文章编号：0253-374X(2019)08-1216-09 DOI:10.11908/j.issn.0253-374x.2019.08.020

基于优化的Inception-ResNet-A模块与Gradient Boosting的人群计数方法

郭瑞琴[1]，陈雄杰[1]，骆 炜[2]，符长虹[1]

(1.同济大学 机械与能源工程学院，上海 201804；2.斯图加特大学 工程与计算力学研究所，斯图加特 70569)

摘要：针对人群计数问题，基于优化Inception-ResNet-A模块，使用集成学习中的Gradient Boosting方法提出了一种可用于稀疏人群和密集人群的人群计数方法，并给出此方法实现的具体细节.通过在三个公开数据集和真实场景(含光照和视角变化)中进行测试，检验了该方法对于光照、人群密度、视角等变化的鲁棒性.实验结果表明，该方法对于以上变化具有较强的鲁棒性，并且相比于之前的人群计数方法在准确性和稳定性方面具有更好的性能.

关键词：人群计数；优化Inception-ResNet-A模块；Gradient Boosting；多尺度特征；感知野

中图分类号：TP181 **文献标志码**：A

A Method of Crowd Counting Based on Improved Inception-ResNet-A Module with Gradient Boosting

GUO Ruiqin[1], *CHEN Xiongjie*[1], *LUO Wei*[2], *FU Changhong*[1]

(1. School of Mechanical Engineering, Tongji University, Shanghai 201804, China; 2. Institute of Engineering and Computational Mechanics, University of Stuttgart, Stuttgart 70569, Germany)

Abstract: To count the pedestrians in the scenarios with the sparse or dense crowd, a network based on the improved Inception-ResNet-A module is proposed, which is trained with the gradient boosting method of ensemble learning, and the details of the proposed method are given. Besides, a dataset collected in a real scenario, which contains illumination and camera view changes, and other three public datasets are used to evaluate the robustness of the proposed method in terms of illumination, population density, and camera view changes. The experimental results show that the proposed method is robust to the aforementioned changes. In addition, the proposed method favorably outperforms the state-of-the-art approaches in terms of accuracy and stability.

Key words: crowd counting; improved Inception-ResNet-A module; gradient boosting; multi-scale features; receptive field

随着科学技术的快速发展，交通工具更加便利，城市化进程不断加快，城市流动人口的数量快速增长，城市繁华街道越来越拥挤，各种大型展览会的参展人员也越来越多.为了保证城市交通通畅，合理控制人群密集场合的人员数量，保证人民群众生命安全，有必要对行人行为和分布规律进行研究[1-11].人群计数技术作为该领域的重要组成部分之一，近年来受到众多国内外研究机构的关注[1-7].人群计数主要有以下难点：第一，在一张图片中，行人的尺度变化；第二，不同场景下的行人分布变化；第三，相同场景下不同时间的行人分布变化.由于不同场景图片中行人数量和大小差别较大，因此要求计数方法对不同环境场景中行人尺寸的多样性具有很强的鲁棒性.为解决这个难题，学者们提出了各种不同的人群计数方法来保证神经网络适应这种尺度的变化[1-6].

在早期的人群计数研究中，大多数计数方法是以目标检测为基础，其检测方法主要分为两类：一类是基于人工设计特征的目标检测[11]；另一类是基于深度神经网络提取特征的目标检测[12-13].目标检测方法首先通过训练得到能够定位目标的检测器，然后使用该目标检测器在图片中找到指定目标，并将检测得到的目标数量作为最后的计数结果.该方法能够比较有效地检测出目标，并对单个目标精确定

收稿日期：2018-11-12
基金项目：中央高校基本科研业务费专项资金(22120180009)
第一作者：郭瑞琴(1962—)，女，副教授，硕士生导师，工学博士，主要研究方向为机器人理论与应用研究、机构及传动系统设计、产品设计与开发.E-mail:07172@tongji.edu.cn
通信作者：符长虹(1986—)，男，助理教授，硕士生导师，工学博士，主要研究方向为基于计算机视觉的无人机目标跟踪、即时定位与地图构建以及智能控制.E-mail:changhongfu@tongji.edu.cn

位，但是对于有行人相互遮挡和尺度变化比较大，且人群密度较高的场景图片，其识别准确率较低.

