跨摄像头下面向跟踪的行人身份对齐

目录

第一章 绪论

1.1 研究背景及意义

随着国民生活水平的提高和科技的发展，人们对周围公共安全问题越来越重视。而由中央政法委牵头，由公安部联合工信部等多个部委共同发起建设的“天网工程”，是世界上最大的视频监控网，监控摄像头超过 2000 万个。但随着视频监控应用的日益丰富，视频监控系统变得越来越庞大，采集的数据也呈现指数级增加，如果通过人工观看视频再进行处理的方式不仅耗时耗力，并且很难满足对多目标、大范围的长时间持续监控和处理。因此，依靠计算机对图像、视觉等相关技术的强大计算能力，来实现对数据集的自动化处理是发展的方向，而智能化的处理技术是视频监控应用未来的必然趋势。

一个智能化的视频监控系统应具备，能够对获取到的监控视频使用相应的算法针对感兴趣的目标进行在线的分析和处理，剔除大量的无用信息，只保留当前关心的信息，同时还需要对视频中出现的特别行为应该能够给用户适当的反馈信息。与智能化的视频监控系统相关的具体应用包括：行人异常行为检测和识别、公共场所人流量统计、针对犯罪分子的识别和跟踪等。智能化的监控系统在提高人们安全保障的同时也大大的减少了人力物力的消耗。因此使用智能化的技术构建一个稳定有效的智能视频监控系统具有十分重要的实际应用价值和广阔的现实需求。

随着针对行人目标的视频监控应用日益丰富，行人检测、跟踪、检索等相关智能化视频处理技术迅速发展。因为应用场景的不断扩大，摄像头的布置范围也越来越大，此时监控网络呈现出以多摄像头监控为主，从经济角度考虑其中非重叠区域的多摄像头监控成为了主流方向。非重叠视野区域的多摄像头行人跟踪也称之为跨摄像头下的行人跟踪。目前，单摄像头下多项技术相对成熟，但多摄像头，尤其是非重叠视野区域情形下的行人跟踪仍面临着诸多挑战。其中的一大挑战在于盲区的存在使得目标的时空信息变得不再可靠。当目标行人离开前一摄像头视野区域后，进入下一摄像头视野区域前，跟踪断续，该时间区间内目标时空信息的缺失增加了目标行人从上一摄像头正确移交给下一摄像头的难度。摄像头下的光照强度、摄像头安防位置、角度、物理性质等差异，导致跨越两个不同摄像头的同一个目标行人外观和姿态发生严重的变化。而一个智能化的监控系统都需要能够处理不同视域目标行人的关联问题，即实现不同摄像头间同一个目标行

人的再识别或正确关联。因此，研究跨摄像头下目标行人身份对齐技术是智能化的监控视频系统中的重要组成之一。

在满足视频监控系统基本功能需求的同时，一个智能化的跨摄像头下面向跟踪的行人身份对齐应该满足以下功能：

1)、目标行人检测：目标行人在摄像头下出现的具体位置；

2)、目标行人跟踪：目标行人在摄像头下的运行轨迹；

3)、目标行人关联：不同摄像头下哪些不同时刻出现的目标属于同一个，即摄像头间的行人再识别。

本文通过对监控视频中目标行人关联相关技术进行深入研究，经过行人检测、行人跟踪和行人关联三个步骤，对多个感兴趣的目标行人完成长时间持续的跟踪，实现了一种跨摄像头下面向跟踪的行人身份对齐，完成智能化的监控任务。

1.2 国内外研究现状

随着视频监控应用的日益丰富，目标跟踪、检索等相关视智能处理技术发展迅速。其研究范畴从单摄像头延展到多摄像头，研究对象也从单个行人等发展到多行人多目标。目前，单摄像头下多项技术相对成熟，但多摄像头，尤其是非重叠视野区域情形下的行人跟踪仍面临着诸多挑战。根据英国 IMS（Intex Management Service）机构发布的调查报告显示，全球的智能视频分析软件的市场规模将在最近五年内快速增长。巨大的市场需求使得智能视频监控技术的研究和开发受到了世界各国企业和学术届的高度重视，如 CMU、INRIA、IBM、Sarnoff、ObjectVideo、IoImage、Sony、MERL、FXPAL、Verient、Vidient、NICE 等科研机构和企业都在该领域进行了大量的研究和开发，并取得了一系列优秀的成果。

国外对智能视频监控技术的研究相对较早，具有较完善的理论基础。其中具有代表性的项目有，1998 年，美国国防高级研究计划局（DARPA）资助，同时由 DARPA、CMU、SARNOFF 研究中心等多个机构共同研发的 VSAM 系统，目标是开发自动视频理解技术，用于军事安全控制，局部战场监控等军事应用场景。1999 年，欧盟成立专项计划 ADVISOR，由 The University of Reading、INRIA、VIGITEC、Kingston University 等多个大学和机构共同研发的 Advisor (Annotated Digital Video for Surveillance and Optimised Retrieval)智能视频监控系统主要针对地铁站内的安全监控问题展开研究，重点研究基于计算机视觉的乘客行为识别、人群监控、人体跟踪以及快速视频检索等。在多人交互行为识别方面的研究的是该系统的一大亮点，系统已于 2003 年完成。2003 年立项，2005 年 10 月初步实现了主要功能的 AVITRACK (Aircraft surroundings, categorized Vehicles and Individuals Tracking for Airport Region Activity model interpretation and Check)系

统，组要组成成员以法国 SILOGIC 为首，包括英国雷丁大学、法国 INRIA、以及挪威、奥地利等国的多个计算机视觉研究机构，系统专门针对飞机场的智能安防进行研究，是一个基于多摄像机的实时多目标跟踪、分类与预警系统。在单摄像机跟踪时采用单高斯模型对背景色彩进行建模以达到较高的检测速度，用 KLT 算法对运动分割后的前景区域进行跟踪，并用多层背景的思想解决目标长时间静止时的检测问题。跟踪结果通过网络传输到控制中心，通过最近邻法进行数据关联，并用 Kalman 滤波做状态估计和融合。2005 年，IBM T.J. Watson 研究中心首次发布了 IBM Smart VideoSurveillance System，该系统具有强大的实时视频分析、事件检测与和报警功能。通过实时目标检测、跟踪、分类，Smart 系统能有效的对视频进行简单的语义描述，并将感兴趣的视频保存在数据库中，用户可以按照事件类型或事件组合等多种方式进行检索，从而大幅度提高了检索效率和对突发事件的快速反应能力。据纽约时代周刊 2007 年 12 月 7 日报道，北京奥组委已经开始部署 IBM 的智能视频监控系统，对全市街道的视频图像进行监控，以应对突发事件和暴力事件，确保 2008 年北京奥运会安全。随着智能视频研究的不断深入，采用多摄像机对目标进行大范围、长时间跟踪成为新的研究热点。2007 年，位于硅谷的美国富士施乐帕拉阿图实验室(FXPAL) 率先构建了一个网络环境下室内多摄像机监控系统 DOTS(Dynamic Object Tracking System) ，采用二十部 AXIS 网络摄像机对办公大楼内的走廊、电梯、出入口、会议室等公共场所进行 24 小时不间断监控。与传统的 IVS 系统不同，DOTS 不仅能在光照变化、运动阴影等条件下稳定检测运动目标，而且通过多摄像机交接，能够完成对特定目标的长时间跟踪。此外，DOTS 具有友好的人机交互界面，主要功能包括基于非线性时间轴的事件标注、多摄像机协同跟踪与最优视角切换等，而且用户可通过网络对整个监控系统进行实时访问和管理。

国内智能视频监控的研究起步较晚，但后来者居上。主要的代表项目有，中科院计算所和上海银晨科技公司共同合作开发的银晨人脸识别系统，系统的目标是能够应用于像机场安检、出入控制和特定对象布控等场景。为了实现大范围视觉监控与语义理解，中科院自动化所研发了 VStar 视频监控系统。中科院自动化所和北京数字奥森公司合作公共完成了 AuthenMetric-F1 系统，实现了对各种条件下高精度、快速人脸识别。

目前，跨摄像头目标跟踪的方法主要分为两种：基于全局检测的目标关联方法和基于机器学习的目标关联方法。

1）、基于全局检测的目标关联模型

由于计算机处理能力的不断提升和摄像头设备分辨率的提高，基于全局检测的目标关联模型在近几年来不仅成为了单摄像头目标跟踪的热门研究方向[23]，而且在跨摄像头领域也有不少研究者提出了基于全局检测的目标关联模型[24]。

HUANG 等[24]提出了一个基于 TLD 算法的跨摄像头跟踪系统。该方法在目标离开某个摄像头视野区域后，检测算法会全局扫描一定范围内摄像头的视频帧，然后通过模板匹配的方法定位目标。在单摄像头目标跟踪时，作者替换了 TLD 中基于特征点检测的光流跟踪方法使用 MeanShift[]和粒子滤波跟踪算法代替。但目标检测模块只采用简单的全局搜索方法，需要很强的计算能力，算法效率较差。

