基于增强特征融合网络的行人重识别方法

刘玉杰¹⁾, 周彩云¹⁾, 李宗民¹⁾, 李华^{2,3)}

¹⁾ (中国石油大学(华东)计算机科学与技术学院 青岛 266580)

²⁾ (中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

³⁾ (中国科学院大学 北京 100049)

(liuyujie@upc.edu.cn)

摘 要: 针对行人重识别技术受遮挡、背景冗余、光照、姿态以及检测误差等问题的影响, 鲁棒的行人特征表达对正确检索行人越来越重要. 为了利用对齐特征和度量学习的优势, 进一步分析局部空间语义特征. 首先, 在特征层面: 一是在 ResNet50 框架中嵌入空间变换结构, 自适应对齐局部区域空间特征, 解决因局部区域不对齐导致的空间语义不一致的问题; 二是通过对齐的局部特征设计一种增强特征融合网络, 充分利用语义信息间的关联性提取图像的细节特征. 然后, 在损失函数层面: 提出一种排序矩阵方法选取区域样本对, 设计了一种局部三元组损失计算方法, 联合正则化分类损失共同训练网络, 充分利用融合的增强特征, 达到高效度量的效果. 最后, 文中方法结合现有的重排算法进一步提高了 Rank-1 与 mAP 检索精度, 在行人重识别基准数据集 Market-1501 上的实验结果, 证明了本文方法的有效性.

关键词: 空间语义特征; 增强特征融合网络; 排序矩阵; 局部三元组损失

中图法分类号: TP391.41

DOI: 10.3724/SP.J.1089.2021.18303

Strong Feature Fusion Networks for Person Re-Identification

Liu Yujie¹⁾, Zhou Caiyun¹⁾, Li Zongmin¹⁾, and Li Hua^{2,3)}

¹⁾ (College of Computer Science and Technology, China University of Petroleum, Qingdao 266580)

²⁾ (Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

³⁾ (University of Chinese Academy of Sciences, Beijing 100049)

Abstract: With the development of deep learning, the performance of person Re-Identification (Re-ID) has been significantly improved. It's still a challenging task due to the challenges coming from large variations on persons such as occlusion, background clutter, pose, illumination and detection failure, etc. To retrieve true pedestrians, robust feature expression is significant. Instead of using external cues, this paper takes advantage of robust alignment features and metric learning. First, from the aspect of feature extraction, there were three contributions. (i) Embedded a spatial transformer network in the network architecture, which is called ResNet_STN in this paper, which can solve the problem of local spatial semantic feature inconsistency, accurately express the main characteristics of the target, and achieve pedestrian alignment. (ii) Designed a strong feature fusion network based on the aligned local features, which is named a Strong Feature Fusion Module (SFFM) and can make full use of the connection between semantic information to extract detailed features of images. Then, from the aspect of metric loss function, one contribution was put forward. (iii) Proposed a Ranking Matrix (RM) method to select local

收稿日期: 2020-04-07; 修回日期: 2020-06-29. 基金项目: 国家自然科学基金(61379106, 61379082, 61227802); 山东省自然科学基金(ZR2013FM036, ZR2015FM011). 刘玉杰(1971—), 男, 博士, 副教授, CCF 会员, 主要研究方向为计算机图形图像处理、多媒体数据分析、多媒体数据库; 周彩云(1995—), 女, 硕士研究生, 主要研究方向为行人重识别、图像检索; 李宗民(1965—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为计算机图形学、图像处理、模式识别、科学计算可视化; 李华(1957—), 男, 博士, 研究员, 博士生导师, CCF 会员, 主要研究方向为计算机图形图像处理.

triplet samples and compute local triplet loss. We combined a regularized classification loss to train the network to unleash the discrimination ability of the learned strong representations of this network architecture. Finally, the proposed method with the existing re-ranking algorithm to further improves Rank-1 and mAP retrieval accuracy. Experimental results on Market-1501 dataset demonstrate the effectiveness of our proposed method.

Key words: spatial semantic feature; strong feature fusion networks; ranking matrix; local triplet loss

随着互联网技术的不断进步,人脸识别^[1]、目标追踪^[2]以及行人再识别^[3]等行人智能认知技术日益发展,并逐渐用于商业应用.行人重识别(person re-identification)任务旨在给定1幅行人的查询图像,在大规模的行人数据中检索出不同摄像头下拍摄的与其最相似的前 k 幅图像,被视为一种跨摄像头的图像检索^[4]任务.由于训练数据与测试数据的行人身份不重合,所以也看做一种零样本学习^[1]问题.该技术可以与行人检测^[5]、目标跟踪^[2]等技术结合应用于智能视频监控、智能安保、智能交通等领域.在现实场景中,如刑侦破案以及特定场景下的识人、寻人问题上具有广泛的应用前景.

