行人再识别技术综述

李幼蛟 [1, 2, 3] 卓 力 [1, 2, 4] 张 菁 [1, 2] 李嘉锋 [1, 2] 张 辉 [1, 2]

摘 要 行人再识别指的是判断不同摄像头下出现的行人是否属于同一行人, 可以看作是图像检索的子问题, 可以广泛应用于智能视频监控、安保、刑侦等领域. 由于行人图像的分辨率变化大、拍摄角度不统一、光照条件差、环境变化大、行人姿态不断变化等原因, 使得行人再识别成为目前计算机视觉领域一个既具有研究价值又极具挑战性的研究热点和难点问题. 早期的行人再识别方法大多基于人工设计特征, 在小规模数据集上开展研究. 近年来, 大规模行人再识别数据集不断推出, 以及深度学习技术的迅猛发展, 为行人再识别技术的发展带来了新的契机. 本文对行人再识别的发展历史、研究现状以及典型方法进行梳理和总结. 首先阐述了行人再识别的基本研究框架, 然后分别针对行人再识别的两个关键技术 (特征表达和相似性度量), 进行了归纳总结, 重点介绍了目前发展迅猛的深度学习技术在行人再识别中的应用. 另外, 本文对行人再识别中代表性的数据集以及在各个数据集上可以取得优异性能的方法进行了分析和比较. 最后对行人再识别技术的未来发展趋势进行了展望.

关键词 行人再识别, 人工设计特征, 深度学习, 特征表达, 相似性度量

引用格式 李幼蛟, 卓力, 张菁, 李嘉锋, 张辉. 行人再识别技术综述. 自动化学报, 2018, **44**(9): 1554−1568

DOI 10.16383/j.aas.2018.c170505

A Survey of Person Re-identification

LI You-Jiao[1, 2, 3] ZHUO Li[1, 2, 4] ZHANG Jing[1, 2] LI Jia-Feng[1, 2] ZHANG Hui[1, 2]

Abstract Person re-identification aims to associate the same person across different views and can be taken as a sub-problem of image retrieval. It has extensive application prospects in many areas such as intelligent video surveillance, security, and criminal investigation. Due to poor illumination condition, image resolution, camera viewpoint, environment, and pedestrian pose, person re-identification has become one of the challenging problems in computer vision. Early person re-identification methods mostly rely on hand-crafted features and researches are conducted on small-scale datasets. In recent years, the emergence of large-scale datasets and rapid development of deep learning techniques provide person re-identification with new opportunities. This survey gives a detailed overview of the history, state of the art, and typical methods in this domain. Firstly, the general framework of person re-identification is presented. Then, feature representation, similarity measurement, and two key aspects of person re-identification, are further summarized, respectively. We also highlight the application of rapid developing deep learning techniques to person re-identification. Moreover, the representative datasets of person re-identification and methods of obtaining excellent performance on each dataset are analyzed and compared. Finally, the future trends of this field are discussed.

Key words Person re-identification, hand-crafted feature, deep learning, feature representation, similarity measurement

Citation Li You-Jiao, Zhuo Li, Zhang Jing, Li Jia-Feng, Zhang Hui. A survey of person re-identification. *Acta Automatica Sinica*, 2018, **44**(9): 1554−1568

行人再识别 (Person re-identification, Re-ID) 起源于多摄像头跟踪, 用于判断非重叠视域中拍摄到的不同图像中的行人是否属于同一个人. 行人再识别涉及计算机视觉、机器学习、模式识别等多个学科领域, 可以广泛应用于智能视频监控、安保、刑侦

收稿日期 2017-09-05 录用日期 2018-01-19
Manuscript received September 5, 2017; accepted January 19, 2018
国家自然科学基金 (61531006, 61372149, 61370189, 61471013), 北京市属高等学校高层次人才引进与培养计划项目 (CIT&TCD20150311, CIT&TCD201404043), 北京市自然科学基金 (4142009, 4163071), 北京市教育委员会科技发展计划项目 (KM201410005002, KM201510005004), 北京市属高等学校人才强教计划资助项目 PHR (IHLB) 资助
Supported by National Natural Science Foundation of China (61531006, 61372149, 61370189, 61471013), the Importation Development of High-Caliber Talents Project of Beijing Municipal Institutions (CIT&TCD20150311, CIT&TCD201404043), Beijing Natural Science Foundation (4142009, 4163071), Science and Technology Development Program of Beijing Education Committee (KM201410005002, KM201510005004), and Funding Project for Academic Human Resources Development in Institutions of Higher Learning under the Jurisdiction of Beijing Municipality
本文责任编委 黄庆明
Recommended by Associate Editor HUANG Qing-Ming

1. 北京工业大学计算智能与智能系统北京市重点实验室 北京 100124 2. 北京工业大学信息学部微电子学院 北京 100124 3. 山东理工大学计算机科学与技术学院 淄博 255000 4. 北京电动车辆协同创新中心 北京 100081
1. Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124 2. College of Microelectronics, Faculty of Information Technology, Beijing University of Technology, Beijing 100124 3. College of Computer Science and Technology, Shandong University of Technology, Zibo 255000 4. Beijing Collaborative Innovation Center of Electric Vehicles, Beijing 100081

等领域. 近年来, 行人再识别技术引起了学术界和工业界的广泛关注, 已经成为计算机视觉领域的一个研究热点. 由于行人兼具刚性和柔性物体的特性, 外观易受穿着、姿态和视角变化以及光照、遮挡、环境等各种复杂因素的影响, 这使得行人再识别面临着巨大的技术挑战.

对再识别的研究可以追溯到 2003 年, Porikli[1] 利用相关系数矩阵建立相机对之间的非参数模型, 获取目标在不同相机间的颜色分布变化, 实现了跨视域的目标匹配. 2006 年, Gheissari 等[2] 首次提出行人再识别的概念, 利用颜色和显著边缘线直方图 (Salient edge histograms) 实现行人再识别. 经过多年的研究, 行人再识别取得了诸多有意义的成果. 2010 年, Farenzena 等[3] 第一次在计算机视觉领域的顶级会议 CVPR (Computer vision and pattern recognition) 上发表了关于行人再识别的文章 *Person re-identification by symmetry-driven accumulation of local features*. 自此以后, 在计算机视觉领域的国际重要会议, 如 CVPR, ICCV (International conference on computer vision), BMVC (British machine vision conference), ECCV (European conference on computer vision), ICIP (International conference on image processing) 和权威期刊, 如 TPAMI (*Transactions on Pattern Analysis and Machine Intelligence*), IJCV (*International Journal of Computer Vision*), *Pattern Recognition* 等, 行人再识别都成为一个重要的研究方向, 涌现了大量的研究成果. 尤其是近年来, 很多学者和研究机构陆续公布了专门针对行人再识别问题的数据集, 极大地推动了行人再识别研究工作的开展.

行人再识别的典型流程如图 1 所示. 对于摄像头 A 和 B 采集的图像/视频, 首先进行行人检测, 得到行人图像. 为了消除行人检测效果对再识别结果的影响, 大部分行人再识别算法使用已经裁剪好的行人图像作为输入. 然后, 针对输入图像中提取稳定、鲁棒的特征, 获得能够描述和区分不同行人的特征表达向量. 最后根据特征表达向量进行相似性度量, 按照相似性大小对图像进行排序, 相似度最高的图像将作为最终的识别结果.

行人再识别包括两个核心部分: 1) 特征提取与表达. 从行人外观出发, 提取鲁棒性强且具有较强区分性的特征表示向量, 有效表达行人图像的特性; 2) 相似性度量. 通过特征向量之间的相似度比对, 判断行人的相似性. 可以看出, 行人再识别与图像检索的思路相同, 可以看作是图像检索的子问题.

根据行人再识别采用的数据源, 可分为基于图像的行人再识别和基于视频的行人再识别. 后者得益于视频中包含更为丰富的时间信息, 可以获得更优的性能.

根据采用的特征提取与表达方法, 行人再识别技术的发展可以分为两个阶段: 1) 2012 年之前的人工设计特征阶段; 2) 2012 年之后的深度特征阶段. 随着深度学习研究的不断深入, 各种基于深度学习的行人再识别方法被不断推出, 并取得了远超过传统方法的性能[4].

本文对基于人工设计特征和基于深度学习的行人再识别技术的研究进展情况进行综述. 第 1 节介绍基于人工设计特征的行人再识别方法研究进展, 重点阐述特征提取与表达、相似性度量的常用方法. 第 2 节介绍基于深度学习的行人再识别方法研究进展, 将其分为端到端式、混合式和独立式分别加以介绍. 第 3 节介绍具有代表性的行人再识别数据集, 并对各个数据集上取得优异性能的方法进行详细分析和比较. 第 4 节对行人再识别技术的未来发展趋势进行展望.

1　基于人工设计特征的行人再识别

基于人工设计特征的行人再识别主要包含特征提取与表达和相似性度量两部分. 特征是整个行人再识别的基础, 特征的好坏直接影响到最终的识别性能, 合理的相似性度量方法将进一步提高识别准确率.

