计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(5):187−194 [doi: 10.15888/j.cnki.csa.009485]

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

基于自适应人体拓扑结构引导的步态识别①

徐　颖, 朱　明

(中国科学技术大学 信息科学技术学院, 合肥 230026)
通信作者: 徐　颖, E-mail: xuying7@mail.ustc.edu.cn

摘　要: 不同于基于外形的步态识别方法, 基于关键点的步态识别方法采取人体关键点作为模型的输入, 能够有效避免数据集带来的背景噪声干扰; 其次, 现有的基于关键点的步态识别方法忽略了人体结构先验知识的利用, 且更倾向于提取局部特征, 从而忽略了全局上的关联性. 本文提出了一个基于关键点的步态识别框架 GaitBody, 能够从步态关键点序列中提取更有分辨性的特征. 首先, 我们设计了带有较大卷积核的多尺度卷积模块来提取多粒度的时序特征; 其次, 我们利用自注意力机制来提取空间特征, 并在此基础上引入了人体结构拓扑信息来进一步利用人体结构的先验知识; 最后, 为了更好使用时序信息, 我们生成最有代表性的时序特征, 并将其引入到自注意模块来融合时序和空间特征. 在 CASIA-B 和 OUMVLP-Pose 数据集上的实验结果表明, 我们的方法在基于关键点的步态识别方法上取得了最优结果, 消融实验也证明了各个模块的有效性.

关键词: 自注意力机制; 多尺度卷积; 先验知识; 基于关键点步态识别; 深度学习

引用格式: 徐颖,朱明.基于自适应人体拓扑结构引导的步态识别.计算机系统应用,2024,33(5):187–194. http://www.c-s-a.org.cn/1003-3254/9485.html

Adaptive Human Body Topology Guidance for Gait Recognition

XU Ying, ZHU Ming

(School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: Unlike appearance-based methods whose input may bring in some background noises, skeleton-based gait representation methods take key joints as input, which can neglect the noise interference. Meanwhile, most of the skeleton-based representation methods ignore the significance of the prior knowledge of human body structure or tend to focus on the local features. This study proposes a skeleton-based gait recognition framework, GaitBody, to capture more distinctive features from the gait sequences. Firstly, the study leverages a temporal multi-scale convolution module with a large kernel size to learn the multi-granularity temporal information. Secondly, it introduces topology information of the human body into a self-attention mechanism to exploit the spatial representations. Moreover, to make full use of temporal information, the most salient temporal information is generated and introduced into the self-attention mechanism. Experiments on the CASIA-B and OUMVLP-Pose datasets show that the method achieves state-of-the-art performance in skeleton-based gait recognition, and ablation studies show the effectiveness of the proposed modules.

Key words: self-attention mechanism; multi-scale convolution; prior knowledge; skeleton-based gait recognition; deep learning

1　引言

步态是一种可以从较远距离进行采集, 且不需要被采集对象配合的一种生物特征[1], 能够被用来识别个体. 因此, 步态识别技术在犯罪预防, 法医鉴定和社会

① 基金项目: 科技创新特区计划 (20-163-14-LZ-001-004-01)
收稿时间: 2023-11-08; 修改时间: 2023-12-11; 采用时间: 2023-12-20; csa 在线出版时间: 2024-03-15
CNKI 网络首发时间: 2024-03-19

安全[2,3]等方面有着较大的应用价值. 然而, 步态识别仍然受到在现实场景中很常见的视角转换, 携带背包和衣物穿着这些外界因素的影响[4,5].

现有步态识别方法按输入类型划分为基于外形的步态识别方法, 和基于关键点的步态识别方法. 基于外形的步态识别方法目前更为主流, 它采用二值轮廓图作为模型输入, 也取得了较好的性能, 但二值轮廓图会引入一些非必要的特征, 如发型, 穿着的衣物等, 这会给模型带来干扰. 而基于关键点的步态识别方法采用的是人体关键点作为模型输入, 从而忽略了如穿着等外形带来的干扰, 由关键点序列组成的信息可以被视作是真正的步态表征. 因此, 基于关键点的步态识别有着更大的潜力. 而现有的步态识别方法要么忽视了人体结构这一先验知识, 要么倾向于提取局部信息而忽略了各结点之间的联系. 针对上述问题, 本文提出了一种基于关键点的步态识别方法 GaitBody, 来提高步态识别的准确性. 本方法利用带有较大尺寸卷积核的时序多尺度卷积模块来提取多粒度的时序特征, 并在此基础上生成最有代表性时序信息, 在空间自注意力模块中引入了人体拓扑结构信息来充分利用人体结构的先验知识和提取全局性的步态特征.

