应用光学
Journal of Applied Optics
SCOPUS,等来源期刊
中文核心期刊、中国科技核心期刊

基于改进ViT的红外人体图像步态识别方法研究

杨彦辰 云利军 梅建华 卢琳

Gait recognition method of infrared human body images based on improved ViT

YANG Yanchen, YUN Lijun, MEI Jianhua, LU Lin

引用本文:
杨彦辰, 云利军, 梅建华, 等. 基于改进ViT的红外人体图像步态识别方法研究[J]. 应用光学, 2023, 44(1): 71-78. DOI: 10.5768/JAO202344.0102002
YANG Yanchen, YUN Lijun, MEI Jianhua, et al. Gait recognition method of infrared human body images based on improved ViT[J]. Journal of Applied Optics, 2023, 44(1): 71-78. DOI: 10.5768/JAO202344.0102002

在线阅读 View online: https://doi.org/10.5768/JAO202344.0102002

文章编号：1002-2082 (2023) 01-0071-08

基于改进 ViT 的红外人体图像步态识别方法研究

杨彦辰[1]，云利军[1,2]，梅建华[1]，卢　琳[3]

（1. 云南师范大学 信息学院，云南 昆明 650500；2. 云南省光电信息技术重点实验室，云南 昆明 650500；
3. 云南省烟草烟叶公司 设备信息科，云南 昆明 650218）

摘　要：针对卷积神经网络在步态识别时准确率易饱和现象，以及 Vision Transformer（ViT）对步态数据集拟合效率较低的问题，提出构建一个对称双重注意力机制模型，保留行走姿态的时间顺序，用若干独立特征子空间有针对性地拟合步态图像块；同时，采用对称架构的方式，增强注意力模块在拟合步态特征时的作用，并利用异类迁移学习进一步提升特征拟合效率。将该模型运用在中科院 CASIA C 红外人体步态库中进行多次仿真实验，平均识别准确率达到 96.8%。结果表明，本文模型在稳定性、数据拟合速度以及识别准确率 3 方面皆优于传统 ViT 模型和 CNN 对比模型。

关键词：步态识别；对称双重注意力机制；迁移学习；红外人体图像；Vision Transformer；卷积神经网络

中图分类号：TN219;TP181　　**文献标志码：**A　　**DOI:** 10.5768/JAO202344.0102002

Gait recognition method of infrared human body images based on improved ViT

YANG Yanchen[1], YUN Lijun[1,2], MEI Jianhua[1], LU Lin[3]

（1. College of Information, Yunnan Normal University, Kunming 650500, China; 2. Yunnan Provincial Key Laboratory of Optoelectronic Information Technology, Kunming 650500, China; 3. Department of Equipment Information, Yunnan Tobacco Leaf Company, Kunming 650218, China）

Abstract: Aiming at the phenomenon that the accuracy of convolutional neural network is easy to be saturated in gait recognition and the problem of low fitting efficiency of vision transformer (ViT) to gait data set, an idea to construct a symmetrical dual attention mechanism model was proposed to retain the time order of walking posture, and fit the gait image blocks with several independent feature subspaces. At the same time, the symmetrical architecture was adopted to enhance the role of attention module in fitting gait features, and the heterogeneous transfer learning was used to further improve the efficiency of feature fitting. The model was applied to CASIA C infrared human body gait database of Chinese Academy of Sciences for many simulation experiments, and the average recognition accuracy was 96.8%. The results show that the proposed model is superior to the traditional ViT model and CNN comparison model in stability, data fitting speed and recognition accuracy.

Key words: gait recognition; symmetrical dual attention mechanism; transfer learning; infrared human body images; vision transformer; convolutional neural network

引言

红外人体步态识别作为最有潜力的非侵入式中远距离生物特征识别技术之一，可在无需被采集者配合的情况下，利用采集到的中远距离低分辨率红外步态图像，识别行人的身份信息[1]。相较人脸、指纹等识别条件相对严格的生物特征识别技术而言，红外步态识别技术应用场景更为广泛，在可见光强度不足、雨雪天气等特殊环境下仍能

收稿日期：2022-03-21；　修回日期：2022-04-28
基金项目：云南省应用基础研究计划重点项目 (2018FA033)；云南师范大学研究生科研创新基金项目 (YJSJJ21-B77)
作者简介：杨彦辰（1997—），男，硕士研究生，主要从事视频图像处理研究。E-mail: 649228448@qq.com
通信作者：云利军（1973—），男，博士，教授，主要从事物联网技术、视频图像处理研究。E-mail: yunlijun@ynnu.edu.cn

