上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

智能视频分析

基于背景建模的视频前景检测和
基于 Faster R-CNN 的视频目标检测

电子系

XX

019XXXXXXXXX

导师：XX

2020 年 5 月

摘 要

本次实验中，我选择了第三次课大作业中使用传统方法进行前景检测以及使用深度学习方法进行目标检测两个题目，并分别使用帧差法、中值滤波法和高斯混合模型实现了基于背景建模的视频前景检测，以及使用了 Faster R-CNN 算法实现了视频目标检测。

首先，我在 CDNET(ChangeDetection.NET)官网上选取了一段用于目标检测的视频，然后用 OpenCV 分别实现了帧差法、中值滤波法以及高斯混合模型，并对这些方法的实现效果进行了比较分析。接着我将视频拆帧，对每张帧图像用 Faster R-CNN 进行目标检测，并将检测后的帧图像合并成一段视频得到目标检测后的视频。最后，我对以上实现方法进行了总结，进一步加深了对目标检测算法的理解。

关键词：视频目标检测；帧差法；中值滤波；高斯混合模型；Faster R-CNN

1 理论基础

1.1 基于背景建模的视频目标检测

运动目标检测是指在序列图像中检测出变化区域并将运动目标从背景图像中提取出来。 通常情况下，目标分类、跟踪和行为理解等后处理过程仅仅考虑图像中对应于运动目标的像素区域，因此运动目标的正确检测与分割对于后期处理非常重要。然而，由于场景的动态变化，如天气、光照、阴影及杂乱背景干扰等的影响，使得运动目标的检测与分割变得相当困难。根据摄像头是否保持静止，运动检测分为静态背景和运动背景两类。大多数视频监控系统是摄像头固定的，因此静态背景下运动目标检测算法受到广泛关注，常用的方法有帧差法、背景减除法、混合高斯模型、光流法等。

1.1.1 帧差法

帧差法是最为常用的运动目标检测和分割方法之一，基本原理是在图像序列相邻两帧或三帧间采用基于像素的时间差分通过阈值化来提取出图像中的运动区域。首先，将相邻帧图像对应像素值相减得到差分图像，然后对差分图像二值化，在环境亮度变化不大的情况下，如果对应像素值变化小于事先确定的阂值时，可以认为此处为背景像素，如果图像区域的像素值变化很大，则认为这是由于图像中运动物体引起的，将这些区域标记为前景像素，利用标记的像素区域可以确定运动目标在图像中的位置。由于相邻两帧间的时间间隔非常短，用前一帧图像作为当前帧的背景模型具有较好的实时性，其背景不积累，且更新速度快、算法简单、计算量小。算法的不足在于对环境噪声较为敏感，阈值的选择相当关键，选择过低不足以抑制图像中的噪声，过高则忽略了图像中有用的变化。对于比较大的、颜色一致的运动目标，有可能在目标内部产生空洞，无法完整地提取运动目标。

1.1.2 背景减除法

背景减除法是一种有效的运动对象检测算法，基本思想是利用背景的参数模型来近似背景图像的像素值，将当前帧与背景图像进行差分比较实现对运动区域的检测，其中区别较大的像素区域被认为是运动区域，而区别较小的像素区域被认为是背景区域。 背景减除法必须要有背景图像，并且背景图像必须是随着光照或外部环境的变化而实时更新的，因此背景减除法的关键是背景建模及其更新。针对如何建立对于不同场景的动态变化均具有自适应性的背景模型，减少动态场景变化对运动分割的影响，研究人员已提出了许多背景建模算法，但总的来讲可以概括为非回归递推和回归递推两类。 非回归背景建模算法是动态的利用从某一时刻开始到当前一段时间内存储的新近观测数据作为样本来进行背景建模。非回归背景建模方法有最简单的帧间差分、中值滤波方法。回归算法在背景估计中无需维持保存背景估计帧的缓冲区，它们是通过回归的方式基于输入的每一帧图像来更新某个时刻的背景模型，这类方法包括广泛应用的线性卡尔曼滤波法、混合高斯模型等。

