ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 3,May 2023,pp. 565－573
DOI:10. 16337/j. 1004-9037. 2023. 03. 006

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

特征分块重构的视频行人重识别算法

王锦华，周　非，白梦林，舒浩峰

（重庆邮电大学通信与信息工程学院，重庆 400065）

摘　要：基于视频的行人重识别是将一段视频轨迹与剪辑后的视频帧进行匹配，从而实现在不同的摄像头下识别同一行人。但由于现实场景的复杂性，采集到的行人轨迹会存在严重的外观丢失和错位，传统的三维卷积将不再适用于视频行人重识别任务。针对这一问题，提出三维特征分块重构模型，利用第一张特征图在水平分块的级别上对后续特征图进行对齐。在保证特征质量的前提下充分挖掘轨迹的时间信息，在特征重构模型后加入三维卷积核，并且将它与现有的三维卷积网络相结合。此外，还引入一种由粗到细的特征分块重构网络，不仅能使模型在两种不同尺度的空间维度上进行特征重构，还能进一步减少计算开销。实验表明，由粗到细的特征分块重构网络在MARS和DukeMTMC-VideoReID数据集上取得了良好的结果。

关键词：视频行人重识别；特征分块；特征重构；三维卷积；由粗到细的特征分块重构网络

中图分类号：TP391.41　　**文献标志码：**A

Video-Based Person Re-identification Algorithm Based on Feature Block Reconstruction

WANG Jinhua, ZHOU Fei, BAI Menglin, SHU Haofeng

(School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: Video-based person re-identification (Re-ID) is to match a video track with a clipped video frame, so as to recognize the same pedestrian under different cameras. However, due to the complexity of the real scene, the collected pedestrian trajectories will have serious appearance loss and dislocation, and the traditional 3D convolution will no longer be suitable for the video pedestrian re-identification task. Therefore, a 3D feature block reconstruction model(3D-FBRM) is proposed, which uses the first feature map to align subsequent feature maps at the level of horizontal blocks. In order to fully mine the time information of the trajectory under the premise of ensuring the quality of the features, a 3D convolution kernel is added after the FBRM, and it is combined with the existing 3D ConvNets. In addition, a coarse-to-fine feature block reconstruction network(CF-FBRNet) is introduced, which not only enables the model to perform feature reconstruction in two different scales of spatial dimensions, but also further reduces computational overhead. Experiments show that the CF-FBRNet achieves state-of-the-art results on the MARS and DukeMTMC-VideoReID datasets.

Key words: video-based person re-identification; feature block; feature reconstruction; 3D convolution;

收稿日期：2022-04-14；**修订日期：**2022-05-19

coarse-to-fine feature block reconstruction network(CF-FBRNet)

引 言

随着智能监控、智能安保的普及,政府和相关部门加大了对人民群众安全保障的力度,无论是在大街小巷还是在居民楼道里,都安装了大量高清摄像头。通过行人重识别技术分析从一组非重叠相机获取的行人数据并在已有的行人库中进行检索,可以高效地定位目标行人,无论是在刑侦安防还是行人行为理解上都有重大意义。在过去十几年里,卷积神经网络在基于图像的行人重识别技术上应用广泛[1-3]。但是单帧的图像所包含的信息有限,仅含有行人的空间信息(身份信息),当面临复杂场景时,容易发生误匹配。而视频片段能够提供行人丰富的时间信息(运动信息),更符合真实的场景。所以,近年来,有关基于视频的行人重识别技术越来越受关注。

