doi:10.3969/j.issn.1001-4616.2020.02.019

视频行人重识别研究进展

李梦静,吉根林

(南京师范大学 计算机科学与技术学院,江苏 南京 210023)

[**摘要**] 视频行人重识别是指在不同摄像头拍摄的视频中检索特定行人的技术. 与图像行人重识别相比,视频行人重识别赋含信息更多,包含了帧与帧之间的时间信息、运动信息等,这有利于提高行人检索的准确率,因此视频行人重识别引起了国内外学者的广泛关注. 本文探讨了视频行人重识别的处理过程,详细介绍了其中特征提取和距离度量的方法,并对各种特征提取方法的特点及应用进行了总结,给出了一些视频行人重识别实验数据集和评价标准,提出了视频行人重识别研究领域面临的挑战及相应的解决方案,最后对视频行人重识别技术未来的研究问题做了展望.

[**关键词**] 行人重识别,视频行人重识别,视频分析,计算机视觉

[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2020)02-0120-11

Research Progress on Video-based Person Re-Identification

Li Mengjing,Ji Genlin

(School of Computer Science and Technology,Nanjing Normal University,Nanjing 210023,China)

Abstract:Video-based person re-identification is a technique for retrieving specific pedestrians from video captured by different cameras. Compared with image-based person re-identification, video-based person re-identification has more information,including time information and motion information between frames,which is more conducive to improving the accuracy of pedestrian retrieval,so it has attracted widespread attention from scholars at home and abroad. This paper discusses the process of video-based person re-identification,introduces the methods of feature extraction and distance metric in detail,summarizes the characteristics and applications of various feature extraction methods,and some video-based person re-identification experimental data sets and evaluation standards are also given,what's more,the challenges and corresponding solutions in the field of video person re-identification are presented. Finally an outlook to the future research problems of video person re-identification technology is given.

Key words:person re-identification,video-based person re-identification,video analysis,computer vision

视频行人重识别(又叫行人再识别)是指在不同摄像头拍摄的行人视频中检索特定行人的技术,即给定一个行人视频,跨摄像头检索该行人,得到其在其它摄像头下的视频. 这种针对特定人的视频检索具有重要的研究意义,在失踪者定位、犯罪跟踪和行人视频检索等方面有着广泛的应用.

行人重识别问题的研究最早可以追溯到跨摄像头多目标跟踪问题上,2005 年,文献[1]提出了当目标行人在某个摄像头视频丢失之后,如何将其在其它摄像头视频中再次匹配的问题. 2006 年,文献[2]第一次提出了行人重识别的概念,将其从跨摄像头多目标跟踪问题中抽离出来,作为一个独立的问题进行研究. 早期的行人重识别研究使用传统方法,例如提取手工设计的特征. 2014 年以后,深度学习得到了迅猛发展,学者们试图将深度学习技术应用在了行人重识别领域,获得了更好的效果.

根据输入数据的不同,行人重识别可以分为图像行人重识别和视频行人重识别,这两者既有相同点也有不同点,相同点是它们都面临着摄像头本身的低分辨率、拍摄场景的多样性、物体遮挡、光照变化等等问题带来的挑战. 不同点是相比于图像,视频数据中蕴含的信息更多,视频中包含了行人的运动信息和时间

收稿日期:2019-12-17.

基金项目:国家自然科学基金资助项目(41971343).

通讯作者:吉根林,博士,教授,博士生导师,研究方向:大数据分析与挖掘技术. E-mail:glji@njnu.edu.cn

信息,数据量更多、计算量更大,更复杂,且视频数据存在高度冗余,如何提取具有鉴别力的部分也更值得研究. 与图像行人重识别相比,视频行人重识别的研究工作较少,但是视频行人重识别更接近真实应用,自 2016 年以后,越来越多的学者开始关注视频行人重识别问题,提出了各种不同的解决方法.

1 视频行人重识别的处理过程

一般来说,视频行人重识别问题处理过程主要分为 3 个阶段:(1)视频数据预处理:将视频按帧切分成图像序列,利用行人检测技术得到行人检测框,并处理图像噪声、光照变化等问题;(2)特征提取:提取描述行人外观的有区别的、稳定的特征;(3)距离度量:找到更有效的行人相似性度量方法,建立一个新的特征空间,使相同行人的特征距离更小,不同行人的特征距离更大.

视频行人重识别传统处理过程如图 1 所示. 训练视频首先经过预处理,再进行特征提取和距离度量,最后通过损失函数反馈训练,不断迭代,直到获得较好的特征学习模型. 在检索过程中,检索视频和多个候选视频分别进行预处理,然后输入到已训练完成的特征提取模型中,得到每个视频的特征,再利用合适的距离度量方法计算检索视频与每个候选视频之间的距离. 最后按照距离的大小,输出所有候选视频的排序结果.

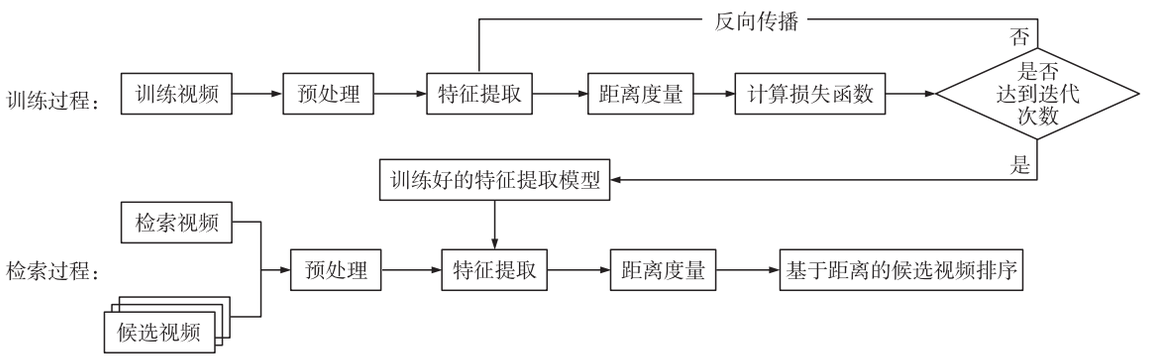

图 1 视频行人重识别的处理过程

Fig. 1 The process of video-based person re-identification

2 特征提取方法

特征提取方法主要分为基于手工设计的特征提取和基于深度学习的特征提取. 如表 1 所示,基于手工设计的特征主要有颜色特征[3-4]、光流特征和属性特征[5],其中前两者属于视觉特征,容易被周围环境干扰,而属性特征属于中层语义特征,更具有鲁棒性. 随着深度学习的发展,2016 年以后使用深度学习模型提取特征成为主流.

