

计算机视觉基础

中国石油大学（华东）

青岛软件学院、计算机科学与技术学院

宫文娟

计算机视觉基础

- 第一章 概论
- 第二章 基础知识
- 第三章 图像分类
- 第四章 图像语义分割
- 第五章 目标检测
- 第六章 识别
- 第七章 目标跟踪
- 第八章 多目视觉
- 第九章 视觉问答

第七章 目标跟踪

- 7.1 目标跟踪的分类
- 7.2 目标跟踪的发展
- 7.3 目标跟踪数据集
- 7.4 目标跟踪评价指标
- 7.5 基于生成式模型的目标跟踪
- 7.6 基于判别式模型的目标跟踪
- 7.7 基于深度学习的目标跟踪

概述

目标跟踪是利用视频或图像序列的上下文信息对目标的表观和运动信息进行建模，从而预测目标运动状态并确定其位置。

- 简单来说，目标跟踪就是在连续的视频序列中，获取目标位置关系从而建立完整的运动轨迹，一般是给定目标在第一帧图像中的位置边框，计算其在后续若干帧图像中的位置。
- ▣ 在运动过程中，目标可能会呈现一些视觉上的变化，比如姿态、形状、尺度、遮挡、亮度等变化，目标跟踪算法围绕着解决这些变化展开。
- 随着计算机处理能力的飞速提升，目标跟踪问题得到了越来越多的关注，各种基于目标跟踪的民用和军用系统纷纷落地，广泛应用于：
 1. 智能视频监控
 2. 智能人机交互
 3. 智能交通
 4. 视觉导航
 5. 无人驾驶等领域。

7.1 目标跟踪的分类

根据跟踪目标数目的不同，目标跟踪可分为：

1. 单目标跟踪

- 单目标跟踪按照目标建模方式的不同可以分为：

- ① 生成类方法。生成类方法首先对目标建立表观模型，然后在视频中搜索与目标对象相似度最高的区域作为目标区域，算法主要对目标自身特征进行细致描述，一般忽略了背景信息的影响，在目标特征变化或者出现遮挡等情况下容易导致跟踪失败。
- ② 判别类方法通过建立分类器模型对目标对象和背景信息进行区分，从而实现目标在视频中的搜索跟踪，由于判别类方法实现了背景信息建模，所以在目标跟踪时表现更为鲁棒，目前已经成为目标跟踪的主流方法。

2. 多目标跟踪

- 多目标跟踪一般是针对特定类别的多个对象的跟踪，涉及各个对象的出现、遮挡、分离等情况。

7.1 目标跟踪的分类

多目标跟踪与单目标跟踪看起来只是对象数量上的差异，但它们采用的方法差别很大：

- 单目标跟踪关注如何对目标进行重定位。
- 多目标跟踪往往更加关注如何对已检测到的目标进行精准匹配。多目标跟踪一般需要建立表观模型和运动模型，表观模型主要是对目标的整体外观特征建模，从而尽可能将目标与背景分离，运动模型主要是对目标的运动特性建模，通过对视频上下文信息建模来获取目标的运动轨迹。