为了解决目标检测方法存在的问题，Chan 等[14]提出了基于回归的计数方法，该方法直接给出人群的目标数量，不需要对每个目标进行检测. 虽然基于回归的方法相对于基于目标检测的方法准确率有所提升，但此种方法没有考虑人群的空间分布，无法理解深层的场景信息，从而影响整体计数结果的准确性. 为了将图片中人群的空间位置信息加以利用以提高人群计数的准确率，Oñoro-Rubio 等[15]提出了基于密度图的计数方法. 该方法通过对人群密度图的像素分析，并对密度图按像素求和，将求和得到的数字作为计数的结果. 其中，多列结构的卷积神经网络是具有代表性的密度计数方法，该方法的基本思想是在不同的列中使用不同大小的卷积，使得每一列网络具有不同大小的感知野，通过提取每列网络的特征，并将这些特征进行融合，解决人群计数中的行人尺度变化问题[1-2]. 该方法的不足之处在于，当多列结构训练参数较多时，训练网络变得困难. 其次，每一列的网络都表现出相似的人群密度变化特性，使得各列之间的差异并不显著，这与设计多列结构的初衷相违背[6].

最近，Zhang 等[3]提出一种端到端的尺度自适应网络结构用于生成密度图，并在多个数据集上证明其有效性，展示出其在准确率和鲁棒性上的优越性. 值得注意的是，该网络在其后端使用了最大池化层和反卷积层，尽管这种结构能增大其感知野，但也会导致细粒度信息的丢失. Li 等[6]随后提出一种使用空洞卷积作为其后端的神经网络，可有效地增大网络的感知野，同时也避免了使用最大池化层带来的细粒度信息丢失等问题. 然而，该网络结构并不包含多尺度信息. 本文实例的实验结果证明，多尺度信息可以改进网络人群计数的准确率.

为了提高人群计数方法的准确率，本文使用优化的 Inception-ResNet-A 模块[16]并结合 Gradient Boosting 集成学习方法提出一种端到端的卷积神经网络结构，即 Gradient Boosting Multi-Scale Counting Net(GB-MSCNet)，该结构能够在有效地增大网络输出层感知野的同时保存图片中的细粒度信息. 而且该网络能够将多尺度特征融合，保证该网络对于行人尺度变化以及人群密度变化的鲁棒性，使生成的密度图更精确，从而准确地给出图片中的行人数目.

1　人群真实密度图的获取

在基于密度图的计数方法中，监督学习的目标就是学习从原图片到相对应真实密度图的映射，因此，真实密度图能否准确反应原图中行人的空间分布，对最后的计数准确率有很大的影响.

对于人群计数数据集，数据集中通常会给出行人在图片中的位置参数，即计数目标在图片中几何中心的位置(以像素点坐标的形式给出). 根据数据集中的位置参数，可得到初始密度图 $\boldsymbol{M}$

$$\boldsymbol{M}=\sum_{i=1}^{J}\boldsymbol{\delta}(P_i) \tag{1}$$

式中：P_i 为第 i 个行人目标在图片中的位置；J 为行人的数量；$\boldsymbol{M}$ 为与原图片尺寸相同但通道数为 1 的 $n\times m$ 矩阵；$\boldsymbol{\delta}(P_i)$表示在 P_i 处的值为 1，而其他位置值为 0 的与 $\boldsymbol{M}$ 相同尺寸、相同通道数的矩阵. 在 $\boldsymbol{M}$ 的基础上，用高斯核 G_{σ_i} 对 $\boldsymbol{M}$ 进行滤波操作，生成真实密度图 $\boldsymbol{F}$

$$\boldsymbol{F}=\boldsymbol{M}*G_{\sigma_i}(x) \tag{2}$$

高斯核函数 $G_{\sigma_i}(x)$定义如下：

$$G_{\sigma_i}(x)=\frac{1}{2\pi\sigma_i^2}\mathrm{e}^{-\frac{(x-P_i)^2}{2\sigma_i^2}} \tag{3}$$

式中：σ_i 为二维高斯分布的标准差；$G_{\sigma_i}(x)$代表矩阵 $\boldsymbol{F}$ 中 x 处关于 P_i 的高斯核的取值，最后 $\boldsymbol{F}$ 在 x 处的取值为所有 P_i 在 x 处的高斯核函数取值(与 σ_i 相关)之和.