表 1 行人特征分类

Table 1 Pedestrian feature classification

行人特征	基于手工设计的特征	颜色特征:颜色直方图、颜色聚合向量 光流特征:图像序列中像素的变化 属性特征:长发/短发;有无背包;身高;衣服颜色等等
	基于深度学习的特征	时空特征:获取帧级空间特征和帧间时间特征 局部特征:考虑行人不同部位外观之间的关系和重要性

基于深度学习的特征主要有:(1)时空特征,视频中含有丰富的时空信息. 空间信息指图像的每一帧不同位置具有的特征,时间信息指视频不同帧之间的联系,空间信息与时间信息是互补的. 如果缺少信息的任何一部分,行人的信息就不能得到充分的表达. (2)局部特征,早期特征提取模型都是一幅图像提取一个特征,不考虑一些局部信息,随着行人重识别数据集越来越复杂,全局特征并不能得到很好的效果,提取更加复杂的局部特征成为一个新的解决方法. 详细的特征提取方法特点及应用整理如表 2 所示.

2.1 时空特征提取

提取时空特征的方法可以分为 3 类:(1)额外给 CNN(Convolutional Neural Network)输入光流等动态光流特征[6-8];(2)先提取帧级空间特征,再将所有帧特征输入到循环神经网络中提取时间特征[9-11]或是

利用时间汇聚或者权重学习得到时间特征[12-14];(3)将视频看作三维数据,通过 3D CNN 等方法提取时空特征[15-16].

2017 年,Zhou 等[13]利用 CNN 网络提取每帧图像的空间特征,然后提出了一种时态注意模型 TAM(Temporal Attention Model)来提取时间特征. 2019 年,Chen 等[10]提出了一种联合关注时空特征聚合网络(Joint Attentive spatial-temporal Feature aggregation Network,JAFN),同时学习质量感知模型和框架感知模型,利用 CNN 学习空间特征,同时引入 LSTM(Long-Short Time Memory)学习时间特征.

上面两种方法都是先提取帧级空间特征,再利用 LSTM 网络或者注意力机制生成视频级的时间特征,都是直接在高层特征上建立时间连接,因此无法捕获图像局部细节上的时间信息. 为了解决这个问题,Li 等[15]受 3D 卷积神经网络在动作识别领域得到的成功所启发,提出了双流 M3D(Multi-Scale 3D)卷积神经网络将多个多尺度三维卷积层插入到二维 CNN 网络中,同时提取时间和空间信息.

2.2　局部特征提取

局部特征是特征提取模型自动关注视频的某些局部区域. 相比于全局特征,局部特征对光照变化的鲁棒性更强,且能减弱遮挡的影响. 提取局部特征的主要方法有:(1)人工将图像划分成块,根据人体固有的特点,将图像划分成头、上身、下身几个部分;(2)提取人体骨架关键点,利用注意力机制关注局部身体部位处的特征;(3)构建人体部位特征邻接图,对不同部位特征之间的关系进行建模.

Liu 等[17]提出一种新的时空特征混合模型,首先将人体水平分割成 N 个部分,包括头部、腰部、腿部等信息. 然后整合每个部分的特征,以实现对每个人更有鉴别力的表达. 2019 年,文献[18]认为行人身份信息主要表现在躯干、肘部、手腕、膝盖、脚踝等身体部位. 首先检测人体关键点,然后获取行人图像中人体关节的重要系数矩阵,根据系数矩阵通过注意力机制整合 CNN 得到的图像外观特征.

这两种方法忽略了人体各部位之间的相关性,而人的各个部位之间的关系有助于降低复杂情况(如遮挡、不对齐和杂乱背景)的影响. 为了利用各部位之间的关系,Wu 等[19]提出了一种新颖的自适应图表示学习方案,首先利用位姿对齐连接和特征关联连接来构造一个自适应结构感知邻接图,利用该邻接图对图节点间的内在关系进行建模. 自适应地捕获行人身体部位特征之间的内在关联结构信息,并进一步传递互补的上下文信息,丰富行人外观特征表示.

表 2　特征提取方法特点及应用

Table 2　Characteristics and application of feature extraction methods

行人特征	特征提取方法	特点	应用实例(见文献)
属性特征	识别人体部位再进行属性分类	相比于底层特征具有一定的鲁棒性	[5],[20],[21]
时空特征	额外添加手动设计的动态光流特征	运动特征单一、受光线变化影响较大	[7],[22],[23],[24]
	先提取帧级空间特征再利用循环神经网络或权重学习提取时间特征	可以获得帧之间的联系,且能获得不同帧的重要性,但是只能获取图像帧全局的时间信息	[10],[21],[25],[26],[27],[28],[29],[30],[31]
	利用 3D 卷积同时提取时间特征和空间特征	可以获得图像局部之间的时间连续性,但模型复杂,计算量大	[16],[23],[32],[33],[34],[35],[36],[37],[38]
局部特征	手动划分成不同的部分	简单易实现,对行人检测准确率要求高	[21]
	生成注意力图,自动关注图像中“重要”的部分	去除背景的干扰,关注行人的部分	[14],[39],[40],[41]
	检测人体关键点,利用注意力机制给予不同部位特征相应的权重	将局部特征更加细化并且能获得不同部位特征的重要性	[14],[18],[42]
	构建人体部位特征邻接图,对不同部位特征之间的关系进行建模	考虑了人体各部位特征之间的关系,更加丰富行人的外观特征表示	[19],[43],[44],[45]

2.3　特征融合

特征融合分为两类:(1)多特征融合:对同一个视频提取多种特征,例如颜色、光流、时空等等,然后将多种特征融合作为最终行人特征;(2)多帧融合:视频本质上属于图像序列,常规方法是对每一帧提取一个特征,再将多帧的特征融合,得到视频级特征.