1.1　特征提取与表达

行人再识别采用的特征可分为低层视觉特征、中层滤波器特征和高层属性特征三类. 另外, 在基于

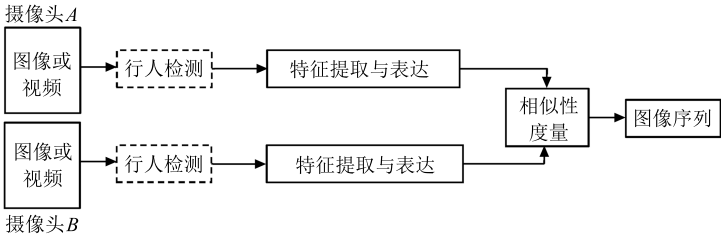


图 1　行人再识别典型流程图

Fig. 1　Typical flowchart of person Re-ID

视频的行人再识别中, 不仅提取空间特征, 而且提取时间特征来反映视频的运动信息, 提高识别精度.

低层特征是指颜色、纹理等基本的图像视觉特征. 低层视觉特征及其组合是行人再识别中常用的特征. 多个低层视觉特征组合起来比单个特征含有更加丰富的信息, 具有更好的区分能力, 因此常将低层视觉特征组合起来用于行人再识别.

中层滤波器特征是指从行人图像中具有较强区分能力的图像块组合中提取出的特征. 滤波器是对行人特殊视觉模式的反映, 这些视觉模式对应不同的身体部位, 可以有效表达行人特有的身体结构信息.

高层属性特征是指服装样式、性别、发型、随身物品等人类属性, 属于软生物特征, 拥有比低层和中层特征更强大的区分能力.

虽然行人再识别技术最初提出的目的是用于视频追踪, 但受限于有限的计算和存储能力, 目前大多数行人再识别方法是基于静止图像的, 各个摄像头下仅拍摄有一张或少数几张行人图像. 然而, 静止图像中包含的信息十分有限, 导致再识别的准确性难以尽如人意. 近年来, 许多学者开始利用视频进行行人再识别. 相对于静止图像, 视频中包含更加丰富的时空信息, 充分利用视频中的时空特征, 可以获得更优的识别性能.

1.1.1 低层视觉特征

行人外观具有丰富的颜色信息, 颜色是行人再识别中最常用的低层视觉特征之一. 颜色直方图是应用最为广泛的一种特征, 可以表征行人图像的整体颜色分布. 此外, 颜色矩、颜色相关图、颜色聚合向量等也是主要的颜色特征. Farenzena 等[5] 提出一种对称驱动的局部特征累计方法 (Symmetry-driven accumulation of local features, SDALF). 该方法首先取得行人的前景图像, 然后分别提取三种互补的颜色特征: 加权颜色直方图、最稳定颜色区域和高重复结构颜色块, 三种特征结合起来用以描述行人外观的颜色特性.

颜色特征对于姿态和视角变化具有鲁棒性, 但易受光照和遮挡的影响, 而且由于着装相似问题, 只利用颜色特征很难有效区分大规模的行人图像. 行人衣着常包含纹理信息, 而纹理特征涉及到相邻像素的比较, 对光照具有鲁棒性, 因此很多研究工作将颜色和纹理特征组合起来使用. 文献 [6] 提出一种 ELF (Ensemble of localized features) 特征, 利用 Adaboost 算法在一组颜色和纹理特征中选择出合适的特征组合, 可以提高识别的准确性.

颜色和纹理特征能够提供行人图像的全局信息, 但是缺乏空间信息. 因此很多行人再识别方法在颜色和纹理特征中加入空间区域信息. 行人图像被分成多个重叠或非重叠的局部图像块, 然后分别从中提取颜色或纹理特征, 从而为行人特征增加空间区域信息. 当计算两幅行人图像的相似度时, 对应的图像块内的特征将分别进行比较, 然后将各个图像块的对比结果融合, 作为最终的识别结果[3, 5]. 或简单地将各个图像块特征级联为一个特征向量, 然后进行对比[6−8].

表 1 是对行人再识别常用的几种典型图像块分割方法进行的归纳和总结. 采用局部分割模型的目的是通过对人体结构的多层次建模, 提高局部特征的判别性和区分性, 尽可能多地过滤掉背景信息. 图 2 是表 1 中四种分割方式的分割结果示意图. 图 2 (a) ∼ 2 (d) 依次为上下半身分割法、条纹分割法、滑动窗分割法和三角形分割法. 其中滑动窗分割方法符合人类的视觉规律, 识别效果最好.

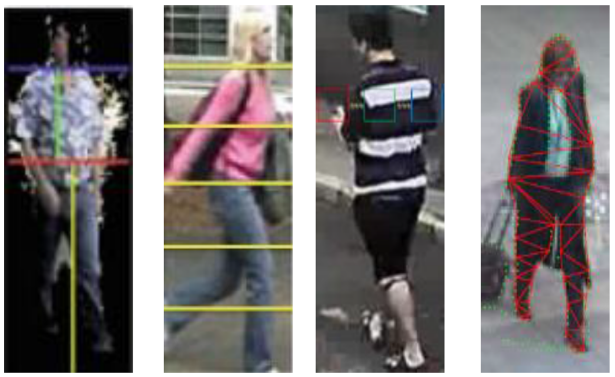
图像块分割方法可以利用行人身体子块位置的先验知识. 采用这种方法实现的识别过程相对简单, 但是无法确保图像子块与身体子块之间的精确匹配, 对于强烈的视角变化, 鲁棒性较差.

另外, 当多种低层特征组合使用时, 随着特征数目的增加, 特征向量维数会呈指数增长. 利用协方差作为行人图像的特征描述, 可以大大降低特征维数[9−10]. 文献 [9] 提出一种 HSCD (Hybrid spatiogram and covariance descriptor) 描述符, 将空间直方图与协方差算子进行融合. 空间直方图由各个

表 1 典型行人图像分割方法

Table 1 Typical segmentation methods of pedestrian image

分割方式	对应文献	主要思想
上下半身分割	[3, 5]	提取行人的前景图像, 分成头部、躯干和腿部三部分. 对后两部分计算垂直对称轴. 对提取的特征根据与垂直对称轴的距离进行加权, 从而减少行人姿态变化的影响. 缺点是分割过程过于复杂.
条纹分割	[6−7]	分成六个水平条, 分别对应于行人头部、水平躯干的上下部、腿部的上下部分. 然后提取水平条内的 ELF 特征, 减少了视角变化对识别的影响. 缺点是会造成水平条内空间细节信息的损失.
滑动窗分割	[8]	利用滑动窗来描述行人图像的局部细节信息, 在每个滑动窗内提取颜色和纹理特征. 缺点是特征维数过大.
三角形分割	[2]	利用局部运动特征对行人图像进行三角形时空分割. 缺点是分割结果不够准确.



(a)　(b)　(c)　(d)

图 2　行人图像块分割方法

Fig. 2　Patch segmentation methods of pedestrian image

图像区域上的多通道颜色直方图累加而成. 协方差算子由相同图像区域中包含了颜色和纹理信息的协方差矩阵构成. 然而, 协方差描述符会去除图像的均值信息, 而这些信息对区分行人是非常重要的. 文献 [11] 提出一种 GOG (Gaussian of Gaussian) 描述符, 利用分层高斯算子将图像分为由多个高斯分布进行描述的不同区域来表示颜色和纹理信息, 每种高斯分布代表一个小的图像块, 每个图像块的特征组合起来得到行人图像的特征向量, 用于识别.

低层特征的提取不需要复杂的训练过程, 可解释性较强. 但是表达能力较弱, 面对复杂的识别环境其泛化能力受到一定制约, 无法针对具体的行人再识别任务进行优化.

1.1.2　中层滤波器特征

中层滤波器特征是利用聚类算法, 从行人图像中学习出一系列有表达能力的滤波器. 每一个滤波器都代表一种与身体特定部位相关的视觉模式, 也称显著区域 (Salient region). 如果在同一行人的多幅图像中存在由若干小的图像块组成的显著区域, 例如提包, 会有助于做出判断, 如图 3[12] 所示, 图中虚线框为显著区域检测结果. 如果提包出现在多张图像中的不同空间位置, 很多行人再识别算法会将其忽略. 这些方法[3, 5, 13−14] 通常只考虑大块的对应上衣和裤子的颜色区域, 小的颜色区域因为不属于身体主要区域会被当作异常值而忽略掉. 因为显著区域对于光照和视角变换具有较强的鲁棒性, 因此合理利用显著区域会有效提高再识别的性能[12, 15−16].

Zhao 等[12] 以非监督的方式得到图像中的显著区域用于行人再识别, 对获得的显著区域进行显著性排序, 并据此分配权重. Cheng 等[17] 采用类似 SDALF 的前景分割方法, 利用人们对于行人外观的先验知识首先提取出行人的前景图像. 然后基于图画结构训练出包含 11 个身体部分的人体结构模型, 在该模型的基础上提取颜色直方图组成行人图像的特征表达, 用于行人再识别.

图 3　行人显著区域示意图

Fig. 3　The illustration of salient region

人体由各个身体部位组成, 具有良好的结构特性, 使用与人体部位对应的滤波器特征能够平衡行人描述符的区分能力和泛化能力. 低层和中层特征结合起来使用能够充分发挥各自的优势, 在一定程度上克服行人再识别中的光照和视角变化问题. 但是, 人体是非刚性目标, 外观易受到姿态、遮挡等各种因素的影响, 仅利用低层和中层特征会导致识别精度不高, 还需要利用其他更高层的特征.

