计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(4):279−287 [doi: 10.15888/j.cnki.csa.009468]

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

基于视频残差神经网络的深度步态识别①

马玉祥, 代雪晶

(中国刑事警察学院 公安信息技术与情报学院, 沈阳 110854)
通信作者: 马玉祥, E-mail: 1467842033@qq.com

摘　要: 步态识别是根据人体的行走方式进行身份识别. 目前, 大多数步态识别方法通过浅层神经网络进行特征提取, 在室内步态数据集表现良好, 然而在近年新公布的室外步态数据集中性能表现不佳. 为了解决室外步态数据集带来的严峻挑战, 提出了一种基于视频残差神经网络的深度步态识别模型. 在特征提取阶段, 基于提出的视频残差块构建深层 3D 卷积神经网络 (3D CNN), 提取整个步态序列的时空动力学特征; 然后, 引入时序池化和水平金字塔映射降低采样特征分辨率并提取局部步态特征; 使用联合损失函数驱动训练过程, 最后通过 BNNeck 平衡损失函数并调整特征空间. 实验分别在公开的室内 (CASIA-B)、室外 (GREW、Gait3D) 这 3 个步态数据集上进行. 实验结果表明, 该模型在室外步态数据集中的准确率以及收敛速度优于其他模型.

关键词: 计算机视觉; 步态识别; 视频残差神经网络; 金字塔映射; 深度学习; 步态轮廓图像

引用格式: 马玉祥,代雪晶.基于视频残差神经网络的深度步态识别.计算机系统应用,2024,33(4):279–287. http://www.c-s-a.org.cn/1003-3254/9468.html

Deep Gait Recognition Based on Video Residual Neural Network

MA Yu-Xiang, DAI Xue-Jing

(School of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China, Shenyang 110854, China)

Abstract: Gait recognition is the process of identifying individuals based on their walking patterns. Currently, most gait recognition methods employ shallow neural networks for feature extraction, which performs well in indoor gait datasets but produces poor performance on the newly released outdoor gait datasets. To address the complicated challenges that arise from outdoor gait datasets, this study proposes a deep gait recognition model based on video residual neural networks. In the feature extraction phase, a deep 3D convolutional neural network (3D CNN) is constructed by the proposed video residual blocks to extract the spatio-temporal dynamics features of the entire gait sequence. Subsequently, temporal pooling and horizontal pyramid mapping are introduced to reduce the feature resolution of sampling data and extract local gait features. The training process is driven by a joint loss function, and finally loss functions are balanced and the feature space is adjusted by BNNeck. The experiments are conducted on three publicly available gait datasets, including both indoor (CASIA-B) and outdoor (GREW, Gait3D) gait datasets. The experimental results verify that the model outperforms other models in accuracy and convergence speed on outdoor gait datasets.

Key words: computer vision; gait recognition; video residual neural network; pyramid mapping; deep learning; gait silhouette image

① 基金项目: 公安部科技强警基础工作专项 (2016GABJC06); 中央高校基本科研业务费 (D2023001)
收稿时间: 2023-10-03; 修改时间: 2023-11-03, 2023-12-04; 采用时间: 2023-12-07; csa 在线出版时间: 2024-01-17
CNKI 网络首发时间: 2024-01-19

步态, 即人们行走或奔跑时的体态, 蕴含着丰富的个体信息[1]. 步态识别可以应用于医学[2]、运动科学、身份验证、视频侦查与安防系统[3]等不同领域. 与指纹、人脸、虹膜识别等生物特征识别相比, 步态识别拥有无需受试人员配合、可以进行远距离身份识别、对图像分辨率要求低, 难以隐藏和伪装等优势.

现有的步态识别方法大致可以分为两类: 基于模型的步态识别方法和基于轮廓的步态识别方法. 基于模型的步态识别方法倾向于将预测的人体结构作为输入, 即利用人体动力学知识建立人体真实三维模型, 将人体运动特征模型化, 根据运动数据得到模型参数, 而后在这些特征模型的基础上进行步态识别. 常见的模型有 2D Pose、3D Pose 以及 SMPL (skinned multi-person linear)[3]模型等. 为了解决视角转变以及人物衣着改变带来的准确率下降的问题, Liao 等人[4]认为 3D Pose 比步态轮廓图信息维度更低且更加紧凑, 提出 PoseGait 模型, 使用人体 3D Pose 以及人体的先验知识构建模型以准确定位身体部位; GaitGraph[5]引入了图卷积神经网络对基于 2D 骨骼图 (2D, skeleton-based) 进行特征处理; HMRGait[6]微调了预训练的 Human mesh recovery 网络构建了一个端到端的 SMPL 模型; 还有一些基于模型的步态识别方法结合多种形式的数据作为输入, 例如 SMPLGait[7]结合步态剪影序列和 SMPL 模型中的 3D 几何信息来提高步态外观特征学习; Peng 等人[8]将 skeletons 与步态轮廓图结合, 提出 Bimodal Fusion (BiFusion) 网络, 用以增强网络获取更具有区分性特征的能力. 虽然这些人体模型对外观变化更加鲁棒, 但该方法通常要求带有特殊功能的摄像装备, 并且由于真实环境中所拍摄视频的分辨率不高, 从而难以从中提取出完整的人体模板, 且准确率较低. 因此该方法在大多真实场景中缺乏实用性.

