doi:10.3772/j.issn.1002-0470.2020.01.004

面向嵌入式系统的高精度实时人群计数算法研究①

金 鑫②* 赵 旭** 赵朝阳** 徐华中* 王金桥**

(*武汉理工大学自动化学院 武汉 430070)

(**中国科学院自动化研究所 北京 100190)

摘 要 针对大部分基于深度学习的方法在一些计算力低的嵌入式设备中难以保证速度和精度上同时满足实际需求的问题,本文提出了一种面向嵌入式设备的基于深度学习的快速人群计数算法。本文首先设计了弱计算力平台加速网络(LPANet),结合单阶段目标检测算法对图像中出现的人体头肩区域进行快速检测,然后引入快速多目标跟踪算法对连续多帧的检测结果进行轨迹关联,进而进行准确的人群计数。本文建立了完备的头肩数据集对算法进行验证,算法以 640 ×480 的分辨率在 ARM 平台(双核 cortex-A72)上获得 20 帧/s 的运行速度和 1.15 的计数平均绝对误差精度。基于深度学习的头肩检测结合跟踪的人群计数算法为嵌入式设备在多种场景下的应用提供了可行方案。

关键词 人群计数;头肩检测;多目标跟踪;弱计算力平台加速网络(LPANet);头肩数据集

0 引 言

基于计算机视觉的人群计数算法即在图像或视频帧中对行人的轨迹进行分析,统计当前时段行人的数量/流量信息。人群计数算法主要分为基于回归模型和基于目标检测两类。

在基于回归模型的方法中,经典方法如 Davies 等人[1]在视频帧中提取前景像素和边缘特征等原始特征,然后从原始特征导出前景区域和总边缘计数之类的整体属性,最后利用线性回归模型建立整体属性与实际人数之间的直接映射。随着深度学习的发展, Zhang 等人[2]提出利用深度特征进行人数估计的方法,利用不同尺度卷积核的卷积层提取图片的深度特征,然后利用深度特征得到人群密度图,最后根据密度图估计图片中行人数量。基于回归方法的人群计数优点在于对密度极高的人群计数会有较强的鲁棒性和较快的速度,然而由于利用图片整体特征或密度图进行估计人数的策略无法获取图像中单人的位置,在一般场景下精度较低。

在基于目标检测的人群计数算法中,行人检测应用最为广泛。Dalal 等人[3]提出了利用梯度方向直方图(histogram of oriented gradient,HOG)特征检测图片中行人的数量,通过提取图像 HOG 特征的方式获取图像内行人的边缘特征,根据边缘特征检测到行人,直接统计行人位置和数量。在 HOG 之后,Dollar 等人[4]提出了聚合通道特征(aggregate channel feature,ACF)检测算法进行行人检测,通过对图片进行快速特征金字塔变换,然后聚合多通道特征,更加准确和高效地提取行人边缘特征,最终获取图片中行人数量。然而 HOG 和 ACF 等经典检测算法精度较低,无法满足实际需求。随着深度学习的发展,快速区域卷积神经网络[5](faster region-based convolutional neural networks,Faster R-CNN)在目标检测领域获得巨大的成功,它首先利用视觉几何组

① 道路交通安全公安部重点实验室开放基金(2018ZDSYSKFKT03),国家自然科学基金青年基金(61806200)和国家自然科学基金面上项目(61876086)资助。

② 男,1993 年生,硕士生;研究方向:模式识别与智能系统;联系人,E-mail: whut_jx@163.com

(收稿日期:2019-01-10)