由于人体结构的非刚性和实际场景中的视频轨迹存在大量噪声等原因,如何准确获取特定行人的时间信息是视频行人重识别的一大挑战。针对这一难题,现有的方法包括结合长短期记忆网络(Long short-term memory, LSTM)、3D卷积或非局部神经网络来获取行人时空特征。其中LSTM是一种特殊的RNN形式,与3D卷积相似,能处理序列的部分时间关系,而非局部神经网络能处理长时间的关系。Chen等[4]采用卷积神经网络(Convolutional neural network,CNN)和LSTM相结合的方法捕获行人的时空特征,并且利用注意力机制来进行特征聚合,提高两种特征的显著性。但考虑到模型的效率和计算成本,文献[5]设计了一种部分上下文感知注意力模型,结合时空域中的上下文信息,从视频帧中提取具有鲁棒性和鉴别性的部分特征,并且将模型插入到不同的卷积层中,再利用多头协作学习方案,在不增加模型结构和计算成本的前提下提升学习的准确性。当今,结合3D卷积的方法是视频行人重识别的研究热点。当3D卷积核通过堆叠帧时,能得到多个相邻帧的特征图,从而获取行人的时间信息。不过3D卷积核会将堆叠帧的相对位置都处理成一个值,所以当存在相邻帧特征不对齐的情况,这种方法会导致模型性能下降。如图1所示,从MARS数据集[6]抽取了4帧轨迹图,当t=1时红色小方框所对应的行人特征为女性的头部,但随着时间的推移,红框里所包含的女性头部信息越来越少,t=4时,仅包含男性的肩膀信息,如果继续采用3D卷积去提取这段轨迹特征,最终会影响模型的检索能力。文献[7]针对这一问题,提出了三维外观保持卷积(Appearance-preserving 3D convolution,AP3D),对相邻特征在像素级别上实现特征外观恢复。该方法将每一帧都视为锚(中心)帧,沿时间轴去对齐相邻帧,并将模块与3D卷积相结合,以此来替换原始3D残差网络里部分3D卷积核。虽然该方法在各大数据集都表现出良好的性能,且模型的参数量和浮点数运算量都比较少,但并未考虑连续帧里存在的噪声。利用锚帧恢复相邻帧的做法会增大轨迹噪声。相比之下,Liu等[8]提出运用重检测和连接模型对抽样的轨迹图进行预处理,通过预训练的检测器生成更紧密的边界框,再填充图像保持轨迹图像大小相同,最终得到对齐的视频轨迹。但这种预处理的方法需要额外的计算成本。

图1 MARS数据集中的一段带噪声的视频轨迹

Fig.1 A noisy video track in the MARS dataset

鉴于现有的方法没有很好地解决以上问题,考虑到图像特征在水平方向的相似性,本文提出一种由粗到细的特征分块重构网络(Coarse-to-fine feature block reconstruction network,CF-FBRNet),将视频轨迹的第一帧作为锚点去恢复后续帧。主要工作有:(1)将特征进行自适应水平分块,利用块间的余弦相似性实现在水平分块级别上的特征重构;(2)提出由粗到细的特征分块重构模块(Coarse-to-fine feature block reconstruction model,CF-FBRM),并与3D卷积(I3D[9]等)相结合,用于替换神经网络中的

2D卷积。本文提出的方法能在一定的模型参数量前提下耗费较低的运算成本,获取视频轨迹中行人的时间线索,并在MARS数据集和DukeMTMC-VideoReID[10]数据集上取得良好的结果。

1 相关工作

1.1 视频时间信息建模

在视频行人重识别任务中,常见的对视频轨迹时间信息进行建模的方法有两种:一种是使用2D卷积提取特征,再将得到的序列特征通过沿时间维度的池化或利用时间注意力等操作来聚合时间信息。其中池化操作包括平均池化和最大池化,这是一种十分简单的建模方法,即取视频轨迹片段特征的平均值或最大值作为最终的特征输出。不过池化操作没有考虑到视频序列的层次特征[11],利用时间注意力的方法可以对不同层次的特征加权,以聚合更充足的时间信息。Rao等[12]提出非局部注意力时间网络,考虑全局的前提下计算所有视频帧的注意力分数,捕获视频帧的长期依赖关系,但基于注意力的方法大多用于高级特征的聚合,缺乏鲁棒性;另一种方法是利用三维卷积神经网络对轨迹的局部时间关系建模,即用3D卷积核替换残差结构中的2D卷积核。Li等[13]提出的双流卷积网络就利用了3D卷积能学习时序特征的特性,采用3D卷积神经网络和2D卷积神经网络并行的结构,提取行人的时间、空间特征。在视频行人重识别任务上,采用3D卷积神经网络能够弥补2D卷积不能聚合连续特征中时间线索的缺点。并且有研究者已证明其具有强大的性能[9]。但仅仅使用传统的3D卷积核会导致模型参数量增加,还会增加GPU的计算开销。所以,如何在改进3D卷积神经网络性能的前提下控制模型参数量和计算成本是本文研究的重点之一。