基于轮廓的步态识别专注于人体的运动带来的形态变换, 该方法将步态视频序列作为输入, 计算并建立相邻帧之间关系, 提取其中的时空特征进行步态识别. 因为人们可以感知一帧所处的大概时间位置, 而无需特意为序列信息进行建模, GaitSet[9]创新性地将步态序列视作一个无序集合, 通过最大值函数压缩帧级空间特征, 是近年来最具影响力的步态识别方法之一; GaitPart[10]提出微运动捕捉模块 (micro motion capture module, MCM) 强调对特征图局部细节进行特征提取; GaitGL[11]认为只提取整体的步态特征方法忽略了局部特征的重要性, 而只提取局部细节的步态特征方法又容易忽视相邻部分之间的关系, 因此提出了全局-局部卷积层 (global and local convolution layer) 分别提取全局与局部的时空步态特征; CSTL[12]受人类可以自适应地关注不同时间尺度的时间序列来区分不同目标的步态特征的启发, 提出了上下文敏感的时间特征学习网络 (context-sensitive temporal feature learning, CSTL); 因为人体各个部位并不是均匀分布以及运动时身体部位的改变, 将特征图平均水平划分的方法无法准确定位身体部位, 因此 3DLocal[13]使用局部 3D 卷积神经网络, 在特征提取使用自适应空间和时间尺度, 从而更好地学习身体部位的时空模式; 徐硕等人[14]提出双分支特征融合网络 (dual branch feature fusion network), 构建特征融合模块, 融合外观特征和姿态特征, 并引入通道注意力机制实现任意尺寸的特征融合; 王晓路等人[15]提出双支路卷积网络 (two-branch convolutional network), 分别使用水平金字塔映射 (horizontal pyramid mapping, HPM)[16]和 MCM 双分支共同提取步态特征. 随着公共区域监控摄像头大规模的部署, 基于轮廓的步态识别方法对图像分辨率要求不高, 并且能够有效地提取步态特征, 该方法已成为目前主流的步态识别方法.

本文所提方法属于基于轮廓的步态识别方法. 大多数步态识别方法通过浅而简单的神经网络提取步态特征[17,18], 在室内步态数据集中有着良好表现, 在被广泛使用的 CASIA-B[19]数据集的测试中平均 Rank-1 准确率均超过 87%. 但是由于模型的学习能力较弱或过拟合等问题, 导致在近年新公布的室外步态数据集 (GREW[20]以及 Gait3D[7]) 中大多方法 Rank-1 准确率都不足 50%, 部分方法性能相较室内数据集下降超过 40 个百分点. 文献[17]中提到当前步态识别领域所面临的问题, 目前现有的步态识别方法大多在室内数据集中验证其有效性 (如表 1 所示), 这将在实际应用中存在许多缺陷. 针对该问题, 本文提出一种基于视频残差神经网络的深度步态识别模型 VRGait. 本文方法的创新之处在于: 1) 设计了一种视频残差主干网络提取步态特征, 学习到步态序列的时空特征表示, 并且解除了步态识别中使用 3D 卷积操作必须固定帧长的这一限制; 2) 借鉴 GaitBase[17]中的特征处理方法, 结合视频残差网络, 提出了一种新的步态识别模型; 3) 在 3 个公开数据集上与主流方法进行实验对比, 验证了该方法

的有效性. 图 1 展示了现有的步态识别方法在不同数据集中的性能表现.

表 1 不同模型所使用的数据集信息汇总

模型	数据集	环境	年份
PoseGait	CASIA-B/CASIA-E	室内	2018
GaitGraph	CASIA-B	室内	2021
HMRGait	CASIA-B/OU-MVLP	室内	2021
SMPLGait	Gait3D	室外	2022
BiFusion	CASIA-B/OU-MVLP	室内	2023
GaitSet	CASIA-B/OU-MVLP	室内	2018
GaitPart	CASIA-B/OU-MVLP	室内	2020
GaitGL	CASIA-B/OU-MVLP	室内	2021
CSTL	CASIA-B/OU-MVLP GREW	室内 室外	2021
3DLocal	CASIA-B/OU-MVLP	室内	2021
双分支融合网络	CASIA-B	室内	2022
双支路卷积网络	CASIA-B	室内	2023
GaitBase	CASIA-B/OU-MVLP GREW/Gait3D	室内 室外	2023

1 VRGait

1.1 主要模块

受 Tran 等人[21]提出的 R3D (3D ResNet) 的启发, 构建了适用于步态识别的视频残差网络 (VideoResNet), 并引入步态识别领域常用的时序池化 (temporal pooling)、水平金字塔映射以及 BNNeck[22]等方法构成特征处理和推理模块, 命名为 VRGait, 其结构如图 2 所示. 图 2 中 DA 为数据增强; S 为一个批次中包含的序列数; c、h、w 指代特征图的通道数、高度以及宽度; Stem 与 Layer 为主干网不同的网络层; TP 为时序池化; HPM 为水平金字塔映射; 推理模块中的符号见第 1.1.3 节. 首先对输入网络的步态轮廓图进行数据增强操作; 之后主干网将提取步态轮廓图中的时空特征并转换为具有高、宽、通道以及序列四维特征图; 然后通过特征处理模块对特征进行聚合并提取局部步态特征; 最后通过推理模块平衡损失函数并调整特征空间得到预测结果.