根据高斯核函数的性质可知，σ_i 越大，第 i 个行人在生成的真实密度图中对应的区域越大，因此 σ_i 应与第 i 个行人在图片中的大小相关，为达成这一目的，Zhang 等[2]提出了一种用于密集人群的真实密度图的生成方法，将 σ_i 与行人之间的平均距离相关联，即对每个行人，以距离该行人最近的 k 个行人与该行人之间的平均距离来代表 σ_i(k 是超参数，可通过实验选择效果最好的 k). 使用该方法在图 1a 所示的 ShanghaiTech 数据集中生成的真实密度图如图 1b 所示，对图片中的所有行人使用同样大小的 σ_i 生成的真实密度图如图 1c 所示. 显然图 1b 中的密度分布随着行人在图片中的大小而变化，相比图 1c 更接近图 1a 中行人的密度分布.

然而 Zhang 等[2]提出的方法不适用于描述相对稀疏人群的行人目标大小. 如果行人之间的平均距离与行人在图片中的大小相关度不高，就无法使用平均距离代表行人在图中的大小. 为解决这一问题，

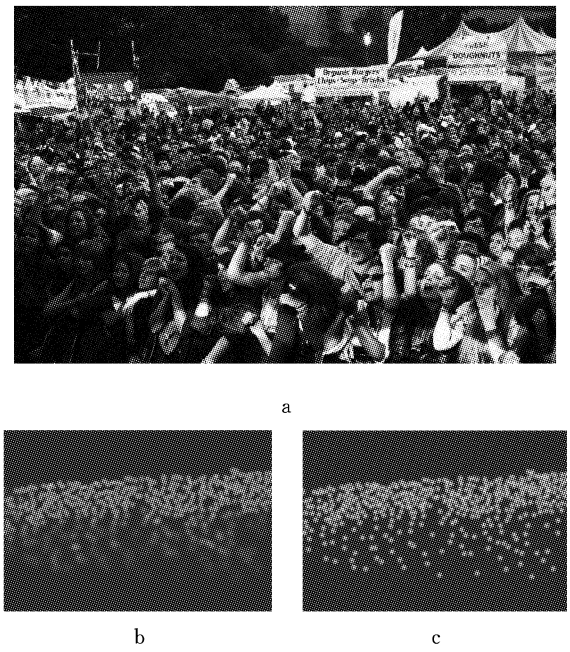


图1 ShanghaiTech数据集中的密集人群图片与用平均距离代表和使用固定不变的 σ_i 生成的真实密度图

Fig. 1 Crowd image with dense population in ShanghaiTech dataset and corresponding density map with average distance between pedestrians and fixed σ_i

Chen等[17]提出一种针对具有固定背景的人群计数数据集的真实密度图生成方法,使用线性拟合的方法模拟行人在图片中的位置与行人大小的关系,并通过数据集中的位置信息近似计算 σ_i 并进一步得到不同位置行人的大小,使获得的真实密度图能更精确地表示行人的空间分布. 如图2所示,使用该方法在UCSD数据集(该数据集中的图片背景固定不变)中生成的真实密度图相比于使用固定的 σ_i 更能代表行人的密度分布.

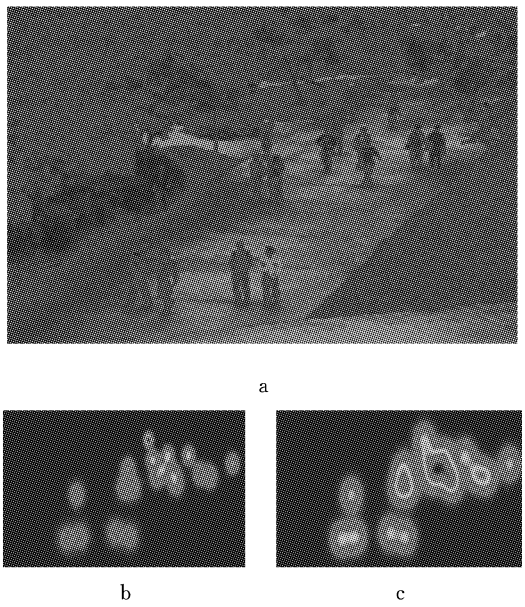


图2 UCSD数据集中的人群图片与使用线性拟合方法近似计算和使用固定不变的 σ_i 生成的真实密度图

Fig.2 Crowd image in UCSD dataset and corresponding density map with linear fitting and fixed σ_i

由于上述原因,在本文中,对于人群密度较大的ShanghaiTech数据集和UCF_CC_50数据集[11],使用Zhang等提出的基于行人平均距离的真实密度图生成方法;而对于人群相对稀疏且图片背景固定不变的UCSD数据集,则使用Chen等人提出的基于线性拟合的真实密度图生成方法.

