## 目标跟踪和人流检测技术调研报告

## 1.MOT多目标跟踪简介

MOT（Multi-Object Tracking），顾名思义，就是在一段视频中同时跟踪多个目标。在事先不知道目标数量的情况下，对视频中的行人、汽车、动物等多个目标进行检测并赋予ID进行轨迹跟踪。不同的目标拥有不同的ID，以便实现后续的轨迹预测、精准查找等工作。MOT主要应用场景是安防监控和自动驾驶等，这些场景中我们往往需要对众多目标同时进行追踪。这是仅用目标检测算法或单目标跟踪算法都无法做到的，人们就自然就开始了对多目标跟踪算法的。

在单目标跟踪中，我们往往会使用给定的初始框，在后续视频帧中对初始框内的物体进行位置预测。而多目标跟踪算法，大部分都是不考虑初始框的，原因就是上面的目标消失与产生问题。取而代之，在多目标跟踪领域常用的跟踪策略是TBD（Tracking-by-Detecton），又或者也可叫DBT（Detection-Based-Tracking）。即在每一帧进行目标检测，再利用目标检测的结果来进行目标跟踪，这一步我们一般称之为数据关联（Data Assoiation）。

多目标跟踪是在事先不知道目标数量的情况下，对视频中的行人、汽车、动物等多个目标进行检测并赋予ID进行轨迹跟踪。不同的目标拥有不同的ID，以便实现后续的轨迹预测、精准查找等工作。

## 2.MOT多目标跟踪任务难点

相对于单目标跟踪，多目标跟踪面临着更加复杂的问题包括频繁的遮挡、轨迹的管理、相似的外观和多目标间的相互影响。在实际的应用场景中，因为识别对象多样性存在各种复杂变化。

### 2.1 目标自身的变化

目标的颜色变化（行人的衣服颜色变化），目标的尺度变化（离摄像头的远近）和目标的形态变化（行人的站立、蹲与躺）等。

### 2.2 外界环境的变化

光线明暗的变化、目标所处环境的多样性、目标的消失与出现和目标的遮挡问题。这些复杂变化会影响跟踪对象与背景环境的区分度，从而进一步影响多目标跟踪算法的跟踪效果和结果的好坏，所以需要恰当地处理这些变化来提高多目标跟踪的准确性。

## 3.MOT与ReID的联系与区别

在MOT任务中，一般常用的特征模型有运动模型和表观模型，其中表观模型以行人重识别(ReID)类算法为主流。Re-ID任务主要解决的是跨摄像头场景下行人的识别与检索，其中存在给定了身份的图片序列query，需要为不同摄像头场景下的多组图片gallery的行人身份进行判定。

|  |  |  |
| --- | --- | --- |
| 指标 | ReID | MOT |
| 研究对象 | 行人/车辆 | 行人/车辆 |
| 场景 | 跨摄像头 | 单一摄像头 |
| 先验信息 | 表观信息 | 时空运动信息，表观信息 |
| 研究任务 | 行人匹配 | 数据关联 |
| 目标顺序是否严格对称 | 是 | 否 |

**目标序列严格对称：是指的待比对的两个视频序列中，目标的id是否一致，如果不一致，则表明两个视频序列各自可能存在新的目标。**

ReID更像是一个图像检索的问题，把检测出来的行人图片提取一个feature，根据feature来判断两张图片的相似性，实现检索的目的，而[tracking](https://www.zhihu.com/search?q=tracking&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A265070848%7D)更多的像一个数据关联的问题，利用ReID特征、时空信息、运动信息等来把两个目标给匹配关联起来，跟踪方法其中有一类方法是tracking by detect（TBD），而ReID做tracking就属于这一类，主要思想就是先把行人目标检测出来，然后判断detect的bounding boxes是否属于同一个行人，把同一个行人的检测框给关联起来就是tracklet。ReID的难点在于是否能够保证相同的人员在不同位置角度摄像机下能被识别出来。Tracking 的问题是在于估计物体的state，它的难点在于如何基于detection的信息对物体的未来或者过去进行预测，来解决detection不可靠的问题。ReID可以不基于tracking来做，通过detection后就可以了，但是因为遮挡的问题，所以基于tracking的结果做re-ID可以提升短时遮挡问题的准确性。