多特征融合:早期的特征提取采用低级特征,2016 年,文献[7]利用颜色和光流信息来捕获图像和运动信息,Zheng 等[46]将颜色直方图与 SURF(Speeded Up Robust Feature)特征相结合提取特征,但随着深度学习在计算机视觉领域取得巨大成功,研究者们开始使用深度学习来学习更具鉴别力的特征. 将手工特

征与深度学习得到的特征相融合,2017 年,Li 等[47]将手工设计的局部特征与 PCN(PCA-based Convolutional Network)生成的深度特征进行融合. 2018 年,Sun 等[22]利用孪生神经网络学习深度特征,将光流特征与深度特征进行融合作为行人特征描述符.

多帧融合:从帧级特征到视频级特征,必然的方法就是进行多帧特征融合,最简单的方法是将所有帧的重要性视为相同,将所有帧级特征进行平均池化得到视频级特征,但是显然因为光照、遮挡等因素的影响,不同的帧等提供的有用信息是不一样的. 基于这样一个事实,文献[48]提出一种时域注意方法,其中采用一个全卷积的时间注意模型来生成注意力分数,它代表每一帧的重要性. Ouyang 等[49]利用深度强化学习剔除质量差、误导和混淆的帧,从视频中挑选出具有代表性的帧,再进行多帧融合. 文献[50]提出的视频协方差方法还研究了视频帧之间的相关性,利用相关性对多帧特征进行整合,提升了特征表示的鉴别力.

3 距离度量方法

距离度量阶段的任务是定义一个距离度量函数计算两个特征向量之间的距离,通过最小化网络的度量损失,得到一个最优的特征空间,使得相同行人的视频之间的距离尽可能小,不同行人视频之间的距离尽可能大. 传统的距离度量学习方法有 LMNN[51]、KISSME[52]、XQDA[53]、LFDA[54].

行人视频是由不同相机拍摄的,光照变化、视角变化都会使视频间产生较大的差异,可以观察到:(1)同一个人的不同视频之间存在严重差异;(2)每个视频内的不同帧或步行周期之间也存在较大差异. 两种差异都会对行人视频之间的匹配带来不利影响. 2018 年,Zhu 等[55]提出了一种同步的视频内和视频间的距离度量学习方法 SI^{2}DL. 如图 2 所示,该方法首先将单个视频内距离拉近,然后拉近类内距离,同时推远类间距离,从而使得不同行人的视频分开.

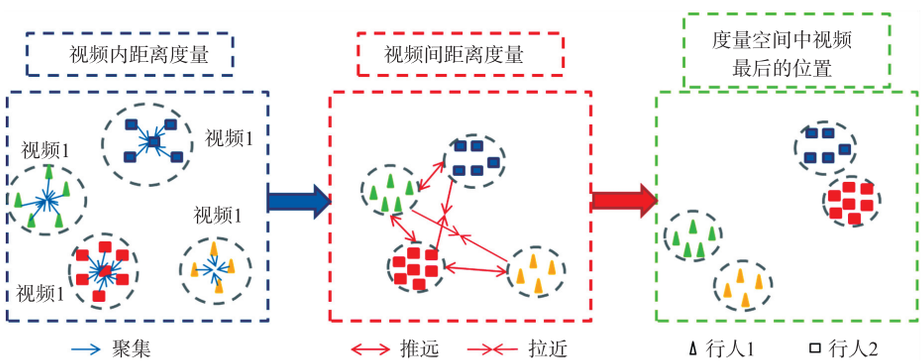

图 2 SI^{2}DL 方法说明

Fig. 2 The Illustration of SI^{2}DL

相似地,为了减少视频内特征的差异,2019 年,Zhang 等[56]引入“均值-体”的概念,定义一个视频内的损失,以解决同一视频时空特征之间的变化,然后结合视频内损失和孪生损失来提高训练速度.

在实际应用时,视频数据是一个流式数据,图像帧是不断加入原有的数据中的,这就要求损失函数不仅能够计算最终特征向量之间的距离,还要在新的数据加入时,对原有距离进行更新. 2019 年,Navaneet 等[57]提出了排名损失. 它可以在新数据加入时,确保距离度量的质量在不断改善,并防止由于质量差的帧加入而导致的退化.

4 视频行人重识别研究面临的挑战

虽然近几年视频行人重识别取得了重大的发展,但是还是面临着许多挑战,在真实场景下,行人重识别问题会遇到跨摄像头导致的姿态变化、遮挡、光照变化等问题.

4.1 行人不对齐且姿态变化

行人重识别数据集中普遍存在的一个关键问题就是图像对之间的不对齐,如图 3 所示,由于背景杂波和位置不对齐,直接比较未对齐的图像对效果非常差. 另一个问题就是姿态的变化,如图 4 所示,人体的姿态总是根据相机的不同视角、行走路径、行为等发生变化,这个问题显著降低模型的性能.

