleton; spatio-temporal graph convolutional network (ST-GCN) ; action recognition; deep learning

1　引言

人类动作识别是计算机视觉领域的一个热门话题, 其广泛地应用在视频监控、人机交互和自动驾驶等领域中[1]. 同时, 动作识别也是视频理解方向很重要的一个问题, 至今为止已经被研究多年[2]. 简单地说, 动作识别问题就是: 对于给定的分割好的视频片段, 按照其中的人物动作, 如: 打球、跑步和挥手, 进行分类. 根据输入模型的模态, 动作识别通常可以划分为: 基于视频和

① 收稿时间: 2022-01-11; 修改时间: 2022-01-30; 采用时间: 2022-02-22; csa 在线出版时间: 2022-06-28

RGB 图片的动作识别和基于骨架数据的动作识别. 基于视频和 RGB 图片的动作识别方法通常从 RGB 图像中提取感兴趣的特征, 如: RGB 图片/视频中具有代表性的人体动作信息, 然后将一个动作视频转换成一个特征向量, 最后将特征向量输入分类器中进行分类. 得益于 Microsoft Kinect v2 深度摄像机的发展和人体关键点检测技术的迅速发展[3], 基于骨架数据的动作识别研究也变得火热. 从生物学角度来说, 骨架数据是一种具有高级语义信息的特征, 即使没有外观信息, 人类也能够通过仅观察骨架关节的运动过程来识别动作类别[4]. 特别地, 相比于 RGB 数据, 骨架数据因其对动态环境和复杂背景的强适应性而受到广泛研究. 本文研究的重点内容是基于骨架数据的动作识别.

1.1 基于骨架的动作识别方法

传统的基于骨架的动作识别方法通过手工设计特征来建模骨架的数据依赖, 如局部占位特征[5]、时间联合协方差[6]和李群曲线[7]. 这些方法是根据物理直觉设计算法来建模人类动作的时空特征, 其不能灵活地应付大型数据集[8]. 深度学习是一种数据驱动的方法, 面对大型骨架序列数据集能更好地展示其优势. 已有的基于深度学习的动作识别方法按照模型类型可以分为 3 大主流方法: 基于 RNN (recurrent neural networks) 的方法、基于 CNN (convolutional neural networks) 的方法和基于 GCN (graph convolution networks) 的方法.

1.2 基于 RNN 的方法

RNN 被广泛地应用于时序任务上, 但 RNN 模型通常只能接受矢量序列作为输入, 不能较好地建模骨架关节间的空间依赖. 为了克服这一缺点, Du 等人[9]提出了端到端分层 RNN 框架, 该方法将骨架划分为多个部位并作为每个 RNN 子网络的输入, 然后将子网络的输出进行分层融合. Zhu 等人[10]提出在 LSTM 网络中使用组稀疏正则化来自动探索骨架关节的共同发生特征. 为了同时建模骨架关节间的空间和时间依赖, 空间-时间 LSTM 网络将深度 LSTM 模型扩展到两个并发域, 即时间域和空间域[11].

1.3 基于 CNN 的方法

CNN 被广泛地应用于图像分类任务[12]. 为了满足 CNN 网络输入的需要 (二维网格), 研究人员将骨架关节编码为多个 2D 伪图像, 然后将其输入 CNN 网络以学习时空特征[13,14]. Wang 等人[15]提出了关节轨迹图 (joint trajectory maps), 该方法通过颜色编码将关节轨迹的空间结构和动力学表示为 3 幅纹理图像. 然而, 该方法较为复杂, 在映射过程中也失去了骨架内部重要意义的空间信息. Li 等人[16]使用了平移比例不变的图像映射策略, 该方法首先根据人体的物理结构将每个帧中的人体骨架关节划分为 5 个主要部分, 然后将这些部分映射到 2D 形式. Li 等人[17]提出了一个共同发生特征学习网络框架 (HCN), 该方法利用 CeN 网络来聚合骨架全局上下文特征并且取得了不错的效果. 基于 GCN 的方法. 最近, Yan 等人[18]提出了时空图卷积网络 (ST-GCN), 该方法将人体骨架数据直接建模为图结构, 其无需要手工设计并划分骨架部位或制作人体骨架关节点遍历规则, 因此该方法比以前的方法取得了更好的性能[8]. 随后, Shi 等人[19]提出了 2s-AGCN 网络, 该方法将自适应拓扑图添加到每个图卷积层中增强图卷积层的远距离空间建模能力. Zhang 等人[20]提出了 SGN 网络, 该方法利用人体关节点和帧的语义信息, 丰富了骨架特征的表达能力, 从而提高模型的识别准确率. 无论如何, RNN 网络和 CNN 网络都不能完全表征骨架数据空间结构, 因为骨架数据不是矢量序列或二维网格, 其具有人体结构自然连接的图的结构. 与前两者方法相比较, 基于 GCN 的方法不需要手工划分骨架为多个部位和设计关节遍历规则, 并且在建模骨架空间和时间依赖过程中可以保留骨架拓扑结构, 因此, 基于 GCN 的动作识别方法建模骨架时空特征更具优势并且逐渐成为该领域的首选框架.

