基于 Faster-RCNN 的智能家居行人检测系统设计与实现

Smart Home Pedestrian Detection System Based on Faster-RCNN

朱 虹 翟 超 吕 志 程 风 （中国科学技术大学工程科学学院，安徽 合肥 230027）

摘要：行人检测是目标检测领域的重要应用，卷积神经网络（CNN）在目标识别中表现优异。设计与实现了运用 Faster-RCNN 算法在智能家居中的行人检测系统，并且通过 GPU 加速实现实时检测，该系统的基本功能有：①运用 Faster-RCNN 算法检测行人并保存下图像；②发送该图像到网络服务器；③将保存的图片推送到手机，对智能家庭的安全性提供保障。实验证明，该系统在复杂的场景中具有良好的性能，且每张图片 78ms 的检测时间可以满足实时性的要求。

关键词：Faster-RCNN，智能家居，行人检测，GPU 加速，实时检测图片推送

Abstract:This paper designs and implements a pedestrian detection system in smart home with Faster-RCNN algorithm and use GPU to accelerate computation,the basic functions of which include 3 aspects: ①using Faster-RCNN algorithm to detect pedestrians and save images;②sending images to web server;③pushing pictures uploaded to server to mobile phone to protect the security of smart home.

Keywords:Faster-RCNN,smart home,pedestrian detection,GPU acceleration,real-time detection,picture push

家庭环境不比公司或者商场等大型环境可以配备专业的人员时刻观察监控视频。因此，如果可以只保存视频中有行人的图片，将给我们带来很大的便利。基于此，智能家居监控视频中的行人检测技术的研究就显得很有必要。行人检测技术是通过图像和计算机视觉处理算法，对所给视频或者图片中的行人进行智能分类识别。目前常用的行人检测的方法主要有光流法、帧间差法、背景差法和基于机器学习等方法。前面的三种检测方法都是常规的基于图像处理技术的行人检测方法，然而这些方法不能解决人体形状和外貌各式各样的难点，人体的不同运动方式的问题，受天气以及光照的随机变化，行人的服饰和姿态改变影响较大。2012 年起，深度学习引起了人们的广泛关注，并在图像识别与检测中取得了良好的识别效果。本文研究了深度学习在目标检测中的应用，并且将目标检测中的 Faster-RCNN[2]算法运用到行人检测当中。

1 Faster-RCNN

在 R-CNN[3]和 Fast-RCNN[4]之后，为了进一步减少检测网络的运行时间，微软 Shaoqing.Ren 等提出了新的目标检测方法 Faster-RCNN。他们设计一种区域建议网络（region proposal network，RPN）来生成区域建议网络（region proposals）替代之前 Selective Search[5]和 EdgeBoxes[6]等方法，它和检测网络共享卷积特征，使得 region proposals 检测几乎不花时间，另外，空间金字塔池化的使用使得网络可以处理任意大小的图片。

RPN 是全卷积的网络结构[7]，它能同时预测输入图片产生的候选框的位置信息和属于真实目标的概率。通过 RPN 和 Fast-RCNN 交替训练的方式运行训练，RPN 和 Fast-RCNN 可以在训练时共享卷积特征。由此可见，Faster-RCNN 的整体结构可以认为是“RPN+Fast-RCNN”的合成。RPN 网络主要用于生成高质量建议区域框，Fast-RCNN 则是提取高质量建议区域的特征并且对建议区域进行分类。Faster-RCNN 在生成建议区域的改进，使得检测效率提升。选取 ZF-Net[8]与 Faster-RCNN 相结合进行行人检测。

2 基于 Faster-RCNN 的行人检测

与 SPPNet 和 Fast-RCNN 相比，Faster-RCNN 方法既降低了生成建议区域的时间瓶颈，又能确保较为理想的识别率。因此，本文以 Faster-RCNN 识别方法为主，对家庭环境中的行人进行检测。

2.1 行人检测的网络训练

Faster-RCNN 方法包含 2 个 CNN 网络：区域建议网络 RPN（Regional Proposal Network）和 Fast-RCNN 检测网络。都需要用 ImageNet 网络进行初始化[9]，两个网络训练阶段的主要步骤如图 1 所示。