随着深度学习的不断发展,在计算机视觉领域更深层的网络^[6]被应用到行人重识别任务中,优化了识别效果.但随着更深层网络的提出,底层细节信息随着卷积层的加深往往会丢失.同时,受背景冗余、遮挡、光照、摄像头拍摄角度、运动模糊和检测误差等问题的影响,导致行人图像间的差异很大,不易区分.由此,提取鲁棒的行人特征表示^[7-9]以及应用高效的相似性度量方法^[10-13]仍是重中之重.

(1) 特征层面.目前,在有监督行人重识别任务中,为了提取更具判别力的行人特征,全局特征通常会联合局部特征共同表征行人的整体信息,但由于背景冗余、遮挡和摄像头拍摄角度不同,会带来极大的局部区域偏差,导致局部语义空间不一致的问题.本文在不引入额外的姿态估计模型以及关键点信息的前提下,通过嵌入空间变换结构解决行人图像间区域不对齐的问题.为进一步表达图像局部区域之间的关联性并提取目标区域的显著性特征,本文利用对齐的相邻区域设计了增强特征融合模块(strong feature fusion module, SFFM),在不引入其他模型的情况下粗定位行人的关键区域,增强网络的判别力.

(2) 损失层面.行人重识别作为一种分类任务,常用交叉熵计算分类损失;同时作为一种检索任务,常用对比损失^[11]、三元组损失^[10,12]以及四元组损失^[13]进行样本间的相似性度量.为充分地利用融合网络提取的对齐增强特征,本文采用分类损失和

三元组损失训练网络,提出一种排序矩阵(ranking matrix, RM),根据特有的监督信息选取局部三元组样本,设计了一种局部三元组损失计算方法.同时考虑行人重识别作为一类特殊的检索问题,即训练数据与测试数据中行人身份不重合,为避免训练过程中过拟合现象的发生,采用一种正则化的分类交叉熵损失,进一步解决类内差异大、类间差异小的问题.

本文根据网络框架输出的对齐特征,设计3层级SFFM方法,以增强特征的鲁棒性和表达能力;同时提出RM方法选取三元组区域样本对,并设计一种局部三元组损失计算方法准确度量样本间的相似性,以达到正确聚类的效果;整合新的网络框架ResNet_STN,在经典的行人重识别数据集Market-1501上,将现有的重排算法引入到本文方法中,优化检索结果并进一步提高了检索精度.

1 相关工作

1.1 行人重识别

行人重识别作为一类特殊的检索问题,一般是通过对数据图像进行特征提取,获得其特征表示,然后投影到特定的度量空间中进行相似性度量,并依据相似性距离进行排序,最终得到检索结果.目前,基于深度学习的行人重识别^[14]发展迅速,并逐步向限制场景下的应用阶段迈进.近3年在计算机视觉会议CVPR, ICCV, ECCV发表的行人重识别方向的论文,多数依旧是在特征表达以及相似性度量方面做工作,但也有少量工作开始转向无监督跨域迁移,利用生成对抗网络解决数据少、遮挡、视角变换等严峻问题,有效提高了行人重识别的检索精度.但是,无论是在有监督的行人重识别还是解决无监督数据域分布问题上,严峻的挑战依旧存在,如行人换衣、姿态各异以及域风格适应问题仍然没有得到很好的解决.因此,如何设计网络结构提取更加鲁棒的行人特征,以及寻找最优的度量学习方法匹配不同行人间的相似程度依旧是行人重识别中研究的重点.