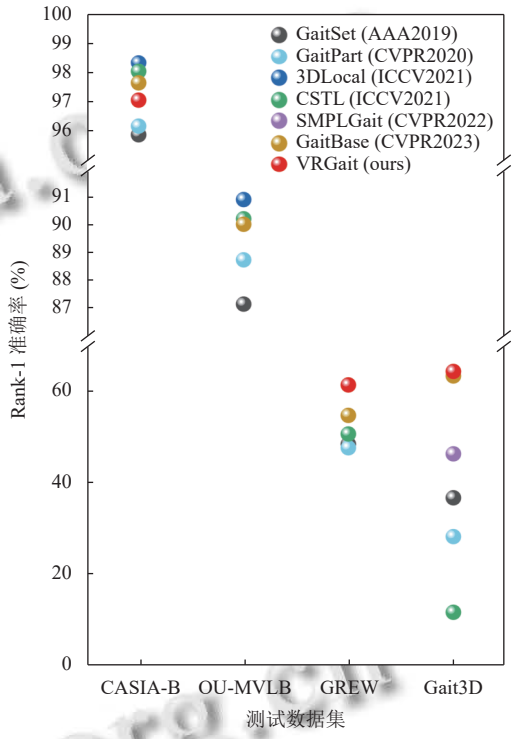

图 1 不同模型在室内外数据集中的性能对比

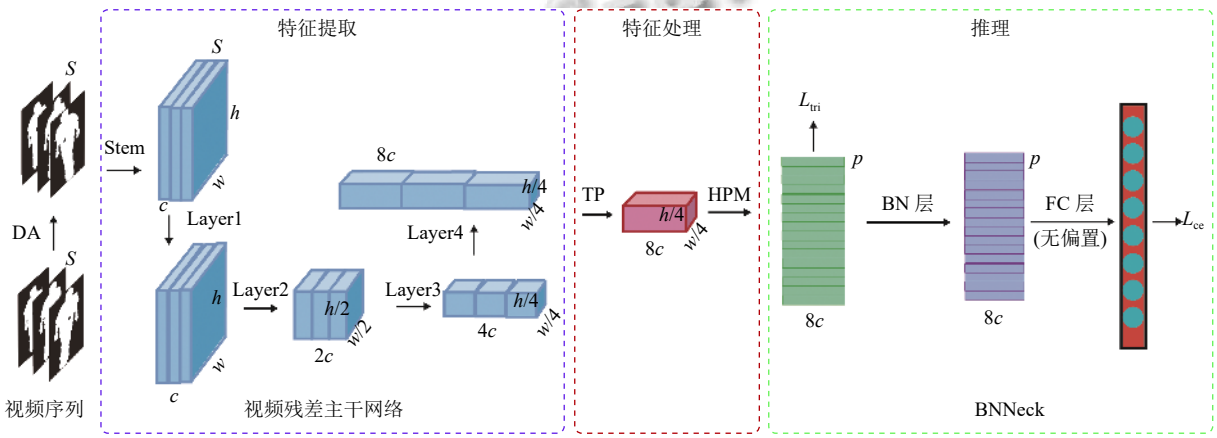

图 2 VRGait 框架图

1.1.1 特征提取模块

特征提取模块由 1 层 Stem 层和 4 层残差层构成. 其中, Stem 层由一个 2D 卷积构成, 一个 2D 批归一化以及 ReLU 激活函数构成, 其作用是二值轮廓图数据

转为拥有丰富的低级结构特征 (边缘、形状信息) 的浮点型特征图.

在行为识别领域中, 相较于 2D 卷积, 3D 卷积能够直接从视频中提取时空特征, 并对时序信息与行为模式进行建模. 而残差学习最早由 He 等人[23]提出并广泛应用于视觉任务中的深层网络. 因为步态识别的图像多为二值轮廓图, 跟视觉任务中的图片有很大的区别. 因此, 本文利用 Tran 等人[21]提出的 3D 残差块构建适用于步态识别的视频残差网络. 残差块结构如图 3.