保证较高的识别准确率，在身份识别领域异军突起[2]。

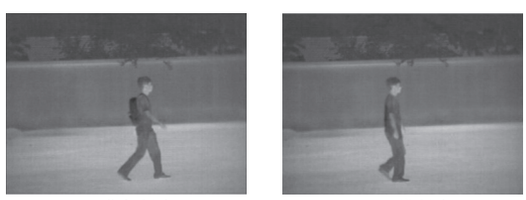
卷积神经网络(convolutional neural network, CNN)作为一种快速、可扩展的端到端学习框架，极大地简化了传统机器学习低效、冗杂的结构，在图像处理的各个领域都取得了不错的成果。He K 等人[3]提出了一种易优化的深度残差网络，通过各残差块之间的跳跃连接，防止网络过深带来的梯度消失问题，并提高了准确率。Wang H 等人[4]以残差网络为基础构建了一种 L-Resnet-50 网络，在维持较高步态识别准确率的前提下减少各部分 50% 的参数量，取得了不错的效果。Huang G 等人[5]采用调节局部特征流动的方法构建了一种处理步态数据的网络结构，该网络通过提取帧级特征和帧间局部特征间的关系，灵活地获取局部和全局中最有判别性的特征，在 CASIA B 中取得了 95.1% 的准确率。由于 CNN 无法捕捉序列化数据中的连续动态时空信息，使得卷积神经网络在自然语言处理 NLP(natural language processing)及一些数据间有顺序关系的领域表现并不是很理想。之后 Wang X 等人[6]设计了 FF-GEI(frame-by-frame GEI)，扩大了步态能量图可用数据量，结合带有长短期记忆的 Conv-LSTM(convolutional long short-term memory)模型，在 CASIA B 和 OU-ISIR 上对泛化能力进行验证，取得了较为优秀的结果。Vaswani A 等人[7]首次在 NLP 领域中提出了完全基于自注意力机制的 Transformer 架构，使模型在拥有简单结构的情况下，对带有时序信息的数据进行特征提取。Dosovitskiy A 等人[8]利用 Transformer 处理图像数据，将图像进行无重叠切片，再进行包含位置信息的数据特征学习，提供了一种全新的模型架构思想，在拥有大量样本的数据集中已经逐步赶超现在流行的一些 CNN 网络模型，但在小样本数据集上的表现仍有很大的提升空间。

本文将构建对称的双重完全注意力机制模型，以中科院 CASIA C 红外步态库作为数据集，经过数据预处理和步态周期划分之后，采用多次实验取平均的方式进行多轮消融实验。首先将本文模型和同尺寸 ViT Base 模型对比，以证明对称双重注意力结构能有效促进模型收敛。然后加入迁移学习，得出其对模型收敛速度的促进效果。最后将加入迁移学习的本文模型同 CNN 模型进行稳定性、收敛速度和准确率对比，证明融合了迁移学习之后的本文模型在保留背包、步速等杂项步态特征的状态下，仍能取得较优的识别准确率。

1 数据处理

1.1 数据集

实验数据集采用中国科学院自动化研究所 CASIA 步态数据库中的 Dataset C 红外步态数据库。该数据库在单人单一角度下对 153 名被采集者正常行走(fn)、快速行走(fq)、慢速行走(fs)、背包行走(fb)的 4 种不同行走状态进行拍摄。固定角度设置为 90°, 大小约有 66.5 MB。图 1 给出了 CASIA C 数据库中的两种红外步态实例。



(a) 背包行走 (b) 正常行走

图 1 CASIA C 数据库中的红外步态实例

Fig. 1 Examples of infrared gait in CASIA C database

1.2 红外步态图像预处理

本文首先采用背景减除法[9-10]来提取行走过程中的人体轮廓特征，再将图像进行二值化处理，进一步强化人体姿态信息，最后剪裁大量无用背景信息，并将被采集者的步态信息居中显示，最后将其调整为 128×128 像素。具体处理结果如图 2 所示。

(a) 二值化后的红外步态视频帧 (b) 剪裁无用背景信息后的图像

图 2 红外步态图像预处理

Fig. 2 Image preprocessing results of infrared gait

1.3 步态周期估计函数

由于 ViT 是对一组带有时间信息的图像数据进行特征学习的模型结构，因此，需要将行人的步态数据按照步态周期进行划分。将步态周期组作为数据输入，可使模型在特征学习过程中不止学

习到人体瞬时姿态特征, 同时又将一段时间内的姿态特征按照时间顺序联系起来, 有助于增加模型的鲁棒性和稳定性。图 3 为背包状态下以左脚迈出为初始状态的双脚步态周期图。