2）、基于机器学习的目标关联模型

目标从一个摄像头到下一个摄像头的过程中并未经过其他的摄像头，则称这两个摄像头是直接连接(directly-connected)。两个直接连接的摄像头的视野区域中包含一个进入区域（目标进入的摄像头）和一个退出区域（目标离开的摄像头）。训练数据由进入区域观测数据集（X）和退出区域观测数据集（Y）两部分组成，每个进入或退出区域的观测实例中包含目标行人的一些特征信息(颜色、纹理特征以及进入或退出该视野区域的时间等)。在摄像头数量较少且摄像头拓扑图已知的情况，采用人工标定每对直接连接的摄像头以及每对直接连接摄像头的进入区域或退出区域比较容易。此时使用监督学习的方法就可以获得目标关联模型。但人工标定训练数据不仅耗时耗力，而且由于其固定了目标进出区域使得算法并不具有普适性，除此之外随着摄像头数量的增加其也变得不可实现。同时，每当摄像头网络中发生变化（增加或减少摄像头数目等）都需要重新标定训练数据。

Chun-Te Chu 等人[]提出一种使用非监督学习框架建立摄像头关联模型来进行跨摄像头行人跟踪。方法用关联矩阵 $P_{(N_1+1)\times(N_2+1)}$ 表示观测数据集 X 和 Y 之间的关联性，其中 N_1 和 N_2 为数据集 X 和 Y 中的样本个数。与监督学习法关联矩阵 P 为人工标定的二值矩阵不同，非监督学习的方法把求解关联矩阵 P 当作一个优化问题。通过算法对 P 的求解获得对应的最优关联解。

Javed 等人[]构建一种根据摄像头间的时空线索和目标外观特征的贝叶斯框架实现跨摄像头目标跟踪。由 Parzen 窗口训练数据中得到摄像头间的时空模型，使用颜色模型之间的距离度量作为行人经过摄像头间外观变化的标准。两个相邻摄像头下的对象是否为同一个对象的概率由时空模型和颜色模型共同决定，称之为一致性概率(Correcpondence Probability)，求解关联模型最优解的过程转化成求解最大后验概率估计。该方法的优点是不需要提前校准摄像头，因此方法具有较好的扩展性。其不足之处是求解一致性概率比较复杂，而且训练数据需要提前人工标记运动轨迹等信息。

Chen 等人[]在 Javaed 工作的基础上，提出和设计了另一种非监督方法构建摄像头间目标转移模型，该模型同样由时空线索和目标外观线索组成。不同之处在于作者将时空模型使用一个互相关函数表示为一个摄像头连接(Link)中进入观测实例和退出观测实例。互相关函数解决了一致性概率求解复杂的问题，但它没有考虑连接的转移时间分布和峰值不明显的问题。文中作者提出通过减小正相关实例的转移时间方差，以消减大量负相关实例带来的影响，从而突出互相关函数的峰值。经过多次迭代后，该方法能够估计任意有效连接的转移时间分布，而且避免了求解非重叠摄像头中目标一致性问题。文中作者使用基于颜色的特征作为目标外观线索，提出了利用颜色特征转移算法（CCT）[]来建立摄像头间外观转换模型。该算法部分解决了摄像头间光照变化对目标外观颜色的不利影响，缺点是不同的该方法使用的颜色特征转移算法只是基于全局颜色信息考虑，并且颜色转移方向（例如从摄像头 x 到 y 与从 y 到 x）对 CCT 的性能具有较大的影响。

虽然，大量的研究者对跨摄像头中的关联模型进行了大量的研究。但对于跨摄像头中存在的摄像头间由于各个摄像头安防位置不同和物理设备差异等造成的光照差异大、同一目标出现在不同摄像头下是姿态变化严重等难题至今依然没有很好的解决办法。由此也造成了目前方法处理在线监控视频数据难以完成即时的行人身份对齐任务，而对于一个能够即时的完成行人身份对齐任务的智能化的监控系统却有着十分重要的实际应用价值。

1.3 本文主要工作

本文主要针对行人作为目标对象，对行人的鉴别行特征学习模型、关照转移模型和细粒度匹配模型进行了相关的研究，以实现跨摄像头场景下面向跟踪的行人身份对齐，主要工作如下：

首先，提出了一种基于孪生网络的行人鉴别行特征学习模型与时时建立最小费用流图求解最优关联解集的方法。通过孪生网络整合行人分类和认证模型形成一个多任务模型，完成行人特征鉴别性的学习和提取，并时时建立最小费用流图来获得当前时刻的最优关联解集，从而完成初步的行人对齐任务。

其次，针对跨摄像头下普遍存在的关照差异问题，提出和设计基于模糊聚类的关照转移模型。需要特别指出的是，与传统的颜色亮度转移方法不同的是，我们引入了隶属度因子来控制每个像素点的颜色亮度转移量。

最后，针对行人属于非刚性物体，容易发生姿态、几何形状变化等问题，引入了注意力机制模型。通过在基础的孪生网络模型中加入注意力机制，使得模型可以对目标进行细粒度的特征学习和匹配，从而缓解由于姿态变化等问题所带来的影响。

同时，本文通过整合以上方法实现了一个完整的跨摄像头下面向跟踪的行人身份对齐模型，在跨摄像头行人跟踪的基准数据集 NLRP 上的实验结果表明，本文的模型具有较好的行人身份对齐性能。

1.4 本文的结构安排

本文内容安排如下：

第一章 绪论。

第二章 基本理论。

第三章 即时对齐模型。

第四章 光照迁移模型。

第五章 注意力机制模型。

结论部分对本文的主要工作进行了总结并对未来工作做出了展望。

第二章 跨摄像头行人身份对齐的关键技术综述

2.1 跨摄像头的行人身份对齐概述

跨摄像头下面向跟踪的行人身份对齐问题可以描述为：在单摄像头下完成目标行人的检测和跟踪，并根据单摄像头下获取的目标行人特征信息以及有限的时空信息，完成摄像头间的目标行人关联任务，维持相同行人的身份标识不变，还原目标行人经过所有摄像头的完整运行轨迹。跨摄像头下的行人身份对齐算法研究可以将任务分为两个步骤：单摄像头下的行人检测、跟踪和跨摄像头的行人身份关联。如果单摄像头的跟踪结果已知时，此时的任务重点是把不同摄像头间的同一个目标行人运行轨迹正确的关联即得到了跨摄像头行人身份对齐的最终结果。

针对目标行人作为处理对象的智能化监控系统中的视频数据有如下基本特点：1)、固定背景：固定的摄像头，其背景也不会发生改变；2)、摄像头间的差异：摄像头视野区域中各摄像头之间多种因素变化的不确定性（如光照变化，安防角度，摄像头物理参数，等)；3)、目标的特性：行人属于非刚性物体，不同摄像头下的行人姿态具有较大的差异，且目标出现的位置和时机具有不确定性。

这些数据特点以及跨摄像头中摄像头的拓扑位置，决定了在跨摄像头中行人目标跟踪、检测和关联存在以下难点：

（1） 不连续的时空关系

在跨摄像头场景中，由于摄像头间盲区的存在，导致目标行人在离开前一个摄像头进入下一个摄像头的过程中必须经过一个盲区，如图 x 所示，此时目标的时空信息都处于无法获取状态。而目标行人时空信息的非连续性，导致目标关联时候的时空信息不再可靠，因此极大的增加了关联任务的难度。

（2）摄像头间光照变化

光照变化是跨摄像头下行人身份对齐中十分常见的情形，是行人跟踪和行人关联最难解决的问题之一。由于摄像头的安放位置、角度和物理参数等差异，导致不同摄像头下同一目标行人的外观具有较大的差别(灰度和颜色等行人特征信息准确度大大降低)，如图 x 所示，使得同一个摄像头下的不同目标的相似性将会大于不同摄像头下的同一目标。而即使是同一个摄像头下，由于不同的视野区域其也具有不同的关照强度，其无疑给单摄像头下的目标检测和跟踪带来了巨大的挑战。

（3）目标形变

行人在跨摄像头中的行为和运动轨迹难以预测，如转身、弯腰等情况，使得目标行人的几何轮廓信息的可信度大大下降。而且摄像头的不同安放位置和角度设置，使得行人在离开摄像头时和再次进入摄像头时的姿态具有较大的变化，如图 x 所示。在关联问题上，目标形变所带来的影响给关联任务带来巨大的挑战。

2.2 目标检测

目标检测通常可以定义为[25]：对给出的一张图像，目标检测的任务不仅要在给定的图像上识别出存在的物体，给出物体的所属类别，还需要将该物体的准确位置通过 Bounding Box 给出，如图 x 所示。如何将图像中的信息转换为计算机可理解的信息，是计算机图像处理、计算机视觉的关键问题。目标检测是计算机视觉领域中一个传统且十分基础的任务。

按不同的处理方法可以将目前主流的目标检测算法分为 3 类：

（1）、传统的目标检测算法：一般分为三个阶段：首先使用一些算法，如 Selective Search 算法[]，在给定的图像上进行区域选择，将那些潜在可能的目标的区域选择出来。然后对这些选择得到的区域进行特征提取。但由于目标的形态、颜色、关照情况、背景的多样性，设计出相关鲁棒的特征是比较困难的。

Cascade+Harr/SVM+HOG/DPM 以及上述方法的改进版本。

（2）、基于深度学习的方法

传统的目标检测存在一些问题，如区域检测的时间复杂度高，且很难针对性的进行目标的搜索。另一方面，手工设计的特征对于物体的多样性变化、复杂物体等鲁棒性并不强，且随着检测任务的推广，设计特征变得越来越复杂。