1.1.3　高层属性特征

人类在辨识行人时会使用离散而精确的特有属性 (Attribute), 例如服装样式、性别、胖瘦等都属于行人的属性特征. 行人图像对应的属性特征通常采用离散的二进制向量表示形式, 例如图 3 中的行人, 假设定义 3 个属性 (是否男性; 是否长发; 是否携带提包), 则对应的属性特征向量为 [1 0 1]. 与其他特征相比, 高层属性特征尽管在提取和表达方面复杂, 属性标定需要大量的人工和时间成本, 但含有更加丰富的语义信息, 而且对于光照和视角变化具有更强的鲁棒性. 因此, 属性特征与低层特征联合使用, 可以有效提高识别性能.

Layne 等[18] 将属性特征用于行人的再识别. 针对服装的样式、发型、随身物品以及性别设计并手工标注了 15 种基于低层特征的行人属性. 在进行基于属性的行人再识别时, 首先利用一组人工标定好属性的样本图像训练支持向量机 (Support vector machines, SVM) 属性分类器, 将属性判别结果用于

行人再识别. 因为训练样本中的某一属性是通过不同摄像头拍摄的图像学习得到的, 因此属性分类器具有一定的视角鲁棒性.

属性的标定费时费力, 因此研究者们开始探索如何扩展已有的属性. 文献 [19] 借助其他非行人再识别专用的大型数据集训练出一组属性, 这些大型数据集带有颜色、纹理和类别标签. 训练好的属性通过非监督的方式直接应用到小型的行人再识别数据集上. 无论是手工标注还是通过低层特征学习得到的属性特征, 彼此之间相互独立. 如果能利用属性特征中包含的语义信息, 将属性特征投影到连续的有关联的属性空间中, 将大大提高属性特征的区分能力. 文献 [20] 利用多任务学习[21] 得到行人属性特征的相关性低秩矩阵, 通过该矩阵转换后的属性特征向量具有较小的类内差和较大的类间差, 因此具有很好的区分性.

属性特征可以对行人图像进行语义层面的解释, 能够有效缩小低层视觉特征与高层语义特征之间的语义鸿沟. 研究结果表明, 与低层特征相比, 在再识别过程中使用高层属性特征, 性能明显提升, 以最常用的 VIPeR 数据集[22] 为例, 平均识别精度可以提高 6 % 左右.

1.1.4 视频时空特征

在基于视频的行人再识别中, 每个行人至少包含两段跨视域的视频序列, 其中包含数量不等的视频帧. 这些视频帧能够提供大量的训练样本, 可以更方便地训练机器学习算法, 从而提高识别的性能.

处理视频最常用的方法是提取每一帧的低层特征, 然后利用平均/最大池化方法将其聚合为一个全局特征向量, 用以反映行人的外观信息[23]. 值得注意的是, 虽然视频数据量巨大, 但是人们感兴趣的信息可能主要集中在某些方面. 另外, 视频中的冗余信息对识别结果有一定的负面影响. 因此, 许多学者致力于从视频中挖掘更有效的信息. Gao 等[24] 提出一种时间对准池化方法, 利用行走的周期特性将视频序列分成独立的行走周期, 选择最符合正弦信号特性的周期代表该视频序列, 提高了识别性能.

与图像相比, 视频序列中的帧与帧之间不仅存在空间依赖关系, 也存在时间次序关系, 合理利用视频的时间特征能够反映行人的运动特性, 提高识别准确率. 因此, 对于基于视频的行人再识别来说, 往往提取视频的时空特征用于识别. 在判别视频帧选择排序 (Discriminative video fragments selection and ranking, DVR)[25] 方法中, 首先通过计算每个行人视频序列的步态能量图像[26] 来提取行人的运动特征, 然后融合 HOG3D[27] 时空特征, 最后通过判别视频帧排序模型进行相似性度量. You 等[23] 采用 HOG3D 时空特征, 并融合行人图像的颜色和纹理特征作为行人的特征表达.

总的来说, 时空特征反映了视频中的运动信息, 是行人外观特征的有效补充. 然而, 时空特征易受视角、尺度和速度等因素的影响, 在新型的大型行人再识别数据集上表现得差强人意. 因为对于大型行人再识别数据集来说, 随着行人的大幅增加, 行人之间的运动相似性也随之增加, 这使得时空特征的区分能力大幅下降. 同时, 大型数据集中摄像头数量多, 使得同一行人的姿态差异增大, 运动差异愈加明显, 这些都限制了时空特征在行人再识别中的作用. 因此, 如何设计更具区分性的时空特征是基于视频的行人再识别需要解决的问题.

1.2 相似性度量

行人再识别利用特征之间的相似性来判断行人图像的相似性, 特征相似的行人图像将被看作是同一个人, 选择合适的相似性度量方法对行人再识别至关重要. 根据度量过程中是否使用标签, 相似性度量可以分为无监督度量和监督度量. 另外, 在基于视频的行人再识别中, 行人除了外观相似之外, 不同行人的运动特性也往往非常相似, 这使得行人再识别成为一个挑战性的难题. 如何设计相似性度量方法, 对特征相似的行人加以区分是提高行人再识别性能需要解决的关键问题.

1.2.1 无监督度量

无监督度量直接利用特征表达阶段获得的特征向量进行相似性度量. 特征向量之间的相似性往往通过特征向量之间的距离进行度量, 特征向量之间的距离越小, 说明行人图像越相似. 早期的行人再识别研究工作通常使用简单的欧氏距离或巴氏距离作为相似性度量方法. 假设 x, y 分别代表两个摄像头下的行人图像特征向量, 则对应的欧氏距离为

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \quad i = 1, 2, \cdots, n \tag{1}$$

巴氏距离[28] 经常在分类任务中用于测量类之间的可分离性, 其计算公式为

$$D_B(x, y) = -\ln(BC(x, y)) \tag{2}$$

其中, $BC(x, y) = \sum \sqrt{x_i y_i}$ 代表巴氏系数. 文献 [3] 中提取了加权颜色直方图、最稳定颜色区域和高重复结构颜色块三种行人特征, 前两种特征采用巴氏距离而最后一种采用欧氏距离进行度量, 三种距离的加权和作为最终的特征距离.

欧氏距离和巴氏距离等简单的几何距离通常将数据的各个维度等同对待, 没有考虑不同维度对识

别效果的影响程度, 因此获得的相似度并不准确. 而监督方式利用带标签的训练集样本, 通过对目标函数的优化, 可以获得能够有效反映样本相似关系的特征空间, 成为目前行人再识别中相似性度量的主要方法[7−8, 23].

1.2.2 监督度量

距离度量学习是基于成对约束的监督度量方法, 基本思路是利用给定的训练样本集学习得到一个能够有效反映数据样本间相似度的度量矩阵, 在减少同类样本之间距离的同时, 增大非同类样本之间的距离. 当特征向量提供的信息足够充足时, 距离度量能够获得比非监督方式更高的区分能力. 但是, 与非监督度量方法相比, 距离度量学习需要额外的学习过程, 在训练样本不足时容易产生过拟合现象, 且图像库和场景变化时需要重新训练.

距离度量学习最常见的是基于马氏距离[29] 的度量. 给定一个 $\mathbf{R}^d$ 空间上的 n 个特征向量 $[x_1, x_2, \cdots, x_n]$, 找到一个半正定矩阵 $M \in \mathbf{R}^{d\times d}$, 则向量对 (x_i, x_j) 之间的马氏距离为

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^{\mathrm{T}} M (x_i - x_j)} \tag{3}$$

式 (3) 可以转化为凸优化问题进行求解[30]. 例如 Zheng 等[31] 提出一种概率相对距离比较 (Probabilistic relative distance comparison, PRDC) 方法, 对行人特征的相对距离函数进行优化. 对于每张行人图像, 选择同一行人样本和不同行人样本组成三元组, 在训练过程通过最小化不同类样本距离与同类样本距离的和, 得到满足相对约束的马氏距离度量矩阵.

经典的度量学习方法有大间隔最近邻 (Large margin nearest neighbor, LMNN)[32]、基于信息论的度量学习 (Information theoretic metric learning, ITML)[33] 和基于逻辑判别的度量学习 (Logistic discriminant metric learning, LDML)[34] 等. 在行人再识别问题中, 行人的特征表达往往包含图像的多种统计信息, 使得行人图像的特征向量结构复杂, 维数较高. 上述方法由于复杂的优化策略对系统资源造成了过高的负担, 因此不适合大规模的行人再识别.

保持简单直接度量算法 (Keep it simple and straightforward metric, KISSME)[35] 不需要通过复杂的迭代算法计算度量矩阵, 因此计算效率更快. 实验结果表明, 对比 ITML 等传统算法, KISSME 算法在识别准确率和算法效率上都更具有优势. KISSME 通过似然比检验的方法将距离度量学习转化为

$$\delta(x_{ij}) = \log \frac{p(x_{ij}|H_0)}{p(x_{ij}|H_1)} \tag{4}$$

其中, $x_{ij} = x_i - x_j$, H_0, H_1 分别为样本对相似与否的假设检定.