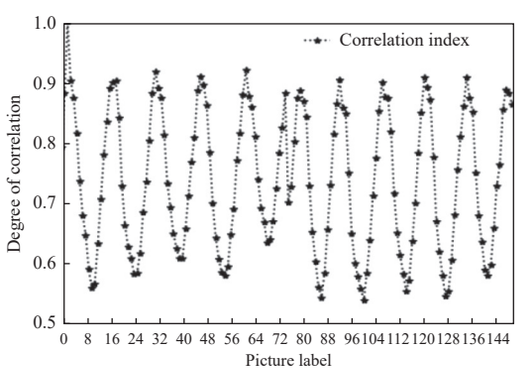
图 3　背包状态双脚步态周期

Fig. 3　Feet gait cycle in backpack state

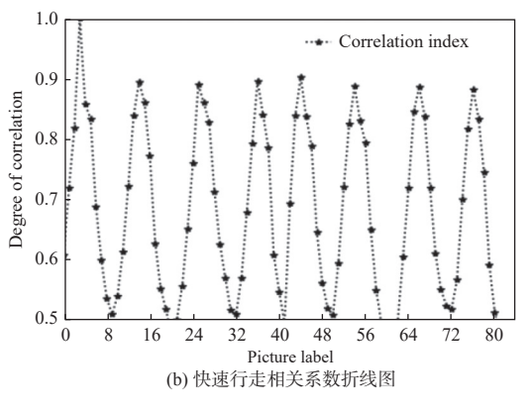
目前常用的图像相似性模板匹配算法有绝对差和(sum of absolute differences, SAD)[11]、归一化交叉相关系数(normalized cross correlation, NCC)[12]与零均值归一化交叉相关系数(zero-normalized cross correlation, ZNCC)[13]3 种, 考虑到红外步态图中对人体行走姿态特征敏感度要求较高, 同时为避免计算绝对差和或误差平方和可能出现的模式匹配错误, 本文采用对人体姿态轮廓识别更精细的 ZNCC 函数作为步态周期的估计函数, 其结果越大表明两张图像的相关性越强。ZNCC 函数可以用(1)式来表示:

$$\mathrm{ZNCC}(x,y)=\frac{1}{n}\sum_{x,y}\frac{1}{\sigma f\sigma t}(f(x,y)-\mu_f)(t(x,y)-\mu_t) \tag{1}$$

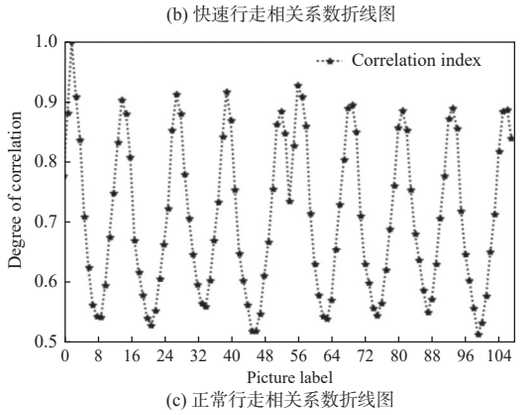
式中: (x,y)为图像中的像素位置坐标; $f(x,y)$是原图像像素值; $t(x,y)$为模板图像像素值; n为模板中像素(元素)的个数; μ_f、μ_t分别为原图像和模板图像的像素均值。将包含有时间顺序的一连串步态图像逐一输入 ZNCC 函数中, 与设定好的初始状态图像进行相关系数计算, 根据相关系数变化图对比得出特征重复周期, 再取最大值, 从而估算得到本文研究的步态周期。以 001 号类别为例, 如图 4 所示, 其中 2 个相邻峰值之间为 1 个单脚步态周期。