近年来，随着计算机的计算性能得到了大幅度的提升，依靠深度学习而发展而来的检测算法也如雨后春笋般的涌现出来。并且占据了各大公开的检测竞赛的榜单。总的来说，基于深度学习的目标检测算法可以分为两大类，一类是将目标检测问题转化为分类问题来解决，其主要思路是先通过启发式方法(selective search)或者 CNN 网络(RPN)产生一系列稀疏的候选框，然后对这些候选框进行分类与回归，该方法的主要优势是算法准确率高，缺点是算法运行效率不高；另一类是将目标检测问题直接当作回归任务处理，其主要思路是均匀的在图片的不同位置进行密集抽样，抽样时可以采用不同尺度和长宽比，然后利用 CNN 提取特征后直接进行分类与回归，该方法的主要优势是算法运行效率高，缺点是正样本和负样本采样及其不均衡，导致算法训练困难和模型准确率不高。

本文将行人作为目标检测对象，使用目前目标检测主流方法中的深度学习方法，并根据行人存在的特点对算法进行了适当的修改和优化，以其能更加符合本

文的应用场景需求。

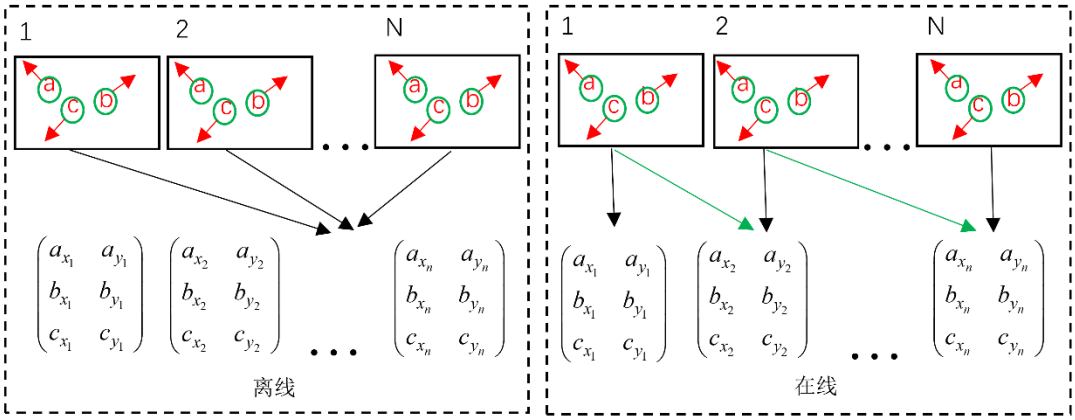


图 2 离线跟踪(左)和在线跟踪(右)

2.3 目标跟踪

目标跟踪任务是在给定输入视频或者连续图像序列的情况下，对目标行人进行定位，并根据时序关系保持目标的身份标识不变和产生他们各自的运行轨迹。行人跟踪作为计算机视觉中的中级任务，行人跟踪作为高级任务的基础方法，如姿态估计[26]、动作识别[27]和行为分析[28]等。对于跟踪来说分为单目标跟踪和多目标跟踪。多目标跟踪是视频分析及监控领域中的基本问题之一，在视频分析、场景剖析、行为事件理解、交通管理及安全防控等应用中都是必须解决的关键问题。与单目标跟踪[][][]仅针对指定的单个目标框进行跟踪不同,多目标跟踪[9][][][]致力于对视频中的所有感兴趣目标进行自动提取，并通过时域关联，得到其运动轨迹信息。因此,多目标跟踪更适合处理包含大量目标的复杂场景。本文将目标行人跟踪作为跨摄像头下行人对齐的研究对象。

多目标跟踪根据对跟踪数据的不同处理方式，分为两种不同类型的的跟踪算法.根据当在第 i 帧的时候,是否可以使用第 i+1~N 帧跟踪数据，分为在线跟踪标和离线跟踪[32]。在线跟踪方法仅依赖于当前帧之前可用的过去信息，而离线跟踪方法无论是当前帧之前或者是未来帧的信息都能够使用。

（1） 在线跟踪

在在线跟踪[][][][]方法中，图像序列是以每帧递进迭代方式(step-wise manner)处理，因此在线跟踪也称为顺序跟踪。如图 x 所示，其中有三个对象(不同的圆圈)a,b 和 c。绿色箭头表示当前帧之前获得的目标信息，结果向量由对象的位置和身份标识组成。在线跟踪是基于最新观测信息，目标运行轨迹是即时输出的。

（2） 离线跟踪

离线跟踪[][][]利用批量的帧对视频进行处理，如图 x 所示，当前帧可使用的

信息包含之前和之后视频帧的观测值，根据所获得的所有信息进行分析估算最后的输出值。

总的来说，在线跟踪比较适用于监控视频数据只能序列化(obtain sequentially)获得的情况。离线跟踪方法的典型用法是在可以使用全局数据的情况。

2.4 目标关联

目前，非重叠视野区域的多摄像头行人跟踪常见的解决机制以单摄像头下的目标检测和跟踪为基础，分为两步骤[5-6]：首先，获得目标在每一个摄像头下的运行轨迹；其次，使用关联算法将摄像头间独立的行人运行轨迹进行对齐合并，获得每个目标完整的运动轨迹。目前主流方法[7-14]均将目标外观特征作为主要线索，并联合时空信息，实现目标关联。

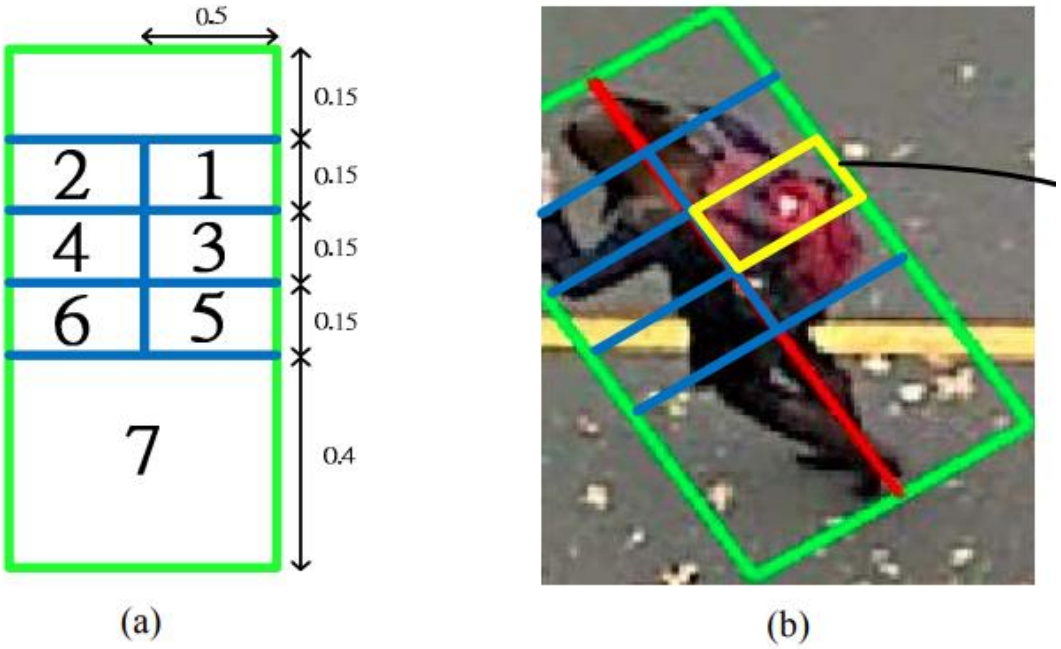
2.4.1 特征提取

特征提取是行人关联的的前提条件，对于提高行人关联的表现的一个重要手段是，提取更具鉴别行的行人特征。对于行人关联被使用最多的特征主要有颜色、纹理、和形状。当一个目标行人经过不同摄像头的时候，由于摄像头的光照和物理性质差异，导致了行人外观发生了剧烈的改变。因此，仅使用单一的特征来识别和标识对象并不能满足复杂的跨摄像头跟踪的环境。将多种互补的特征进行融合来提高关联模型的判别能力，是目标跨摄像头下行人特征提取的主流方向。如，zhang 等[33]将颜色和纹理特征整合为一个描述子作为行人的特征提取器。Zhao 等[34]融合颜色直方图和 SIFT 描述子来进一步提升特征提取器的性能。然而，当对多种特征进行融合的时候，并不是所有的特征的重要性权值是完全均等的，如何对不同的特征设置适当的权重成了特征融合的关键问题。除了形状和外观特征外，表示人体结构的空间信息也有助于提高行人识别关联的性能。为了获得更加有效的人体结构的空间信息，一些研究者将人体按区域划分为了不同的部件，由此来提高关联模型的判别性能。例如，zheng 等[35]为了更加有效利用人体结构的空间信息，他们将人体水平的划分为了 6 个不同部件，并对每个部件分别提取 29 个特征通道(features channels)。Chun-Te 等[36]将人体结构强制划分 8 个部分且未使用头部特征信息，如图 x 所示，给剩下的 6 个部件分别分配不同的权重，在关联的时候仅对相应的相同部件进行匹配关联。其主要的不足就是分割过于生硬，往往会造成区域错误匹配，导致最后细粒度划分的匹配结果并不是很理

想。

2.4.2 相似性度量

距离度量学习通常用于通过学习模型测量来自两个不同摄像头的两个对象之间的特征距离。度量学习目标是使模型能够度量出相同的目标行人具有较小距离，不同的行人将具有较大的特征距离。使用较广的相似性度量方法主要有RankSVM[]、Rankboost[]、PRDC(Probabilistic relative distance comparison)[]和



RDC(relative distance comparison)[]等。总之，距离度量学习的基本思想可以大致分为两种学习模式，第一种学习模式，比如 RankSVM [37]，就是要制作相同的对象不同的相机有较小的距离和制造不同相机的不同物体有更大的不同距离，即最大化不同类别的判别间隔。另一种学习模式，如 RDC [36]，就是简单让同一类别的特征距离小于不同类别，即仅涉及相对距离比较。