7.2 目标跟踪的发展

1. 早期的目标跟踪算法主要分为两类：

1.1 基于特征建模的目标跟踪，通过对目标的表观特征建模，然后在视频中搜索目标，常用方法包括：

- 1) 特征匹配
- 2) 特征点跟踪
- 3) 光流法
- 4) 主动轮廓跟踪等

1.2 基于搜索的目标跟踪，将预测算法加入跟踪中，在预测区域附近进行目标搜索，减少了搜索范围，常用预测算法有：

- 1) 卡尔曼滤波
- 2) 粒子滤波
- 3) 核方法
 - a) Meanshift
 - b) Camshift 等。

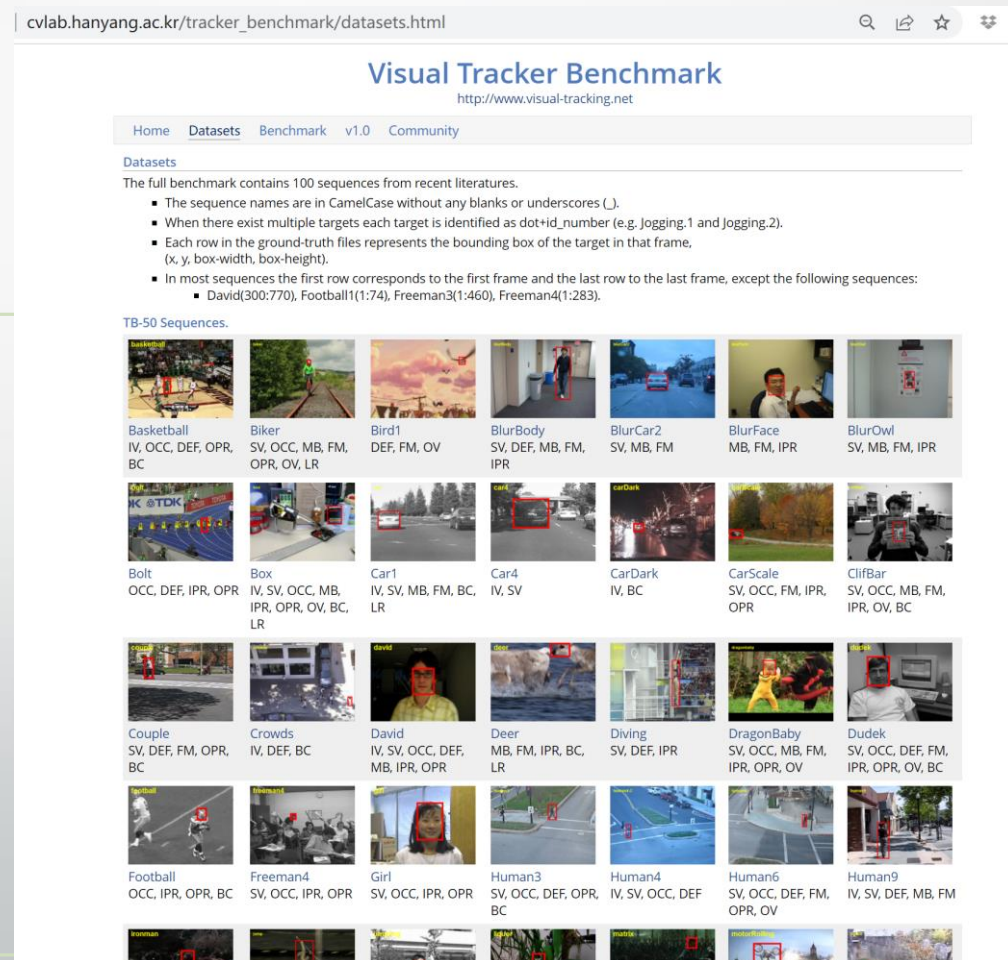
2. 基于深度学习的方法

7.3 目标跟踪数据集

各类数据集的建立直接推动了目标跟踪算法的发展，当前使用最多的几个数据集：

1. OTB50

OTB50是一套较为全面的目标跟踪数据库，它由50个完全标注的视频序列组成，包含51个不同尺寸的目标，总计超过29000帧图像。OTB50标注了11种常见的视频属性：光照变化、尺度变化、遮挡、形变、运动模糊、快速运动、平面内旋转、平面外旋转、移出视野、背景干扰和低分辨率，每一帧图像至少含有2种标注属性。此外，OTB50在统一输入输出格式的基础上整合了29个流行跟踪算法，便于大规模算法性能评估。



7.3 目标跟踪数据集

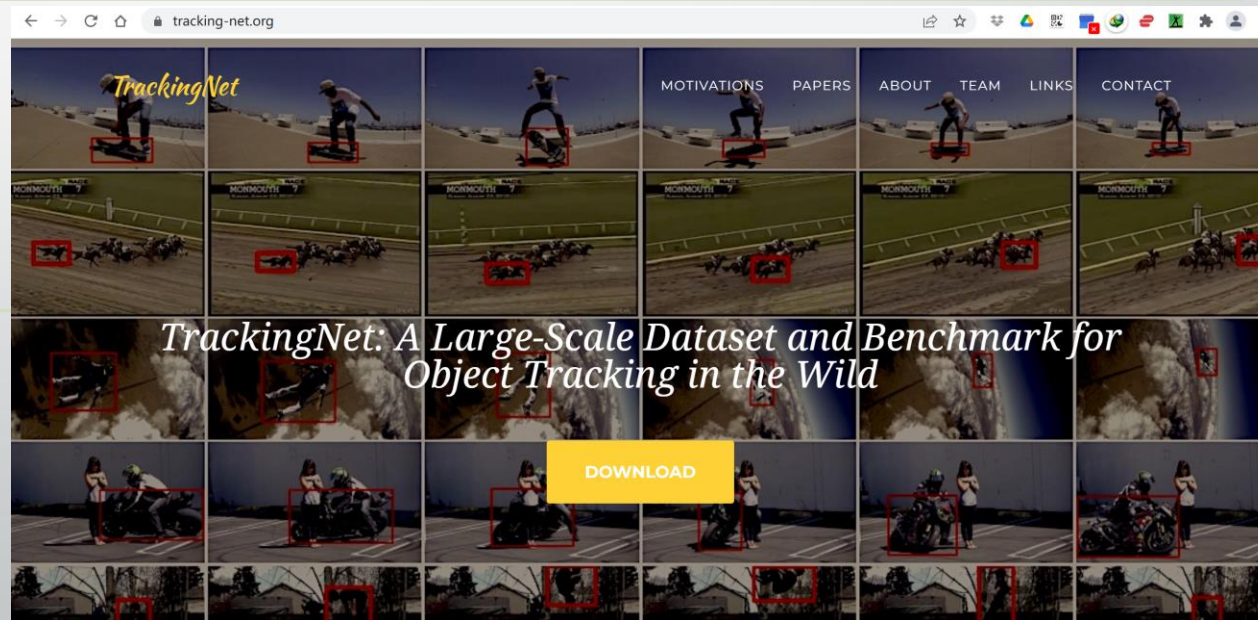
各类数据集的建立直接推动了目标跟踪算法的发展，当前使用最多的几个数据集：

1. OTB50

OTB50是一套较为全面的目标跟踪数据库，它由50个完全标注的视频序列组成，包含51个不同尺寸的目标，总计超过29000帧图像。OTB50标注了11种常见的视频属性：光照变化、尺度变化、遮挡、形变、运动模糊、快速运动、平面内旋转、平面外旋转、移出视野、背景干扰和低分辨率，每一帧图像至少含有2种标注属性。此外，OTB50在统一输入输出格式的基础上整合了29个流行跟踪算法，便于大规模算法性能评估。

2. TrackingNet

TrackingNet由YouTube上的30000多个视频构成，包含了更加广泛的跟踪目标，标注了共计1420万个目标框，该数据库的设计更偏重于挖掘目标的时序信息，使得其更适合开发接近真实世界中的目标跟踪任务。TrackingNet包含各种序列长度的视频，可用来评价短程和长程目标跟踪算法，其中Long-term Tracking in the Wild是专门针对长程目标跟踪算法构建的数据库，共包含14小时的366个视频序列，其中每个视频的平均时长超过2分钟，并带有频繁的目标消失，增加了跟踪难度。



7.3 目标跟踪数据集

各类数据集的建立直接推动了目标跟踪算法的发展，当前使用最多的几个数据集：

1. OTB50

OTB50是一套较为全面的目标跟踪数据库，它由50个完全标注的视频序列组成，包含51个不同尺寸的目标，总计超过29000帧图像。OTB50标注了11种常见的视频属性：光照变化、尺度变化、遮挡、形变、运动模糊、快速运动、平面内旋转、平面外旋转、移出视野、背景干扰和低分辨率，每一帧图像至少含有2种标注属性。此外，OTB50在统一输入输出格式的基础上整合了29个流行跟踪算法，便于大规模算法性能评估。

2. TrackingNet

TrackingNet由YouTube上的30000多个视频构成，包含了更加广泛的跟踪目标，标注了共计1420万个目标框，该数据库的设计更偏重于挖掘目标的时序信息，使得其更适合开发接近真实世界中的目标跟踪任务。TrackingNet包含各种序列长度的视频，可用来评价短程和长程目标跟踪算法，其中Long-term Tracking in the Wild是专门针对长程目标跟踪算法构建的数据库，共包含14小时的366个视频序列，其中每个视频的平均时长超过2分钟，并带有频繁的目标消失，增加了跟踪难度。

