



文章编号:1001-9081(2019)08-2223-07

doi:10.11772/j.issn.1001-9081.2018122505

## 基于深度多视图特征距离学习的行人重识别

邓 轩<sup>1</sup>, 廖开阳<sup>1,2\*</sup>, 郑元林<sup>1,3</sup>, 袁 晖<sup>1</sup>, 雷 浩<sup>1</sup>, 陈 兵<sup>1</sup>

(1. 西安理工大学 印刷包装与数字媒体学院, 西安 710048; 2. 陕西省印刷包装工程技术研究中心, 西安 710048;  
3. 陕西省印刷包装工程重点实验室, 西安 710048)  
(\*通信作者电子邮箱 liaokaiyang@xaut.edu.cn)

**摘要:**传统手工特征很大程度上依赖于行人的外观特征,而深度卷积特征作为高维特征,直接用来匹配图像会消耗大量的时间和内存,并且来自较高层的特征很容易受到行人姿势背景杂波影响。针对这些问题,提出一种基于深度多视图特征距离学习的方法。首先,提出一种新的整合和改善深度区域的卷积特征,利用滑框技术对卷积特征进行处理,得到低维的深度区域聚合特征并使其维数等于卷积层通道数;其次,通过交叉视图二次判别分析方法,从深度区域聚合特征和手工特征两个角度出发,提出一种多视图特征距离学习算法;最后,利用加权融合策略来完成传统特征和卷积特征之间的协作。在Market-1501和VIPeR数据集上的实验结果显示,所提融合模型的Rank1值在两个数据集上分别达到80.17%和75.32%;在CUHK03数据集新分类规则下,所提方法的Rank1值达到33.5%。实验结果表明,通过距离加权融合之后的行人重识别的精度明显高于单独的特征距离度量取得的精度,验证了所提的深度区域特征和算法模型的有效性。

**关键词:**行人重识别;卷积神经网络;区域聚合特征;加权融合策略;距离度量

中图分类号: TP183; TP391.4 文献标志码:A

### Person re-identification based on deep multi-view feature distance learning

DENG Xuan<sup>1</sup>, LIAO Kaiyang<sup>1,2\*</sup>, ZHENG Yuanlin<sup>1,3</sup>, YUAN Hui<sup>1</sup>, LEI Hao<sup>1</sup>, CHEN Bing<sup>1</sup>

(1. College of Printing, Packaging Engineering and Digital Media Technology, Xi'an University of Technology, Xi'an Shaanxi 710048, China;  
2. Printing and Packaging Engineering Technology Research Centre of Shaanxi Province, Xi'an Shaanxi 710048, China;  
3. Key Laboratory of Printing and Packaging Engineering of Shaanxi Province, Xi'an Shaanxi 710048, China)

**Abstract:** The traditional handcrafted features rely heavily on the appearance characteristics of pedestrians and the deep convolution feature is a high-dimensional feature, so, it will consume a lot of time and memory when the feature is directly used to match the image. Moreover, features from higher levels are easily affected by human pose or background clutter. Aiming at these problems, a method based on deep multi-view feature distance learning was proposed. Firstly, a new feature to improve and integrate the convolution feature of the deep region was proposed. The convolution feature was processed by the sliding frame technique, and the integration feature of low-dimensional deep region with the dimension equal to the number of convolution layer channels was obtained. Secondly, from the perspectives of the deep regional integration feature and the handcrafted feature, a multi-view feature distance learning algorithm was proposed by utilizing the cross-view quadratic discriminant analysis method. Finally, the weighted fusion strategy was used to accomplish the collaboration between handcrafted features and deep convolution features. Experimental results show that the Rank1 value of the proposed method reaches 80.17% and 75.32% respectively on the Market-1501 and VIPeR datasets; under the new classification rules of CUHK03 dataset, the Rank1 value of the proposed method reaches 33.5%. The results show that the accuracy of pedestrian re-identification after distance-weighted fusion is significantly higher than that of the separate feature distance metric, and the effectiveness of the proposed deep region features and algorithm model are proved.

**Key words:** person re-identification; Convolutional Neural Network (CNN); regional integration feature; weighted fusion strategy; distance metric

## 0 引言

行人重识别问题是通过多个摄像机视图判断行人是否为同一目标的过程,当前已广泛应用于跟踪任务的视频分析和

行人检索中。但是在实际生活中,由于行人重识别受到视角、光照、姿态、背景杂波和遮挡等因素的影响,使得行人图像在不重叠的摄像机视图中的差异性较大,如何减少和降低这种差异性对行人重识别的影响,是当前行人重识别中存在的巨

收稿日期:2018-12-19;修回日期:2019-02-28;录用日期:2019-03-21。

基金项目:国家自然科学基金资助项目(61671376, 61771386);陕西省教育厅科学项目(18JK0556)。

作者简介:邓轩(1995—),女,山东菏泽人,硕士研究生,主要研究方向:深度学习、图像处理;廖开阳(1976—),男,湖北荆州人,讲师,博士,主要研究方向:机器视觉、人工智能;郑元林(1975—),男,山东泰安人,副教授,博士,主要研究方向:色彩管理、彩色图像质量评估、颜色科学;袁晖(1993—),男,江西赣州人,硕士研究生,主要研究方向:深度学习、图像处理;雷浩(1995—),男,湖南岳阳人,硕士研究生,主要研究方向:深度学习、图像处理;陈兵(1997—),男,陕西安康人,主要研究方向:图像处理。