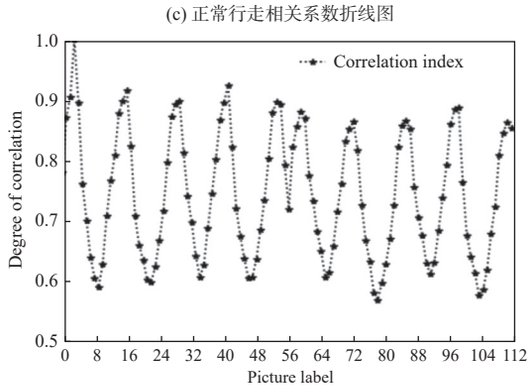
(a) 慢速行走相关系数折线图



(b) 快速行走相关系数折线图



(c) 正常行走相关系数折线图



(d) 背包行走相关系数折线图

图 4　4 种不同状态下的相关系数周期图

Fig. 4　Periodogram of correlation coefficient in four different states

2　网络模型

2.1　卷积神经网络对比模型

卷积神经网络 (CNN)[14] 是一种典型的前馈网

络结构，主要由输入层、隐藏层和输出层 3 个部分组成。输入层将图像输入 CNN 中；隐藏层通过对输入的图像卷积、池化等操作进行特征学习，其中利用池化层来压缩数据和参数量，去掉特征图中不重要的信息，突出重要特征；卷积层则是利用卷积核对感受野内的局部特征数据进行计算，其参数权重是共享的，这也使得 CNN 具有图像上的空间局部相关性；输出层根据最终的图像特征来给出图像分类结果。传统 CNN 通过增加隐藏层规模来提升识别准确率，但随着卷积的逐渐深入，丢失的输入图像细节和位置特征也越多，导致模型不易训练，准确度出现饱和甚至下降[15]。为解决这种网络退化问题，Resnet 网络[3,16] 中构建了一种残差结构。通过不同残差块之间的跳跃连接，实现了一种短路机制，使得网络可以在一定条件下通过恒等映射规则跳过一些残差块，以此来适当地调节网络深度，一定程度上解决了准确度饱和以及不易训练的问题。

考虑到模型对小样本数据集的拟合特点，本文基于 Resnet 网络构建了一种浅层双路残差网络。如图 5 所示，将 Conv1 Block 和 Block1 这 2 个浅层块组并联，通过 AVG 模块，将双路特征信息进行直接融合后取平均，然后依次输入 Block2、Block3 和 Block4 中进一步拟合特征，之后输入 AdaptiveAvgPool2D（二元自适应平均池化层）调整数据格式，最后经由 FC 层输出结果。Block 块中利用恒等映射方法来调节隐藏层深度，每个 Block 都可由 shortcut 块跳过，以保证在接收到屏蔽该残差块信号之后仍可以将图像按照相应尺寸输出。采用 RELU 作为本网络的激活函数，Block 块结构均相同但参数不同，其中 In_ch 为输入通道数，O_ch 为输出通道数，B_num 为该 Block 循环次数，Stride 为步长。

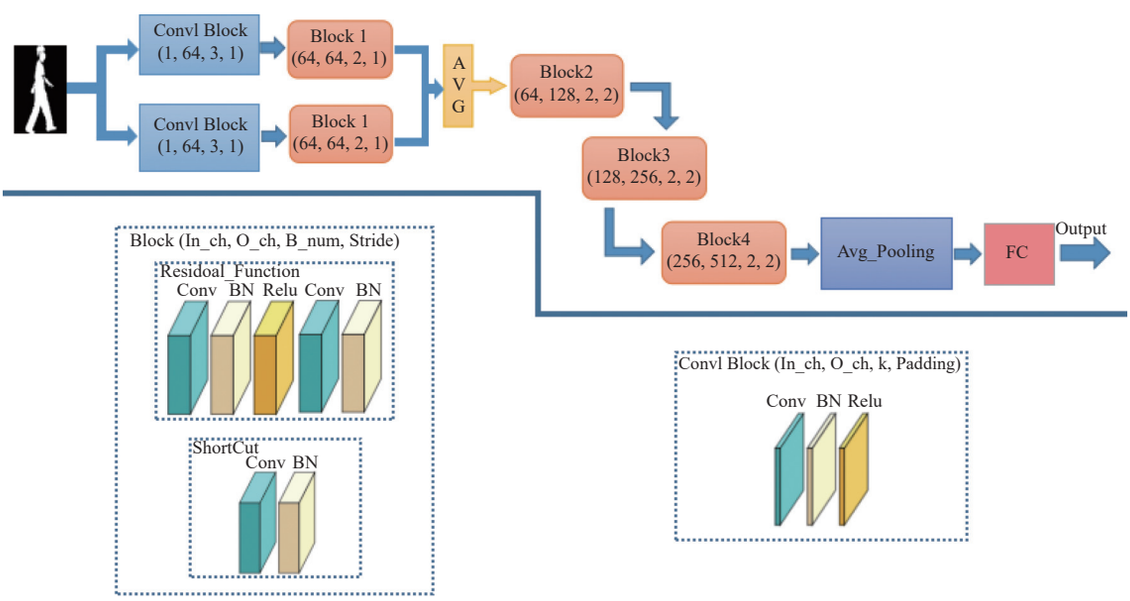


图 5 双路 CNN 步态识别模型

Fig. 5 Gait recognition model of double channel CNN

2.2 对称双重注意力机制模型

注意力机制（Attention）[17] 是一种仿照人类视觉关注重点，捕获输入信息重点特征的结构。如今常将其与 CNN 结合作为卷积的补充，通过构建“查询向量 (Query)”、“值向量 (Value)”和“键向量 (Key)”来进行缩放点积注意力操作，使得网络对不同特征分配不同的注意力，可以用（2）式、（3）式、（4）式表示：

$$f(\boldsymbol{Q},\boldsymbol{K})=\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}} \tag{2}$$

$$\boldsymbol{X}=\mathrm{softmax}(f(\boldsymbol{Q},\boldsymbol{K})) \tag{3}$$

$$\mathrm{Attention}(\boldsymbol{X},\boldsymbol{V})=\boldsymbol{X}\times\boldsymbol{V} \tag{4}$$