Tracking 也可以不基于re-ID 来做，比RFS，即使同时考虑所有association，做monte carlo的概率分析，从而在不考虑ID的前提下预测目标数量和估计目标状态。也可以使用labelled RFS 从而预测ID，并把ID反馈给系统，来提升整体系统的正确率。

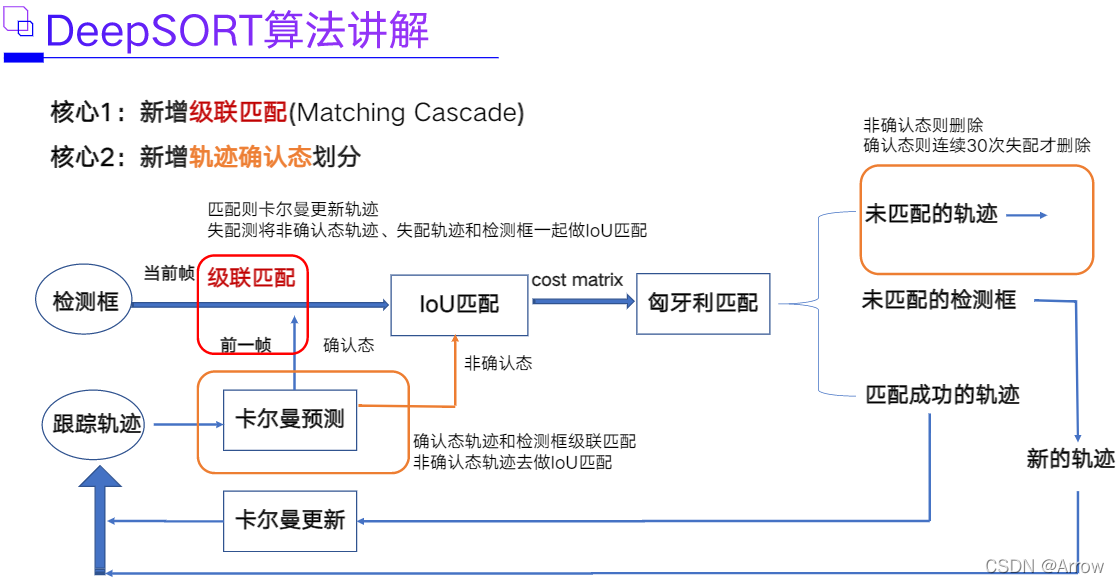


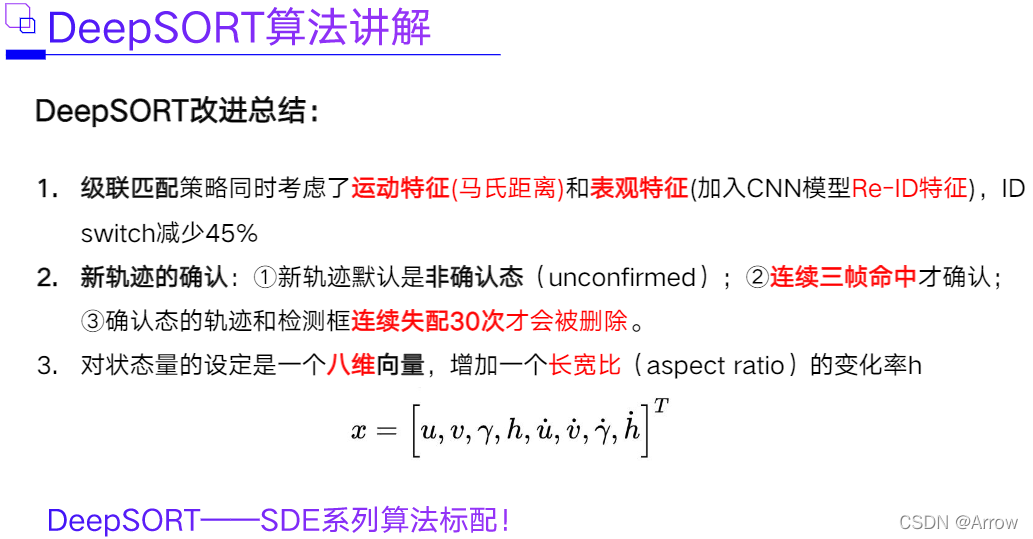
## 4.MOT主流算法及最新成果介绍

### 在这里插入图片描述4.1.基于Tracking-by-detection的MOT

基于Tracking-by-detaction框架的MOT算法是先对视频序列的每一帧进行目标检测，根据包围框对目标进行裁剪，得到图像中的所有目标。然后，转化为前后两帧之间的目标关联问题，通过IoU、外观等构建相似度矩阵，并通过匈牙利算法、贪婪算法等方法进行求解。