大问题和面临的严峻挑战。

特征表示和度量学习是行人重识别系统中的两个基本要素,而且由于特征表示是构成距离度量学习的基础,使其在行人重识别系统中显得尤为重要。虽然度量学习具有一定的有效性,但它很大程度上取决于特征表示的质量。因此,当前许多研究致力于开发更加复杂和具有鲁棒性的特征,用以描述可变条件下的视觉外观,可以将其提取的特征划分为两类:传统特征和深度特征。

部分学者对传统特征的研究多集中于设计具有区分性和不变性特征,着手于不同外观特征的拼接,克服了重识别任务中的交叉视图的外观变化,使得识别更加可靠。Liao 等<sup>[1]</sup>提出局部最大出现特征 (Local Maximal Occurrence Feature, LOMO) 来表示每个行人图像的高维特征,不仅从图像中提取尺度不变的局部三元模式 (Scale Invariant Local Ternary Pattern, SILTP) 和 HSV (Hue, Saturation, Value) 颜色直方图以形成高级描述符,还分析局部几何特征的水平发生,并最大化出现以稳定地表示行人图像。

当前深度学习提供了一种强大的自适应方法来处理计算机视觉问题,而无需过多地对图像进行手工操作,广泛应用于行人重识别领域。部分研究侧重于通过卷积神经网络 (Convolutional Neural Network, CNN) 框架学习特征和度量,将行人重新编码作为排序任务,将图像对<sup>[2]</sup>或三元组<sup>[3]</sup>输入 CNN。由于深度学习需要依赖于大量的样本标签,因而使得该方法在行人重识别领域中具有应用的局限性。

度量学习旨在开发一种判别式匹配模型来测量样本相似性,例如针对类内样本数目少于类间样本数目的情况,丁宗元等<sup>[4]</sup>提出了基于距离中心化的相似性度量算法。Koestinger 等<sup>[5]</sup>通过计算类内和类间协方差矩阵之间的差异,设计了简单而有效的度量学习方法,但是所提出的算法对特征表示的维度非常敏感。作为一种改进,Liao 等<sup>[1]</sup>通过同时学习更具辨别的距离度量和低维子空间提出了一种交叉视图二次判别分析 (Cross-view Quadratic Discriminant Analysis, XQDA) 方法。从实验结果来看,XQDA 是一种可以实现高性能的鲁棒性方法。

卷积神经网络提取的特征对图像具有较强的描述能力,通常可以提取三维的卷积特征以及单维的全连接特征向量,但卷积层特征比全连接层特征更适合用来识别图像,故本文使用微调过的 Resnet-50 模型作为研究的网络模型,提取其卷积层的特征。由于卷积特征是三维特征,如果将其展成一维的特征向量,其维数必然很高,使用高维特征在数据库中的图像进行匹配,必然会花费大量的时间,增加计算的复杂度。因此如何将三维特征变成一维,并能够保证特征的简单化是本次研究的一个核心问题,基本思路是将通过滑框操作,将三维的卷积特征压缩成一维的特征向量。由于来自较高层的特征具有大的感受野,容易受到人类姿势和背景杂波的污染,不能充分地应用于行人的重识别。而手工制作的不同的外观特征,旨在克服重新识别任务中的跨视图外观变化,有时会更加独特和可靠。所以本次研究的另一个核心问题是通过操作完成深度特征和传统手工特征的融合,使之相互影响、互相协作,进而提高识别的准确度。于是,本次研究利用区域特征向量聚合的方法,在微调卷积神经网络的基础上,提出了一个新的低维深度特征向量,并提出了一种深度多视图特征距离

学习的算法模型,从深度区域聚合特征和传统手工特征两个角度出发,利用加权策略,以一种有效的方式完成深度特征与传统手工特征之间的协作,用参数加权融合来调整两个特征的相对重要性。

本文的工作主要体现在以下两个方面:

1) 提出新的区域特征向量聚合的方法,将高维卷积特征向量变成低维的全局特征向量,并提高了图像局部信息的描述能力。

2) 提出了深度多视图特征距离学习的新方案,从深度区域聚合特征和传统手工特征两个角度出发,通过 XQDA 度量学习完成传统特征和深度特征之间的协作,利用参数加权融合的方式来判断传统特征和深度特征的相对重要性。

## 1 相关工作

特征表示是行人重识别的基本问题。许多现有的研究集中于开发强大和复杂的特征来描述在显著不同的条件下产生的高度可变的视觉外观。

手工制作的特征经常用于行人识别,例如通过利用对称和不对称的感性主体,Farenzena 等<sup>[6]</sup>提取了三种特征类型来模拟人类外观,包括最大稳定颜色区域、加权颜色直方图和经常性高结构色块。LOMO 通过分析局部特征的水平发生,并最大化出现以稳定地表示重新识别图像。这些方法对解决低分辨率和遮挡图像以及姿态、照明和视点变化带来的识别问题都很有效。由于传统的颜色信息不是描述颜色的最有效的方法,张耿宁等<sup>[7]</sup>将颜色标签特征与颜色和纹理特征融合,并通过区域和块划分的方式提取直方图来描述图像特征。不同外观组合成的传统特征向量通常维数较高,为了解决这个问题,孙金玉等<sup>[8]</sup>提出典型相关分析 (Canonical Correlation Analysis, CCA) 方法进行特征投影变换,通过提高特征匹配能力来避免高维特征运算引起的维数灾难问题,