1.4 骨架几何特征用于动作识别

与骨架坐标特征相比较, 骨架几何特征具有视觉不变性的优势. 早期, 骨架几何特征被研究人员进行大量研究, 如, Geometric Pose Descriptor[21]、Fusing Geometric Features[22,23]和 DD-Net[24]. Chen 等人[21]通过手工设计了多组骨架几何特征 (关节-关节距离、关节-关节角度和关节-关节平面等等) 用于表征人类动作信息. Zhang 等人[22]提出了多组简单的骨架几何特征, 然后将每组特征分别送入一个 3 层 LSTM 框架. Li 等人[23]将多组骨架几何特征分别输入到 LSTM 和 CNN 中, 再将多个流最后的输出进行融合. Yang 等人[24]提出了 DD-Net, 该方法分别对 fast motion 特征、slow motion 特征和 JCD (joint collection distances) 特征进行

嵌入学习, 再将 3 种特征进行早期融合, 最后将融合特征输入到 1D CNN 网络. 事实上, 骨架几何特征 (关节-关节距离和关节-关节角度等等) 是高效的和无需参数学习的特征, 然而, 目前基于 GCN 的动作识别方法[18–25] 忽视了这些骨架几何特征. 为此, 本文在 ST-GCN 网络框架上研究了每帧骨架中关节间的距离特征, 将其作为 ST-GCN 网络的特征补充, 并利用骨架几何建模模块和早期特征融合方法构建了融合几何特征时空图卷积网络框架 (GEO-GCN).

2 融合几何特征时空图卷积网络框架

2.1 时空图卷积网络框架

骨架序列能够高效和简洁地表征人类动作的动态信息. 基于深度学习的骨架动作识别的算法种类繁多, 而图卷积网络[26] 作为后起之秀, 因其可以更好地建模非规则数据, 因此, 本文采用 ST-GCN 网络框架[18] 作为本文的基准网络框架.

一般地, 原始骨架序列数据每帧中的位置信息由向量表示. 每个向量表示相应人体关节的二维或三维坐标. 一个完整的人类动作包含多个帧, 对于不同的动作序列样本具有不同的帧数. 本文遵循 ST-GCN 网络框架, 使用时空人体拓扑图来建模骨架关节之间的空间和时间信息. 图 1 展示了 ST-GCN 构建的时空人体拓扑图, 其中每个圆点表示为时空图的顶点, 人体的自然连接表示为每帧骨架空域图的空域边. 对于时间维度, 两相邻帧间对应关节的连接表示为时域边. 每个关节的坐标向量为对应图顶点的属性. 为了建模时空骨架图的时空特征, ST-GCN 中每层 GCN Layer 通过交替堆叠 GC-block 和 TC-block 来构建而成, 其中, GC-block 和 TC-block 分别沿着关节维度 (V) 和时间维度 (T) 聚合特征. 对于空间维度上建模, GC-block 可以表示为:

$$Y=\sum_{k=1}^{K}\Lambda_k^{-\frac{1}{2}}A_k\Lambda_k^{-\frac{1}{2}}XW \tag{1}$$

其中, X和Y分别表示输入和输出特征. W表示可学习矩阵. 对于每个骨架空间配置, A是骨架拓扑图的邻接矩阵, Λ是用于归一化的对角节点度矩阵. 根据 ST-GCN 的空间配置, K表示 GC-block 中人体拓扑图的数量, 特别地, 原始 ST-GCN 设置每个 GC-block 的拓扑图数目 $K=3$. 此外, 节点i的阶数由$\Lambda^{ii}=\sum_j A^{ij}+\alpha$计算所得, 其中$A^{ij}$表示元素在$A$中的第$i$行和第$j$列中加上一个常数$\alpha$, 以避免$A$为全零的问题.

图 1 ST-GCN 时空拓扑图

对于时间维度上建模, 由于每个顶点的邻域数固定为 2 (两个连续帧中的对应关节), 因此应用类似经典卷积运算的图卷积是较为简单的. 具体地说, TC-block 是内核大小为$K_t\times 1$的普通卷积层.

