基于深度学习的行人多目标跟踪方法

徐 涛[1,2],马 克[1,2],刘才华[1,2]

(1. 中国民航大学 计算机科学与技术学院,天津 300300;2. 中国民航大学 中国民航信息技术科研基地,天津 300300)

摘 要:综合了近年来基于检测跟踪的主流行人多目标跟踪方法,介绍了基于检测的行人多目标跟踪方法概念,从目标检测、特征提取和数据关联与跟踪三个阶段对行人多目标跟踪方法进行了概述,比较并评价了这些方法在MOTChallenge系列数据集上的性能,阐述了多目标跟踪的未来研究方向。

关键词:计算机视觉;多目标跟踪;目标检测;特征提取;数据关联

中图分类号:TP391 **文献标志码**:A **文章编号**:1671-5497(2021)01-0027-12

DOI:10.13229/j.cnki.jdxbgxb20200509

Multi object pedestrian tracking based on deep learning

XU Tao[1,2],MA Ke[1,2],LIU Cai-hua[1,2]

(1. *School of Computer Science and Technology, Civil Aviation University of China, Tianjin* 300300, *China*; 2. *Information Technology Research Base of Civil Aviation Administration of China, Civil Aviation University of China, Tianjin* 300300, *China*)

Abstract: A survey of the mainstream multi object tracking methods based on tracking by detection in recent years is carried out. Then, the concept of detection based multi object tracking is introduced. The multi object tracking methods are summarized in object detection, feature extraction and data association & tracking. The performance of some multi object tracking(MOT) methods are compared and evaluated on the MOTChallenge series datasets. The future development direction of multi object tracking is discussed.

Key words: computer vision; multi object tracking; object detection; feature extraction; data association

0 引 言

多目标跟踪(Multi object tracking,MOT)利用计算机视觉技术对视频中的目标进行持续跟踪,通过对输入的视频图像进行处理,得到多个目标,并对目标的外观特征、位置、运动状态等信息进行计算分析,最终得到连续的运动轨迹。在深度学习技术成熟之前,传统的多目标跟踪方法[1]建立高斯混合模型或隐马尔可夫随机场模型以构建生成外观模型,然后使用基于贝叶斯理论的模型或基于局部/全局数据关联的模型进行持续的目标跟踪。随着深度学习技术的发展和深入,多

收稿日期:2020-07-06.

基金项目:天津市自然科学基金项目(18JCYBJC85100);中央高校基本科研业务基金项目(3122018C024);中国民航大学科研启动项目(2017QD16X).

作者简介:徐涛(1962-),男,教授,博士.研究方向:智能信息处理,图像处理.E-mail:txu@cauc.edu.cn

通信作者:刘才华(1987-),女,讲师,博士.研究方向:机器学习,计算机视觉.E-mail:chliu@cauc.edu.cn

目标跟踪方法取得突破性发展,典型的深度学习方法如基于卷积神经网络(Convolutional neural network,CNN)的方法[2,3]、基于循环神经网络(Recurrent neural network,RNN)的方法[4,5]等在解决多目标跟踪问题上均取得了良好效果。

围绕基于深度学习的多目标跟踪研究进展,Xu等[6]在2019年总结分析了多目标跟踪领域相关动态,将基于深度学习的多目标跟踪大致分为三类:使用深度特征的多目标跟踪增强、具有深度网络嵌入的多目标跟踪和通过端到端深度神经网络学习进行多目标跟踪。Sun等[7]在2020年梳理了使用多目标跟踪的主流方法——基于检测的跟踪(Tracking by detection,TBD)的研究方法,并详细分析了这些方法的性能表现。本文按照基于检测的跟踪方法流程展开综述,不同于上述文献,本文从方法类型角度归纳分析目前主流多目标跟踪的解决方案,重点关注近年来最新的研究进展,并对领域未来的发展方向进行展望。

1 基于检测的跟踪基本概念

根据对跟踪目标初始化的方式,本文将多目标跟踪方法分为3类:基于检测的跟踪、无检测的跟踪和使用强化学习的跟踪。无检测的跟踪在视频第一帧手动选定跟踪目标,使用强化学习的跟踪同时进行目标检测和预测目标的跟踪轨迹。基于检测的跟踪是目前多目标跟踪领域的主流方法,其跟踪基本框架如图1所示,这类方法把多目标跟踪问题看作多序列多变量估计问题,通过目标检测、特征提取、数据关联与跟踪三个阶段进行持续的目标跟踪。

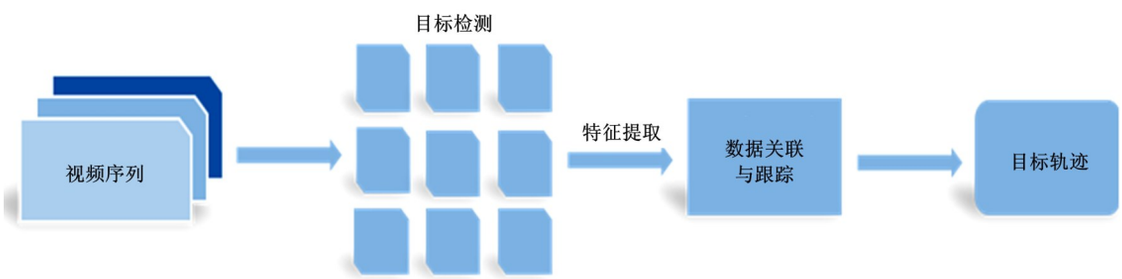

图1 基于检测的跟踪基本框架图

Fig. 1 Basic framework of tracking by detection

目标检测是多目标跟踪的首要任务,即检测输入视频中的目标,根据跟踪任务保留特定类别的、准确率较高的检测结果。一般都使用诸如Faster R-CNN[8]、SSD[9]、YOLO[10]等目标检测器进行检测。

特征提取方法的优劣是影响多目标跟踪算法性能的重要因素之一,要求提取的特征既能较好地描述跟踪目标又能快速进行计算。目前,主流的方法是使用如CNN、Siamese网络、长短时记忆网络(Long short-term memory,LSTM)等深度神经网络进行特征提取。