2 相关工作

(1) 基于外形的步态识别算法

根据模型输入信息的使用方式, 还可以继续细分为以下 3 类: 基于步态能量图的步态识别[6,7], 基于集合的步态识别[8]和基于序列的步态识别[9–15]. 步态能量图 (gait energy image, GEI)[6]是通过将二值轮廓图在时序尺度上进行平均池化操作, 将一个步态序列信息压缩到一帧图片中而产生的. 这些方法在简化模型处理过程的同时, 也会使模型损失一部分具有分辨性的时序特征. 而在基于集合的步态识别方法中, Chao 等人提出了 GaitSet[8], 将步态序列看成是一个无序的集合, 因此该方法专注于建模空间特征, 而忽略时序特征之间的依赖. 基于序列的步态识别方法将步态序列按帧输入, 并按顺序逐帧进行处理, 提取各帧之间的潜在联系, 在提取步态空间特征的同时, 更加注重于捕捉步态序列的时序特征. Fan 等人提出 GaitPart[9], 通过捕捉输入外形的局部特征, 并利用微动作捕捉模块来建模时序依赖来提取步态特征. 而 Lin 等人提出了 GaitGL[10], 认为全局步态表征忽略了局部细节, 而局部特征不能捕捉相邻区域之间的关系, 因此设计全局局部卷积层来提取步态特征; Huang 等人提出 CSTL[11], 注重于通过多尺度时序卷积来提取时序上下文信息. Ma 等人提出 DANet[12], 利用动态自注意力机制来选择具有代表性的局部运动特征, 从而进一步学习鲁棒的全局运动特征.

(2) 基于关键点的步态识别算法

Liao 等人提出 PoseGait[16], 根据输入关键点, 计算一些手工特征作为模型的先验知识, 如关节角度, 骨骼长度和关节运动信息来克服穿着上变化带来的影响. 但是手动设置的特征在模型迭代中灵活性较差, 且在实际生活运用中, 可能会因为摄像头角度变化带来的影响从而降低识别精度. Teepe 等人提出 GaitGraph[17], 在步态识别领域第 1 次提出利用图卷积网络 (graph convolutional network, GCN) 来建模人体关节点的结构信息, 采取将人体拓扑结构[18]随模型迭代而更新的策略来建模潜在的人体结构模型[17–19], 取得了较好的性能. 但是利用图卷积操作会使模型更倾向于提取局部关键点特征, 从而忽略了全局特征建模. 针对这个问题, Pinyoanuntapong 等人提出 GaitMixer[20], 利用自注意力机制来提取各关键点之间的潜在联系, 并利用大尺度的时序卷积提取时序特征, 进一步提高了识别的精度, 但是忽视了人体结构模型的建模, 并且采用单一尺度时序卷积提取步态特征, 对较为复杂的步态特征提取有不利影响. Huang 等人提出了 CAG[21], 设计了一种条件自适应图卷积网络来动态适应各种行走方式和视角变化来提取步态特征, 但是模型在大数据集上的泛化能力较弱. 综上所述, 本研究旨在设计一个基于关键点的步态识别算法, 能够自适应提取步态关键点之间潜在的空间联系, 建模人体拓扑结构, 同时能建模多尺度时序特征, 且具有较好的模型泛化能力.

3 基于关键点的步态识别算法

3.1 网络结构

本文提出的步态识别方法的整体框架如图 1 所示. GaitBody 将有 T 帧 J 个关键点的步态关键点序列作为模型的输入, 可以表示为 $X \in R^{C_0 \times T \times J}$, 其中 C_0 是通道的维度. 首先 X 被传入到位置编码模块, 先利用线性映射层进行映射, 并在映射结果上加上位置编码来生成编码后的特征 $E_0 \in R^{C_1 \times T \times J}$, 这里 C_1 表示 embedding block 的输出维度. 随后, 将特征 E_0 传入到 N 个由空间自注意力模块和时序多尺度模块交替堆叠形成的模块,

来提取步态的时间空间特征. 这里在空间自注意力模块引入人体结构的先验知识用来提取每帧中各个节点之间的隐藏关联, 该模块的输出特征可以表示为$E_i^0 \in R^{C_i^0 \times T \times J}$, $i \in 1,2,\cdots,N$, 是N个堆叠模块的索引, 0代表的是空间模块. 利用时序多尺度卷积模块来捕捉多粒度的步态特征, 该模块由 4 个并行的卷积核尺寸不同的卷积组成, 并最后在输出上额外添加一个池化操作, 该模块的输出特征可以表示为$E_i^1 \in R^{C_i^1 \times T \times J}$和$E_i^{\text{Salient}} \in R^{C_i^1 \times 1 \times J}$, 其中$E_i^1$多尺度时序特征, E_i^{Salient}表示最有代表性的时序特征, 1表示时序模块. 最后将时空特征在时间和空间维度上进行二维自适应平均池化操作得到最终步态特征. 最后将特征分别进行线性映射传入到损失函数中. 本文采用 triplet loss 和 cross entropy loss 作为模型最终的损失函数.

