DOI: 10.11992/tis.201706084
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20171109.1534.032.html

行人重识别研究综述

宋婉茹,赵晴晴,陈昌红,干宗良,刘峰
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要: 行人重识别是智能视频分析领域的研究热点,得到了学术界的广泛重视。行人重识别旨在非重叠视角域多摄像头网络下进行的行人匹配,即确认不同位置的摄像头在不同的时刻拍摄到的行人目标是否为同一人。本文根据研究对象的不同,将目前的研究分为基于图像的行人重识别和基于视频的行人重识别两类,对这两类分别从特征描述、度量学习和数据库集 3 个方面将现有文献分类进行了详细地总结和分析。此外,随着近年来深度学习算法的广泛应用,也带来了行人重识别在特征描述和度量学习方面算法的变革,总结了深度学习在行人重识别中的应用,并对未来发展趋势进行了展望。

关键词: 行人重识别;特征表达;度量学习;深度学习;卷积神经网络;数据集;视频监控
中图分类号: TP181 **文献标志码:** A **文章编号:** 1673−4785(2017)06−0770−11

中文引用格式: 宋婉茹, 赵晴晴, 陈昌红, 等. 行人重识别研究综述[J]. 智能系统学报, 2017, 12(6): 770–780.
英文引用格式: SONG Wanru, ZHAO Qingqing, CHEN Changhong, et al. Survey on pedestrian re-identification research[J]. CAAI transactions on intelligent systems, 2017, 12(6): 770–780.

Survey on pedestrian re-identification research

SONG Wanru, ZHAO Qingqing, CHEN Changhong, GAN Zongliang, LIU Feng
(College of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: The intelligent video analysis method based on pedestrian re-identification has become a research focus in the field of computer vision, and it has received extensive attention from the academic community. Pedestrian re-identification aims to verify pedestrian identity in image sequences captured by cameras that are orientated in different directions at different times. This current study is classified into two categories: image-based and video-based algorithms. For these two categories, using feature description, metric learning, and various benchmark datasets, detailed analysis is performed, and a summary is presented. In addition, the wide application of deep-learning algorithms in recent years has changed pedestrian re-identification in terms of feature description and metric learning. The paper summarizes the application of deep learning in pedestrian re-identification and looks at future development trends.

Keywords: pedestrian re-identification; feature representation; metric learning; deep learning; convolutional neural networks; datasets; video surveillance

在人的感知系统所获得的信息中,视觉信息大约占到 80%~85%。图像与视频等相关的应用在国民日常生活的地位日益突出。图像处理学科既是科学领域中具有挑战性的理论研究方向,也是工程领域中的重要应用技术。行人重识别(person re-identification)是近几年智能视频分析领域兴起的一项新技术,属于在复杂视频环境下的图像处理和分析范畴,是许多监控和安防应用中的主要任务[1-3],并且在计算机视觉领域获得了越来越多的关注[4-8]。

收稿日期: 2017−06−27. **网络出版日期:** 2017−11−09.
基金项目: 国家自然科学基金项目 (61471201).
通信作者: 宋婉茹. E-mail: songwanruu@163.com.

1 行人重识别概述

1.1 背景与研究意义

行人重识别是指在已有的可能来源与非重叠摄

像机视域的视频序列中识别出目标行人。以图 1 为例, 因为这些镜头是无重叠的, 所以视域完全不同, 假设我们要对在摄像头 2 中拍摄到的目标个体 1 在其他镜头中进行重识别, 需要在其他的摄像头中定位到这个目标, 除了目标本身在不同镜头下外观上的不同, 还会受到其他个体的影响, 比如在摄像头 2 中目标个体 1 需要与摄像头 1 中的 4 个目标个体都进行比较。因此研究该问题对公共安全和刑侦有着非常重要的现实意义。

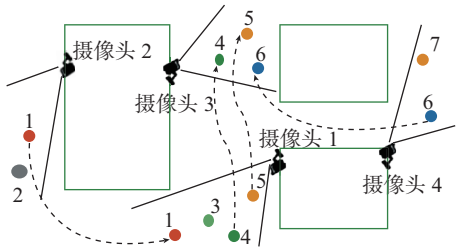

图 1　多镜头监控中的行人重识别

Fig. 1　Person Re-identification under different cameras

行人重识别的研究面临着诸如图像分辨率低、视角变化、姿态变化、光线变化以及遮挡等带来的诸多挑战。比如, 1)监控视频的画面一般比较模糊, 分辨率也比较低, 如图 2(a) 所示, 所以利用人脸识别等方式无法进行重识别的工作, 只能利用头部之外的人体外观信息进行识别, 而不同行人的体型和衣着服饰有可能相同, 这为行人重识别的准确度带来了极大的挑战; 2)行人重识别的图像往往采自于不同的摄像机, 由于拍摄场景、摄像参数不同, 行人重识别工作一般存在光照变化及视角变化等问题, 如图 2(b)、(c) 所示, 这导致同一个行人在不同摄像机下存在较大的差异, 不同行人的外貌特征可能比同一个人的外貌特征更相似; 3)进行重识别的行人图像可能拍摄于不同的时间, 行人姿态、衣着会有不同程度的改变。此外在不同的光照条件下, 行人的外观特征也会有很大的差异, 如图 2(c)。此外实际视频监控下的场景非常复杂, 很多监控场景人流量大, 场景复杂, 画面很容易出现遮挡等情况, 如图 2(d), 这种时候靠步态等特征就很难进行重识别。以上情况都给行人重识别的研究带来了巨大的挑战, 因此目前的研究距离实际应用层面还有很大的距离。