KISSME 包含两个主要阶段: 1) 进行主成分分析 (Principal component analysis, PCA) 降维; 2) 利用 PCA 子空间上行人类内差和类间差的协方差矩阵学习距离函数. 然而, 这种两阶段的处理方式在低维空间中很可能无法求得最优解. 因为在经过第一阶段之后, 隶属于不同类的样本会变得杂乱无章. 跨视域二次判别分析方法 (Cross-view quadratic discriminant analysis, XQDA)[8] 对该方法进行了改进, 能够同时学习基于跨视域数据的子空间和低维空间上的距离度量, 通过学习行人类内差和类间差的协方差矩阵的核度量来建立距离度量函数.

总的来说, 由于具有去耦合和量纲无关两种优良的性质, 使得基于马氏距离的距离度量学习方法在行人再识别中应用得最为广泛. 在传统的小型数据集上, 为了获取更加丰富的行人信息, 行人描述符的维度远远超过训练样本的数量, 造成距离学习过程中的小样本问题 (Small sample size, SSS). 为了解决该问题, 往往需要对行人特征进行降维和正则化处理, 导致距离学习函数只能获得次优解. 最近, 大型行人再识别数据集的出现有效缓解了距离度量学习的小样本问题. 然而, 目前的距离度量算法大都是基于成对约束的, 约束的数量是训练样本数量的平方, 导致大样本时约束数量将变得非常巨大. 因此, 构建合理的训练约束库, 设计更加快速有效的训练机制, 将是距离度量学习下一步需要深入研究的问题.

1.2.3 基于视频的距离度量

在基于视频的行人再识别方法中, 大多沿用基于马氏距离的度量方法. 例如顶推距离学习模型 (Top-push distance learning model)[23] 是专门为基于视频的行人再识别设计的度量方法, 通过对样本对之间最大的干扰项施以较大的惩罚来快速有效地增大类间差异. 顶推距离学习比较的不是正样本对与所有相关的负样本对之间的距离, 而是正样本对与所有相关负样本对的最小距离.

与顶推距离学习模型采用马氏距离矩阵不同, Karanam 等[36] 提出一种 SRID (Sparse re-id) 方法, 利用字典学习进行相似性度量, 通过求解共同嵌入空间上的块稀疏恢复问题来确定行人类别. 类似地, Karanam 等[37] 提出的 DVDL (Discriminative dictionary learning) 方法利用与 SRID 相同的特征, 学习出一个矩阵以及对应的稀疏编码, 通过优化

稀疏编码的欧氏距离来提升字典的区分能力.

综上所述, 基于人工设计特征的行人再识别经过多年研究, 已经取得很大的研究进展, 积累了许多成功经验. 然而, 人工设计特征的优劣很大程度上依赖于设计者的先验知识和手工调参. 由于在特征设计中允许出现的参数个数十分有限, 有效特征的产生需要一个漫长的过程, 特征的泛化能力较弱.

2 基于深度学习的行人再识别

自从 Krizhevsky 等[38] 获得 ILSVRC′12 分类竞赛冠军以来, 基于卷积神经网络 (Convolutional neural network, CNN) 的深度学习被广泛应用于计算机视觉领域[39−42]. 深度学习与传统方法的最大不同在于其特征是从大数据中自动学习得到的, 通过建立类似于人脑的分层模型结构, 能够从大量数据中逐级提取由底层到高层的特征, 获得适合于分类或者识别的深度特征.

2014 年, Li 等[43] 率先将深度学习应用到行人再识别中, 提出一种基于 CNN 的 FPNN (Filter pairing neural network) 网络模型. FPNN 的第一层是带最大池化操作的卷积层, 然后加入块匹配层对跨视域的滤波器响应进行匹配. FPNN 能够在一个统一的框架下解决未对准、遮挡和背景等因素对识别性能的影响.

在行人再识别中用到的深度学习模型往往包含三个基本的网络层: 卷积层、池化层和全连接层, 如图 4[43] 所示. 首先将不同视域中的行人图像作为网络的输入, 然后将这些图像分解为不同的颜色通道子图分别进行处理. 对于每幅子图, 在接下来的卷积层中对其实施卷积滤波操作, 得到不同局部图像块的响应, 作为局部特征. 这些局部特征组合起来, 形成特征图, 作为该卷积层的输出. 卷积层的作用是提取图像的各种信息, 例如边缘和形状. 在接下来的池化层中对产生的特征图进行最大/平均池化操作. 池化层的作用是对卷积后的特征信号进行抽象, 从而大幅减少训练参数, 另外还可以减少过拟合现象的出现. 卷积层和池化层可以出现多次, 获得行人抽象的、多层次描述. 全连接层的作用是将池化层得到的特征图投影到一维的特征空间, 形成行人图像的特征向量. 在最后的 Softmax 层, 通过二值函数判断输入的图像对是否输入同一个行人.

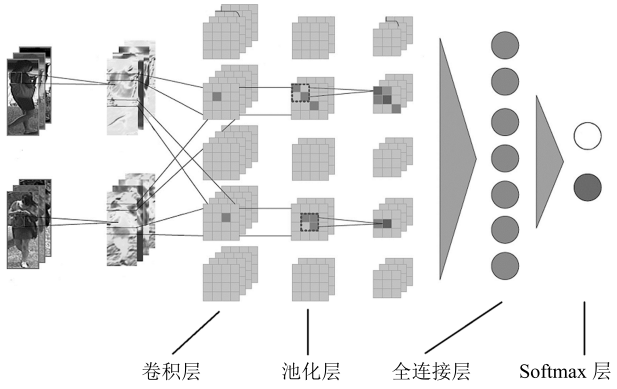
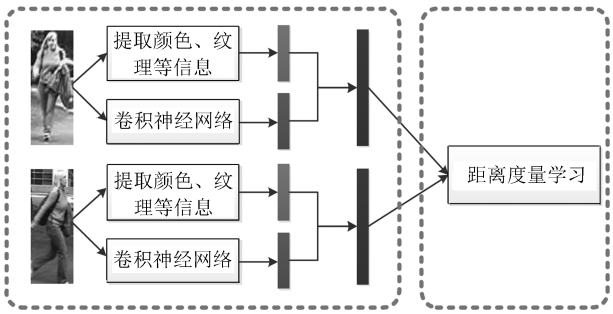


图 4 深度学习模型各网络层示意图

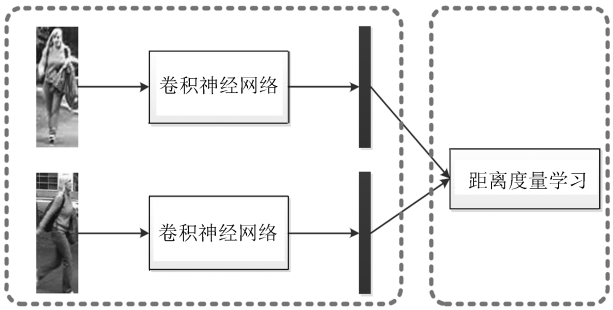
Fig. 4 Illustration of the network layers in deep learning model

深度学习模型可以将特征表达和相似性度量两个环节整合在一起, 通过二者的联合优化获取远超传统方法的性能. 按照两个环节整合方式的不同, 将基于深度学习的行人再识别方法分为端到端式、混合式和独立式三种 (如图 5 所示).



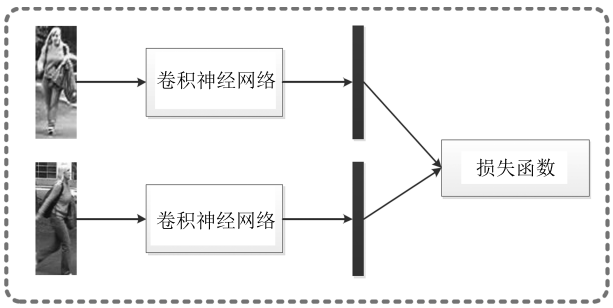
(a) 混合式

(a) Hybrid



(b) 独立式

(b) Standalone



(c) 端到端式

(c) End-to-end

图 5 基于深度学习的行人再识别方法的三种方式

Fig. 5 Three ways of deep learning-based person re-identification

2.1 端到端式的深度行人再识别

端到端式的行人再识别利用深度学习模型, 将特征提取和相似性度量这两个主要环节整合到一个统一的框架下进行联合优化, 形成一种端到端的行人再识别方案. 例如 Ahmed 等[44] 利用多输入的近邻差和图像块简要特征, 提出一种增强的深度学习框架, 用以学习跨视域特征之间的关系. 文献 [43−44] 均以 Siamese 网络为原型, 以不同视域的行人图像对作为网络输入, 网络输出是二进制变量, 用以指示输入的图像对是否属于同一个人. 采用 Siamese 网络的重要原因是行人再识别数据集规模普遍较小, 无法充分发挥深度学习的优势, 而采用图像对的输入形式可以大大增加训练样本.

在基于视频的行人再识别中, 如何有效利用深度网络提取时间特性是提升识别效果的关键. 而 CNN 假设输入数据是互相独立的, 因此无法对时间序列数据进行建模. 为此, 最近的研究工作是在 Siamese 网络的基础上引入递归神经网络 (Recurrent neural network, RNN) 和长短期记忆网络 (Long short-term memory, LSTM), 将视频的时间特性与深度特征相融合, 提升识别准确度.