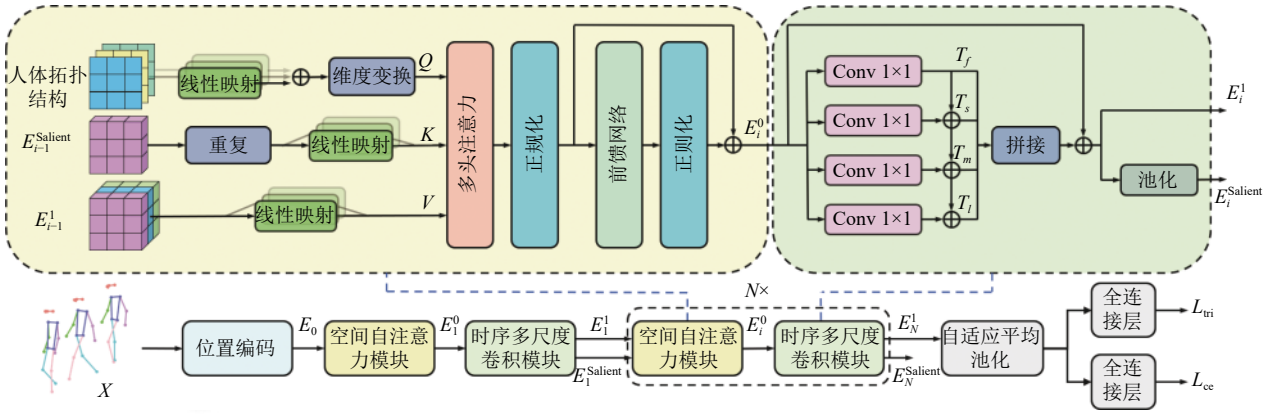

图 1 基于关键点的多视角步态识别算法流程图

3.2 时序多尺度卷积模块

为了捕捉多粒度的时序特征, 本文提出了一个由 4 个并行的卷积核尺寸大小不同的卷积分支组成的多尺度卷积模块, 其中 4 个卷积分支的卷积核大小为 1×1, 11×1, 21×1和31×1. 与传统方法不同的是, 我们采用了更大的卷积核尺寸来提取全局的时序特征. 由上述多尺度卷积分支提取的多粒度时序信息由下述公式进行融合:

$$T_f = \text{Conv}_{1\times1}\left(E_i^0\right) \tag{1}$$

$$T_s = \text{Conv}_{11\times1}\left(E_i^0\right) + T_f \tag{2}$$

$$T_m = \text{Conv}_{21\times1}\left(E_i^0\right) + T_s \tag{3}$$

$$T_l = \text{Conv}_{31\times1}\left(E_i^0\right) + T_m \tag{4}$$

$$E_i^1 = \text{Concat}\left(T_f + T_s + T_m + T_l\right) + \text{Residual}\left(E_i^0\right) \tag{5}$$

$$E_i^{\text{Salient}} = \text{Pool}\left(E_i^1\right) \tag{6}$$

其中, T_f, T_s, T_m, T_l表示帧级, 短期, 中期和长期时序特征. Concat表示拼接操作, Residual表示残差连接, Pool表示平均池化操作. 考虑到各个最有代表性的关节点特征会出现不同帧中, 在时序维度上提取最有代表性的时序特征比直接使用时序特征更加适合来捕捉步态时序信息. 因此, 本文在时序维度上利用平均池化操作来提取最有代表性的时序特征作为后续的空间自注意力模块的一个输入.

3.3 空间自注意力模块

本文采用自注意力机制来捕捉人体步态的全局信息. 然而, 相较于图片, 步态关键点损失了人体的结构化信息. 为了充分利用人体结构化信息, 我们引入了人体拓扑结构[18]来建模人体空间结构. 人体拓扑结构[18]即采用邻接图的方式来保存人体各个结点之间的位置信息, 可以保存结点之间空间位置关系. 在利用拓扑信息前, 本文先利用线性映射层将拓扑信息映射到高维空间, 这样拓扑信息可以随模型的学习而不断迭代, 从而建模隐性的人体结构模型. 与 GaitGraph[17]和 GaitGraph2[19]使用图卷积的方式来利用人体拓扑结构不同, 本文将映射后的人体拓扑结构作为自注意力模块的 query, 从而避免了让模型局限于捕捉局部性特征, 与此同时, 将人体拓扑结构作为 query 能够有效引导模型捕捉潜在的人体结构. 本文将映射后的人体结构拓扑信息作为自注意力模块的 query, 最有代表性的时序表征E_i^{Salient}作为模块的 key 来生成注意力图谱, 将时序特征E_{i-1}^1作为 value, 与生成的注意力图谱进行融合. 至于

第 1 个空间自注意力模块, 由于没有提取的时序特征作为输入, 我们直接将编码后特征E_0模块的 key 和 value 进行输入. 最后, 本文将生成的注意力结果传递给多层感知机, 来进一步在通道维度上进行融合.

