经典目标跟踪方法

2010 年以前，目标跟踪领域大部分采用一些经典的跟踪方法，比如 Meanshift、Particle Filter 和 Kalman Filter，以及基于特征点的光流算法等。Meanshift 方法是一种基于概率密度分布的跟踪方法，使目标的搜索一直沿着概率梯度上升的方向，迭代收敛到概率密度分布的局部峰值上。首先 Meanshift 会对目标进行建模，比如利用目标的颜色分布来描述目标，然后计算目标在下一帧图像上的概率分布，从而迭代得到局部最密集的区域。Meanshift 适用于目标的色彩模型和背景差异比较大的情形，早期也用于人脸跟踪。由于 Meanshift 方法的快速计算，它的很多改进方法也一直适用至今。

粒子滤波（Particle Filter）方法是一种基于粒子分布统计的方法。以跟踪为例，首先对跟踪目标进行建模，并定义一种相似度度量确定粒子与目标的匹配程度。在目标搜索的过程中，它会按照一定的分布（比如均匀分布或高斯分布）撒一些粒子，统计这些粒子的相似度，确定目标可能的位置。在这些位置上，下一帧加入更多新的粒子，确保在更大概率上跟踪上目标。Kalman Filter 常被用于描述目标的运动模型，它不对目标的特征建模，而是对目标的运动模型进行了建模，常用于估计目标在下一帧的位置。另外，经典的跟踪方法还有基于特征点的光流跟踪，在目标上提取一些特征点，然后在下一帧计算这些特征点的光流匹配点，统计得到目标的位置。在跟踪的过程中，需要不断补充新的特征点，删除置信度不佳的特征点，以此来适应目标在运动中的形状变化。本质上可以认为光流跟踪属于用特征点的集合来表征目标模型的方法。

在深度学习和相关滤波的跟踪方法出现后，经典的跟踪方法都被舍弃，这主要是因为这些经典方法无法处理和适应复杂的跟踪变化，它们的鲁棒性和准确度都被前沿的算法所超越，但是，了解它们对理解跟踪过程是有必要的，有些方法在工程上仍然有十分重要的应用，常常被当作一种重要的辅助手段。

检测与跟踪相结合的方法

在 2010 年左右，目标跟踪的方法通常分成基于产生式模型的方法和基于鉴别式模型的方法。前面介绍的经典跟踪方法都可以归类为产生式模型的方法，而基于鉴别式模型的方法是指利用分类来做跟踪的方法，即把跟踪的目标作为前景，利用在线学习或离线训练的检测器来区分前景目标和背景，从而得到前景目标的位置。虽然此时通用物体的检测率还非常低（ImageNet 的检测率不超过 20%），因为物体检测主要基于手工设计的特征，但是，通过在新更新检测器的模型和各种底层特征的提出，鉴别式跟踪方法更能适应跟踪过程中的复杂变化，所以利用检测来做跟踪（Tracking By Detection）逐渐成为主流。

人们开始采用一种在线学习的方法来训练跟踪的检测器，比如，在第一帧目标出现的位置建立初始的检测模型，并在后续帧更新模型以适应目标的变化，这里介绍当时 Surrey 大学博士生 Z.Kalal 提出的一种长效跟踪的 TLD（Tracking-Learning-Detection）方法。

TLD 方法的整个框架有三个模块，其中跟踪器（Tracking）负责用经典的方法跟踪目标，论文中采用基于光流的特征点统计方法确定目标在下一帧的跟踪位置，检测器（Detection）负责在一定的范围内检测目标，并且与跟踪结果综合得到最终的输出结果。在确定目标的最佳位置之后，学习模块（Learning）负责对跟踪结果和检测器进行修正，在目标的周围选取更多的正负样本在线更新检测器的模型。

作者利用了多级分类器的方式来提高检测器的检测能力，降低误检，级联了一个在线学习的随机森林分类器和最近邻的分类器。为了平衡整个检测器的稳定性和学习能力，作者提出了一种半监督的学习方法，利用 P-N 限制条件更新模型，使检测器能够收敛到一个稳定的状态。从整个框架的设计上看，TLD 的工程意义要大于算法本身，作者对每个模块都做了很多细节上的调整，比如利用 Backward-Forward 来过滤不可靠的光流特征点，对物体的变化做了尺度上的估计，利用 P-N 样本集更新随机森林分类器等。