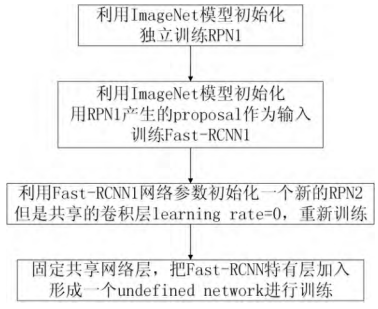

图 1 行人检测联合网络训练过程

1）RPN 网络训练。区域生成网络可以用随机梯度下降（stochastic gradient descent，SGD）进行端到端（end-to-end）的训练。采用采样方法进行训练，每个 mini-batch 从一张图片中采样多个正样本及负样本锚点，从每张图片中随机选取 256 个锚点计算每个 mini-batch 的损失函数。其中正/负样本的比例最高容许为 1:1。如果一张图片中少于 128 个正样本，用负样本填充 mini-batch。

训练中，用零均值，标准差为 0.01 的高斯分布随机初始化所有新加的层。所有其他层（例如，共享的卷积层）会由 ImageNet 分类任务预训练的模型初始化。对于 ZF 网络，调整所有层的参数。

2）Fast-RCNN 检测网络训练。Fast-RCNN 检测网络也利用 ImageNet 预训练的网络初始化。对输入图像进行 5 层卷积网络的特征提取，第 5 层特征图（CONV5）是一个 256×256 的特征图，将 256 个通道内的全部特征串联成一个高维（4096 维）特征向量，后面添加另一个 4096 维的特征层，形成 FC7。由 FC7 特征层可预测：①候选区域框属于每个类别的概率；②候选区域对应的目标对象的更合适的位置，用它相对于候选区域框的 2 个平移和 2 个放缩共 4 个参数表示。通过预先标记的信息利用反向传播算法对该检测网络进行微调。

3)2 个网络的 CNN 共享和联合调优。利用 2)得到的检测网络初始化区域生成网络的训练，并且固定住所有共享的卷积层，只微调区域生成网络独有的参数层。这样两个网络就可以共享卷积层了。最后，保持共享的卷积层不变，微调检测网络独有的层。经过这样四步，两个网络就可以共享卷积层，并可以被合并为一个统一的网络进行测试。同样的交替优化可以被进行多轮，但是更多的轮数并不能带来更多效果上的提升。

2.2 检测识别过程

由上面的训练可知，2 个网络最终可共用同一个 5 层的卷积神经网络，这使整个检测过程只需完成系列卷积运算即可完成检测识别过程，彻底解决了原来区域建议步骤时间长开销大的瓶颈问题。检测识别的过程如图 2 所示，其实现步骤为：①对整个图像进行系列卷积运算，得到特征图 CONV5；②由区域建议网络在特征图上生成大量候选区域框；③对候选区域框进行非最大值抑制，保留得分较高的前 2000 个框；④取出特征图上候选区域框内的特征形成高维特征向量，由检测网络计算类别得分，并预测更合适的目标外围框位置。

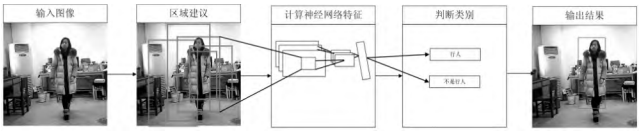

图 2 检测识别过程

2.3 CUDA 编程对非极大值抑制部分的加速

在物体检测中，非极大值抑制[10]应用十分广泛，主要目的是为了消除多余的框，找到最佳的物体检测的位置。由网络产生的候选框个数多并且重叠比例高，所以根据置信度的高低，可以去掉一些重叠比高、置信度低的候选框。原理如图 3 所示。

图 3 非极大值抑制抑制原理图

CUDA 技术[11]通过大量可并行执行的线程对算法进行加速，并由 GPU 动态调度和执行。CUDA 的结构图如图 4 所示。其中 Thread 为线程，多个 Thread 组成一个 Block，多个 Block 又组成一个 Grid。相同计算过程不同变量的同时计算可以大大减少运算所需要的时间。