3.4 模型输出与损失函数

为了得到模型识别结果, 我们将最后一个时序多尺度卷积模块的输出E_N^1在时间和空间维度上进行二维自适应平均池化操作, 并将池化结果输入到一层线性映射层中, 从而得到模型的输出, $E_{\text{output}} \in R^{1\times D}$, D 表示特征向量的维度.

本文采取将 triplet loss 和 cross entropy loss 组合的方式来训练模型. 本文将模型输出E_{output}分别输入到两个独立的线性映射层中, 并分别得到两个输出$E_{\text{tri}} \in R^{1\times D}$和$E_{\text{ce}} \in R^{1\times D}$, 将$E_{\text{tri}}$用于 triplet loss, E_{ce}用于 cross entropy loss. 最后组合损失函数L_{combined}可以表示为:

$$L_{\text{combined}} = L_{\text{tri}} + L_{\text{cse}} \tag{7}$$

其中, L_{tri}表示 triplet loss, L_{cse}表示 cross entropy loss.

4 实验分析

4.1 数据集

(1) CASIA-B

CASIA-B 数据集[5]是由中国科学院自动化研究所提出的开源步态数据集. 数据集中包含 124 个个体, 每个个体都有 11 个不同的视角 (0, 18°, ⋯, 180°), 3 种不同的行走状态, 如正常行走 (NM), 背包行走 (BG) 和穿大衣行走 (CL), 其中正常行走状态下步态序列共有 6 条 (NM01, ⋯, NM06), 背包行走状态共有 2 条 (BG01, BG02), 穿大衣行走共有 2 条 (CL01, CL02), 即每个个体共 110 条步态序列.

(2) OUMVLP 和 OUMVLP-Pose

OUMVLP[4]是由大阪大学提供的大型开源步态识别数据集, 其中包含 10307 个个体, 每个个体包括 14 个不同的视角 (0, 15°, ⋯, 90°; 180°, ⋯, 270°). OUMVLP-Pose[22]是从大型数据集 OUMVLP 中提取而来, 并且与 OUMVLP 有相同的个体数和帧数. OUMVLP-Pose[22]包含两个数据集, 它们分别是由两个预训练的姿态估计模型, OpenPose[23]和 AlphaPose[24]提取而来.

4.2 实验设计

(1) 数据划分

针对 CASIA-B 数据集[5], 因为没有官方的数据划分策略, 我们采用与 Wu 等人提出的[25]中一致的划分方式. 本文将前 74 个个体作为训练集, 剩下的 50 个个体作为测试集. 在测试阶段, 将正常行走状态下的前 4 条步态序列作为注册集 (Gallery set), 剩下的 6 条步态序列作为验证集 (Probe set). 针对 OUMVLP-Pose[22]数据集, 本文将数据集中前 5 153 个个体作为训练集, 剩下的 5 154 个个体作为测试集. 在测试阶段, 索引为#01 的步态序列被视作是注册集, 索引为#00 的步态序列被视作是验证集.

(2) 训练

针对在 CASIA-B[5]上的训练, 本文采用与 GaitGraph[17]相同的数据增强策略, 针对 OUMVLP-Pose[22], 本文采用与 GaitGraph2[19]中相同的数据增强策略. 针对 CASIA-B, 本文将迭代次数设置为 300 次, 针对 OUMVLP-Pose, 本文将迭代次数设置为 1 000 次. 本文将学习率设置为 5E–3, 采用 1-cycle 学习率调度器, 将权重消失设置为 1E–5, 并采用与 GaitMixer[20]中相同的数据采样策略. 本文将 CASIA-B 的批大小设置为 74×4, 其中 74 是个体数, 4 是单个个体的步态序列数量, 将 OUMVLP-Pose 的批大小设置为 200×4.

(3) 测试

本文将传递到 triplet loss 中的特征作为模型的最终输出, 并采用余弦相似度来衡量注册集和验证集之间的距离来进行步态匹配.