图 3　对齐和不对齐的示例

Fig. 3　Examples of alignment and misalignment

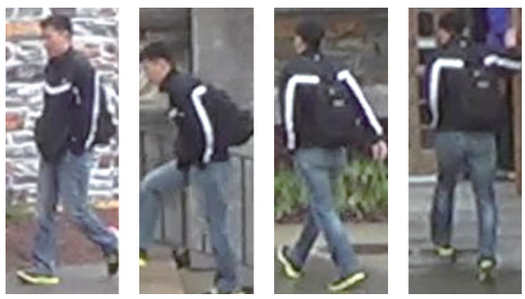

图 4　姿态变化示例

Fig. 4　Examples of pose variation

针对这两个问题,近几年,姿态估计对齐方法[58-59]得到了广泛的应用. Chen 等[60]针对图像不对齐的问题提出了基于位置引导的空间转换子网络 STSN(pose-guided Spatial Transformer Sub-Network). STSN 首先对输入图像的各种转换参数进行回归,然后使用仿射变换将图像转换为对齐的图像. 针对姿态变化的问题,提出了一种新的训练策略,称为关键帧选择,它可以选择具有最大转换贡献值的帧作为关键帧,然后用这些关键帧来训练网络,从而减少姿态变化的影响.

上述方法的缺点是需要额外的姿态标注. Wu 等[61]提出了一种半监督的方法,将训练好的姿态估计模型直接应用到行人重识别数据集上,避免了在行人重识别数据集上标注姿态的麻烦.

4.2　遮挡问题

现实场景中,行人的任何部分都可能被其他行人或环境物体(如车和指示牌)遮挡,如图 5 所示,这会导致行人外观的巨大变化.

图 5　遮挡示例

Fig. 5　Examples of occlusion

相比于基于图像的行人重识别,视频行人重识别已经弱化了遮挡的影响,因为一般来说视频中只有一部分图像帧存在遮挡问题. 但这显然不能得到很好的效果. 后来,学者们提出了基于注意力机制的方法,2017 年,Zhou 等[13]提出通过时间注意模型自动选择出视频中最具鉴别力的帧,对质量好的帧进行特征提取. 同年,Xu 等[62]设计了注意力时间池化从图像序列中选择信息帧. 2018 年,Li 等[14]同样使用时间注意模型从所有帧中提取有用的信息.

虽然这些方法一定程度上解决了部分遮挡的问题,但是丢弃遮挡图像的方法并不理想,一方面,丢弃帧的剩余可见部分可能存在有用的信息;另一方面,丢弃帧中断了视频的时间信息. 针对以上的问题,2019 年,Hou 等[63]提出时空补全网络 STCnet(Spatial-Temporal Completion network),试图恢复被遮挡部分的外观来解决部分遮挡的问题.

4.3　光照变化

突然的光照变化会严重降低行人重识别模型的性能,因为现有的大部分解决方法十分依赖颜色特征,而光照变化会带来图像颜色的巨大变化,如图 6 所示.

2010 年,Farenzena 等[64]遵循水平垂直对称原则对人体轮廓进行分割,并从每个分割的身体部位积累局部颜色特征,在对数色度空间中,通过考虑不同光照条件下的颜色分布,建立了基于颜色的不变特征.

2014 年,Ma 等[65]提出了生物协方差描述符(gBiCov)来处理光照变化,该方法使用了 Gabor 滤波器来增强模型对光照变化的鲁棒性,然后使用协方差描述符计算相邻尺度下特征的相似性,大部分光照变化带来的影响会被协方差矩阵吸收. 文献[66]使用 Retinex 算法对图像进行预处理. 通过考虑光照变化和颜色感知来生成一致的彩色图像. 它消除了阴影区域,产生具有增强色彩信息的图像.

图 6　光照变化示例

Fig. 6　Examples of illumination change

处理光照变化的一类方法就是提取对光照变化具有鲁棒性的特征;第二类在正常图像和光照变化图像之间建立一种联系,通过协方差矩阵等方法处理两者之间的不同;第三类就是采用合适的方法对视频进行预处理,使颜色变化平缓.

4.4　跨模态检索

现有大多数行人重识别模型单独关注图像或视频行人重识别问题. 事实上,从图像到视频的行人重识别在失踪者定位、犯罪跟踪和行人视频检索等方面具有重要的意义. 在图像-视频行人重识别任务中,由于图像与视频存在着巨大的跨模态差异,如何融合图像特征与视频特征,以及如何在外观图像特征与时空视频特征之间进行准确的匹配是该问题的关键挑战.

针对融合问题,2018 年,文献[67]提出了 3 种融合方案,包括早期融合,乘积规则融合和自适应查询的后期融合,并分析了 3 种时期融合的效果. 早期的融合方案将手工特征 KDES 和深度特征结合起来并反馈到 SVM 模型中,基于乘积规则融合方案根据乘积计算相似度融合特征,而后期的融合方案则是通过计算特征的得分曲线评估一个特征的有效性,并在融合时分配不同的权重. 实验表明,在 3 种融合方案中,后两种延迟融合方案效果优于第一种. 2019 年,他们在特征提取阶段添加了 GOG 特征和 ResNet 学习到的深度学习特征[68],提升了特征的表达能力.

针对匹配问题,2018 年,Zhu 等[69]通过学习图像与视频间的异构字典将图像和视频特征转化为具有相同维数的编码系数,再利用编码系数进行匹配. Wang 等[70]学习了一种图像到视频的距离度量方式来完成两者之间的匹配,此方法在 MARS-P2S 数据集上 Rank-1 只有 55.25%,还有很大的提升空间,值得进一步研究.

5　实验数据集与评价标准

5.1　实验数据集介绍

为了使研究场景更加接近真实情况,人们采集了若干视频行人重识别数据集,为了解决特定应用场景下的行人重识别问题,学者们也收集了一些针对特定问题的数据集. 数据集信息如表 3 所示. 最常用的 4 个数据集如下,

(1)PRID2011 数据集[71]于 2011 年发表,由两个摄像头拍摄,A 相机下有 385 组行人序列,B 相机下有 749 组行人序列,每个序列长度大约是 100 到 150 帧,但是只有 200 人同时在两个相机下出现过.

(2)iLIDS-VID 数据集[72]于 2014 年发表,拍摄于机场到达大厅. 它由 300 个随机采样的人的 600 个图像序列组成,每个人在两个摄像机视图中都有一对图像序列. 每个图像序列的长度可变,从 23 到 192 不等,平均数量为 73. 该数据集存在严重的遮挡问题,极具挑战性.