1) 递归神经网络

递归神经网络通过添加跨越时间点的自连接隐藏层, 提取视频的时间特性, 将时间特性引入到行人特征中. 原理是利用建立在隐含层的状态向量, 隐式地记录视频序列的历史信息. RNN 的输入不仅包括当前的视频帧, 还包括上一个视频帧, 输出是最后一帧的特征向量. 文献 [45] 中, 视频帧及其光流图作为 CNN 的输入, 通过两个对称的 CNN 模型提取视频序列的深度特征, 深度特征再经过一个 RNN 层投影到一个低维特征空间, 并与前一时刻的信息进行组合, 从而获取视频中的时间信息. 最后利用时间池化层对深度特征进行聚合, 降低特征维度, 并提升特征的鲁棒性. 该方法利用视频帧刻画行人的外观信息, 光流图可以直接对行人的运动特性 (例如步态) 进行编码, 通过两种信息的整合来提高行人视频特征的区分能力.

2) 长短时记忆网络

RNN 结构可以记忆历史信息, 但记忆时间极短. LSTM 网络是对 RNN 的改进, 可以记忆长时信息, 对视频中复杂的动态信息建模. LSTM 同样可以利用历史信息更新当前状态, 它与传统 RNN 的不同在于神经元的构造, LSTM 中的神经元在下一个加权的时间段内与自身连接, 因此可以复制自己的实际状态并累加外部信号, 最后一个 LSTM 节点的输出信息囊括了整个视频序列的信息. 文献 [46] 提出一种递归特征聚合网络 (Recurrent feature aggregation network, RFA-Net), 利用 LSTM 记录行人身体部位随时间的变化信息. 与文献 [45] 不同, RFA-Net 将低层人工特征, 而不是 CNN 深度特征, 作为时间节点的输入, 目的是为了避免在小数据上训练 CNN 带来的过拟合现象. LSTM 可以对性能优异的特征进行记忆和传播, 同时忽略掉较差特征, RFA-Net 利用 LSTM 的这一优势, 建立视频的全局特征表达.

端到端式的行人再识别方法沿用了在图像分类中常见的深度网络模型, 一经提出即引起广泛关注. 然而, 早期的端到端方法在进行相似性比较时, 往往采用简单的欧氏距离或余弦距离, 缺少距离学习的过程, 影响了识别准确率. 常见的解决方法是在深度网络训练过程中加入损失函数约束, 使得同类样本距离变小, 异类样本距离变大, 达到距离学习的效果[47]. 另外, 随着训练样本的增加, Siamese 网络模型会变得过于复杂, 需要漫长的训练过程.

2.2 混合式的深度行人再识别

人工特征经过多年研究已经较为成熟, 在一些应用中取得了较好的结果. 例如 LBPH 特征适用于人脸识别、HoG 特征比较适用于行人检测等. 为此, 人们尝试将深度特征和人工特征相结合, 利用距离度量学习进行相似性度量, 实现行人再识别, 本文称之为混合式的深度行人再识别方法. 该方法可以采用较为成熟的人工特征表达行人的局部特性, 采用浅层的网络结构提取行人的全局特征, 二者结合可以充分发挥各自优势, 在一定程度上弥补训练数据的不足, 同时可以在一定程度上避免深度网络模型过于复杂、网络训练速度慢的缺点.

Wu 等[48] 将 CNN 特征和 ELF[6] 特征相结合, 提出了一种特征融合网络, 利用反向传播算法对 CNN 特征的提取过程进行约束. 融合后的特征结合传统的距离度量方法, 在三个小型行人再识别数据集上取得了良好的识别效果. 文献 [49] 尝试将 PCANet 网络[50] 特征与 LOMO 特征[8] 结合, 提出一种基于多种统计信息的金字塔级联行人描述符, 在小型数据集 VIPeR 上获得了优越的性能. Zheng 等[51] 提出一种查询自适应融合方法对不同特征的性能进行衡量, 并将 CNN 特征和另外五种人工特征相结合用于行人再识别. 研究结果表明, 混合式的深度行人再识别方法能够将人工设计特征和深度特征的互补性发挥出来, 使结合后的行人描述符对于视角和光照变化等问题具有更强的鲁棒性.

混合式的行人再识别方法适用于中小型的数据集, 结合先验知识, 已经证明其有效性的人工特征, 只采用浅层的网络结构即可达到较高的识别准确率, 大大简化了网络训练过程.

2.3 独立式的深度行人再识别

独立式的深度行人再识别方法的框架与基于人工特征的方法相似, 不同的是采用深度神经网络提取行人图像的深度特征, 然后结合距离度量学习方法完成行人再识别. 随着大型行人再识别数据集的不断提出, 提取行人的深度特征成为可能, 并能获得很好的识别性能. 例如文献 [52] 提出一种姿态融合网络, 将原始图像、姿态校正图像和姿态估计置信打分作为深度残差网络 (Deep residual network)[53] 的输入, 再结合 KISSME[35] 度量学习方法提高识别率. 文献 [54] 首先利用 CNN 特征学习出一个身份嵌入模型, 然后利用置信加权度量学习进行相似性度量, 可以有效提高识别性能.

随着深度学习技术的快速发展, 性能优异的深度网络模型不断涌现, 而独立式的深度行人再识别方法设计思路简单, 借助于新的网络模型, 能够有效提高行人再识别的性能. 表 2 是在 Market-1501 数据集上[55] 采用不同网络模型获得的首轮识别率. 从表 2 可以看出, 随着网络深度的不断增加, 识别准确率也相应提升. 但是与此同时, 网络训练复杂度也随之增加. 因此, 如何根据应用需求, 获得网络深度与识别准确率之间的最优折中是一个值得研究的问题.

表 2 Market-1501 数据集上不同深度模型对首轮识别率的影响

Table 2 Rank-1 matching rates of different deep models in Market-1501

模型名称	提出时间	首轮识别率 (%)
AlexNet[38]	2012 年	56.03
VGG-16[56]	2014 年	64.34
Residual-50[53]	2016 年	72.54

上述三类方法的性能表现与数据集的规模和特点紧密相关, 表 3 列出了三类方法在常用数据集上所能取得的最好效果.

从表 3 可以看出, 在大型数据集 Market-1501 和 MARS 上, 基于端到端式的深度学习方法取得了最好的效果. 因为目前的距离度量学习方法绝大多数是基于成对约束的, 当大型数据集中摄像头的数量超过两个时, 距离度量学习的效果将大大减弱. 而端到端式的方法利用三元损失函数取代距离度量学习, 取得了超过另外两种方法的效果. 在中型数据集 iLIDS-VID 和 CUHK01 上, 混合式和独立式的方法借助距离度量学习, 通过二次学习的过程提高了识别的准确率, 都取得了显著的效果. 而在以 VIPeR 为代表的小型数据集上, 无法提供深度模型所需的训练样本规模, 通过结合人工特征, 提升识别精度.

另外, 为了提升小型数据集上的识别性能, 可以采用预训练 (Pre-training) + 细调 (Fine-tuning) 的策略, 即在大型数据集上对网络参数进行预训练, 然后利用小型数据集中的样本对网络参数进行细调, 将其泛化到小型数据集上. 类似的思路也被应用到基于属性的行人再识别中, 例如文献 [61] 提出一种半监督的深度属性学习模型, 首先利用带标签的行人再识别数据集训练 CNN 网络, 并对目标数据集进行属性预测. 然后利用目标数据集对网络参数进行细调. 最后组合两种数据集重新训练网络, 更新属性标签, 提高属性的分类准确率. 文献 [62] 利用人体部位的先验知识提高属性标签分类的准确性. 首先将行人图像分为重叠的图像块, 然后送入带有多个属性标签的 CNN 网络进行训练, 属性标签的设计跟图像块的位置紧密相关, 得到一个能够同时预测多个行人属性的深度网络模型.

综上所述, 基于深度学习的行人再识别方法可以模拟人脑的抽象和迭代过程, 获得行人图像的分层特征表达. 深度网络的低层特征是从像素中学习得到刻画身体局部特性的边缘和纹理特征; 中间层特征则通过将各种边缘滤波器的组合来描述不同的人体部位; 高层特征描述的是整个行人的全局特征. 因此, 基于深度学习的行人再识别方法仅经过极少的预处理就可以得到从原始像素到高层语义的有效特征表达. 另外, 在行人图像中, 各种复杂的因素, 包括姿态、性别、着装等, 往往以非线性的方式组合在一起, 而深度学习可以通过多层非线性映射将这些因素分开, 利用不同的神经元代表不同因素, 使其变成简单的线性关系, 不再相互影响, 从而提升识别效果.

表 3 基于深度学习的方法目前所取得的最好效果

Table 3 The best results of deep learning-based methods

整合方式	方法	取得最好效果的数据集	提出时间	首轮识别率 (%)
端到端式	TriNet[57]	Market-1501, MARS	2017 年	84.9, 79.8
混合式	HIPHOP[58]	VIPeR, CUHK01[59]	2017 年	54.2, 78.8
独立式	LCAR[60]	iLIDS-VID[25]	2017 年	60.02

从目前行人再识别采用的深度网络结构来看, 网络层数越来越多. 通过增加网络深度, 可以提升网络的非线性表达能力, 使其更好地拟合目标函数, 获得具有更优泛化能力的分布式特征表达. 但是, 这种做法增加了网络整体的复杂度, 使网络变得难以优化, 这时需要大规模的数据集作为支撑, 否则过拟合将不可避免. 研究结果表明, 行人再识别的精度随数据集规模的增加而增加.