网络[6](visual geometry group network,VGGNet)等卷积神经网络提取目标的深度特征,然后在深度特征的基础上利用区域候选网络(region proposal network,RPN)获得候选框,利用感兴趣区域特征提取(region of interest pooling,ROIPooling)操作和全连接层或高维度卷积层构成的子网络对候选框内深度特征进一步分析,最终预测目标检测结果。由于 Faster R-CNN[5]对于检测精度的显著提升,被广泛应用在行人检测领域。例如,Mao 等人[7]通过在 RPN 之前的通道中融入图像分割特征提升行人特征在特征图中的响应,从而提升行人检测的精度。Zhang 等人[8]利用注意力模块加强特征提取过程中被遮挡行人的可见部分特征响应,减少被遮挡行人的漏检。在减少由于行人之间相互遮挡导致部分行人检测框被非极大值抑制策略(non-maximum suppression,NMS)抑制,最终形成漏检的问题上,Wang 等人[9]提出了互斥损失函数,使得候选框在回归过程中不仅距离其匹配的目标距离减少,同时距离其他目标的距离增大,避免了候选框由于回归后相互之间仍有较大交并比而被 NMS 抑制的情况。然而,行人检测目前对于人群中的互相遮挡情况依然较为敏感,当人群中互相遮挡较为严重时,此类方法会产生较大的误差。Li 等人[10]选取视频序列中的头肩区域作为检测目标,并采用 FasterR-CNN[5]算法对每一个视频帧进行检测,然后采用核化相关滤波器(kernelized correlation filter,KCF)[11]算法对 FasterR-CNN[5]检测到的头肩结果进行跟踪,从而获取视频序列中每个头肩的跟踪轨迹,最终根据跟踪轨迹的数量确定人数。基于文献[5]的方法虽然获得了较高的精度,但是由于双阶段检测算法诸如 Faster R-CNN[5]和特征金字塔网络(feature pyramid networks,FPN)[12],在特征层经过 ROIPooling 后会采用维度很高的卷积层或者全连接层对 RPN 过程产生的每个候选框进一步地特征提取,导致模型时间复杂度急剧升高。采用 VGGNet[6]为骨干网络,该网络时间复杂度过高。因此,基于文献[5]的方法在高级精简指令集机器(advanced reduced instruction set computing machine,ARM)上运行效率极低。

目前在提升检测器执行效率的研究中,单阶段多边框目标检测(single shot multibox detector,SSD)[13]和统一实时对象检测(you only look once,YOLO)[14,15]等单阶段检测器采用直接利用卷积预测分类和回归的方式避免了计算量过高的问题。轻量级神经网络的研究例如 MobileNets[16,17]等可以在很大程度上缓解由于骨干网络计算量过大导致的运行速度过慢的问题。但实际上即使在融合了 MobileNets[16,17]等轻量级网络后,检测器依然难以在中低端 ARM 设备上进行高效部署。主要原因是在 ARM 设备上执行头肩检测等任务时,MobileNets[16,17]计算量依旧过大。

本文提出了一种基于目标检测人群计数算法,选取人体头肩区域作为检测目标,针对目标检测计算量过大的问题,提出了一种速度快、精度高的轻量级弱计算力平台加速网络(low-computing-power platforms acceleration network,LPANet),并基于此对文献[13]中检测算法进行改进。LPANet 主要包含 2 个模块:特征图快速缩减模块和多尺度感受野扩增模块。特征图快速缩减模块可以使检测器在功耗较低的 ARM 平台上获得实时的速度;多尺度感受野扩增模块可以扩大每个先验框关联的感受野,利用多尺度感受野的特征减少误检。与此同时,在检测器设计上,本文优化了先验框与目标的匹配策略,解决了匹配过程中小目标匹配到的先验框过少导致检出率过低的问题,以及密集人群中由于目标拥挤造成的误检。得益于特征图快速下降模块、多尺度感受野扩增模块和优化的匹配策略,检测器不仅可以在头肩检测数据集上获得高精度,同时可以在 ARM 平台上获得实时的运行速度。在检测之后,头肩检测器的结果偶尔会有一些误检,对此,本文采用多目标跟踪算法对多帧结果进行轨迹关联。通过对轨迹长度的分析,可以消除视频帧头肩检测中的部分误检。同时本文建立了一个较为完备的头肩检测数据集,在该数据集上的实验结果证明了本文提出的人群计数方法的有效性。

1　人群计数算法

本文提出的人群计数算法主要基于头肩检测结

合多目标跟踪的方式进行。算法框架如图 1 所示。在自主设计的轻量级网络基础上,采用单阶段目标检测算法对头肩目标进行提取。在头肩检测后,采用多目标匹配跟踪算法获得每个目标的跟踪轨迹,抑制误检,最终由跟踪轨迹的数量确定视频帧中的人数。