鉴于 CNN 的成功,使用 CNN 学习深度特征最近受到关注。目前有许多研究在寻求行人外观的独特和有效特征的组合,并且证明利用集成编码的补充信息来发现完整数据表示的多视图特征是可行的。Wu 等<sup>[2]</sup>提出的特征融合网络 (Feature Fusion Network, FFN) 将卷积神经网络 (CNN) 深度特征和手工提取的特征 (RGB、HSV、YCbCr、Lab、YIQ 五种颜色空间提取的颜色特征和多尺度多方向 Gabor 滤波器提取的纹理特征) 相结合,认为传统的直方图特征可以补充 CNN 特征,并将两者融合,得到了一个更具辨别性和紧密性的新的深度特征表示。Tao 等<sup>[9]</sup>提出用折衷参数来完成深度特征与传统特征的协作,文中所构建的网络模型卷积层数少,深度特征采用全连接层特征。相比之下,对于特定任务的识别来说,较高的卷积层特征更适合用于图像识别,故本文采用微调的深度网络模型,并对卷积特征采用滑框操作,形成区域特征向量,并利用加权融合策略来判断深度区域特征和 LOMO 特征的相对重要性。

## 2 算法模型

在本章中,将介绍本文提出的算法模型,以完成深度区域聚合特征和传统特征之间的协作。对于区域特征向量,即利用微调的 Resnet-50 模型提取三维卷积特征向量,并利用滑框技术对卷积特征进行处理,卷积层激活映射在各自对应的滑



框尺度和步长的处理下,滑框每移动一个步长就会得到一个局部区域的激活映射并累加聚合成一个区域向量。在图像检索领域中,Gong等<sup>[10]</sup>提出的多目标规划(Multiple Objectives Programming, MOP)算法采用一种多尺度的滑框对原图进行处理,本文采用类似的方法对卷积特征平面用滑框进行处理得到区域特征向量,然后将区域向量通过加和方式得到全局特征。如图1所示,本文的算法模型主要分为以下三个部分:

1)聚合区域特征向量,得到低维深度特征。使用微调的Resnet-50模型提取图像的三维卷积特征,设计不同尺度的滑框,并将滑框作用于网络的最后一个卷积层,每个滑框内的元素直接相加求和得到多个局部特征向量,经过L2归一化,最后直接相加得到低维的深度全局特征向量。

2)对于参考集和测试集中的行人图像分别提取传统LOMO特征。

3)分别通过度量方法XQDA训练两个特征获得两个距离,并通过参数加权融合获得最终距离,再根据最终距离得到匹配的等级。

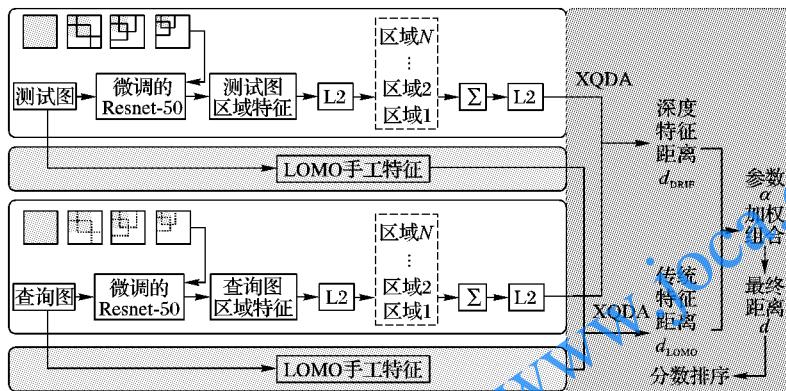


图1 本文算法的结构

Fig. 1 Framework of the proposed algorithm

## 2.1 区域聚合

本节介绍如何利用CNN卷积层的激活映射来获得图像区域的表达,将特征平面分成多个不同区域进行单独计算,再聚合区域向量以产生在行人重识别阶段中使用的低维全局特征向量,并增强图像深度特征的局部描述能力。

算法选取的是图像在微调模型Resnet-50上最后一个卷积层产生的激活映射。提取行人单幅图像的激活映射并定义为 $A_{i,j,k}$ , $i,j,k$ 分别代表激活映射的宽( $W$ )、长( $H$ )、通道( $C$ )。为了提高局部特征在算法中的比重,将滑框技术应用到提取的卷积激活映射的长宽截面上。滑框的使用将会被分为 $L$ 个尺度等级进行,即用 $L$ 个不同比例的区域进行采样。

在最大尺度( $L=1$ )时,区域尺寸被确定为尽可能大,即滑框的尺寸等于 $\min(W,H)$ ,利用全部的激活映射从长、宽两个方向进行累加整合成一个特征向量。在滑框与滑框之间,都有一定的重叠区域,采用相加的方式生成全局特征,可以认为对那些重叠的区域赋予了较大的权重。并且每个滑框都是正方形的,对区域进行均匀采样,使得连续区域之间的重叠尽可能接近40%。滑框的大小由特征平面的短边决定,滑框边长的表达式如下:

$$l = 2 \times \min(W, H) / (L + 1) \quad (1)$$

如图1所示,当 $L=4$ 时,即有4个不同尺度的滑框对激活映射进行操作,对激活映射的各个区域块以宽的方向进行编

号,定义为区域1,区域2,…,区域 $N$ ,将滑框内的元素直接相加。若定义 $f_{a,b}$ 为第 $a$ 个滑框尺度等级下的第 $b$ 个向量,那么单幅图像在通过滑框操作后累加所有尺度下特征向量的和,就能得到最终的全局特征向量,具体表达式如下:

$$\mathbf{F} = \sum_a \sum_b f_{a,b} \quad (2)$$

通过计算经过滑框操作的每个区域相关的特征向量,并用L2归一化对这些特征向量进行后处理。将后处理之后的区域特征向量聚合到单个图像向量中,通过将多个区域特征向量相加并最终进行L2归一化。以上步骤使维度保持较低,并使维数等于特征通道的数量。这些步骤能够提取单幅图像的卷积特征,并将图像特征用一个等于通道数的向量维度表示。最终得到的深度区域聚合特征用深度区域聚合特征(Deep Regional Integration Feature, DRIF)来表示。