图 2 展示了 ST-GCN 网络框架, 其由 10 层 GCN Layer 堆叠而成. 整体 ST-GCN 网络可以被划分为 3 个阶段, 第 1 个阶段包含了 4 层 GCN Layer, 而第 2 个和第 3 个阶段都包含了 3 层 GCN Layer. 骨架坐标特征通过每个阶段, 其通道维度数量变为原来的两倍, 而时间维度特征数量减少至原来的一半, 这样做的目的是: 增强骨架特征表达能力, 同时保持张量数据的总参数量不变. 模型最终输出的时空特征经过全局池化层 (GAP), 再被输入到 Softmax 分类器, 以获得动作预测结果.

图 2 中下方展示了 GCN Layer 内部结构, 其包含了一个 GC-block 和一个 TC-block. 根据上述可知, 骨架坐标特征输入 GCN Layer 后, GC-block 首先对输入骨架坐标特征进行空间建模, 跟随其后的是一个 BN (batch normalization) 层[27] 和一个 ReLU 激活层, 分别对特征起到正则化和非线性激活作用. 骨架坐标特征被空间建模后, TC-block 对其进行时间建模, 同样地, BN 层和 ReLU 激活层跟随其后. 此外, 每个 GCN Layer 都包含残差连接 (skip connect)[12], 其起到稳定网络训练的作用.

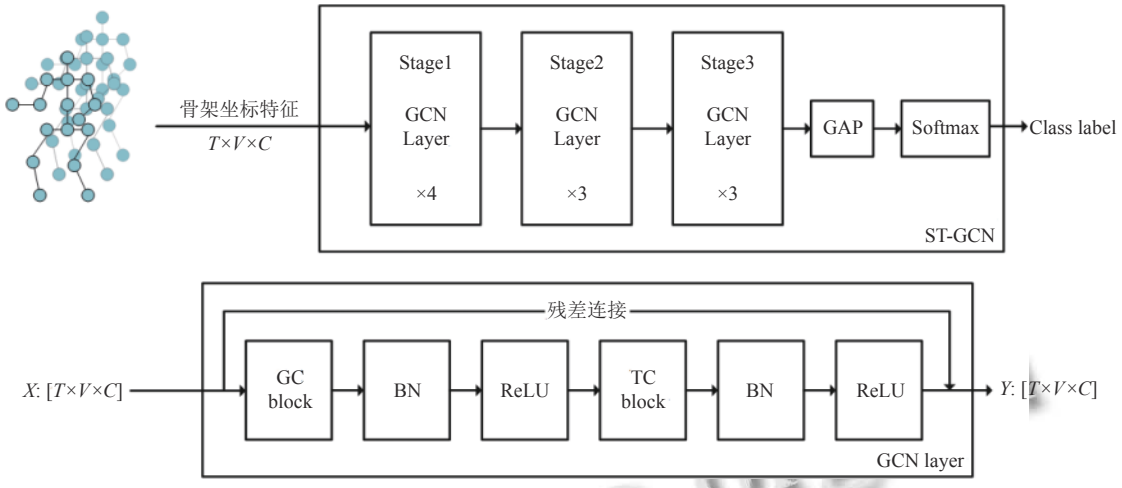

图 2 ST-GCN 网络框架和 GCN Layer 结构图

2.2 骨架关节几何特征

在现实场景中, 人类的动作可能会被任意的相机视角观察. 为了应对视角变化所带来的挑战, Zhang 等人[22] 将多组具有视觉不变性的骨架几何特征序列分别输入子 LSTM 网络中. Yang 等人[24] 利用帧内关节间距离集合的下三角矩阵作为 JCD 特征, 将其与 fast motion 特征和 slow motion 特征进行早期特征融合. 然而, 上述方法提出的几何特征很大程度上依赖于人的先验经验, 这不利于在不同数据集间泛化. 此外, 手工获取的骨架几何特征被转换为矢量序列, 不能单独考虑每个关节点的几何特征, 这不利于模型提取有判别力的时空特征. 为了缓解这些问题, 本文引入骨架关节几何特征, 即, 每帧内关节与关节间的欧几里得距离, 其具有视觉不变性, 而且骨架关节几何特征可以依靠时空拓扑图进行信息交换.

更详细地说, 给定一副骨架序列$X \in \mathbb{R}^{T\times V\times C}$, 其中$T$表示骨架序列总帧数 (本文默认设置$T$ =48), 每帧骨架总共有V个关节点, C表示骨架数据所处的是三维笛卡尔坐标系或者二维笛卡尔坐标系. 在第 t 帧骨架中, 第v个关节点的三维笛卡尔坐标表示为$P_v^t=(x,y,z)$, 而二维笛卡尔坐标表示为$P_v^t=(x,y)$.