3. 其他数据库

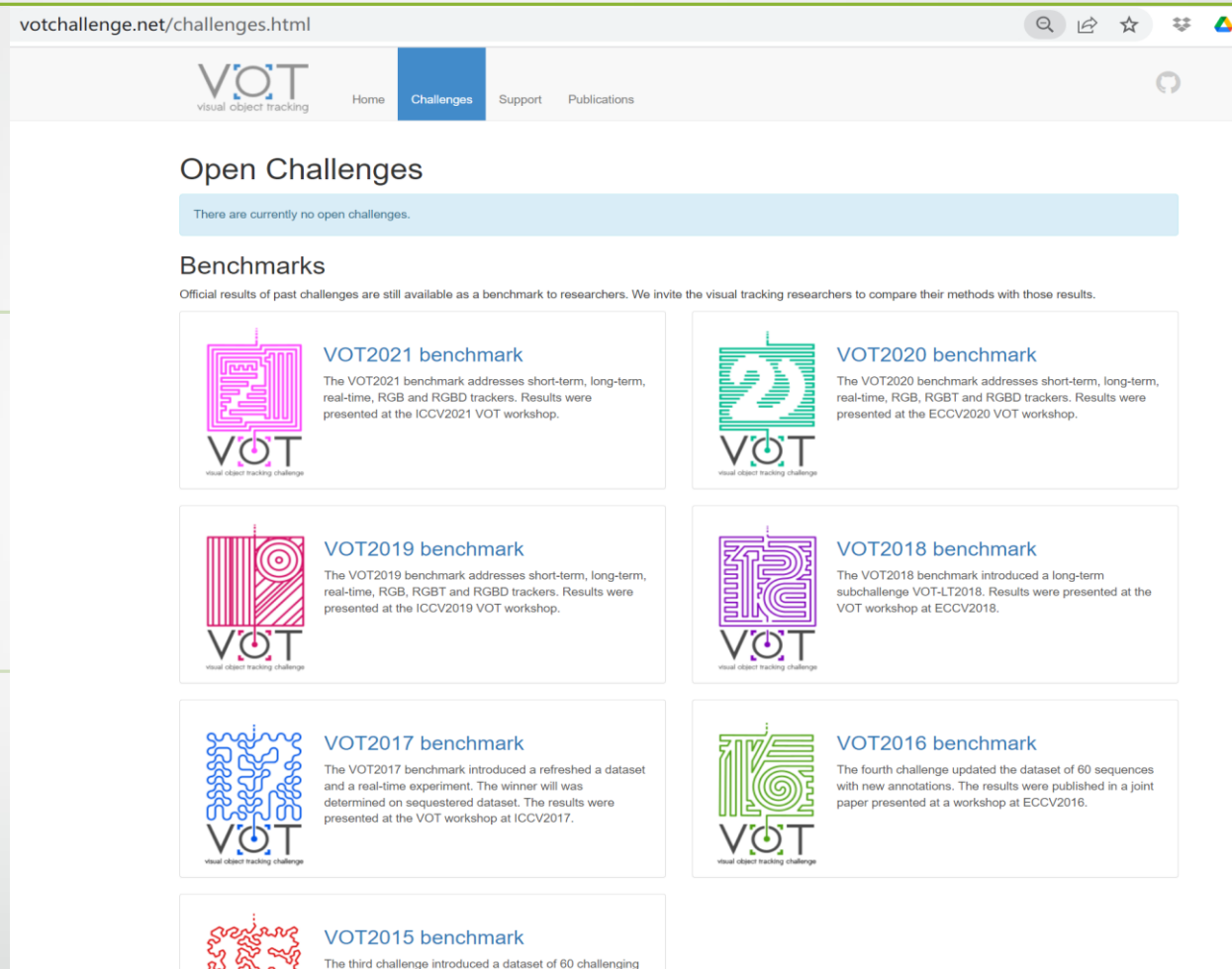
TempleColor128包含128个彩色视频序列；ALOV++数据集旨在尽可能地囊括现实世界中存在的各种干扰因素，如亮度变化、相似物干扰、遮挡等各种情况；UAV123数据集采用无人机以低空的航拍角度拍摄了123个高清序列；LaSOT数据集不仅标注了目标边框，而且增加了丰富的自然语言描述。

4. 视觉目标跟踪竞赛VOT

VOT2013仅包含16个视频序列，VOT2014将视频增至25个，并采用多边形区域方式重新标注了样本，较之OTB数据集的对齐标注更准确，VOT2015、VOT2016和VOT2017等进一步扩充到60个视频，并增加了热成像跟踪子系列，VOT2018进一步增加了长程目标跟踪任务。

7.3 目标跟踪数据集

各类数据集的建立直接推动了目标跟踪算法的发展，当前使用最多的几个数据集：



4. 视觉目标跟踪竞赛VOT

VOT2013仅包含16个视频序列，VOT2014将视频增至25个，并采用多边形区域方式重新标注了样本，较之OTB数据集的对齐标注更准确，VOT2015、VOT2016和VOT2017等进一步扩充到60个视频，并增加了热成像跟踪子系列，VOT2018进一步增加了长程目标跟踪任务。

7.4 目标跟踪评价指标

目标跟踪的评价标准通常包含两个基本参数：

1. 中心位置误差

中心位置误差是跟踪目标的中心位置和人工标注的准确位置之间的平均像素距离，通常采用一个序列中所有帧的平均中心位置误差来评价跟踪算法对该序列的总体性能。

然而，当跟踪器丢失目标时，预测的跟踪位置是随机的，此时平均误差值无法准确评估跟踪器的性能。因此将其进一步扩展为精确度曲线图，统计在不同阈值距离下的成功跟踪比例，并采用阈值为20个像素点所对应的数值作为代表性的精确度评价指标。