(a) 低分辨图像 (b) 视角变化　(c) 光线变化　(d) 遮挡

图 2　行人重识别的困难与挑战

Fig. 2　Difficulty and challenge to person reidentification

1.2　研究现状

相对于行人检测来说, 行人重识别的研究还不算成熟, 但早在 1996 年, 就有学者关注行人重识别问题[9], 在 2006 年, 行人重识别的概念第一次在 CVPR 上提出后[10], 相关的研究不断涌现。2007 年 Gray 提出一个对于行人重识别的研究具有重大意义的数据库 VIPeR[11]。此后越来越多的学者开始关注行人重识别的研究。近些年, 每年在国际顶级的会议以及顶级期刊上关于行人重识别的工作不在少数, 如图 3。2012 年, 第一个行人重识别研讨会在 ECCV 会议上召开; 2013 年, Gong 等[12]出版第一本行人重识别的专著; 2014 年后, 深度学习被应用到行人重识别领域; 2016 年, 行人重识别迎来井喷式的增长, 在各大计算机视觉的会议中出现了几十篇相关论文, 尤其是基于深度神经网络的方法引起了广泛的关注; 同时, 相关数据集在不断地扩充, 在各个数据集上的结果也获得很大的提升, 到目前, 行人重识别问题已成为计算机视觉的一个热点问题。

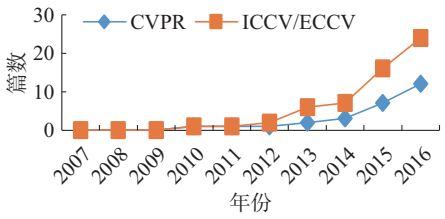

图 3　顶级会议收录行人的论文篇数

Fig. 3　Percentage of person re-ID papers on top conferences over the years

传统的行人重识别从特征提取和距离度量学习两个方面进行研究。2014 年后, 越来越多的研究者尝试将行人重识别的研究与深度学习结合在一起[13-15], 深度学习不仅应用于提取高层特征, 也为度量学习的研究带来了革新。即使深度学习在规模较小的数据集上的结果没有很明显的提升, 但随着研究方法的成熟以及较大规模的数据集的出现, 深度学习在行人重识别领域越来越受研究者们青睐。行人重识别最开始是在基于图片的情况下[13, 16-19], 即在每个数据集中每个摄像机视角下只有一幅或者几幅行人图像。但是视频相较于图像而言拥有更多信息, 并且基于视频的研究更符合视频监控环境下的现实情况, 因此我们很自然地考虑处理基于视频的行人重识别问题。从 2010 年后, 很多学者开始对基于视频的行人重识别进行研究[20-24]。我们将按照基于图像和基于视频的行人重识别研究进行介绍。

1.3　评价标准

在研究中为了评价所提出的行人重识别方法的

性能, 通常将数据库中的行人分为训练集和测试集两个部分, 在测试时, 第 1 个摄像机所拍摄的数据作为查找集, 而第 2 个摄像机中的行人数据为候选集。目前常用的评价标准主要是 CMC 曲线 (cumulated matching characteristic), 当查找的对象在候选集中进行距离比较之后, 将候选集中的行人按照距离的远近由小到大进行排序, 要查找的行人排序越靠前, 则算法的效果越好。假设总共有 N 个行人, 即共进行 N 次查询和排序, 每次查询中目标行人的排序结果用$r=(r_1,r_2,\cdots,r_N)$表示, 那么 CMC 曲线可以表示为

$$\mathrm{CMC}(R)=\frac{1}{N}\sum_{i=1}^{N}\begin{cases}1, & r_i\leqslant R\\ 0, & r_i>R\end{cases} \tag{1}$$

在近几年, Zheng 等[18]在论文中提出用平均正确率均值 (mean average precision, mAP) 来进行算法的评价标准, 指出同时使用 mAP (mean average precision) 作为评价标准能更好地比较方法的优劣, 目前已有文献[20]将 CMC 曲线和 mAP 结合作为评价标准。

2　基于图像的行人重识别研究

行人重识别算法大致可分为基于特征描述的方法和基于距离度量学习的方法两类。基于特征描述的方法关注的是找到较好的描述行人外貌特征的表观模型, 基于度量学习的方法关注的是找到有效的行人特征相似度的度量准则。下面将分别介绍这两类。

2.1　特征表达方法

基于特征表示的方法重点在于设计鲁棒可靠的行人图像特征表示模型, 即能够区分不同行人, 同时能够不受光照和视角变化的影响, 将其主要分为以下几类进行介绍, 典型特征总结见表 1。

表 1　典型特征的总结

Table 1　A summary of typical features

作者	年份	图像特征	时间信息	表征
D. Gray等[4]	2008	颜色、纹理	无	ELF (RGB, YCbCr, HSV, Gabor filters)
A.Krizhevsky等[38]	2012	CNN颜色、形状	无	CNN
Zhao R等[7]	2013	颜色	无	dColorSIFT (Dense Color, Dense SIFT)
B. Ma等[47]	2014	外观、纹理 生物激励特征	无	gBiCov(BIF, Gabor, Covariance描述符)
Xiang Li等[48]	2015	颜色、形状、纹理	无	Color, LBP, HOG
Gou M等[34]	2016	颜色、局部、纹理、轨迹	有	Color&LBP, HOG3D, DynFV
T. Matsukawa等[49]	2016	局部、形状、颜色、梯度	无	GOG(区域Gaussian分布、LAB, HSV, nRGB)
McLaughlin等[39]	2016	颜色、轨迹、CNN	有	卷积神经网(CNN)循环神经网络(RNN)