## 2.2 LOMO 特征提取

大多数行人重识别数据集在许多类中只有有限个数量的样本。例如,广泛使用的VIPeR数据集<sup>[11]</sup>仅包含来自632个人的1264个样本。因此,大多数CNN方案不能很好地处理这个问题,传统特征更适合这种情况。

LOMO主要着重解决光照和视角问题。在特征提取之前先采用Retinex算法进行图像增强。Retinex算法是一种常见的图像增强算法,它可以在动态范围压缩、边缘增强和颜色恒常性三个方面达到平衡,因此可以对各种不同类型的图像进行自适应的增强。在图像增强之后采用HSV直方图来提取图像的颜色特征,SILTP直方图则用来提取光照尺度不变的纹理特征,使用滑动窗口来描述行人图像的局部信息。具体来说,使用大小为 $10 \times 10$ 、步长为5的窗口来定位大小为 $128 \times 48$ 图像中的局部块。

在每个子窗口中,分别提取两个尺度的SILTP直方图和一个HSV直方图。为了进一步考虑多尺度信息,构建一个三尺度金字塔,它通过两次 $2 \times 2$ 局部平均混合操作对原图像进行下采样,进而得到LOMO特征向量。LOMO通过分析局部特征的水平发生并使事件最大化,以获得对视点变化稳健的表示。LOMO在许多行人重识别任务中取得了最先进的性能。因此,本文以一种有效的方式利用LOMO传统特征和提出的区域聚合特征向量之间的协作来度量距离,用参数加权来评估两个特征的相对重要性。

## 2.3 XQDA 距离学习

XQDA是在保持直接简单原则的度量和贝叶斯人脸方法基础上提出的。该方法用高斯模型分别拟合类内和类间样本特征的差值分布,再根据两个高斯分布的对数似然比推导出马氏距离。其中类内协方差矩阵、类间协方差矩阵分别定义为:

$$\Sigma_I = \frac{1}{N_I} \sum_{y_{ij}=1} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \quad (3)$$

$$\Sigma_E = \frac{1}{N_E} \sum_{y_{ij}=0} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \quad (4)$$

其中: $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 分别代表两个样本, $y_{ij}$ 是 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 的指示变量,若 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 属于同一个行人,则 $y_{ij} = 1$ ,否则 $y_{ij} = 0$ ;而 $N_I$ 代表相



似样本对的数量,  $N_E$  代表不相似样本对的数量。

子空间  $\mathbf{W}$  通过学习优化广义瑞利熵来得到:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \boldsymbol{\Sigma}_E \mathbf{w}}{\mathbf{w}^T \boldsymbol{\Sigma}_I \mathbf{w}} \quad (5)$$

其中具有交叉视图数据的子空间  $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_r)$ , 表示在  $r$  维子空间中去学习交叉视图相似性度量的距离函数。

来自不同摄像机下的一对行人样本数据  $(\mathbf{x}_i, \mathbf{x}_j)$  在子空间  $\mathbf{W}$  的距离函数如式(6)所示:

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W} \times ((\mathbf{W}^T \boldsymbol{\Sigma}_I \mathbf{W})^{-1} - (\mathbf{W}^T \boldsymbol{\Sigma}_E \mathbf{W})^{-1}) \times \mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j) \quad (6)$$

## 2.4 加权融合策略

由于提出的深度特征学习模型与实际问题直接相关,但来自较高层的特征可能受到行人姿势背景杂波等显著变化的污染,不能充分地定位于行人的重识别;并且深度网络依赖大量的样本标签,而传统的 LOMO 特征与样本数量无关,在克服重新识别任务中的跨视图外观变化时会更加可靠。所以整合两种特征的编码补充信息以克服它们各自的缺陷是有效的。

具体而言,从深度区域聚合特征和 LOMO 特征这两个角度考虑,XQDA 从这两个特征分别学习测试库和查询库图像之间的距离。基于 LOMO、本文提出的 DRIF 两个特征,采用式(6)定义的距离函数可分别获取每个特征优化的距离度量,如式(7)所示:

$$d_k(\mathbf{x}_{ik}, \mathbf{x}_{jk}) = (\mathbf{x}_{ik} - \mathbf{x}_{jk})^T \mathbf{W}_k \times ((\mathbf{W}_k^T \boldsymbol{\Sigma}_I \mathbf{W}_k)^{-1} - (\mathbf{W}_k^T \boldsymbol{\Sigma}_E \mathbf{W}_k)^{-1}) \times \mathbf{W}_k^T (\mathbf{x}_{ik} - \mathbf{x}_{jk}) \quad (7)$$

式中: $k$  分别代表 LOMO 和 DRIF 两个不变特征。

为了更好地表达传统和深度学习功能之间的协作,最终用于排序的距离可以通过以下加权平均方案将深度特征得到的距离与传统特征得到的距离融合:

$$d = \alpha d_{LOMO} + (1 - \alpha) d_{DRIF} \quad (8)$$

其中参数  $0 \leq \alpha \leq 1$  用来调整区域聚合深度特征和传统特征的相对重要性。

## 3 实验结果与分析

### 3.1 数据集和评估协议

本文使用三个重识别数据集进行实验,包括:Market-1501<sup>[12]</sup>、CUHK03<sup>[13]</sup> 和 VIPeR,它们的具体信息如表 1 所示。其中:#ID 表示数据集中含有的行人身份;#image 表示数据集中含有的行人的图像的数量;#camera 表示该数据集使用的相机的数量;evaluation 表示实验中对数据集所用的评估方法。所有行人图像的大小调整为  $224 \times 224$ ,用调整后的行人图像来微调网络,提取卷积特征。