首先检测目标，然后链接到轨迹中。这种策略也通常被称为“tracking-by-detection”。给定一个序列，在每帧中进行特定类型的目标检测或运动检测（基于背景建模，得到目标假设， 然后进行顺序或批量跟踪，将检测假设连接到轨迹中。有两个问题值得注意：第一，由于提前训练目标检测器，TBD大部分关注特定的目标类型，如行人、车辆或人脸。第二，TBD的性能非常依赖于所采用的目标检测器的性能代表方法：**SORT、DeepSORT**

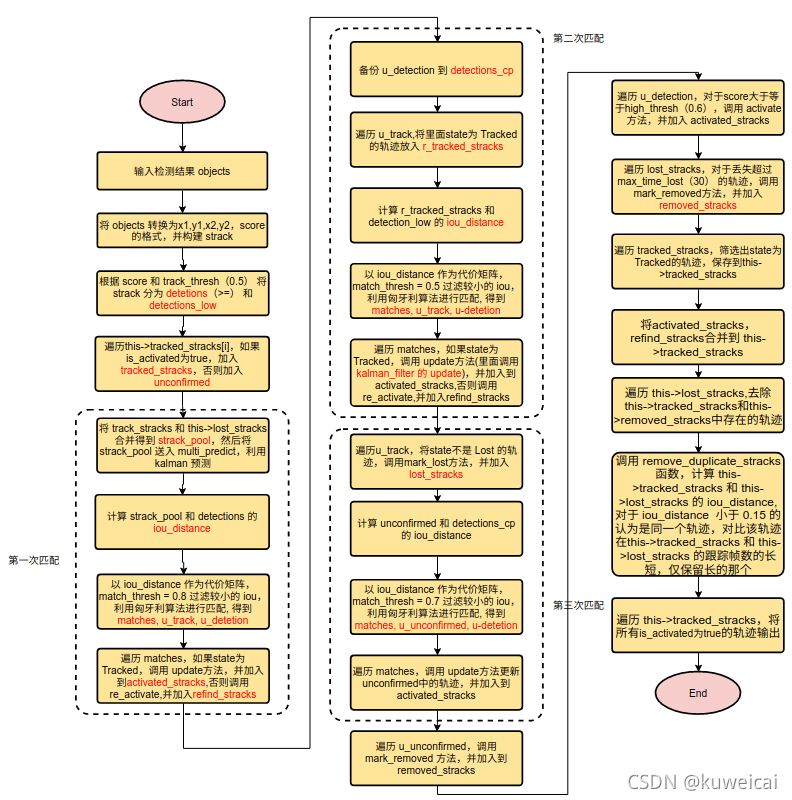




**ByteTrack**

ByteTrack 是基于 tracking-by-detection 范式的跟踪方法。作者提出了一种简单高效的数据关联方法 BYTE。它和之前跟踪算法的最大区别在于，并不是简单的去掉低分检测结果，正如论文标题所述，Assiciating Every Detection Box。利用检测框和跟踪轨迹之间的相似性，在保留高分检测结果的同时，从低分检测结果中去除背景，挖掘出真正的物体（遮挡、模糊等困难样本），从而降低漏检并提高轨迹的连贯性。速度到 30 FPS（单张 V100），各项指标均有突破。就我个人 demo 测试来看，相比 deep sort，ByteTrack 在遮挡情况下的提升非常明显。但是需要注意的是，由于ByteTrack 没有采用外表特征进行匹配，所以跟踪的效果非常依赖检测的效果，也就是说如果检测器的效果很好，跟踪也会取得不错的效果，但是如果检测的效果不好，那么会严重影响跟踪的效果。ByteTrack 的核心在于 BYTE，也就是说可以套用任何你自己的检测算法，把你的检测结果输入跟踪器即可，和 deepsort 类似，这种方式相比 JDE 和 FairMOT，在工程应用上更为简洁。