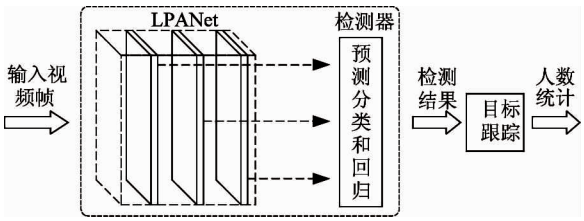

图 1　人群计数算法流程

1.1　头肩检测

作为人群计数算法的核心组件,头肩检测算法的检测精度和效率决定了整体算法的实用性。本文针对 ARM 嵌入式平台运算能力低下的特点,首先设计了一种轻量化的神经网络(LPANet),其后基于该网络设计了一种改进的单阶段检测算法,快速提取和定位头肩目标。LPANet 网络结构如图 2 所示,其中 Conv 表示卷积,Conv 下标注依次表示卷积核宽、高和通道数,s 表示卷积步长,d 表示膨胀卷积中膨胀率,该网络由特征图快速缩减模块和多尺度感受野扩增模块 2 部分组成。特征图快速缩减模块设计思路主要在于利用快速缩减特征图空间尺寸的方式提升网络运行速度;多尺度感受野扩增模块则对目标相关联感受野进行扩增,并结合多尺度感受野的形式为头肩目标特征提供丰富的上下文语义信息。受文献[13]中单阶段目标检测算法的启发,本文中头肩检测算法采用分 3 层对各尺度目标进行提取,此外,由于文献[13]算法中先验框和图像中目标的匹配策略设计没有考虑到小目标、互相遮挡的目标,导致小目标的先验框匹配率低,互相遮挡目标匹配到的先验框中包含大量其他目标信息,从而产生小目标的漏检和互相遮挡目标间的误检。有鉴于此,本文提出更为均衡的先验框设计和分配策略。下面将对各模块进行详细介绍。

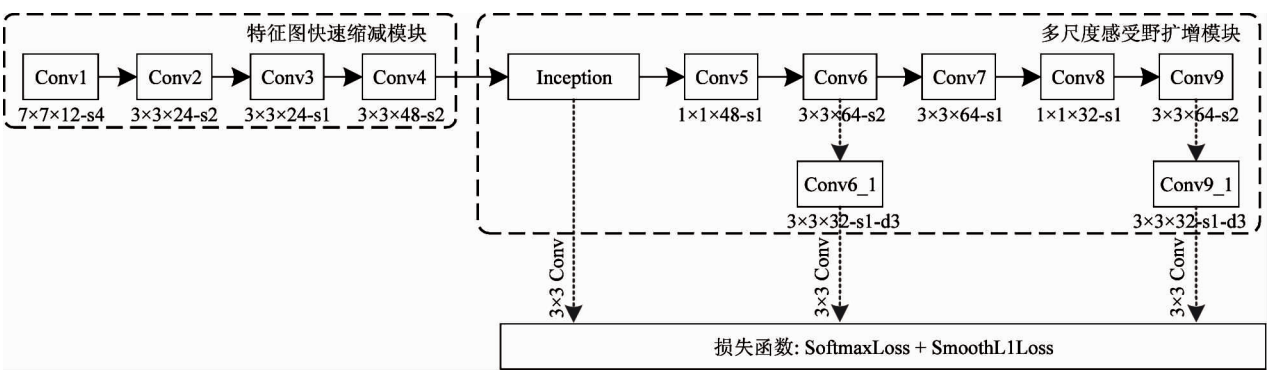

图 2　基于 LPANet 的轻量化头肩检测网络结构示意

(1)特征图快速缩减模块。特征图快速缩减模块的结构如图 2 所示,首先,不同于文献[6]和文献[16,17]第 1 层采用 3 ×3 卷积核尺寸、步长为 1 或 2 的卷积,网络采用 7 ×7 卷积核尺寸、步长为 4 的卷积快速下降输入尺寸,从而大幅度缩小后续卷积层处理的特征图大小,进而减小计算量。同时,7 ×7 的卷积核参数量和感受野相对较大,提取到的特征更为丰富,从而可以减少特征图尺寸快速下降带来的特征信息损失。

在第 1 层卷积之后,网络采用 3 ×3 卷积核尺寸、步长为 2 的卷积层 Conv2 进一步缩小特征图尺寸。然而在特征图尺寸快速下降的过程中,目标的细节特征信息也会有一定程度的损失。因此,为了减少特征信息过度损失,网络在 Conv2 之后采用 3 ×3 卷积核尺寸、步长为 1 的卷积,一方面减小特征信息的过速流失,另一方面加深网络深度使网络提取到更精确的深度特征。最终,在 Conv4 卷积层,网络将特征图尺寸快速缩小为输入的 1/16。特征图

快速缩减模块在大幅提升速度的同时减轻了特征信息损失带来的精度下降问题。不仅可以加快模型的运行速度,还使模型保持较高的精度。此外,网络分别将 Conv1、Conv2、Conv3 和 Conv4 的卷积核数量设置为 12、24、24 和 48,降低了参数冗余,进一步提升了运行效率。