1) 底层视觉特征: 这种方法基本上都是将图像划分成多个区域, 对每个区域提取多种不同的底层视觉特征, 组合后得到鲁棒性更好的特征表示形式。最常用的就是颜色直方图, 多数情况下行人的衣服颜色结构简单, 因此颜色表示是有效的特征, 通常用 RGB、HSV 直方图表示。把 RGB 空间的图像转化成 HSL 和 YCbCr 颜色空间, 观察对数颜色空间中目标像素值的分布, 颜色特征在不同光照或角度等行人识别的不适环境中具有一定的不变性。形状特征如方向梯度直方图[14](histogram of oriented gradients, HOG)以及局部特征, 如局部不变特征–尺度不变特征变换(scale-invariant feature transform, SIFT)[15], SURF[25]和 Covariance 描述子 ELF (ensemble of localized features)方法中, 结合 RGB、YCbCr、HS 颜色空间的颜色直方图, 具有旋转不变性的 Schmid 和 Gabor 滤波器计算纹理直方图。还有纹理特征、Haar-like Represention[26]、局部二值模式(LBP)[27]、Gabor 滤波器[28]、共生矩阵(Co-occurrence Matrics)[29]。

2) 中层语义属性: 可以通过语义信息来判断两张图像中是否属于同一行人, 比如颜色、衣服以及携带的包等信息。相同的行人在不同视频拍摄下, 语义属性很少变化。Layne 等[30]采用 15 种语义来描述行人, 包括鞋子、头发颜色长短、是否携带物品等, 分类器用 SVM 定义每幅行人图像的以上语义属性。结合语义属性重要性加权以及与底层特征融合, 最终描述行人图像。Shi 等[31]对图像超像素划分, 最近分割算法对图像块定义多种特征属性, 颜色、位置和 SIFT 特征, 效果有提高。

3) 高级视觉特征: 特征的选择技术对行人再识别的识别率的性能进行提升, 如 Fisher 向量[32]编码; 提取颜色或纹理直方图, 预先定义块或条纹形状的图像区域; 或者编码区域特征描述符来建立高级视觉特征[33]。Gou 等[34]用某种描述符对密集轨迹、纹

理、直方图进行编码, 突出重要信息。受到多视角行为识别研究和 Fisher 向量编码的影响, 一种捕获软矩阵的方法, 即 DynFV(dynamic fisher vector)特征和捕获步态和移动轨迹的 Fisher 向量编码的密集短轨迹时间金字塔特征被提出。Fisher 向量编码方法是首先用来解决大尺度图像分类的方法, 也能改善行为识别的性能。Karanam 等[35]对行人的 n 幅图像的每个图像分成 6 个水平条带, 在每个条带上计算纹理和颜色直方图。在 YCbCr、HSV、白化的 RGB 颜色空间计算直方图建立颜色描述符, 并用 local fisher disrciminant analysis(LFDA)降维。Sugiyama 等[36]学习出的矩阵把特征转换到新的空间, LFDA 能在嵌入过程中使特征的局部结构适用于图像遮挡, 背景变化和光照变化的情况, 最后把计算变换空间中的特征向量的均值作为这个行人最终的特征向量表示。T. Matsukawa 等[37]提出 GOG(Gaussian Of Gaussian), 把一幅图像分成水平条带和局部块, 每个条带用一个高斯分布建模。每个条带看作一系列这样的高斯分布, 然后用一个单一的高斯分布总体表示。GOG 特征提取的方法好表现在用像素级特征的一个局部高斯分布来描述全局颜色和纹理分布, 并且 GOG 是局部颜色和纹理结构的分层模型, 可以从一个人的衣服的某些部分得到。

此外, 深度学习也被应用于行人重识别的特征提取中, 在 AlexNet-Finetune 中, 开始在 ImageNet 数据集上预训练的基于 AlexNet 结构的 CNN, 并用这个数据集对数据进行微调[38]。在微调过程中, 不修改卷积层的权重, 训练后两个全连接层。McLaughlin 等[39]采用了类似的方法, 对图像提取颜色和光流特征, 采用卷积神经网络(CNN)处理得到高层表征, 然后用循环神经网络(RNN)捕捉时间信息, 然后池化得到序列特征。T. Xiao 等[40]对来自各个领域的数据训练出同一个卷积神经网络(CNN), 有些神经元学习各个领域共享的表征, 而其他的神经元对特定的某个区域有效, 得到鲁棒的 CNN 特征表示。

2.2　度量学习方法

由于摄像机的视角、尺度、光照、服饰与姿态变化、分辨率不同以及存在遮挡, 不同摄像头间可能会失去连续的位置和运动信息, 使用欧氏距离、巴氏距离等标准的距离度量来度量行人表观特征的相似度不能获得很好的重识别效果, 因此, 研究者们提出通过度量学习的方法。该方法获得一个新的距离度量空间, 使得同一行人不同图像的特征距离小于与不同人的距离。距离度量学习方法一般是基于马氏距离(Mahalanobis distance)而进行。 2002 年, Xing 等[41]提出以马氏距离为基础的度量学习算法, 根据样本的类别标签, 将具有相同标签的样本组成正样本对, 反之组成负样本对, 并以此作为约束训练得到一个马氏矩阵, 通过这样学习到的距离尺度变换, 使得相同的人的特征距离减小, 而不同的人特征距离增大, 以此开创了行人重识别中距离度量学习的先河。

目前在行人重识别研究中有一些普遍用于比较的度量学习算法, 见表 2。Weinberger 等[42]提出 LMNN 算法, 通过学习一种距离度量, 使在一个新的转换空间中, 对于一个输入 x_i 的 k 个近邻属于相同的类别, 而不同类别的样本与 x_i 保持一定大的距离。Dikmen 等[43]对 LMNN 进行改进提出 LMNN-R 方法, 用所有样本点的平均近邻边界来代替 LMNN 中不同样本点所采用的各自近邻边界, 相较于 LMNN 方法具有更强的约束效果。同一年, Guillaumin 等[44]基于概率论提出了 LDML 算法。LDML 算法基于逻辑回归的思想, 使用 S 型函数来表示样本对是否属于等值约束的概率。Prosser 等[28]将重识别问题抽象为相对排序问题, 提出 RankSVM 学习到一个子空间, 在这个子空间中相匹配的图像有更高的排序。

