# 基于频域注意力时空卷积网络的步态识别方法

赵国顺[1,2],方建安[1,2],瞿斌杰[1,2],Samah A.F.Manssor[1,2],孙韶媛[1,2]

(1.东华大学 信息科学与技术学院,上海 201620;
2.数字化纺织服装技术教育部工程研究中心,上海 201620)

**摘 要**:为了解决步态信息冗余多、特征重要性分布不均匀以及步态的时空特征难以学习的问题,提出了基于频域注意力的时空卷积网络进行步态识别。该方法改进了三维卷积网络(C3D)学习时空特征,同时提出了一种频域注意力卷积操作,既减少了冗余计算,注意力的调整又提高了学习效果。网络首先将步态信息划分为五组,然后通过改进的卷积进行时空特征抽取,最后通过 Softmax 层进行分类。在中科大数据集 CASIA dataset B 中进行测试,在 Bag 状态与 Coat 状态下准确率分别为 88.5%、92.8%,分别较传统深度卷积网络(Deep CNN)提升 3%左右,同时注意力在网络学习中重新分布,各个角度下的准确率也平均提升 2%左右。

**关键词**:频域;注意力;三维卷积;步态识别;生物特征;深度学习

**中图分类号**:U491.1 **文献标识码**:A **DOI**:10.19358/j.issn.2096-5133.2020.06.003

**引用格式**:赵国顺,方建安,瞿斌杰,等.基于频域注意力时空卷积网络的步态识别方法[J].信息技术与网络安全,2020,39(6):13-18.

## Gait recognition method based on frequency domain attention spatio-temporal convolutional network

Zhao Guoshun[1,2], Fang Jianan[1,2], Qu Binjie[1,2], Samah A.F.Manssor[1,2], Sun Shaoyuan[1,2]

(1.School of Information Science and Technology, Donghua University, Shanghai 201620, China;
2.Engineering Research Center of Digitized Textile & Apparel Technology, Ministry of Education, Shanghai 201620, China)

**Abstract**: In order to solve the problems of redundant gait information, uneven distribution of feature importance, and difficulty in learning the spatiotemporal features of the gait, a spatiotemporal convolutional network based on attention in the frequency domain was proposed for gait recognition. In the experiment, the spatial and temporal characteristics of three-dimensional convolutional network(C3D) learning were improved. At the same time, a frequency-domain attention convolution operation was proposed, which not only reduced redundant calculations, but also adjusted the attention and improved the learning effect. The network firstly divides the gait information into five groups, then extracts the spatiotemporal features through improved convolution, and finally classifies them through the Softmax layer. Tested in the CASIA dataset B of the Chinese University of Science and Technology, the accuracy rates in the Bag state and Coat state are 88.5% and 92.8% respectively, which are about 3% higher than traditional deep convolutional networks(Deep CNN). At the same time, attention is redistributed in network learning, the accuracy rate of each angle is increased by about 2% on average.

**Key words**: frequency domain; attention; 3D convolution; gait recognition; biometrics; deep learning

## 0 引言

步态特征,通俗来说就是人行走时的姿态外观,具体包括手臂、大腿、小腿等身体轮廓的变化,由于步态的采集不需要与被识别者有物理上的接触,也不需要近距离的接触,因此应用场景比较完善。医学研究表明,每一个人的步态都有自己的形态,具有唯一性[1],使用步态识别具有一定的安全性,不会导致信息的错误。将步态识别技术应用于当今智能监控领域,可以在多场景下对人员进行监控,防止意外情况发生,也有利于锁定犯罪嫌疑人,节省人

www.bcachina.com

力物力。

目前,关于步态识别的方法主要有两种。一种是基于步态模板的方法,主要是通过构建步态特征,比如关节点的位置变化、重心的起伏周期等几何数字特征,将一个人的行走视频序列压缩成一个模板,然后通过匹配待预测行人的步态与模板的相似度进行识别[2-4]。另一种方法是通过深度学习直接抽取原始图像序列的步态信息,通过深度神经网络学习高维时空信息来匹配行人的步态,这种方法不需要大量精细的特征构建,是一种端到端的识别方法[5-7]。