2.4.3 关联模型

Chen 等[7]提出了一个改进的相似度度量对单摄像机跟踪和跨摄像头间目标相似性进行处理，并在全局图模型中进行优化求解。文献[8,10,14]考虑了两个不重叠的摄像头下所有可能的情况，通过外观特征和时空线索定义了一个对应矩阵，计算成对目标关联的概率，最后采用匈牙利算法解决目标间关联问题。Chen 等[9]将所有单摄像头下目标的运动轨迹关联转化为一个全局的最大后验问题，映射成一个费用流网络，并通过一个最小费用流算法来解决。Cheng 等[11]将摄像头间的轨迹关联问题定义为一个多分类问题，将前序摄像机中每个轨迹视为一个

类，则后继摄像头中的每一个轨迹都需要被分到其中一个类中。进而定义一个能量函数，它编码了部分感知的外观变化、群体活动、轨迹间的相互排斥信息等，最后使用马尔可夫随机场模型解决该多分类问题。Gao 等[12]根据单摄像头下跟踪器的跟踪结果，利用时空相关性，实现跟踪一致性并在多个跟踪结果之间建立成对关联。Zhang 等[13]将多个相互作用目标的跟踪作为网络流问题，通过 k-最短路径算法得到解决方案。并利用目标之间的时空关系识别群体合并和分离事件。

2.5 跨摄像头行人身份对齐的评价指标

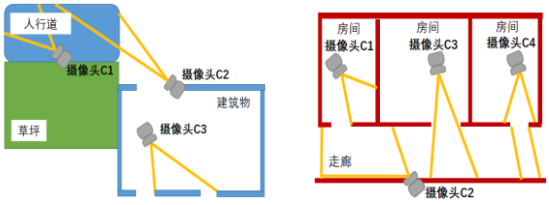
本文在跨摄像头行人跟踪的现实场景数据集 NLPR_MCT[7]上进行实验，该数据集来自现实生活中的监控摄像头网络所记录的一系列室内外跨摄像头场景，包含四个子数据集，每个子数据集来自 3~5 个非重叠视野区域的摄像头不等，目标行人数量不同，存在不同程度的遮挡。表 2 列出了数据集 NLPR_MCT 的具体信息，图 6 为每个子数据集中的摄像头视野区域和拓扑分布，从图中可知每个摄像头的视野区域光照强度不同，增加了跨摄像头行人对齐难度。

表 2 数据集 NLPR_MCT 细节信息

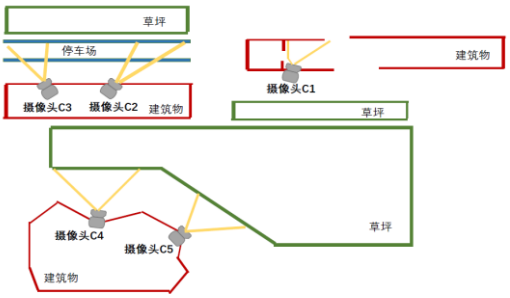
数据集	Dataset1	Dataset2	Dataset3	Dataset4
视频时长	20 分钟	20 分钟	3。5 分钟	24 分钟
分辨率	320×240	320×240	320×240	320×240
行人数目	235	255	14	49
GT^s	71853	88419	18187	42615
GT^c	334	408	152	256
摄像头数	3	3	4	5

本文以跨摄像头下行人跟踪为具体实验场景。通常跨摄像头行人跟踪采用 MCTA[7]评价指标，该指标如式(3)所示：

$$MCTA = Detection * Tracking^{SCT} * Tracking^{ICT}$$
$$= \left(\frac{2 * precision * recall}{precision + recall}\right)\left(1 - \frac{\sum_t mme_t^s}{\sum_t tp_t^s}\right)\left(1 - \frac{\sum_t mme_t^c}{tp_t^c}\right) \quad (3)$$



a. Dataset1/Dataset2 室内外场景　　b. Dataset3 室内场景



c. Dataset4 室内外场景

图 6 数据集中摄像头分布拓扑和视野区域

式(3)中，*MCTA* 评价可细分为三个部分，分别用于表征行人检测，单摄像头下行人跟踪和跨摄像头行人关联的精度。其中，*Detection* 对应目标检测算法中的 $F_1-score$；该评价将新进入的目标行人作为 tp_t^c 的一个真值当做一个默认设置。mme_t^* 和 tp_t^* 分别表示在第 t 帧算法的错误跟踪数目和当前帧的真值数，mme_t^s、tp_t^s 和 mme_t^c、tp_t^c 分别表示单摄像头和跨摄像头的值。*MCTA* 最终值在 0~1 之间，值越大表示精度越高。

2.6 本章小结

跨摄像头下的行人身份对齐是一个复杂的任务，主要涉及的问题有单摄像头下的跟踪和检测，以及其核心问题摄像头间的目标关联。本章主要分为四个部分，分别对跨摄像头中行人身份对齐涉及到的任务进行了分析和介绍。其中，每一部分首先对每个任务进行描述性介绍，之后介绍了目前主流的解决方法并对相应的算法进行优缺点分析。最后，对本文所采用的算法评价标准和跨摄像头下行人身份对齐基准数据集进行了介绍。

第三章 跨摄像头的行人身份即时对齐

3.1 引言

行人即时对齐问题记录每个时刻离开任一摄像头下离开其视野域的目标行人等待关联，同时首次检测到进入任一摄像头视野域的新目标行人时，先与摄像头网络中等待关联的行人匹配，如匹配成功，则找到其前序运动轨迹，该目标行人从前序摄像头移交至当前摄像头并延续跟踪。此时，由于目标行人在当前摄像头下为首次检测到，尚未形成跟踪轨迹，因此，较传统延后关联而言，即时对齐的最大挑战在于基于目标单帧图像完成关联任务。

假设摄像头网络由 n 个视野区域不重叠的摄像头 $C_1,C_2,...,C_n$ 组成，在 t_p 时刻，摄像头 C_i 中获取到 $m_i^{t_p}$ 个目标，记为 $O_i^{t_p}=\left\{o_{i,1}^{t_p},o_{i,2}^{t_p},...,o_{i,m_i^{t_p}}^{t_p}\right\}$，其中 $o_{i,k}^{t_p}$ 记为三元组 $\left(feature_{i,k}^{t_p},boundbox_{i,k}^{t_p},track_{i,k}^{t_p}\right)$，为获取到的目标信息，由目标外观特征描述子 $feature_{i,k}^{t_p}$、目标代表性检测框 $boundbox_{i,k}^{t_p}$ 和目标时空信息 $track_{i,k}^{t_p}$ (目标的进出位置和时间等)组成。跨摄像头下的目标关联即根据两个摄像头下观察到的分离的目标外观相似程度，在满足时空约束关系的前提下，将属于现实中同一目标的两个观察对象进行合并，得到目标跨越多个摄像头的连续完整运动轨迹。

令 $OUT_i^{t_p}=\left\{x\middle|x\in O_i^{t_p-1}\wedge x\notin O_i^{t_p}\right\}$ 表示 t_p 时刻刚离开摄像头 C_i 的目标集合，$IN_i^{t_p}=\left\{x\middle|x\notin O_i^{t_p-1}\wedge x\in O_i^{t_p}\right\}$ 表示 t_p 时刻新进入摄像头 C_i 的目标集合，$LEFT_i^{t_p}=LEFT_i^{t_p-1}\cup\left\{OUT_i^{t_p}\right\}$ 表示截止 t_p 时刻离开摄像头 C_i 后尚未在其它摄像头中找到匹配的目标集合。

令 $k_{i,a}^{j,b}$ 表示 $\left(o_{i,a}^{t_q},o_{j,b}^{t_p}\right)$ 在 t_p 时刻一次成功的目标关联，其含义为 $o_{i,a}^{t_q}\in LEFT_i^{t_p}$ 与 $o_{j,b}^{t_p}\in IN_j^{t_p}$。$(t_q<t_p)$为现实世界中 t_q。时刻离开摄像头 C_i 且在 t_p 时刻进入摄像头 C_j 的同一目标。

假设 t_p 时刻，若 $\bigcup_{i=1}^{n}LEFT_i^{t_p}\neq\varnothing$ 且 $\bigcup_{i=1}^{n}IN_i^{t_p}\neq\varnothing$，则触发行人即时对齐。即，寻找一个最优关联集合 $K=\left\{k_{i,a}^{j,b}\right\}$，$k_{i,a}^{j,b}\in K$ 当且仅当 $\left(o_{i,a}^{t_q},o_{j,b}^{t_p}\right)$（$o_{i,a}^{t_q}\in\bigcup_{i=1}^{n}LEFT_i^{t_p}$，$o_{j,b}^{t_p}\in\bigcup_{i=1}^{n}IN_i^{t_p}$)属于现实世界中两次连续出现的同一行人$(t_q<t_p)$，且满足[15]：

$$\forall k_{i,a}^{j,b},k_{u,c}^{v,d}\in K,k_{i,a}^{j,b}\neq k_{u,c}^{v,d}\Rightarrow(i,a)\neq(u,c)\wedge(j,b)\neq(v,d)\quad(1)$$