(2)多尺度感受野扩增模块。相较于人脸检测和行人检测,头肩目标细节特征较少,多数情况下头肩检测更依赖于头肩的轮廓特征。当视频帧中某些区域的轮廓与头肩轮廓相似度较高时,检测器很可能会将这些区域检测为头肩,造成误检,影响人群计数算法的精度。同时,特征图快速缩减模块网络层数较少,每个先验框所对应的特征表达力不足,且相关联的感受野较小,这些因素限制了检测精度的提升。不同于 Zhao 等人[18]在主干网络之后加入不同膨胀率的膨胀卷积来丰富特征图的上下文信息,本文提出多尺度感受野扩增模块为头肩数据分布设计了更为适配的膨胀率并且大幅缩减了由于多分支膨胀卷积层带来的速度损失。多尺度感受野扩增模块如图 2 虚线框中所示,其中的 Inception 模块如图 3 所示。Inception 模块采用 3×3 卷积核尺寸,膨胀率为 3 的膨胀卷积增大感受野的同时获取到多尺度的感受野,使得检测器可以利用更多头肩周围的行人特征区分正负样本,减少与头肩轮廓相似物体的误检。此外,多尺度感受野扩增模块在预测目标类别和坐标回归之前增加 3×3 卷积核尺寸,膨胀率为 3 的卷积进一步扩大感受野。在感受野尺度扩增和多尺度生成之后,每个先验框尺寸和所关联的感受野尺寸如表 1 所示。

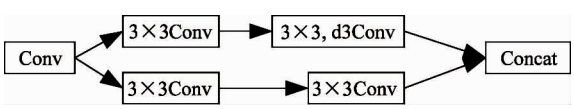

图 3　Inception 结构

表 1　先验框和感受野尺寸

先验框关联层	先验框尺寸(像素)	感受野尺寸(像素)
Inception	16×16,32×32	79×79,143×143
Conv6 _ 1	64×64	303×303,367×367
Conv9 _ 1	128×128,256×256	623×623,687×687

(3)先验框设计与分配策略。先验框尺寸设计如表 1 中所示。由于在头肩检测中,头肩目标宽高比趋近于 1:1,因此,为了先验框能更高效地回归目标以及节约计算量,本文仅设计宽高比为 1:1 的先验框。如图 2 中所示,检测算法采用在 3 个不同分辨率的卷积层分别预测的方式对头肩目标进行预测。在 Inception 后的第 1 个预测层先验框设计中,采用 16 像素和 32 像素的先验框尺寸;在 Conv6 _ 1 后的第 2 层预测层先验框设计中,采用 64 像素的先验框尺寸;在 Conv9 _ 1 后的第 3 层预测层先验框设计中,采用 128 和 256 像素的先验框尺寸。分层预测结合多尺度先验框设计可以有效地提升检测器的鲁棒性。

(4)先验框与目标匹配策略。为了保证检测算法执行速度,算法选择在特征图尺寸缩小为原图 1/16大小的卷积层上预测小目标,这意味着该预测层的先验框在原图分布的步长为 16 个像素。对于该层预测小目标的先验框来说,其步长和尺度之间的比例过大。此时如果按照文献[14]中的先验框与目标交并比阈值为 0.5 的匹配规则,会导致小目标的先验框匹配数量过低,使得小目标检出率大幅下降。为了提升小目标的检出率,简单直接的方法就是降低匹配过程中的阈值,但是当阈值降低以后,目标匹配到的先验框质量大幅下降,在这种情况下,很大一部分匹配到目标的先验框会包含较多背景特征或者其他目标的特征,导致判别力减弱。同时,当存在 2 个及以上目标距离较近或者互相遮挡情况时,目标匹配到的先验框不但会与当前目标有较高的交并比,还会与相邻的目标有较高的交并比,在这种情况下,先验框中包含了太多其他目标的特征,导致在测试过程中相邻或遮挡目标间产生误检。为了解决以上问题,本文设计的分配策略步骤如下:1)在目标与先验框匹配的过程中,若先验框与目标交并比大于 0.5,并且与该目标的交并比大于与其他目标交并比,则将该先验框分配给此目标。同时,若该先验框与其他目标交并比大于 0.3,则定义此先验框为不合格先验框,最终选取所有已匹配先验框为正样本,选取所有未匹配先验框和不合格先验框为负样本。2)将步骤 1)中匹配先验框数量小于 1)

中平均数量的目标匹配阈值降为 0.35,并在匹配过程之后,将与这些目标匹配到的先验框按照交并比大小进行排序,选取得分前 k 数量的先验框作为该目标的最终匹配到的正样本,k 的值选取为步骤 1)中每个目标所匹配先验框的平均数量,从而既保证每个目标匹配到的先验框数量,又保证其质量。通过本文提出的先验框匹配策略,小目标的检出率有明显的提升,同时由于目标之间距离过近导致的误检明显下降。