表 1 实验数据集的具体信息

Tab. 1 Details of datasets used in experiments

数据集	#ID	#image	#camera	evaluation
Market-1501	1501	32 668	6	CMC/mAP
CUHK03	1467	13 164	2	CMC/mAP
VIPeR	632	1 264	2	CMC

Market-1501 是目前最大的基于图像的行人基准数据集。它包含 32 668 个标记的边界框,其中包含从不同视点捕获的 1501 个身份,每个行人身份的图像最多由 6 台摄像机拍摄。根据数据集设置,将数据集分为两部分:训练集有 751 人,包含 12 936 张图像;测试集有 750 人,包含 19 732 张图像。实验时,从测试集中人为选取 750 人,共有 3 368 幅图像作为查询

集合,对于查询集中给定的行人样本,需要在测试集中找出和该样本一样的行人,最后根据相似度排名给出识别结果。

CUHK03 包含 1 467 位行人的 13 164 幅图像。每个行人都是由 CUHK 校园的两台摄像机拍摄的,每位行人在一个摄像头下平均有 4.8 张图像。该数据库提供了 labeled 和 detected 两个数据集,本文对这两个数据集分别进行了实验。

VIPeR 是行人重识别中使用最广泛的数据集,包含 632 个人,每个行人包含不同视点中的两幅图像,这使得难以从两个不同的视点中匹配同一个人。此外,诸如拍摄地点、照明条件和图像质量等其他变化也使得匹配相同行人更加困难。因此,VIPeR 数据集具有挑战性。

本文将行人重识别作为图像检索问题来处理,使用 Rank1 即第一次命中的匹配准确率和平均准确率 (mean Average Precision, mAP) 两个评估指标来评估 Market-1501 和 CUHK03 数据集上重识别方法的性能,而使用累计匹配曲线 (Cumulative Matching Characteristic, CMC) 来评估数据集 VIPeR 上重识别方法的性能。

### 3.2 在 Market-1501 上的实验

本文首先在最大的基于图像的重识别数据集上评估提出的算法模型。在此数据集中,在微调的 ResNet-50 上提取最后一层卷积特征,并对产生的卷积映射进行滑框操作,产生多个局部特征向量,经过 L2 归一化处理,并直接相加操作后,得到低维的深度特征,其向量维数等于卷积层通道数,故得到的新的深度特征向量维数为 2 048 维。接着使用 LOMO 特征和本文提出的区域整合特征向量通过参数  $\alpha$  的加权融合完成二者的协作。在此数据集上,本文设置滑框尺度  $L = 4$ ,加权参数  $\alpha = 0.5$ ,用参数  $\alpha$  加权来评估传统 LOMO 特征和区域聚合深度特征的相对重要性。与本文提出的算法模型 Fusion Model 进行比较的算法包括:对称驱动的局部特征积累 (Symmetry-Driven Accumulation of Local Features, SDALF)<sup>[6]</sup>, 词袋模型 (Bag-of-Words model, BOW)<sup>[12]</sup>, LOMO<sup>[1]</sup>, CAN<sup>[14]</sup>, ID 判别嵌入 (ID-discriminative Embedding, IDE)<sup>[15]</sup> (其中 IDE(C) 表示所用模型为 Caffe, IDE(R) 表示所用模型为 Resnet-50), 姿态不变嵌入 (Pose Invariant Embedding, PIE)<sup>[16]</sup> 和 Spindle Net<sup>[17]</sup>。表 2 的结果显示,本文的 Fusion Model 与 PIE (Res50) 相比 Rank 1 性能要高 1.52 个百分点。本文提出了深度区域聚合特征向量 (DRIF),在此特征向量的基础上提出了距离融合模型,所以将本文的 Fusion Model 与 DRIF 相比较,Rank1 的值提高了 3.77 个百分点,mAP 值提高了 5.46 个百分点,说明本文提出的算法模型是有效的。

表 2 Market-1501 数据集上不同算法的实验结果对比 %

Tab. 2 Experimental result comparison of different algorithms on Market-1501 dataset %

方法	Rank1	mAP
SDALF <sup>[6]</sup>	20.53	8.20
BOW <sup>[12]</sup>	34.40	14.09
LOMO + XQDA <sup>[1]</sup>	43.79	22.22
CAN <sup>[14]</sup>	48.24	24.43
IDE(R) + XQDA <sup>[15]</sup>	71.41	48.89
PIE (Res50) <sup>[16]</sup>	78.65	53.87
Spindle Net <sup>[17]</sup>	76.90	—
DRIF	76.40	56.04
Fusion Model	80.17	61.50

为了说明通过微调网络得到的新的区域整合特征向量的



鲁棒性,将 XQDA 度量应用于新的区域整合向量与另外几种已有的特征包括 BOW<sup>[12]</sup>、LOMO<sup>[1]</sup>、IDE(C)<sup>[15]</sup>、IDE(R)<sup>[15]</sup>进行比较,结果如表 3 所示,本文提出的 DRIF 特征与 IDE(R) 特征相比 Rank1 值提高了 4.99 个百分点,mAP 值提高了 7.15 个百分点。

表 3 Market-1501 数据集上不同特征的实验结果对比 %

Tab. 3 Experimental result comparison of different features on Market-1501 dataset %

方法	Rank1	mAP
BOW <sup>[12]</sup>	41.39	19.72
LOMO <sup>[1]</sup>	43.56	22.44
IDE(C) <sup>[15]</sup>	57.72	35.95
IDE(R) <sup>[15]</sup>	71.41	48.89
DRIF	76.40	56.04