1.2 特征图重构

对图像或特征进行重构是图像配准的过程[14],将检测到的行人特征进行特征匹配,再通过人工设定的映射函数实现图像转换。常见的特征检测和特征匹配方法有两种:一种是基于区域的方法,该方法用于特征不明显的情况。如图2所示,使用一对相同大小的窗口去覆盖相邻特征图(f_1,f_2)的局部区域,滑动f_2上的窗口,并计算窗口对所包含特征的相似性矩阵,最后挑选出矩阵中的最大值所对应的f_2窗口作为a区域所对应的窗口,但这种基于区域的方法有较大的计算复杂度,并且不适用于特征复杂的图像;另一种方法是基于特征的方法,通过特征提取器准确地提取图像特征,但得到的特征包含着图像的高层信息,不能直接处理,需要利用特征的空间关系或不变描述符等方法找到图像对之间的特征对应关系。Ghorbel等[15]提出一种改进的基于傅里叶的不变描述符,提取一对特征的傅里叶描述符并进行平均操作,以便计算中间描述符,提升了配准速度。

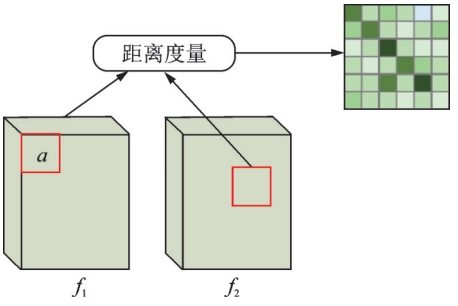

图2 对f_1的a区域生成相似性矩阵

Fig.2 Generating similarity matrix for the a region of f_1

本文采用的特征图重构的方法类似于基于特征的图像配准方法,利用特征的空间特性,将同一身份的不同特征图进行叠加,从而实现特征对齐。

2 本文方法

2.1 水平分块模块

通过特征提取器捕获的行人特征具有较高的特征级别,但当利用这些复杂的深层特征去实现配准过程时会产生大量的计算。为减少匹配过程中的计算开销,并且考虑到视频帧在同一水平方向具有相

似的特征，本文采用分块的方法将特征沿水平方向切分成M块，将原始复杂的特征X转换成M个较为粗略的特征向量：$\boldsymbol{x}_n$（$n=1,\cdots,M$）。这种方法能降低重构操作的计算量，比如，在单通道情况下，若采用逐像素的方法计算两个相邻特征（特征尺寸$\boldsymbol{X}\in\mathbf{R}^{H\times W}$）的相似性矩阵时，需要$O(H^2W^2)$的计算量，而在分块级别上去计算相似性矩阵时，仅需要$O(M^2)$的计算量，其中$M\ll HW$。

2.2 分块重构模块

在利用2.1节的分块特征进行重构操作前，本文选取抽样视频轨迹的第一帧作为锚帧，如图3所示，将锚帧扩展到与原始序列相同的时间长度T，本文所设置的时间长度为4。这种方法相较于AP3D[7]的好处在于所有帧都是在基于第一帧的基础上进行特征重构，不仅能减少视频序列中无关因素的干扰，还能解决第一帧与最后一帧图像存在行人错位严重的问题。并且，在视频帧的抽样阶段，为了避免产生多余的计算成本，本文并未对第一帧的质量进行判断。

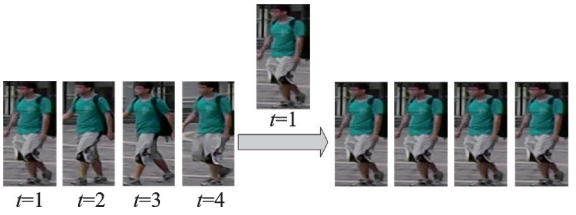

图3 取轨迹的第一帧作为锚帧

Fig.3 Taking the first frame of the trajectory as the anchor frame

本文利用原始特征$\boldsymbol{X}$和锚帧$\boldsymbol{Y}$的线性余弦相似性矩阵来重构$\boldsymbol{X}$。对齐模型如图4所示，首先利用2.1节的方法，对两种特征进行水平分块得到粗略的特征表示，把$\boldsymbol{X}$的原始特征向量从$C\times T\times H\times W$水平分成$M$块，块集由式(1)所示，其中$\boldsymbol{x}_i\in\mathbf{R}^{C\times T\times M\times 1}$。同样，$Y$也分成$M$块。