式中：矩阵$\boldsymbol{Q}$、$\boldsymbol{V}$和$\boldsymbol{K}$分别为查询向量矩阵、值向量矩阵和键向量矩阵；d_k指的是k的维度，此处以$\sqrt{d_k}$为缩放因子，然后使用 softmax 函数来获取注意力权重，最后将得到的权重与对应矩阵$\boldsymbol{V}$相乘，得到最终的注意力权重，如（3）式和（4）式所示。

Transformer[7] 是一种完全基于注意力机制的特征提取网络，其通过构建一种多头注意力机制（multi-head attention），在 Attention 模块的基础上进

一步完善了自注意力层, 增强了模型专注于不同位置的能力, 并为不同的注意力层加入了不同的“独立子空间”。经过多头注意力机制后, 每个头都会有独立的权重矩阵, 使得网络对每个不同位置的特征都有不同的权重参数。

$$\text{head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}}, \boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}}, \boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}) \tag{5}$$

$$\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \cdots, \text{head}_n)\boldsymbol{W}^{\boldsymbol{O}} \tag{6}$$

式中的$\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$、$\boldsymbol{W}_i^{\boldsymbol{V}}$分别为$\text{head}_i$之前时刻$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$的叠加权重矩阵, 即其他时刻的输入同样对当前结果产生影响, 从而构成了数据之间的时序相关性。(6)式是将每个注意力通道的值连接起来并与$\boldsymbol{W}^{\boldsymbol{O}}$点积, 线性转换得到 MultiHead 值。

ViT 不同于传统卷积对整张图像进行操作, 其将一张图像以固定尺寸无交叉分割成为一组图像块, 并将每个小块转化为一维张量。由于这种分割的图像块仍未标识每个小块的位置关系, 因此采用 Positional Encoding 方式为其添加位置嵌入, 如(7)式所示:

$$\text{PE}_{(\text{pos},2i)} = \sin(pos/10\,000^{2i/d_{\text{model}}}) \tag{7}$$

式中: pos 表示 token 在全局中的位置序列号; i 取$[0, \cdots, d_{\text{model}}/2]$; d_{model}取 512。这种位置嵌入方式可以适应不同尺寸分割块。

本文通过多次模型消融实验, 发现导致 ViT 模型对小样本步态数据集拟合速度较低、效果较差的原因, 是由于模型在特征拟合过程中对注意力机制模块不够重视导致的。另外, 这种分割不同位置的特征提取方式, 虽加入了位置嵌入, 但对人体步态整体特征学习能力仍较弱。因此, 本文构建了一种对称双重注意力机制步态模型, 通过设计对称双重注意力机制块, 抵消模型在拟合步态特征时速率较低的缺陷; 另外, 为了避免该模块影响过大导致准确率震荡反而不易收敛的问题, 本文设计特征融合模块(feature fusion, FF), 其先将 2 个通道获得的特征信息进行等尺寸融合, 然后再通过设置影响因子来对影响效果进行控制。

虽然这种基于完全注意力的模型可以在保留特征位置信息的情况下对一组顺序数据进行学习, 但在较小样本的数据集上会导致收敛较慢, 对模型正则化或数据增强(AugReg)的依赖性增加[18]。而迁移学习[19]可以很好地解决这种问题, 在大型数据集上进行预训练后, 将训练权重参数迁移到小样本数据集上, 使各层参数无需从初始值开始收敛, 加快数据拟合效率。

具体模型结构如图 6 所示。首先将步态周期组中每一时刻的图像按序输入 Embedding 层, 再将尺寸为 128×128 像素的步态图像分割成 16 个尺寸为 32×32 像素的图像块, 然后将图像块分别重构成一维张量, 利用(7)式中的正弦位置嵌入方法计算并添加位置嵌入后, 输入 Encoder Block 中, 经过并联的多头注意力模块, 对输入的不同人体位置姿