1.1.3 高斯混合模型(GMM)

混合高斯模型就是指对样本的概率密度分布进行估计，而估计的模型是几个高斯模型加权之和[2]。

$$p(x)=\sum_{k=1}^{K}p(k)p(x|k)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k)$$

对样本中的数据分别在几个高斯模型上投影，就会分别得到在各个类上的概率。然后选取概率最大的 类作为判决结果。理论上可以通过增加模型的个数，用 GMM 近似任何样本的概率分布。一般混合高斯模型使用 k 个（基本为 3 到 5 个）高斯模型来表征图像中各个像素点的特征，在新一帧图像获得后更新混合高斯模型, 用当 前图像中的每个像素点与混合高斯模型匹配。如果成功则判定该点为背景点, 否则为前景点。由于是对运动目标的背景提取建模，因此需要对高斯模型中的方差和均值两个参数进行实时更新。为提高模型的学习能力，一些改进方法对均值和方差的更新采用不同的学习率进行了学习。

1.1.4 基于 VIBE 的背景建模

Vibe 算法全称为 Visual Background Extractor，属于一种新的背景建模方法，是由 Olivier Barnich 和 Marc Van Droogenbroeck 于 2009 年首次提出[3]。Vibe 是一种类似于混合高斯模型的像素级背景建模方法，其特点在于引入了随机变量，通过对每一个像素点随机地选取其临近点的像素值，作为该点的有限样本集中的样本值。Vibe 算法的优点在于具有较强的抗噪能力，算法简单，而且其占用内存少，初始化快，易于实现。Vibe 算法与其他算法最重要的区别在于，它不对背景像素的概率密度进行估计，而是对每个像素建立一组样本，将每个像素的当前值与样本集进行比较，并根据新像素值样本与集中样本点的接近程度，来判定该点是否属于背景像素。

1.1.5 光流法

光流法的主要任务就是计算光流场，即在适当的平滑性约束条件下，根据图像序列的时空梯度估算运动场，通过分析运动场的变化对运动目标和场景进行检测与分割。通常有基于全局光流场和特征点光流场两种方法。最经典的全局光流场计算方法是 L-K(Lueas&Kanada) 法和 H-S(Hom&Schunck)法[4]，得到全局光流场后通过比较运动目标与背景之间的运动差异对运动目标进行光流分割，缺点是计算量大。特征点光流法通过特征匹配求特征点处的流速，具有计算量小、快速灵活的特点 但稀疏的光流场很难精确地提取运动目标的形状。 总的来说，光流法不需要预先知道场景的任何信息，就能够检测到运动对象，可处理背景运动的情况，但噪声、多光源、阴影和遮挡等因素会对光流场分布的计算结果造成严重影响，而且光流法计算复杂，很难实现实时处理。

1.2 基于深度学习的目标检测

随着深度学习在计算机视觉领域的发展，目标检测领域也取得了长足的进步。目前主流的基于深度学习模型的目标检测算法可以分成两大类。目标检测模型的主要性能指标是检测准确度和速度，其中准确度主要考虑物体 的定位以及分类准确度。一般情况下，Two-Stage 算法在准确度上有优势， 而 One-Stage 算法在速度上有优势。

1.2.1 Two Stage 目标检测算法

Two-Stage 目标检测算法，这类检测算法将检测问题划分为两个阶段， 第一个阶段首先产生候选区域（Region Proposals），包含目标大概的位置信息，然后第二个阶段对候选区域进行分类和位置精修，这类算法的典型代表有 R-CNN[5]，Fast R-CNN[6]，Faster R-CNN[7]等。

本次实验我主要采用了 Faster R-CNN 来对视频进行目标检测。Faster R-CNN 是基于深度学习对象检测的一个典型案例，图 1 是 Faster R-CNN 的主要架构：