数据关联与跟踪通常是一个高度复杂的离散组合问题,即将检测到的目标和已生成的轨迹离散点(或下一帧中的相同目标)进行组合匹配,正确的匹配即为跟踪的结果,错误的匹配则需舍去或重新进行匹配计算。任何可能的目标匹配都应该满足一对一的约束,以避免将同一轨迹分配给不同的目标。

2 行人多目标检测与跟踪

2.1 目标检测

2.1.1 基于Faster R-CNN的目标检测

Faster R-CNN[8]在Fast R-CNN基础上引入了区域建议网络(Region proposal network,RPN)用于推荐检测的候选区域,其工作流程如图2所示:首先提取输入图像的特征并得到特征图。然后由RPN完成前景和背景的区别判断,最后在ROI池化层将建议区域映射为一个固定尺度的特征向量,并对候选区域进行不同类别的分类。Faster R-CNN通过引入一种有效定位目标区域的方法,并按区域在特征图上进行检测,大幅降低了卷积计算的时间消耗。

SORT[11]使用Faster R-CNN作为基础目标检测器,并使Faster R-CNN只关注行人分类而忽略其他所有类别,仅将输出概率大于50%的检测结果传递给跟踪框架,使跟踪结果的准确性得到很大提高。Yu等[12]认为,具有较高质量的检测结果可以降低对复杂跟踪算法的要求。他们提出的

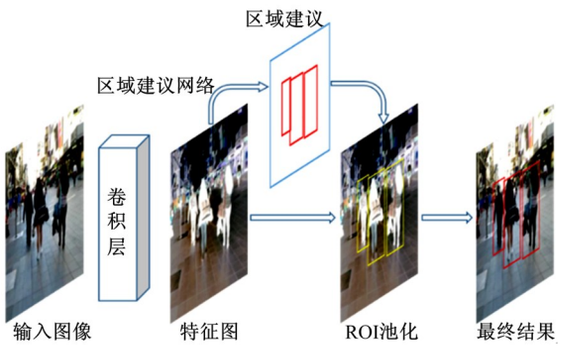

图 2　Faster R-CNN 网络结构

Fig. 2　Faster R-CNN network structure

方法基于 Faster R-CNN，采用随机采样动态尺度的多尺度训练策略，组合不同尺度和水平层次的特征。

2.1.2　基于 SSD 的目标检测

SSD[9]（Single shot multi-box detector）是基于单个 CNN 直接进行目标检测的目标检测器，通过设置不同尺度、不同分辨率的多个特征图处理各种大小的对象，提升检测的质量。SSD 均匀地在图片上的不同位置进行密集抽样，采用不同的长宽比，提取特征后直接进行分类和回归计算。但是均匀密集采样会使训练的正负样本极不均衡，导致 SSD 的准确性降低。

为了克服 SSD 的缺点，Kieritz 等[13]提出联合检测多目标跟踪（JDMOT）方法。通过在 JDMOT 的其他步骤中获取信息来缓解 SSD 的样本不均衡问题，并使用跟踪和检测之间的亲和力得分取代 SSD 中原有的标准非最大抑制（Non-maximumsuppression，NMS）。Dawei 等[14]使用 SSD 搜索场景中的行人和车辆，并且使用基于 CNN 的相关过滤器使 SSD 生成更准确的边界框。通过这种方式，Dawei 等[14]的方法能够从较小的检测中提取出更多有价值的语义信息。

2.1.3　基于 YOLO 的目标检测

YOLO[10]（You only look once）将对象检测建模为回归问题，“仅需看一次”即可同时、快速、直接地从完整图像中预测多个边界框和物体分类的概率。YOLO 网络结构如图 3 所示，卷积层用于提取输入图像的特征，全连接层用于计算物体对象的位置及所属类别的概率。YOLO 的缺点在于当检测的物体对象之间距离相隔较近，并且物体较小时，检测效果较差。Redmon 等[15]在 2018 年提出了 YOLOv3，使用新的分类网络，可以较为准确地检测出较小的物体，处理速度也得到了提升，但是识别物体的精确度有所下降。

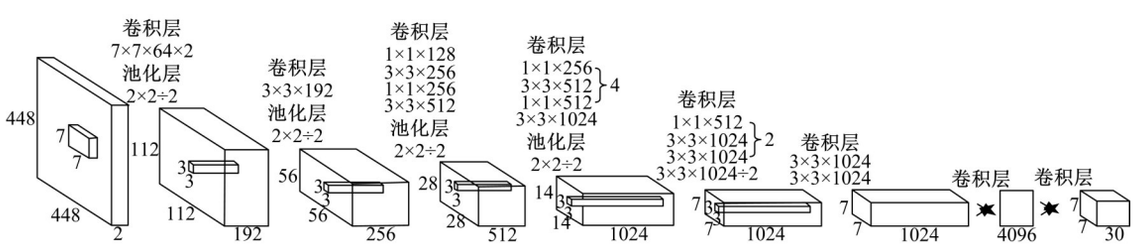

图 3　YOLO 网络结构

Fig. 3　YOLO network structure

Jiang 等[16]提出基于强化学习的 MADRL 方法，使用 YOLOv3 作为目标检测器，每一帧的图像被视为一个环境，每个单独的对象被视为一个主体，形成多主体系统，再使用多主体深度强化学习获得跟踪结果。He 等[17]对 YOLOv3 做出改进，使用双线性插值方法调整图像大小，使得 YOLO 网络检测小目标的能力提升，更加适合行人多目标跟踪的应用场景，提升了跟踪效率。为了有效解决行人被遮挡的问题，Zhou 等[18]提出深度对齐网络，使用 YOLOv3 检测目标对象获得置信度分数，使用深度对齐网络矫正错误检测。

2.1.4　其他检测器

除了上述几个著名的检测器之外，以下检测器也被应用在多目标跟踪的方法之中：

AVOD[19]：聚合视图对象检测网络 AVOD（Aggregate view object detection network）是一种适用于自动驾驶的网络，由 RPN 和第二阶段检测器网络两个子网构成。Baser 等[20]提出的 FANTrack 方法使用 AVOD 作为基础 3D 检测器，使提取的局部低级特征更具区别性，高级特征更加趋近语义特征。

OpenPose[21]：基于 CNN 的开源库，可以实现人体动作、面部表情等姿态估计，适用于单人或多人检测，拥有极好的鲁棒性。Ristani 等[22]使用 OpenPose 方法进行目标检测，并且通过裁剪和水平翻转等操作进行数据增强，以补偿检测器的定