4.3 与现有方法的对比实验

(1) CASIA-B

1) 如表 1 所示, 本方法在所有状态下都比其他基于关键点方法的性能要高, 这证明了本方法的有效性. 2) 相较于基于外形的方法, 本方法展现了一定的竞争力, 在平均精度上, 本方法比 GaitSet 和 GaitPart 分别高了 5.9% 和 1.9%, 并进一步缩短了与 GaitGL 之间的差距. 特别的, 在穿外套 (CL) 条件下, 本方法性能大幅度超越了基于外形的方法, 比 GaitSet, GaitPart 和 GaitGL 分别高了 16.8%, 8.5% 和 3.6%, 这证明了在外形轮廓受到干扰时, 基于关键点步态识别算法的潜力. 3) 本方法在状态发生变化时, 模型性能更加鲁棒. 如表 2 所示, 在模型保持较高性能的同时, 本方法在行走状态变化时的标准差是最小的. 这能在真实场景中保持良好的模型稳定性.

(2) OUMVLP 和 OUMVLP-Pose

如表 3 所示, 本文将提出的 GaitBody 与 Gait-

Graph2[19], CAG[21]和基于外形的 GaitGL[10]进行比较. 其中, GaitBody, GaitGraph 和 CAG 是基于关键点的方法, 在 OUMVLP-Pose 数据集上测试, GaitGL 是基于外形的方法, 在 OUMVLP 数据集上进行测试, 但是因为 CAG 没有在 OpenPose[23]提取的数据集上测试, 所以只在 AlphaPose[24]提取的数据集上与其比较. 从结果可以得知, 在基于关键点的方法中, 本方法在准确度上有着较大提升, 这也显示了本方法在大型数据集上的泛化能力. 与基于外形的方法相比, 虽然本方法精度比基于外形的步态识别方法低, 但是相较于以往的基于关键点的步态识别方法, 本方法进一步缩小了与基于外形的步态识别方法之间的差距.

表 1 模型在 CASIA-B 上的结果 (%)

行走状态	分类	模型	0	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	平均值
NM	基于外形的方法	GaitSet[8]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
		GaitPart[9]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
		GaitGL[10]	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	基于关键点的方法	PoseGait[16]	55.3	69.9	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
		GaitGraph[17]	85.3	88.5	91.0	92.5	87.2	86.5	88.4	89.2	87.9	85.9	81.9	87.7
		GaitMixer[20]	94.4	**94.9**	94.6	**96.3**	95.3	**96.3**	**95.3**	94.7	95.3	94.7	92.2	94.9
		Ours	**94.6**	**94.9**	**95.3**	95.3	**95.8**	96.2	94.8	**95.3**	**95.9**	**95.3**	**93.9**	**95.2**
BG	基于外形的方法	GaitSet[8]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
		GaitPart[9]	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
		GaitGL[10]	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
	基于关键点的方法	PoseGait[16]	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
		GaitGraph[17]	75.8	76.7	75.9	76.1	71.4	73.9	78.0	74.7	75.4	75.4	69.2	74.8
		GaitMixer[20]	83.5	85.6	88.1	89.7	85.2	87.4	84.0	84.7	84.6	87.0	81.4	85.6
		Ours	**88.4**	**89.1**	**89.6**	**91.8**	**89.7**	**91.1**	**89.8**	**90.1**	**90.6**	**90.4**	**86.1**	**89.7**
CL	基于外形的方法	GaitSet[8]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
		GaitPart[9]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
		GaitGL[10]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	基于关键点的方法	PoseGait[16]	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	36.0
		GaitGraph[17]	69.6	66.1	68.8	67.2	64.5	62.0	69.5	65.6	65.7	66.1	64.3	66.3
		GaitMixer[20]	81.2	83.6	82.3	83.5	84.5	84.8	86.9	88.9	87.0	85.7	81.5	84.5
		Ours	**81.9**	**85.7**	**88.1**	**90.1**	**87.9**	**86.7**	**88.6**	**89.9**	**90.7**	**88.0**	**81.8**	**87.2**

表 2 模型在 CASIA-B 上平均精度和标准差 (%)

分类	方法	Probe			平均	标准差
		NM	BG	CL		
基于外形的方法	GaitSet[8]	95.0	87.2	70.4	84.2	10.3
	GaitPart[9]	96.2	91.5	78.7	88.8	7.4
	GaitGL[10]	97.4	94.5	83.6	91.8	5.9
基于关键点的方法	PoseGait[16]	68.7	44.5	36.0	49.7	13.9
	GaitGraph[17]	87.7	74.8	66.3	76.3	8.8
	GaitMixer[20]	94.9	85.6	84.5	88.3	4.7
	Ours	95.2	89.7	87.2	90.7	3.3

4.4 消融实验

(1) 空间自注意力模块

为了证明提出的空间自注意力模块的有效性, 本文对比了不同配置下的模型性能, 性能结果如表 4 所示. 根据前 3 组实验, 我们可以知道引入拓扑信息或者引入最有代表性的时序信息能够有效提高模型的性能, 同时引入两部分信息, 并且即将拓扑信息作为自注意力机制的 query, 将最有代表性的时序信息作为 key 能够最大提高模型性能. 这证明了人体拓扑结构和最有代表性时序特征有助于模型的特征提取.