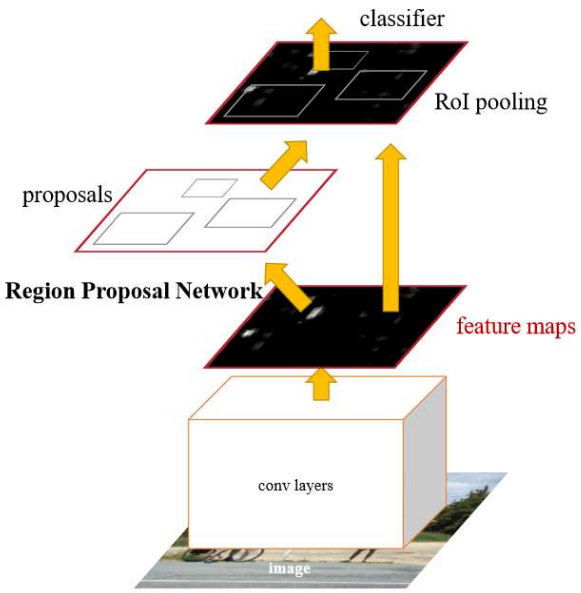

图 1. Faster R-CNN 的主要架构

Faster R-CNN 用一个快速神经网络代替了运算速度很慢的选择性搜索算法：通过插入区域提议网络（ RPN ），来预测来自特征的建议。 RPN 决定查看“哪里”，这样可以减少整个推理过程的计算量。RPN 快速且高效地扫描每一个位置，来评估在给定的区域内是否需要作进一步处理，其实现方式如下：通过输出 k 个边界框建议，每个边界框建议都有 2 个值——代表每个位置包含目标对象和不包含目标对象的概率。一旦我们有了区域建议，就直接将它们送入 Fast R-CNN 。 并且，Faster R-CNN 还添加了一个池化层、一些全连接层、一个 softmax 分类层以及一个边界框回归器。总之，Faster R-CNN 的速度和准确度更高。值得注意的是，虽然以后的模型在提高检测速度方面做了很多工作，但很少有模型能够大幅度的超越 Faster R-CNN 。

1.2.2 One Stage 目标检测算法

One-Stage 目标检测算法，这类检测算法不需要 Region Proposal 阶段， 可以通过一个 Stage 直接产生物体的类别概率和位置坐标值，比较典型的算法有 YOLO[8]和 SSD[9]等。

本次实验中我没有采用 One Stage 目标检测算法，这里只做简单的介绍。YOLO 舍去了候选框提取分支（Proposal 阶段），直接将特征提取、候选框回归和分类在同一个无分支的卷积网络中完成，使得网络结构变得简单，检测速度较 Faster RCNN 也有近 10 倍的提升。算法将待检测图像缩放到统一尺寸，为了检测不同位置的目标，将图像等分成的网格，如果某个目标的中心落在一个网格单元中，此网格单元就负责预测该目标。这使得深度学习目标检测算法在当时的计算能力下开始能够满足实时检测任务的需求。

SSD 对 YOLO 进行了改进，达到了和两阶段方法相当的精度，同时又保持了较快的运行速度。SSD 也采用了网格划分的思想，和 Faster RCNN 不同的是它将所有的操作整合在一个卷积网络中完成。为了检测不同尺度的目标，SSD 对不同卷积层的特征图像进行滑窗扫描；在前面的卷积层输出的特征图像中检测小的目标，在后面的卷积层输出的特征图像中检测大的目标。它的主要特点是基于多尺度特征图像的检测，在多个尺度的卷积特征图上进行预测，以检测不同大小的目标，一定程度上提升了小目标物体的检测精度。并且借鉴了 Faster R-CNN 中的 Anchor boxes 思想，在不同尺度的特征图上采 样候选区域，一定程度上提升了检测的召回率以及小目标的检测效果。

2 实验操作及结果

为了巩固上述前景检测和目标检测算法的理解，我通过编程实现了基于帧差法、中值滤波、高斯混合模型的视频前景检测，以及基于 Faster R-CNN 的目标检测。首先，我在 CDNET(ChangeDetection.NET)官网上找了一个 40s 左右适用于目标检测的短视频，视频是在车站月台场景下拍摄的，包含了来来往往的行人，背景是一辆未移动的火车和月台的地面。图 2 是视频中选取的某些帧图像。