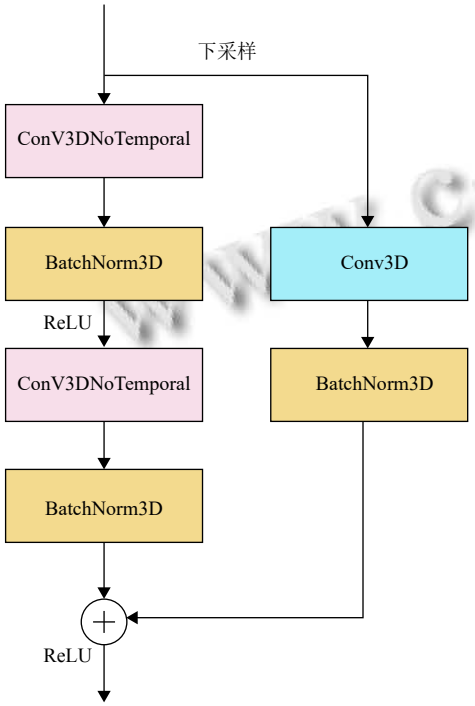

图 3 VideoResNet 残差块结构

图 3 中的“⊕”指对应元素相加; 图中 Con3DNoTemporal 使用特殊的参数配置, 使得卷积操作后不改变特征图中序列数, 因此解决了步态识别中使用 3D 卷积操作必须固定帧长的这一问题. 该卷积与图 3 中 Conv3D 的区别在于两者参数不同, Conv3D 不带有 padding, 而 Con3DNoTemporal 中的 padding 均为 (0, 1, 1). 此外 Conv3D 位于下采样分支, 因此使用不同名称用来区分两者. 另外在每次卷积操作后均带有 3D 批归一化 (使用 PyTorch 中默认参数) 以及 ReLU 激活函数. 经过下采样的特征将与图 3 中左侧分支的特征进行逐元素相加最终得到该层输出. 每层卷积操作的具体参数如表 2 所示 (Stem 中 2D CNN 的 padding 为 (1, 1)).

1.1.2 特征处理模块

为了从特征图中得到最具判别力时空特征, 使用了 Fan 等人[17]提出的 Baseline 中特征处理结构, 该模块分别包含时序池化层和水平金字塔映射.

表 2 VideoResNet 架构

层	卷积	卷积核, 通道	步长
Stem	Plain 2D CNN	[3×3, 64]	[1×1]
Layer1	Conv3DNoTemporal	$\begin{bmatrix}1\times3\times3,64\\1\times3\times3,64\end{bmatrix}$	$\begin{bmatrix}1\times1\times1\\1\times1\times1\end{bmatrix}$
Layer2	Conv3DNoTemporal	$\begin{bmatrix}1\times3\times3,64\\1\times3\times3,128\end{bmatrix}$	$\begin{bmatrix}1\times2\times2\\1\times1\times1\end{bmatrix}$
—	Conv3D	[1×1×1, 64]	[1×2×2]
Layer3	Conv3DNoTemporal	$\begin{bmatrix}1\times3\times3,128\\1\times3\times3,256\end{bmatrix}$	$\begin{bmatrix}1\times2\times2\\1\times1\times1\end{bmatrix}$
—	Conv3D	[1×1×1, 128]	[1×2×2]
Layer4	Conv3DNoTemporal	$\begin{bmatrix}1\times3\times3,256\\1\times3\times3,512\end{bmatrix}$	$\begin{bmatrix}1\times1\times1\\1\times1\times1\end{bmatrix}$

为了获取序列中时序步态特征最具有代表性的一帧, 时序池化层对特征的时间维度使用 max 函数压缩帧级特征. 式 (1) 中 X_{in} 代表特征提取模块所提取到的特征图; $F^{t}_{\max}$ 指时序池化操作; F_{TP} 为经过时序池化层后的输出特征.

$$F_{\text{TP}} = F^{t}_{\max}(X_{\text{in}}) \tag{1}$$

在行人重识别 (person re-identification) 研究领域中, HPM 可以消除人体关键部位缺失或者步态序列未对齐所带来的负面影响. HPM 对同一特征图按所设尺度进行重复水平分割, 再进行池化操作后将其依次堆叠, 形成金字塔状的特征图. HPM 结构如图 4 所示.

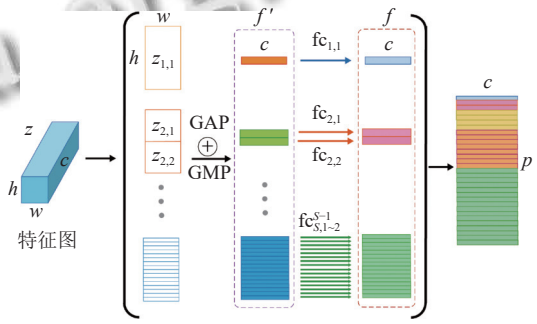

图 4 水平金字塔映射结构

图 4 中, h, w 分别代表特征图的高与宽; c 代表通道数; p 为水平特征向量的数量; S 代表 HPM 的尺度数目; $Z_{s,t}$ 为特征块在尺度 S 下的索引, 其中 $t \in 1, 2, \cdots, 2^{s-1}$; “⊕”指对应元素相加; GAP 与 GMP 指全局平均池化 (global average pooling, GAP) 和全局最大池化 (global max pooling, GMP); fc 指独立的全连接操作; f' 与 f 指代全连接操作前后的特征向量. 式 (2) 为 GAP 与 GMP