位误差。

2.2 特征提取

2.2.1 基于CNN的特征提取

CNN是应用最为广泛的特征提取方法,Kim等[23]将预先训练的CNN视觉特征整合至经典算法中,提出多假设跟踪MHT算法,得到了较好的跟踪结果。Bewley等[24]提出DeepSORT,通过自定义的残差CNN网络提取视觉信息,在目标检测阶段引入运动特征和外观特征,使用余弦距离进行长期遮挡后的身份恢复。

Chu[2]等引入时空注意力机制,提出STAM方法,使用ROI池以及空间注意力机制提取候选检测的特征。Chu等[2]指出,ROI池会忽视输入的视频帧中目标被遮挡的问题,而引入空间注意力机制可以更加注意未被遮挡的区域。

通常行人外观的上、下部分有明显差异,为此He等[17]使用调整过的VGGNet,将行人分为上、下两个部分,分别提取特征,并且使用三重损失函数进行训练以获得行人外观的特征向量。

除了外观特征,还可以提取运动特征。Mahmoudi等[25]提出基于CNN的方法,引入运动位置特征。使用全连接层提取外观特征,利用最后一帧的平均速度和位置预测下一帧中每个目标的位置,得到运动位置特征。结合两种特征进行数据关联操作。

Sheng等[26]使用GoogleNet的卷积部分对每个检测的目标提取256维的特征,使用余弦距离计算检测对之间的亲和力得分,并将结果和运动预测整合以计算总体亲和力,该亲和力在图问题中作为边缘成本以供后续使用。

Chen等[27]提出的AP_HWDPL方法使用Faster R-CNN进行特征提取。该方法首先使用Faster R-CNN的顶层卷积层提取高级语义特征得到前景概率,再结合目标的历史外观及新提取的特征计算目标属于同一行人的概率。

Peng等[28]提出链式跟踪方法,使用两个ResNet-50提取相邻两帧图像的高级语义特征,再使用特征金字塔网络生成用于后续预测的多尺度特征表示,并将相邻两帧图像的特征图连接在一起,输入至预测网络以回归边界框。

He等[29]创新性提出通过生成动画跟踪方法,将输入图像转换为动画后进行多目标跟踪。该方法使用全连接层进行特征提取,然后利用可微分的渲染器进行动画转换,通过空间转换网络压缩和移动外观特征、合成图层、重建动画帧三个阶段将输入图像动画化。

Braso等[30]探索了在图域上进行多目标跟踪的方法,提出消息传递网络(Message passing network,MPN),将检测视为图的节点,检测间的关联视为图的边。使用CNN提取外观特征,并利用节点的边界框计算基于坐标的几何特征,然后跨图传递这些特征中包含的信息。

Porzi等[31]利用ResNet-50和特征金字塔网络组成的网络提取5种分辨率的多尺度特征,并将特征输入至区域分割模块(Region segmentation head,RSH)以预测检测对象的候选边界框,使用ROI Align提取边界框区域中特定于实例的特征以预测目标类的概率。

2.2.2 基于Siamese的特征提取

Siamese网络可以学习一组最具区别性的特征向量,其结构如图4所示,两个相同的分支网络用于提取输入图像的特征,共享相同的权重,简化了代码实现的难度,全连接层用于提取候选对象的特征向量。研究者可以从不同角度出发,使用Siamese网络提取所需特征。

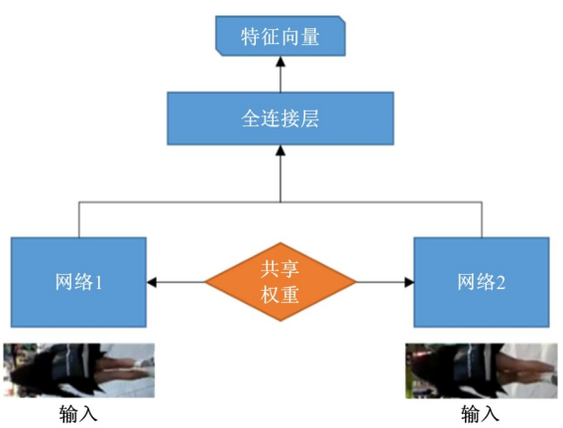

图4 Siamese网络结构

Fig. 4 Siamese network structure

Kim等[32]使用Siamese网络将外观相似性与时间几何信息相结合,同时学习外观特征和几何特征,可以有效地处理空间上相距遥远但共享相似特征的对象对(如两个相隔较远但衣着相似的行人)。

Lee等[33]提出特征金字塔Siamese网络,使用相同的参数从两个不同的图像中提取外观特征,然后采用上采样和合并策略为金字塔的每个阶段创建特征向量,同时引入时空运动特征以克服缺

少运动信息的缺点。

Leal-Taixé 等[34]使用 Siamese 网络对两组堆叠的图像进行训练,学习图像间的局部时空结构,聚合像素值和光流信息,并输出两个图像属于同一个人的概率,然后使用此概率对网络进行训练以得到最具代表性的行人特征。

Zhu 等[35]引入空间注意力网络 SAN,采用 Siamese 网络结构,使用预训练的 ResNet-50 作为基本网络结构,从网络最后的卷积层中提取出空间注意力图,以便利用提取的特征进行噪声检测和模糊处理。

Tang 等[36]提出 LMP 方法融合行人姿态信息进行行人重识别步骤,将行人重识别的网络框架分为 3 个分支,其中第二分支 SiameseNet 使用全连接层提取特征,然后通过两个全连接层将提取的特征串联,并使用 softmax 函数估计待跟踪行人是否属于同一个人的概率。

Yin 等[37]设计一种三元组网络,将网络中的正样本分支和锚分支视为 Siamese 网络,用于目标跟踪和运动预测。利用 AlexNet 提取正、锚分支的具有区别性的身份特征,同时提出任务特定的注意力模块(Task-specific attention,TSA)用于提高跟踪的准确性以及捕获细粒度的局部语义特征。