图 2 是三个示例图片的查询结果,对应每一幅查询图片,右边第一行和第二行分别是使用 IDE 和 DRIF 特征得到的排名结果,框中的行人图片表示与查询图属于同一个行人。由图 2 可以看出,对于 DRIF 特征,在排名列表顶部能够得到更多正确匹配的行人,而正确匹配的行人图片在 IDE 的排名列表中被遗漏,进一步说明本文提出的 DRIF 特征是具有判别性的。



图 2 在 Market-1501 上三幅查询图的实验效果图

Fig. 2 Experimental results of three probes on Market-1501 dataset

### 3.3 在 CUHK03 上的实验

在 CUHK03 上数据集上,本文使用类似于 Market-1501 的新训练/测试协议重新评估性能。新协议将数据集分为训练集和测试集,分别由 767 个行人和 700 个行人组成。在测试中,从每个摄像机中随机选择一个图像作为每个图像的查询,并使用其余图像构建测试。新协议将数据集均匀地划分为训练集和测试集,有利于避免重复训练和测试。对于 CUHK03 数据集的新的分类情况如表 4 所示。

本文设置滑框尺度  $L = 4$ 、加权参数  $\alpha = 0.5$ ,表 5 的结果表明,本文提出的 Fusion Model 比 LOMO 特征<sup>[1]</sup>、IDE(C)<sup>[15]</sup>

和 IDE(R)<sup>[15]</sup>得到的性能要好。本文提出的 Fusion Model 与 DRIF 相比:在训练集“Labeled”上的 Rank1 值要高 2.3 个百分点,mAP 值要高 2.8 个百分点;在测试集“Detected”上的 Rank1 值要高 2.5 个百分点,mAP 值要高 2.0 个百分点。这说明本文提出的加权融合模型(Fusion Model)是有效的。

表 4 对 CUHK03 数据集的新的分类

Tab. 4 New classification of CUHK03 dataset

数据集分类	训练集 Labeled		测试集 Detected	
	Train	Gallery	Query	1400
Train	7368	7365		
Gallery	5328	5332		
Query			1400	1400

表 5 CUHK03 数据集新分类规则下不同算法的实验结果对比 %

Tab. 5 Experimental result comparison of different algorithms on CHUK03 dataset with new classification rules %

方法	训练集 Labeled		测试集 Detected	
	Rank1	mAP	Rank1	mAP
LOMO + XQDA <sup>[1]</sup>	14.8	13.6	12.8	11.5
IDE(C) <sup>[15]</sup>	21.9	20.0	21.1	19.0
IDE(R) <sup>[15]</sup>	32.0	29.6	31.1	28.2
DRIF	31.2	28.4	30.3	28.5
Fusion Model	33.5	31.2	32.8	30.5

### 3.4 在 VIPeR 上的实验

VIPeR 数据集包含的行人图像样本数量少,因此无法用该数据集的图像作为标签来微调网络,所以本文采用的仍然是使用 Market-1501 数据集的图像微调过的 Resnet-50 模型。实验结果表明,在提取同一层卷积特征,并都对卷积特征进行区域特征提取的条件下,使用微调过后的模型比未改进的模型效果要好,说明使用行人重识别数据集微调网络使模型对于判别不同身份的行人是非常有效的。

本文采用了广泛使用的 CMC 方法对性能进行定量评估。对于 VIPeR,随机选取大约一半人(316 人)进行训练,其余人员用于测试。使用单次评估方法,并将 VIPeR 测试集划分为参考集和测试集。在 VIPeR 数据集上,设置滑框尺度  $L = 4$ ,加权参数  $\alpha = 0.8$ ,对比算法包括深度多视图特征(Deep Multi-View Feature, DMVFL)<sup>[9]</sup>、Deep Feature Learning<sup>[3]</sup>、LOMO 特征<sup>[1]</sup>、CNN<sup>[18]</sup>。当  $\alpha = 0.8$  时,Rank1 的值为 75.32%。排名前 1、5、10、20(即 Rank1、Rank5、Rank10 和 Rank20)的结果见表 6。由表 6 可知,本文提出的加权融合策略在此数据集上得到的性能最好;而且与另外两大数据集 Market-1501 和 CHUK03 相比,该算法模型在 VIPeR 数据集上得到的效果是最显著的。

表 6 VIPeR 数据集上不同算法的实验结果对比 ( $P = 316$ ) %

Tab. 6 Experimental result comparison of different algorithms on VIPeR dataset %

方法	Rank1	Rank5	Rank10	Rank20
DMVFL <sup>[9]</sup>	46.41	74.92	86.14	95.03
Deep Feature Learning <sup>[3]</sup>	40.50	60.80	70.40	84.40
LOMO + XQDA <sup>[1]</sup>	40.00	67.40	80.51	91.08
CNN <sup>[18]</sup>	22.50	49.50	63.20	79.80
DRIF	32.71	55.63	74.45	86.78
Fusion Model	75.32	94.62	98.42	99.37



为了进一步说明文中所提融合模型在 VIPeR 数据集上的有效性,在使用同一度量方法 XQDA 的前提下,使用融合模型、LOMO 特征、Resnet-50 模型和微调的 Resnet-50 模型得到的性能如表 7 所示。表 7 中的结果表明本文所提算法模型效果显著,由于 LOMO 特征强调 HSV 和 SILTP 直方图,因此它在特定照明条件下表现效果更佳。视角和光照的多样性是 VIPeR 数据集的特点,表 7 中的结果表明本文提出的加权融合模型在背景、照明和视点等方面有大幅变化的数据集上效果最明显,能够显著提高行人重识别的性能。本文提出的融合模型得到的性能优于 LOMO 特征,与单独使用 LOMO 特征以及使用 F-Resnet-50 模型提取卷积特征并对特征进行滑框操作得到的区域聚合特征进行距离度量相比,融合 LOMO 特征和深度区域聚合特征(DRIF)这两个特征距离能够得到较高的识别率,说明这两个特征距离的融合具有强烈的判别能力,并进一步表明本文提出的区域聚合特征是 LOMO 特征的互补特征。