1.2 增强特征融合

为提取行人的显著性特征以及图像中具有判别力的属性特征(衣服 LOGO、配饰、背包等),大部分工作开始融合全局特征与局部特征,但由于数据集中的行人图像存在背景冗余以及遮挡的情况,所以局部特征的引入会带来特定区域的不对齐问题,导致特征表达上的空间语义不一致.语义一般针对图像数据而言,一致是要求提取的特征具有一定的联系性.在行人重识别任务中,由于拍摄角度、场景变化、运动模糊等带来的不同图像中同一行人的部位不对齐会导致不同图像在空间上的语义特征表达不一致,进而影响行人重识别的性能.由此,Wei 等^[15]通过引入额外的姿态估计模型进行语义空间对齐,虽然融合对齐后的局部特征提高了精度,但额外模型的引入也带来了新的模型误差以及计算量.Li 等^[16]采用注意力机制的方法定位行人的关键区域,增强目标的特征表达,提取显著性特征来实现图像局部空间一致性,但忽略了有用的边缘细节信息,缺少语义信息间的关联性,且这类方法一般需要额外的标注信息.2017 年,Zhang 等^[17]直接采用分块的方法学习局部区域特征并提出利用最短路径解决局部空间语义不一致的问题.同年,在 Sun 等^[7]提出的基于局部区域的卷积基线(part-based convolutional baseline, PCB)网络中,虽然利用简单的分块方法 PCB 取得了不错的检索效果,但其前提条件是假设数据集中的行人图像对齐,这忽略了行人图像中存在的背景冗余以及遮挡等问题带来的影响;提出的局部池化优化(refined part pooling, RPP)方法需要进一步微调,且不能端到端训练网络.

针对以上问题,本文在 ResNet50 框架中嵌入空间变换结构实现对齐,主要针对输出的对齐特征设计了一种增强特征融合模块.该模块分为 3 层级特征:第 1 级特征提取行人的整体全局特征;第 2 级特征提取行人空间分块区域的局部特征;第 3 级特征利用对齐局部相邻区域的融合提取行人图像中的关联性特征,充分利用了语义信息间的关联性达到弱注意力机制的效果,既不需要大量的标注信息,也不需要引入额外的姿态关键点模型.

1.3 相似性度量

相似性度量旨在采用一种抽象的距离方式去刻画图像之间的相似程度.在行人重识别任务中,主要是将图像特征编码到欧几里得空间中,根据样本间的欧几里得度量优化损失函数,通过将函

数最小化来控制正样本对 $d_{a,p}$ 以及负样本对 $d_{a,n}$ 之间的距离,最终达到拉近正样本对、推开负样本对的效果,即 $d_{a,p} < d_{a,n}$ .2014 年,Yi 等^[18]在行人重识别问题上首次提出用一个统一的框架联合特征提取与度量学习进行端到端学习.Schroff 等^[10]首次将三元组损失引入到人脸识别领域并取得了突破性进展.为提高网络的泛化能力以及收敛速度,Hermans 等^[12]选取困难样本三元组来更好地表达类内、类间的结构分布,从而加快网络的训练速度并提升相似样本的区分能力.目前,度量学习在计算机视觉任务中非常有效,在目标检索^[4]、零样本学习^[1]和人脸验证^[10]等方向均有广泛应用.

本文考虑三元组损失带来的过拟合、泛化能力差的问题,采用困难样本三元组损失^[12]作为相似性度量优化函数.在充分拟合的基础上,加入局部融合特征并计算其之间的相似性距离,最终将设计的局部三元组损失联合类别约束进一步提升了网络对行人图像的区分能力.

1.4 重排算法

重排算法(re-ranking, RK)旨在再次对检索排序结果进行重新排序.本文采用基于 k 互邻居方式的重排算法^[19],在局部查询扩展方法基础上,计算原始距离与杰卡德距离的加权集合作为最终的检索距离排序列表,有效地降低了图像错误匹配的情况,具体算法详见文献[19].

2 本文方法

针对行人重识别面临的挑战,本文进一步探索局部语义信息之间的关联性,提出一种基于增强特征融合网络的行人重识别方法.搭建网络框架并在卷积层中引入空间变换结构,以提取语义一致的局部特征;设计新型融合模块,以利用邻近区域的关联性信息提取到更强判别力的特征;引入局部三元组损失计算方法,并采用矩阵存储的形式,以保留邻近区域的监督信息;最终对 3 层级特征分别计算三元组损失和分类损失来约束样本间的相似程度,进而达到正确检索行人的目的.整体框架如图 1 所示,主要由 2 部分构成:嵌入空间变换结构的残差网络框架 ResNet_STN 和增强特征融合模块 SFEM,后者是本文的主要创新.使用在 ImageNet 上预训练的 ResNet50 模型,包括 5 个 ResBlock 模块和全局平均池化层(global average pooling, GAP),每个 ResBlock 模块包括卷积层、归一化层、激活层(激活

函数采用 ReLU)以及最大池化层,通过移除最后的全连接层并根据行人类别数目增加对应映射的全连接层来预测行人类别概率.在框架 ResNet_STN 中,利用输出的 ResBlock₄ 的特征图作为空间变换结构

的输入,输出的 M 个对齐局部特征在 SFFM 中得到进一步的分析和利用.在测试阶段将融合后的特征作为最终的行人检索特征,进行相似性度量获得检索结果.