1.2　目标跟踪和计数

在头肩检测之后,当前帧每个头肩目标都会对应一个检测结果框,然而,受限于检测器的检测精度以及训练数据的丰富度,检测结果中时常会出现漏检和误检。因此,为了提高人群计数算法的精度以及鲁棒性,本文在检测基础上加入多目标跟踪算法修正检测结果并获取每个头肩目标在连续视频帧中的跟踪轨迹。如图 4 所示,本文采用由 Bewley 等人[19]提出的简单实时在线跟踪(simple online and realtime tracking,SORT)算法对头肩目标检测结果进行跟踪,当同一目标在连续 3 帧中被检测到时,则开始跟踪该目标,并且如果在最后一次检测后连续 10 帧内没有检测到该目标则结束此次跟踪。最终,根据跟踪轨迹的数量确定视频帧中的人数。相较于仅采用基于单帧图像目标检测的人群计数策略,跟踪结合检测的人群计数策略进一步提升了人群计数的精度和鲁棒性。

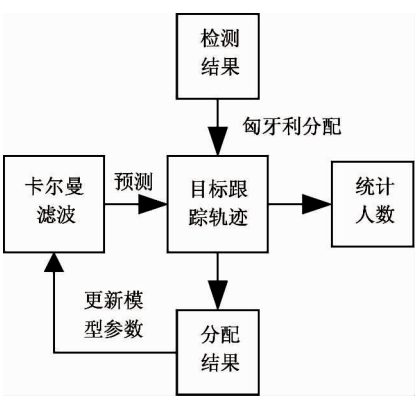

图 4　头肩目标跟踪与人数统计流程

1.3　算法实现细节

人群计数算法实现主要包含 2 个部分:头肩检测部分和跟踪部分。其中头肩检测算法包含了数据扩增、困难样本挖掘、训练参数设置以及测试参数设置。

(1)对头肩数据进行扩增。将原始图片以边长 1.5 倍进行边缘填充,然后在图片中随机截取[0.3,1]的正方形区域,并将其缩放到 640×640 分辨率。缩放过程中,去除像素值少于 16 个像素的目标。

(2)困难样本挖掘。训练过程中大部分的先验框为负样本,导致负样本数量与正样本数量严重的不均衡,从而使训练过程不稳定以及收敛速度降低。因此,本文将训练过程中的所有负样本根据其损失值进行排序,并挑选损失最大的前 n(n 设置为正样本数量的 3 倍)个负样本参与训练。

(3)训练参数设置。在训练过程中,将原始图片转为灰度图,在对灰度图数据扩增以后,首先采用热身学习策略[20]以线性提升学习率的方式迭代 4 000次,可以有效避免学习率过高导致梯度爆炸的问题,然后采用随机梯度下降策略、0.9 动量、0.0005梯度衰减和 32 批量训练。采用分段学习策略,以 0.01 学习率迭代 26 000 次,再以 0.001 学习率迭代10 000 次,然后以 0.0001 学习率迭代 8 000 次,最后以 0.00001 学习率迭代2 000 次。本文实验均在 Caffe 深度学习框架上进行。

(4)测试参数设置。在测试过程中,均采用视频图形阵列(video graphics array,VGA)分辨率(640×480)在瑞芯微 RK3399 的 2 个大核(cortex-A72 1.8 GHz)上测试,为了充分发挥 RK3399 的计算力,本文借助开源加速库 ncnn[21]将算法进行移植。

2　实　验

2.1　头肩数据集

参考 Zhang 等人[22-26]的方法,本文收集了包含服装店、电脑城和汽车 4S 店等超过 10 个场景共计 23 963 张图片建立头肩数据集,该数据集涵盖了常规摄像头架设场景下可能出现的头肩在画面中呈现的视角。本文将所有图片随机划分为训练集和测试集,测试集包含 1 639 张图片,其余作为训练集。测试集和训练集头肩目标尺度分布分别如图 5 和图 6