表 7 VIPeR 数据集上不同模型的实验结果对比 ( $P = 316$ ) %Tab. 7 Experimental result comparison of different models on VIPeR dataset ( $P = 316$ ) %

方法	Rank1	Rank5	Rank10	Rank20
LOMO + XQDA <sup>[1]</sup>	40.00	67.40	80.51	91.08
F-Resnet-50	42.70	69.78	81.14	91.99
Fusion Model	75.32	94.62	98.42	99.37

### 3.5 微调策略分析

使用 MatConvNet (Convolutional neural Networks for Matlab) 工具,利用 ImageNet 模型训练 Market-1501 数据集。对于网络 Resnet-50 使用默认参数设置,并从 ImageNet 预先训练好的模型中进行微调。图像在被送入网络之前被调整  $224 \times 224$  大小;初始学习率设置为 0.001,并在每次迭代后减少至上一次的 1/10,训练迭代 36 次之后完成。为了证明微调策略的有效性,在 VIPeR 数据集上用微调的模型进行实验,分别用 Resnet-50 网络以及微调过后的 Resnet-50 网络(Fine-tuning Resnet-50, F-Resnet-50)提取卷积层特征,并对两个模型提取的同一层三维卷积特征利用区域聚合特征方法变成 2048 维的特征向量,并进行距离度量。表 8 中的结果表明,利用行人重识别数据集微调过后的网络模型提高了区分能力,减少了错误检测对背景的影响,并提高了识别率。

表 8 微调 Resnet-50 模型对重识别性能的影响 ( $P = 316$ ) %Tab. 8 Impact of fine-tuning of Resnet-50 model on re-identification performance ( $P = 316$ ) %

算法模型	Rank1	Rank5	Rank10	Rank20
Resnet-50	29.30	59.68	74.46	86.93
F-Resnet-50	42.70	69.78	81.14	91.99

### 3.6 参数分析

如图 1 区域聚合部分所示,在滑框与滑框之间,都存在一定的重叠区域,而最终采用简单的加和方式把局部的区域特征向量整合成全局特征,其中那些重叠的区域可以认为是给予了较大的权重。因此,并不是将特征平面分得越细越好。在本文中,滑框之间的重叠率取 40%。在实验中,使用  $L$  种

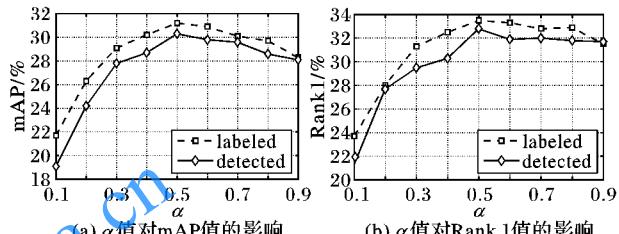
不同尺度的滑框处理特征平面在数据集 CUHK03 上进行实验,结果如表 9 所示,可以看出,当  $L = 4$  时,将提出的新的区域聚合特征向量用于度量时的效果最好。

为了克服传统特征和深度特征各自的缺陷,用参数  $\alpha$  加权评估深度区域特征向量和传统 LOMO 特征的相对重要性。其中  $0 \leq \alpha \leq 1$ ,由图 3 可知,当  $\alpha = 0.5$  时,在 Labeled 和 Detected 两个数据集上的 Rank1 和 mAP 值都最高,即在 CUHK03 数据集得到的性能最好。

表 9 采用不同尺度滑框处理的性能

Tab. 9 Performance with different scales of sliding windows

L	训练集 Labeled		测试集 Detected	
	Rank1/%	mAP/%	Rank1/%	mAP/%
2	27.7	25.7	21.1	19.0
3	27.5	26.0	28.8	26.3
4	31.2	28.4	30.3	28.5
5	29.5	26.0	28.1	26.3

图 3 CHUK03 数据集上不同的  $\alpha$  值对 mAP 和 Rank1 值的影响Fig. 3 Impact of parameter  $\alpha$  on mAP and Rank1 on CHUK03 dataset

### 3.7 运行时间分析

如表 10 所示是 VIPeR 数据集上单个图像的平均特征提取时间。可以看出,本文提出的深度区域特征提取方法比一些手工特征提取方法更快,例如基于生物启发特征的协方差描述符(Covariance descriptor based on Bio-inspired features, gBiCov)<sup>[19]</sup>;与 LOMO 手工特征、CNN 特征、FFN 特征相比,本文提出的 DRIF 特征的维度是 2048 维,具有更低的维度,并且其维度等于卷积层通道数。通过在速度和维度复杂性之间取得平衡,本文提出的区域特征向量提取算法可以实际应用。

表 10 单幅图像提取特征的平均时间

Tab. 10 Average time of extracting features of a single image

方法	提取时间/s	特征维度
gBiCov <sup>[19]</sup>	13.62	5940
LOMO <sup>[1]</sup>	0.27	26960
CNN <sup>[18]</sup>	0.18	4096
FFN <sup>[2]</sup>	0.76	4096
DRIF	0.42	2048