式(1)保证最优关联集合 K 中，任一目标行人至多被关联一次。

3.2 基于最小费用流模型的行人即时对齐

给定 $o_{i,a}^{t_q}\in\bigcup_{i=1}^{n}LEFT_i^{t_p}$，$o_{j,b}^{t_p}\in\bigcup_{i=1}^{n}IN_i^{t_p}$，令 $W(k_{i,a}^{j,b})$ 为 $\left(o_{i,a}^{t_q},o_{j,b}^{t_p}\right)$ 的关联适配度，其值越大，则意味着两者为同一目标的可能性越大。由此，t_p 时刻下的行人即时对齐问

题本质上为求解具有最大关联适配度 $arg\ max\sum W(k_{i,a}^{j,b})$ 的优化问题。

该优化问题可转化为最小费用流模型求解。首先，$\bigcup_{i=1}^{n} LEFT_i^{t_p} \neq \varnothing$ 和 $\bigcup_{i=1}^{n} IN_i^{t_p} \neq \varnothing$ 中每一个目标在网络中对应进出两个节点 x^- 和 x^+。为保证最终求解结果满足约束条件(1)，两节点间存在一条容量为 1，费用为 0 的有向 $e\left(x^-, x^+\right)$，其次，若 $W(k_{i,a}^{j,b}) > 0$，则节点 $o_{i,a}^-$ 和节点 $o_{j,b}^+$ 间存在一条容量为 1，费用为 $W\left(k_{i,a}^{j,b}\right)$ 的有向边 $e\left(o_{i,a}^+, o_{j,b}^-\right)$。最后，添加一个源点和一个汇点，且源点与每个行人的入节点连接，每个行人的出节点与汇点连接，相应有向边均容量为 1，费用为 0。通过使用最小费用流算法[27]对该费用流图进行最大费用流求解，即可获得最优关联集合 K。

然而，与延时关联方式下，将有单摄像头下目标轨迹对应两个节点，即可一次性构建费用流图，求解后即可获得全局的目标关联的方法[7, 15]显著不同的是，行人即时对齐意味着每一时刻集合 $\bigcup_{i=1}^{n} IN_i$ 和 $\bigcup_{i=1}^{n} OUT_i$ 处于动态变化之上，相应的费用流图也需要即时更新。

为此，本文采用最小费用流模型即时更新策略。给定 t_p-1 时刻下的完成即时对齐后费用流图 $G_{t_p-1}^-$，当 t_p 时刻，为集合 $\bigcup_{i=1}^{n} IN_i^{t_p}$ 和 $\bigcup_{i=1}^{n} OUT_i^{t_p}$ 中每一个目标新增进出两个节点，更新新增节点与源点、汇点间有向边连接；进一步，计算集合 $\bigcup_{i=1}^{n} IN_i^{t_p}$ 和 $\bigcup_{i=1}^{n} OUT_i^{t_p}$ 中两两目标间的关联适配度，并更新相应节点间有向边。由此，得到 t_p 时刻新的费用流图 G_{t_p}。求解完成后，删除所有对齐目标节点以及集合 $\bigcup_{i=1}^{n} IN_i^{t_p}$ 剩余未对齐的目标节点，并将集合 $\bigcup_{i=1}^{n} IN_i^{t_p}$ 剩余结点标记为新进入的目标行人，得到费用流图 $G_{t_p}^-$，等待下一时刻更新对齐。

显然，该算法的对齐结果很大程度上依赖于行人间关联适配度的定义。本文认为，关联适配度应由三个部分组成，第一个部分是观察到的两个行人外观上的适配度 P_{app}，第二个部分是两个行人空间上的转移关系 P_{pos} 是否成立，第三个部分是两个行人时间上的转移关系 P_{st} 是否成立。即，

$$W\left(k_{i,a}^{j,b}\right)=P_{app}\left(o_{i,a}^{t_q^-}, o_{j,b}^{t_p^+}\right)\times P_{pos}\left(o_{i,a}^{t_q^-}, o_{j,b}^{t_p^+}\right)\times P_{st}\left(o_{i,a}^{t_q^-}, o_{j,b}^{t_p^+}\right) \quad (2)$$

式(2)中三部分适配度具体定义见表 1。

本文为目标从摄像头 C_i 转移到 C_j 设置了时间间隔分布 $[\varepsilon_1^{i,j}, \varepsilon_2^{i,j}]$，其取值与文献[16]类似使用高斯分布来获得，由此考察两个关联目标时间上的转移关系是否成立。

参数	描述	定义
P_{app}	外观适配度	$P_{app} = Sim\left(o_{i,a}^{t_q}, o_{j,b}^{t_p}\right)$.
P_{pos}	空间适配度	$P_{pos}\left(o_{i,a}^{t_q}, o_{j,b}^{t_p}\right)=\begin{cases}1, & \text{外围进入和离开,}\\ 0, & \text{其他.}\end{cases}$

P_{st}	时间适配度	$P_{st}\left(o_{i,a}^{t_q}, o_{j,b}^{t_p}\right)=\begin{cases}1, \varepsilon_1^{i,j} < t_q - t_p < \varepsilon_2^{i,j}, \\ 0, \quad \text{其他.}\end{cases}$

表 1 行人关联适配度参数定义

为考察两个关联目标空间上的转移是否符合摄像头拓扑结构，文献[7-8, 16]中采用了限制目标行人进出区域的方法。然而，摄像头网络拓扑结构未知的情况下，此类方法不具有求解普适性。为此，本文将单个摄像头的跟踪视野区域划分为核心区域和临界区域，如图 3 中黄色和绿色部分所示。显然，任何目标的进入和离开都必先经过临界区域，因此，本文只考察临界区域内的行人具有合理空间转移关系，以最大程度地保证后续行人对齐求解的普适性。

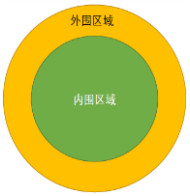


图 3 摄像头视野区域划分

3.3 零样本行人即时对齐与鉴别行特征学习

由式(2)可知，行人外观适配度是关联适配度的最终表征。即时对齐无法从完整的单摄像头下的跟踪轨迹中获取全面的外观特征描述，势必要求能从单帧中获取目标更加本质的特征刻画。因此，本文拟通过深度学习的高层抽象特征学习能力对于不同的行人 $o_{i,k}^{t_p}$，由 $boundbox_{i,k}^{t_p}$ 提取特征 $feature_{i,k}^{t_p}$，并度量两两间外观适配度 P_{app}。

需要特别指出的是，使用深度学习度量行人外观适配度面临着零样本问题。因为现实场景中我们无法完备的采集到所有出现在摄像头视野区域内的行人数据。因此，训练是极其不完备的，训练数据集中的行人并不会出现在真实采集到的视频中，而真实采集到的视频中需要对齐的行人也从未出现在训练数据集中。

为此，面向实际应用中零样本行人即时对齐的挑战，本文整合行人分类模型和身份认证模型，基于深度卷积孪生网络构建了一个行人单帧图像鉴别性特征学习模型，以提高行人关联准确性。

3.1 零样本行人身份即时对齐基本思路

借助深度学习求解行人外观适配度有两种不同的思路.

第一种是基于行人分类的思路， 借助 CNN 特征学习能力， 通过大型的行人图像数据集， 构建一个行人分类模型， 其中的最后一层卷积层即为特征. 给定

两张行人检测框图像，抽取各自卷积特征后，计算其欧式距离，即为特定特征嵌入空间内两行人间的外观相似度，可表征外观适配度.

第二种是基于行人身份认证的思路，成对输入行人图像，判定是否属于同一行人，通过深度网络的度量学习能力，最终构建一个二分类的鉴别模型. 给定成对输入的两张行人检测框图像，可通过部分匹配[17-18]或者对比损失[19]直接计算两个嵌入向量之间的欧氏距离，同样可用于表征外观适配度.

然而，上述两种思路各有所缺. 基于行人分类的思路中，输入是相互独立的，并没有显式地考虑两个行人之间的相似性度量，至多通过引入交叉熵损失学习嵌入向量之间存在的隐式关系. 基于行人身份鉴别的思路中，倒是直接学习了两个行人之间的外观相似性度量，但是学习局限于是否同属于同一行人的弱标签，只考虑了有限行人图像对之间的关系，缺乏全局层面类别区分特征的学习.

尽管基于行人分类和身份认证的思路都能学习到一定程度上的高层特征，但其嵌入空间对训练图像数据集的依赖性较高；考虑到实际应用中，行人即时对齐的零样本特性，其中的特征学习模型的表征能力仍需进一步提升.

为此，本文提出将分类模型与身份认证模型合二为一，融合鉴别模型在相似性度量学习上的优势和分类模型在类别区分特征学习上的优势，互相提升，从而获取更具有鉴别性的行人外观特征描述向量. 具体地，本文基于孪生网络模型，引入 Square 层[20]结合认证和分类两个模型，同时使用行人的类别标签信息与相似度信息，以学习到更具有区分性的特征.

3.2 鉴别性特征学习模型

本文提出的鉴别性特征学习模型整体结构如图 4 所示. 该模型为一个卷积孪生网络，对于给定的一对行人检测框对象，将同时预测两个行人的预类别 ID 和两张图片的相似性值.

图 4 中，上下两个网络为卷积网络，通过权重共享的方式组成，在两个基础网络输出的高层特征后，各自使用核大小为$1\times1\times4096$的卷积对基础网络输出的高层特征进行计算. 经过卷积层计算后的结果分别作为两个分类损失和 Square 层的输入值. 分类损失计算完后将分别得到两个行人的 ID. Square 层用于结合鉴定和分类损失，其定义为 $f_s=(f_1-f_2)^2$. 输入 Square 层的值经过计算后再次输入一个核大小为$1\times1\times4096$的卷积，而经过卷积层计算后的值最终作为认证损失的输入值，并计算输出两个行人的相似性值.