所示,其中横坐标表示像素值,纵坐标表示数量,此外,本文收集了一段和训练数据集完全无交叉的测试视频进行算法泛化性测试。视频信息如表 2 所示。本文分别对检测部分(包含特征图快速缩减模块、多尺度感受野扩增模块和先验框分配策略)和跟踪部分进行探究性实验。实验中,检测算法精度采用 11 点插补法平均精度(average precision, AP)衡量,*AP* 计算公式如式(1)、(2)、(3)、(4)所示。

$$p = \frac{TP}{TP + FP} \tag{1}$$

$$r = \frac{TP}{TP + FN} \tag{2}$$

$$p_{interp}(r) = \max_{\tilde{r}:\tilde{r}\geq r} p(\tilde{r}) \tag{3}$$

$$AP = \frac{1}{11} \sum_{r\in\{0,0.1,\cdots,1\}} p_{interp}(r) \tag{4}$$

式中,*TP* 为检测结果中正确的个数,*FP* 为错误的个数,*FN* 为未检测出的目标个数,*p* 为准确率,*r* 为检出率,$p(\tilde{r})$ 表示在检出率为 $\tilde{r}$ 时的准确率。速度采用帧率(frames per second,FPS)衡量,人群计数整体算法采用平均绝对误差(mean absolute error, MAE)和均方误差(mean square error, MSE)衡量,*MAE* 和 *MSE* 如式(5)和(6)所示。