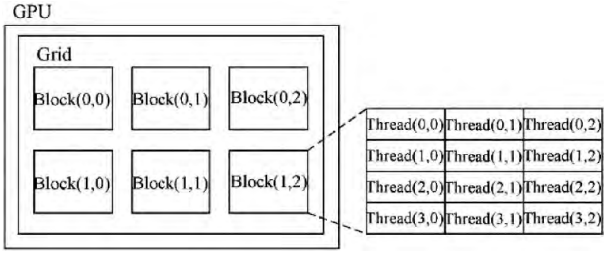

图 4 CUDA 结构图

表 1 nms 部分耗时统计

使用 CPU 计算耗时（ms）	使用 GPU 计算耗时（ms）
1155	14

当计算很多个不同的候选框之间的重叠比时，进行的就是大量重复的简单操作，所以该过程完全可以用 CUDA 进行加速。本文方法在执行时将每个候选框分别安排一个线程，每个 Block 里面的多个 Thread 可以同时计算，也就是同事可以计算多组候选框之间的重叠比。通过多个线程并行处理，最终输出结果。在 RPN 的非极大值抑制计算部分采用了此种加速方法，RPN 网络中对按照置信度从高到低留下的 3000 个候选框进行非极大值抑制的计算。测试时间如表 1 所示。实验效果表明，利用 CPU 计算 3000 个候选框的各自重叠比是 GPU 计算时间的 128 倍多，利用 CUDA 编程可以有效地加速运算。

2.4 检测结果

部分检测结果的图片如图 5 所示，由检测结果可知：即使在干扰很多的复杂背景下，该方法也能够准确地检测出视频中的行人，并且不受行人姿势的影响。对于人数较多的场景也有较好的检测效果。

图 5 部分检测结果示意图

测试时，当识别出的外围框与标记的外围框重叠面积达到标记外围框的 90%以上时，视为一次成功的识别。本次试验中，用正确率和召回率来评判识别的准确性，其中正确率为目标类别标记正确的外围框个数除以所有标记出的外围框个数；召回率为目标类别标记正确的外围框个数除以所有标准的外围框个数。表 2 记录了 Faster-RCNN 算法的准确率和召回率。

表 2 Faster-RCNN 的识别率和召回率

Faster-RCNN 正确率	87.1%
Faster-RCNN 召回率	91.2%

本文的实验都是在基于 GPU 实现，非极大值抑制部分也是使用 GPU 来实现，显卡使用 Nivdia 公司的 GEFORCEGTX1060 (3G 显存)，表 3 记录了 Faster-RCNN 进行行人检测时所需花费的时间，从中可以看出，由于卷积特征的共用，使得区域建议的时间几乎可以忽略不计，检测时间可以在 70ms 内完成。结果表明，采用深度学习方法 Faster-RCNN 算法可以实现对图片中行人的实时检测。

表 3 Faster-RCNN 各部分计算开销

平均时间	区域建议数量	关键步骤计算时间	关键步骤计算时间	关键步骤计算时间
78ms	平均 17900，取置信度高的前面 2000 个	卷积+区域建议时间 42ms	非极大值抑制时间 14ms	Fast-RCNN 计算时间 25ms

3 系统框架与系统实现

系统用 Visual Studio 编程实现，完整的系统框图如图 6 所示，由于系统既有图像的处理又有网络间的通信，对于多路上传图片到服务器的部分，服务器端采用多线程编程。主线程主要用于视频的显示，图片中有无行人的检测，子线程用于上传图片到服务器和发送到手机。图片上传到服务器的部分采用 socket 实现[12]，上传之后根据每张图片的 URL 和每个手机 App 各自独立的 appId、appKey、masterSecret、cid 将每个线程保存的图片通过个推发送到指定的手机中。当一个线程对应的用户保存到图片之后，将会给对应的手机 App 发送一条消息推送，效果如图 7 所示，点开推送消息之后就可以在网页上看到上传到服务器的图片。