虽然基于步态模板的方法取得了一定的准确率,但是这种特征构造方法复杂,而且受角度、环境、穿着变化影响较大,同时这种特征缺失了时空信息的抽取,在精度上具有一定的限制性。深度学习方法是一种端到端的学习方法,鲁棒性强,易于操作,但是由于模型参数巨大,如何保证准确性与实时性成了关键。

本文基于深度学习的方法,改良了三维卷积网络(C3D)的网络结构,提出频域注意力卷积操作,主要通过划分频域空间,引进频域卷积。同时另一个创新主要是注意力机制的引入,这使得网络更加关注不同步态之间的不同,调整步态分布的重要性,提升网络学习效果。经由中科大数据集 CASIA dataset B 检测,本文方法在跨视角实验和方法对比实验中具有提升。

## 1 结合注意力机制的频域卷积

针对图像的高低频域所对应信息特点,引入频域卷积的思想,同时根据注意力机制的思想,针对不同信息给予不同注意力。通过频域卷积思想,可以进一步减少冗余信息的干扰,同时注意力的加入有利于增强模型的学习能力,达到精简结构、提升网络能力的效果。

### 1.1 频域卷积

卷积神经网络(CNN)生成的特征图的空间维度上存在大量冗余,其中每个位置独立存储自己的特征描述符,而忽略了可以一起存储和处理的相邻位置之间的公共信息。自然图像可以分解为描述平滑变化的结构的低频分量和描述快速变化的精细细节的高频分量[8-9]。同样,如果卷积层的输出特征图也可以分解为不同空间频率的特征,高频和低频特征如何映射到不同频率特征的组这一信息将会被 CNN 学习到,通过在相邻位置之间共享信息以减少空间冗余,可以安全地降低低频组的空间分辨率。

为此,本文引入频域卷积(OctConv),流程如图 1 所示。频域卷积将输入和输出分为高频、低频两个部分,在进行卷积操作的时候分别进行交互,输出的特征图同样分为高频与低频两个部分,尺寸与输入相同。具体公式如下:

$$\mathrm{HIGH}=f(X^{\mathrm{H}};W^{\mathrm{H}\to\mathrm{H}})+f(X^{\mathrm{L}};W^{\mathrm{L}\to\mathrm{H}}) \tag{1}$$

$$\mathrm{LOW}=f(X^{\mathrm{H}};W^{\mathrm{H}\to\mathrm{L}})+f(X^{\mathrm{L}};W^{\mathrm{L}\to\mathrm{L}}) \tag{2}$$

$$\mathrm{OUT}=\mathrm{concat}(\mathrm{HIGH},\mathrm{LOW}) \tag{3}$$

### 1.2 注意力机制

注意力已经被广泛用作序列建模的计算模块,因为它具有捕获长距离交互的能力,针对视觉任务已经提出了多种注意力机制,以解决卷积的弱点。例如,Gather-Excite[10]和 Squeeze-and-Excitation[11]使用从整个特征图抽取信号重新赋值特征通道,而 BAM[12]和 CBAM[13]在通道和空间维度上独立地定义卷积特征。在非局部神经网络中[14],通过在卷积架