$$\boldsymbol{X}'=\{\boldsymbol{x}_i|i=1,\cdots,M\} \tag{1}$$

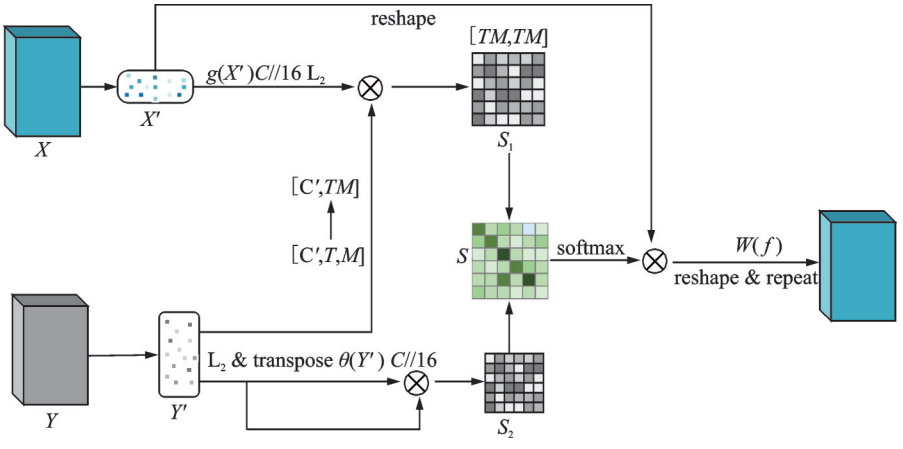

图4 特征分块重构模块

Fig.4 Feature block reconstruction module

然后，使用$1\times1\times1$的映射函数θ、g将特征块集投到低维嵌入空间中进行计算，通过式(2)的线性组合，得到相似性矩阵。其中，$\boldsymbol{S}_1$表示特征块集$\boldsymbol{X}'$、$\boldsymbol{Y}'$的余弦相似性，$\boldsymbol{S}_2$表示$\boldsymbol{Y}'$自身的相关性因素，并且使用L_2范式来约束$\boldsymbol{S}_1$、$\boldsymbol{S}_2$、ω表示权重因子，用于增大具有较高相似性的水平块特征之间的相似性系数。由式(3)得到特征块$\boldsymbol{x}_i$、$\boldsymbol{y}_j$的余弦相似性。

$$\boldsymbol{S}=\omega\boldsymbol{S}_1+\boldsymbol{S}_2 \tag{2}$$

$$\boldsymbol{S}_1(\boldsymbol{x}_i,\boldsymbol{y}_j)=\frac{g(\boldsymbol{x}_i)^{\mathrm{T}}\cdot\theta(\boldsymbol{y}_j)}{\|g(\boldsymbol{x}_i)\|^{\mathrm{T}}\|\theta(\boldsymbol{y}_j)\|} \tag{3}$$

最后,利用Softmax函数归一化的相似性矩阵$\boldsymbol{S}$去填充特征块集$\boldsymbol{X}'$。如式(4)所示,考虑到$\boldsymbol{x}_i$可能在$\boldsymbol{Y}'$中对应多个相似的块,因此,每一个重构后的特征块f_i为$\boldsymbol{x}_i$与所有$\boldsymbol{y}_j$的相似性加权求和,从而实现校准。将重构特征图f通过映射函数W(1×1×1卷积)映射到原始特征空间中,在恢复特征尺寸的同时扩大特征的时间维度,保证重构后的特征能顺利通过3D卷积运算。

$$f_i=\sum_j\frac{\boldsymbol{x}_i e^{s(x_i,y_j)}}{\sum_j e^{s(x_i,y_j)}} \tag{4}$$