(2) 时序多尺度卷积模块

为了验证多尺度时序卷积模块中卷积分支数量以及卷积核大小设置的合理性, 本文针对模块在不同设置下的模型性能设计实验, 实验结果如表 5 所示. 根据前 4 组实验可知, 卷积核尺寸设置为 31 时, 模型精度最高, 这可能是因为步态特征是一种全局性的表征, 而随着卷积核尺寸增大, 模型更加倾向于注重全局信息, 从而提高模型的识别精度. 当卷积核设置为 41 时, 模型精度大幅度下降, 这可能是因为卷积核设置过大从而使模型无法有效提取时序特征. 根据前 11 组实验, 我们发现采用两个并行卷积分支的性能比采用卷积核

大小为 31 的单分支卷积性能差. 这可能是因为两个分支信息差距较大, 信息融合时会互相干扰. 并且当双分支模型中一个分支卷积核大小为 41 时, 随着另一个分支卷积核的增大, 模型的识别精度大幅下降, 这也从另一方面证明了局部特征的必要性. 根据最后 3 组实验, 我们发现当采用 4 个并行卷积分支时, 模型性能都好于单分支或双分支结构. 这可能是因为采取 4 个并行卷积能够包含更多时序粒度, 同时通过采取较短时序特征加上较长时序特征的融合方式能够减小特征之间的差距, 减小特征之间的相互干扰. 最后验证结果是, 当设置 4 个卷积分支, 并将卷积核分别设置为 1, 11, 21 和 31 时, 模型取得最高性能.

表 3 模型在 OUMVLP 和 OUMVLP-Pose 上的结果 (%)

视角	OUMVLP-Pose[22]					OUMVLP[5]
	OpenPose[23]		AlphaPose[24]			GaitGL[10]
	GaitGraph2[19]	Ours	GaitGraph2[19]	CAG[21]	Ours	
0	32.9	44.0	54.3	45.4	65.7	84.9
15°	47.7	61.1	68.4	61.2	77.1	90.2
30°	53.9	64.7	76.1	64.7	79.3	91.1
45°	56.8	68.7	76.8	67.6	80.9	91.5
60°	53.9	68.3	71.5	67.0	80.6	91.1
75°	54.7	65.1	75.0	63.5	78.1	90.8
90°	45.4	56.9	70.1	57.7	75.1	90.3
180°	29.0	47.2	52.2	39.9	59.1	88.5
195°	35.7	54.9	60.6	48.3	67.2	88.6
210°	34.3	50.8	57.8	44.0	61.5	90.3
225°	44.3	64.0	73.2	61.0	78.2	90.4
240°	46.2	63.4	67.8	60.8	77.7	89.6
255°	46.4	59.3	70.8	57.1	75.3	89.5
270°	38.4	51.2	65.3	52.1	71.3	88.8
平均	44.3	58.5	67.1	56.4	73.4	89.7

(3) 可视化

为了证明将人体结构的先验知识引入模型的重要性, 我们对模型的特征图进行了可视化操作. 特征图选自自适应平均池化层前的输出, 并将该输出在通道维度进行池化操作得到用来可视化的特征图. 如图 2 所示, 当引入了拓扑信息后, 模型更加倾向于注重人体的关键节点, 如手腕, 膝盖, 脚踝等, 这些节点是最能体现人体步态信息的. 而没有引入拓扑信息的模型更倾向于平均地关注各个节点, 如图 3 所示, 相较于前者, 这显然是次优的. 根据可视化结果, 我们推断出引入拓扑结构后, 模型能够更好地建模人体结点之间的隐性关联.