(3)Mars 数据集[73]于 2016 年发表,拍摄于清华大学校园,是第一个可以用于深度学习的大型视频行人重识别数据集. 它由 6 台摄像机拍摄,总共有 1261 个不同的行人,20715 个图像序列,每个行人至少被 2 个摄像机捕获.

(4)DuckMTMC-VideoReID 数据集拍摄于杜克大学,是多摄像头跟踪数据集 DukeMTMC 的子集,包括 702 个用于训练的身份,702 个用于测试的身份,以及 408 个干扰项. 总共有 2 196 个视频用于训练,2 636 个视频用于测试.

除此之外,一些针对特定问题的数据集如下:

(1)EgoReID 数据集[74]的视频为可穿戴相机或手机等拍摄的第一人称视角视频,具有自我运动、模糊、视角扭曲等特点. 其中包含 900 个不同的行人,10 200 个视频.

(2)Motion-ReID 数据集[75]专门为长期场景中的行人重识别问题收集,拍摄于办公楼,由两个独立的监控摄像头拍摄,一共收集了 30 个人的 240 个视频片段,每个行人在两个摄像机下记录的间隔时间较长,至少一周.

(3)HLVID 数据集[76]针对监控视频中普遍存在的分辨率低问题收集,拍摄于公共交通道路,由两个摄像头拍摄,总共包含 200 个不同行人,每个视频长度从 56 到 236 不等,平均长度为 126. 其中,高分辨率图像分辨率范围从 44 * 120 到 173 * 258,平均分辨率为 105 * 203;低分辨率图像分辨率范围从 8 * 19 到 19 * 31,平均分辨率为 11 * 21.

(4)CGVID 数据集[77]针对现实应用中既存在彩色视频又存在灰度单色视频的问题收集,拍摄于武汉大学,由两个摄像头拍摄,总共包含 200 个不同行人的 52 723 幅图像,每个视频长度从 58 到 262 不等,平均长度为 130.

表 3　视频行人重识别实验数据集

Table 3　Video-based person re-identification data sets

数据集名称	行人数量	视频数量	行人检测框个数	摄像机个数	检测方法	评价指标	发表时间	场景	特点
PRID2001	200	400	40 000	2	手动	CMC	2011	—	干扰项多
Ilids-vid	300	600	43 800	2	RT FCL	CMC	2014	机场	穿着相似、遮挡问题严重
Mars	1 261	20 715	1 067 516	6	DPM+GMMCP	CMC+Map	2016	校园	第一个大型视频行人重识别数据集
DuckMTMC-VideoReID	1 404	4 832	815 420	8	—	CMC+Map	—	校园	时空连续性强
EgoReID	900	10 200	176 000	3	Yolo9000+FCDSC	CMC+Map	2018	—	第一人称视频视频
Motion-ReID	30	240	—	2	—	CMC	2018	办公楼	拍摄间隔时间长
CGVID	200	—	52 723	2	手动	CMC+Map	2018	校园	彩色图像与灰度图像
HLVID	200	—	50 656	2	手动	CMC	2019	交通道路	高分辨率与低分辨率

5.2　评价标准

评价行人重识别方法的性能指标主要有两个,一是累计匹配曲线 CMC(Cumulative Match Characteristic)曲线;二是 mAP(mean Average Precision).

CMC 曲线:反映的是 top-k 的击中概率,主要用来评估排序结果的正确率. 具体含义是指,在候选视频中检索查询视频,前 k 个结果中包含正确匹配结果的比例. 如图 7 所示,它表示在所有的查询视频中,30%的查询视频返回的 top-1 结果是正确的. 计算方式如下:

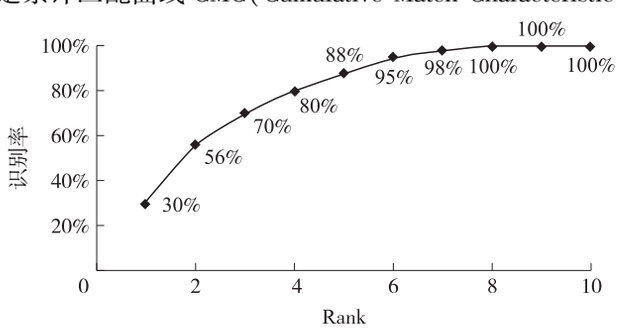

图 7　累计匹配曲线

Fig. 7　Cumulative Match Characteristic curve

$$\mathrm{CMC}_k = \frac{\sum_{i=1}^{|P|} I(r(g_{p_i}) \leqslant k)}{|P|},$$

式中,$|P|$ 为查询集的大小,即 $P=\{p_1,p_2,\cdots,p_{|P|}\}$,$p_i$ 为查询集中的第 i 个人,一一计算其与候选集 $g_i \in G$ 的距离($G=\{g_1,g_2,\cdots,g_n\}$),并进行排序. 正确的目标记为 g_{p_i},其在排序中的位置记为 $r(g_{p_i})$,$I(g_{p_i})$ 为示性函数.

mAP:是每一个查询视频结果与正确结果的匹配程度,与 CMC 曲线不同的是,它更加重视结果的排序,即正确的结果位置越前越好. 计算公式如下:

$$\mathrm{mAP}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{AP}_i,\quad \mathrm{AP}=\int_0^1 \mathrm{CMC}(k)\,\mathrm{d}k$$

式中,AP 为平均准确率,实际上就是 CMC 曲线与坐标轴的面积. 每次查询的结果对应一个 AP,mAP 是总的查询结果 AP 的算术平均值.

6 未来研究问题

在实际生活中存在拍摄视频分辨率差距大、彩色视频与灰色视频混杂或者行人更换服饰等现象,现有的大部分视频行人重识别研究大多假定不存在上述现象,在实际应用中有一定的局限性. 为了进一步拓展视频行人重识别技术的应用场景,未来仍有三个问题值得进一步研究.