的计算公式:

$$f' = maxpool(Z_{s,t}) + avgpool(Z_{s,t}) \tag{2}$$

在原始论文中$S \in 1,2,3,\cdots,s$, 但 Fan 等人[17]通过进一步进行消融实验得知, 在移除 HPM 的多尺度机制后, 模型甚至超越了原来的性能, 并且减少了超过 80% 的训练权重, 因此本文选择只设置一个尺度 S=4, 特征图将会被水平分为 2^4=16 份. 之后进行池化操作, 结果为 GAP 与 GMP 之和. 因为只使用 GAP 可能会丢失一些非常具有区分性的特征, 例如一个人的某个部分具有显著特征, 但被背景包围, 从而导致该部分特征的低响应, 因此加入 GMP 解决该问题[16]. 通过分割与池化操作后, 3D 特征块将转为 1D 特征向量 (特征块的 H=16), 然后将特征向量在高度维度上进行级联. 本文将原有的 HPM 中 1×1 卷积层替换为独立的全连接层, 将水平特征向量 f'互不干扰地映射到深度判别空间得到 f. 独立的全连接层移除偏置项, 并使用 Xavier[24] 初始化对全连接层进行初始化以加快训练速度并提升网络性能. 该操作使身体各部分之间的耦合弱化, 对于衣着、遮挡等复杂条件下的步态特征有着更强的鲁棒性. 最终 f 将作为三元组损失 (triplet loss) 函数的输入.

1.1.3 推理模块

推理模块由 BNNeck 构成. BNNeck 的具体结构如图 2 中推理模块所示: p 为水平特征向量数量; L_{tri} 与 L_{ce} 为三元组损失与交叉熵损失; BN 层与 FC 层为批归一化层与全连接层. BNNeck 最早用于解决行人重识别领域在训练过程中使用联合损失函数产生一个损失函数下降而另一个损失函数震荡甚至上升的这一问题. 具体而言, 通过在两个损失函数中间添加一个 BN 层, 减轻两个损失函数之间的耦合且同时保持了同一人特征的紧凑分布. 此外, BNNeck 使用 HPM 中同样结构的全连接层, 并添加标签平滑 (label smooth) 操作防止过拟合, 结果作为交叉熵损失 (cross-entropy loss) 函数的输入.

1.2 联合损失函数

三元组损失由 Google 研究团队在文献[25]中最早提出并应用于人脸识别任务中. 三元组损失的优势是数据集的细节区分, 能够在输入数据相近时对细节更好地进行建模. 输入为 (a, p, n) 三元组形式, 分别对应 3 种步态序列. a 为基准样本, p 为与 a 同类别的样本, n 为与不 a 同类别的样本. 式 (3) 展示了三元组损失的计算公式:

$$L_{tri} = \max(d(a,p) - d(a,n) + margin, 0) \tag{3}$$

其中, L_{tri} 代表三元组损失值; d 代表两个特征之间的欧氏距离; $margin$ 为边界的距离, 目的是使得 a 与 p 之间的距离更小, 与 n 之间的距离更大.

在步态识别领域中, 多数方法采用三元组损失进行训练. 虽然三元组损失增强了欧氏空间中的类内 (intra-class) 紧密度和类间 (inter-class) 可分性, 但是不能提供全局最优的约束, 导致在训练过程中模型难以收敛且容易过拟合.

交叉熵损失是在图像分类任务中普遍使用的一种损失函数, 度量预测分布 p 与真实分布 q 之间的接近程度. 交叉熵损失公式为式 (4):

$$L_{ce} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_{ij})\log(q(x_{ij})) \tag{4}$$

其中, L_{ce} 为交叉熵损失; m 为样本数; n 为类别数; $p(x_{ij})$ 为样本真实分布中第 i 个样本为类别 j 的概率; $q(x_{ij})$ 为样本预测分布中样本 i 为类别 j 的概率.

在受试人员的数量非常多, 而每个人训练样本个数较少、类内距离与类间距离差距较小的情况下, 只使用交叉熵将导致训练难以收敛. 但是三元组损失可以有效地拉开类间距离并且缩小类内距离. 因此, 在训练阶段, 结合使用两种损失函数进行训练有助于模型学习到更具辨识性的特征. 以此构造联合损失函数, 公式如式 (5):

$$L_{combine} = \alpha L_{tri} + \beta L_{ce} \tag{5}$$

其中, $L_{combine}$ 为联合损失函数; L_{tri} 为三元组损失; L_{ce} 为交叉熵损失; α 与 β 为超参数, 用来平衡模型的收敛性和模型的收敛性, 合适的超参数可以加速模型收敛并提高模型性能[15].

2 实验及结果分析

2.1 数据集介绍

本文实验共使用 3 个公开步态数据集进行训练, 其中包括 1 个室内步态数据集 CASIA-B 以及 2 个室外步态数据集 GREW 和 Gait3D.

2.1.1 CASIA-B

CASIA-B 数据集是中国科学院自动化研究所于 2005 年 1 月提供的一个大规模、多视角的步态数据

集, 是目前使用最广泛的步态数据集, 共包含 124 名受试人员. 从 0–180°视角每间隔 18°设置一个采集视角, 共采集 11 个视角的步态序列. 另外, 该数据集在每一个采集角度中又设置了 3 种行走时不同的步态条件, 分别为 6 个正常行走 (normal walking, NM)、2 个穿着外套行走 (walking with a coat, CL) 以及 2 个携包行走 (walking with a bag, BG) 的步态序列. 该数据集官方并未划分训练集以及测试集. 一般的划分方法有 3 种, 分别为小样本训练 (small-sample training, ST), 中等样本训练 (medium-sample training, MT) 以及大样本训练 (large-sample training, LT). 在本文, 训练数据划分采用的是 LT 划分. 在测试时, 将测试集前 4 个 NM 序列划分为图库集 (gallery set), 其余共 6 个序列作为探针集 (probe set), 也即 NM 序列、CL 序列以及 BG 序列各 2 个.