2.2.3　基于 LSTM 的特征提取

LSTM 可以有效整合目标的外观特征、运动特征和其他信息,但是在实际应用中不能存储长时信息序列。Kim 等[5]进一步探索了 LSTM 学习外观特征的能力,提出了双线性 LSTM(Bilinear LSTM,bLSTM)网络,使用乘法方式与输入耦合以取代传统的加法耦合,其网络结构如图 5 所示。Kim 等[5]使用传统 LSTM 方法提取运动特征,使用 bLSTM 提取外观特征,结合二者构建递归网络,当遇到检测不稳定或者丢失的情况,只需更新 bLSTM 的内部存储,便可对后续的跟踪计算做出快速反应。

为了表示不同的检测对象,Lu 等[38]提出关联 LSTM 直接回归对象的位置和类别,由 LSTM 过滤并计算关联特征。关联特征包括类别特征,即在 SSD 检测步骤中预测使用的,包含检测对象时空信息的特征,随后这些特征被用于亲和力计算。

Sadeghian 等[39]设计了一种基于 RNN 的网络结构,采用 3 种不同的 RNN 计算各种类型的特征。第一个 RNN 用于提取外观特征,其输入是由 VGGNet 提取的视觉特征向量,用于行人的重识别。第二个 RNN 是经过训练的 LSTM,可以预测每个跟踪对象的运动模型。最后一个 RNN 通过训练,学习场景中不同对象之间的相互影响。

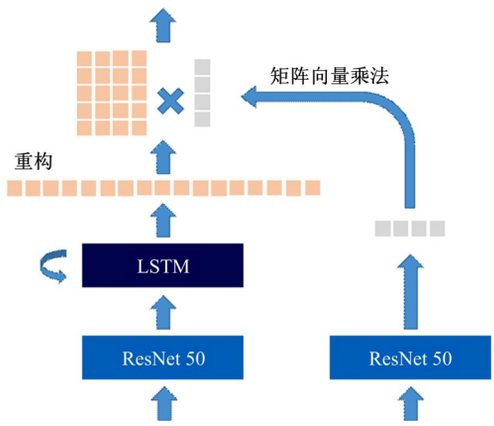

图 5　双线性 LSTM 网络结构

Fig. 5　Bilinear LSTM network structure

Chen 等[40]提出循环度量网络(Recurrent metric network,RMNet),利用 ResNet-18 的卷积部分建立自定义模型,并在卷积层顶部堆叠一个 LSTM 单元。RMNet 首先利用 ResNet-18 提取特征,然后将特征输入至 LSTM 单元,以便同时计算得到特征向量和相似性分数。

2.2.4　其他特征提取方法

为了避免网络过拟合,并且使外观特征准确度提高,Son 等[3]提出了一种新的 CNN 网络结构 Quad-CNN,将 CNN 的最后一个卷积层的输出分为两个部分并分别学习。同时 Quad-CNN 学习一组位置特征,同外观特征一起计算距离度量,该度量用于评估两种特征的相对重要性。

Rosello 等[41]提出基于强化学习的 MARLMOT 方法,训练一组有助于特征提取的主体。MARLMOT 没有采用任何视觉特征,仅使用了运动特征,并通过卡尔曼滤波学习运动模型。通过学习,主体学会使用一组信息(包括开始或停止跟踪)来决策卡尔曼滤波器应采取哪个动作(包括忽略预测,忽略新度量)。

2.3　数据关联与跟踪

2.3.1　匈牙利算法

匈牙利算法[42]是一个经典的在多项式时间内求解任务分配问题的组合优化算法,在多目标跟踪领域可用于处理前、后两帧的目标匹配问题。He 等[17]将数据关联算法分为约束问题和距离计

算问题,将约束矩阵和距离矩阵的总和用于匈牙利算法进行目标匹配。Zhou等[18]使用YOLOv3检测行人,结合空间、外观、运动信息建立关联矩阵,应用匈牙利算法逐帧解决分配问题。Rosello等[41]结合强化学习和匈牙利算法解决数据关联问题。SORT[11]和Kieritz等[13],或Deep SORT[24]和CNNMTT[25]等研究或方法均使用匈牙利算法解决数据关联问题。

Xu等[43]受多目标跟踪评估指标MOTA、MOTP(详见3.1节)的启发,提出端到端的数据关联方法:深度匈牙利网络(Deep hungarian net,DHN)。DHN由两个连续的双向RNN(Bidirectional RNN)构成,使用一个可微分的函数在真实目标和预测的轨迹之间进行匹配。作者声称他们的方法推进了多目标跟踪的端到端研究方法[35]的进步。

2.3.2 循环神经网络

循环神经网络RNN以顺序方式工作,给定先前状态和可能附加输入的情况下,在每个时间步进行预测。RNN的核心是大小为n的隐藏状态h,它充当预测输出的主要控制机制。RNN在运动预测和状态更新的任务上表现良好,使用RNN网络或其衍生网络可以很好地解决数据关联问题。

Fang等[4]基于RNN构建了循环自回归网络RAN,整合外观信息和运动信息以解决数据关联问题。RAN利用边界框计算运动动态特征,利用检测图像提取外观特征,根据条件概率结合二者计算检测和目标的关联得分,得分最大的一组目标即和检测进行关联。

Kim等[5]提出的bLSTM方法利用MHT解决数据关联问题。bLSTM生成多个跟踪建议,然后选择最可能的跟踪结果。同时还引入RNN作为门控网络,选择一些特定的检测更新和扩展跟踪建议(包含目标的运动信息和外观信息)。

Zhu等[35]提出对偶匹配注意力网络(DMAN)解决数据关联问题。DMAN使用跟踪分数衡量跟踪结果的可靠度,利用运动信息选择候选检测并计算目标相似度,然后使用时空双重注意力机制解决轨迹的遮挡和未对准问题。

Milan等[44]提出基于LSTM的网络结构RNN_LSTM,学习一对一的目标分配问题。RNN_LSTM首先构建时间循环神经网络学习动态模型,直接预测目标运动和状态的更新,然后使用LSTM预测每一次分配的目标,以得到预期的分配结果。Milan等[44]声称这是第一次把RNN网络应用在多目标跟踪中,并且是完全无模型的,不需要任何先验知识[44]。

Yoon等[45]使用深度神经网络架构解决数据关联问题。架构包含编码器、解码器两个部分。其中解码器由带有投影层的双向LSTM(Bidirectional LSTM)网络组成。网络最终输出一个关联矩阵,反映跟踪与检测之间的匹配分数。