3　数据集及性能比较

目前已经公布了许多专门用于行人再识别的数据集. 深度学习出现之前, 大部分行人再识别方法都是采用的人工特征, 并在小型数据集上验证方法的性能. 随着对深度学习研究的深入, 出现了许多基于深度学习的行人再识别方法以及大型数据集. 其中常用的六个数据集及其主要参数如表 4 所示.

VIPeR、CUHK01 和 Market-1501 均为基于图像的行人再识别数据集. VIPeR 使用最为广泛, 包含 632 个行人, 每个人拍摄有两幅照片, 均存在不同程度的光照、视角和姿态等变化, 具有非常大的挑战性. CUHK01 数据集采用手工抠图, 包含 971 个行人的 3 884 幅图像, 均归一化到 160 像素 × 60 像素, 图像质量比较好. Market-1501 数据集包含 32 668 幅图像, 由 6 个相机在清华大学校园内拍摄 1 501 个行人得到. 行人边框采用部分变形模型 (Deformable part model, DPM) 自动检测得到, 很多图像只包含了行人的部分身体.

PRID-2011、iLIDS-VID 和 MARS 均为基于视频的行人再识别数据集. PRID-2011 数据集中的视频对通过两个固定的监控摄像头进行采集, 摄像头 A 包含 385 个行人, 摄像头 B 中包含 749 个行人. 这些行人中, 只有 200 个行人曾经出现在两个摄像头中. iLIDS-VID 是在 PRID-2011 之后公布的数据集, 与 PRID-2011 相比, 数据更加整齐, 也更有挑战性. iLIDS-VID 数据集是通过机场到达大厅的 CCTV 监控视频采集得到的, 包含 300 个行人在两个摄像头下的 600 段视频. 视频中存在严重的着装相似、光照和视角变化、复杂背景和遮挡现象, 因此识别难度大. MARS 数据集是 Market-1501 数据集的扩充版, 图像数量由 32 668 扩展到了 1 191 003 幅.

从表 3 可以看出, 随着行人再识别研究工作的不断推进, 数据集的规模越来越大, 摄像头数目越来越多. 其中 Market-1501 是迄今为止规模最大的行人再识别图像数据集, 可以提供足够多的样本数据供深度网络进行训练. 另外, 小型数据集中的行人边框是手工标定的, 而大型数据集中的行人边框则普遍利用 DPM 之类的行人检测方法自动获得.

接下来对在各个数据集上取得优异性能的行人再识别方法展开分析和比较. 在衡量算法性能时, 有两种常见的度量准则: 累计匹配性能曲线 (Cumulative match characteristic, CMC) 和 Rank-N 表格.

累计匹配性能曲线在识别性能评估中被广泛使用, 计算公式如下:

$$CMC(k)\sum_{r=1}^{k}p(r),\quad k=1,2,\cdots,m \tag{5}$$

从式 (5) 可以看出, 该曲线反映的是在前 k 个匹配结果中找到正确结果的概率. 横坐标表示 k, 纵坐标是识别率. 一般会将几种方法的识别结果画在一个坐标系中, 以便能直观地比较性能. 图 6 是五种方法在 VIPeR 数据集上的实验结果[9]. 当 $k=1$ 时, 表示首轮识别率 (即传统意义上的分类准确率): 准确匹配的目标所占的比例. CMC 曲线中纵坐标数值越大, 表明识别效果越好. 随着候选目标的增加, 准确率必然上升. 因此, CMC 曲线随着横坐标的增加呈递增的趋势.

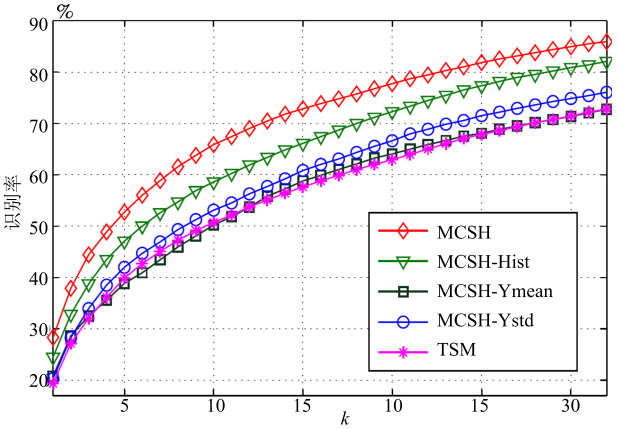


图 6　CMC 曲线示意图

Fig. 6　The illustration of CMC curve

在对不同的行人再识别方法进行直观比较时, 如果不同方法的性能差别不大, 则很难作出清晰的判断. 为此, 人们往往采用另外一种更加准确的度量方法 — Rank-N 表格, 以数值形式给出关键匹配点的累积匹配准确率. 常见的有 Rank-1, Rank-5, Rank-10 和 Rank-20. 其中 Rank-5 代表在前 5 幅图像中可以正确匹配的概率, 概率值越大表示效果越好.

下面以 Rank-N 表格的形式给出上述六个数据集上取得优异性能的方法. 表 5 是在三个行人再识别图像数据集上取得优异性能的方法对比.

表 5 中, SCSP 方法利用多项式特征图对全局特征与多种局部特征进行整合. Zhang 等[63] 针对传

统距离度量学习中存在的特征维数过高问题, 对再识别过程进行改进. WARCA 算法将加权逼近排序合成损失函数与满足线性正交投影的正则化项组合以提高识别准确率. 上述三种方法是目前公布的三个数据集上采用人工设计特征能得到的最好结果.

表 5 中的其他方法均属于基于深度学习的方法. FFN 网络中同时利用深度 CNN 特征和人工设计特征 ELF 对行人图像进行描述. HIPHOP 利用深度网络整合行人的全局和局部特征. Zheng 等[64] 融合了身份认证模型和身份分类模型的损失函数, 并参考损失函数的梯度对深度网络进行优化. SOMAnet 网络模型利用稀疏的图像对架构获取深度网络模型具体的学习内容.

从表 5 可以看出, 基于人工设计特征的行人再识别算法利用多年积累的先验知识, 在中小型数据集上还可以取得类似基于深度学习算法的性能. 但是后者在每类行人仅有少数几幅图像 (两幅或四幅) 的情况下取得了 50 % 以上的首轮识别率, 显示出强大的特征提取优势. 在大型数据集上, 基于深度学习的方法在性能上远超传统方法, 平均准确率差距接近 30 %.

表 6 是在三个行人再识别视频数据集上取得优

表 4 常用行人再识别数据集及其参数

Table 4 Popular person re-identification datasets and their parameters

数据库名称	发布时间	图像/视频	人数	图像/视频片段数量	摄像头数量
VIPeR	2007 年	图像	632	1 264	2
CUHK01	2012 年	图像	971	3 884	2
Market-1501	2015 年	图像	1 501	32 668	6
PRID-2011[65]	2011 年	视频	200	400	2
iLIDS-VID	2014 年	视频	300	600	2
MARS[4]	2016 年	视频	1 261	20 715	6

表 5 行人再识别图像数据集上取得优异性能的方法对比

Table 5 Comparison of state-of-the-art methods on image-based person re-identification datasets

数据集	算法	人工设计/深度学习	rank-1 (%)	rank-5 (%)	rank-10 (%)	rank-20 (%)	年份
VIPeR	SCSP[66]	人工	53.5	82.6	91.5	96.6	2016 年
	FFN[50]	深度	51.1	81	91.4	96.9	2016 年
	HIPHOP[58]	深度	54.2	82.4	91.5	96.9	2017 年
CUHK01	Zhang 等[63]	人工	65	85	89.9	94.4	2016 年
	FFN	深度	55.5	78.4	83.7	92.6	2016 年
	HIPHOP	深度	78.8	92.6	95.3	97.8	2017 年
Market-1501	Zheng 等[64]	深度	85.8	94.4	96.4	97.5	2016 年
	SOMAnet[67]	深度	81.3	92.6	95.3	97.1	2017 年
	WARCA[68]	人工	45.1	68.1	76	84	2016 年

表 6 行人再识别视频数据集上取得优异性能的方法对比

Table 6 Comparison of state-of-the-art methods on video-based person re-identification datasets

数据集	算法	人工设计/深度学习	rank-1 (%)	rank-5 (%)	rank-10 (%)	rank-20 (%)	年份
PRID-2011	zhang 等[60]	深度	83.3	93.3	–	96.7	2017 年
	McLaughlin 等[45]	深度	70	90	95	97	2016 年
	TAPR[24]	人工	68.6	94.6	97.4	98.9	2016 年
iLIDS-VID	Zhang 等[60]	深度	60.2	85.1	–	94.2	2017 年
	McLaughlin 等[45]	深度	58	84	91	96	2016 年
	TAPR	人工	55	87.5	93.8	97.2	2016 年
MARS	Zhang 等[60]	深度	55.5	70.2	–	80.2	2017 年
	CNN+XQDA[4]	深度	65.3	80.2	–	89	2016 年
	LOMO+XQDA[4]	人工	30.7	46.6	–	60.9	2016 年

异性能的方法对比. TAPR 方法通过选择出视频中最符合正弦曲线的行走周期应对视频中噪声对识别结果的影响. Zhang 等[60] 首先利用流动能量轮廓法找到视频中最有代表性的 4 帧, 然后利用预训练的基于 VGG 结构的深度模型提取深度特征. McLaughlin 等[45] 利用递归神经网络捕捉视频中的时间信息. 文献 [4] 以 XQDA 作为距离度量方法, 分别采用深度 CNN 特征和人工 LOMO 特征进行行人再识别, 给出了在该数据集上的性能对比结果.