2. 区域重叠面积比率

区域重叠面积比率计算跟踪算法得到的边界框区域和人工标注边界框区域的交集与并集之比，它解决了中心位置误差无法评价目标在跟踪过程中的尺度变化问题。

7.5 基于生成式模型的目标跟踪

生成式模型通过提取目标的特征来构建表观模型，并据此在图像中搜索与模板最匹配的区域作为跟踪结果：

7.5.1 基于光流的目标跟踪

7.5.2 基于稀疏表示的目标跟踪

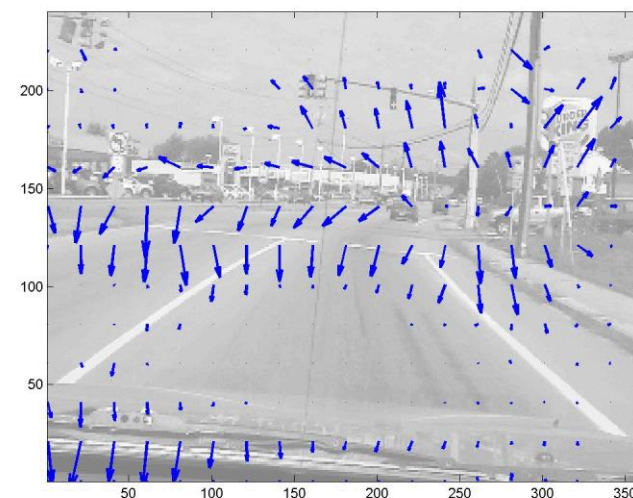
7.5 基于生成式模型的目标跟踪

7.5.1 基于光流的目标跟踪

- 光流法通过视频序列在相邻帧之间的像素关系，寻找像素的位移变化来判断目标的运动状态，从而实现对运动目标的跟踪。
- 光流法的适用范围需要满足三个假设：
 - ✓光照不变性，即需要图像的光照强度保持几乎不变；
 - ✓空间一致性，即每个像素在不同帧中相邻点的位置不变，这样便于求得最终的运动矢量；
 - ✓时间连续性，即时间的变化不会引起目标位置的剧烈变化，也即相邻两帧之间的目标位移不能太大。
- 基于光流的目标跟踪算法由于需要对视频中所有的像素进行计算，使得算法的实时性很差，设计可以缩小计算范围的算法, 对于目标跟踪问题具有重大的意义。



(a) 从行驶中的车辆上拍摄的视频

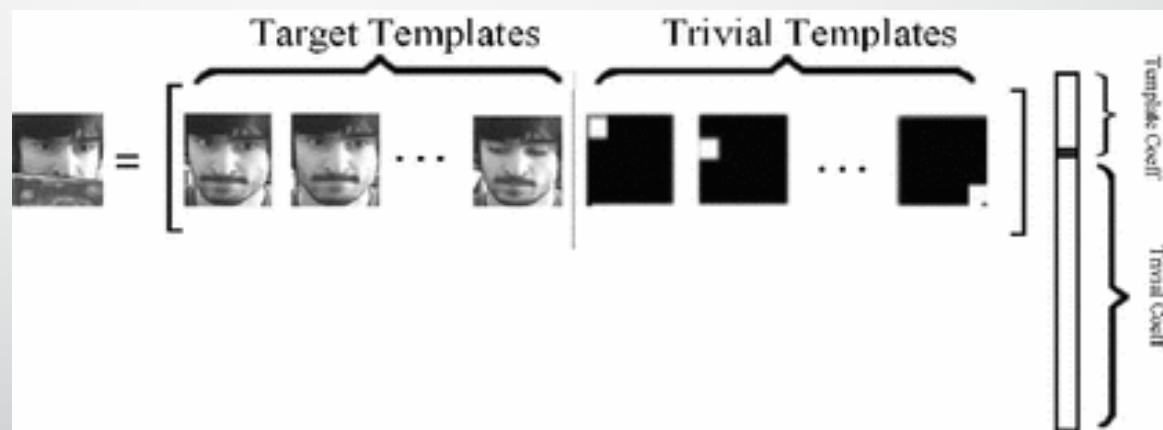


(b) 可视化从视频计算得到的光流

7.5 基于生成式模型的目标跟踪

7.5.2 基于稀疏表示的目标跟踪

- 稀疏表示在很多计算机视觉应用领域都得到广泛研究，包括图像去噪、图像去模糊、图像超分辨等图像恢复问题，还可以将稀疏表示用于解决目标跟踪中的特征表示问题。
- 实例：下图给出了最早的基于稀疏表示的目标跟踪方法^[1]的流程。



基于稀疏表示的目标跟踪方法[1]

7.5 基于生成式模型的目标跟踪

7.5.2 基于稀疏表示的目标跟踪

- 假设被跟踪的目标能由目标模板和遮挡模板组成的字典集合稀疏表示，
➤ 图7-1分别显示了好的和差的候选区域的稀疏表示系数。
 - 左侧显示当前视频帧的图像，紫红色曲线框内显示当前存储的目标模板，红色和蓝色框分别表示一个好的和一个差的目标候选，右侧显示好的和差的候选图像区域的稀疏表示系数。
 - 可以看出，好的候选表示系数明显稀疏于差的候选表示系数，这说明稀疏约束有利于定位更好的目标候选，从而获取精确的跟踪结果。
 - 如何合理地选择目标模板和遮挡模板建模跟踪目标，以及如何设计快速有效的跟踪算法是这类算法的关键。
- 除了利用稀疏表示对目标的表观建模外，稀疏表示还用于特征提取、特征选择、以及运动状态建模等方面。