2.3.3 强化学习

强化学习(Reinforcement learning)的目标是学习一个最优策略(policy),使主体(agent)在特定环境(environment)中根据当前状态(state)做出相应动作(action),从而获得最大回报(reward)。使用强化学习在跟踪过程中做出决策可解决数据关联的问题,如学习一组策略以预测目标的位置,然后决定何时做出删除错误的检测,更新正确检测的轨迹,或保留丢失检测等动作。

Jiang等[16]指出,多目标跟踪中学习数据关联的相似性函数相当于在MADRL中学习相关策略。目标检测的结果被视为多主体系统,主体的状态表示与动作历史记录关联。多主体系统使用Independent Q-Learners学习每个主体独立的策略,结合MADRL学习获得多个主体的联合动作得到跟踪结果。

Rosello等[41]使用多主体强化学习进行目标的跟踪管理,提出了信任区域策略优化(Trust-region policy optimization,TRPO)方法训练主体之间的共享策略,确定何时开始和停止跟踪并更新卡尔曼滤波器的操作,然后使用匈牙利算法进行数据关联。

Chu等[46]在其方法中使用了三种不同的CNN。第一个CNN用于区分背景和跟踪对象;第二个CNN用于区分不同的目标;第三个CNN用于检验关联结果的好坏。该方法采用强化学习进行训练,并使用匈牙利算法恢复被遮挡的目标。

2.3.4 其他数据关联方法

Peng等[28]在其提出的链式跟踪方法中,将数据关联问题转换为成对的对象检测问题。该方法计算前、后两帧中同一目标的IoU值,并应用Kuhn-Munkres算法[47]匹配目标前后两帧的边界框,生成目标轨迹。

He 等[29]将输入图像转换为动画进行跟踪，并提出了重优先化注意式跟踪方法（Reprioritized attentive tracking，RAT），共享跟踪目标之间的参数避免过拟合，并使用注意力实现显示的数据关联以解决跟踪中断的问题，并且根据场景中呈现的对象数量调整计算的时间以提高计算效率。

3　评估标准与数据集

3.1　评价指标

通常采用建立一套完整的度量评价指标来公平地测试和比较多目标跟踪算法的性能。目前研究者常用的指标由 Wu 和 Nevatia 定义的传统指标[48]，CLEAR MOT 指标[49]和 ID 指标[50]构成，表 1 列出了当前最常使用的多目标跟踪评价指标，其中向上的箭头（向下的箭头）表示该指标越高（越低），性能越好。

表 1　多目标跟踪评价指标

Table 1　Evaluation metrics for multi object tracking

度量名称	期望分值	简述
MOTA ↑	100%	多目标跟踪准确度
MOTP ↑	100%	多目标跟踪精度
MT ↑	100%	最多跟踪的目标
ML ↓	0%	最少丢失的目标
Frag ↓	0	跟踪被打断的总次数
IDSW ↓	0	身份切换的总次数
FP ↓	0	错误正样本数量
FN ↓	0	错误负样本数量
IDF1 ↑	100%	识别 F 值
Hz ↑	正无穷	处理速度（以 FPS 为单位，但不包括检测器的处理速度）

传统指标定义了多目标跟踪算法可能产生的错误类型，通常有：最多跟踪（Mostly tracked，MT）——跟踪器输出的轨迹覆盖至少 80% 的真实值的比率；最多丢失（Mostly lost，ML）——跟踪器输出的轨迹覆盖最多 20% 的真实值的比率。同时，研究者还会评估多目标跟踪算法的处理速度 Hz（即每秒帧数 FPS，Frames per second）。

CLEAR MOT 指标是 2006 年和 2007 年举办的 Classification of Events，Activities and Relationships（CLEAR）研讨会制定的，常用的评估指标如下：

中断数（Fragmentation，Frag）：一个跟踪被丢失的检测中断的次数。

错误正样本（False positives，FP）：整个视频中被预测为正的负样本数量。

错误负样本（False negatives，FN）：整个视频中被预测为负的正样本数量。

ID 切换总数（Identity switches，IDSW）：正确地跟踪了对象，但错误更改了该对象的身份标识的总次数。

多目标跟踪准确度（Multi-object tracking accuracy，MOTA）：以错误正样本（FP），错误负样本（FN）和 ID 切换总数（IDSW）为依据的跟踪准确度：

$$\mathrm{MOTA}=1-\frac{\sum(\mathrm{FN}+\mathrm{FP}+\mathrm{IDSW})}{\mathrm{GT}} \tag{1}$$

式中：GT 为所有真实值的数量。

多目标跟踪精度（Multi-object tracking precision，MOTP）：根据真实值和预测之间的边界框重叠的情况，计算的总体跟踪精度。该评估标准不考虑有关跟踪的信息，而是侧重于检测的质量。

MOTA 主要考虑跟踪器做出错误决定的次数，但是在某些特定应用场景（如考虑公共场所的安全），人们可能更关心跟踪器能否在尽可能长的时间内跟踪物体。因此，Ristani 等[50]定义了 ID 指标，作为上述指标的补充：识别精确度（Identification precision，IDP）、识别召回率（Identification recall，IDR）和识别 F 值（Identification F-Score，IDF1）。

3.2　数据集

本节将介绍多目标跟踪领域的重要基准——MOTChallenge，然后介绍其数据集的发展，最后介绍一些其他通用的数据集。

MOTChallenge：上传并公布多目标跟踪方法研究成果的公共平台，拥有最大的公开行人跟踪数据集。这些数据集都提供了训练集的标注，训练集与测试集的检测，以及数据集的目标检测结果。

MOT15：MOTChallenge 于 2015 年提供的数据集[51]，视频序列都是由静态或动态摄像机在不受约束的环境中拍摄的。

MOT16：MOTChallenge 于 2016 年提供的数据集[52]，使用了基于组件的可变性模型 v5[53]（Deformable Part-based Model v5，DPM）提供预先的检测结果，获得了更好的检测性能。