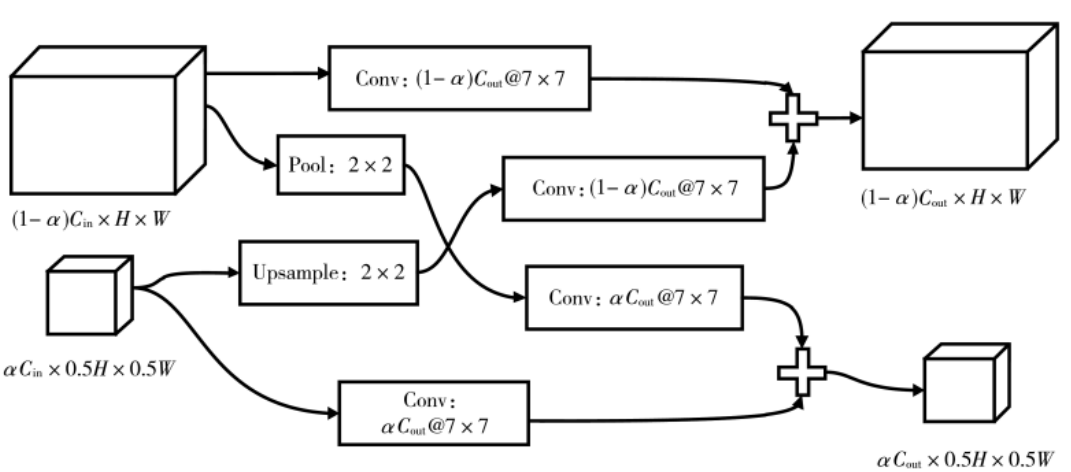


图 1 频域卷积示意图

www.pcachina.com

构中使用一些采用自注意力的非局部残差块,使视频分类和目标检测得到了改进。但是,非局部模块仅在 ImageNet 预训练之后才添加到网络结构中,并以不破坏预训练的方式进行初始化。

给定尺寸为$(H, W, F_{in})$的输入张量,将其展平为矩阵$\boldsymbol{X} \in \mathbf{R}^{HW \times F_{in}}$,按照 Transformer 体系结构中的多头注意力结构[15],单头 h 的自我注意机制的输出可以表示为:

$$\boldsymbol{O}_h = \text{softmax}\left(\frac{(\boldsymbol{X}\times\boldsymbol{W}_q)(\boldsymbol{X}\times\boldsymbol{W}_k)^T}{\sqrt{d_k^h}}\right)(\boldsymbol{X}\times\boldsymbol{W}_v) \tag{4}$$

其中$\boldsymbol{W}_q$、$\boldsymbol{W}_k \in \mathbf{R}^{F_{in}\times d_k^h}$和$\boldsymbol{W}_v \in \mathbf{R}^{F_{in}\times d_v^h}$是学习到的线性变换,将输入$\boldsymbol{X}$映射到查询$\boldsymbol{Q}$,键$\boldsymbol{K}$和值$\boldsymbol{V}$。

然后将所有头部的输出连接并再次拼接,如下所示:

$$\text{MHA}(\boldsymbol{X}) = \text{concat}(\boldsymbol{O}_1, \boldsymbol{O}_2, \cdots, \boldsymbol{Q}_{N_h}) \tag{5}$$

其中$\boldsymbol{W}_o \in \mathbf{R}^{d_v\times d_v}$是学习的线性变换。然后将 MHA($\boldsymbol{X}$)重塑为$(H, W, d_v)$形状的张量以匹配原始空间尺寸。多头注意力会导致$O((HW)^2 d_k)$的复杂性和$O((HW)N_h^2)$的存储成本提高,因为它需要存储每个头的注意力图。

### 1.3 频域注意力融合卷积模块

注意力是一种对于图像局部信息重要性的一种度量,在图像特征中对于高频和低频应该有着不同的注意力,同时对于不同时间的步态应该也有着不同的注意力。在引入注意力机制后,有利于学习步态的时间和空间信息,而且有利于进一步减少冗余信息的学习。因此,本文在频率卷积的基础上结合注意力机制提出频率注意力卷积模块。

改进的频率注意力模块结构如图 2 所示,输入 HIGH:$H\times W\times\alpha C$,LOW:$0.5H\times 0.5W\times(1-\alpha)C$,由 Oct-Conv 卷积生成$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$频率特征图,其尺寸分别为:

$$\boldsymbol{Q}\text{-HIGH}: H\times W\times\alpha d_q^h \tag{6}$$

$$\boldsymbol{Q}\text{-LOW}: 0.5H\times 0.5W\times(1-\alpha)d_q^h \tag{7}$$

$$\boldsymbol{K}\text{-HIGH}: H\times W\times\alpha d_k^h \tag{8}$$

$$\boldsymbol{K}\text{-LOW}: 0.5H\times 0.5W\times(1-\alpha)d_k^h \tag{9}$$

$$\boldsymbol{V}\text{-HIGH}: H\times W\times\alpha d_v^h \tag{10}$$

$$\boldsymbol{V}\text{-LOW}: 0.5H\times 0.5W\times(1-\alpha)d_v^h \tag{11}$$

其中,$\alpha$为高频通道所占比例;$d_k^h$、$d_q^h$、$d_v^h$分别为各个头部通道数,且$d_k^h = d_q^h$。$\boldsymbol{Q}$-HIGH 表示注意力$\boldsymbol{Q}$的高频输出分量,$\boldsymbol{K}$-HIGH 表示注意力$\boldsymbol{K}$的高频输出分量,$\boldsymbol{V}$-HIGH 表示注意力$\boldsymbol{V}$的高频输出分量。$H$表示输入特征图高度,$W$表示输入特征图宽度,$\boldsymbol{Q}$-LOW 表示注意力$\boldsymbol{Q}$的低频输出分量,$\boldsymbol{K}$-LOW 表示注意力$\boldsymbol{K}$的低频输出分量,$\boldsymbol{V}$-LOW 表示注意力$\boldsymbol{V}$的低频输出分量。同时将$\boldsymbol{Q}$、$\boldsymbol{K}$频率特征图进行对应相乘,然后将高频、低频分别除以各自的通道数量$H\times W\times(\alpha\times d_v^h)$并进行计算注意力 Softmax 的得分,最后将注意力得分特征图与$\boldsymbol{V}$的频率特征图进行对应相乘,得到经过注意力机制学习的并具有频率学习信息的头部,其尺寸为:

$$\text{head-h}: H\times W\times(\alpha\times d_v^h)$$

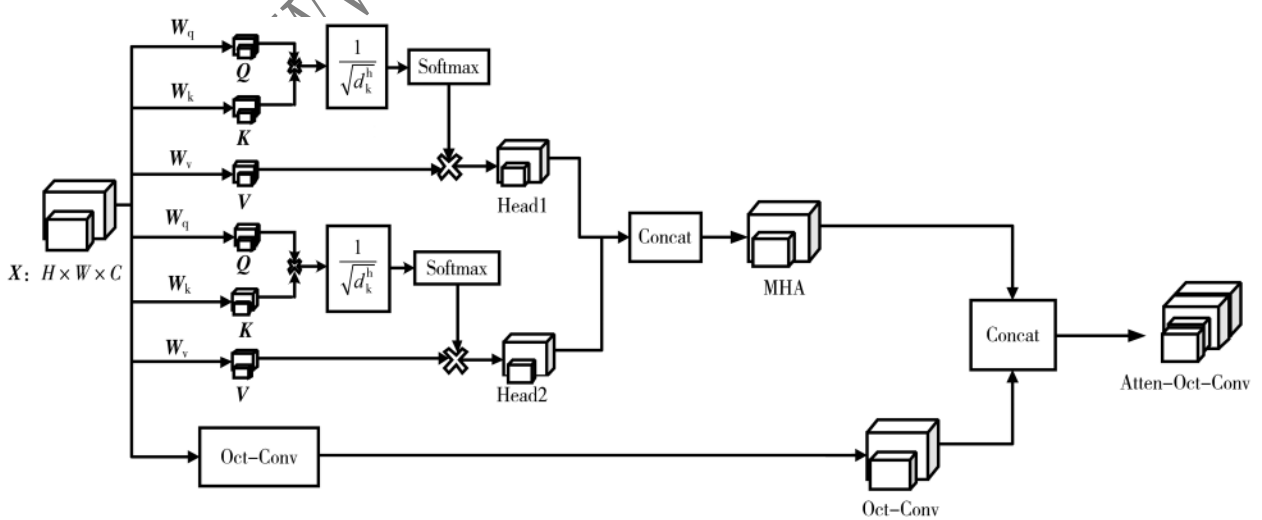


图 2 频域注意力卷积流程示意图及模块结构

www.bcachina.com

$$\text{head}-\text{l}: H\times W\times(1-\alpha)\times d_{\text{v}}^{\text{h}}$$