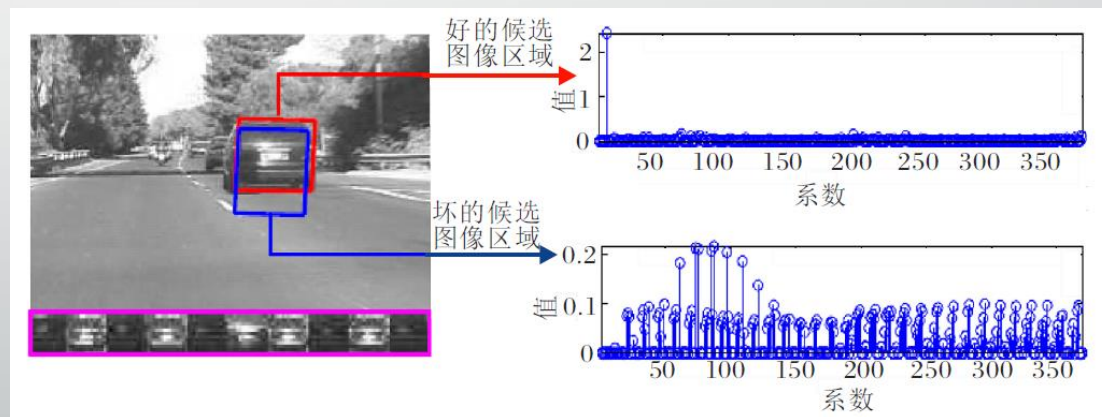


图7-1 候选目标区域在目标模板和遮挡模板组成的字典集合上的稀疏表示系数

7.6 基于判别式模型的目标跟踪

不同于生成式模型，判别式模型同时考虑目标和背景信息，它将目标跟踪建模成分类或者回归问题，目的是寻找一个判别函数将目标从背景中分离出来，从而实现对目标的跟踪。

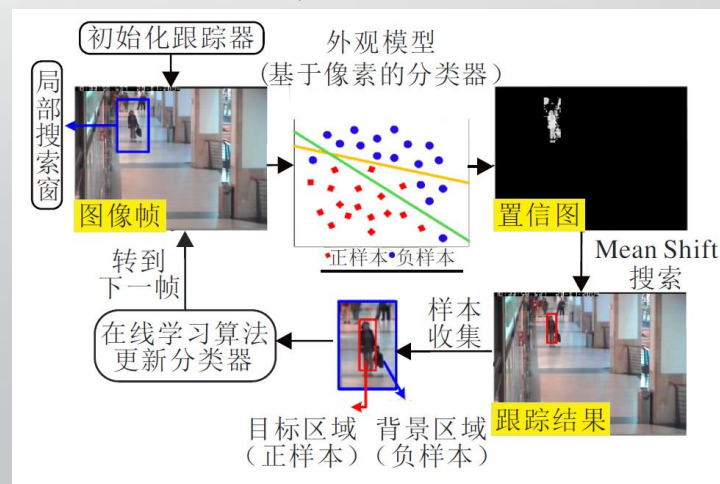
7.6.1 基于分类器的目标跟踪

7.6.2 基于相关滤波的目标跟踪

7.6.1 基于分类器的目标跟踪

7.6 基于判别式模型的目标跟踪

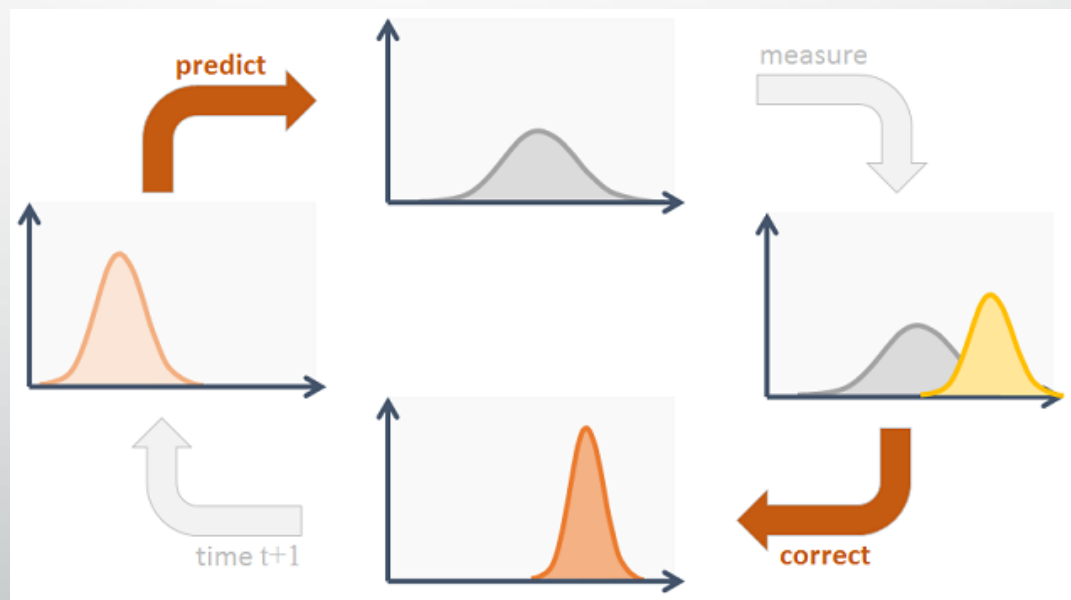
- 基于判别学习的目标跟踪算法将目标跟踪建模成一个目标与背景的二分类问题，目标跟踪的准确性和稳定性很大程度上依赖于特征空间上目标与背景的可分性。
 - 在跟踪过程中，被跟踪的目标和周围背景的表现一般都会发生变化，导致判别性的特征集合也会发生变化。
 - ◆ 在线建立能适应目标和背景表现变化的判别模型是这类算法的关键。
 - 基于判别学习的目标跟踪算法包含：
 1. 基于像素或超像素分类的跟踪。基于像素或超像素分类的跟踪算法通过分类当前图像帧中搜索区域内的像素或超像素，形成一个关于跟踪目标的置信图，然后利用Mean Shift等模态搜索算法精确定位目标。各类分类器被引入到这类方法中，先后提出了基于支持向量机、AdaBoost、在线随机森林的目标跟踪算法。
 2. 基于检测的目标跟踪。
- 基于分类器的跟踪算法流程图如右图所示。



7.7 基于深度学习的目标跟踪

7.6.2 基于相关滤波的目标跟踪

- 主要思想是使用目标图像训练得到的滤波器对图像进行滤波处理，在响应图像中寻找最大值位置，此位置即是图像中对应的目标位置。
- 在这种情况下可把目标跟踪的过程近似地看成对搜索区域图像进行相关滤波的过程，寻找目标位置也就是寻找滤波器响应图像的最大值位置。