6.1 尺度失配

在实际生活中,由于某些相机质量差或者行人与摄像头之间的距离太远,通常会导致采集的行人视频为低分辨率视频,从而导致视频中有用信息的丢失,所以在低分辨率和高分辨率的视频之间进行重识别是未来的一个研究问题.

现有的方法十分依赖图像的颜色特征,而且需要先检测行人,所以行人轮廓是非常重要的,但低分辨率图像无法提供高质量的像素,行人轮廓模糊,且行人服装与背景混为一体. 针对这些问题,2017 年,Zheng 等[78]首先对图像进行预处理,消除图像之间的混色差异,从而使同一个人的色彩特征相等,虽然行人轮廓信息丢失,但人的头部、夹克等某些宽区域的颜色和垂直位置并没有发生明显的变化. 因此将人的图像分成这些大的区域,并从每个区域提取颜色特征.

上述方法解决了处理低分辨率的问题,但是没有考虑低分辨率图像与高分辨率图像之间的匹配问题. 2019 年,Ma 等[76]提出了一种基于半耦合映射的集对集距离学习方法 SMDL(Semi-coupled Mapping based set-to-set Distance Learning),发现高分辨率图像与低分辨率图像之间的映射关系,得到的映射矩阵可以补偿低分辨率图像的损失信息.

6.2 成像风格失配

现实场景中,因为相机故障或者相机为灰色模式会导致采集的行人视频为灰色单色视频,颜色信息会大量丢失,这需要采取有效的方法在彩色和灰度视频之间进行行人重识别,称之为 CGVPR 任务(Color to Gray Video Person Re-identification). 为了解决 CGVPR 任务,2020 年,Ma 等[77]认为同一个人的彩色和灰度视频之间存在着内在关系,提出了一种基于非对称视频内投影的半耦合字典对学习方法 SDPL(Semi-coupled Dictionary Pair Learning),该方法分别学习一对视频内投影矩阵、一对彩色和灰度视频字典以及半耦合映射矩阵. 学习到的字典对和映射矩阵可以一起弥合真彩色和灰度视频的特征之间的差距.

6.3 长时视频行人重识别

现有的行人重识别方法依赖视频中行人的外观特征,如颜色特征,所以大多假设行人短时间内没有显著的外貌变化,不能解决行人换装的问题. 然而,在许多现实场景中,行人可能在长时间间隔后重新出现在监控视频里,但是衣着不一样. 由于着装的改变,利用行人的外观特征进行视频之间的匹配不再适用. 2018 年,Zhang 等[75]认为同一个人即使换装,但步态、身体动作等特征不会发生变化,提出一种基于动态线索的精细运动编码模型,从视频中提取行人的动态运动模式,根据运动模式的不同来区别不同的行人,一定程度上解决了这个问题,但还值得更深入的研究.

7 结论

本文探讨了视频行人重识别的处理过程,详细描述了处理过程中最重要的两个阶段:特征提取和距离度量. 介绍了视频行人重识别现有实验数据集和评价标准,然后提出了该研究领域目前面临的四大挑战,包括姿态变化、遮挡、光照变化、跨模态检索等,给出了相应的解决方案,并展望了视频行人重识别未来的研究问题.

[参考文献]

[1] WOJCIECH Z,ZORAN Z,BEN J A K. Keeping track of humans:have I seen this person before? [C]//IEEE International Conference on Robotics and Automation,Barcelona,Spain,2005.

[2] NILOOFAR G,THOMAS B S,RICHARD I. Hartley. person re-identification using spatiotemporal appearance[C]//IEEE Conference on Computer Vision and Pattern Recognition,New York,USA,2006.

[3] LORIS B, MARCO C, ALESSANDRO P, et al. Multiple-shot person re-identification by hpe signature [C]//International