2.1.2 GREW

GREW 数据集是由清华大学于 2021 年公布的室外步态数据集. GREW 包括 26 345 个受试人员以及 128 671 组步态序列, 采集于室外开放环境中的 882 个摄像头. 另外, 该数据集还包括步态研究界的第 1 个干扰物集, 其中包含了 233 857 个步态序列, 使得数据集更接近真实环境. GREW 采集角度多样, 没有固定拍摄角度, 而且提供了更为多样的行走条件, 包括人物携带不同的箱包以及穿着方式. 具体而言, 箱包携带条件共有 5 种, 分别为无携带、双肩背包、单肩包、手提包以及拉杆箱等 5 种. 穿戴风格有长袖、短袖、背心、长裤、短裤以及裙子等 6 种. 该数据集将年龄分为 5 组: 成人以 14 岁为间隔分为 3 组 (即 16–30 岁、31–45 岁及 46–60 岁). 儿童 (16 岁以下) 和老人 (60 岁以上) 各分为一组, 且每个年龄组中, 性别分布均衡. 官方将 GREW 划分为训练集、验证集以及测试集: 其中训练集包含 20 000 个受试人员及 102 887 个步态序列, 测试集包含 6 000 个受试人员及 24 000 个步态序列, 验证集包含 345 个受试人员的 1 784 个步态序列. 这 3 个数据集中包含了相同的受试人员, 且皆由不同的摄影机记录所得. 另外在测试集中每个受试人员包含 4 个步态序列, 其中 2 个作为图库集, 另外两个作为探针集.

2.1.3 Gait3D

Gait3D 是杭州电子科技大学于 2022 年发布的第一个基于大规模 3D 表示的步态数据集. Gait3D 在无限制的室外环境通过 39 个摄像头采集了 4 000 个受试人员的超过 25 000 个步态序列, 可支持密集人体体型、三维视角和步态动态信息建模. 官方将前 3 000 个受试人员的步态序列划分为训练集, 将余下 1 000 个受试人员的步态序列划分为测试集. 对于测试集进一步划分为包含 1 000 个步态序列的图库集, 剩余的 5 369 个步态序列为探针集.

2.2 数据集信息汇总

实验所用数据集信息汇总于表 3.

表 3 数据集信息汇总表

数据集	训练集		测试集		条件
	Id	Seq	Id	Seq	
CASIA-B	74	8 140	50	5 500	NM, BG, CL
GREW	20 000	102 887	6 000	24 000	多样
Gait3D	3 000	18 940	1 000	6 369	多样

2.3 实验环境与参数配置

为排除不同实验设备对实验数据造成的影响, 本文所涉代码复现以及实验均在服务器 (2×NVIDIA GeForce RTX 3090, Inter(R) Xeon(R) Platinum 8350C CPU @2.60 GHz, 84 RAM) 上进行, 操作系统为 Ubuntu 16.04; 深度学习框架 PyTorch 版本为 2.0; CUDA 版本为 11.8; Python 版本 3.10.9; 所使用的测试平台为 OpenGait[17].

由于 CASIA-B 数据集中的每张图像中步态轮廓图不在统一位置, 本文使用 Chao 等人[9]所提出的数据预处理方法将步态序列中人体对齐. 将 3 个数据集中的步态轮廓图统一裁剪为 64×44 大小. 同时, 为了增强模型的泛化能力并提升模型的鲁棒性, 对数据集使用数据增强操作. 每个样本输入序列帧数固定为 30 帧.

在训练阶段, 针对不同数据集的参数设置如表 4 所示 (该部分参数设置以及下方超参数设置均沿用 GaitBase[17]中默认参数设置): 表 4 中批次大小 (batch size)(n×p) 表示每次随机选取 n 个人, 每个人随机选取 p 个步态序列; 多步学习率调整 (multistep scheduler) 是指在迭代次数达到该值时对学习率将乘 0.1; R、P、H、E 分别指: 随机旋转 (random rotate)、随机透视 (random perspective)、 随机水平翻转 (random horizontal flip) 以及随机擦除 (random erasing);

本文选择的优化器算法为 SGD (stochastic gradient descent), 学习率初始值设置为 0.1. 为了加速训练, 帮助模型在参数空间内跨越局部极小值, 设置动量 (momentum) 为 0.9. 同时, 为了减小模型复杂度并防止过拟合,

设置权重衰减为 5E–4. 在损失函数中的参数设置中, 其中三元组损失中的*margin*设置为 0.2, 损失函数中的两个超参数均设置为 1.0. 在 HPM 中设置 S=4. 在训练阶段, 每训练 20 000 次进行一次验证, 评价指标统一选择 Rank-1 准确率.