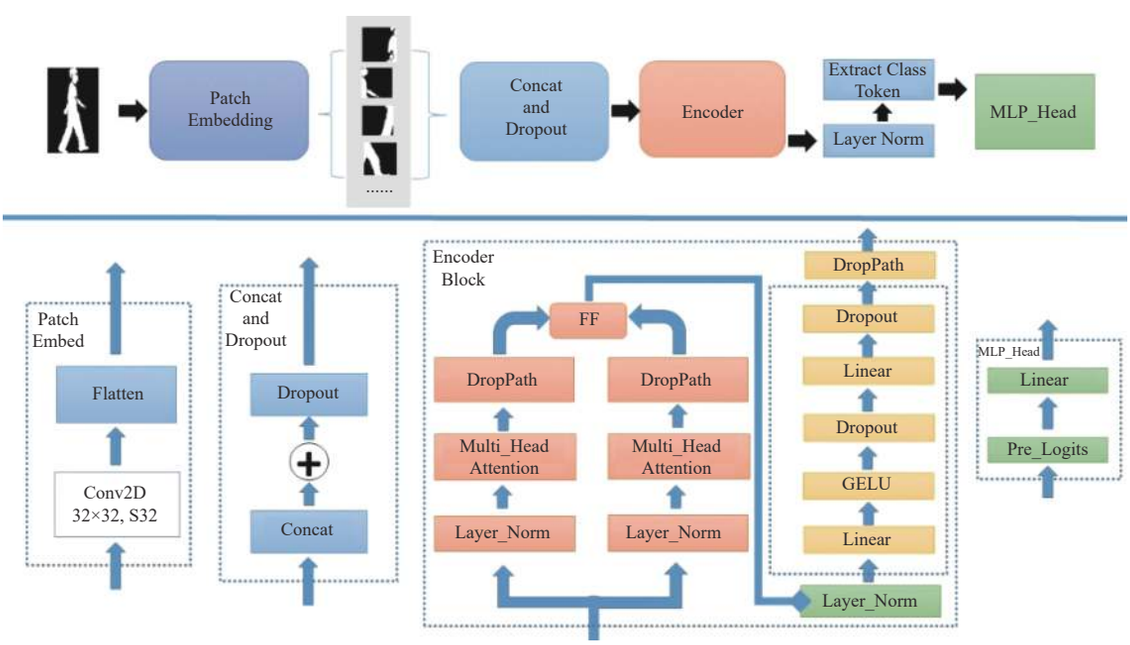


图 6　双重对称注意力机制步态模型

Fig. 6　Dual symmetrical attention mechanism gait model

态图像(如:手臂、腿等)进行姿态细节特征提取后,通过特征平均融合模块拟合特征权重,再经过LayerNormalization层后进入MLP模块中,最后传入MLP_Head块得到分类结果。为防止模型过分拟合小样本数据,在对称的多头注意力模块和MLP模块中采用DropPath来代替传统的Dropout,随机将网络中的多分支结构随机删除。

3 仿真实验与结果分析

3.1 实验设计

本文在Pytorch 1.7、Python 3.8环境下进行模型搭建。采用CASIA C红外步态数据库作为数据集,包含153名被采样者的4种行走状态,共100346张红外图像。运用2.2节中的预处理方法剔除无用特征,突出人体姿态细节,然后再按估算的单脚步态周期划分为步态周期组,并按照7∶3的比例将其分割为训练集和测试集。

按照图5构建本文的CNN对比模型,将数据无序输入模型中,设置初始学习率为1×10^{-2}。使用Adam优化器加强有效收敛,并采用categorical_crossentropy多分类交叉熵损失函数计算Loss值。设置Batchsize为14,训练迭代次数为16次。按照图6构建本文模型,设置Embedding尺寸为32,影响因子为1/2。采用相同的ViT Base模型作为对比,由于ViT是处理带有时序性数据的模型结构,因此,要保持人体行走姿态的顺序性,故将划分好的步态周期组作为模型的数据输入,再将图像按照相应的Embedding尺寸切割成多个部分,按照从左到右、从上到下的顺序输入Encoder中。采用消融实验思想,设置初始学习率为1×10^{-3},Multi_Head Attention数量为12个,以Adam作为优化器,使用categorical_crossentropy多分类交叉熵损失函数计算Loss。

3.2 结果分析

模型效果通过模型分类准确率曲线进行比较说明。首先将本文模型与相同尺寸的ViT Base模型在10个Epoch内进行效果对比。经过4次试验,将得到的结果取平均值如图7所示。由图7中圆点线和三角点线可以明显看出,Embedding尺寸为32的本文模型始终快于同尺寸的ViT对比模型,且在5个Epoch前,二者都有较快的收敛速度,之后模型对人体姿态细节的掌握程度随训练迭代增多逐渐加深,学习速度放缓,收敛加速度减少。最终,传统ViT Base32模型和本文模型在第10个Epoch时多次测试的平均识别准确率分别为60.3%和75.3%。经分析得出,该结果是因为传统ViT Base模型对注意力权重不够重视导致的,适当地加强注意力机制的影响程度,可以有助于提高准确率饱和上限,使之不易出现准确率震荡。可见,本文构建的对称双重注意力机制模型可以在一定程度上加快数据收敛速度,能够更好地拟合小样本数据特征。