$$MAE = E(|k_j - k'_j|) \tag{5}$$

$$MSE = E((|k_j - k'_j|)^2) \tag{6}$$

式中,k_j 和 k'_j 分别表示第 j 帧中头肩目标的实际数量和预测数量。

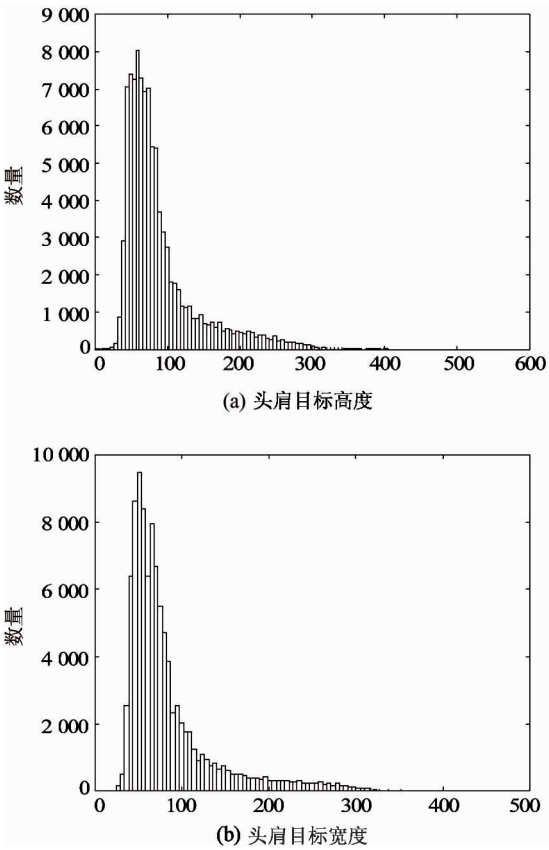

图 5　测试集头肩分布

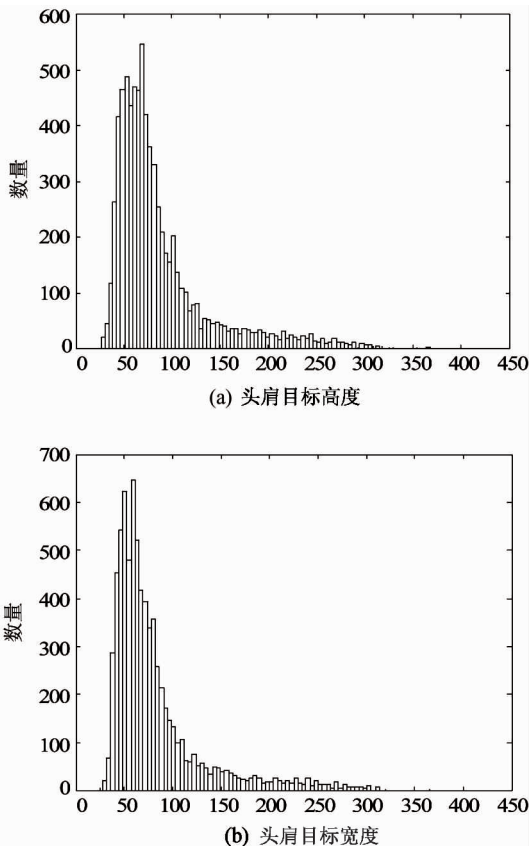

图 6　训练集头肩分布

表 2　视频子数据集信息

总帧数	分辨率	帧率	每帧目标数量	目标总数量
500	640 × 480	25	8 ~ 22	8 521

2.2　实验结果分析

本文对人群计数算法检测部分、跟踪部分以及检测部分各个模块分别进行实验探究,下面将对实验结果进行详细介绍。

(1)对特征图快速缩减模块、多尺度感受野扩

增模块和先验框匹配策略进行探究。为了验证特征图快速缩减模块的有效性,本节将特征图快速缩减模块替换为 5 个通道数与特征图快速缩减模块相同的卷积层,并将 5 个卷积层的卷积核尺寸和步长值分别设为 3 ×3 和 2 进行对比实验。为了验证多尺度感受野扩增模块的有效性,本文去掉 Inception 模块包含有膨胀卷积的分支,并将另一分支通道数加倍进行对比实验。测试结果如表 3 所示。表 3 中 M1 为特征图快速缩减模块,M2 为多尺度感受野扩增模块,M3 为先验框匹配策略。由表 3 数据可知,特征图快速缩减模块提升 20% 检测速度的同时提升了 2.6% 的精度,使得模型可以在 RK3399 上获得实时的检测速度和较高的精度。多尺度感受野扩增模块有效提升了每个先验框关联的感受野尺寸,并抑制了与头肩目标特征相似的误检,并且膨胀卷积在增大感受野的同时并没有带来计算量的增大,最终使得模型获得 0.9% 的精度提升。先验框匹配策略是必不可少的,它大幅地提升了数据集中小目标的检出率,并且抑制了拥挤场景下相邻目标间出现的误检,最终使得模型获得了 6.7% 的精度提升。

表 3　LPANet 各模块性能测试

模块	精度	速度(fps)
M1 + M2 + M3	83.8%	24
M2 + M3	81.2%	20
M1 + M3	82.9%	24
M1 + M2	77.1%	24

(2)对人群计数算法整体检测部分在图片测试集上进行探究。为了对比头肩检测器在速度和精度上的优势,本文在头肩数据集上复现了 2 个计算量与头肩检测器相似的检测算法,其一是在人脸检测方面表现出众的 Faceboxes[27] 检测器,其二是将文献[16]中网络缩减 3/4 通道数,并将其融合到文献[13]的检测器中(记为 SSD + 0.25MobileNets)。2 个检测器训练过程与头肩检测器保持一致,实验对比结果如表 4 所示。表 4 中 M1、M2 与 M3 含义和表 3 中一致。从表 4 中数据可知,头肩检测器在保证精度的同时获得了大幅度检测速度提升。部分头肩检测效果如图 7 所示。

表 4　头肩检测器性能对比测试

检测器	精度	速度(fps)
SSD + 0.25MobileNet	84.5%	7.7
FaceBoxes[27]	71.0%	20.0
SSD + M1 + M2 + M3	83.8%	24.0

图 7　头肩检测效果图

(3)对人群计数算法跟踪部分在视频子测试集上进行探究。探究实验结果如表 5 所示。表 5 中 M1、M2 与 M3 含义和表 3 中一致,Tracker 为本文所采用的多目标跟踪算法。由表 5 中数据可知,本文提出的人群计数算法在 ARM 平台 RK3399 上可以获得实时的执行速度并且在与训练集场景差异很大

表 5　跟踪模块测试

方法	*MAE*	*MSE*	速度(fps)
SSD + M1 + M2 + M3	1.15	1.32	24
SSD + M1 + M2 + M3 + Tracker	1.06	1.13	20

的视频子数据集上有较高的精度。加入跟踪模块在提升算法精度和鲁棒性的同时,算法执行速度并没有受到很大的影响。

3　结　论

本文提出了一种面向嵌入式平台的基于深度学习的快速人群计数方法。算法采用头肩目标检测结合多目标跟踪的方式来进行人群计数。在算法检测部分,本文针对头肩检测算法计算量大、小目标多、目标拥挤等问题,设计了轻量网络 LPANet 作为算法骨干网络,并改进先验框设计和分配策略,在大幅提升检测速度的同时,保证了目标检测的精度。同时本文引入多目标跟踪模块来进行轨迹关联和误检抑制。本文提出的方法能够在 ARM 嵌入式平台上对连续视频帧中的人群进行实时准确计数统计。在包含复杂场景的头肩人群计数数据集上的对比实验表明了本文方法的有效性。

参考文献

[1] Davies A C, Yin J H, Velastin S A. Crowd monitoring using image processing[J]. *Journal of Electronics & Communication Engineering*, 1995, 7(1):37-47

[2] Zhang Y Y, Zhou D S, Chen S Q, et al. Single-image crowd counting via multi-column convolutional neural network[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 589-597

[3] Dalal N, Triggs B, et al. Histograms of oriented gradients for human detection[C]//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005: 886-893

[4] Dollar P, Appel R, Belongie S, et al. Fast feature pyramids for object detection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(8): 1532-1545

[5] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2015, 39(6):1137-1149

[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. *arXiv*: 1409.1556V6, 2015