图1 本文方法框架图

2.1 网络模块

(1) ResNet_STN. 选择在动作识别、图像检索等计算机视觉任务中取得显著成果的 ResNet50 网络结构作为本文模型的基本框架.为解决引入局部特征带来的问题以及更好地利用本文提出的增强特征融合模块,在该网络中嵌入可微分的空间变换结构^[20].如图1所示的灰色虚线框,在 ResNet_STN 中引入了空间变换结构.空间变换结构作为一种轻量级的通用模块,可以无缝集成到任何网络架构中,使神经网络能够主动变换特征图,实现对网络内的数据仿射变换的操作,该结构不需要额外关键点的标定,能够根据分类任务自适应地将数据进行空间变换和对齐.空间变换结构主要包含定位网络(localization net, LN)、网格变换以及线性采样3个组成部分,将 ResNet50 输出的 ResBlock₄ 的特征图作为输入,利用 LN 得到的 M 个空间位置参数,根据空间变换计算得到对应的仿射变换参数 $\theta=[\theta_1, \theta_2, \dots, \theta_M]$,最后利用双线性插值方法实现局部区域位置的对准.在本文实验中, LN 由1个卷积层和2个全连接层构成,其他部分设置参考文献[20],最终提取到局部空间语义一致的对齐特征,并在核心方法 SFFM 中得到利用.

(2) SFFM. 如图2所示,对于数据集中的多数样本而言,同类目标虽然在外观上呈现多样性,但具有相同的深层语义特征;不同类目标之间,即使外观有一定的相似性,其语义却不相关,因而利用分类模型便能很好地将行人 A、B、C 进行归类识别.但对于数据集中极为相似的样本,如图2中行人 C 的分类样本,红色方框中的图像与行人 C 外观相似、网络提取的语义特征也相似,故将其错误分类.

由此,利用全局语义特征进行分类很大程度上会受限于数据集样本情况.

为进一步提取样本间的显著性特征提高网络的区分能力,设计了图1中红色虚线框中的 SFFM 方法,如图3所示.根据 ResNet_STN 输出的 M 个对齐局部特征②,除了将其全部融合作为全局特征①之外,为表达图像局部区域之间的关联性,本文进一步分析和利用 M 个局部特征,将相邻局部区域特征再次融合作为增强特征③,最后融合特征①②③作为最终的检索行人特征.在第3层级特征中,增强特征融合后经过一个平均池化层 GAP 得到 M' 个列向量,其中, $M'=M/2$. 随后经过 ϕ 操作降低特征维度, ϕ 由1个 1×1 卷积层、归一化层和激活层(激活函数采用 ReLU)构成.最后,特征向量输入到由全连接层 FC 和 Softmax 函数构成的分类器中,从而预测行人类别概率.具体地,损失部分见第2.2节.第3.4节实验中,在全局特征融合 M 个局部特征的基础上,验证了第3层级增强特征的进一步融合切实有效地提高了检索精度.

图2 利用全局特征分类行人图像示意图

图 3 增强特征融合模块结构图

2.2 损失函数

为了充分利用增强的融合特征来度量行人图像间的相似性, 本文采用正则化的分类损失以及三元组损失联合训练网络, 并使用梯度下降算法使得损失函数不断下降来进行优化, 最后在特征空间实现类内距离足够小、类间距离足够大的效果, 从而准确分类并检索出不同摄像头下的相似行人。

2.2.1 正则化分类损失

在训练过程中, 对每个行人的图像 x 做预测分类, 计算分类损失。本文采用基于 Softmax 激活函数的交叉熵损失训练分类器, Softmax 激活函数表示为

$$P_j = \frac{e^{z_j}}{\sum_{n=1}^N e^{z_n}} \quad (1)$$

其中, N 为训练集类别总数; z 表示分类层输出得分, $z_j = W_j^T x$, $z_n = W_n^T x$ 。对于行人 j , $j \in 1, 2, 3, \dots, N$, 其类别的概率为 P_j 。交叉熵分类损失为

$$L^{\text{lsid}} = -\sum_{j=1}^N \phi \log(P_j) \quad (2)$$

本文考虑到行人重识别中训练与测试数据的行人身份不重合, 为避免过拟合现象的发生, 在原有的标签平滑(label-smoothing, LS)^[21]方法基础上采用 ϕ 函数, 即

$$\phi = \begin{cases} \frac{\delta}{N}, & y \neq j \\ 1 - \delta + \frac{\delta}{N}, & y = j \end{cases} \quad (3)$$