ByteTrack检测流程如下：

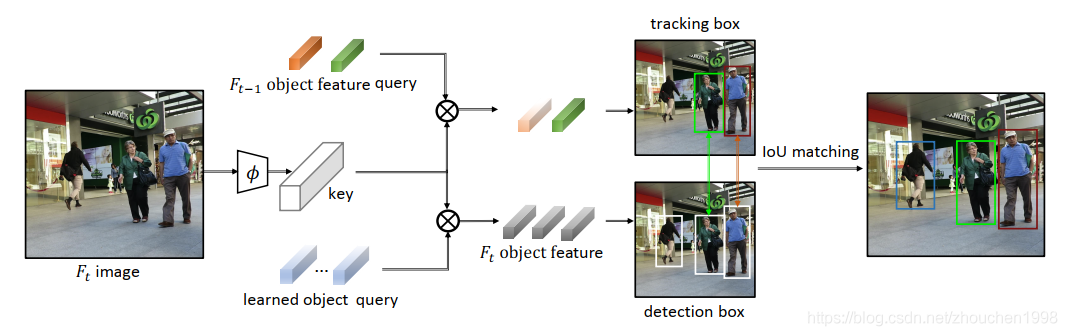


### 4.2基于检测和跟踪联合的MOT

代表方法： FairMOT、CenterTrack、ChainedTracker

FairMOT：它采用了一个非常简单的包含两个分别检测物体和提取ReID特征的均匀分支的网络框架。检测分支是以无锚的方式实现的，它估计物体的中心和大小，以位置感知图来表示。同样，re-ID分支为每个像素估计一个re-ID特征，来描述以该像素为中心的物体。注意到两个完全平等的分支和之前将检测和re-ID级联的方法完全不同。所以FairMOT消除了检测分支不公平的优势，有效地学习高质量的re-ID特征并且对于更好的MOT结果取得了检测re-ID的平衡。

### **4.3**基于注意力机制的MOT

**代表方法：基于transformer的TransTrack,TrackFormor**

pipeline的设计，它基于Query-Key目前最火热的Transformer架构构建。最中间的key来自骨干网络对当前帧图像提取的特征图，而query按照两个分支的需求分别来自上一帧的目标特征query集和一个可学习的目标query集。这两个分支都很有意思，我们先看下面这个检测分支，这里这个learned object query思路来自于DETR，是一种可学习的表示，它能逐渐学会从key中查询到目标的位置从而完成检测，想知道得更明白得可以去看看DETR论文。可以很明显地看明白，这个检测分支完成了当前帧上所有目标的检测得到detection boxes。然后我们看上面这个跟踪框分支，这个object feature query其实就是上一帧的检测分支产生的目标的特征向量，这个object feature query从key中查询目标当前帧中位置，用CenterTrack的思路来理解，这可以认为是一个位移预测分支，它最终得到tracking boxes。最后，由于跟踪框和检测框都在当前帧上，简单的IOU匹配后就能完成跟踪了。