2 GB-MSCNet计数网络

GB-MSCNet是基于Inception-Res-Net-A模块的端到端的全卷积网络结构,其具体结构如图3所示;GB-MSCNet中各模块详细结构如图4所示.

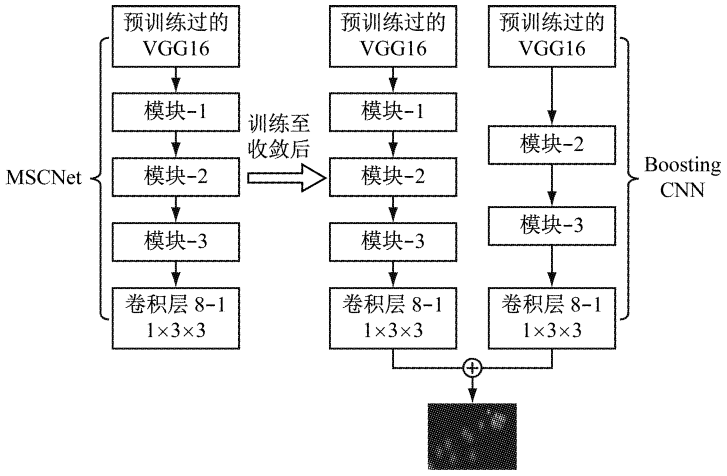


图3 GB-MSCNet网络结构

Fig.3 Architecture of GB-MSCNet

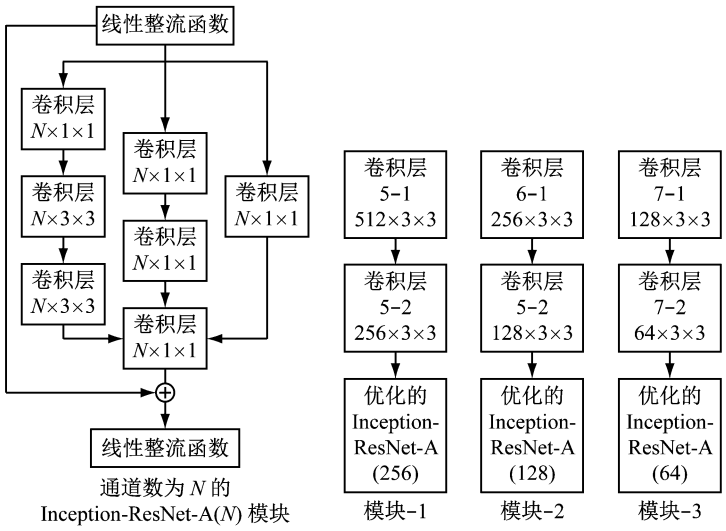


图 4　GB-MSCNet 中各组件的结构

Fig. 4　Architecture of componets of GB-MSCNet

该网络能在有效增大网络输出层感知野的同时避免丢失图片中的浅层特征，并将不同尺度的特征融合，保证了该网络对于行人尺度变化的鲁棒性，使生成的密度图更精确.

2.1　网络结构设计

根据之前的研究，在人群计数中，网络提取特征的能力、感知野大小、多尺度特征是影响准确率的主要因素[17]. 因此，GB-MSCNet 使用在 ImageNet 上预训练过的 VGG16[16]的前 10 层卷积层作为前端网络，以增强网络提取特征的能力；而在网络的后端，使用多个优化的 Inception-ResNet-A 模块，因为该模块能在增大网络的感知野的同时将不同尺度的特征融合.

为了适应 VGG16 的网络结构，在各个 Block 中使用优化后的 Inception-ResNet-A 模块，优化后的模块与原模块的主要区别在于通道数，实验证明优化后的模块能够提高人群计数的精度[17]. 相比于第 1 列的网络，第 2 列删去了 Block-1，这种结构设计的好处在于既能避免参数量过多(Block 1 中的参数占 MSCNet 参数量的 80%)，又能使用 Gradient Boosting 方法提高计数的准确率.

2.2　网络训练方法

在本文中，使用 Gradient Boosting 方法对 GB-MSCNet 进行训练. Gradient Boosting 是集成学习方法的一种，它主要的思想是每一次学习都是建立在前一次模型所犯的错误的基础上，以纠错的方式来提升整体模型的性能. 在训练 GB-MSCNet 时，首先单独训练第 1 列的卷积网络，至其收敛时将第 1 列的权重固定，再将第 2 列的卷积网络与第 1 列并联，使其学习第 1 列产生的密度图与真实密度图之间的残差.