Conference on Pattern Recognition, Istanbul, Turkey, 2010.
[4] MICHELA F, LORIS B, ALESSANDRO P, et al. Person re-identification by symmetry-driven accumulation of local features[C]// IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 2010.
[5] XIU Z, FEDERICO P, BIR B. Attributes co-occurrence pattern mining for video-based person re-identification[C]// Advanced Video and Signal Based Surveillance, Lecce, Italy, 2017.
[6] LIU H, JIE Z, JAYASHREE K, et al. Video-based person re-identification with accumulative motion context[J]. IEEE transactions on circuits and systems for video technology, 2017, 28(10): 2788-2802.
[7] NIALL M, JESUS M D R, PAUL M, et al. Recurrent convolutional network for video-based person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vergas, USA, 2016.
[8] DAHJUNG C, KHALID T, EDWARD J D. A two stream siamese convolutional neural network for person re-identification[C]// International Conference on Computer Vision, Venice, Italy, 2017.
[9] YAN Y C, NI B B, SONG Z C, et al. Person re-identification via recurrent feature aggregation[C]//European Conference on Computer Vision, Amstordam, The Netherlands, 2016.
[10] CHEN L, YANG H, GAO Z Y. Joint attentive spatial-temporal feature aggregation for video-based person re-identification[J]. IEEE access 7, 2019: 41230-41240.
[11] ZHANG W, MA B P, LIU K, et al. Video-based pedestrian re-identification by adaptive spatio-temporal appearance model[J]. IEEE transactions on image processing, 2017, 26(4): 2042-2054.
[12] LIU H, JIE Z Q, KARLEKAR J, et al. Video-based person re-identification with accumulative motion context[J]. IEEE transactions on circuits and system for video technology, 2017, 28(10): 2788-2802.
[13] ZHOU Z, HUANG Y, WANG W, et al. See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017.
[14] LI S, SLAWOMIR B, PETER C, et al. Diversity regularized spatiotemporal attention for video-based person re-identification[C]// IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018.
[15] LI J N, ZHANG S L, HUANG T J. Multiscale 3d convolution network for video based person reidentification[C]//The Association for the Advance of Artificial Intelligence, Honolulu, USA, 2019.
[16] LIU J W, ZHA Z J, CHEN X J, et al. Dense 3D-convolutional neural network for person re-identification in videos[J]. ACM transactions on multimedia computing, communications, and applications, 2019, 15: 1-19.
[17] LIU J, SUN C, XI X, et al. A spatial and temporal features mixture model with body parts for video-based person re-identification[J]. Applied intelligence, 2019, 49(9): 3436-3446.
[18] YU B Z, XU N, ZHOU J. Cross-media body-part attention network for image-to-video person re-identification[J]. IEEE access 7, 2019: 94966-94976.
[19] WU Y M, OMAR E F B, LI X. Adaptive graph representation learning for video person re-identification[EB/OL]. arXiv: 1909.02240, 2019.
[20] XU S M, HU S Q. Video-based person re-identification by region quality estimation and attributes[C]//International Conference on Cognitive Systems and Signal Processing, Beijing, China. 2018.
[21] SONG W R, ZHENG J Y, WU Y H, et al. A two-stage attribute-constraint network for video-based person re-identification[J]. IEEE access 7, 2019: 8508-8518.
[22] SUN R, HUANG Q H, XIA M M, et al. Video-based person re-identification by an end-To-end learning architecture with hybrid deep appearance-temporal feature[J]. Sensors, 2018, 18(11): 3669-3689.
[23] NEERAJ M, GAURAV S. Video person re-identification using learned clip similarity aggregation[EB/OL]. arXiv: 1910.08055, 2019.
[24] CHEN P X, DAI P Y, WU Q, et al. Video-based Person re-identification with two-stream convolutional network and co-attentive snippet embedding[EB/OL]. arXiv: 1905.11862, 2019.
[25] XIE Z W, LI L, ZHONG X, et al. Image-to-video person re-identification by reusing cross-modal embeddings[EB/OL]. arXiv: 1810.03989, 2018.
[26] BHASWATI S, SAI R K, JAYANTA M, et al. Video based person re-identification by re-ranking attentive temporal information in deep recurrent convolutional networks[C]//IEEE International Conference on Image Processing, Athems, Greece, 2018.
[27] CHEN D P, LI H S, XIAO T, et al. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding[C]//IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018.
[28] ZHANG D Y, WU W X, CHENG H, et al. Image-to-video person re-identification with temporally memorized similarity learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(10): 2622-2632.

[29] DAI J,ZHANG P P,WANG D,et al. Video person re-identification by temporal residual learning[J]. IEEE Transactions on Image Processing,2019,28(3):1366-1377.
[30] OUYANG D Q,ZHANG Y H,SHAO J. Video-based person re-identification via spatio-temporal attentional and two-stream fusion convolutional networks[J]. Pattern recognition letters,2019,117:153-160.
[31] SONG W R,WU Y H,ZHENG J Y,et al. Extended global-local representation learning for video person re-identification[J]. IEEE access 7,2019:122684-122696.
[32] LIU Z,CHEN J X,WANG Y H. A fast adaptive spatio-temporal 3D feature for video-based person re-identification[C]// International Conference on Image Processing,Phoenix,USA,2016.
[33] WU L,WANG Y,SHAO L,et al. 3D person VLAD:learning deep global representations for video-based person re-identification[EB/OL]. CoRR abs/1812.10222,2018.
[34] CHENG L,JING X Y,ZHU X K,et al. A hybrid 2D and 3D convolution based recurrent network for video-based person re-identification[C]//International Conference on Neural Information Processing,Siem Reap,Cambodia,2018.
[35] NAOKI K,KOHEI H,MASAMOTO T,et al. Video-based person re-identification by 3d convolutional neural networks and improved parameter learning [C]//International Conference on Image Analysis and Recognition, Povoa de Varzim, Portugal,2018.
[36] LIAO X Y,HE L X,YANG Z W,et al. Video-based person re-identification via 3D convolutional networks and non-local attention[C]//Asian Conference on Computer Vision,Perth,Australia,2018.
[37] LIU Y H,YUAN Z X,ZHOU W G,et al. Spatial and temporal mutual promotion for video-based person re-identification[C]// AAAI Conference on Artificial Intelligence,Honolulu,USA,2019.
[38] LI J N,ZHANG S L,HUANG T J. Multi-scale 3D convolution network for video based person re-identification[C]//AAAI Conference on Artificial Intelligence,Honolulu,USA,2019.
[39] YANG X,ZHANG B,DONG Y,et al. Spatiotemporal attention on sliced parts for video-based person re-identification[C]// Visual Communications and Image Processing,Taichung,China,2018.
[40] WU L,WANG Y,GAO J B,et al. Where-and-when to look:deep siamese attention networks for video-based person re-identification[J]. IEEE transactions on multimedia,2019,21(6):1412-1424.
[41] XI J L,ZHOU Q,ZHAO Y R,et al. Fine-grained fusion with distractor suppression for video-based person re-identification[J]. IEEE access 7,2019:114310-114319.
[42] LI J N,WANG J D,TIAN Q,et al. Global-local temporal representations for video person re-identification[EB/OL]. arXiv: 1908.10049,2019.
[43] YE M,LI J W,ANDY J M,et al. Dynamic graph co-matching for unsupervised video-based person re-identification[J]. IEEE transactions on image processing,2019,28(6):2976-2990.
[44] ABHIMANYU S,ANANDA S C. A graph-theoretic framework for summarizing first-person videos[C]//Graph Based Representations in Pattern Recognition,Tours,France,2019.
[45] SUN L C,ZHOU Y,LIU J L,et al. Graph regularized and label-matched dictionary learning for video-based person re-identification[C]//Visual Communications and Image Processing,Taichung,China,2018.
[46] ZHENG Y,CHEN Z H,SENEM V,et al. Person detection and re-identification across multiple images and videos obtained via crowdsourcing[C]//International Conference on Distributed Smart Cameras,Paris,France,2016.
[47] LI Y J,ZHUO L,LI J F,et al. Video-based person re-identification by deep feature guided pooling[C]//IEEE Conference on Computer Vision and Pattern Recognition,Honolulu,USA,2017.
[48] TANZILA R,MRIGANK R,WANG Y. Convolutional temporal attention model for video-based person re-identification[C]// IEEE International Conference on Multimedia and Expo,Shanghai,China,2019.
[49] OUYANG D Q, SHAO J, ZHANG Y H, et al. Video-based person re-identification via self-paced learning and deep reinforcement learning framework[C]//ACM Multimedia,Seoul,Korea,2018.
[50] BASSEM H,WALID A,MOHAMED A. Multi-shot person re-identification using a novel video covariance approach[C]// ACM Symposium on Applied Computing,Marrakech,Morocco,2017.
[51] WEINBERGER K Q,SAUL L K. Fast solvers and efficient implementations for distance metric learning[C]//International Conference on Machine Learning,Helsinki,Finland,2008.
[52] OSTINGER,HIRZER M,WOHLHART P,et al. Large scale metric learning from equivalence constraints[C]//IEEE Conference on Computer Vision and Pattern Recognition,Providence,USA,2012.
[53] 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.