并在本文网络中取得很好的效果。其中, ϕ 表示目标概率; 若 y 表示的真值标签与 j 相等, 则目标概率值较为接近 1, 反之接近 0; δ 为常数, 在实验中设置 $\delta = 0.1$ 。在控制损失较小的情况下, 通过 ϕ 函

数的引入, 一是提高模型的泛化能力, 防止过拟合现象的发生; 二是通过 δ 值控制分类预测值大小, 在一定程度上缩小了不同类别之间的差距, 提高分类任务的挑战性。

如图 3 所示, 训练过程中, SFFM 方法中的全局特征以及 $(M + M')$ 个局部特征分别计算分类损失, 最终的正则化分类损失为

$$L^{\text{cla}} = L_G^{\text{lsid}} + L_m^{\text{lsid}} + L_{m'}^{\text{lsid}} = \sum_{j=1}^N -\phi \log(P_j) + \sum_{j=1}^N \sum_{m=1}^M -\phi \log(P_j^m) + \sum_{j=1}^N \sum_{m'=1}^{M'} -\phi \log(P_j^{m'}) \quad (4)$$

2.2.2 三元组损失

在三元组集合 $S = \{X_a, X_p, X_n\}$ 中, (X_a, X_p) 构成一对正例, 即表示同一行人的 2 幅图像, (X_a, X_n) 构成一对负例, 表示不同行人的 2 幅图像。相似性度量旨在优化函数达到让相似正样本对 X_a 与 X_p 之间的距离小于负样本对 X_a 与 X_n 之间距离的效果。三元组损失公式为

$$L = \frac{1}{N_{\text{trip}}} \sum_{X_a, X_p, X_n} \left(\|f(X_a) - f(X_p)\|_2^2 - \|f(X_a) - f(X_n)\|_2^2 + \gamma \right)_+ \quad (5)$$

其中, N_{trip} 表示三元组总数; $X_a, X_p, X_n \in S$; $f(\cdot)$ 为网络输出的特征; $(\cdot)_+ = \max(\cdot, 0)$ 。通过最小化 L , 达到拉近正样本、推开负样本的效果。

考虑三元组损失中简单与困难样本的选取对网络模型具有重要作用, 且不同的采样策略对全局特征与局部特征的表征分布会有不同影响。由此, 为使三元组损失函数在训练过程中更有效, 加快网络训练速度的同时提升对相似图像的区分能力, 采用困难样本三元组损失函数^[12], 即通过欧几里得度量选取相同标签下距离最远的一对为正例, 以及不同标签下最近的一对为负例, 它们作为困难三元组样本^[22]。在一个批次处理中选取 P 个行人的 K 幅图像, $N'_{\text{trip}} = P \times K$ 。由于硬件条件限制, 最大取 $N'_{\text{trip}} = 64$, 困难样本三元组损失函数计算公式为

$$L^{\text{trihard}} = -\sum_{i=1}^P \sum_{a=1}^K \left[\max_{p=1, \dots, K} \|f(X_a^i) - f(X_p^i)\|_2^2 - \min_{\substack{n=1, \dots, K \\ l=1, \dots, P \\ l \neq i}} \|f(X_a^i) - f(X_n^l)\|_2^2 + \gamma \right]_+ \quad (6)$$

在 SFFM 方法中,选用第1级特征和第3级特征做三元组损失,为了避免网络训练过程中收敛速度慢,没有用到第2级特征进一步联合损失计算.如图3所示,第1级特征计算全局三元组损失 L^g ,即式(6).图3中,对于全局特征①,对号表示利用 ϕ 操作之前的特征向量计算相似性,而非叉号表示的归一化后的向量计算.针对第3级特征,提出一种 RM 方法选取三元组区域样本对,并计算局部三元组损失 L^l ,对号表示利用第3级增强融合的特征计算 L^l ,而不是利用叉号表示的第2级特征.