表 4 训练参数表

数据集	批次大小	多步学习率调整	数据增强	迭代
CASIA-B	(8, 16)	(20k, 40k, 50k)	R: 0.3; E: 0.3	60k
GREW	(32, 4)	(80k, 120k, 150k)	P: 0.2; H: 0.2; R: 0.2	100k
Gait3D	(32, 4)	(20k, 40k, 50k)	P: 0.2; H: 0.2; R: 0.2	60k

2.4 实验结果与分析

将 VRGait 模型的测试结果与近年来提出的 GaitSet、GaitPart 等 7 种较先进的模型在 4 个步态数据集 CASIA-B、OU-MVLP[26]、GREW 和 Gait3D 上的测试数据进行对比 (使用 Rank-1 准确率作为评价指标). 对比结果如表 5 所示 (NM、BG 以及 CL 分别指 CASIA-B 数据集中测试集包含的正常、携包以及穿着外套步态序列; OU-MVLP 为日本大阪大学公开的室内数据集; 其中带“*”数据为使用原论文的代码及参数复现的实验数据).

表 5 不同方法的准确率对比 (%)

模型	数据集					
	CASIA-B			OU-MVLP	GREW	Gait3D
	NM	BG	CL			
VRGait	97.0	91.9	76.0	—	**61.3**	**64.3**
GaitBase	97.6	94.0	77.4	90.0	54.8*	63.3*
GaitSet	95.8	90.0	75.4	87.1	48.4	36.7
GaitPart	96.1	90.7	78.7	88.7	47.6	28.2
GaitGL	97.4	94.5	83.8	89.7	47.3	29.7
CSTL	98.0	95.4	**87.0**	90.2	50.6	11.7
3Dlocal	**98.3**	**95.5**	84.5	**90.9**	—	—
SMPLGait	—	—	—	—	—	46.3

表 5 显示, VRGait 在室内步态数据集 CASIA-B 中 NM、BG 序列中的 Rank-1 准确率达到 97.0% 和 91.9%, 相较于在 CASIA-B 中识别率较高的 3Dlocal 模型性能相差 1.3% 和 3.6%, 其中在 CL 序列中的测试结果与 CSTL 模型相差 11.0%, 说明 VRGait 在室内步态数据集中的表现尚未达到最优水平, 尤其在 CL 序列的测试中与目前先进方法仍有着较大差距, 这是由于 VRGait 拥有更深的网络层数, 而较深的网络模型往往需要大量的多样化的数据进行训练才能达到较好的效果. CASIA-B 仅有 74 个受试者, 并且数据集采集于实验室, 受到许多限制, 与在真实环境中采集到的数据有着很大的差异, 例如缺少真实场景中常常出现的遮挡、不佳的光照环境及视角的任意性等, 这些因素极大地限制了训练样本的多样性, 难以充分发挥深层网络模型的优势. 此外, Fan 等人[27]也通过实验说明: 模型的网络层数越深, 越容易在简单的室内步态数据集中出现过拟合, 从而导致性能下降.

现阶段在室内数据集中的表现较好的模型多数使用浅层网络提取步态特征, 由表 5 可知其性能在室外步态数据集中的退化十分严重, 其中 CSTL 在 Gait3D 数据集的测试中性能下降超过 80%, 说明仅使用浅层网络在面临复杂数据集时无法准确提取步态特征. VRGait 使用深层视频残差网络进行特征提取, 尽管在室内步态数据集中表现一般, 但是该模型在室外步态数据集中的表现却超越了目前先进的模型.

在训练阶段, 模型在训练集上的 Softmax 损失与准确率变化如图 5 所示, 尽管训练准确率在不断提升, 但是测试准确率与其并非成正相关. 其中 VRGait 在 GREW 数据集上训练 100 000 次后测试准确率达到最高 (61.27%), 之后测试准确率不断下降, 在 180 000 次训练后准确率下降至 60.82%. 推测该模型在训练 100 000 前后出现过拟合. 最终选取第 100 000 次迭代后的网络参数作为 GREW 数据集的训练模型.

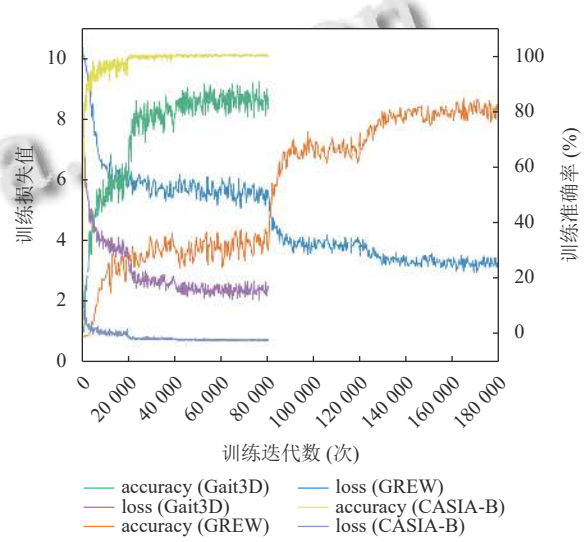

图 5 Softmax 损失及准确率随迭代次数变化

2.5 消融实验

本文提出使用视频残差神经网络进行取步态特征提取, 为了验证该网络的有效性, 本节使用 Gait3D 与 GREW 室外步态数据集进行消融实验. 实验将 VRGait

的主干网替换为 GaitBase 中的 2D 残差网络, 其余结构及参数保持于本文相同. 结果如图 6 所示. 图 6 中“VRGait_GREW”指 VRGait 模型在 GREW 测试集上的实验结果, 以此类推. VRGait 较 GaitBase 在 GREW 以及 Gait3D 中测试准确率为+6.5% 和+1.0%. 此外, VRGait 模型在室外步态数据集中具备较高的准确度的同时还拥有相对较快的收敛速度. 在 Gait3D 数据集的测试中, GaitBase 训练 20 000 次共用时 4h8'17'', VRGait 仅用时 3h15'14''. 在 GREW 数据集中 VRGait 仅迭代 20 000 次, 其结果便已经超越 GaitBase 模型 160 000 次迭代后的最高准确率 4.96 个百分点.