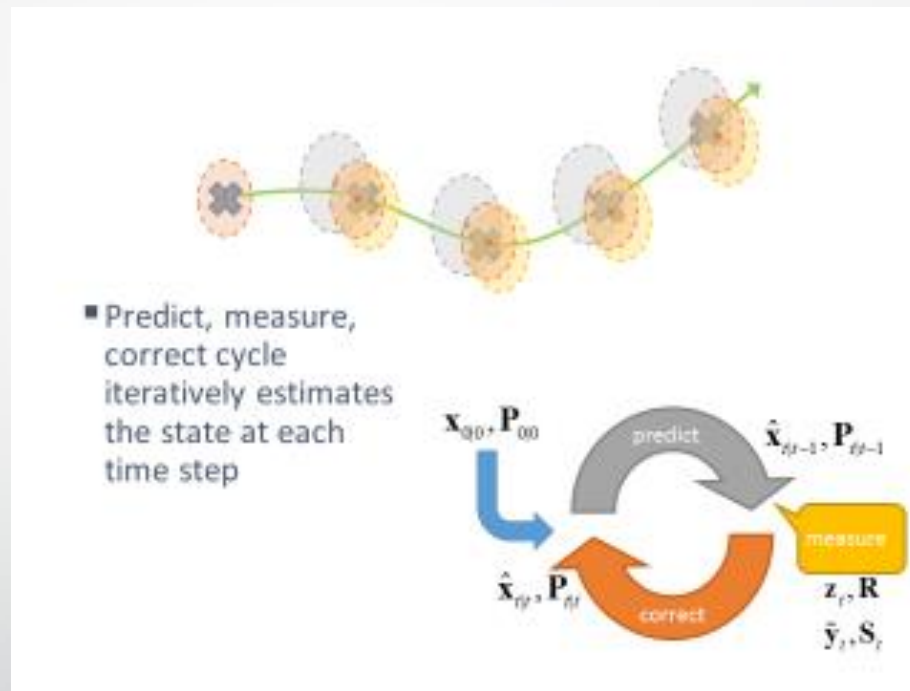


卡尔曼滤波循环过程示意图。其整个过程就是在传播更新高斯分布。首先，根据目前的状态和运动模型估计下一个状态（灰色），然后，加入从图像得到的带噪音的测量值（浅黄色），两个结合得到最终的状态（橙色）。

7.7 基于深度学习的目标跟踪

7.6.2 基于相关滤波的目标跟踪

➤ 过程示意图：



利用当前值和估计值，我们将估计下一个时刻的状态值（灰色）。接下来，我们利用一个带噪音的测量值（浅黄色）去更新估计的状态值，得到修改后的状态值。

7.7 基于深度学习的目标跟踪

基于深度学习的目标跟踪方法包含：

7.7.1 基于深度网络的目标跟踪

7.7.2 基于孪生网络的目标跟踪

7.7.3 基于强化学习的目标跟踪

7.7.4 基于深度学习的多目标跟踪

7.7.1 基于深度网络的目标跟踪

1. 基于卷积神经网络的目标跟踪

- 目前基于卷积神经网络的目标跟踪方法分为两种：

1. “离线训练+在线微调”。“离线训练+在线微调”是从大量离线训练数据中学习通用特征表示，耗时且特征针对性不强。
2. 构建简化版的卷积神经网络，力求摆脱离线训练，达到完全在线运行的要求。

- 实例：

- Wang^[2]设计了基于全卷积网络的目标跟踪算法，由于浅层特征包含较多位置信息，深层特征包含较多语义信息，该算法分别针对Conv5-3与Conv4-3特征构造捕捉类别信息的GNet以及区分背景的SNet，然后将选择的特征输入到定位网络中得到热度图，将两个定位网络的热力图进行综合得到最终的跟踪结果。

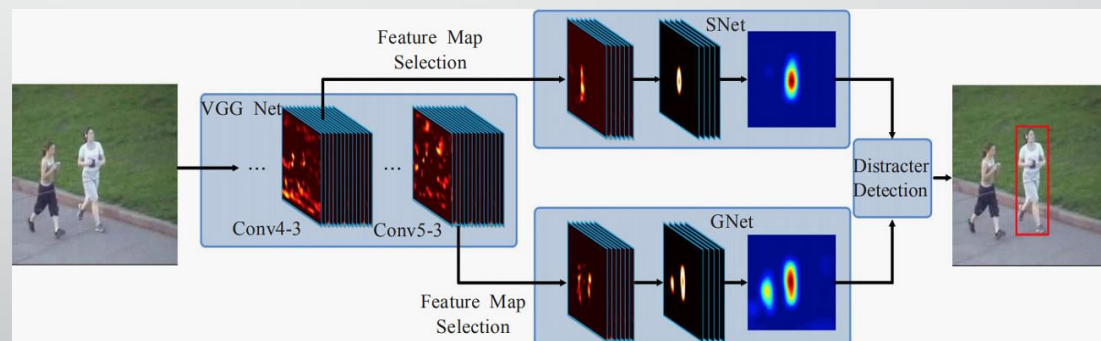


图7-3 基于全卷积网络的目标跟踪算法

7.7 基于深度学习的目标跟踪

7.7 基于深度学习的目标跟踪

7.7.1 基于深度网络的目标跟踪

1. 基于卷积神经网络的目标跟踪

□ 尽管基于卷积神经网络的目标跟踪方法已经取得了很大进展，但在时间连续性和空间信息建模方面还有待进一步改善，主要原因有：

1. 每次只能对当前帧的跟踪目标进行建模，没有考虑当前帧和历史帧之间的关联性；
2. 提取出来的深度特征往往随着网络层数的加深变得高度抽象，丢失了目标自身的结构信息；
3. 池化操作会降低特征图的分辨率，损失了目标的空间位置和局部结构信息；
4. 只关注目标本身的局部空间区域，忽视了对目标周边区域的上下文信息进行建模。