(3) RM. 针对 SFFM 方法中的第3级特征,采用经过 ϕ 操作之后的归一化特征 $\|f_{x'} - f_{y'}\|_2 \in [0,1]$ 计算一个批次大小 N'_{trip} 中 M' 个局部特征之间的欧几里得度量,归一化后的特征距离 d 构成 RM, 即

$$d_{x',y'} = \frac{e^{\|f_{x'} - f_{y'}\|_2} - 1}{e^{\|f_{x'} - f_{y'}\|_2} + 1} \quad \forall x', y' = \{1, 2, \dots, M'\} \quad (7)$$

RM 表示为

$$\mathbf{M}_{RM} = \begin{bmatrix} d_{11} & \dots & d_{1M'} \\ \dots & d_{x',y'} & \dots \\ d_{M'1} & \dots & d_{M'M'} \end{bmatrix}.$$

RM 按序存储,由此会保留一定的监督信息.选出 $\mathbf{M}_{RM}$ 中正样本对与负样本对下的对应区域的最小值,即对角线区域 d_{mm} , $m \in 1, 2, \dots, M'$, 计算局部三元组损失 L^l , 其公式为

$$L^l = \frac{1}{M'} \sum_{m=1}^{M'} \left[d_{a,p}^{mm} - d_{a,n}^{mm} + \beta \right]_+ \quad (8)$$

最终三元组损失 L^{trip} 公式为

$$L^{trip} = L^g + \lambda L^l \quad (9)$$

其中, L^g 中的 γ 与 L^l 中的 β 是常量超参阈值,均用于控制三元组类内类间距离; λ 表示局部三元组损失所占比重.本文设置 $\gamma = \beta = 0.01$, $\lambda = 2$.

2.2.3 总优化损失函数

联合分类损失以及三元组损失作为整个网络的损失函数,最终的总优化损失函数为

$$L^{\text{total}} = L^{\text{cla}} + L^{trip} \quad (10)$$

3 实验结果与分析

3.1 数据集介绍

本文选择在经典的行人重识别数据集 Market-1501^[23]上进行评估.

Market-1501 是行人重识别领域的基准数据集,一共有来自 6 个不同的摄像头 1501 个行人的 32 668 幅图像.数据集被分为训练集和测试集,其中 751 个行人的 12 936 幅图像作为训练集,剩余的 750 个行人的 19 732 幅图像作为测试集,在测试集的 750 个行人中随机选择每个摄像头中的一幅图像作为查询图像,共 3 368 幅查询图像.对于每幅查询图像,旨在从测试集中检索出最相似的行人图像序列.因 Market-1501 数据集涉及不同视角、不同场景以及各类行人不同风格的图像,相较来说比较接近真实场景应用,故选择该数据集以验证本文方法的有效性.

3.2 评价标准

在 Market-1501 数据集中,采用 Rank-1 准确率和平均准确率(mean average precision, mAP)作为评价标准.

3.3 实验设置

实验基于深度学习框架 PyTorch 实现,在 Windows 10 操作系统、NVIDIA GeForce GTX TITAN X 以及 4 块 NVIDIA Tesla P100 的服务器上进行.

在数据预处理阶段,针对 Market-1501 数据集的分布特点采用了色彩抖动、随机擦除、水平翻转进行数据预处理做数据增强,初始化图像大小为 256×128 .在训练过程中,使用在 ImageNet 数据上预训练的 ResNet50 模型为底层框架,采取学习率更新策略,并将最后一个残差块的步长设置为 1,增大行人的显著性特征.实验设置局部区域 $M=6$, $M'=3$,批量大小为 64,即 16 个行人各选取 4 幅图像,训练迭代总次数为 200 次.在测试阶段,利用融合后的特征向量作为行人的整体特征,实现精准检索.

3.4 实验分析

表 1 所示为不同方法在 Market-1501 数据集上的评估结果,其中 SFFM①②表示去掉第3级局部特征融合,仅利用第1,2级局部特征的融合作为最终的检索特征;+SFFM- L^l 表示在增强特征融合模块的基础上去掉局部三元组损失的优化;+RK 表示在本文方法上加入重排算法优化检索排序的结果.可以看出,本文方法较 Res50-baseline^[6]结果有明显提升,在重排算法优化下,其在 Market-1501 数据集上 Rank-1 达到 94.22%, mAP 达到 85.03%.为进一步验证本文方法中增强特征融合模块及损失优化带来的效果,表 1 中分别比较 Res50-baseline, ResNet_STN, Res50-baseline+SFFM 以及本文方法

分别去掉第 3 级局部融合特征和局部三元组损失、本文方法 ResNet_STN+SFFM 以及本文方法结合重排后的效果. 对比发现, 第 3 级特征的进一步融合在 mAP 上提升了 7.7%, 在 Rank-1 上提升了 2.7%; 此外, 局部三元组损失的加入有效地提高了精度, 在 mAP 指标上提高了 4.39%. 同时, 相比于 Res50-baseline, ResNet_STN 的精度在 Rank-1 指标上也提高了 2.29%.