图 2. 测试视频中的 3 张帧图像

2.1 基于背景建模的前景检测

2.1.1 帧差法

对于帧差法的实现，我主要借助了 OpenCV 库中的 cv2.absdiff 函数，通过相邻两帧图像的差异捕捉到运动前景。图 3 是我使用帧差法获得帧图像，完整的检测视频见附件中的 PPT。从效果上来看，当运动目标离镜头较近，或者移动速度较快时，检测出的前景效果比较显著，但是如果物体离镜头较远，或者移动比较微小，算法就会把他当作是背景而未能识别出来。

图 3. 使用帧差法获得的帧图像

2.1.2 中值滤波

中值滤波是背景减除法中一个典型的算法，对于该算法的实现，我主要借助了 OpenCV 库中的 cv2.medianBlur 函数。图 4 是我使用中值滤波法获得帧图像，完整的检测视频见附件中的 PPT。中值滤波法和帧差法获得的检测效果相近，同样也是当运动目标离镜头较近，或者移动速度较快时，检测出的前景效果比较显著。因为中值滤波只是一种单纯的数学统计模型，所以相对于帧差法效果的提升并不明显。

图 4. 使用中值滤波获得的帧图像

2.1.3 高斯混合模型

最后，我还使用了高斯混合模型对视频进行前景目标的检测。高斯混合模型是对样本的概率密度分布进行估计，在实现过程中，我同样是借助了 OpenCV 库，除了检测出移动的前景目标，我还在这些目标上增加了矩形框从而使检测效果更加显著。图 5 是我使用高斯混合模型获得帧图像，从效果上看，相比于之前的帧差法和中值滤波法，GMM 生成的帧图像检测出的前景更加显著，即使是离镜头较远的目标也可以清

晰地识别出来，并且在有光照的情况下还可以识别出目标的影子，但是 GMM 仍然存在缺陷，当目标被静态物体遮挡，比如说月台的栏杆，被遮挡的部分就不能被识别出来，从而造成检测存在误差，而且由于现实中复杂的环境，GMM 算法生成的帧图像中仍然存在一定的误差。

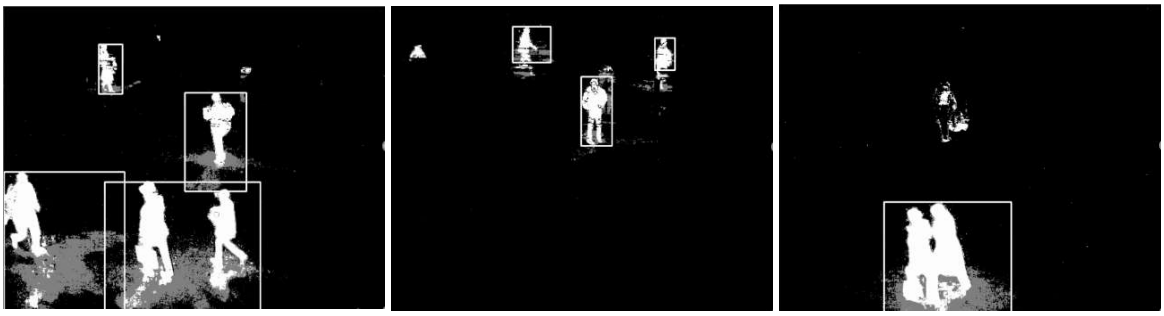

图 5.使用高斯混合模型获得的帧图像

2.2 基于 Faster R-CNN 的目标检测

由于 Faster R-CNN 原本是对图像进行目标检测的，实验中我首先使用 OpenCV 将原本 40s 的短视频进行拆帧，共得到了 1199 张帧图像，每张图像的大小为 720*576 像素，然后我使用 Faster R-CNN 分别对这 1199 张图像进行目标检测，在 Linux 系统中 2 核 CPU、8G 内存的环境下共运行了约 120min，图 6 是其中的 3 张帧图像经过目标检测后取得的效果。

