# 基于深度学习的多目标跟踪算法分类

多目标跟踪算法按照轨迹生成的顺序可以分为离线的多目标跟踪和在线的多目标跟踪算法。

离线的多目标跟踪

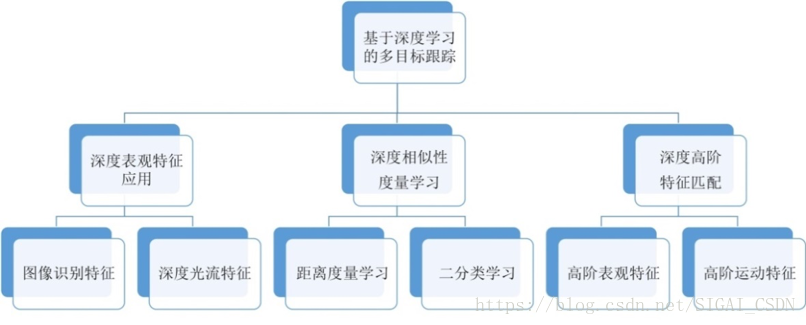
离线方式的多目标跟踪算法通常构造为目标检测关系的图模型，其中设计和计算检测之间的相似度或者距离度量是决定图模型构造正确性的关键。

在线的多目标跟踪

在线方式的多目标跟踪算法根据当前检测观测，计算与已有轨迹的匹配关系，计算合适的匹配度量决定了匹配的正确性。

无论是离线方式的多目标跟踪还是在线方式的多目标跟踪算法，学习检测结果的特征并计算匹配相似度或者距离度量都是多目标跟踪算法的关键步骤。

基于深度学习的多目标跟踪算法的主要任务是优化检测之间相似性或距离度量的设计。

表观特征的深度学习 基于相似性度量的深度学习 基于高阶匹配特征的深度学习

https://blog.csdn.net/qq\_23225317/article/details/78061230

图像目标检测任务在过去几年深度学习的发展背景下取得了巨大的进展，检测性能得到明显提升。但在视频监控、车辆辅助驾驶等领域，基于视频的目标检测有着更为广泛的需求。由于视频中存在运动模糊，遮挡，形态变化多样性，光照变化多样性等问题，仅利用图像目标检测技术检测视频中的目标并不能得到很好的检测结果。如何利用视频中目标时序信息和上下文等信息成为提升视频目标检测性能的关键。

对于视频目标检测来说，一个好的检测器不仅要保证在每帧图像上检测准确，还要保证检测结果具有一致性/连续性（即对于一个特定目标，优秀的检测器应持续检测此目标并且不会将其与其他目标混淆，称之为视频目标检测时序一致性）。

视频目标检测算法目前主要使用了如下的框架:

* 将视频帧视为独立的图像，利用图像目标检测算法获取检测结果；
* 利用视频的时序信息和上下文信息对检测结果进行修正；
* 基于高质量检测窗口的跟踪轨迹对检测结果进一步进行修正。

### **单帧图像目标检测**

此阶段通常将视频拆分成相互独立的视频帧来处理，通过选取优秀的图像目标检测框架以及各种提高图像检测精度的技巧来获取较为鲁棒的单帧检测结果。

#### **1. 训练数据选取**

需要注意，同一个视频片段背景单一，相邻多帧的图像差异较小。所以要训练现有目标检测模型，训练集可能存在大量数据冗余，并且数据多样性较差，有必要对其进行扩充。

#### **2. 网络结构选取**

不同的网络结构对于检测性能也有很大影响。同样的训练数据，基于 ResNet101 的 Faster R-CNN 模型的检测精度比基于 VGG16 的 Faster R-CNN 模型的检测精度高12%左右。今年比赛前几名的队伍基本上也是使用ResNet/Inception的基础网络。

### **三、改进分类损失**

目标在某些视频帧上会存在运动模糊，分辨率较低，遮挡等问题，即便是目前最好的图像目标检算法也不能很好地检测目标。幸运的是，视频中的时序信息和上下文信息能够帮助我们处理这类问题。比较有代表性的方法有T-CNN中的运动指导传播（Motion-guided Propagation, MGP）和多上下文抑制（Multi-context suppression, MCS）。

#### **1. MGP**

单帧检测结果存在很多漏检目标，而相邻帧图像检测结果中可能包含这些漏检目标。所以我们可以借助光流信息将当前帧的检测结果前向后向传播，经过MGP处理可以提高目标的召回率。如图1所示将T时刻的检测窗口分别向前向后传播，可以很好地填补T-1和T+1时刻的漏检目标。

#### **2. MCS**

使用图像检测算法将视频帧当做独立的图像来处理并没有充分利用整个视频的上下文信息。虽然说视频中可能出现任意类别的目标，但对于单个视频片段，只会出现比较少的几个类别，而且这几个类别之间有共现关系（出现船只的视频段中可能会有鲸鱼，但基本不可能出现斑马）。所以，可以借助整个视频段上的检测结果进行统计分析：对所有检测窗口按得分排序，选出得分较高的类别，剩余那些得分较低的类别很可能是误检，需对其得分进行压制（如图2）。经过MCS处理后的检测结果中正确的类别靠前，错误的类别靠后，从而提升目标检测的精度。