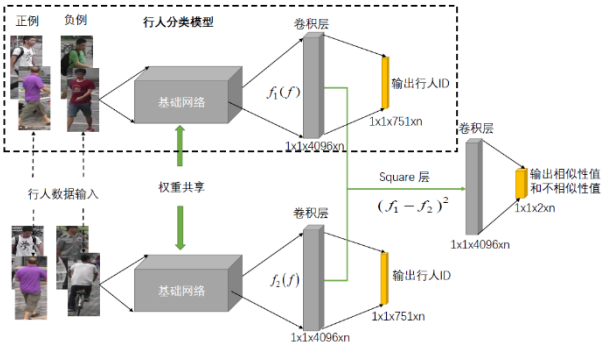


图 4 行人鉴别性特征学习模型

考虑到 ResNet 网络[21]通过明确地将层作为输入学习残差函数，不仅较好的控制了参数的数量，并且通过瓶颈结构和特征图逐层递进的方法，保证了输出特征的表达能力. 因此，图 4 中的基础网络以 ResNet 网络作为改进的基本结构. 改进后的网络结构称为 R-ResNet，具体结构如图 5 所示.

为了提取到更加有效的行人特征，R-ResNet 移除了 ResNet 最后所有的全连接层，得到图 5 中的 ResNet⁻网络. 经过 ResNet⁻网络得到的特征图将同时作为 3 个网络层的输入数据. 网络层 C1、网络层 C2 和网络层 C3 都是由 1024 个特征图组成的卷积层，每个卷积层的特征图大小为3×3 、5×5 和 7×7 . 之后将 3×1024 个特征图，通过连接层进行通道合并. 经过一个大小为 4×4 的邻域连接的池化层，池化层后为特征图大小为 2×2 的 2048 个特征图的卷积层 C4.

训练时，本文在 Market-1501[36]行人重识别基准数据集上进行训练，数据集包含 1306 个身份的 32668 个带注释边界框.

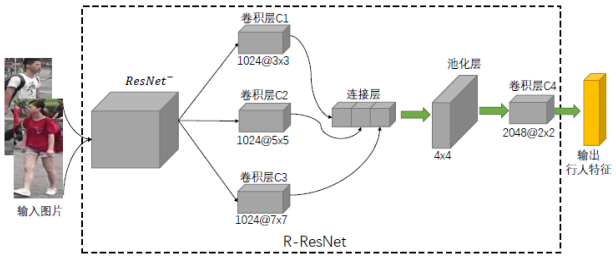


图 5 R-ResNet 网络结构

3.4 实验结果与分析

表 3 中给出了实验一的对照结果，从中可知，本文的即时行人对齐方法已经优于大多数延时关联下的算法，仅略低于效果最好 UW_IPL[16].

事实上，采用延后关联方法可以通过对单摄像头下的目标运行轨迹中获得的大量特征进行选择性的学习来不断提高外观模型的鲁棒性，降低噪声特征带来的影响从而提高对齐精度. 相较之下，本文提出的即时行人对齐方法仅使用

目标的单帧画面学习特征，在单摄像头的行人信息利用上具有较大的劣势．然而得益于采用了鉴定性行人判别模型对目标特征进行学习，使得本文的行人特征模型可以获得具有较强的识别性和区分性的行人特征，从而降低了不同摄像头配置和外部环境差异带来的影响，提高了行人外观相似性度量的准确率，间接的提高了最终对齐精度.

为了更加充分地说明鉴别性特征学习模型的有效性，本文同时给出了分别使用单摄像头下首帧和尾帧进行特征提取和相似性度量下的对齐结果．实验结果表明，无论是使用首帧还是尾帧的情况下，本文都获得了较好的实验结果，进一步验证了本文方法提取到的行人外观特征具有较强的鉴别性.

表 4 中给出了实验二的对照结果，可以看到基于本文提出的行人即时对齐下的跨摄像头跟踪性能已经超过当前的大部分基于行人延后关联的跨摄像头跟踪算法.

特别需要指出的是，与实验一中各方法均使用标定好的行人检测框和跟踪轨迹完成行人对齐不同，实验二中各方法的行人检测和单摄像头跟踪精度各不相同．本文的行人即时对齐需要基于单摄像头下行人首帧检测框完成．从表 4 中可以看出，本文最终实现的跟踪算法中 *Detection* 性能显著弱于 USC-Vision 和 UW_IPL，但是本文的行人即时对齐下的 $Tracking^{ICT}$ 性能仍然与 USC-Vision 和 UW_IPL 较为接近，甚至 Dataset3 中在 USC-Vision 算法获得了较高的单摄像头下的检测和跟踪精度的情况下，本文的 $Tracking^{ICT}$ 性能显著优于其.

也正是由于本文即时对齐方法具有较好的 $Tracking^{ICT}$ 性能，才能使得本文最终实现的跟踪算法在 *Detection* 和 $Tracking^{SCT}$ 性能都不敌 USC-Vision 算法的情况下，最终平均跨摄像头跟踪精度仅仅略低于 USC-Vision 的结果.

数据集	评价指标	本文 首帧	本文 尾帧	CRIPAC-MCT	EGTracker	USC-Vision	Hfutdspmct	UW_IPL
Dataset1	mme^c	25	54	113	55	27	86	13
	MCTA	0.925	0.838	0.667	0.835	0.915	0.742	0.961
Dataset2	mme^c	49	96	167	121	34	141	30
	MCTA	0.880	0.764	0.591	0.703	0.913	0.654	0.926
Dataset3	mme^c	55	21	44	39	70	40	32
	MCTA	0.638	0.862	0.711	0.741	0.516	0.736	0.789
Dateset4	mme^c	68	100	110	157	72	155	62
	MCTA	0.734	0.609	0.570	0.384	0.705	0.394	0.758
平均		0.795	0.768	0.633	0.666	0.762	0.632	0.858

数据集	评价指标	本文方法	CRIPAC-MCT	EGTracker	USC-Vision	Hfudspmcr	UW_IPL
Dataset1	*precision*	0.673	0.148	0.796	0.691	0.711	0.772

	recall	0.476	0.215	0.592	0.606	0.346	0.608
	Detection	0.558	0.175	0.679	0.646	0.465	0.681
	$Tracking^{SCT}$	0.956	0.995	0.974	0.998	0.922	0.998
	$Tracking^{ICT}$	0.893	0.711	0.622	0.928	0.653	0.885
	MCTA	0.476	0.124	0.412	0.695	0.281	0.601
	precision	0.793	0.143	0.797	0.695	0.746	0.833
	recall	0.506	0.193	0.633	0.784	0.366	0.709
	Detection	0.618	0.164	0.705	0.737	0.491	0.766
Dataset2	$Tracking^{SCT}$	0.994	0.994	0.977	0.999	0.934	0.999
	$Tracking^{ICT}$	0.837	0.751	0.694	0.869	0.612	0.884
	MCTA	0.514	0.107	0.479	0.626	0.281	0.676
	precision	0.485	0.085	0.820	0.475	0.334	0.659
	recall	0.514	0.120	0.534	0.662	0.009	0.726
	Detection	0.523	0.099	0.646	0.553	0.018	0.691
Dataset3	$Tracking^{SCT}$	0.967	0.971	0.974	0.990	0.968	0.986
	$Tracking^{ICT}$	0.259	0.114	0.295	0.101	0.243	0.546
	MCTA	0.131	0.011	0.186	0.055	0.035	0.372
	precision	0.728	0.060	0.835	0.522	0.772	0.875
	recall	0.508	0.099	0.619	0.794	0.121	0.860
	Detection	0.598	0.074	0.710	0.629	0.222	0.867
Dataset4	$Tracking^{SCT}$	0.957	0.976	0.927	0.995	0.986	0.997
	$Tracking^{ICT}$	0.498	0.295	0.430	0.544	0.294	0.627
	MCTA	0.285	0.021	0.284	0.340	0.060	0.542
平均		0.352	0.066	0.341	0.405	0.164	0.548

3.5 本章小结

本文针对目前延后关联算法时序约束弱化的本质缺陷， 提出了跨摄像头下多行人即时对齐的解决思路和基于实时最小费用流图行人即时对齐模型， 并针对行人即时对齐的零样本特性， 提出了基于孪生卷集网络的鉴别性特征学习模型， 将行人分类和身份认证合二为一， 在习得更具有区分能力的鉴别性特征的同时解决了行人外观相似性度量问题. 跨摄像头行人跟踪的数据集 NLPR_MCT 上的组合实验结果表明， 本文所提出的即时对齐方法已经优于绝大部分采用行人

延后关联的算法，仅略稍逊色于当前精度最高的 UW_IPL.