其中,head-h 表示每个头部的高频特征输出尺寸,head-l 表示每个头部的低频特征输出尺寸。经过试验发现,多头注意力相比较单注意力机制具有更好的表现能力,故采用 Nh=2 的多头注意力机制,将两个频率注意力头部进行拼接操作,得到多头频率注意力特征图 MFHA(Multiple Frequency Head Attention)尺寸为:

$$\textbf{MFHA}-\text{HIGH}: H\times W\times\alpha(2\times d_{\text{v}}^{\text{h}}) \tag{12}$$

$$\textbf{MFHA}-\text{LOW}: 0.5H\times 0.5W\times(1-\alpha)\times(2\times d_{\text{v}}^{\text{h}}) \tag{13}$$

为了提高精度,防止由于注意力机制忽略对于某些位置信息的学习,将原始输入经过 Oct-Conv 得到频率特征图:

$$\text{Oct}-\text{Conv}-\text{High}: H\times W\times\alpha\times(F_{\text{out}}-2d_{\text{v}}^{\text{h}}) \tag{14}$$

$$\text{Oct}-\text{Conv}-\text{Low}: 0.5H\times 0.5W\times(1-\alpha)\times(F_{\text{out}}-2d_{\text{v}}^{\text{h}}) \tag{15}$$

其中,Oct-Conv-High 表示频率卷积高频输出特征图尺寸,$F_{\text{out}}$ 表示频率卷积输出通道数,Oct-Conv-Low 表示频率卷积低频输出特征图尺寸。将其与 MFHA 进行拼接,得到结合频率信息与注意力信息的特征图,输出尺寸为:

$$\text{HIGH}: H\times W\times(\alpha\times F_{\text{out}}) \tag{16}$$

$$\text{LOW}: 0.5H\times 0.5W\times(1-\alpha)\times F_{\text{out}} \tag{17}$$

其中,HIGH 表示 Atten-Oct-CNN 高频输出尺寸,LOW 表示 Atten-Oct-CNN 低频输出尺寸。

## 2 改进的 C3D 步态识别网络

给定 3DCNN 块的局部时空特征,可以认为 3DCNN 模块关注了值得注意的时空的显著空间信息区域。另一方面,由于频域注意力卷积会将特征图划分为高频和低频两个部分,故当首次输入到频域注意力卷积 First-3D 模块中时,内部会将低频通道进行池化以缩小特征图尺寸。当输入 Media-3D 中时,将进行正常频域卷积操作,而输入 Last-3D 中会将低频部分上采样进行尺寸恢复。本文使用 C3D 网络作为主体,引入硬线层[16]划分输入,使用三个 3DCNN 层,同时将里面的三维 3DCNN 更换为本文改进的卷积模块,全连接改为 FC-512、FC-124,具体网络结构如图 3 所示。

## 3 实验

### 3.1 实验环境配置

本文实验使用 Pytorch 框架,实验的软硬件环境如表 1 所示。

表 1 实验环境

| CPU | 内存 | GPU | OS | CUDA |
|---|---|---|---|---|
| Intel i5 | 8 GB | GTX1060 | Win10 | 9.0 |

### 3.2 实验过程

实验使用中科院自动化所公开的大型步态数据库 CASIA dataset B。该数据库包含 124 个行人,每个行人有 11 个视角,分别从 0°~180°每 18°产生一个步态序列,每个视角下有三种步态状态,分别为正常状态(Normal)、穿大衣状态(Coat)以及背包状态(Bag)。

实验都是在同种状态下进行对比,Normal 状态下有 6 个序列 01~06,其中 01~04 序列用来训练模型,05~06 用来测试网络模型;Bag 状态有两个序列 01~02,01 序列作为训练集,02 序列作为测试样本;