2.3 由粗到细的特征分块重构模块

为了不影响图像重构算法的准确性,同时进一步减少模型的计算量,本文将输入特征$\boldsymbol{X}\in\mathbf{R}^{C\times T\times H\times W}$沿通道维度切分成两块,第一块保持原始特征尺度$\boldsymbol{X}_1\in\mathbf{R}^{C/2\times T\times H\times W}$,第二块采用下采样操作得到粗尺度$\boldsymbol{X}_2\in\mathbf{R}^{C/2\times T\times H/2\times W/2}$,然后将切分的张量分别输入到重构模块中,最后将粗尺度特征上采样后与细尺度沿通道维度连接起来,恢复原始特征维度,如图5所示。由于进行了下采样操作,本文还考虑两个重构模块中的水平分块个数,具体在3.3.2节叙述。当粗尺度对齐分支的水平分块数为细尺度的一半时,所需计算量为原始分块对齐模型的5/8。

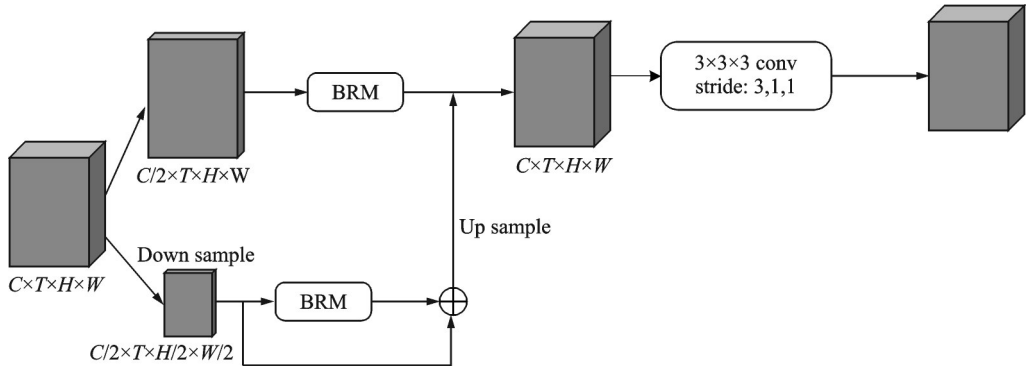

图5 由粗到细的3D特征分块重构模块

Fig.5 Coarse-to-fine 3D feature block reconstruction model

在重构特征后添加步长为3×1×1,核为3×3×3的卷积。用改进后的重构模块替换I3D残差块中的3D卷积核,如图6(a)所示。同时,本文还考虑了另一种3D卷积结构——P3D[16],P3D将原始的

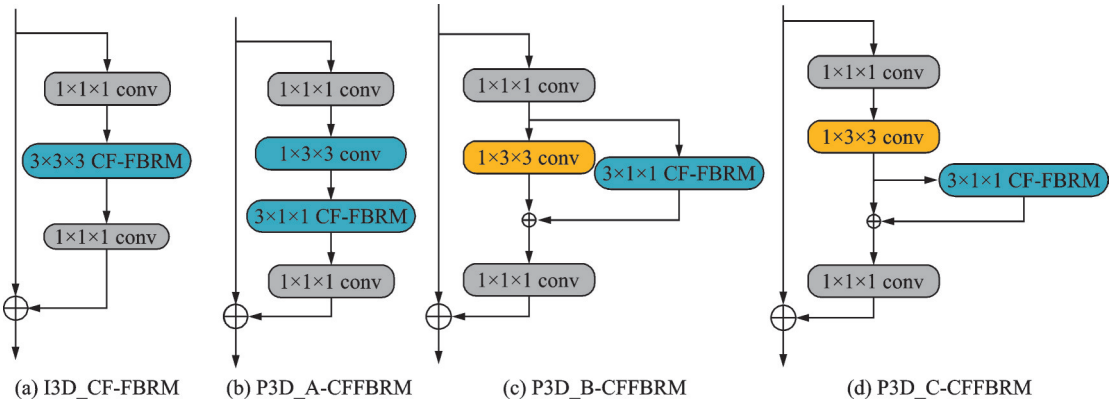

图6 I3D、P3D残差块和对应改进的I3D_CFFBRM、P3D_CFFBRM

Fig.6 I3D, P3D residual blocks and corresponding improved I3D_CFFBRM, P3D_CFFBRM

3D卷积核转换成2个串行或并行传输的卷积核,增加了结构多样性,提升神经网络的学习能力。在重构特征后添加核大小为$3\times1\times1$的卷积,以此来替换原始P3D中的$3\times1\times1$卷积核,对P3D的3种模式进行了改进,如图6(b~d)所示。