## 4 结语

本文构建了一个完整的行人重识别的算法模型,通过微调的 Resnet-50 网络提取三维卷积特征,并把不同尺度的滑框作用于卷积激活映射,得到了低维的区域聚合特征向量;从深度区域聚合特征和传统手工特征 LOMO 两个角度出发,用参数加权来评估各自的相对重要性,并利用有效的加权融合方式得到最终用于计算的距离,用最终距离进行识别排序。在



Market-1501、CHUK03 和 VIPeR 三个数据集上进行测试,在重新训练网络的情况下,大量实验表明本文提出的算法模型在指标 Rank1 和 mAP 上均具有较明显的提升,展示了所提算法模型的有效性。下一步的研究方向是提取出更鲁棒性的特征,使其能够更具有判别性,使多视图特征融合方法能够显著提高行人重识别的性能。

#### 参考文献 (References)

- [1] LIAO S, HU Y, ZHU X, et al. Person re-identification by local maximal occurrence representation and metric learning [ C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 2197 – 2206.
- [2] WU S, CHEN Y-C, LI X, et al. An enhanced deep feature representation for person re-identification [ C]// Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision. Washington, DC: IEEE Computer Society, 2016: 1 – 8.
- [3] DING S, LIN L, WANG G, et al. Deep feature learning with relative distance comparison for person re-identification [ J]. Pattern Recognition, 2015, 48(10): 2993 – 3003.
- [4] 丁宗元,王洪元,陈付华,等.基于距离中心化与投影向量学习的行人重识别[J].计算机研究与发展,2017,54(8):1785 – 1794.  
(WANG Z Y, WANG H Y, CHEN F H, et al. Person re-identification based on distance centralization and projection vectors learning [ J]. Journal of Computer Research and Development, 2017, 54 (8): 1785 – 1794.)
- [5] KÖSTINGER M, HIRZER M, WOHLHART P, et al. Large scale metric learning from equivalence constraints [ C]// Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2012, 1: 2288 – 2295.
- [6] FARENZENA M, BAZZANI L, PERINA A, et al. Person re-identification by symmetry-driven accumulation of local features [ C]// Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2010: 2360 – 2367.
- [7] 张耿宁,王家宝,李阳,等.基于特征融合与核局部 Fisher 判别分析的行人重识别[J].计算机应用,2016,36(9):2597 – 2600.  
(ZHANG G N, WANG J B, LI Y, et al. Person re-identification based on feature fusion and kernel local Fisher discriminant analysis [ J]. Journal of Computer Applications, 2016, 36 (9): 2597 – 2600.)
- [8] 孙金玉,王洪元,张继,等.基于块稀疏表示的行人重识别方法 [ J].计算机应用,2018,38(2):448 – 453. (SUN J Y, WANG H Y, ZHANG J, et al. Person re-identification method based on block sparse representation [ J]. Journal of Computer Applications, 2018, 38(2): 448 – 453.)
- [9] TAO D, GUO Y, YU B, et al. Deep multi-view feature learning for person re-identification [ J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(10): 2657 – 2666.
- [10] GONG Y, WANG L, GUO R, et al. Multi-scale orderless pooling of deep convolutional activation features [ C]// Proceedings of the 2014 European Conference on Computer Vision, LNCS 8695. Cham: Springer, 2014: 392 – 407.
- [11] GRAY D, BRENNAN S, TAO H. Evaluating appearance models for recognition, reacquisition, and tracking [ C]// Proceedings of the 2007 IEEE International Workshop on Performance Evaluation for Tracking and Surveillance. Piscataway, NJ: IEEE, 2007: 41 – 49.
- [12] ZHENG L, SHEN L, TIAN L, et al. Scalable person re-identification: a benchmark [ C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Washington, DC: IEEE Computer Society, 2015: 1116 – 1124.
- [13] LI W, ZHAO R, XIAO T, et al. DeepReID: deep filter pairing neural network for person re-identification [ C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2014: 152 – 159.
- [14] LIU H, FENG J, QI M, et al. End-to-end comparative attention networks for person re-identification [ J]. IEEE Transactions on Image Processing, 2017, 26(7): 3492 – 3506.
- [15] ZHENG L, ZHANG H, SUN S, et al. Person re-identification in the wild [ C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2017: 3346 – 3355.
- [16] ZHENG L, HUANG Y, LU H, et al. Pose invariant embedding for deep person re-identification [ J]. arXiv E-print, 2018: arXiv: 1701.07732.
- [17] ZHAO H, TIAN M, SUN S, et al. Spindle net: person re-identification with human body region guided feature decomposition and fusion [ C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2017: 907 – 915.
- [18] PAISITKRIANGKRAI S, SHEN C, van den HENGEL A. Learning to rank in person re-identification with metric ensembles [ C]// Proceedings of the 2015 Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 1846 – 1855.
- [19] MA B, SU Y, JURIE F. Covariance descriptor based on bio-inspired features for person re-identification and face verification [ J]. Image and Vision Computing, 2014, 32(6/7): 379 – 390

This work is partially supported by the National Natural Science Foundation of China (61671376, 61771386), the Scientific Research Project of Shaanxi Provincial Department of Education (18JK0556).

**DENG Xuan**, born in 1995, M. S. candidate. Her research interests include deep learning, image processing.

**LIAO Kaiyang**, born in 1976, Ph. D., lecturer. His research interests include machine vision, artificial intelligence.

**ZHENG Yuanlin**, born in 1975, Ph. D., associate professor. His research interests include color management, evaluation of quality of color image, color science.

**YUAN Hui**, born in 1993, M. S. candidate. His research interests include deep learning, image processing.

**LEI Hao**, born in 1995, M. S. candidate. His research interests include deep learning, image processing.

**CHEN Bing**, born in 1997. His research interests include image processing.