通过距离公式, 可以计算每帧内任意两个关节点间的欧几里得距离, 具体公式如下:

$$D_{i,j}^t=\sqrt{\left(x_i^t-x_j^t\right)^2+\left(y_i^t-y_j^t\right)^2+\left(z_i^t-z_j^t\right)^2} \tag{2}$$

通过式 (2), 可求得第t帧第i个关节点与第t帧内所有关节点的欧几里得距离特征为$D_i^t=\mathbb{R}^{V\times V}$, 特别地当$i=j$时, 特征值为 0. 因此, 对于给定的一副骨架序列数据, 通过距离公式, 可求得该骨架序列的骨架关节几何特征为$D\in\mathbb{R}^{T\times V\times V}$. 特别地, 每帧骨架关节几何特征不需要转为矢量序列.

2.3 早期特征融合与几何特征建模

骨架几何特征和骨架坐标特征是不同的模态. 模态融合方法[28] 可以分为: 早期融合和晚期融合. 在基于视频的动作识别领域中, Simonyan 等人[29] 提出了晚期融合的双流模型, 该方法利用双流模型分别对 RGB 图像和光流数据进行建模, 对各流模型的最后输出特征进行融合, 但双流模型方法会导致总模型的参数量成倍数增加. Yang 等人[24] 提出的 DD-Net 利用早期特征融合方法对 3 种骨架几何特征进行融合, 该方法利用骨架几何特征提高了模型的准确率同时不会大幅度增加总网络的参数量. 本文借鉴 DD-Net 的早期特征融合方法, 使得 ST-GCN 框架融合骨架关节几何特征D同时不大幅度增加总网络的参数量. 然而, DD-Net 方法的嵌入学习模块不能较好地建模骨架关节几何特征的时空依赖, 为此, 本文探索了 3 种骨架关节几何特征建模方法分别为: 直接融合方法、特征嵌入方法和 GCN 建模方法.

(1) 直接融合方法. 为了验证骨架关节几何特征的有效性, 本文提出直接将距离公式计算所得的骨架关节几何特征D与 ST-GCN 网络第 1 阶段输出的时空特征在通道维度上进行拼接融合, 利用一层1×1卷积层对融合特征进行降维操作, 然后将其作为 ST-GCN 剩余网络的输入. 值得注意的是, 该方法可视为一层单元层.

(2) 特征嵌入方法. 一方面, 骨架关节几何特征D具有一定的先验经验, 而先验经验不利于模型的泛化性. 另外一方面, 骨架关节几何特征和骨架坐标特征是不同的模态, 上述方法是通过特征拼接方式对两种模态进行融合, 这在一定程度上不利于 ST-GCN 网络提取有判别力的时空特征. 为了减少先验经验带来的影响同时让骨架关节几何特征更好地融合到 ST-GCN 网络, 本文参考 DD-Net[24] 对骨架几何特征处理方法, 利用两层全连接层 (fully connected layer) 对骨架关节几何特征进行特征嵌入学习, 再将所得的骨架关节几何嵌入特征和 ST-GCN 网络第一阶段输出的时空特征在通道维度上进行拼接融合, 再利用一层 1×1 卷积层对融合特征进行降维操作并将输出作为 ST-GCN 剩余阶段网络的输入.

(3) GCN 建模方法. 然而, 上述两种方法都忽视了对骨架关节几何特征D时间维度上的建模. ST-GCN 网络第 1 阶段输出的是时空特征, 为了让每帧骨架的几何特征具备时间维度上的依赖, 本文利用两层 GCN Layer 对骨架序列的几何特征进行时空建模, 其目的是使骨架关节几何特征与 ST-GCN 第 1 阶段所建模的时空特征更具有一般性. 最后, 被 GCN 模块建模的骨架关节几何特征如上述两种方法一样被拼接融合和降维操作, 再将其输入 ST-GCN 的第 2 和第 3 阶段进行时空建模.

2.4 融合几何特征时空图卷积算法框架

图 3 展示了本文提出的融合几何特征时空图卷积网络框架 (GEO-GCN). 骨架坐标特征作为 ST-GCN 网络第一阶段的输入, 同时, 通过距离公式计算所得的骨架关节几何特征D作为骨架几何建模模块的输入. 两模块的输出在通道维度上进行拼接融合, 融合特征被一层 1×1 卷积层进行降维操作, 其目的是与 ST-GCN 网络第 2 阶段的输入适配. 值得注意的是, 骨架几何特征建模模块在最终模型中使用的是 GCN 建模方法. 通过早期特征融合方法, GEO-GCN 的参数量不会成倍数地增加, 同时可以使得 ST-GCN 网络在保持自身建模能力的情况下, 增强了剩余阶段网络对融合骨架关节几何特征的时空特征的建模能力, 从而增强模型性能.