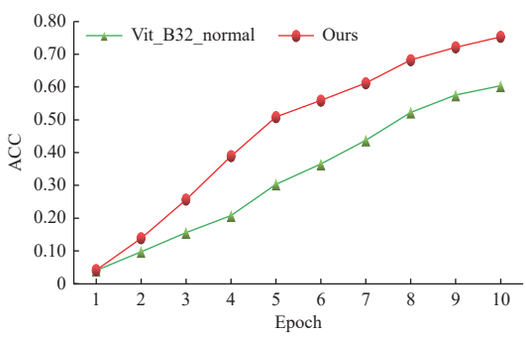


图7 本文模型与同尺寸Vit对比

Fig. 7 Comparison between proposed model and ViT of same size

为进一步解决ViT在小样本数据集上的应用存在的收敛速度过慢、不易训练等诸多问题,对本文模型采用异类迁移学习方法,先将ViT模型在ImageNet21K数据集上进行预训练,并将训练的各层参数权重进行剪裁后应用于本文构建的模型中,经过3次试验,将结果取平均值后如图8所示。由图8中圆点线和菱点线对比可见,利用迁移学习后的本文网络在训练初期便得到了较高的准确率(ACC),这是因为异类迁移为模型各层设置了初始权重,极大缩短了模型训练的时间。同时,为

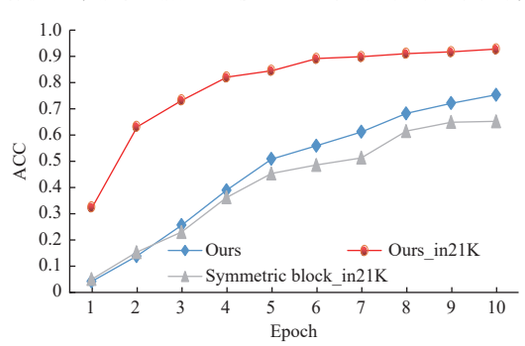


图8 加入迁移学习的本文模型与研究过程中其他尝试的对比

Fig. 8 Comparison between proposed model with transfer learning and other attempts in research process

证明并不是单纯将某一部分进行并联堆叠都有效, 本文将图 6 中的整个 Encoder Block 并联, 同样利用迁移学习进行 3 次消融实验, 结果取平均值, 如图 8 中三角点线, 发现其效果甚至不如未采用迁移学习的本文模型。原因是注意力强化模块学到的特征在经过 Layer_Norm 层、MLP 和 DropPath 层处理后已丢失了部分关键特征信息, 此时对这些残缺信息进行加强, 反而会起到消极作用。

最后, 将加入了迁移学习的本文模型与本文构建的 CNN 模型进行对比。进行 2 次试验, 将结果取平均值后如图 9 所示, 本文模型的准确率在第 7 个 Epoch 就率先达到 90%, 远超对比 CNN 模型 5 个百分点。在 11 个 Epoch 时, CNN 模型出现了准确率饱和的情况, 而本文模型的准确率却始终呈现稳定提升, 并在 16 个 Epoch 时达到 96.8%。分析表明, 加入迁移学习的本文模型不但有效缩短了模型各层权重的拟合时间, 也进一步提高了准确率上限, 使该模型在稳定性、数据拟合速度以及识别准确率 3 方面皆优于 CNN 对比模型。