3 实验结果与分析

3.1 数据集和参数设置

本实验使用的数据集是视频行人重识别中常用的MARS和DukeMTMC-VideoReID。MARS数据集是由6个摄像头采集的1 261个行人轨迹,平均每个行人有13个轨迹,每个轨迹有59帧,其中625个行人用于训练,636个行人用于测试。DukeMTMC-VideoReID数据集包含1 404个行人身份,其中702个用于训练,702个用于测试,视频轨迹采用每12帧进行采样。

硬件配置为GTX 2080Ti GPU,在基于PyTorch的深度学习框架上搭建网络模型。输入帧尺寸调整为256×128,采用随机翻转和随机擦出作为数据增强。模型共训练200个epoch,使用Adam优化网络参数,ε设为5×10^{-4},初始学习率为3×10^{-4},学习率每40个epoch乘以0.1。在训练阶段,每批图像包含8个行人身份,每个行人包含4张随机抽取的视频帧。在测试阶段,将每个视频序列分成几个32帧的视频片段,通过CF-FBRnet提取特征,取视频片段特征的平均表示,最后使用余弦距离来度量Query和Garrery之间的特征,并完成检索。

3.2 损失函数及实验评估

本文采用交叉熵损失与三元组损失共同优化特征,其中,三重损失采用余弦距离来度量。

3.2.1 评估指标

本文采用平均精度均值(Mean average precision,mAP)和累计匹配特征(Cumulative matching characteristics ,CMC)作为评估指标。

3.2.2 对比实验

为了验证本文方法的有效性,用I3D、P3D以及改进后的I3D和P3D分别去替换2D卷积神经网络的阶段2、阶段3的5个2D残差块。在MARS数据集的实验结果如表1所示,表中还对比了模型的参数量(Parameter)和浮点数运算量(Floating point operation,FLOP)。在进行模型的参数量和浮点数运算量计算时,保持输入张量$\boldsymbol{X}\in\mathbf{R}^{3\times4\times256\times128}$。发现,在用3D残差块替换2D残差块后,会增大模型的参数量和浮点数运算量,但模型的性能明显降低。而本文提出的方法能在3D卷积神经网络的基础上增加较少的模型参数量,并提升网络性能。与传统的I3D比较,改进的方法在Rank-1和mAP上分别提升了0.8%、1.1%;与传统的P3D比较,对3种模式下的改进都有所提升,其中采用P3D_B-CFFBRM的效果最佳,在Rank-1和mAP上分别提升了1.7%、3.1%。因此后续的实验都是基于P3D_B-CFFBRM的基础上进行的。

表1 本文方法与基线参数量和浮点数运算量比较

Table 1 Comparison of the proposed method with the baseline number of parameters and floating point operations

方法	MARS		Parameter/10^3	FLOP/10^6
	Rank-1/%	mAP/%		
C2D	88.9	83.4	24.79	16.31
I3D	88.5	83.1	28.92	19.33
P3D_A	88.9	83.2	25.48	16.85
P3D_B	88.5	83.1	25.48	16.85
P3D_C	88.8	83.0	25.48	16.85
I3D_CFFBRM	89.3	84.2	28.94	19.32
P3D_A-CFFBRM	89.1	84.2	25.63	16.84
P3D_B-CFFBRM	90.2	86.2	25.62	16.82
P3D_C-CFFBRM	89.8	85.6	25.63	16.87

表2是用P3D_B-CFFBRM去替换ResNet50中不同阶段的2D残差块的结果,n表示替换的模块数量。由于第一阶段只能获取图像浅层特征,所以实验中并未对该阶段进行改进。可以看出,替换越多的块效果越好,因为P3D_B-CFFBRM能使网络更好地聚合时空信息,但太多又会导致过拟合。最好的效果是替换