从表 6 可以看出, 利用视频进行行人再识别时, 由于每类行人都包含几十幅以上的视频帧, 能够提供足够的训练样本. 因此, 在三种规模的数据集上, 深度学习方法的性能都大大超过了传统方法. 随着数据集规模的扩大, 深度学习的优势也愈加明显.

4 总结与展望

行人再识别是当今计算机视觉领域的核心难点问题, 其解决具有重要的理论意义和良好的应用前景. 总的来说, 目前对于行人再识别尚处于研究探索阶段. 由于人体结构和外部环境的复杂性, 基于人工特征的方法在性能上还无法令人满意. 随着数据规模的不断扩大, 基于深度学习的方法展现出巨大的优势, 取得了不错的效果[69]. 虽然识别准确率在不断提高, 但是距离实用还存在一定的差距. 将来的研究工作可以从以下几方面展开:

1) 长时行人再识别. 目前大多数行人再识别算法假设行人图像或视频是在较短时间间隔内拍摄得到的, 不存在换装问题. 而在实际情况中, 不同行人图像之间拍摄时间间隔越大, 目标更换服装和随身物品的可能性就越大, 识别难度也随之加大. 因此, 长时行人再识别将是一个值得深入研究的问题.

2) 结合多模态生物线索的行人再识别. 生物线索包括人脸、步态、整体外观等信息, 具有良好的区分能力. 受限于环境条件, 目前的行人再识别方法过于依赖行人的整体外观信息, 而使用单一的生物线索很难达到理想的识别效果. 因此, 结合多模态生物线索将大大促进行人再识别.

3) 密集场景与低分辨率环境下的行人再识别. 在现实的复杂监控环境下, 行人检测框内往往包含两个甚至更多的行人, 这种样本会导致身份匹配上的混乱. 另外, 受到拍摄距离、设备分辨率等因素的影响, 部分行人图像的分辨率较低, 导致行人再识别的难度增加. 如何克服复杂环境因素的干扰仍需进一步探索.

4) 设计鲁棒的语义级行人特征表达. 目前的行人再识别性能还远无法令人满意, 其根本原因是行人特征的表达能力不足. 因此, 构建有效的图像特征空间与高层语义空间之间的映射关系, 实现对行人图像的语义级描述, 将大大提升行人特征的区分性和描述性, 在这方面还有很大的研究空间.

5) 基于深度属性的行人再识别. 基于深度学习的特征表达具有强大的数据描述能力, 并且在识别精度和泛化能力上都比传统方法更胜一筹. 在深度网络的训练过程中加入属性信息的指导, 加强神经元对于不同属性的选择性, 将有助于提高深度网络对于高层语义信息的表达能力. 目前的研究难点在于如何选择最具代表性的、具有较好语义表达能力的属性, 各个属性的组合规律也尚未定论. 另外, 属性的标注需要大量的人工成本, 导致现有数据集属性丰富性比较欠缺. 因此, 开展基于深度属性的行人再识别将极有可能产生突破性的成果, 并最终促进该领域的发展.

References

1 Porikli F. Inter-camera color calibration by correlation model function. In: Proceedings of the 2003 International Conference on Image Processing. Barcelona, Spain: IEEE, 2003. II-133-6

2 Gheissari N, Sebastian T B, Hartley R. Person reidentification using spatiotemporal appearance. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2006. 1528−1535

3 Farenzena M, Bazzani L, Perina A, Murino V, Cristani M. Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2010. 2360−2367

4 Zheng L, Bie Z, Sun Y F, Wang J D, Su C, Wang S J, et al. MARS: a video benchmark for large-scale person re-identification. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016. 868−884

5 Bazzani L, Cristani M, Murino V. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 2013, **117**(2): 130−144

6 Gray D, Tao H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer, 2008. 262−275

7 Zheng W S, Gong S G, Xiang T. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(3): 653−668

8 Liao S C, Hu Y, Zhu X Y, Li S Z. Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 2197−2206

9 Zeng M Y, Wu Z M, Tian C, Zhang L, Hu L. Efficient person re-identification by hybrid spatiogram and covariance descriptor. In: Proceedings of the 2015 IEEE Conference

on Computer Vision and Pattern Recognition Workshops. Boston, MA, USA: IEEE, 2015. 48−56

10 Ma B P, Su Y, Jurie F. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 2014, **32**(6−7): 379−390

11 Matsukawa T, Okabe T, Suzuki E, Sato Y. Hierarchical Gaussian descriptor for person re-identification. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 1363−1372

12 Zhao R, Ouyang W L, Wang X G. Person re-identification by saliency learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(2): 356−370

13 Kviatkovsky I, Adam A, Rivlin E. Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(7): 1622−1634

14 Qi Mei-Bin, Tan Sheng-Shun, Wang Yun-Xia, Liu Hao, Jiang Jian-Guo. Multi-feature subspace and kernel learning for person re-identification. *Acta Automatica Sinica*, 2016, **42**(2): 229−308
(齐美彬, 檀胜顺, 王运侠, 刘皓, 蒋建国. 基于多特征子空间与核学习的行人再识别. 自动化学报, 2016, **42**(2): 229−308)

15 Zhao R, Ouyang W L, Wang X G. Unsupervised salience learning for person re-identification. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2013. 3586−3593

16 Zhao R, Ouyang W L, Wang X G. Learning mid-level filters for person re-identification. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 144−151

17 Gong S G, Cristani M, Yan S C, Loy C C. *Person Re-Identification*. London: Springer, 2014. 139−160

18 Layne R, Hospedales T M, Gong S G. Person re-identification by attributes. In: Proceedings of the 2012 British Machine Vision Conference. Surrey, UK: BMVA Press, 2012.

19 Shi Z Y, Hospedales T M, Xiang T. Transferring a semantic representation for person re-identification and search. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 4184−4193

20 Su C, Yang F, Zhang S L, Tian Q, Davis L S, Gao W. Multi-task learning with low rank attribute embedding for person re-identification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 3739−3747

21 Caruana R A. Multitask learning: a knowledge-based source of inductive bias. In: Proceedings of the 10th International Conference on Machine Learning. Amherst, USA: Elsevier, 1993. 41−48

22 Gray D, Brennan S, Tao H. Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of the 10th International Workshop on Performance Evaluation for Tracking and Surveillance. Rio de Janeiro, Brazil: IEEE, 2007. 1−7

23 You J J, Wu A C, Li X, Zheng W S. Top-push video-based person re-identification. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 1345−1353

24 Gao C X, Wang J, Liu L Y, Yu J G, Sang N. Temporally aligned pooling representation for video-based person re-identification. In: Proceedings of the 2016 International Conference on Image Processing. Phoenix, AZ, USA: IEEE, 2016. 4284−4288

25 Wang T Q, Gong S G, Zhu X T, Wang S J. Person re-identification by video ranking. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 688−703

26 Man J, Bhanu B. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, **28**(2): 316−322

27 Klaser A, Marszałek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of the 19th British Machine Vision Conference. Leeds, UK: British Machine Vision Association, 2008, 275: 1−10

28 Bhattachayya A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin Calcutta Mathematical Society*, 1943, **35**: 99−109

29 De Maesschalck R, Jouan-Rimbaud D, Massart D L. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 2000, **50**(1): 1−18

30 Xing E P, Ng A Y, Jordan M I, Russell S J. Distance metric learning, with application to clustering with side-information. In: Proceedings of the 15th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2002. 521−528

31 Zheng W S, Gong S G, Xiang T. Person re-identification by probabilistic relative distance comparison. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO, USA: IEEE, 2011. 649−656

32 Weinberger K Q, Saul L K. Fast solvers and efficient implementations for distance metric learning. In: Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008. 1160−1167

33 Davis J V, Kulis B, Jain P, Sra S, Dhillon I S. Information-theoretic metric learning. In: Proceedings of the 24th International Conference on Machine Learning. Corvalis, Oregon, USA: ACM, 2007. 209−216

34 Guillaumin M, Verbeek J, Schmid C. Is that you? Metric learning approaches for face identification. In: Proceedings of the 12th International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 498−505

35 Köestinger M, Hirzer M, Wohlhart P, Roth P M, Bischof H. Large scale metric learning from equivalence constraints. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 2288−2295

36 Karanam S, Li Y, Radke R J. Sparse re-id: block sparsity for person re-identification. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston, MA, USA: IEEE, 2015. 33−40

37 Karanam S, Li Y, Radke R J. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 4516−4524

38 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: Curran Associates Inc., 2012. 1097−1105