表 1 不同方法在 Market-1501 数据集上的评估结果 %

方法	Rank-1	mAP
Res50-baseline ^[6]	85.14	68.59
ResNet_STN	87.43	70.38
Res50-baseline(+SFFM)	91.78	75.17
ResNet_STN(+SFFM①②)	90.17	74.10
ResNet_STN(+SFFM- L^k)	91.83	77.41
ResNet_STN(+SFFM)	92.87	81.80
+RK	94.22	85.03

为了进一步验证局部特征之间对齐的相似程度, 随机选取 10 对正例样本统计归一化后的相似性距离 $d \in [0, 1]$. 如图 4 所示, 对任意一对正例数据 $\{I_1, I_2\}$, 其对应的横坐标表示图像 I_1 的 6 个局部区域, 纵坐标表示图像 I_2 的 6 个局部区域, 计算两两局部特征间距离. 距离越小, 表示相似程度越高, 在图 4 中表示的颜色越浅; 距离越大, 表示相似程度越低, 在图 4 中表示的颜色越深. 根据随机选择的 10 对样本数据计算各部分平均距离的统计示例来看, $d_{\min} = (I_{2\zeta}, I_{1\zeta})$, $\zeta = \{1, \dots, 6\}$, 即正例数据相对应的局部特征间的平均距离最小, 相似程度最高, 说明本文的网络模型提取到的局部特征达到了较好的对齐效果.

图 4 局部特征相似性度量统计示例

在 Market-1501 数据集中, 将本文方法与现有的行人重识别方法进行比较, 如表 2 所示. 对比方法包括传统方法(BoW+kissme^[23], LOMO+XQDA^[24])和深度学习方法(MSCAN^[8], TripNet^[12], GLAD^[15], HA-CNN^[16], AlignedReID^[17], PCB^[7], SRB^[14]), 分析如下:

(1) 相比于传统方法, 基于深度学习方法 of 准确率有大幅度的提升, 这也证明了深度学习在特征表示以及度量学习方面的优越性.

(2) 相比于基准的分类识别方法 MSCAN^[8], TripNet^[12], 本文方法在 Market-1501 数据集上的准确率结果都有大幅度的提高. 对比局部特征学习的方法 GLAD^[15], 本文方法不需要额外的姿态估计模型以及关键点信息, 这证明了本文方法有效地增强了特征表达, 从而提高行人重识别的性能.

(3) 相比于现有的行人重识别方法 SRB^[14], 本文方法达到了相近的结果, 这主要是结合了鲁棒特征和度量学习的优势. SRB 方法利用行人的全局特征进行识别分类, 但是考虑行人图像本身的多样性以及相似性, 局部信息的引入往往会更加有助于区分性特征的表达. 以图 2 中行人 C 的样本图像为例, 红色方框中的图像与其他行人图像外观相似、网络提取的全局特征也相似, 计算样本对特征之间的距离 $d_{a,p} - d_{a,n} > 0$, 所以容易将其分错.

表 2 13 种方法在 Market-1501 数据集上的对比 %

方法	Rank-1	mAP
BoW+kissme ^[23]	44.42	20.76
LOMO+XQDA ^[24]	22.22	43.79
MSCAN ^[8]	80.31	57.53
TripNet ^[12]	84.92	69.14
TripNet(RK) ^[12]	86.67	81.07
GLAD ^[15]	89.90	73.90
HA-CNN ^[16]	91.20	75.70
AlignedReID ^[17]	91.80	79.30
PCB ^[7]	92.40	77.30
PCB+RPP ^[7]	93.80	81.60
SRB ^[14]	94.50	85.90
本文方法	92.87	81.80
本文方法(RK)	94.22	85.03

本文方法通过融合增强的局部特征作为最终的行人特征表示, 如图 5 所示. 联合多层次局部区域信息提取到更具有区分性的细节特征(背包肩带、鞋子颜色), 计算得到样本间的欧几里得度量 $d_{a,n} > d_{a,p}$, 因此能准确衡量正负样本间的相似程度, 最终优化目标函数达到拉近正例、推开负例的效果.