阶段2、阶段3中的5个2D残差块。

为了探讨不同权重因子ω对模型性能的影响,原始分支水平分16块,下采样分支分8块,将ω设置为2、4、6、8,分别在MARS数据集上训练模型。结果如表3所示,当权重因子为4时模型的效果达到最佳,如果继续扩大会使模型的性能逐渐下降。分析可能的原因是由于较小的权重因子不足以将相似度较低的特征区分开来,而较大的权重因子会导致一些比较相似的特征被区分开。

表4是研究不用的分块个数对模型性能的影响。考虑到ResNet50第3阶段的输出特征尺寸$\boldsymbol{X}\in\mathbf{R}^{16\times 8}$,所以进行降维操作后变为$\boldsymbol{X}_2\in\mathbf{R}^{8\times 4}$,为了使重构模块能替换任意阶段的卷积核,即原始分支最大水平分块个数为16,下采样分支最大采样个数为8。因此,对比了4种不同分块搭配,并在MARS数据集上训练模型。结果显示,当两个分支采取最大分块个数时,模型效果达到最佳。

表2 用P3D_B-CFFBRM替换不同阶段不同数量的残差块的结果

Table 2 Results of replacing a different number of residual blocks at different stages with P3D_B-CFFBRM %

Stage	n	MARS		DukeMTMC-VideoReID	
		Rank-1	mAP	Ranl-1	mAP
Stage_2	1	89.6	85.3	96.9	96.3
Stage_3	1	89.5	85.0	96.7	96.2
Stage_4	1	89.0	84.7	96.7	96.2
$\text{Stage}_{2,3}$	5	90.2	86.2	97.6	96.9
$\text{Stage}_{2,4}$	5	89.9	85.8	96.9	96.4
$\text{Stage}_{3,4}$	5	89.5	85.4	96.5	96.0
$\text{Stage}_{2,3,4}$	6	89.7	85.5	97.0	96.4

表3 不同的尺度因子对模型性能的影响

Table 3 Effect of different scaling factors on model performance %

ω	MARS	
	Rank-1	mAP
2	89.4	84.6
4	90.2	86.2
6	89.7	85.4
8	89.5	85.5

表4 不同分块对模型性能影响

Table 4 Effect of different blocks on model performance %

分块	MARS	
	Rank-1	mAP
(8,4)	88.9	84.9
(8,8)	89.0	85.0
(16,4)	89.2	85.0
(16,8)	90.2	86.2

3.2.3 与先进方法比较

将提出的方法与当今一些先进的方法进行比较,如表5所示。尺度因子ω设置为4,原尺度分支水

表5 不同方法在MARS、DukeMTMC-VideoReID数据集上的性能比较

Table 5 Performance comparison of different methods on MARS and DukeMTMC-VideoReID datasets %

方法	MARS		DukeMTMC-VideoReID	
	Rank-1	mAP	Rank-1	mAP
VRSTC[17]	88.5	82.3	95.0	93.5
NVAN[18]	90.0	82.8	96.3	94.9
TCLNet[19]	89.8	85.1	96.9	96.2
AP3D[7]	90.1	85.1	96.3	95.6
MG-RAFA[20]	88.8	85.9	—	—
STGCN[21]	90.0	83.7	97.3	95.7
DL+CF-AAN[8]	91.3	86.5	96.7	96.2
STMN[22]	89.9	83.7	96.7	94.6
STRF[23]	90.3	86.1	97.4	96.4
Ours	90.2	86.2	97.6	96.9

平分为16块,降维后的尺度分支分为8块。在MARS这个最具挑战性的数据集上,本文所提出的方法在Rank-1可达90.2%,mAP可达86.2%;在Duke MTMC-VideoReID大型数据集上,Rank-1可达97.6%,mAP可达96.9%。由此可见,本文方法在不消耗额外计算成本的前提下能达到良好的水平。

4 结束语

本文提出了一种由粗到细的特征分块重构算法,将特征分块、图像配准、尺度融合相结合,在水平特征的角度上去重构图像。本文模型不仅能解决3D卷积无法处理不对齐行人的问题,还能够轻易地与现有的卷积神经网络相结合,提升网络学习能力。并通过在两个最具代表性的视频行人重识别数据集上进行对比试验,证明了方法的有效性。后续将进一步改进本文算法,提高算法的识别效率,以适用于更为复杂的现实场景。

参考文献:

[1] ZHANG Z, LAN C, ZENG W, et al. Relation-aware global attention for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Seattle:IEEE, 2020: 3186-3195.