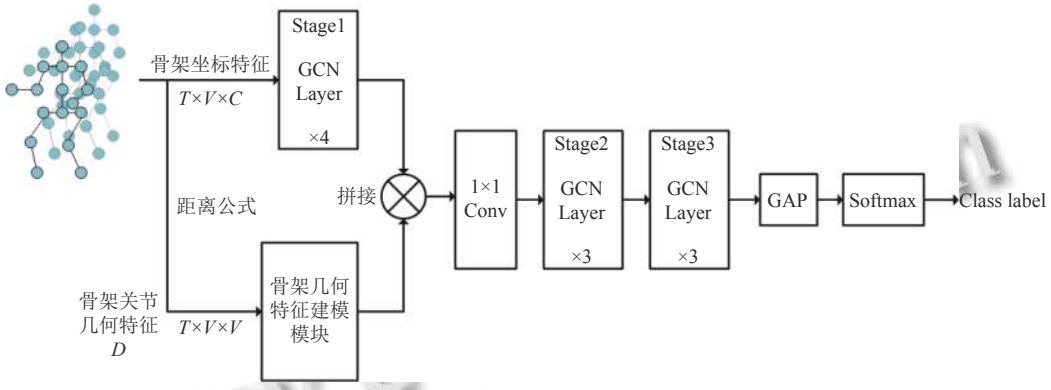

图 3 融合几何特征的时空卷积网络框架 (GEO-GCN)

3 实验分析

本文在两个大规模的基于骨架的动作识别基准数据集, 即, NTU-RGB+D 数据集[30] 和 NTU-RGB+D 120 数据集[31], 对 GEO-GCN 网络框架进行评估. 本文首先通过充分的消融实验以验证骨架关节几何特征能作为 ST-GCN 网络的特征补充, 然后验证不同骨架几何建模模块对 GEO-GCN 网络框架的影响. 最后, 将 GEO-GCN 网络框架与其他动作识别模型进行准确率的比较.

3.1 实验环境和实验数据集

本文所有实验都是在一个 RTX 2080 TI GPU 上进行的并且该 GPU 采用 PyTorch 深度学习框架和 Python 编程语言.

NTU-RGB+D 是一个大规模的人体动作识别数据集, 包含 4 种模态, 即 RGB 视频、深度序列、红外视频和 3D 骨架数据. 3D 骨架序列数据由 Microsoft Kinect v2 摄像头捕获. 它总共有 56 880 个视频, 由 3 台摄像机从不同角度拍摄. 这些动作涵盖 60 种人类动作类别, 包括类别 1 到类别 49 的单人动作和类别 50 到类别 60 的双人交互动作. 数据集的发布方推荐了两个评估基准, 即, 交叉对象 (cross-subject) 评估和交

叉视角 (cross-view) 评估. 在 X-Sub 评估基准中, 训练集包含了来自 20 名受试者的 40 320 个视频, 其余 16 560 个视频片段用于测试. 在 X-View 评估基准中, 它包含 37 920 个从第 2 摄像头和第 3 摄像头拍摄的视频, 用于训练. 从第一个摄像头拍摄的视频包含 18 960 个视频, 用于测试.

NTU-RGB+D 120 是 NTU-RGB+D 的扩展, 其中类别的数量扩大到 120, 样本的数量扩大到 114 480. 还有两种推荐的评估基准, 即交叉主体 (C-subject) 评估和交叉设置 (C-setup) 评估. 在 X-Sub 评估基准中, 来自 53 个受试者的 63 026 个视频片段被用于训练, 其余受试者则被用于测试. 在 X-Set 评估基准中, 54 471 个具有偶数集合设置 ID 的视频片段被用于训练, 其余具有奇数设置 ID 的片段被用于测试.

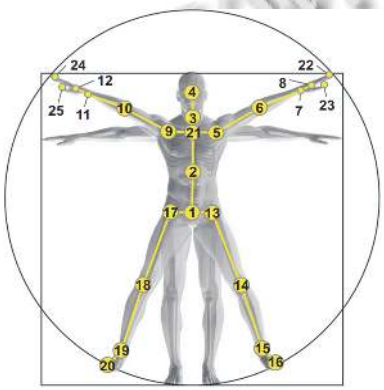

图 4 NTU 数据集人体结构图

3.2 实验细节

为了更加公平地与 ST-GCN 网络[18] 进行比较, 本文对 ST-GCN 网络进行复现, 同时使得本文的实验分析更加可靠. 原始的 ST-GCN 网络一共包含了 9 层 GCN Layer, TC-block 的卷积核大小为 9. 网络框架每个阶段的输出通道数量分别为 64, 128, 256. 模型输入样本帧数为 300. 本文参照 2s-AGCN[19] 官方公布的代码, 所搭建的复现模型一共包含了 10 层 GCN Layer, TC-block 的卷积核大小为 5. 此外, 实验数据预处理方法参照了 SGN 模型[20] 所提出的方法, 并且固定输入模型的每个样本帧数为 48.