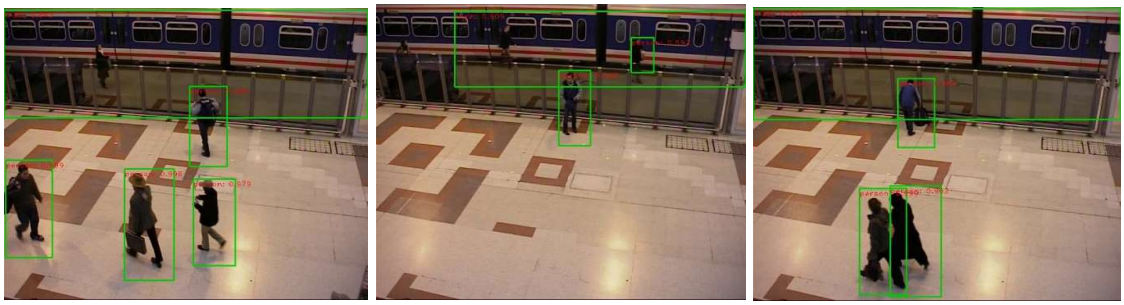

图 6.使用 Faster R-CNN 获得的帧图像

最后，我再次使用 OpenCV 将经过目标检测后的这些帧图像进行合并，获得经过目标检测的视频，合并后的视频见附件的 PPT 中。从实现效果上来看，Faster R-CNN 基本可以将目标准确的识别出来，生成的 bounding box 随着目标的移动而移动，而且除了移动的行人，背景中静止的火车也可以检测出来，但是算法仍然存在缺陷，比如离镜头较远且被遮挡的行人有时不能被识别出，而且静止的火车四周的 bounding box 也一直在抖动。

3 实验总结

本次实验中，我分别使用帧差法、中值滤波法和高斯混合模型实现了基于背景建模的视频前景检测，并使用了 Faster R-CNN 算法实现了视频目标检测。

从实现效果来看，帧差法和中值滤波法都是实现较为简单的算法，两者的检测效果相近，对于离镜头较近或者移动较为明显的目标检测效果比较显著，而对于离镜头较远并且移动微小的目标检测效果不是很理想。而相比于帧差法和中值滤波法，使用高斯混合模型进行前景检测的效果得到了明显提高，即使是离镜头较远的目标也可以清晰地识别出来，并且可以检测出由于光照引起的目标的影子，但是对于被遮挡的目标识别效果不是很好，还存在一些噪声。

基于 Faster R-CNN 的视频目标检测，检测效果较为理想，视频中移动的目标可以用矩形框精准地定位出，对于被静止物体遮挡的目标也可以准确识别出来，但是对于离镜头较远的小目标，算法有事存在抖动，检测效果不佳。

总的来说，通过这次实验，无论是从理论还是实践上都加深了我对视频目标检测的理解，帮助我对智能视频分析这门课程有了更深一步的了解。

参考文献

[1] 郑世宝．智能视频分析2020春第三章课件．上海．

[2] C.Stauffer, W.E.L.Grimson. Adaptive background mixture models for real-time tracking [C]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999:246252．

[3] O. Barnich and M. Van Droogenbroeck. “Vibe: A powerful random technique to estimate the background in video sequences.” In IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pages 945 –948, 2009.

[4] H.K.P.Horn, B.G.Schunck. Determining optical flow[J]. Artificial Intelligence. 1981, 17(1-3):185- 204．

[5] R.Girshick, J.Donahue, T.Darrell, and J.Malik, Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR 2014.

[6] R.Girshick, Fast R-CNN. In ICCV, 2015.

[7] S.Ren, K.He, R.Girshick, and J.Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In CVPR, 2016.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. SSD: Single shot multibox detector. In ECCV, 2016.