### **四、利用跟踪信息修正**

上文提到的MGP可以填补某些视频帧上漏检的目标，但对于多帧连续漏检的目标不是很有效，而目标跟踪可以很好地解决这个问题。

1.使用图像目标检测算法获取较好的检测结果；

2.从中选取检测得分最高的目标作为跟踪的起始锚点；

3.基于选取的锚点向前向后在整个视频片段上进行跟踪，生成跟踪轨迹；

4.从剩余目标中选择得分最高的进行跟踪，需要注意的是如果此窗口在之前的跟5.踪轨迹中出现过，那么直接跳过，选择下一个目标进行跟踪；

6.算法迭代执行，可以使用得分阈值作为终止条件。

### **五、网络选择与训练技巧**

对于视频目标检测，除了要保证每帧图像的检测精度，还应该保证长时间稳定地跟踪每个目标。

评价指标：图像目标检测mAP评测对象是每个检测窗口是否精准，而视频时序一致性评测对象是目标跟踪轨迹是否精准；图像目标检测中如果检测窗口跟Ground Truth类别相同，窗口IoU大于0.5就认定为正例。而评价时序一致性时，如果检测得到的跟踪轨迹和Ground Truth（目标真实跟踪轨迹）是同一个目标（trackId相同），并且其中检测出的窗口与Ground Truth窗口的IoU大于0.5的数量超过一个比例，那么认为得到的跟踪轨迹是正例；跟踪轨迹的得分是序列上所有窗口得分的平均值。分析可知，如果一个目标的轨迹被分成多段或者一个目标的跟踪轨迹中混入其他的目标都会降低一致性。

那么如何保证视频检测中目标的时序一致性呢？

* 保证图像检测阶段每帧图像检测的结果尽量精准；
* 对高质量检测窗口进行跟踪并保证跟踪的质量（尽量降低跟踪中出现的漂移现象）；
* 前面两步获取到的跟踪结果会存在重叠或者临接的情况，需针对性地进行后处理。

人工智能在视频上的应用主要一个课题是视频理解，努力解决****“语义鸿沟”****的问题，其中包括了：

视频结构化分析：即是对视频进行帧、超帧、镜头、场景、故事等分割，从而在多个层次上进行处理和表达。

目标检测和跟踪：如车辆跟踪，多是应用在安防领域。

人物识别：识别出视频中出现的人物。

动作识别：Activity Recognition，识别出视频中人物的动作。

情感语义分析：即观众在观赏某段视频时会产生什么样的心理体验。

## 基于单帧的识别方法

一种最直接的方法就是将视频进行截帧，然后基于图像粒度(单帧)的进行deep learninig 表达， 如图2所示，视频的某一帧通过网络获得一个识别结果。图2为一个典型的CNN网络，红色矩形是卷积层，绿色是归一化层，蓝色是池化层 ，黄色是全连接层。然而一张图相对整个视频是很小的一部分，特别当这帧图没有那么的具有区分度，或是一些和视频主题无关的图像，则会让分类器摸不着头脑。因此，学习视频时间域上的表达是提高视频识别的主要因素。当然，这在运动性强的视频上才有区分度，在较静止的视频上只能靠图像的特征了。

## 基于CNN扩展网络的识别方法

它的总体思路是在CNN框架中寻找时间域上的某个模式来表达局部运动信息，从而获得总体识别性能的提升。图3是网络结构，它总共有三层，在第一层对10帧 (大概三分之一秒)图像序列进行MxNx3xT的卷积(其中 MxN是图像的分辨率，3是图像的3个颜色通道，T取4，是参与计算的帧数，从而形成在时间轴上4个响应)，在第2、3层上进行T=2的时间卷积，那么在第3层包含了这10帧图片的所有的时空信息。该网络在不同时间上的同一层网络参数是共享参数的。

它的总体精度在相对单帧提高了2%左右，特别在运动丰富的视频，如摔角、爬杆等强运动视频类型中有较大幅度的提升，这从而也证明了特征中运动信息对识别是有贡献的。在实现时，这个网络架构可以加入多分辨的处理方法，可以提高速度。

## **双路CNN的识别方法**

这个其实就是两个独立的神经网络了，最后再把两个模型的结果平均一下。上面一个就是普通的单帧的CNN，而且文章当中提到了，这个CNN是在 ImageNet的数据上pre-train，然后在视频数据上对最后一层进行调参。下面的一个CNN网络，就是把连续几帧的光流叠起来作为CNN的输入。 另外，它利用multi-task learning来克服数据量不足的问题。其实就是CNN的最后一层连到多个softmax的层上，对应不同的数据集，这样就可以在多个数据集上进行 multi-task learning。网络结构如图4所示。