表 2　行人重识别研究中常用的度量学习的方法

Table 2　A summary of metric learning

年份	作者	方法
2009	Weinberger等[42]	大间隔最近邻居(large margin nearest neighbor, LMNN)
2009	Guillaumin等[44]	逻辑判别距离度量学习(logistic discriminant metric learning, LDML)
2010	Prosser等[28]	RankSVM, 对每种特征学习一个独立的权重
2011	Zheng等[45]	概率相对距离比较(probabilistic relative distance comparison, PRDC)算法
2012	Köstinger等[27]	保持简单有效原则下的距离测度学习算法(Keep It Simple and Straightforward metric learning, KISSME
2013	Zheng等[46]	相对距离比较算法(relative distance comparison, RDC)
2013	Pedagadi等[17]	局部Fisher判别分析(local fisher discriminant analysis, LFDA)
2015	Liao等[47]	XQDA(cross-view quadratic discriminative analysis)

Zheng 等[45]提出 PRDC 算法，相同人的图像组成同类样本对，不同行人目标之间组成异类样本对，获得度量函数对应的系数矩阵，优化目标函数使得同类样本对之间的匹配距离小于异类样本对之间的距离，对每一个样本，选择一个同类样本和异类样本与其形成三元组，在训练过程通过最小化异类样本距离减去同类样本距离的和，得到满足约束的距离度量矩阵。算法的基本思想在于增加正确匹配之间会拥有较短距离的可能性。2013 年，Zheng 等[46]在 PRDC 的基础上提出了一种相对距离比较算法 RDC，RDC 采用 Adaboost 算法来减少对标注样本的需求。

Köstinger 等[27]提出 KISSME 算法，认为所有相似样本对和不相似样本对的差向量均满足一个高斯分布，因此可以通过相似和不相似训练样本对分别大致计算出均值向量和协方差矩阵。给定两个样本组成的样本对，作者分别计算该样本对属于相似样本对的概率和该样本属于不相似样本对的概率，并用其比值表示两个样本之间的距离，并把该距离变幻成马氏距离的形式，而马氏距离中的矩阵正好等于相似样本对高斯分布协方差矩阵的逆减去不相似样本对高斯分布协方差矩阵的逆。因此，该方法不要用迭代优化过程，适合用于大尺度数据的距离度量学习。

Pedagadi 等[17]提出 LFDA 算法进行度量学习，该方法在进行特征提取的时候，首先提取不同特征的主要成分，然后拼接成特征向量。在距离度量学习上，该方法考虑不是对所有样本点都给予相同的权重，考虑到了局部样本点，应用局部 Fisher 判别分析方法为降维的特征提供有识别能力的空间，提高度量学习的识别率。

Liao 等[47]提出了 XQDA 算法，这是 KISSME 算法在多场景下的推广。XQDA 算法对多场景的数据进行学习，获得原有样本的一个子空间，同时学习一个与子空间对应的距离度量函数，该距离度量函数分别用来度量同类样本和非同类样本。

此外，2015 年，Zheng 等[18]在之前研究的基础上，提出了非对称的距离度量模型 CVDCA，解决了不重叠的摄像机下的环境不同所导致特征变换不同的问题。核方法 (kernel method) 是目前机器学习领域内的研究焦点之一，引入核方法可以更好地解决行人重识别的距离度量中的非线性问题。上文中作者将核方法引入距离度量学习中，提出 KCVDCA 算法[18]，使得重识别结果有所提升。同样，LFDA 需要对高维散列矩阵进行 PCA 降维，降低了特征的表达能力，因此 Xiong 等[19]在 LFDA 的基础上同样引入核方法，提出了核局部 Fisher 判别分析 (kernel local fisher discriminant analysis, kLFDA) 算法，可避免求解高维的散列矩阵，既减少了运算量，又提高了重识别的准确率。深度学习的发展同样带来了度量方法的变革。Yi 等[48]基于孪生卷积神经网络提出了一种深度度量学习方法，取得了不错的效果。Liu 等[49]基于邻域成分分析和深度置信网络提出一种深度非线性度量学习方法。邻域变换分析的作用是通过数据变换使训练数据中每类数据的可识别样本数目最大化。为了扩展邻域变换分析中的数据变换，采用深度置信网络来学习非线性特征变换。Li 等[50]提出了一种深度学习框架来学习滤波器组，该滤波器组旨在对不同视角下的 photometric 变换进行自动编码。Ding 等[51]在损失函数和学习算法上做了改进，提出了一种基于深度神经网络的可扩展距离驱动特征学习框架，取得了不错的效果。

2.3　数据集

目前已存在很多基于图像的行人重识别库，具体见表 3。

表 3　常见的行人重识别数据集

Table 3　Common dataset in person re-identification based on image

数据库	时间	行人	图片	相机
VIPeR[30]	2007	632	1 264	2
iLIDS[52]	2009	119	476	2
GRID[53]	2009	250	1 275	8
CAVIAR[54]	2011	72	610	2
CUHK01[55]	2012	971	3 884	2
CUHK02[56]	2013	1816	7 264	10
CUHK03[50]	2014	1467	13 164	2
RAiD[57]	2014	43	1 264	4
PRID450S[58]	2014	450	900	2
Market-1501[59]	2015	1501	32 668	6

VIPeR 数据集是行人重识别中使用最为普遍的数据集，也是最具挑战性的数据集之一。VIPeR 基于图像，包含 632 个行人，1 264 幅图片，具有两个相机视角，每个相机视角下包含一个行人的一副图片。数据集中同一行人的两个相机下的成像视角差距较大，大部分在 90°以上。数据集中所有的图像都归一化到相同的分辨率 128×48。

CUHK01 也是具有较高的挑战性的数据集。该数据集包含 3 884 幅图像，971 个行人。每个行人对应从两个相机视角拍摄的 4 幅图像，每个相机 2 幅。所有图像分辨率均归一化到 160×60。