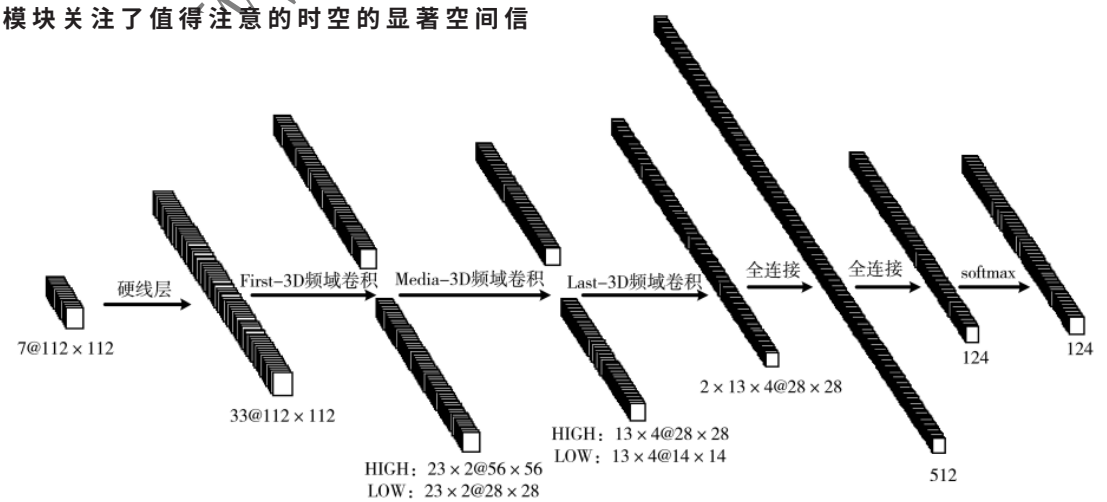


图 3 改进的 C3D 步态识别网络结构图

www.pcachina.com

Coat 状态有两个序列 01~02,01 序列作为训练样本,02 序列作为测试样本。

### 3.3 相同状态下的实验结果与分析

实验采用了 GenI、GPPE、STIPS 与 Deep CNN 作为对比方法,在 Normal、Bag 与 Coat 三种状态下,这几种方法 Normal 状态识别准确率最高均为 90%以上。而初始准确率最低的穿大衣状态起初只有 50%多的准确率,随着方法的创新,Deep CNN 已经达到 89%的准确率。最后在本文方法下,正常状态下准确率提高了 0.3%,背包状态下提升了 4.7%,穿大衣状态下提高了 3.7%。实验结果对比如表 2 所示。

表 2 不同状态下各方法实验结果对比

(%)

| 算法 | 正常 | 背包 | 穿大衣 |
|---|---|---|---|
| GenI[17] | 92.3 | 65.3 | 55.1 |
| GPPE[18] | 93.4 | 62.2 | 55.1 |
| STIPS[19] | 95.4 | 73.0 | 70.6 |
| Deep CNN[20] | 98.3 | 83.8 | 89.1 |
| 本文 | 98.6 | 88.5 | 92.8 |

### 3.4 相同状态下各视角实验结果与分析

进一步地,检查每个角度下单独的准确率可以发现,Normal 状态比较均衡,在 126°时有略微下降,在 18°~108°之间本文方法具有显著提升,这表明注意力比原先对于这两个角度有了较多关注。同时Bag 状态提升较为明显,在 28°、54°、90°与 144°分布与 Deep CNN 分布不同,注意力重新划分。Coat 状态下的分布规律较为一致,同时在频域卷积的强大学习力下准确率有很大的提升。各视角准确率曲线如图 4 所示。