第四章 跨摄像头的光照迁移

4.1 引言

对于跨摄像头跟踪来说，行人外观是最经常被使用到的信息之一。对于外观线索，通常是使用一种或者融合多种特征来表示一个目标的外观信息。然而，跨摄像头下的行人外观线索很容易受到一些外界因素的影响，不同的摄像头由于不同摄像头放置的位置、视角和设备的物理性质差异造成了所采集到的监控视频图像的光照强度具有较大的差异。据我们所知，基于颜色的行人外观特征是跨摄像头行人跟踪中最经常被使用的信息之一。但颜色特征却很容易受到光照的影响，往往会造成同一个摄像头中出现的不同类别行人的相似性大于不同摄像头中的同类行人。

从上述可以，消除光照带来的影响成为了跨摄像头下行人对齐的重要且必须的步骤。目前对如何消除光照所带来的影响，主要有两种思路方向。第一种是通过简单的学习光照迁移函数(BTFs)[37]，当一个目标离开前一个摄像头再次进入下一个摄像头时候，通过关照转移函数对目标对不同摄像头下的关照进行迁移。Javed 等[]，证明了所有的关照转移函数都是基于在低维度的子空间中处理从一个摄像头到另一个摄像头时目标的关照变化的过程，并且在这个低维度的子空间中还可以计算两个目标间的相似性值。他们通过使用概率主成分分析算法在训练集中为每对摄像头都学习一个光照迁移函数。然而，他们的方法依赖于亮度范围变化不大的训练数据集，以给出一个准确的平均颜色迁移值。为了扩展[]的工作，Prosser 等人，他们通过累计统计的方法对目标的整体亮度进行计算，通过使用所有的摄像头下的数据集进行计算，而不是仅针对每对摄像头对转移函数进行训练。

虽然，这些关照迁移函数可以提高最终的行人识别精度，但他们的精度只有在给出的训练集在一个合适的亮度范围内和训练集之间具有相关性，才可以取得较好的效果。当一个较小的训练集或者数据间没有较高相关性，此时算法的性能会有巨大的下降。此外，当有一对摄像头并未在训练阶段获得转移函数时，其就无法在算法测试阶段完成摄像头间的光照迁移。并且对于已经训练好的每对摄像头的迁移函数，在摄像头的光照发生变化的时候，需要重新采集数据并对这对摄像头重新训练。因此，这些方法并不适用于在实际场景中光照不可控的应用系统中。

另一种方法的主要思路是通过校正每个摄像头下采集到的图片亮度，即通过光源恒量方法对图片亮度进行计算改变。然而，这些颜色恒常算法很复杂，需要

一个具有已知光源的图像数据集用于校准，在跨摄像头下很难得到满足。而只有不需要训练的颜色恒定算法具有已知光源的数据集可以应用于跨摄像头环境，如Gray-world[]、max-RGB[]、shades of gray[]和gray-edge[]等。

因此，本章通过针对目前已有的算法进行分析他们的各自优缺点，提出了我们的基于模糊聚类的光照迁移算法。其不仅不需要提前对模型进行预训练，而且通过隶属度因子改善了传统的迁移方法中对图像元素非此即彼的硬划分方法缺点。

4.2 光照迁移模型

首先把两个不同摄像头下采集到的目标图像转换到$l\alpha\beta$颜色空间；根据模糊聚类算法对两个目标图像的颜色特征进行分析，并将根据分析结果将图片划分为不同的特征聚类域；通过分别计算两个目标图像每个聚类域的匹配权值，根据匹配权值对两个目标的每个聚类域进行匹配；最后使用隶属度控制因子对目标图像亮度传输程度进行控制，并获得最终的光照迁移结果图[38]。

4.2.1 匹配聚类域的划分

与传统整张图片仅根据均值和标准差进行直接传输不同，我们先对目标图片按相似性进行聚类分析，进一步提高亮度转移的可靠性。经典的聚类算法有很多，例如，最长被使用的无监督算法K-means、ISODATA算法等。虽然，这些算法那在对图像进行按不同类别进行分类的时候可以获得较好的效果，但是这些算法对目标图像的像素进行严格的划分归属类别。而对于亮度迁移这种特定的任务来说，有些像素是存在不确定性和混合元素问题，并不能简单的进行严格硬划分。

我们使用模糊聚类算法(FCM)对图像按不同聚类域进行划分，由此引入隶属度控制因子来适应光照迁移这种复杂的场景图像聚类域划分。在FCM中隶属度因子被定义为0~1之间的实数，使得在对图像每个像素进行划分时不再只是非此即彼的进行硬划分聚类。

由于图像的RGB颜色空间的三通道间具有具有较大的相关性，当对其中一个通道进行改变时候，其他通道也要跟着改变。因此，我们在对图像进行聚类域划分时先将图片的颜色空间从RGB转换到$l\alpha\beta$空间中。其转换公式如下所示：将RGB的颜色空间转换成LMS的颜色空间，转换公式如（1）所示。

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.3811 & 0.5783 & 0.0402 \\ 0.1967 & 0.7244 & 0.0782 \\ 0.0241 & 0.1288 & 0.8444 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

为了 LMS 颜色空间所带来的数据偏差，我们通过对将 RGB 转换为 LMS 颜色空间的数据，取对数来减小存在的偏差。其次，通过将 LMS 颜色空间再次转换为$l\alpha\beta$颜色空间，转换公式如（2）所示。

$$\begin{bmatrix} l \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{6}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -2 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} L \\ M \\ S \end{bmatrix} \quad (2)$$

当对光照迁移处理结束后，我们需要将颜色空间$l\alpha\beta$转换为 RGB 颜色空间，以便我们接下来的使用。和开始的转换步骤相反，通过将转化为 LMS 颜色空间，再通过将 LMS 颜色空间转换为 RGB 颜色空间，得到光照迁移后的 RGB 图像。转换公式分别如（5），（6）公式所示。

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -2 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 & 0 \\ 0 & \frac{\sqrt{6}}{6} & 0 \\ 0 & 0 & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} l \\ \alpha \\ \beta \end{bmatrix} \quad (5)$$

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 4.4679 & -3.5873 & 0.1193 \\ -1.2186 & 2.3809 & -0.1624 \\ 0.0497 & -0.2439 & 1.2045 \end{bmatrix} \begin{bmatrix} L \\ M \\ S \end{bmatrix} \quad (6)$$

使用 FCM 进行聚类域划分的方法主要如下：假设目标图像大小为$W \times H$，则需要进行匹配聚类分析的像素个数为$N = W \times H$，N 个像素用符号记为$\{p_1, p_2, ..., p_N\}$，在$l\alpha\beta$颜色空间中像素$p_i = \{p_i^l, p_i^\alpha, p_i^\beta\}$。假设把目标图像划分为 C 类，每个类别的聚类中心$V = \{v_1, v_2, ..., v_C\}$。定义聚类中心为$v_j$的像素$p_i$隶属度表示为$u_{ji}$，则目标图像的隶属度矩阵为$U = [u_{ji}]_{C \times N}$。对两个目标图像的匹配聚类划分过程，如算法 x 所示。

Algorithm 1. FCC

Input: source and target images in $l\alpha\beta$ space

Initialization:

Number of categories C, weighted index m, maximum number of iterations LOOP

Cluster centers U, membership degree V, minimal error $\sum$

while T < LOOP **do**

  for k=1 to N **do**

    if $\forall v_i \Rightarrow p_k \neq v_i$, i = 1, 2, ⋯, C **then**

$$d_{ik} = \sqrt{(l_i - l_k)^2 + (\alpha_i - \alpha_k)^2 + (\beta_i - \beta_k)^2}$$

$$u_{ik} = \left[\sum_{j=1}^{C} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}} \right]^{-1}$$

    end if

    if $p_k = v_i$ **then**

      for j=1 to C **do**

        if $p_k = v_j$ **then**

$$u_{jk} = 1$$

        else

$$u_{jk} = 0$$

        end if

      end for

    end if

  end for

  for i=1 to C **do**

$$v_i' = \left(\sum_{k=1}^{N} (u_{ik})^m p_k \sum_{k=1}^{N} (u_{ik})^m \right)$$

  end for

  for i=1 to C **do**

    if $\|v_i - v_i'\| < \sum$ **then**

      break

    end if

  end for

end while

Outputs: result image of color brightness transfer in space $l\alpha\beta$.

4. 2. 2 匹配聚类域的选择

在使用 FCM 对两个目标图像按不同类别聚类后，将对两张图片的聚类域分别

进行匹配，为了更加精确的匹配，将为每个聚类域设置对应的权值参数 w，根据两个目标图像相似的权值参数设置为一对匹配对象。通过对每个聚类域中 l,α,β 3 个通道的标准差求解加权平均值作为权值参数 w 的值。对于第 k 个聚类域其具体计算公式如下所示：

$$w=\frac{1}{3}\sigma_k^l+\frac{1}{3}\sigma_k^\alpha+\frac{1}{3}\sigma_k^\beta,$$

$$\sigma_k^t=\sqrt{\sum_{p_i\in C_k}u_{ki}\left(t_{p_i}-u_k^t\right)^2\Big/Z},t=l,\alpha,\beta.$$

其中，C_k 表示第 k 个聚类域，Z 为规范化加权因子，$Z=\sum_{p_i\in C_k}u_{ki}$ 。

4.2.3 基于 FCM 的光照迁移模型

假设将两个需要进行关照迁移的对象，分别设置为目标图和源图，当目标图中有像素 p_i 的归属域和源图中的聚类域 h 是一对匹配域，则通过关照迁移模型后得到的新值 $p_1^\dagger$ 可以经过下步骤计算得到：