Market-1501 数据集包含 1 501 个行人, 超过 30 000 幅图像, 视频图像来源于 6 个摄像机。在大数据化的今天, 以往的行人重识别数据集规模比较小, Market-1501 的提出, 弥补了这点

3　基于视频的行人重识别研究

研究者们将行人重识别分为 single-shot 和 multi-shot 两种。single-shot 行人再识别是指每个行人在每个场景中只有一幅图像, 而 multi-shot 行人重识别主要是指每个行人在一个摄像机场景中对应一个视频或者图像序列中每个行人在每个场景有多幅图像或图像序列。与 single-shot 相比, 该类方法可利用的信息较多, 同时研究工作也更具有挑战性: 一方面, multi-shot 包含较多冗余信息, 如何提取行人图像序列的关键部分是该类问题的难点; 另一方面, 如何有效地利用行人序列特征设计度量模型, 也是该类问题需要考虑的部分。下面将介绍基于视频序列的 multi-shot 行人重识别的方法。

3.1　传统方法

由于摄像机拍摄的数据大多都是视频信息, 可以提供更多的信息帮助我们更好地进行行人匹配与再识别, 因此随着图像的深入研究, 基于视频序列的行人再识别问题也应运而生。不少方法尝试去提取视频中的三维数据来进行外貌表征, 如 HOG3D[23] 以及 3DSIFT[60]等特征都是从广泛使用的 2-D 扩展而来的。不少工作拿步态来研究基于视频的行人再识别问题[61]。然而步态的获取需要行人轮廓信息域者身体部位信息等, 而没有考虑行人的外貌信息。在遮挡较多、背景较复杂的监控环境下, 如何提取到精确的行人轮廓或身体部位信息, 仍是一个比较棘手的问题。Simonnet 等[62]提出了用动态时间弯曲距离, 对视频序列进行度量学习。Wang 等[21]提出一种基于时空描述子对行人进行重识别的方法, 融合了 HOG3D、步态能量图 (GEI)[63], 提取视频中光流强度值 (FEP) 进行步态周期检测, 进而提取出运动特征。提出通过运动能量强度, 将视频在时间上分割为不同的片段, 并在匹配的过程中通过学习的方法训练一个排序模型, 自动地选择最具判定性的片段。You 等[24]提出 top-push distance learning model (TDL), 在特征提取上融合了颜色特征、LBP 特征和 HOG3D 特征, 并通过改进了 LMNN 算法提出 TDL 算法。LMNN 的目标是缩小附近正样本间的差异, 惩罚附近所有的负样本; 而 TDL 的目标是缩小正样本间的差异, 惩罚离得最近的负样本; 所以 TDL 比 LMNN 有更强的约束。

3.2　结合深度学习方法

近些年来, 随着深度学习发展, 在基于视频的行人重识别也有所应用。以往的数据集规模不大, 因此 Zheng 等[22]建立了一个更大规模的基于视频序列的行人重识别数据集 MARS, 并用深度学习的方法在此数据集上进行实验, 获得了不错的结果。未来的研究中, 包括 MARS 在内的越来越多的大规模数据集将会作为基准数据集使用, 将深度学习的方法引入到研究中, 可以获得较好的重识别结果。在基于视频的行人重识别任务中, 数据集是由行人序列构成, 仅采用和基于图像相同的研究方法不能很好地利用数据的时间信息。然而, 由于行人重识别的数据集本身较为模糊, 具有很大的挑战性, 传统的光流、HOG3D 以及步态等提取图像运动信息的方法已经很难取得突破性进展。随着 CNN 在基于图像的任务中应用的成熟, 部分研究者把其运用到了基于视频的领域中, 此外, 为了弥补 CNN 只能处理空间维度信息的缺陷, 获取更多的时间信息, 研究者们开始将 RNN 以及其改进模型 LSTM 等用于序列建模。不同于 CNN 的深度体现在网络层数及参数规模上, RNN/LSTM 的深度主要体现在时间节点上的深度。Yan 等[64]提出提出了一种 recurrent feature aggregation network (RFA-Net), 先提取图像的颜色特征和 LBP 特征, 获得基于图像的特征, 然后与 LSTM 结合, 获得基于序列的特征, 充分利用序列数据集的信息。Mclaughlin 等[39]提出将输入的信息分为外观特征和光流信息, 将 CNN 和 RNN 网络相结合, 在 CNN 的基础上加入 RNN 使得该网络可以处理视频序列, 而在 RNN 层上加入时域池化层使得该网络可以处理任意长度的视频, 进行联合调参。Zhou 等[65]提出利用深度神经网络将特征学习和度量学习统一在一个框架下, 进行端到端的训练和推理。在特征学习阶段, 我们利用基于时序的注意模型(temporal attention model)来自动识别具有判别力的帧, 使其在特征学习阶段具有较大的权重; 度量学习阶段, 我们首先逐个位置计算一对视频片段的相似度量, 然后利用基于空间的循环神经网络模型(spatial recurrent model)来考虑空间位置的信息, 使得相似度度量融合进了上下文信息而变得鲁棒, 目前都取得了不错的效果。Liu 等[66]提出基于是累积运动上下文的视频人重识别, 采用了时间和空间分离的两路卷积网络结构, 之后将获得的表观特征和运动特征融合, 作为 RNN 的输入, 和目前现有的方法相比, 该方法的 rank-1 非常高。

3.3　数据集

现已存在不少基于视频序列的行人重识别数据库, 主要数据集见表 4 与图 4。

iLIDS-VID[21]数据集也是基于视频情况下的行

人重识别最为常用的数据集之一。该数据集包含319 个行人。每个视频序列包含 23~192 个行人图像，平均帧数为 73 帧。由于该数据集在一个机场大厅拍摄，很多行人的外观特征比较接近，两个摄像机的成像效果比较差，成像视角和光照强度都存在较大差异，每个图像中存在遮挡等不少干扰信息，因此是很有挑战性的数据集，见图 2(a)。