[7] Mao J Y, Xiao T T, Jiang Y N, et al. What can help pedestrian detection? [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu,USA, 2017: 6034-6043

[8] Zhang S S, Yang J, Schiele, B. Occluded pedestrian detection through guided attention in CNNs[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6995-7003

[9] Wang X L, Xiao T T, Jiang Y N, et al. Repulsion loss: detecting pedestrians in a crowd[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7774-7784

[10] Li Z Q, Zhang L, Fang Y K, et al. Deep people counting with faster R-CNN and correlation tracking[C]//Proceedings of the 6th International Conference on Internet Multimedia Computing and Service, New York, USA, 2016: 57-60

[11] Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 37(3):583-596

[12] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of IEEE Conferenceon Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 936-944

[13] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//Proceedings of 14th European Conference on Computer Vision, Amsterdam, Netherlands, 2016: 21-37

[14] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//Proceedings of IEEE Conferenceon Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 779-788

[15] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6517-6525

[16] Howard A G, Zhu M L, Chen B, et al. Mobile nets: efficient convolutional neural networks for mobile vision application[J]. *arXiv*1704.04861V1, 2017

[17] Sandler M, Howard A, Zhu M L, et al. MobilenetV2:

inverted residuals and linear bottlenecks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 4510-4520

[18] Zhao X, Zhao C Y, Zhu Y S, et al. Improved single shot object detector using enhanced features and predicting heads[C]// Proceedings of 4th International Conference on Multimedia Big Data, Xi'an, China, 2018: 1-5

[19] Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking[C]// Proceedings of 23th IEEE International Conference on Image Processing, Phoenix, USA, 2016: 3464-3468

[20] Goyal P, Dollár P, Girshick R, et al. Accurate, large minibatch SGD: training imagenet in 1 hour[J]. *arXiv*1706.02677V2, 2017

[21] Ni H. ncnn[EB/OL]. https://github.com/Tencent/ncnn, 2018

[22] Zhang X Y, Wang S P, Yun X C. Bidirectional active learning: a two-way exploration into unlabeled and labeled data set[J]. *IEEE Transactions on Neural Networks & Learning Systems*, 2017, 26(12):3034-3044

[23] Zhang X Y, Shi H C, Li C S, et al. Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision[C]// Proceedings of the 33th Association for the Advancement of Artificial Intelligence Conference, Honolulu, USA, 2019: 1-8

[24] Zhang X Y. Interactive patent classification based on multi-classifier fusion and active learning[J]. *Neurocomputing*, 2014, 127:200-205

[25] Zhang X Y, Wang S P, Zhu X B, et al. Update vs. upgrade: modeling with indeterminate multi-class active learning[J]. *Neurocomputing*, 2015, 162:163-170

[26] Zhang X Y, Shi H C, Zhu X B, et al. Active semi-supervised learning based on self-expressive correlation with generative adversarial networks[J]. *Neurocomputing*, 2019, 345:103-113

[27] Zhang S F, Zhu X Y, Lei Z, et al. Faceboxes: a CPU real-time face detector with high accuracy[C]// Proceedings of the 3th IEEE International Joint Conference on Biometrics, Denver, USA, 2017: 1-9

Real-time crowd counting for embedded systems with high accuracy

Jin Xin*, Zhao Xu**, Zhao Chaoyang**, Xu Huazhong*, Wang Jinqiao**

(*College of Automation, Wuhan University of Technology, Wuhan 430070)

(**Institute of Automation, Chinese Academy of Sciences, Beijing 100190)

Abstract

On some embedded devices that have limited computing resources, most of the methods which adopt deep learning are difficult to meet the requirements of practical application in running speed and the precision. To solve this problem, this paper proposes a deep learning based fast crowd counting method for embedded devices. Firstly, this paper designs a low-computing-power platforms acceleration network (LPANet) and embeds it into a single-stage head-shoulder detector. The detector quickly detects the head-shoulder part of people in each frame of the videos. Then, this paper adopts a fast multi-target tracking method to obtain the trajectory of each detected header-shoulder region. Finally, the method gets the number of people according to the number of trajectories. This paper proposes a head-shoulder dataset to evaluate the effectiveness of the proposed method. The method gets a 1.15 mean absolute error(MAE) and runs 20 frames per second on ARM platform (dual-core cortex-A72) at 640×480 resolution. The deep learning based crowd counting method consisting of head-shoulder detection and tracking provides a feasible solution for the use of embedded devices in various scenes.

Key words: crowd counting, head-shoulder detection, multi-target tracking, low-computing-power platforms acceleration network (LPANet), head-shoulder dataset