表 1 展示了在 NTU-RGB+D 的 X-View 评估基准上, ST-GCN 网络的复现结果. 其中 ST-GCN 代表原论文所展示的准确率, ST-GCN*代表复现结果, 而带自适应拓扑图的 ST-GCN*是参照了 2s-AGCN 提出的方法. 最后, 我们选用带自适应拓扑图的 ST-GCN*网络作为本文所有实验的基准模型. 除非有必要的说明, 本文所有消融实验都是在 NTU-RGB+D 数据集 X-View 评估基准上进行的.

表 1 不同骨架几何特征建模模块的 GEO-GCN 模型在 NTU-RGB 数据集 X-View 评估上的准确率比较

方法	自适应邻接矩阵	模型参数量 (M)	几何特征建模模块	X-View (%)
ST-GCN[18]	—	—	—	88.3
ST-GCN*	—	2.074	—	90.5
ST-GCN*	√	2.092	—	92.9
GEO-GCN	√	2.099	—	93.6
GEO-GCN	√	2.110	嵌入	93.8
GEO-GCN	√	2.168	GCN	**94.0**

本文所有模型使用随机梯度下降 (stochastic gradient descent, SGD) 优化器进行训练, 并且设置动量为 0.9, 权重衰减为 0.000 1. 训练 epochs 设置为 65, 在前 5 个 epochs 中使用 warmup strategy [12], 以使训练过程更加稳定. 设置初始学习率为 0.1, 并在第 30 个 epoch 和第 55 个 epoch 时以 0.1 的系数进行学习率衰减, batch size 大小设置为 64.

3.3 不同骨架几何特征建模模块的比较

从表 1 可得出, 在 NTU-RGB 数据集中 X-View 评估基准上, 本文提出的 3 种骨架几何特征建模模块所构建的 GEO-GCN 模型的准确率都比带自适应的 ST-GCN*模型的准确率要高, 实验结果说明了本文提出的骨架关节几何特征D能有效地融合到 ST-GCN 模型中, 从而提高 ST-GCN 基准模型的识别率. 特别地, 采用 GCN 建模模块的 GEO-GCN 模型比基准模型的准确率要高出 1%. 对于 3 种不同的几何特征建模模块, 可以发现: 采用直接融合方法的 GEO-GCN 模型的性能提升幅度是最小的, 而采用 GCN 建模模块的 GEO-GCN 模型的准确率取得了最优效果. 综上, 可得出结论: 骨架关节几何特征D (帧内关节点间的距离) 能丰富 ST-GCN 模型所建模的时空特征, 并且基于 GCN 建模模块的 GEO-GCN 模型是有效方法, 该网络框架具有高效性和参数量较少的优势. 最后, 我们选取基于 GCN 建模模块的 GEO-GCN 模型作为后续实验的基准网络.

3.4 不同骨架几何特征数量的比较

为了进一步分析骨架几何特征D对 GEO-GCN 网络的影响, 本文对每个关节的几何特征数量进行了消

融实验. 图 4 展示了 NTU 数据集的人体结构关节点的序号. 在第 2.2 节中, 实验配置对每帧骨架内每个关节计算其与该帧上所有关节间的距离, 具体来说, 对于 NTU 数据集来说, 其关节点数量为 25, 因此每帧每个关节点共有 25 个距离几何特征. 为此, 本文设置关键关节点集合J_i, 其中J_i的下标表示集合内包含元素的个数, J_i中每个元素表示 NTU 人体结构图所对应的关节序号. 在给定J_i的情况下, 在计算每帧每个关节的几何特征时候, 只计算集合中内包含的元素所对应的关节点. 表 2 展示了不同J_i的元素组成, 对于每个关键关节点集合, 元素被选取的依据是: 在“直觉上”与动作信息相关性较大, 如, 序号 7 (左手腕) 关节点, 与人类执行动作过程的相关性较大.

表 2 不同关键关节点集合 J_i 的元素组成

i	元素
5	{1, 8, 12, 15, 19}
10	{1, 4, 6, 8, 10, 12, 15, 16, 19, 20}
20	{1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 21, 22, 24}
25	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25}

从表 3 实验结果可看出, 随着关键关节点集合J_i元素的逐渐增加, GEO-GCN 模型的准确率不是呈线性递增的, 即, 骨架关节几何特征数量与 GEO-GCN 模型的性能不是正相关, 其原因可能是: J_i中关键关节点的选取依赖于人的经验和直觉, 这导致通过距离公式所得的骨架关节几何特征包含了一定程度的先验知识, 从而影响了 GEO-GCN 模型的性能. 特别地, 当i =10 和 i =25 时候, 模型的性能近乎是一致的. 这说明了i =25 时, 骨架关节几何特征存在特征冗余问题. 最后, 考虑到i =25 时, GEO-GCN 模型取得了最优的准确率, 因此, 选取关键关节点集合J_{25}作为最终基准并与其他模型进行比较.