MOT17：与 MOT16 使用相同的视频序列，但是数据集中的每个视频使用 3 组检测：Faster

R-CNN,DPM和规模依赖池[54](Scale dependent pooling,SDP),因此检测更加准确。

表2节选了MOT15、MOT16和MOT17的部分视频序列,描述了视频序列的主要属性,同时列出了光照和天气条件等外在条件,以及每帧人物的平均密度。

MOT19:MOTChallenge于2019年提供的新版本数据集[55],拥有8个行人密度极高的视频序列,平均每帧可以达到150多名行人。

MOT20:在无限制的环境中拍摄的人群密度更高的数据集,包含夜间场景下的广场与体育场入口,拥挤的室内火车站。

KITTI[56]、PETS200911[57]、TUD10[58]和UA-DETRAC[59]:这些数据集是在现在的研究工作中较少使用的数据集,其中KITTI可以对行人和车辆进行跟踪,PETS200911和TUD10主要针对行人的跟踪,UA-DETRAC则专注于通过交通摄像头跟踪道路上的车辆。

表2 MOT系列数据集的视频序列及其主要属性

Table 2 Video sequences and their main properties included in MOT datasets

数据集	视频来源	长度	轨迹数量	FPS	相机状况	视点	密度	天气
MOT2015	TUD-Crossing	201	13	25	静止	水平	5.5	多云
	PETS2009-S2L2	436	42	7	静止	高	22.1	多云
	ETH-Crossing	219	26	14	移动	低	4.9	多云
	ADL-Rundle-1	500	32	30	移动	水平	18.6	晴
	KITTI-16	209	17	10	静止	水平	8.1	晴
MOT2016	MOT16-01	450	23	30	静止	水平	14.2	多云
	MOT16-03	1500	148	30	静止	高	69.7	夜晚
	MOT16-06	1194	221	14	移动	低	9.7	晴
	MOT16-12	900	86	30	移动	水平	9.2	室内
MOT2017	MOT17-01	450	24	30	静止	水平	14.3	晴
	MOT17-03	1500	148	30	静止	高	69.8	夜晚
	MOT17-06	1194	222	14	移动	水平	9.9	晴

4 实验分析

本文介绍的数种多目标跟踪方法使用的数据集多为MOT15和MOT16数据集。按目标检测、特征提取、数据关联与跟踪3个阶段分别分析不同方法的性能表现。表3、4列举了不同多目标跟踪方法在MOT15和MOT16数据集上的实验结果。

(1)目标检测:将Faster R-CNN作为检测器的多目标跟踪方法取得了较好的跟踪结果,如POI[12]中使用改良的Faster R-CNN取得了MOT16中排名较高的MOTA得分。虽然使用YOLO作为检测器的方法NSH[17]也获得了较高的MOTA得分,但是相较于Faster R-CNN依然有可以提升的空间。

(2)特征提取:STAM[2]、QuadMOT[3]、MHT_DAM[23]、CTracker[28]、SiameseCNN[33]、LMP[36]、AMIR[39]都设计了独特的特征提取方法,性能表现各有千秋。影响这些方法MOTA得分的因素可能是不同的方法在计算时采用了不同的处理方法,得到的FN、FP和IDSW有所不同,如STAM[2]的IDSW比QuadMOT[3]小,故STAM的MOTA得分略高于QuadMOT。但虽然CTracker[28]的IDSW较高却依然取得更好的MOTA得分的原因可能是其拥有更优秀的特征提取、数据关联方法。

(3)数据关联与跟踪:虽然已有研究者将强化学习应用于数据关联与跟踪[16,41],但是可以发现,虽然MARLMOT[41]的FN数值在所列方法中最低,但其MOTA得分仍有很大的提升空间。根据现有的实验结果观察,尚未证实强化学习能有效提升多跟踪算法的性能。

综合来看,虽然每个方法使用的神经网络、处理方法各不相同,但是方法之间依然存在共性:

(1)同一种方法如STAM[2]、QuadMOT[3]、MPN[30]和AMIR[39]等在MOT16上的表现比MOT15更好,原因可能是MOT16的检测使用了基于组件的可变性模型,拥有更好的检测结果,因此影响了最终的跟踪结果。

(2)由3.1节可知,影响MOTA得分的主要指标是FN、FP和IDSW,而FN在数量上比另外两个指标高出一个数量级。因此若能大幅减少

表 3 MOT15的多目标跟踪方法的通用指标评估结果

Table 3 Evaluation results of general metrics of MOT15´s multi object tracking methods

方法	MOTA	MOTP	FP	FN	IDSW
RNN_LSTM[44]	19.0	71.0	11 578	36 706	1 490
MARLMOT[41]	27.7	72.5	6 092	21 976	767
SiameseCNN[38]	29.0	71.2	5 160	37 798	639
MHT_DAM[23]	32.4	71.8	9 064	32 060	435
QuadMOT[3]	33.8	73.4	7 898	32 061	703
STAM[2]	34.3	70.5	5 154	34 848	348
RAN[4]	35.1	70.9	6 771	32 717	381
AMIR[39]	37.6	71.7	7 933	29 397	1026
AP_HWDPL[27]	38.5	72.6	4 005	33 203	586
MPN[30]	51.5	76.0	7 620	21 780	375

表 4 MOT16的多目标跟踪方法的通用指标评估结果

Table 4 Evaluation results of general metrics of MOT16´s multi object tracking methods

方法	MOTA	MOTP	FP	FN	IDSW
DAN[18]	40.8	74.4	15 143	91 792	1 051
MHT_bLSTM[5]	42.1	75.9	11 637	93 172	753
QuadMOT[3]	44.1	76.4	6 388	94 775	745
MHT_DAM[23]	45.8	76.3	6 412	91 758	590
STAM[2]	46.0	74.9	6 895	91 117	473
DMAN[35]	46.1	73.8	7 909	89 874	532
AMIR[39]	47.2	75.8	2 681	92 856	774
LMP[36]	48.8	79.0	6 654	86 245	481
MPN[30]	58.6	78.9	4 949	70 252	354
DeepSORT[24]	61.4	79.1	12 852	56 668	781
NSH[17]	63.9	78.5	9 829	55 000	913
POI[12]	66.1	79.5	5 061	55 914	805
CTracker[28]	67.6	78.4	8 934	48 305	1897

FN值,则可能得到性能更加优秀的方法。通过表3、表4可看出,POI[12]、NSH[17]、DeepSORT[24]的MOTA得分优秀的原因之一就是它们的FN值较低。另外Leal-Taixé[60]等指出,MOTA与FN值之间可以通过皮尔森相关系数(Pearson correlation coefficient)联系在一起。因此,虽然一些通用检测器也在减少FN值上有所改进,但是更加行之有效的办法仍然是设计专用的目标检测器,并尽可能地降低FN值。