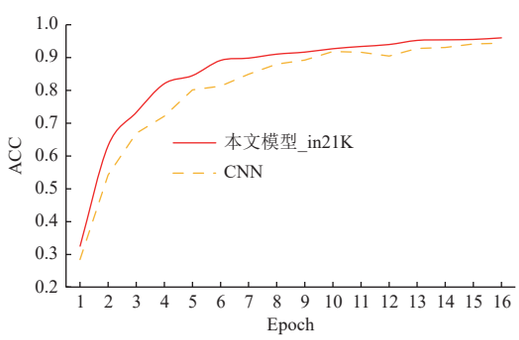


图 9　加入迁移学习的本文模型同 CNN 模型对比

Fig. 9　Comparison between proposed model with transfer learning and CNN model

4　结论

本文构建了对称双重注意力机制模型, 并将其应用于红外步态识别领域。在中科院自动化所提供的 CASIA C 红外数据库中进行 3 组对比模拟仿真实验。保留红外数据库中行人装饰(背包)、行走速度(正常、快、慢)等行走特征, 将数据集按照 ZNCC 函数估计的步态周期, 划分成多个顺序性的组, 然后将人体不同位置分割开来, 使得本文网络不是如传统卷积神经网络那样对整张图像进行学习, 而是利用独立特征子空间, 拟合行走过程中人体不同位置的姿态特征, 使得模型学习更加有针对性; 同时, 为了提高模型的学习效率, 使得模型在小样本数据集上也有较好的效果, 本文还采用了异类迁移学习的思想。经实验证明, 加入迁移学习后的模型在数据拟合速度、稳定性、平均识别准确率等方面可以明显超越 CNN 以及传统 ViT 模型, 进一步使其在本领域应用成为可能, 亦为 ViT 在小样本数据集上的应用提供了新的思路。

参考文献:

[1] MORI A, MAKIHARA Y, YAGI Y. Gait recognition using period-based phase synchronization for low frame-rate videos[C]//2010 20th International Conference on Pattern Recognition. USA: IEEE, 2010: 2194-2197.

[2] KONG K, TOMIZUKA M. A gait monitoring system based on air pressure sensors embedded in a shoe[J]. IEEE/ASME Transactions on mechatronics, 2009, 14(3): 358-370.

[3] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. USA: IEEE, 2016: 770-778.

[4] WANG H, YUAN H. Gait Recognition based on lightweight CNNs[C]//Proceedings of the 2021 international workshop on modern science and technology. [S. l.]: The International Center of National University Corporation Kitami Institute of Technology, 2021: 185-190.

[5] HUANG G, LU Z, PUN C M, et al. Flexible gait recognition based on flow regulation of local features between key frames[J]. IEEE Access, 2020, 8: 75381-75392.

[6] WANG X, YAN W Q. Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory[J]. International Journal of Neural Systems, 2020, 30(1): 1950027.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. [S. l.] : arXiv, 2017: 5998-6008.

[8] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2021-06-03)[2022-03-21]. https://arxiv.org/abs/2010.11929.

[9] 徐剑, 丁晓青, 王生进, 等. 一种融合局部纹理和颜色信息的背景减除方法[J]. 自动化学报, 2009, 35(9): 1145-1150.
XU Jian, DING Xiaoqing, WANG Shengjin, et al. A

background subtraction method integrating local texture and color information[J]. Journal of Automation, 2009, 35(9): 1145-1150.

[10] 刘仲民, 何胜皎, 胡文瑾, 等. 基于背景减除法的视频序列运动目标检测[J]. 计算机应用, 2017, 37(6): 1777-1781.
LIU Zhongmin, HE Shengjiao, HU Wenjin, et al. Moving target detection in video sequences based on background subtraction method[J]. Computer Application, 2017, 37(6): 1777-1781.

[11] 张丽红, 何树成. 基于差值绝对值之和和置信传播的快速收敛立体匹配算法[J]. 计算机应用, 2014, 34(3): 824-827.
ZHANG Lihong, HE Shucheng. Fast convergent stereo matching algorithm based on sum of absolute values of difference and confidence propagation[J]. Computer Applications, 2014, 34(3): 824-827.

[12] 陈丽芳, 刘渊, 须文波. 改进的归一互相关法的灰度图像模板匹配方法[J]. 计算机工程与应用, 2011, 47(26): 181-183.
CHEN Lifang, LIU Yuan, XU Wenbo. Improved normalized cross correlation method for grayscale image template matching method[J]. Computer Engineering and Application, 2011, 47(26): 181-183.

[13] DI STEFANO L, MATTOCCIA S, TOMBARI F. ZNCC-based template matching using bounded partial correlation[J]. Pattern Recognition Letters, 2005, 26(14): 2129-2134.

[14] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[15] RONALD M, POULOSE A, HAN D S. iSPLInception: an inception-resnet deep learning architecture for human activity recognition[J]. IEEE Access, 2021, 9: 68985-69001.

[16] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks[C]//European conference on computer vision. Switzerland: Springer, Cham, 2016: 630-645.

[17] QIUXIA L A I, KHAN S, NIE Y, et al. Understanding more about human and machine attention in deep neural networks[J]. IEEE Transactions on Multimedia, 2020, 23: 2086-2099.

[18] STEINER A, KOLESNIKOV A, ZHAI X, et al. How to train your ViT? data, augmentation, and regularization in vision transformers[EB/OL]. (2021-06-18)[2022-03-21]. https://arxiv.org/abs/2106.10270v1.

[19] ZHAI X, KOLESNIKOV A, HOULSBY N, et al. Scaling vision transformers[EB/OL]. (2021-06-08)[2022-03-21].https://arxiv.org/abs/2106.04560.