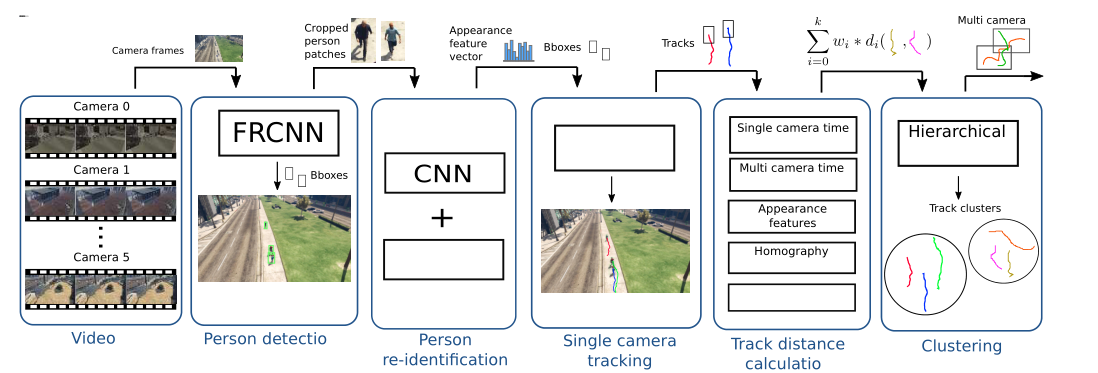
## 5.MTMCT多目标多摄像头跟踪

### 5.1 MTMCT介绍

多目标多摄像机（MTMC）跟踪，即通过可能不重叠的摄像机网络跟踪多人，是现代安全、运动分析或零售系统的一个重要元素。MTMC跟踪方法的开发是一个复杂的问题，因为它们涉及许多任务，这些任务本身就是具有挑战性的计算机视觉问题，即：**人像检测**、**单摄像机多目标跟踪**和**人员重新识别**。所有这些组件都受到挑战的影响，例如分辨率或摄像机距离的变化、视角的变化、不重叠的摄像机视图、拥挤区域的遮挡或照明变化。然而，MTMC跟踪方法开发中更根本的挑战是缺乏合适的数据集。为了允许MTMC开发和评估，数据集需要提供具有这些挑战性特征的图像，以及全面的地面实况，尤其是所有摄像机的一致ID。此类数据不仅难以注释，而且其收集还存在侵犯当前或未来数据保护权利的风险。

把对相关方法的讨论局限于主要任务，即通过跨摄像机数据关联进行多摄像机多目标跟踪能够大大简化整个流程。数据关联的特征为了执行数据关联，通常需要计算不同摄像机的检测、轨迹或轨迹之间的特征、距离或限制。

### 5.2 一种基于加权距离聚合的MTMCT

一般来说，它由5个主要组件组成。首先，人员检测模块预测人员边界框，为其计算人员外观特征。然后将边界框和相应的人物嵌入转发到单摄像机跟踪阶段，该阶段分别计算并输出每个摄像机视图的轨迹。然后，通过计算轨迹之间的一组不同特征距离，将生成的小轨迹传递到框架的核心，即轨迹比较。随后，使用分层聚类方法，基于轨迹距离的加权聚合来合并小轨迹。

由于在单摄像机跟踪组件中使用了检测跟踪方法，因此需要提前在每个视频的每一帧中检测人员。文献中已经存在许多能够非常稳健和准确地识别行人的物体检测器。这就是为什么重点不是在框架的这个阶段，而是应用最先进的检测器。**当将级联R-CNN与ResNext-101结合使用时，获得了69.5%AR和67.0%AP的最佳结果**。

人员重新识别是根据其外观查找人员的问题。这意味着将待查询人的图像用作从大量gallary图像中检索显示相同身份的其他图像的查询。在多摄像机跟踪人的情况下也会出现类似的问题。例如，单个摄像机轨迹可能会因遮挡而中断，或者人员可能会离开场景，稍后再次出现。在这种情况下，可以使用已完成轨迹中的人物外观将其与新轨迹进行比较，以便重新分配人物身份。为了使人在图像中的外观具有可比性，使用CNN将人图像嵌入到特征空间中。通过在该学习嵌入空间中的距离计算，可以确定人员的相似性。使用了ABD,AWG,Strong Baseline三种网络进行检测，ABD在MTA ReID数据集上的效果最佳，但是在域传输时的信息过于特定，ABD的效果反而导致最差的性能。

框架背后的想法是计算单摄像机轨迹，并在摄像机视图内和跨摄像机视图连接它们，以获得多摄像机轨迹。单摄像机跟踪组件将人员检测和（如有必要）人员嵌入作为输入，并提供单摄像机轨迹作为输出，这些轨迹在下一步中进行聚类，以组合相同人员的轨迹。已经存在多种通过检测进行跟踪的单摄像机跟踪器，适用于该框架，从简单的联合交叉（IoU）方法到更复杂的方法，这些方法额外利用从人物图像中提取的嵌入，例如深度排序DeepSort。DeepSORT方法的性能大大优于IoU跟踪器。其中一个主要原因是使用人员嵌入，以便不仅考虑检测，而且利用行人的外观特征。