(3)特征提取是一项关键步骤,提取特征的差异影响最终的跟踪结果,良好的行人特征会很大程度上提高跟踪器的性能。CTracker[28]的MOTA得分较高的原因之一就在于提取了表征能力更强的特征。另外,由于CTracker没有使用行人重识别的特征进行训练,因此其IDSW较高。同时可以发现CTracker的MOTP比POI[12]的低,原因可能是POI使用了额外的训练数据,而Ctracker仅使用了MOT16的训练数据。近几年的工作开始关注不同类型的特征融合(如外观特征和运动特征的融合),如引入运动特征[3,24],然后使用深度网络对不同类型特征进行有效融合。

5　结束语与展望

经过近几年的迅速发展,基于深度学习的多目标跟踪技术已经取得了巨大的进步,深度学习使得多目标跟踪方法的准确度大幅提升。研究者已经对多目标跟踪有了大量的研究,但仍存在许多新颖的研究方向值得探索:①数据关联中应用深度学习技术:目前只有少数工作在数据关联中应用了深度学习技术,仍需进一步探索是否能将深度学习技术应用于数据关联与跟踪,以提升方法的性能;②跨摄像头多目标跟踪:已经有一部分工作[22,61-63]开始研究跨摄像头多目标跟踪的问题,如多个摄像头记录不同的场景[22]。但是相较于单摄像头多目标跟踪,这仍然是一个很少被研究和讨论的领域;③基于3D的多目标跟踪:3D场景可以提供更加精准的跟踪结果和位置信息,可能涉及到场景的布局、3D建模和跨摄像头转换问题;④极高人群密度的多目标跟踪:极高的人群密度对于目标检测、特征提取和数据关联与跟踪都是更高的挑战;⑤场景感知的多目标跟踪:在诸如高峰期的地铁站,或火车站等场景中,场景感知可以通过对背景的分析提供一定的上下文信息或场景结构。除却本文关注的行人跟踪外,若将多目标跟踪的研究拓展至其他类型的目标(如车辆[64],细胞[65-66],动物[67]等)和场景(如无人机航拍,交通监控等),同样存在更多亟需解决的问题和更多的挑战。

参考文献:

[1] Fan L, Wang Z, Cail B, et al. A survey on multiple object tracking algorithm[C] //IEEE International Conference on Information and Automation(ICIA), Ningbo, China, 2016: 1855-1862.

[2] Chu Q, Ouyang W, Li H, et al. Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism[C] //Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 4836-4845.

[3] Son J, Baek M, Cho M, et al. Multi-object tracking with quadruplet convolutional neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017: 5620-5629.

[4] Fang K, Xiang Y, Li X, et al. Recurrent autoregressive networks for online multi-object tracking[C]// IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, 2018: 466-475.

[5] Kim C, Li F, Rehg J M. Multi-object tracking with neural gating using bilinear lstm[C]//European Conference on Computer Vision(ECCV), Munich, Germany, 2018: 208-224.

[6] Xu Y, Zhou X, Chen S, et al. Deep learning for multiple object tracking: a survey[J]. IET Computer Vision, 2019, 13(4): 355-368.

[7] Sun Z, Chen J, Liang C, et al. A survey of multiple pedestrian tracking based on tracking-by-detection framework[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020(99):1-10.

[8] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[9] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//Proceedings of the European Conference on Computer Vision(ECCV), Amsterdam, Netherlands, 2016: 21-37.

[10] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 779-788.

[11] Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking[C]//IEEE International Conference on Image Processing (ICIP), Phoenix, USA, 2016: 3464-3468.

[12] Yu F, Li W, Li Q, et al. POI: multiple object tracking with high performance detection and appearance feature[C]// Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, Netherlands, 2016: 36-42.

[13] Kieritz H, Hubner W, Arens M. Joint detection and online multi-object tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, USA, 2018: 1540-1548.

[14] Dawei Z, Hao F, Liang X, et al. Multi-object tracking with correlation filter for autonomous vehicle[J]. Sensors, 2018, 18(7): 2004-2011.

[15] Redmon J, Farhadi A. YOLOv3:an incremental improvement[EB/OL]. [2018-04-08]. https://arxiv.org/pdf/1804.02767.pdf

[16] Jiang M, Hai T, Pan Z, et al. Multi-agent deep reinforcement learning for multi-object tracker[J]. IEEE Access, 2019, 7: 32400-32407.

[17] He M, Luo H, Hui B, et al. Fast online multi-pedestrian tracking via integrating motion model and deep appearance model[J]. IEEE Access, 2019, 7: 89475-89486.

[18] Zhou Q, Zhong B, Zhang Y, et al. Deep alignment network based multi-person tracking with occlusion and motion reasoning[J]. IEEE Transactions on Multimedia, 2018, 21(5): 1183-1194.

[19] Ku J, Mozifian M, Lee J, et al. Joint 3d proposal generation and object detection from view aggregation [C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 2018: 1-8.

[20] Baser E, Balasubramanian V, Bhattacharyya P, et al. Fantrack: 3d multi-object tracking with feature association network[C] // IEEE Intelligent Vehicles Symposium(IV), Paris, France, 2019: 1426-1433.

[21] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 1302-1310.

[22] Ristani E, Tomasi C. Features for multi-target multi-camera tracking and re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6036-6046.

[23] Kim C, Li F, Ciptadi A, et al. Multiple hypothesis tracking revisited[C]//Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4696-4704.

[24] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]// IEEE International Conference on Image Processing (ICIP), Beijing, 2017: 3645-3649.

[25] Mahmoudi N, Ahadi S M, Rahmati M. Multi-target tracking using CNN-based features: CNNMTT[J]. Multimedia Tools and Applications, 2019, 78(6): 7077-7096.

[26] Sheng H, Zhang Y, Chen J, et al. Heterogeneous as-

sociation graph fusion for target association in multiple object tracking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(11): 3269-3280.

[27] Chen L, Ai H, Shang C, et al. Online multi-object tracking with convolutional neural networks[C] // IEEE International Conference on Image Processing (ICIP), Beijing, 2017: 645-649.