TLD 的代码也公开了，人们可以很容易重现论文的结果。这种跟踪和检测相结合的框架和思想，极大地推动了跟踪算法的发展。那时对于跟踪算法的评测还没有公认的数据集和评测标准，很难对不同跟踪方法的进行比较。所以，在 2013 年左右，人们开始建立公共的测试视频集，定义了跟踪的初始条件和评价指标，比如 OTB 数据集（Object Tracking Benchmark）和 VOT 竞赛（Visual Object Tracking），其中 OTB 数据集至今已经有 100 多段视频，以短时跟踪为主，里面包含了各种复杂的变化场景，VOT 竞赛每年举办一次，迄今已经举办了 4 年，自此跟踪的模型和算法日新月异，比如 2013 年的冠军 Struck，利用结构化 SVM 方法在线学习的一个分类器，2014 年的冠军 DSST，一种改进的基于相关滤波的多尺度跟踪方法，之后陆续又出现了很多基于深度学习的方法。

基于相关滤波的跟踪算法

相关滤波的跟踪算法始于 2012 年 P.Martins 提出的 CSK 方法，作者提出了一种基于循环矩阵的核跟踪方法，并且从数学上完美解决了密集采样（Dense Sampling）的问题，利用傅立叶变换快速实现了检测的过程。在训练分类器时，一般认为离目标位置较近的是正样本，而离目标较远的认为是负样本。回顾前面提到的 TLD 或 Struck，他们都会在每一帧中随机地挑选一些块进行训练，学习到的特征是这些随机子窗口的特征，而 CSK 作者设计了一个密集采样的框架，能够学习到一个区域内所有图像块的特征。

密集采样的过程可以通过作者提出的循环矩阵来实现，并且整个分类器的学习过程可以用快速傅立叶变换转化成频域内的计算，不受 SVM 或 Boost 等学习方法的限制。无论是用线性分类还是核分类，整个学习过程十分高效，证明过程完备，易于理解。在论文中作者给出的算法实现只需要十几行代码。利用快速傅立叶变换，CSK 方法的跟踪帧率能达到 100～400fps，奠定了相关滤波系列方法在实时性应用中的基石。

相关滤波系列的方法发展很快，比如 CSK 作者引用的 MOSSE 方法，后续他又提出了基于 HOG 特征的 KCF 方法。后续还有考虑多尺度或颜色特征（Color Name 表）的方法以及用深度学习提取的特征结合 KCF 的方法（比如 DeepSRDCF 方法）等。从它的发展过程来看，考虑的尺度越来越多，特征信息也更加丰富，当然计算时间也会相应增加，但总体上说，相关滤波系列的跟踪方法在实时性上优势明显，采用哪种改进版本的方法视具体的应用而定。相关滤波的方法也有一些缺陷，比如目标的快速移动，形状变化大导致更多背景被学习进来等都会对 CF 系列方法造成影响。虽然后续的研究也有一些针对性的改进，比如改进边界效应，改善背景更新策略或提高峰值响应图的置信度等，但普适性还需要进一步研究，特别是对不同的应用针对性地调整。

基于深度学习的跟踪方法

最近三年是深度学习技术的高速发展期，深度学习技术也被成功应用在计算机视觉的各个应用领域，跟踪也不例外。在大数据背景下，利用深度学习训练网络模型，得到的卷积特征输出表达能力更强。在目标跟踪上，初期的应用方式是把网络学习到的特征，直接应用到相关滤波或 Struck 的跟踪框架里面，从而得到更好的跟踪结果，比如前面提到的 DeepSRDCF 方法。本质上卷积输出得到的特征表达，更优于 HOG 或 CN 特征，这也是深度学习的优势之一，但同时也带来了计算量的增加。目前很多研究跟踪的框架和方法往往会同时比较两种特征，从而验证跟踪方法或框架的改进提高，一种是传统的手工特征，另一种就是深度网络学习的特征。网络不同层的卷积输出都可以作为跟踪的特征，对于如何有效地利用深度学习的特征，Martin.D 也做了大量的工作，提出了一系列相关的方法，比如 C-COT 和 ECO 等。

GOTURN 方法利用 ALOV300+视频序列集和 ImageNet 检测数据集训练了一个基于图像对输入的卷积网络，输出在搜索区域内相对于上一帧位置的变化，从而得到目标在当前帧上的位置。为了得到网络训练所需的大数据集，作者不仅利用了视频序列集中的随机连续帧对图，而且利用了更多的单张图片集进行数据增强。对于数据的扰动分布，作者也做了大量的实验，使得跟踪的位置变化更加平滑。