表 4 模型在 CASIA-B 数据集上不同配置空间自注意力模块的结果 (%)

模型结构	Probe			平均
	NM	BG	CL	
标准的自注意力机制	94.3	84.7	83.9	87.6
+Topology as query	94.1	87.0	83.7	88.3
+Salient as key	94.8	85.4	86.1	88.8
+Both	95.2	89.7	87.2	90.7
+Both (swap)	95.4	87.6	86.4	89.8

表 5 模型在 CASIA-B 数据集上不同配置时序多尺度模块的结果 (%)

卷积核大小					Probe			平均
1	11	21	31	41	NM	BG	CL	
—	√	—	—	—	93.0	85.0	82.6	86.9
—	—	√	—	—	94.4	86.5	85.4	88.8
—	—	—	√	—	94.6	87.4	85.9	89.3
—	—	—	—	√	93.6	82.0	73.5	83.3
√	—	—	√	—	95.1	86.3	83.2	88.2
—	√	—	√	—	95.2	87.8	84.9	89.3
—	—	√	√	—	95.0	85.5	83.8	88.1
√	—	—	—	√	**95.3**	86.9	85.3	88.2
—	√	—	—	√	94.8	87.1	84.5	88.8
—	—	√	—	√	93.8	83.0	77.8	84.9
—	—	—	√	√	91.9	78.8	70.2	80.9
√	√	√	√	—	95.2	**89.7**	**87.2**	**90.7**
√	√	√	—	√	95.2	87.4	85.3	89.3
—	√	√	√	√	94.9	88.8	87.0	90.2

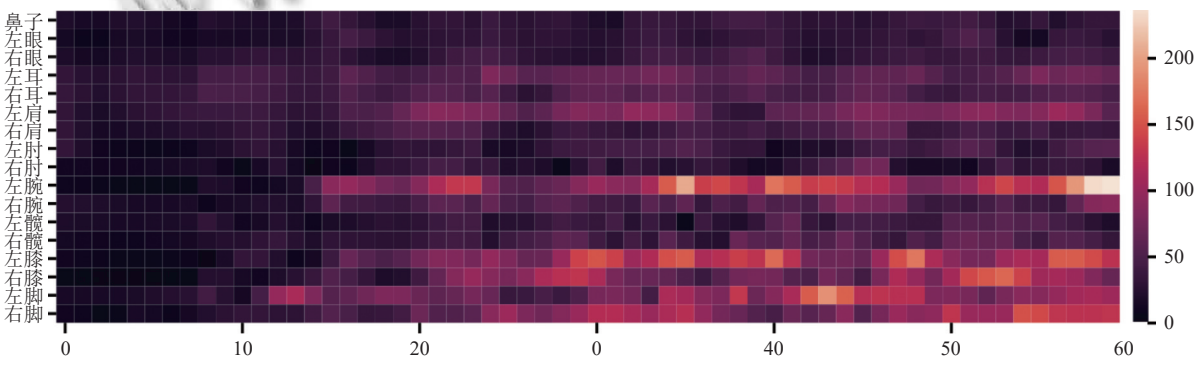

图 2 拍摄视角 72°, 添加人体拓扑结构信息的特征图

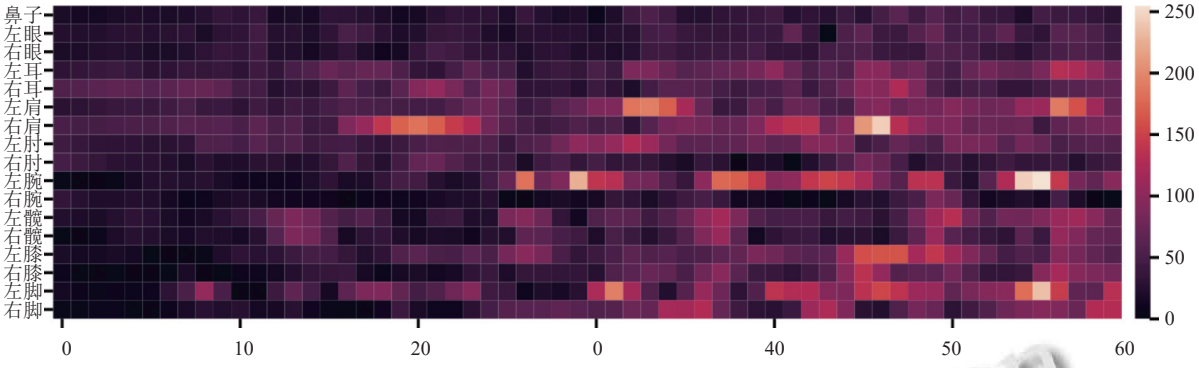

图 3 拍摄视角 72°, 未添加人体拓扑结构信息的特征图

5 结论与展望

本文提出了一个基于关键点的步态识别框架, GaitBody, 来提取步态序列中的关键步态表征. 为了充分利用人体结构的先验知识, 本文将人体结构的拓扑信息引入到自注意力机制中, 并作为 query. 除此之外, 为了利用每帧中最有代表性的关键点, 本文从时序特征中沿时序维度提取最有代表性的时序信息, 并作为自注意力机制中的 key 来提取步态信息的空间特征. 因为步态运动的复杂性, 本文采用带有较大卷积核的时序多尺度卷积模块来提取多粒度的时序特征. 在 CASIA-B 和 OUMVLP-Pose 数据集上的实验结果也证明了所提出框架的有效性.