3.5 与其他模型方法比较

表 4 和表 5 中展示了 GEO-GCN 模型与其他模型在 NTU RGB+D 和 NTU RGB+D 120 上准确率的比较. 从表 4 实验结果可看出, 在 NTU RGB+D 两评估基准上, GEO-GCN (joint) 单模态模型的准确率高于非 GCN 的方法, 而与基于 GCN 的方法 (ST-GCN、AS-GCN 和 SGN) 性能相当. 特别地, 2s-AGCN[19] 采用了模型集成方法, 即, 关节坐标特征 (joint) 和关节骨头特征 (bone) 分别作为输入模态, 同样地, 本文展示了 GEO-GCN 模型集成方法的准确率. 从表 4 可看出, GEO-GCN (joint+bone) 集成方法的准确率高于 2s-AGCN 的准确率, 在 X-Sub 评估基准上高了约 2%. 这说明了本文提出的利用骨架关节几何特征作为 ST-GCN 框架特征补充的方法是高效的. 从表 5 实验结果可看出, 在 NTU RGB+D 120 两评估基准上, GEO-GCN 单模态模型的准确率都比 2s-AGCN 和 SGN 的准确率高. 这说明了本文提出的骨架关节几何特征在大型数据集上能更好地提高模型的性能. 特别地, 在 X-Sub 和 X-Set 评估基准上, GEO-GCN 集成模型的准确率比 2s-AGCN 分别高了 4.1% 和 3.4%. 综上, 可得出结论: 本文提出的 GEO-GCN 网络框架, 其充分利用了骨架关节几何特征作为 ST-GCN 模型的特征补充, 提高了框架的准确率同时不会使框架总参数量成倍数地增加, 是一种非常高效的网络框架.

表 3 不同关键关节点集合 J_i 的 GEO-GCN 在 NTU-RGB 数据集 X-View 评估上的准确率比较

方法	模型参数量 (M)	J_i	X-View (%)
GEO-GCN	2.161	i = 5	93.6
GEO-GCN	2.163	i = 10	93.9
GEO-GCN	2.167	i = 20	93.8
GEO-GCN	2.168	i = 25	**94.0**

表 4 不同算法在 NTU-RGB+D 上的准确率比较 (%)

方法	X-Sub	X-View
Lie Group [7]	50.1	82.8
ST-LSTM [11]	81.8	88.0
HCN[17]	86.5	91.1
ST-GCN[18]	81.5	88.3
AS-GCN[25]	86.8	94.2
2s-AGCN[19]	88.5	95.1
SGN[20]	89.0	94.5
GEO-GCN (joint)	88.2	94.0
GEO-GCN (joint+bone)	**90.1**	**95.4**

表 5 不同算法在 NTU-RGB+D 120 上的准确率比较 (%)

方法	X-Sub	X-Set
2s-AGCN[19]	82.9	84.9
SGN[20]	79.2	81.5
GEO-GCN (joint)	83.6	84.7
GEO-GCN (joint+bone)	**87.0**	**88.3**

4 结论与展望

本文提出了融合几何特征的图卷积网络框架, 其称为 GEO-GCN 网络框架. 该框架利用骨架序列中所蕴含的距离几何特征作为 ST-GCN 基准网络的特征补充. 然后, 本文利用 GCN 建模模块对骨架关节几何特征进行建模, 充分提取有判别力的时空特征, 并且利用早期特征融合方法, 将骨架关节几何特征高效地融合到 ST-GCN 网络中, 与双流模型方法相比较, 本文提出的 GEO-GCN 网络框架的参数量保持一个合适的范围内. 最后, 在 NTU-RGB+D 数据集和 NTU-RGB+D 120 数据集上, 本文进行了充分实验. 实验结果表明: 与 ST-GCN、2s-AGCN 和 SGN 等动作识别模型相比, 本文所提出的 GEO-GCN 网络框架取得了更好准确率的效果. 下一步的研究将会引入时间维度上的注意力模块, 提高网络建模时空特征能力.

参考文献

1 Pareek P, Thakkar A. A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications. Artificial Intelligence Review, 2021, 54(3): 2259–2322. [doi: 10.1007/s10462-020-09904-8]

2 Lo Presti L, La Cascia M. 3D skeleton-based human action classification: A survey. Pattern Recognition, 2016, 53: 130–147. [doi: 10.1016/j.patcog.2015.11.019]

3 Cao Z, Hidalgo G, Simon T, *et al*. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 172–186. [doi: 10.1109/TPAMI.2019.2929257]

4 Johansson G. Visual perception of biological motion and a model for its analysis. Perception & Psychophysics, 1973, 14(2): 201–211.