7.7 基于深度学习的目标跟踪

7.7.1 基于深度网络的目标跟踪

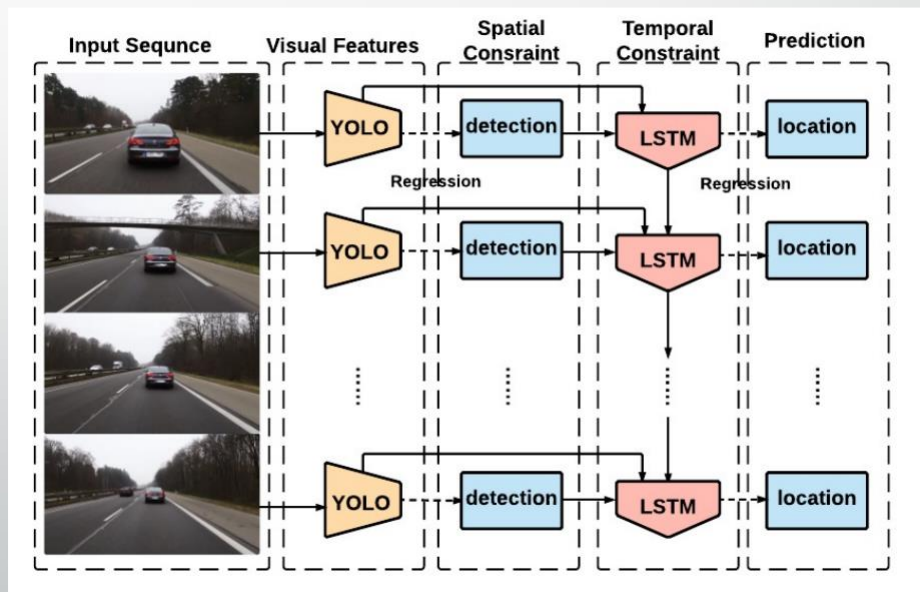
1. 基于卷积神经网络的目标跟踪

2. 基于递归神经网络的目标跟踪

□ 目前基于卷积神经网络的目标跟踪方法都是把目标跟踪建模成分类任务，导致这些跟踪方法很容易受相似物体的干扰

■ 引入递归神经网络提取物体的自身结构信息，结合卷积神经网络增强模型对相似物体的抗干扰能力。

➤ 实例：ROLO^[3]
方法流程如右图所示。



基于递归神经网络的方法[3]流程图。递归神经网络主要负责从YOLO输出的特征中提取时序和视觉特征。

7.7 基于深度学习的目标跟踪

7.7.1 基于深度神经网络的目标跟踪

1. 基于卷积神经网络的目标跟踪

2. 基于递归神经网络的目标跟踪

3. 基于生成式对抗网络的目标跟踪

□ 基于深度分类网络的目标跟踪方法存在以下两个方面问题：

1. 每一帧中的正样本空间上高度重合，不能获取丰富的表观信息；
2. 正负样本的比例严重不平衡。

■ 出现了基于生成式对抗网络的目标跟踪方法

➤ 实例：VITAL^[4]流程图如下图所示。

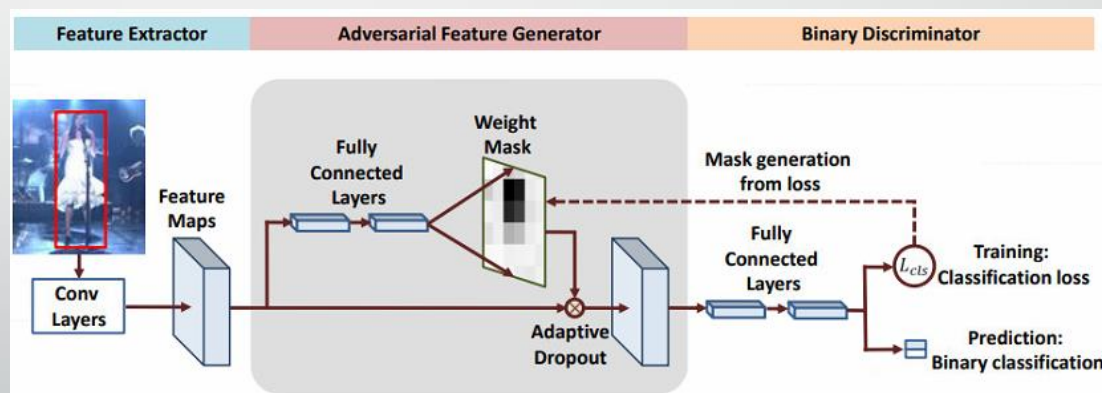


图7-4 VITAL生成对抗目标跟踪网络结构图

7.7 基于深度学习的目标跟踪

7.7.2 基于孪生网络的目标跟踪

- 对于采用深度学习的目标跟踪方法而言，由于计算参数量过于庞大，虽然在精度上有着无可比拟的优势，但是在实际应用以及相关实践上却难以发挥功效。
- 对于深度学习的方法来说，所耗费的大部分计算时间都集中在在线更新时所需要计算的反向传播的过程上，而离线训练避免微调只计算前向传播的过程所花费的时间在跟踪领域是完全可以接受的。
- 基于相关性思想的孪生全卷积网络也成为了近几年研究的热点方向。针对相关滤波处理模型快速变化能力差，同时其采用循环矩阵所造成的边界效应难以解决，基于深度网络思想，人们提出了交叉相关的思想，使用卷积操作来代替滑动窗口检测。

7.7 基于深度学习的目标跟踪

7.7.2 基于孪生网络的目标跟踪

- 基于孪生全卷积网络SiamFC的网络结构如图7-5所示：
 1. 将一对图像通过具有同样结构的两个CNN 网络得到特征的降维映射，
 2. 然后通过卷积实现相关性的计算, 得到目标位置的响应。
 3. 在最后的响应图上插值回原图大小实现目标的定位。
 4. SiamFC采用了跟踪问题中常采用的尺度自适应方法，在3个尺度上达到86fps，在5个尺度上也达到了58fps的效果。
- 由于采用了相关性计算的方法和离线训练的方式，避免在线微调，使得网络速度达到了令人满意的效果。