图5 不同特征表示下的行人相似程度示意图

图6分别是本文方法(w/ SFFM)与去除增强特征融合模块(w/o SFFM)在Market-1501数据集上得到的Rank-5查询结果, query表示待查询图像,由黄色框标出,后面紧跟着Rank-5的检索结果,其中正确的检索结果由绿色框标出,错误的检索结果由红色框标出。根据可视化的结果不难发现,本文方法引入的增强特征融合模块能进一步提高检索排序结果,提高准确性,也进一步说明了特征表达在行人重识别中的重要性。

查询图 Rank-5
a. w/ SFFM

查询图 Rank-5
b. w/o SFFM

图6 可视化Rank-5检索结果

4 结 语

本文提出一种基于增强特征融合网络的行人重识别方法。主要思想是在整合网络框架实现行人局部特征对齐的基础上设计一种3层级增强特征融合网络,利用语义信息间的关联性提取图像的增强特征,提高网络对相似行人的判别性。同时对融合后的特征提出一种RM存储方法,并计算局部三元组损失,达到高效度量的效果。最后,本文方法结合现有的重排算法进一步提高检索精度,在经典的行人重识别数据集上的实验结果验证了其有效性。在未来的工作中,将致力于在无监督行人重识别上有所突破,不再受限于数据本身,提高网络的实用性。

参考文献(References):

- [1] Wu C Y, Manmatha R, Smola A J, *et al.* Sampling matters in deep embedding learning[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2859-2867
- [2] Ristani E, Tomasi C. Features for multi-target multi-camera tracking and re-identification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6036-6046
- [3] Leng Q M, Ye M, Tian Q. A survey of open-world person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(4): 1092-1108
- [4] Zheng L, Yang Y, Tian Q. SIFT meets CNN: a decade survey of instance retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1224-1244
- [5] Han C C, Ye J C, Zhong Y S, *et al.* Re-ID driven localization refinement for person search[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 9814-9823
- [6] He K M, Zhang X Y, Ren S Q, *et al.* Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778
- [7] Sun Y F, Zheng L, Yang Y, *et al.* Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2018: 501-518
- [8] Li B W, Chen X T, Zhang Z, *et al.* Learning deep context-aware features over body and latent parts for person re-identification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 7398-7407
- [9] Wang G S, Yuan Y F, Chen X, *et al.* Learning discriminative

- features with multiple granularities for person re-identification[C] //Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM Press, 2018: 274-282
- [10] Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 815-823
- [11] Varior R R, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2016: 791-808
- [12] Hermans A, Beyer L, Leibe B, *et al.* In defense of the triplet loss for person re-identification[OL]. [2020-04-07]. <https://arxiv.org/abs/1703.07737>
- [13] Chen W H, Chen X T, Zhang J G. Beyond triplet loss: a deep quadruplet network for person re-identification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 1320-1329
- [14] Luo H, Gu Y Z, Liao X Y, *et al.* Bag of tricks and a strong baseline for deep person re-identification[OL]. [2020-04-07]. <https://arxiv.org/abs/1903.07071?context=cs.CV>
- [15] Wei L H, Zhang S L, Yao H T, *et al.* GLAD: global-local-alignment descriptor for pedestrian retrieval[C] //Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM Press, 2017: 420-428
- [16] Li W, Zhu X T, Gong S G. Harmonious attention network for person re-identification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2285-2294
- [17] Zhang X, Luo H, Fan X, *et al.* AlignedReID: surpassing human-level performance in person re-identification[OL]. [2020-04-07]. <https://arxiv.org/abs/1711.08184>
- [18] Yi D, Lei Z, Liao S C, *et al.* Deep metric learning for person re-identification[C] //Proceedings of the IEEE International Conference on Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 34-39
- [19] Zhong Z, Zheng L, Cao D L, *et al.* Re-ranking person re-identification with k -reciprocal encoding[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3652-3661
- [20] Jaderberg M, Simonyan K, Zisserman A, *et al.* Spatial transformer networks[C] //Proceedings of the International Conference on Neural Information Processing Systems. New York: ACM Press, 2015: 2017-2025
- [21] Szegedy C, Vanhoucke V, Ioffe S, *et al.* Rethinking the inception architecture for computer vision[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2818-2826
- [22] Zhang L, Xiang T, Gong S G. Learning a discriminative null space for person re-identification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1239-1248
- [23] Zheng L, Shen L Y, Tian L, *et al.* Scalable person re-identification: a benchmark[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 1116-1124
- [24] Liao S C, Hu Y, Zhu X Y, *et al.* Person re-identification by local maximal occurrence representation and metric learning[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 2197-2206