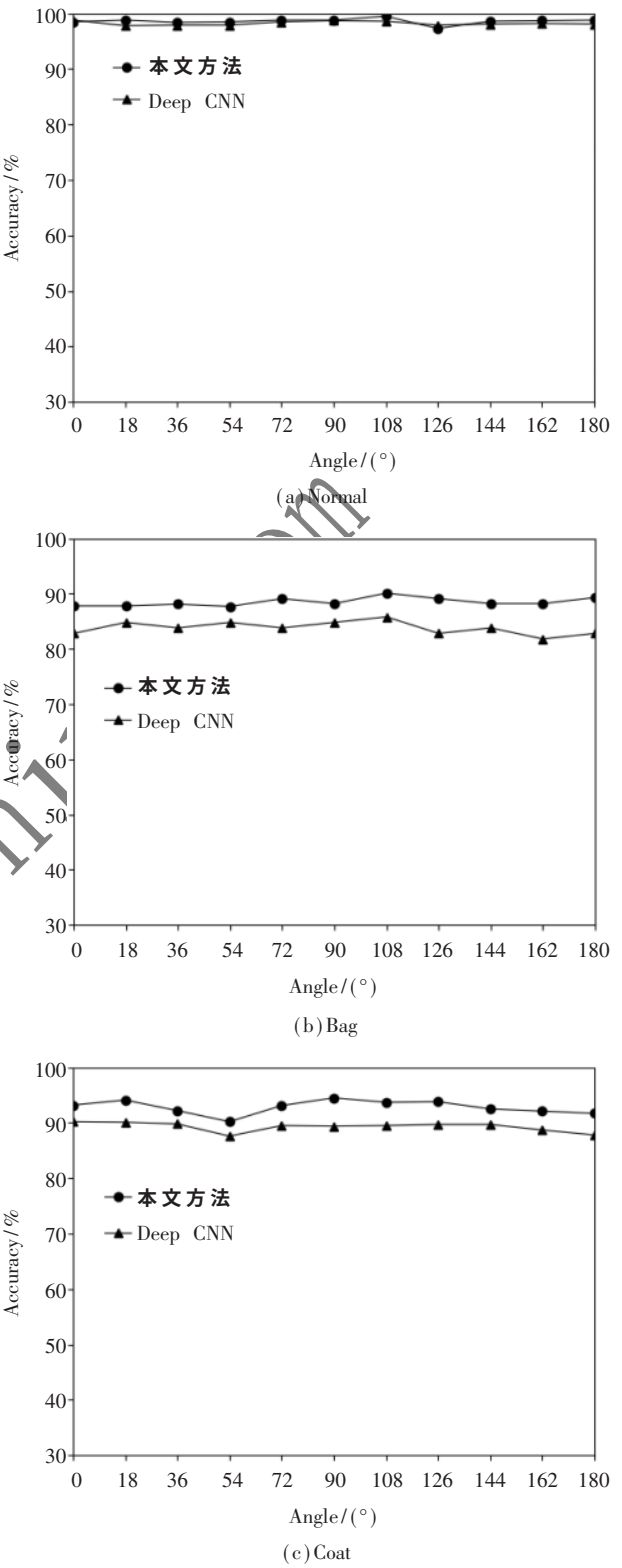


图 4 各视角准确率曲线

## 4 结论

由于传统基于步态模板的识别方法对于图像的预处理较多,大多使用步态能量图(GEI)模板,而 GEI 对于时间维度上步态的先后顺序没有要求,缺失了时间维度上的重要信息,本文通过硬线层将输入划分为 5 组图像序列进行输入,结合三维卷积的时空信息抽取能力,实现了一种端到端的学习方法。同时针对普通网络参数量大、步态特征的重要性分布不均提出一种频域注意力卷积的操作方法,通过合理分配网络学习的注意力分布,同时划分频域特征,大大减小了冗余信息。通过在 CASIA dataset B 的数据集进行实验,发现本文方法在相同状态下的识别准确率具有显著提升,同时对于每种状态的不同角度也有一定的准确率提升。

www.pcachi.com

### 参考文献

[1] 王曙光.指纹识别技术综述[J].信息安全研究,2016,

2(4):343-355.

[2] HE Y W,ZHANG J P,SHAN H M,et al.Multi-task GANs for view-specific feature learning in gait recognition[J].IEEE Transactions on Information Forensics & Security,2019,14(1):102-113.

[3] TAKEMURA N,MAKIHARA Y,MURAMATSU D,et al.On input/output architectures for convolutional neural network-based cross-view gait recognition[J].IEEE Transactions on Circuits & Systems for Video Technology,2019,29(9):2708-2719.

[4] WU Z F,HUANG Y Z,WANG L,et al.A comprehensive study on cross-view gait based human identification with deep CNNs[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(2):209-226.

[5] WOLF T,BABAEE M,RIGOLL G.Multi-view gait recognition using 3d convolutional neural networks[C].2016 IEEE International Conference on Image Processing.New York:IEEE,2016:4165-4169.