5 Wang J, Liu ZC, Wu Y, *et al*. Mining actionlet ensemble for action recognition with depth cameras. Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 1290–1297.

6 Hussein ME, Torki M, Gowayyed MA, *et al*. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing: AAAI Press, 2013. 2466–2472.

7 Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3D skeletons as points in a lie group. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 588–595. [doi: 10.1109/CVPR.2014.82]

8 Wang L, Huynh DQ, Koniusz P. A comparative review of recent Kinect-based action recognition algorithms. IEEE Transactions on Image Processing, 2019, 29: 15–28. [doi: 10.1109/TIP.2019.2925285]

9 Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1110–1118. [doi: 10.1109/CVPR.2015.7298714]

10 Zhu WT, Lan CL, Xing JL, *et al*. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix: AAAI, 2016. 3697–3703.

11 Liu J, Shahroudy A, Xu D, *et al*. Spatio-temporal LSTM with trust gates for 3D human action recognition. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 816–833.

12 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

13 Ding ZW, Wang PC, Ogunbona PO, *et al*. Investigation of different skeleton features for CNN-based 3D action recognition. Proceedings of 2017 IEEE International Conference on Multimedia & Expo Workshops. Hong Kong: IEEE, 2017. 617–622. [doi: 10.1109/ICMEW.2017.8026286]

14 Xu YY, Cheng J, Wang L, *et al*. Ensemble one-dimensional convolution neural networks for skeleton-based action recognition. IEEE Signal Processing Letters, 2018, 25(7): 1044–1048. [doi: 10.1109/LSP.2018.2841649]

15 Wang PC, Li WQ, Li CK, *et al*. Action recognition based on joint trajectory maps with convolutional neural networks. Knowledge-Based Systems, 2018, 158: 43–53. [doi: 10.1016/j.knosys.2018.05.029]

16 Li B, Dai YC, Cheng XL, *et al*. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. Proceedings of 2017 IEEE International Conference on Multimedia & Expo Workshops. Hong Kong: IEEE, 2017. 601–604.

17 Li C, Zhong QY, Xie D, *et al*. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. Proceedings of the 27th International Joint Conference on Artificial Intelligence.

Stockholm: IJCAI.org, 2018. 786–792. [doi: 10.24963/ijcai.2018/109]

18 Yan SJ, Xiong YJ, Lin DH. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 7444–7452.

19 Shi L, Zhang YF, Cheng J, *et al*. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12018–12027. [doi: 10.1109/CVPR.2019.01230]

20 Zhang PF, Lan CL, Zeng WJ, *et al*. Semantics-guided neural networks for efficient skeleton-based human action recognition. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1109–1118. [doi: 10.1109/CVPR42600.2020.00119]

21 Chen C, Zhuang YT, Nie FP, *et al*. Learning a 3D human pose distance metric from geometric pose descriptor. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(11): 1676–1689. [doi: 10.1109/TVCG.2010.272]

22 Zhang SY, Yang Y, Xiao J, *et al*. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks. IEEE Transactions on Multimedia, 2018, 20(9): 2330–2343. [doi: 10.1109/TMM.2018.2802648]

23 Li CK, Wang PC, Wang S, *et al*. Skeleton-based action recognition using LSTM and CNN. Proceedings of 2017 IEEE International Conference on Multimedia & Expo Workshops. Hong Kong: IEEE, 2017. 585–590. [doi: 10.1109/ICMEW.2017.8026287]

24 Yang F, Wu Y, Sakti S, *et al*. Make skeleton-based action recognition model smaller, faster and better. Proceedings of the ACM Multimedia Asia. Beijing: ACM, 2019. 31. [doi: 10.1145/3338533.3366569]

25 Li MS, Chen SH, Chen X, *et al*. Actional-structural graph convolutional networks for skeleton-based action recognition. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3590–3598. [doi: 10.1109/CVPR.2019.00371]

26 Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017. 1–14.

27 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org, 2015. 448–456.

28 何俊, 张彩庆, 李小珍, 等. 面向深度学习的多模态融合技术研究综述. 计算机工程, 2020, 46(5): 1–11. [doi: 10.19678/j.issn.1000-3428.0057370]

29 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 568–576.

30 Shahroudy A, Liu J, Ng TT, *et al*. NTU RGB+D: A large scale dataset for 3D human activity analysis. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1010–1019. [doi: 10.1109/CVPR.2016.115]

31 Liu J, Shahroudy A, Perez M, *et al*. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(10): 2684–2701. [doi: 10.1109/TPAMI.2019.2916873]

(校对责编: 孙君艳)