参考文献

1 Wang L, Tan TN, Ning HZ, *et al*. Silhouette analysis-based gait recognition for human identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(12): 1505–1518. [doi: 10.1109/TPAMI.2003.1251144]

2 Bouchrika I, Goffredo M, Carter J, *et al*. On using gait in forensic biometrics. Journal of Forensic Sciences, 2011, 56(4): 882–889. [doi: 10.1111/j.1556-4029.2011.01793.x]

3 Macoveciuc I, Rando CJ, Borrion H. Forensic gait analysis and recognition: Standards of evidence admissibility. Journal of Forensic Sciences, 2019, 64(5): 1294–1303. [doi: 10.1111/1556-4029.14036]

4 Takemura N, Makihara Y, Muramatsu D, *et al*. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Transactions on Computer Vision and Applications, 2018, 10(1): 4. [doi: 10.1186/s41074-018-0039-6]

5 Yu SQ, Tan DL, Tan TN. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. Proceedings of the 18th International Conference on Pattern Recognition. Hong Kong: IEEE, 2006. 441–444.

6 Han J, Bhanu B. Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(2): 316–322. [doi: 10.1109/TPAMI.2006.38]

7 Shiraga K, Makihara Y, Muramatsu D, *et al*. GEINet: View-invariant gait recognition using a convolutional neural network. Proceedings of the 2016 International Conference on Biometrics. Halmstad: IEEE, 2016. 1–8.

8 Chao HQ, Wang K, He YW, *et al*. GaitSet: Cross-view gait recognition through utilizing gait as a deep set. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(7): 3467–3478.

9 Fan C, Peng YJ, Cao CS, *et al*. GaitPart: Temporal part-based model for gait recognition. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 14225–14233.

10 Lin BB, Zhang SL, Yu X. Gait recognition via effective global-local feature representation and local temporal aggregation. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 14648–14656.

11 Huang XH, Zhu DW, Wang H, *et al*. Context-sensitive temporal feature learning for gait recognition. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 12909–12918.

12 Ma K, Fu Y, Zheng DZ, *et al*. Dynamic aggregated network for gait recognition. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 22076–22085.

13 Wu HQ, Tian J, Fu YJ, *et al*. Condition-aware comparison scheme for gait recognition. IEEE Transactions on Image Processing, 2021, 30: 2734–2744. [doi: 10.1109/TIP.2020.3039888]

14 Huang Z, Xue DX, Shen X, *et al*. 3D local convolutional neural networks for gait recognition. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 14920–14929.

15 Zhang ZY, Tran L, Yin X, *et al*. Gait recognition via disentangled representation learning. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4710–4719.

16 Liao RJ, Yu SQ, An WZ, *et al*. A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognition, 2020, 98: 107069. [doi: 10.1016/j.patcog.2019.107069]

17 Teepe T, Khan A, Gilg J, *et al*. GaitGraph: Graph convolutional network for skeleton-based gait recognition. Proceedings of the 2021 IEEE International Conference on Image Processing. Anchorage: IEEE, 2021. 2314–2318.

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 Press, 2018. 912.

19 Teepe T, Gilg J, Herzog F, *et al*. Towards a deeper understanding of skeleton-based gait recognition. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 1569–1577.

20 Pinyoanuntapong E, Ali A, Wang P, *et al*. GaitMixer: Skeleton-based gait representation learning via wide-spectrum multi-axial mixer. Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023. 1–5.

21 Huang XH, Wang XG, Jin ZDQ, *et al*. Condition-adaptive graph convolution learning for skeleton-based gait recognition. IEEE Transactions on Image Processing, 2023, 32: 4773–4784. [doi: 10.1109/TIP.2023.3305822]

22 An WZ, Yu SQ, Makihara Y, *et al*. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2020, 2(4): 421–430. [doi: 10.1109/TBIOM.2020.3008862]

23 Cao Z, Simon T, Wei SE, *et al*. Realtime multi-person 2D pose estimation using part affinity fields. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 7291–7299.

24 Fang HS, Xie SQ, Tai YW, *et al*. RMPE: Regional multi-person pose estimation. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2334–2343.

25 Wu ZF, Huang YZ, Wang L, *et al*. A comprehensive study on cross-view gait based human identification with deep CNNs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(2): 209–226. [doi: 10.1109/TPAMI.2016.2545669]

(校对责编: 孙君艳)