[54] 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.
[55] ZHU X K,JING X Y,YOU X G,et al. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics[J]. IEEE transactions image processing,2018,27(11):5683-5695.
[56] ZHANG W,LI Y M,LU W Z,et al. Learning intra-video difference for person re-identification[J]. IEEE transactions circuits system video technology,2019,29(10):3028-3036.
[57] NAVANEET K L,VASUDHA T,VENKATESH B R,et al. All for one:frame-wise rank loss for improving video-based person re-identification[C]//IEEE International Conference on Acoustics,Speech,and Signal Processing,Brighton,UK,2019.
[58] WEI L,ZHANG S,YAO H,et al. Glad:global-local-alignment descriptor for pedestrian retrieval[C]//ACM multimedia,Mountain View,USA,2017.
[59] ZHAO H,TIAN M,SUN S,et al. Spindle net:person re-identification with human body region guided feature decomposition and fusion[C]//IEEE Conference on Computer Vision and Pattern Recognition,Honolulu,USA,2017.
[60] CHEN Y Z,HUANG T D,NIU Y Z,et al. Pose-guided spatial alignment and key frame selection for one-shot video-based person re-identification[J]. IEEE access 7,2019:78991-79004.
[61] WU J J,JIANG J G,QI M B,et al. Independent metric learning with aligned multi-part features for video-based person re-identification[J]. Multimedia tools application,2019,78(20):29323-29341.
[62] XU S J,CHENG Y,GU K,et al. Jointly attentive spatial-temporal pooling networks for video-based person re-identification[C]//International Conference on Computer Vision,Venice,Italy,2017.
[63] HOU R B,MA B P,CHANG H,et al. VRSTC:occlusion-free video person re-identification[C]//IEEE Conference on Computer Vision and Pattern Recognition,Long Beach,USA,2019.
[64] FARENZENA M,BAZZANI L,PERINA A,et al. Person re-identification by symmetry-driven accumulation of local features[C]//IEEE Conference on Computer Vision and Pattern Recognition,San Francisco,USA,2010.
[65] MA B P,SU Y,FRéDéRIC J. Covariance descriptor based on bio-inspired features for person re-identification and face verification[J]. Image vision computation,2014,32(6-7):379-390.
[66] APARAJITA N,PANKAJ K S,DUSHYANT S C,et al. A person re-identification framework by inlier-set group modeling for video surveillance[J]. Journal ambient intelligence and humanized computing,2019,10(1):13-25.
[67] THUY B N,THI T L,DINH D N,et al. A reliable image-to-video person re-identification based on feature fusion[C]//Asian Conference on Intelligent Information and Database Systems,Pong Hol City,Vietnam,2018.
[68] THUY B N,THI L,NAM P N. Fusion schemes for image-to-video person re-identification[J]. Journal information telecommunication,2019,3(1):74-94.
[69] ZHU X K,JING X Y,YOU X G,et al. Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix[J]. IEEE transactions information forensics and security,2018,13(3):717-732.
[70] WANG G C,LAI J H,XIE X H. P2SNet:can an image match a video for person re-identification in an end-to-end way[J]. IEEE transactions circuits system video technology,2018,28(10):2777-2787.
[71] MARTIN H,CSABA B,PETER M R,et al. Person re-identification by descriptive and discriminative classification[C]//Scandinavian Conference on Image Analysis,Ystad,Sweden,2011.
[72] WANG T Q,GONG S G,ZHU X T,et al. Person re-identification by video ranking[C]//European Conference on Computer Vision,Zwrich,Switzerland,2014.
[73] ZHENG L,BIE Z,SUN F Y,et al. MARS:a video benchmark for large-scale person re-identification[C]//European Conference on Computer Vision,Amstordam,the Netherlands,2016.
[74] EMRAH B,YONATAN T T,MUBARAK S. EgoReID:person re-identification in egocentric videos acquired by mobile devices with first-person point-of-view[EB/OL]. arXiv:1812.09570,2018.
[75] ZHANG P,WU Q,XU J S,et al. Long-term person re-identification using true motion from videos[C]//IEEE Winter Conference on Applications of Computer Vision,Lake Tahoe,USA,2018.
[76] MA F,JING X Y,YAO Y F,et al. High-resolution and low-resolution video person re-identification:a benchmark[J]. IEEE access 7,2019:63426-63436.
[77] MA F,JING X Y,YAO Y F,et al. True-color and grayscale video person re-identification[J]. IEEE transactions information forensics and security,2020,15:115-129.
[78] ZHENG M X,KENTARO T,NOBUHIRO M,et al. Privacy-conscious person re-identification using low-resolution videos[C]//Asian Conference on Pattern Recognition,Nanjing,China,2017.

[责任编辑:陆炳新]