于传统的颜色亮度迁移函数[]相似，首先先对源图和目标图求解各自对应的均值和标准差，而本文的方法则在 $l\alpha\beta$ 三个通道中分别计算各个聚类域的均值和标准差。然后需使用目标求得的各个聚类域中的均值减去每个聚类中所对应的像素值，之后乘以源图聚类域和目标图聚类域的标准差比值，并加上源图聚类域的均值。与传统关颜色亮度迁移不同的是，在上面步骤之后，我们为了对迁移的亮度程度进行了控制，通过使用 FCM 的隶属度因子。具体的计算公式如下所示：

$$\begin{cases} l'=\sum_{k=1}^{C}{}_T u_{ki}\left(\dfrac{\sigma_{S_h}^l}{\sigma_{T_k}^l\left(\iota_T-\mu_{T_k}^l\right)}+\mu_{S_h}^l\right) \\ \alpha'=\sum_{k=1}^{C}{}_T u_{ki}\left(\dfrac{\sigma_{S_h}^\alpha}{\sigma_{T_k}^\alpha\left(\alpha_T-\mu_{T_k}^\alpha\right)}+\mu_{S_h}^\alpha\right) \\ \beta'=\sum_{k=1}^{C}{}_T u_{ki}\left(\dfrac{\sigma_{S_h}^\beta}{\sigma_{T_k}^\beta\left(\beta_T-\mu_{T_k}^\beta\right)}+\mu_{S_h}^\beta\right) \end{cases}$$

4.3 面向关联的光照迁移模型

实验结果与分析

Dataset	Evaluatin metric	Ours		CRIPAC-MCT	EGTracker	USC-Vision	Hfutdspmct	UW_IL
		Origin	+F-CCT					
Dataset1	mme^c	25	20	113	55	27	86	13
	MCTA	0.925	0.940	0.667	0.835	0.915	0.742	0.961
Dataset2	mme^c	49	40	167	121	34	141	30
	MCTA	0.880	0.902	0.591	0.703	0.913	0.654	0.926
Dataset3	mme^c	55	50	44	39	70	40	32
	MCTA	0.638	0.671	0.711	0.741	0.516	0.736	0.789
Dateset4	mme^c	68	66	110	157	72	155	62
	MCTA	0.734	0.742	0.570	0.384	0.705	0.394	0.758
Average	*MCTA*	0.795	0.814	0.633	0.666	0.762	0.632	0.858

4.4 面向跟踪的关照迁移模型

实验结果与分析

4.5 本章小结

第五章 跨摄像头的行人细粒度关联

5.1 引言

5.2 基于注意力机制的行人鉴别模型

5.3 行人细粒度的即时关联模型

5.4 面向跟踪的行人即时对齐模型

5.5 本章小结

1、原始网络+应分割（比例分块） and 注意力机制模型
2、整个网络的建立流程(加上光照跑一个实验结果)

5.1 引言[22]

Local representations are computed typically by partitioning the person bounding box into cells, e.g., dividing the images into horizontal stripes [18-19] or grids[20-21], and extracting deep features over the cells. These solutions are based on the assumption that the human poses and the spatial distributions of the human body in the bounding box are similar. In real cases, for example, the bounding box is detected rather than manually labeled and thus the human may be at different positions, or the human poses are different, such an assumption does not hold. In other words, spatial partition is not well aligned with human body parts. Thus, person re-identificatin, even with subsequent complex matching techniques(e.g.[20-21])to eliminate the misalignment, is often not quite reliable. Figure 1 providesillusttrative example.

5.2 行人特征提取

5.3 行人相似性度量

5.4 行人局部特征匹配模型

[17]Unlike holistic color histograms, part representations can not only capture appearance differences but also the spatial layout of a person's look.

Moreover, we are also interested in checking which parts paly the most important roles for trajectories association across non-overlapping cameras. To this end, we separately remove the head, left arm, right arm, torso, left leg, and right leg to check the association accuracies variations. The results are demonstrated in Fig.6. As can be seen. Torso and arms are more important than others. Legs are the less important one since it might be the moving parts of a person, which change dramatically in shape and pose. Also, due to the resolution, the head and the lower parts of the body may provide less reliable features, and hence contribute little to the preson association process.

5.4.1 面向关联的行人局部特征匹配模型

实验结果与分析

5.5 本章小结

第六章 跨摄像头下行人身份对齐

结论

参考文献

[1] Chen W, Cao L, Chen X, et al. An equalized global graph model-based approach for multicamera object tracking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(11): 2367-2381.

[2] C. Huang, B. Wu, and R. Nevatia, “Robust object tracking by hierarchical association of detection responses,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2008, pp. 788–801.

[3] C.-H. Kuo, C. Huang, and R. Nevatia, “Inter-camera association of multi-target tracks by on-line learned appearance affinity models,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2010, pp. 383–396.

[4] Fleuret F, Berclaz J, Lengagne R, et al. Multicamera people tracking with a probabilistic occupancy map[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(2): 267-282.

[5] Yu S I, Yang Y, Hauptmann A. Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3714-3720.

[6] Rosenhahn B, Pons-Moll G, Leal-Taixe L. Branch-and-price global optimization for multi-view multi-target tracking[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 1987-1994.

[7] Hofmann M, Wolf D, Rigoll G. Hypergraphs for joint multi-view reconstruction and multi-object tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3650-3657.

[8] Pflugfelder R, Bischof H. People tracking across two distant self-calibrated cameras[C]//Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on. IEEE, 2007: 393-398.

[9] Hu W, Hu M, Zhou X, et al. Principal axis-based correspondence between multiple cameras for people tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(4): 663-671.

[10] Cai Y, Medioni G. Exploring context information for inter-camera multiple target tracking[C]//Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on. IEEE, 2014: 761-768.

[11] Matei B C, Sawhney H S, Samarasekera S. Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 3465-3472.

[12] Piccardi M, Cheng E D. Multi-frame moving object track matching based on an incremental major color spectrum histogram matching algorithm[C]//Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on. IEEE, 2005: 19-19.

[13] Zhao R, Ouyang W, Wang X. Learning mid-level filters for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 144-151.

[14] Raftopoulos K A, Ferecatu M. Noising versus smoothing for vertex identification in unknown shapes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

2014: 4162-4168.
[15] Wang X, Doretto G, Sebastian T, et al. Shape and appearance context modeling[J]. 2007: 1-8
[16] Hamdoun O, Moutarde F, Stanciulescu B, et al. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences[C]//2nd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC-08). 2008: -.
[17] Cheng D, Gong Y, Wang J, et al. Part-aware trajectories association across non-overlapping uncalibrated cameras[J]. Neurocomputing, 2017, 230: 30-39.
[18] Yi D, Lei Z, Liao S, et al. Deep metric learning for person re-identification[C]//Pattern Recognition (ICPR), 2014 22nd International Conference on. IEEE, 2014: 34-39.
[19] Cheng D, Gong Y, Zhou S, et al. Person re-identification by multi-channel parts-based cnn with improved triplet loss function[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1335-1344.
[20] Ahmed E, Jones M, Marks T K. An improved deep learning architecture for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3908-3916.
[21] Li W, Zhao R, Xiao T, et al. Deepreid: Deep filter pairing neural network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 152-159.
[22] Zhao L, Li X, Zhuang Y, et al. Deeply-Learned Part-Aligned Representations for Person Re-identification[C]//ICCV. 2017: 3239-3248.
[23] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 34(7): 1409.
[24] HUANG Y F, Chu-Yang L I, Cai-Rong Y A N. Object Tracking for Multiple Non-overlapping Cameras Based on TLD Framework[J]. DEStech Transactions on Engineering and Technology Research, 2016 (ssme-ist).
[25] 苏松志，李绍滋，陈淑媛，等．行人检测技术综述[J]. 电子学报，2012, 40(4):814-820.
[26] T. Pfister, J. Charles, and A. Zisserman, “Flowing convnets for human pose estimation in videos,” in Proc. IEEE Int. Conf. Comput.Vis., 2015, pp. 1913–1921.
[27] W. Choi and S. Savarese, “A unified framework for multi-target tracking and collective activity recognition,” in Proc. Eur. Conf. Comput. Vis., 2012, pp. 215–230.
[28] W. Hu, T. Tan, L. Wang, and S. Maybank, “A survey on visual surveillance of object motion and behaviors,” IEEE Trans. Syst. Man Cybern. Part C-Appl. Rev., vol. 34, no. 3, pp. 334–352, Mar. 2004.
[29] X. Wang, “Intelligent multi-camera video surveillance: A review,” Pattern Recognit. Lett., vol. 34, no. 1, pp. 3–19, Jan. 2013.
[30] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, “Understanding transit scenes: A survey on human behavior-recognition algorithms,” IEEE Trans. Intell. Transp. Syst., vol. 11, no. 1, pp. 206–224, Jan. 2010.
[31] H. Uchiyama and E. Marchand, “Object Detection and Pose Tracking for Augmented Reality: Recent Approaches,” in Proc. Korea-Japan Joint Workshop Frontiers Comput. Vis., 2012, pp. 721–730.
[32] Luo W, Xing J, Milan A, et al. Multiple object tracking: A literature review[J]. arXiv preprint

arXiv:1409.7618, 2014.

[33] Zhang Y, Li S. Gabor-LBP based region covariance descriptor for person re-identification[C]//Image and Graphics (ICIG), 2011 Sixth International Conference on. IEEE, 2011: 368-371.

[34] Zhao R, Ouyang W, Wang X. Unsupervised salience learning for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3586-3593.

[35] Zheng W S, Gong S, Xiang T. Reidentification by relative distance comparison[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(3): 653-668.

[36] Chu C T, Hwang J N, Yu J Y, et al. Tracking across nonoverlapping cameras based on the unsupervised learning of camera link models[C]//Distributed Smart Cameras (ICDSC), 2012 Sixth International Conference on. IEEE, 2012: 1-6.

[37] Javed O, Shafique K, Shah M. Appearance modeling for tracking in multiple non-overlapping cameras[C]//Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 2: 26-33.

[38] 钱小燕，肖亮，吴慧中．模糊颜色聚类在颜色传输中的应用[J].计算机辅助设计与图像图形学学报, 2006, 18(9): 1332-1336.