表 4　常见基于视频序列的行人重识别数据集

Table 4　Common dataset in person re-identification based on video

数据库	时间	行人数	相机数
ETHZ[67]	2007	85	1
3DPES[68]	2011	192	8
PRID2011[69]	2011	200	2
iLIDS-VID[21]	2014	300	2
MARS[22]	2016	1 261	6

(a) 来自 iLIDS-VID

(b) 来自 PRID2011

图 4　不同摄像机下的行人

Fig. 4　Sample person under different cameras

PRID2011[69]数据集也是基于视频的情况下行人重识别最为常用的数据集之一。该数据集由两个摄像机拍摄，cam_a 视角下有 385 组行人序列，cam_b 视角下有 749 组行人序列，其中两个视角下有 200 个行人相同，每个视频序列包含 5-675 帧图像，平均帧数 100。与 iLIDS-VID 不同的是，该数据集的背景比较干净，图像中较少存在遮挡这种干扰信息，图像的成像效果比较好。和 iLIDS-VID 类似，两个摄像机成像视角和光照强度也存在很大的差异，见图 2(b)。

随着深度学习在行人重识别中的应用，小规模的数据集逐渐难以满足需求，因此近些年，在基于视频序列的行人重识别研究中，也有大规模的数据集提出，如 MARS[22]。

4　发展趋势

由于智能监控系统在国防建设、人民日常生活中的巨大应用前景，以及其所涉及的领域广泛性、研究的巨大挑战性，因此国内外很多研究者对该研究方向越来越重视。同时行人再识别问题也是很多知名的学术会议和国际期刊的重点研究方向之一，例如 2016 年，在 CVPR 上有关于行人重识别的文章就高达 12 篇。由于不断对方法进行革新，行人重识别的研究在各大数据集上都取得了不错的进展。

在基于图像的行人重识别研究中，VIPeR 作为最广泛被采用的数据集，rank-1 的准确率从 2008 年的 12.0%[4]提高了 2015 年的 63.9%[70]；同时，CUHK01 上的 rank-1 自 2010—2016 年，也取得了 56.7% 的提升。由于这些数据集的规模都不大，因此，即使使用了深度学习的方法，依然和手工设计出的特征以及度量方法取得的最好结果近似。但是在 Market-1501 上，深度学习的应用明显提高了 rank-1 的准确率，从 2015 年该数据集刚开始应用到行人重识别的研究中时，rank-1 的准确率从 44.42%[20]提高到了 2016 年的 76.04%[71]。

基于视频的行人重识别研究起步相较图像稍晚一点，但是近几年来引起了很大的重视。早期的 ETHZ 数据集由于情况简单，相对 iLIDS-VID 来说，情况复杂了很多，但 rank-1 准确率从 2014 年的 23.3%[21]到在 2016 年 McLaughlin 等[39]提出的方法，可达到 58%，在 2017 年的出现的文章，有研究者提出基于是累积运动上下文以及联合 CNN、RNN 的 AMOC 方法[66]，rank-1 可以达到 68.7%，具体结果可见图 5；同样 Zheng 等[22]利用对从数据集 MARS 上获得的 CNN 特征进行微调运用到 PRID2011 上，使得其 rank-1 准确率可以达到 77.3%。MARS 数据集被提出，rank-1 准确率可达到 68.3%，同时，作者提出了另外一钟补充评价标准 mAP。2017 年，Zhou 等[65]提出利用深度神经网络将特征学习和度量学习统一在一个框架下的方法，在 iLIDS-VID、PRID2011 以及 MARS 上的 rank-1 准确率分别达到了 55.2%、79.4% 以及 70.6%，在 MARS 上的 mAP 也有所提高。

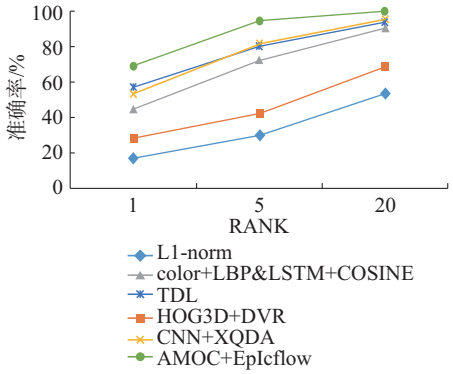

图 5　几种重要方法在数据集 iLIDS-VID 上的结果对比

Fig. 5　Person re-ID accuracy on iLIDS-VID by several principal methods

5　结束语

我们看出行人重识别的研究取得了一定的成果, 研究日益趋向成熟。但是也可以看出, 时至今日, 行人重识别的研究依然很难实现很好的结果, 对于环境恶劣的数据集, rank-1 准确率以及 mAP 均不高, 距离实际应用有更长的路要走。 因此本文通过对已存在的行人重识别的方法进行总结与比较, 对常用数据集进行研究, 我们认为: 1) 大规模行人视频数据库较少。有学者提出深度学习与传统模式识别方法的最大不同在于它所采用的特征是从大数据中自动学习得到, 而非采用手工设计。深度学习可以从大数据中自动学习特征的表示, 可以包含成千上万的参数。采用手工设计出有效的特征往往需要 5~10 年时间, 而深度学习可以针对新的应用从训练数据中很快学习到新的有效的特征。然而在 VIPeR 等常用数据集上, 因为规模限制, 即使结合深度学习等方法, 目前来说没有相较于传统方法有较大突破。为了更好地结合 CNN、RNN 等方法, 在今后的发展中大规模的数据集将会成为研究者的研究重点, 另外更多的有实际研究价值的大规模数据集会被提出, 适应研究发展的需要。2) 在新技术的应用方面还非常不足。虽然引入了深度学习进行特征提取或分类, 但多集中于深度判别式学习, 而很少用到深度生成式模型。生成式模型的目的是找到一个函数可以最大的近似数据的真实分布。如果我们用 $f(X;\theta)$ 来表示这样一个函数, 那么找到一个使生成的数据最像真实数据的 θ 就是一个最大化概率估计的过程。深度网络结构可以表达这样一个复杂的函数, 含有隐变量单元的生成式模型是使得模型更好地理解由训练数据所决定的真实的世界的有效方式。DeepMind 研究员们最近在 arXiv 上传了一篇论文[71], 提出了一种新的深度学习模型——记忆生成时序(generative temporal models with memory, GTMM), 对广泛使用于语音识别、图像识别、语义理解等领域的循环神经网络(RNN)性能带来了显著提升。该模型是在变分推理框架下开发的, 提供了实用训练方法和深入了解模型运作的方法, 模型从序列的早期阶段开始存储信息, 对不可预测的元素标示不确定性, 并能有效地对已存储的信息进行再利用。对于行人重识别来讲, 本身行人序列是时序的, 但是由于视角、服饰、姿态、分辨率、遮挡、光线等诸多因素的影响, 又有很多不确定因素, 而且不能保证数据的充足性, 这些问题采用 GTMM 模型都可以很好地解决。因此可以构建记忆生成时序模型 GTMM 对行人序列进行建模和再识别。