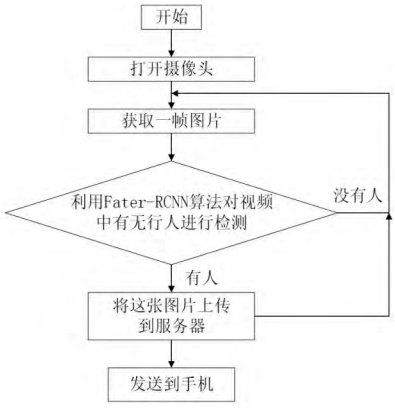

图 6 系统框图

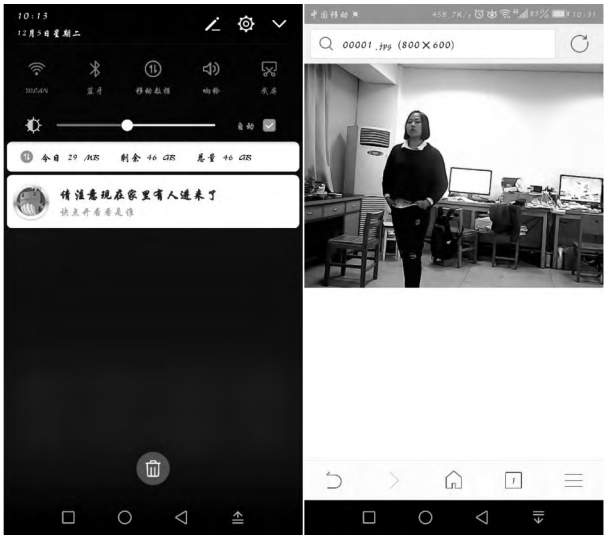

图 7 手机推送消息示意图

4 结束语

本文方法需要通过 GPU 对非极大值抑制部分进行加速，使得单张图片的检测时间大约为 78ms，才可以满足视频实时检测的要求，这就对硬件要求比较高。近年来，很多学者通过压缩深度学习的网络模型、对输入数据和权值数据进行量化等方法，希望在一般的 CPU 环境下也可以对行人进行实时的检测，这是行人检测应用发展的一个方向，也是我们下一步研究的重点。

参考文献

[1]徐渊，许晓亮，李才年，等.结合 SVM 分类器与 HOG 特征提取的行人检测[J].计算机工程，2016，42(1)：56-60，65

[2]Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [C]// 29th Annual Conference on Neural Information Processing Systems, NIPS 2015, Montreal, QC,Canada,December 7-12, 2015. Canada: NIPS Conference, 2015:91-99

[3]Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proceeding of the IEEE conference on computer vision and pattern recognition, Columbus, OH, United states, June, 23-28, 2014. USA: IEEE, 2014:580-587

[4]Girshick R.Fast R-CNN[C]//Proceedings of the IEEE international conference on computer vision, Santiago, Chile, December 11-18, 2015. USA: IEEE, 2015:1440-1448

[5]Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition [J].International journal of computer vision, 2013, 104(2):154-171

[6]Zitnick C L, Dollar P.Edge boxes: Locating object proposals from edges[C]//13th European Conference on Computer Vision, ECCV 2014, Zurich, Switzerland, September, 6-12, 2014. Zurich: Springer, Cham, 2014:391-405

[7]Long J, Shelhamer E,Darrell T.Fully convolutional networks for semantic segmentation [C]//Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, United states, June 7-12, 2015. USA: IEEE, 2015: 3431-3440

[8]Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C]//European conference on computer vision, Zurich, Switzerland, September, 6-12, 2014. Zurich: Springer, Cham, 2014:818-833

[9]Ren S, He K. Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [C]// Neural Information Processing Systems, 2015

[10]Parikh D, Zitnick C L.Finding the weakest link in person detectors [C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR).Providence,RI:IEEE,2011:1425-1432

[11]彭景维，童基均.基于 GPU_CPU 异构并行加速的人头检测方法[J].计算机系统应用，2017，26(11)：95-100

[12]刘运强，王汇源.Socket 和多线程在视频传输的应用[J].山东大学学报(工学版)，2004，34(2)：45-50

[收稿日期：2018.1.8]