[2] ZHENG Z, YANG X, YU Z, et al. Joint discriminative and generative learning for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach:IEEE, 2019: 2138-2147.

[3] EOM C, HAM B. Learning disentangled representation for robust person re-identification[J]. Advances in Neural Information Processing Systems, 2019. DOI:10.48550/arXiv.1910.12003.

[4] CHEN L, YANG H, GAO Z. Joint attentive spatial-temporal feature aggregation for video-based person re-identification[J]. IEEE Access, 2019, 7: 41230-41240.

[5] WU D, YE M, LIN G, et al. Person re-identification by context-aware part attention and multi-head collaborative learning[J]. IEEE Transactions on Information Forensics and Security, 2021, 17: 115-126.

[6] ZHENG L, BIE Z, SUN Y, et al. MARS: A video benchmark for large-scale person re-identification[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2016: 868-884.

[7] GU X, CHANG H, MA B, et al. Appearance-preserving 3D convolution for video-based person re-identification[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 228-243.

[8] LIU C T, CHEN J C, CHEN C S, et al. Video-based person re-identification without bells and whistles[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 1491-1500.

[9] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu:IEEE, 2017: 6299-6308.

[10] WU Y, LIN Y, DONG X, et al. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE, 2018: 5177-5186.

[11] 郑伟诗,吴岸聪.非对称行人重识别:跨摄像机持续行人追踪[J].中国科学:信息科学,2018,48(5): 545-563.
ZENG Weishi, WU Ancong. Asymmetric person re-identification: Continuous person tracking across cameras[J]. Chinese Science: Information Science, 2018, 48(5): 545-563.

[12] RAO S, CAO P, RAHMAN T, et al. Non-local attentive temporal network for video-based person re-identification[C]//Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Taipei, China: IEEE, 2019: 1-8.

[13] LI J, ZAHNG S, HUANG T. Multi-scale 3D convolution network for video based person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]:AAAI,2019,1: 8618-8625.

[14] ZITOVA B, FLUSSER J. Image registration methods: A survey[J]. Image and Vision Computing, 2003, 21(11): 977-1000.

[15] GHORBEL E, GHORBEL F, SAKLY I, et al. Fast blending of planar shapes based on invariant invertible and stable descriptors[C]//Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR). Milan: IEEE, 2021:

10259-10265.

[16] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5533-5541.

[17] HOU R, MA B, CHANG H, et al. VRSTC: Occlusion-free video person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7183-7192.

[18] LIU C T, WU C W, WANG Y C F, et al. Spatially and temporally efficient non-local attention network for video-based person re-identification[EB/OL].[2019-12-30].http://arxiv.org/pdf11908.01683.pbf.

[19] HOU R, CHANG H, MA B, et al. Temporal complementary learning for video person re-identification[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 388-405.

[20] ZHANG Z, LAN C, ZENG W, et al. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 10407-10416.

[21] YANG J, ZHENG W S, YANG Q, et al. Spatial-temporal graph convolutional network for video-based person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 3289-3299.

[22] EOM C, LEE G, LEE J, et al. Video-based person re-identification with spatial and temporal memory networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 12036-12045.

[23] AICH A, ZHENG M, KARANAM S, et al. Spatio-temporal representation factorization for video-based person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 152-162.

作者简介:

王锦华(1997-),**通信作者**,男,硕士研究生,研究方向:结合深度学习的行人重识别算法,E-mail:S200131232@stu.cqupt.edu.cn。

周非(1977-),男,教授,硕士生导师,研究方向:视频信号处理、信息安全、无线定位。

白梦林(1998-),女,硕士研究生,研究方向:结合深度学习的行人重识别算法。

舒浩峰(1998-),男,硕士研究生,研究方向:结合深度学习的行人重识别算法。

(编辑:夏道家)