参考文献:

[1] LI Y, WU Z, KARANAM S, et al. Real-world re-identification in an airport camera network[C]//International Conference on Distributed Smart Cameras. Venice, Italy, 2014: 35.

[2] GONG S, CRISTANI M, YAN S, et al. Person re-identification [M]. London, UK: Springer, 2014.

[3] CAMPS O, GOU M, HEBBLE T, et al. From the lab to the real world: Re-identification in an airport camera network [J]. IEEE transactions on circuits and systems for video technology, 2016, (99): 540–553.

[4] GRAY D, TAO H. Viewpoint invariant pedestrian recognition with an ensemble of localized features[C]//European Conference on Computer Vision. Marseill, France, 2008: 262–275.

[5] PROSSER B, ZHENG W S, GONG S, et al. Person re-identification by support vector ranking[C]//The British Machine Vision Conference. Aberystwyth, British, 2010: 1–21.

[6] JURIE F, MIGNON A. PCCA: a new approach for distance learning from sparse pairwise constraints[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2012: 2666–2672.

[7] ZHAO R, OUYANG W, WANG X. Unsupervised salience learning for person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Oregon, USA, 2013: 3586–3593.

[8] ZHENG W S, LI X, XIANG T. Partial person re-identification[C]//IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4678–4686.

[9] CAI Q, AGGARWAL J K. Tracking human motion using multiple cameras[C]//International Conference on Pattern Recognition. Vienna, Austria, 1996: 68–72.

[10] GHEISSARI N, SEBASTIAN T B, HARTLEY R. Person re-identification using spatiotemporal appearance[C]//IEEE Conference on Computer Vision and Pattern Recog-

nition. New York, USA, 2006: 1528–1535.

[11] GRAY D, BRENNAN S, TAO H. Evaluating appearance models for recognition, reacquisition, and tracking[J]. International journal of computer vision, 2007, 89(2): 56–68.

[12] GONG S G, CRISTANI M, YAN S C, et al. Person re-identification[J]. Advances in computer vision and pattern recognition, 2013, 42(7): 301–313.

[13] YI D, LEI Z, LI S Z, Deep metric learning for practical person re-identification[C]//International Conference on Pattern Recognition. Stockholm Waterfront, Sweden, 2014.

[14] OREOFEJ O, MEHRAN R, SHAH M. Human identity recognition in aerial images[C]//IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 709–716.

[15] JUNGLING K, BODENSTEINER C, ARENS M. Person re-identification in multi-camera networks[C]//Computer Vision and Pattern Recognition Workshops. Colorado, USA, 2010: 709–716.

[16] ZHENG W S, GONG S G, XIANG T. Re identification by relative distance comparison[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(3): 653.

[17] PEDAGADI S, ORWELL J, VELASTIN S, et al. Local fisher discriminant analysis for pedestrian re-identification [C]//IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 3318–3325.

[18] CHEN Y C, ZHENG W S, LAI J H, et al. An asymmetric distance model for cross-view feature mapping in person re-identification[J]. IEEE transactions on circuits and systems for video technology, 2016(99): 1661–1675.

[19] XIONG F, GOU M, CAMPS O, et al. Person re-Identification using kernel-based metric learning methods[C]// European Conference on Computer Vision. Zurich, Switzerland, 2014:1–16.

[20] ZHENG L, SHEN L, TIAN L, et al. Scalable person re-identification: a benchmark[C]//IEEE International Conference on Computer Vision. Santiago, Chile, 2015; 1116–1124.

[21] WANG T, GONG S G, ZHU X, et al. Person re-identification by video ranking[C]//European Conference on Computer Vision. Zurich, Switzerland, 2014: 688–703.

[22] ZHENG L, BIE Z, SUN Y, et al. MARS: A video benchmark for large-scale person re-identification[M]//European Conference on Computer Vision. Springer International Publishing, 2016: 868–884.

[23] KLASER A, MARSZALEK M, SCHMID C. A spatio-temporal descriptor based on 3D-gradients[C]//British Machine Vision Conference 2008. Nottingham, British, 2008: 152–159.

[24] YOU J, WU A, LI X, et al. Top-push video-based person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las vegas, USA, 2016: 1345–1353.

[25] ZHAO R, OUYANG W, WANG X R. Unsupervised salience learning for person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 3586–3593.

[26] BAK S, CORVEE E, BREMOND F, et al. Person re-identification using haar-based and DCD-based signature[C]// IEEE International Conference on Advanced Video and Signal Based Surveillance. Boston, USA, 2010: 1–8.

[27] KOESTINGER M, HIRZER M, WOHLHART P, et al. Large scale metric learning from equivalence constraint [C]//IEEE Conference on Computer Vision and Pattern Recognition. Providence, Rhode island, 2012: 2288–2295.