39 Guan Hao, Xue Xiang-Yang, An Zhi-Yong. Advances on application of deep learning for video object tracking. *Acta Automatica Sinica*, 2016, **42**(6): 834−847
(管皓, 薛向阳, 安志勇. 深度学习在视频目标跟踪中的应用进展与展望. 自动化学报, 2016, **42**(6): 834−847)

40 Chang Liang, Deng Xiao-Ming, Zhou Ming-Quan, Wu Zhong-Ke, Yuan Ye, Yang Shuo, et al. Convolutional neural networks in image understanding. *Acta Automatica Sinica*, 2016, **42**(9): 1300−1312
(常亮, 邓小明, 周明全, 武仲科, 袁野, 杨硕, 等. 图像理解中的卷积神经网络. 自动化学报, 2016, **42**(9): 1300−1312)

41 Duan Yan-Jie, Lv Yi-Sheng, Zhang Jie, Zhao Xue-Liang, Wang Fei-Yue. Deep learning for control: the state of the art and prospects. *Acta Automatica Sinica*, 2016, **42**(5): 643−654
(段艳杰, 吕宜生, 张杰, 赵学亮, 王飞跃. 深度学习在控制领域的研究现状与展望. 自动化学报, 2016, **42**(5): 643−654)

42 Jin Lian-Wen, Zhong Zhuo-Yao, Yang Zhao, Yang Wei-Xin, Xie Ze-Cheng, Sun Jun. Applications of deep learning for handwritten Chinese character recognition: a review. *Acta Automatica Sinica*, 2016, **42**(8): 1125−1141
(金连文, 钟卓耀, 杨钊, 杨维信, 谢泽澄, 孙俊. 深度学习在手写汉字识别中的应用综述. 自动化学报, 2016, **42**(8): 1125−1141)

43 Li W, Zhao R, Xiao T, Wang X G. DeepReID: deep filter pairing neural network for person re-identification. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 152−159

44 Ahmed E, Jones M, Marks T K. An improved deep learning architecture for person re-identification. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 3908−3916

45 McLaughlin N, Martinez Del Rincon J, Miller P. Recurrent convolutional network for video-based person re-identification. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 1325−1334

46 Yan Y C, Ni B B, Song Z C, Ma C, Yan Y, Yang X K. Person re-identification via recurrent feature aggregation. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016. 701−716

47 Cheng D, Gong Y H, Zhou S P, Wang J J, Zheng N N. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 1335−1344

48 Wu S X, Chen Y C, Li X, Wu A C, You J J, Zheng W S. An enhanced deep feature representation for person re-identification. In: Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision. Lake Placid, NY, USA: IEEE, 2016. 1−8

49 Li Y J, Zhuo L, Hu X C, Zhang J. A combined feature representation of deep feature and hand-crafted features for person re-identification. In: Proceedings of the 2016 International Conference on Progress in Informatics and Computing. Shanghai, China: IEEE, 2016. 224−227

50 Chan T H, Jia K, Gao S H, Lu J W, Zeng Z N, Ma Y. PCANet: a simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, 2015, **24**(12): 5017−5032

51 Zheng L, Wang S J, Tian L, He F, Liu Z Q, Tian Q. Query-adaptive late fusion for image search and person re-identification. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 1741−1750

52 Zheng L, Huang Y J, Lu H C, Yang Y. Pose invariant embedding for deep person re-identification. arXiv preprint, arXiv: 1701.07732, 2017.

53 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 770−778

54 Zheng L, Zhang H H, Sun S Y, Chandraker M, Yang Y, Tian Q. Person re-identification in the wild. arXiv preprint, arXiv: 1604.02531, 2016.

55 Zheng L, Shen L Y, Tian L, Wang S L, Wang J D, Tian Q. Scalable person re-identification: a benchmark. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1116−1124

56 Simonyan K, Zisserma A. Very deep convolutional networks for large-scale image recognition. arXiv preprint, arXiv: 1409.1556, 2014.

57 Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv preprint, arXiv: 1703.07737, 2017.

58 Chen Y C, Zhu X T, Zheng W S, Lai J H. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(2): 392−408

59 Li W, Zhao R, Wang X G. Human reidentification with transferred metric learning. In: Proceedings of the 11th Asian Conference on Computer Vision. Daejeon, Korea: Springer, 2012. 31−44

60 Zhang W, Hu S N, Liu K. Learning compact appearance representation for video-based person re-identification. arXiv preprint, arXiv: 1702.06294, 2017.

61 Su C, Zhang S L, Xing J L, Gao W, Tian Q. Deep attributes driven multi-camera person re-identification. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016. 475−491

62 Zhu J Q, Liao S C, Yi D, Lei Z, Li S Z. Multi-label CNN based pedestrian attribute learning for soft biometrics. In: Proceedings of the 2015 International Conference on Biometrics. Phuket, Thailand: IEEE, 2015. 535−540

63 Zhang L, Xiang T, Gong S G. Learning a discriminative null space for person re-identification. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 1239−1248

64 Zheng Z D, Zheng L, Yang Y. A discriminatively learned CNN embedding for person re-identification. arXiv preprint, arXiv: 1611. 05666, 2016.

65 Hirzer M, Beleznai C, Roth P M, Bischof H. Person re-identification by descriptive and discriminative classification. In: Proceedings of the 17th Scandinavian Conference on Image Analysis. Ystad, Sweden: Springer, 2011. 91−102

66 Chen D P, Yuan Z J, Chen B D, Zheng N N. Similarity learning with spatial constraints for person re-identification. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 1268−1277

67 Barbosa I B, Cristani M, Caputo B, Rognhaugen A, Theoharis T. Looking beyond appearances: synthetic training data for deep CNNs in re-identification. arXiv preprint, arXiv: 1701.03153, 2017.

68 Jose C, Fleuret F. Scalable metric learning via weighted approximate rank component analysis. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016. 875−890

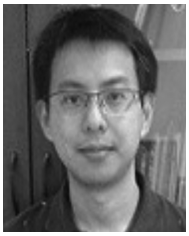
69 Yu D, Li J. Recent progresses in deep learning based acoustic models. *IEEE/CAA Journal of Automatica Sinica*, 2017, **4**(3), 396−409

李幼蛟 北京工业大学信息学部博士研究生. 山东理工大学讲师. 主要研究方向为计算机视觉, 深度学习.
E-mail: liyoujiao@emails.bjut.edu.cn
(**LI You-Jiao** Ph. D. candidate at the Faculty of Information Technology, Beijing University of Technology and lecturer at Shandong University of Technology. His research interest covers computer vision and deep learning.)

卓 力 北京工业大学教授. 1992 年获得电子科技大学无线电技术系工学学士学位, 1998 年和 2004 年分别获得东南大学信号与信息处理专业硕士学位和北京工业大学模式识别与智能系统专业博士学位. 主要研究方向为图像/视频编码和传输, 多媒体内容分析, 多媒体信息安全. 本文通信作者.
E-mail: zhuoli@bjut.edu.cn
(**ZHUO Li** Professor at Beijing University of Technology. She received her bachelor degree in radio technology from University of Electronic Science and Technology in 1992, master degree in signal and information processing from Southeast University in 1998, and Ph. D. degree in pattern recognition and intellectual system from Beijing University of Technology in 2004. Her research interest covers image/video coding and transmission, multimedia content analysis, and multimedia information security. Corresponding author of this paper.)

张 菁 北京工业大学教授. 2008 年获得北京工业大学博士学位. 美国德州大学圣安东尼奥分校计算机科学系访问学者. 主要研究方向为图像处理, 图像识别, 图像检索. E-mail: zhj@bjut.edu.cn
(**ZHANG Jing** Professor at Beijing University of Technology, visiting scholar in the Department of Computer Science, University of Texas at San Antonio, USA. She received her Ph. D. degree from Beijing University of Technology in 2008. Her research interest covers image processing, image recognition, and image retrieval.)

李嘉锋 北京工业大学信号与信息处理实验室讲师. 2009 年获得中国农业大学信息与电气工程学院学士学位, 2012 年和 2016 年获得北京航空航天大学模式识别与智能系统专业硕士学位与博士学位. 2014～2015 年美国匹兹堡大学访问学者. 主要研究方向为计算机视觉/图像增强, 图像复原.
E-mail: lijiafeng@bjut.edu.cn
(**LI Jia-Feng** Lecturer at Signal and Information Processing Laboratory, Beijing University of Technology. He received his bachelor degree from the College of Information and Electrical Engineering, China Agriculture University in 2009, master degree and Ph. D. degree in pattern recognition and intelligence system from Beihang University in 2012 and 2016. He is a visiting scholar in the Department of Neurosurgery, University of Pittsburgh, USA from 2014 to 2015. His research interest covers computer vision, image enhancement, and image restoration.)



张 辉 北京工业大学信息学部讲师. 2010 年获得北京理工大学信号与信息处理专业博士学位. 主要研究方向为计算机视觉, 机器学习在多媒体内容分析, 视觉追踪, 目标检测中的应用.
E-mail: huizhang@bjut.edu.cn
(**ZHANG Hui** Lecturer at the Faculty of Information, Beijing University of Technology. He received his Ph. D. degree in signal and information processing from Beijing Institute of Technology in 2010. His research interest covers computer vision and machine learning techniques applied to multimedia content analysis, visual tracking and object detection.)