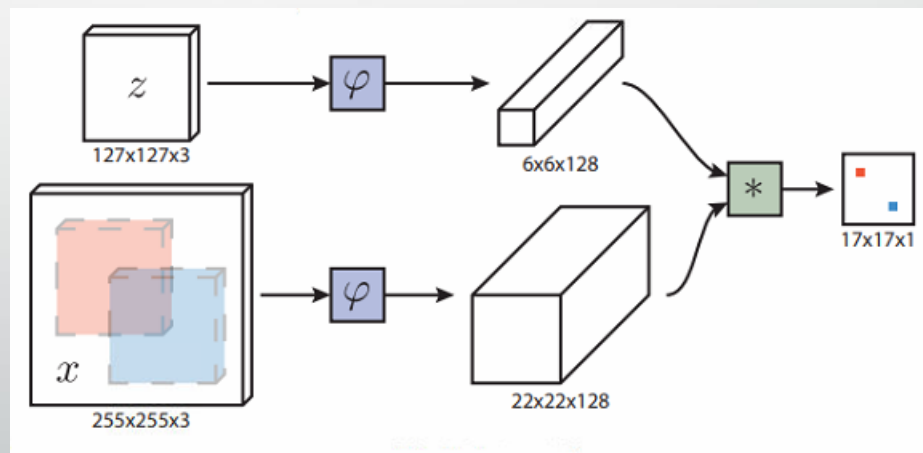


图7-5 SiamFC网络结构图

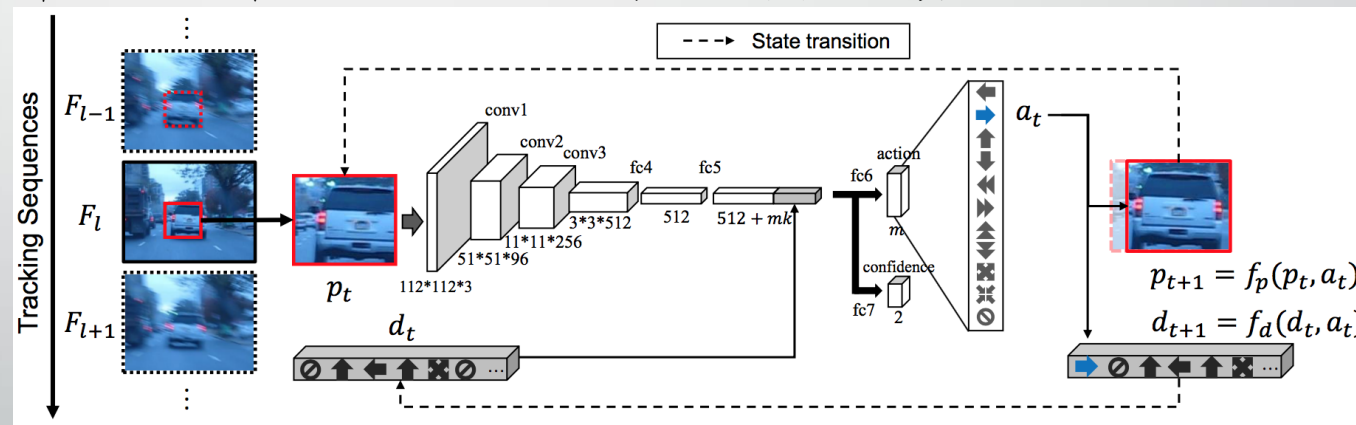
7.7 基于深度学习的目标跟踪

7.7.3 基于强化学习的目标跟踪

- 将强化学习的决策策略引入到目标跟踪任务中，以优化深度网络的参数、网络深度、或预测目标运动状态等信息。

➤ 实例：ADNet^[5]如下图所示

- 取马尔可夫决策过程的基本策略，将目标移动定义为离散化的动作，特征以及观察的历史状态形成当前状态，认为目标跟踪是一系列动作预测和状态变化的过程。在学习过程中采用深度神经网络并使用基于矩形框交并比的奖惩机制。
- 在训练阶段。分为监督学习和强化学习两个阶段。
 - ①在监督学习阶段利用视频序列优化目标位移及尺度变化等动作；
 - ②在强化学习阶段利用监督学习阶段训练的网络作为初始化，然后采取包含采样状态、动作、激励在内的训练序列进行跟踪仿真。



ADNet网络结构图

7.7 基于深度学习的目标跟踪

□ 与单目标跟踪相比，多目标跟踪的研究进展则缓慢得多，可用的数据集并不多，深度学习在该问题上的潜力也尚未被很好地挖掘出来。

- 常见的多目标跟踪算法一般可分为：

1. 基于检测的跟踪（DBT, Detection-Based Tracking），DBT要求由一个目标检测器首先将每帧图像中的目标检测出来。
2. 无检测的跟踪（DFT, Detection-Free Tracking），DFT要求已知每个目标首次出现的位置，再对每个目标分别进行跟踪，这一点可以看作是在同一个视频中进行的多个单目标跟踪。

- 多目标跟踪算法也可分为：

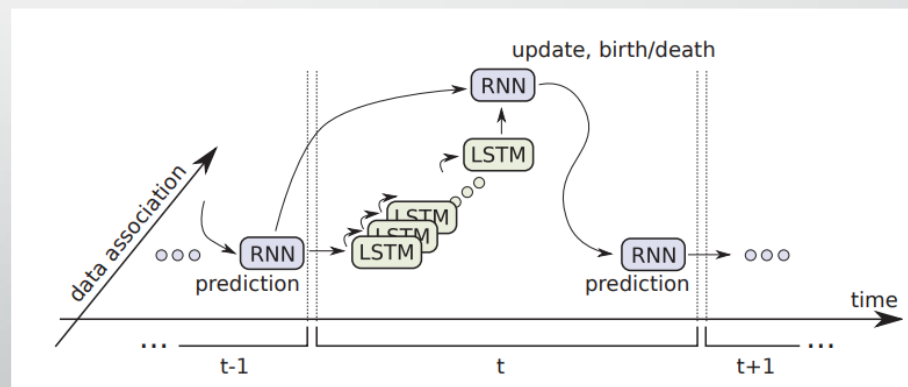
1. 在线跟踪（Online Tracking），在线跟踪要求处理每一帧时，决定当前帧的跟踪结果只能利用当前帧和之前帧中的信息，也不能根据当前帧的信息来修改之前帧的跟踪结果。
2. 离线跟踪（Offline Tracking），离线跟踪则允许利用之后的帧的信息从而获得全局最优解。

- 离线追踪的设定不太适合实际应用场景，但是以一种批处理的离线跟踪（每次得到若干帧，在这些帧中求全局最优）也是可行的，只是会导致一点延迟。

7.7 基于深度学习的目标跟踪

7.7.4 基于深度学习的多目标跟踪

- 作为使用深度学习解决多目标跟踪的一次尝试，Anton Milan等^[6]提出了使用 RNN 进行目标状态预测与数据关联的方法（如下图所示）。
- 这是第一个尝试以端到端的方式完成在线多目标跟踪的工作，目标的状态预测是一个连续空间，而数据关联又是一个离散空间，如何把这两个问题放到神经网络里是值得深入探究的问题，尤其是数据关联问题存在着诸多限制，比如需要满足输出的结果不能出现一对多的情况。



RNN用来进行时序信息预测和更新，以及跟踪管理。
每帧中的数据关联是利用LSTM实现的。