[6] HAN J,BHANU B.Individual recognition using gait energy image[J].IEEE Transactions on Pattern Analysis & Machine Intelligence,2006,28(2):316-322.

[7] LIAO R J,CAO C S,GARCIA E B,et al.Pose-based temporal-spatial network(PTSN) for gait recognition with carrying and clothing variations[C].Chinese Conference on Biometric Recognition.Springer,Cham,2017:474-483.

[8] CAMPBELL F W,ROBSON J G.Application of Fourier analysis to the visibility of gratings[J].The Journal of Physiology,1968,197(3):551-566.

[9] BOYNTON R M.Spatial vision[J].Annual Review of Psychology,1962,13(1):171-200.

[10] HU J,SHEN L,ALBANIE S,et al.Gather-excite:exploiting feature context in convolutional neural networks[C].Advances in Neural Information Processing Systems,2018:9401-9411.

[11] HU J,SHEN L,ALBANIE S,et al.Squeeze-and-excitation networks[J/OL].(2017-09-05).https://ui.adsabs.harvard.edu/abs/2017arXiv170901507H.

[12] PARK J,WOO S,LEE J Y,et al.BAM:Bottleneck attention module[J].arXiv:1807.06514,2018.

[13] WOO S,PARK J,LEE J Y,et al.CBAM:Convolutional block attention module[C].15th European Conference on Computer Vision,2018.Munich,Germany:Springer Verlag,2018.

[14] WANG X L,GIRSHICK R,GUPTA A,et al.Non-local neural networks[C].31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2018.Salt Lake City,United States:IEEE Computer Society,2018.

[15] GRAVES A.Generating sequences with recurrent neural networks[J].arXiv e-prints,arXiv:1308.0850,2013.

[16] JI S W,XU W,YANG M,et al.3D convolutional neural networks for human action recognition[C].Proceedings of 27th International Conference on Machine Learning(ICML),2010.Haifa,Israel,2010:221-231.

[17] BASHIR K,XIANG T,GONG S G.Cross-view gait recognition using correlation strength[C].2010 21st British Machine Vision Conference(BMVC),2010.Aberystwyth,United Kingdom:British Machine Vision Association,2010:1-11.

[18] JEEVAN M,JAIN N,HANMANDLU M,et al.Gait recognition based on gait pal and pal entropy image[C].2013 20th IEEE International Conference on Image Processing(ICIP),2013.Melbourne,VIC,Australia:IEEE Computer Society,2013.

[19] KUSAKUNNIRAN W.Recognizing gaits on spatio-temporal feature domain[J].IEEE Transactions on Information Forensics & Security,2014,9(9):1416-1423.

[20] ALOTAIBI M,MAHMOOD A.Improved gait recognition based on specialized deep convolutional neural network[J].Computer Vision & Image Understanding,2017,164:103-110.

(**收稿日期**:2020-03-29)

**作者简介:**

**赵国顺**(1995-),**通信作者**,**男**,**硕士研究生**,**主要研究方向:图像识别、计算机视觉**。E-mail:260486471@qq.com。

**方建安**(1966-),**男**,**博士**,**教授**,**主要研究方向:复杂系统建模、分析与控制,网络控制系统分析与综合,数字图像处理**。

**瞿斌杰**(1995-),**男**,**硕士**,**主要研究方向:图像处理、无人驾驶**。

www.pcachina.com

# 版权声明

经作者授权，本论文版权和信息网络传播权归属于《信息技术与网络安全》杂志，凡未经本刊书面同意任何机构、组织和个人不得擅自复印、汇编、翻译和进行信息网络传播。未经本刊书面同意，禁止一切互联网论文资源平台非法上传、收录本论文。

截至目前，本论文已经授权被中国期刊全文数据库（CNKI）、万方数据知识服务平台、中文科技期刊数据库（维普网）、JST 日本科技技术振兴机构数据库等数据库全文收录。

对于违反上述禁止行为并违法使用本论文的机构、组织和个人，本刊将采取一切必要法律行动来维护正当权益。

特此声明！

《信息技术与网络安全》编辑部

中国电子信息产业集团有限公司第六研究所