[28] ENGEL C, BAUMGARTNE P, HOLZMANN M, et al. Person re-identification by support vector ranking[C]// British Machine Vision Conference 2010. Aberystwyth, UK, 2010: 1–11.

[29] SCHWARTZ W R, DAVIS L S. Learning discriminative appearance-based models using partial least squares[C]// XXII Brazilian Symposium on Computer Graphics and Image Processing. Gramado, Brazil, 2010: 322–329.

[30] LAYNE R, HOSPEDALES T M, GONG S G. Person Re-identification by Attributes[C]//The British Machine Vision Conference. Nottingham, Park, 2014, 2(3): 8.

[31] SHI Z, HOSPEDALSE T M, XIANG T. Transferring a semantic representation for person re-identification and search[C]//Computer Vision and Pattern Recognition. Boston, USA, 2015: 4184–4193.

[32] MA B, SU Y, JURIE F. Local descriptors encoded by fisher vectors for person re-identification[C]//International Conference on Computer Vision. Barcelona, Spain, 2012: 413–422.

[33] CHEN D, YUAN Z, HUA G, et al. Similarity learning on an explicit polynomial kernel feature map for person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1565–1573.

[34] GOU M, ZHANG X, RATES-BORRAS A, et al. Person re-identification in appearance impaired scenarios[C]// British Machine Vision Conference. [S.l.], 2016: 1–48.

[35] KARANAM S, LI Y, RADKE R J. Person re-identification with discriminatively trained viewpoint invariant dictionaries[C]//IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4516–4524.

[36] SUGIYAMA, MASASHI. Local fisher discriminant analysis for supervised dimensionality reduction[J]. Machine learning, 2010, 78(1/2): 35–61.

[37] MATSUKAWA T, OKABE T, SUZUKI E, et al. Hierarch-

ical gaussian descriptor for person re-identification[C]// IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1363–1372.

[38] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//International Conference on Neural Information Processing Systems. Doha, Qatar, 2012: 1097–1105.

[39] MCLAUGHLIN N, RINCON J M, MILLER P. Recurrent Convolutional Network for Video-based Person Re-Identification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2012: 51–58.

[40] XIAO T, LI H, OUYANG W, et al. Learning deep feature representations with domain guided dropout for person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1249–1258.

[41] XING E P, NG A Y, JORDAN M I, et al. Distance metric learning, with application to clustering with side-information[C]//International Conference on Neural Information Processing Systems. Vancouver: MIT Press, 2002: 521–528.

[42] WEINBERGER K Q, SAUL K L. Distance metric learning for large margin nearest neighbor classification[J]. Journal of machine learning research, 2009, 10(1): 207–244.

[43] DIKMEN M, AKBAS E, HUANG T S, et al. Pedestrian recognition with a learned metric[J]. Lecture notes in computer science, 2010, 6495: 501–512.

[44] GUILLAUMIN M, VERBEEK J, SCHMID C. Is that you? Metric learning approaches for face identification[C]// Proceedings of the 12th International Conference on Computer Vision. Kyoto, Japan, 2009: 498–505.

[45] ZHENG W, GONG S, XIANG T. Person re-identification by probabilistic relative distance comparison[C]// IEEE conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 649–656.

[46] ZHENG W S, GONG S, XIANG T. Re-identification by relative distance comparison[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(3): 653.

[47] LIAO S, HU Y, ZHU X, et al. Person re-identification by local maximal occurrence representation and metric learning[C]//IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2197–2206.

[48] YI D, LEI Z, LI S Z. Deep metric learning for practical person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 34–39.

[49] LIU H, MA B, QIN L, et al. Set-label modeling and deep metric learning on person re-identification[J]//Neurocomputing, 2015 (151): 1283–1292.

[50] LI W, ZHAO R, XIAO T, et al. Deepreid: Deep filter pairing neural network for person re-identification[C]// IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 152–159.

[51] 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.

[52] ZHENG W S, GONG S, XIANG T. Associating groups of people[C]//Proceedings of the British Machine Vision Conference. London, UK, 2009: 251–259.

[53] CHEN C L, XIANG T, GONG S. Multi-camera activity correlation analysis[C]//IEEE conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1988–1995.

[54] DONG S C, CRISTANI M, STOPPA M, et al. Custom pictorial structures for re-identification[C]//British Machine Vision Conference. Dundee, British. 2011: 159–165.

[55] LI W, ZHAO R, WANG X. Human re-identification with transferred metric learning[C]//Asian Conference on Computer Vision. Daejeon, Korea, Springer-Verlag, 2012: 31–44.

[56] LI W, WANG X. Locally aligned feature transforms across views[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 3594–3601.

[57] LI W, ZHAO R, XIAO T, et al. DeepReID: deep filter pairing neural network for person re-identification[C]// IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 152–159.

[58] DAS A, CHAKRABORTY A, ROY-CHOWDHURY A K. Consistent re-identification in a camera network[C]// European Conference on Computer Vision. Springer International Publishing, 2014: 330–345.

[59] ROTH P M, HIRZER M, KOSTINGER M, et al. Mahalanobis distance learning for person re-identification[M]. London: Person re-identification, 2014: 247–267.

[60] SCOVANNER P, ALI S, SHAH M. A 3-dimensional sift descriptor and its application to action recognition[C]// 15th ACM International Conference on Multimedia. New York, USA, 2007: 357–360.

[61] BEDAGKAR-GALA A, SHAH S K. Gait-assisted person re-identification in wide area surveillance[C]//Asian Conference on Computer Vision. Singapore: Springer International Publishing, 2014: 633–649.

[62] SIMONNET D, LEWANDOWSKI M, VELASTIN S A, et al. Re-identification of pedestrians in crowds using dynamic time warping[C]//International Conference on Computer Vision. Springer-Verlag, 2012: 423–432.