[28] Peng J, Wang C. Chained-tracker: chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking[EB/OL]. [2020-10-13]. https://arxiv. org/pdf/2007. 14557. pdf

[29] He Z, Li J, Liu D, et al. Tracking by animation: unsupervised learning of multi-object attentive trackers [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA,2019: 1318-1327.

[30] Brasó G, Leal-Taixé L. Learning a neural solver for multiple object tracking[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA,2020: 6247-6257.

[31] Porzi L, Hofinger M, Ruiz I, et al. Learning multi-object tracking and segmentation from automatic annotations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA,2020: 6846-6855.

[32] Kim M, Alletto S. Similarity mapping with enhanced siamese network for multi-object tracking[EB/OL]. [2020-10-13]. https://arxiv. org/pdf/1609. 09156. pdf

[33] Lee S, Kim E. Multiple object tracking via feature pyramid Siamese networks[J]. IEEE Access, 2018, 7: 8181-8194.

[34] Leal-Taixé L, Canton-Ferrer C, Schindler K. Learning by tracking: siamese CNN for robust target association[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, USA,2016: 418-425.

[35] Zhu J, Yang H, Liu N, et al. Online multi-object tracking with dual matching attention networks[C]// Proceedings of the European Conference on Computer Vision(ECCV), Munich, Germany, 2018: 366-382.

[36] Tang S, Andriluka M, Andres B, et al. Multiple people tracking by lifted multicut and person re-identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA,2017: 3539-3548.

[37] Yin J, Wang W, Meng Q, et al. A unified object motion and affinity model for online multi-object tracking [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA,2020: 6768-6777.

[38] Lu Y, Lu C, Tang C K. Online video object detection using association LSTM[C]//Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 2344-2352.

[39] Sadeghian A, Alahi A, Savarese S. Tracking the untrackable: Learning to track multiple cues with long-term dependencies[C]//Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 300-311.

[40] Chen L, Peng X, Ren M. Recurrent metric networks and batch multiple hypothesis for multi-object tracking [J]. IEEE Access, 2019, 7: 3093-3105.

[41] Rosello P, Kochenderfer M J. Multi-agent reinforcement learning for multi-object tracking[C]//Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, Stockholm, Sweden, 2018: 1397-1404.

[42] Munkres J. Algorithms for the assignment and transportation problems[J]. Journal of the Society for Industrial and Applied Mathematics, 1957, 5(1): 32-38.

[43] Xu Y, Osep A, Ban Y, et al. How to train your deep multi-object tracker[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA,2020: 6787-6796.

[44] Milan A, Rezatofighi S H, Dick A, et al. Online multi-target tracking using recurrent neural networks [C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, USA, 2017:4225-4232.

[45] Yoon K, Kim D Y, Yoon Y C, et al. Data association for multi-object tracking via deep neural networks [J]. Sensors, 2019, 19(3): 559-574.

[46] Chu P, Fan H, Tan C C, et al. Online multi-object tracking with instance-aware tracker and dynamic model refreshment[C]//IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, USA, 2019: 161-170.

[47] Kuhn H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics, 2005, 52 (1): 7-21.

[48] Wu B, Nevatia R. Tracking of multiple, partially occluded humans based on static body part detection

[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR′06), New York, USA, 2006: 951-958.

[49] Keni B, Rainer S. Evaluating multiple object tracking performance: The CLEAR MOT Metrics[J]. Eurasip Journal on Image & Video Processing, 2008 (1): 246309.

[50] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]//European Conference on Computer Vision, Amsterdam, Netherlands, 2016: 17-35.

[51] Leal-Taixé L, Milan A, Reid I, et al. Motchallenge 2015: Towards a benchmark for multi-target tracking [EB/OL]. [2015-04-08]. https://arxiv. org/pdf/1504. 01942. pdf

[52] Leal-Taixé L, Milan A, Reid I, et al. MOT16: A benchmark for multi-object tracking[EB/OL]. [2016-05-03]. https://arxiv. org/pdf/1603. 00831. pdf

[53] Felzenszwalb P F, Girshick R B, Mcallester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Software Engineering, 2010, 32(9): 1627-1645.

[54] Yang F, Choi W, Lin Y. Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers[C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Las Vegas, USA, 2016: 2129-2137.

[55] Dendorfer P, Rezatofighi H, Milan A, et al. CVPR19 tracking and detection challenge: how crowded can it get?[EB/OL]. [2019-06-10]. https://arxiv. org/pdf/1906. 04567. pdf

[56] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite [C]//IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012: 3354-3361.

[57] Ferryman J, Shahrokni A. Pets2009: dataset and challenge[C]//Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Snowbird, USA, 2009: 1-6.

[58] Andriluka M, Roth S, Schiele B. Monocular 3d pose estimation and tracking by detection[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 2010: 623-630.

[59] Wen L, Du D, Cai Z, et al. UA-DETRAC: a new benchmark and protocol for multi-object detection and tracking[J]. Computer Vision and Image Understanding, 2020, 193: 102907.

[60] Leal-Taixé L, Milan A. Tracking the trackers: an analysis of the state of the art in multiple object tracking[EB/OL]. [2020-10-13]. https://arxiv. org/pdf/1704. 02781. pdf

[61] Yoon K, Song Y, Jeon M. Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views[J]. IET Image Processing, 2018, 12(7): 1175-1184.

[62] Liang Y, Zhou Y. Multi-camera tracking exploiting person re-id technique[C]//The 24th International Conference on Neural Information Processing, Guangzhou, China, 2017: 397-404.

[63] Yoo H, Kim K, Byeon M, et al. Online scheme for multiple camera multiple target tracking based on multiple hypothesis tracking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 27(3): 454-469.

[64] Tang Z, Naphade M, Liu M Y, et al. Cityflow: a city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 8797-8806.

[65] Imani Y, Teyfouri N, Ahmadzadeh M R, et al. A new method for multiple sperm cells tracking[J]. Journal of Medical Signals & Sensors, 2014, 4(1): 35-42.

[66] Meirovitch Y, Mi L, Saribekyan H, et al. Cross-classification clustering: an efficient multi-object tracking technique for 3-D instance segmentation in connectomics[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 8425-8435.

[67] Mittek M, Psota E T, Pérez L C, et al. Health monitoring of group-housed pigs using depth-enabled multi-object tracking[C]//Proceedings of International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 2016: 9-12.