从网络结构中可以看出，GOTURN 先从输入的图像对中分别提取前一帧和当前帧的卷积特征，再送入三个全连接层，最后得到预测的位置变化。其中卷积层是通过 ImageNet 预训练得到，并且在跟踪网络训练中固定这部分参数（防止过拟合），学习全连接层的参数。本质上说，GOTURN 也利用了深度学习强大的特征表达能力，但是，它建立了第一个基于深度学习的端到端输出的跟踪框架，而且在 GPU GTX680 上跟踪速度在能达到 100fps。

基于深度学习的跟踪框架目前还在不断发展中，比如牛津大学的 Luca Bertinetto 提出的端到端的跟踪框架，从 SiameseFC 到今年的 CFNet。虽然相比于相关滤波等传统方法，在性能上还非常慢，但是这种端到端输出可以与其他的任务一起训练，特别是和检测分类网络相结合，在实际应用中有着十分广泛的前景。

特定目标的跟踪

前面介绍的跟踪方法都是对通用目标的跟踪，没有目标的类别先验。在实际应用中，还有一个重要的跟踪是特定物体的跟踪，比如人脸跟踪、手势跟踪和人体跟踪等。特定物体的跟踪与前面介绍的方法不同，它更多地依赖对物体训练特定的检测器。人脸跟踪由于它的明显特征，它的跟踪就主要由检测来实现，比如早期的 Viola-Jones 检测框架和当前利用深度学习的人脸检测或人脸特征点检测模型。手势跟踪在应用主要集中在跟踪特定的手型，比如跟踪手掌或者拳头。设定特定的手型可以方便地训练手掌或拳头的检测器。

实际上，跟踪和检测是分不开的，比如传统 TLD 框架使用的在线学习检测器，或 KCF 密集采样训练的检测器，以及当前基于深度学习的卷积特征跟踪框架。一方面，跟踪能够保证速度上的需要，而检测能够有效地修正跟踪的累计误差。不同的应用场合对跟踪的要求也不一样，比如特定目标跟踪中的人脸跟踪，在跟踪成功率、准确度和鲁棒性方面都有具体的要求。

另外，跟踪的另一个分支是多目标跟踪（Multiple Object Tracking）。多目标跟踪并不是简单的多个单目标跟踪，因为它不仅涉及到各个目标的持续跟踪，还涉及到不同目标之间的身份识别、自遮挡和互遮挡的处理，以及跟踪和检测结果的数据关联等。从 2015 年开始，MOT 竞赛每年会举办一次，在基于深度学习的目标检测成功率大幅提高的背景下，MOT 的成功率也处于快速发展的通道

在介绍MOT问题solutions的一般思路之前，先了解一些术语。

**目标**。 https://images2015.cnblogs.com/blog/1027162/201610/1027162-20161030161319546-1278421302.png 在一幅图像中，明显区别于周围环境的闭合区域往往被称为目标，当然这些目标最好具有一定的物理意义。  
**检测**。 https://images2015.cnblogs.com/blog/1027162/201610/1027162-20161030161320109-1794660034.png 给出目标在图像中的位置的过程称为检测。检测方法很多，比如模板匹配，对于视频中运动目标检测，可以通过帧间差分法、背景减除法、光流法等等实现。  
**跟踪**。 https://images2015.cnblogs.com/blog/1027162/201610/1027162-20161030161320437-113379383.png 在连续图像序列中完成对目标的检测，并把物理意义下的同一目标相关联的过程。  
**检测响应**。 https://images2015.cnblogs.com/blog/1027162/201610/1027162-20161030161321109-981906159.png也称为“检测假设”、“检测观测量”。（detection response,detection hypotheses,detection observations）。是**检测**过程的输出量。  
**轨迹(Trajectory)**。https://images2015.cnblogs.com/blog/1027162/201610/1027162-20161030161321687-1440329968.pngMOT系统的输出量，一条轨迹对应这一个目标在一个时间段内中的位置序列。  
**小段轨迹(Tracklet)**。 形成Trajectory过程中的轨迹片段。完整的Trajectory是由属于同一物理目标的Tracklets构成的。  
**数据关联**。https://images2015.cnblogs.com/blog/1027162/201610/1027162-20161030161322421-759632157.png 数据关联是多目标跟踪任务中经常使用的典型的处理方法，用于解决目标间的匹配问题，这里的目标可以是detection responses，也可以是tracklets。  
**注：** 所谓的“物理目标”，就是具有相同物理意义的目标，比如两幅图像都出现了“张三”、“李四”，那么两个“张三”就是同一物理目标，虽然有可能两幅图像中“张三”的形状、表观都发生了很大的变化