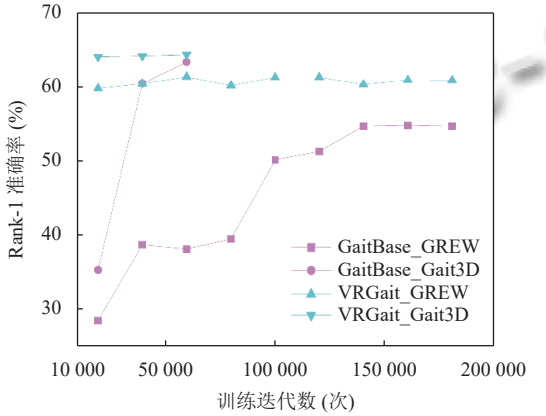

图 6 Rank-1 准确率随迭代次数变化

3 结束语

本文提出一种基于视频残差网络的深度步态识别模型, 构建了 VideoResNet 主干网络, 解除了 3D CNN 网络需要固定长度的帧作为输入的限制并提取步态序列的时空动力学特征; 采用 TP 与 HPM 降低特征采样分辨率并提取局部步态特征; 最后引入 BNNeck 通对两个损失函数进行解耦并调整特征空间. 在室外步态数据集的测试结果表明, 该方法准确率高于目前先进方法, 同时网络收敛速度也有较大的提升.

本文所提方法在室外步态数据集中的识别率相较于目前先进方法虽有显著提升, 但准确率仍有待提高. 因此, 如何提高室外步态数据集即复杂条件下步态识别的准确率仍是未来研究的重点.

参考文献

1 Deligianni F, Guo Y, Yang GZ. From emotions to mood disorders: a survey on gait analysis methodology. IEEE Journal of Biomedical and Health Informatics, 2019, 23(6): 2302–2316. [doi: 10.1109/JBHI.2019.2938111]

2 Jarchi D, Pope J, Lee TKM, *et al*. A review on accelerometry-based gait analysis and emerging clinical applications. IEEE Reviews in Biomedical Engineering, 2018, 11: 177–194. [doi: 10.1109/RBME.2018.2807182]

3 Lin BB, Zhang SL, Bao F. Gait recognition with multiple-temporal-scale 3D convolutional neural network. Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020. 3054–3062.

4 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]

5 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 (ICIP). Anchorage: IEEE, 2021. 2314–2318.

6 Li X, Makihara Y, Xu C, *et al*. End-to-end model-based gait recognition using synchronized multi-view pose constraint. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 4089–4098.

7 Zheng JK, Liu XC, Liu W, *et al*. Gait recognition in the wild with dense 3D representations and a benchmark. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 20196–20205.

8 Peng YJ, Ma K, Zhang Y, *et al*. Learning rich features for gait recognition by integrating skeletons and silhouettes. Multimedia Tools and Applications, 2023: 1–22.

9 Chao HQ, He YW, Zhang JP, *et al*. GaitSet: Regarding gait as a set for cross-view gait recognition. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019. 8126–8133.

10 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. 14213–14221.

11 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. 14628–14636.

12 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. 12889–12898.
13 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. 14900–14909.
14 徐硕, 郑锋, 唐俊, 等. 双分支特征融合网络的步态识别算法. 中国图象图形学报, 2022, 27(7): 2263–2273.
15 王晓路, 千王菲. 基于双支路卷积网络的步态识别方法. 计算机应用. http://kns.cnki.net/kcms/detail/51.1307.TP.20230911.1331.012.html. (在线出版)(2023-09-13).
16 Fu Y, Wei YC, Zhou YQ, *et al*. Horizontal pyramid matching for person re-identification. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2019. 8295–8302.
17 Fan C, Liang JH, Shen CF, *et al*. OpenGait: Revisiting gait recognition toward better practicality. arXiv:2211.06597v2, 2023.
18 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]
19 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 (ICPR 2006). Hong Kong: IEEE, 2006. 441–444.
20 Zhu Z, Guo XD, Yang T, *et al*. Gait recognition in the wild: A benchmark. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 14769–14779.
21 Tran D, Wang H, Torresani L, *et al*. A closer look at spatiotemporal convolutions for action recognition. Proceedings of the 2018 IEEE/CVF conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6450–6459.
22 Luo H, Gu YZ, Liao XY, *et al*. Bag of tricks and a strong baseline for deep person re-identification. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2019. 1487–1495.
23 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.
24 Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia: PMLR, 2010. 249–256.
25 Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 815–823.
26 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]
27 Fan C, Hou SH, Huang YZ, *et al*. Exploring deep models for practical gait recognition. arXiv:2303.03301, 2023.

(校对责编: 牛欣悦)