在训练过程中，首先使用 Adam 优化器，学习率设置为 1×10^{-6}，待其收敛时使用 SGD 优化器，学习率设置为 1×10^{-8}. 同时在训练过程中，还使用了权重衰减(weight decay)、动量(momentum)等方法以加速收敛过程.

本文中使用真实密度图与人群计数网络生成的密度图之间每个像素的欧式距离之和来衡量两者之间的距离，并将这种距离作为训练中的优化函数，其具体定义如下：

$$L(\theta)=\frac{1}{2K}\sum_{i=1}^{K}\|D_i(I_i,\theta)-D_{i,\mathrm{gt}}\|^2 \tag{4}$$

式中：K 为人群图片总数；θ 为计数网络中的参数；I 为输入到网络中的人群图片；$D_i(I_i,\theta)$为网络生成的密度图；$D_{i,\mathrm{gt}}$代表真实密度图.

3　实验验证

3.1　计数性能评价指标

目前，评价人群计数网络的性能一般采用平均绝对误差(MAE)e_1 和均方误差(MSE)e_2 两个指标值，本文采用 MAE 和 MSE 指标对所构建的网络系统进行评价. MAE 和 MSE 的定义如下：

$$e_1=\frac{1}{M}\sum_{i=1}^{M}|C_i-C_{i,\mathrm{gt}}| \tag{5}$$

$$e_2=\sqrt{\frac{1}{M}\sum_{i=1}^{M}|C_i-C_{i,\mathrm{gt}}|^2} \tag{6}$$

式(5)、(6)中：M 为测试的图片数量；C_i 为网络给出

的第 i 张图片的计数结果;$C_{i,\mathrm{gt}}$ 为第 i 张图片的真实计数结果. 当网络给出的计算结果 $C_{i,\mathrm{gt}}$ 越接近准确计数结果 C_i 时,网络结构的性能越好,相对应的式(5)和式(6)的计算结果,MAE 和 MSE 值越小. 因此,MAE 和 MSE 越小,表示网络的计数准确率越高、鲁棒性越强.

3.2 GB-MSCNet 训练过程展示

图 5 是实验过程中 GB-MSCNet 的损失函数曲线以及各阶段对应的密度图以及测试得到的 MAE 数值,从图 5 中可以看出,随着损失函数函数值的减小,密度图越来越接近于真实密度图,各训练阶段生成的密度如图 6a～6e 所示. 从第 1 个轮数完全随机产生的密度图开始,训练到第 10 个轮数已经能较准确地感应出行人的位置,但仍然存在许多对与行人无关的背景的感应;接下来从 100 到 500 再到 1 000 轮数的过程就是逐渐将背景信息过滤的过程,观察图 6d 可以看出,第 500 个轮数时还存在一些不明显的背景干扰,但到 1 000 轮数时背景干扰基本消失. 同时,计数产生的误差值也随着训练逐渐减小,这充分说明使用公式(4)中的函数作为损失函数是较为合理的. 另外,使用该公式作为损失函数的另一个重要原因是该函数是凸函数,而这对于神经网络的损失函数来说是一个必不可少的属性,否则会极大地加大训练的难度.

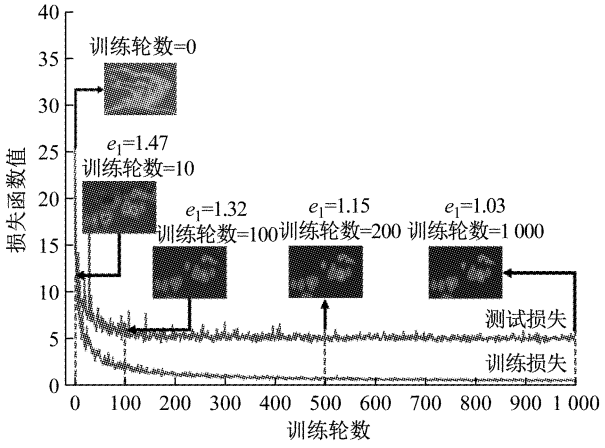


图 5 GB-MSCNet 训练集与验证集损失函数曲线

Fig.5 Training and testing loss of GB-MSCNet

图 7 中给出了测试时估计行人数量与真实行人数量随测试视频图片帧数变化的曲线. 从图 7 中可以看出,估计行人数量的曲线与真实行人数量的曲线非常接近,最大的计数误差不超过 8,说明 GB-MSCNet 能够较好地完成人群计数的任务.

在训练GB-MSCNet时,首先训练第1列,再用

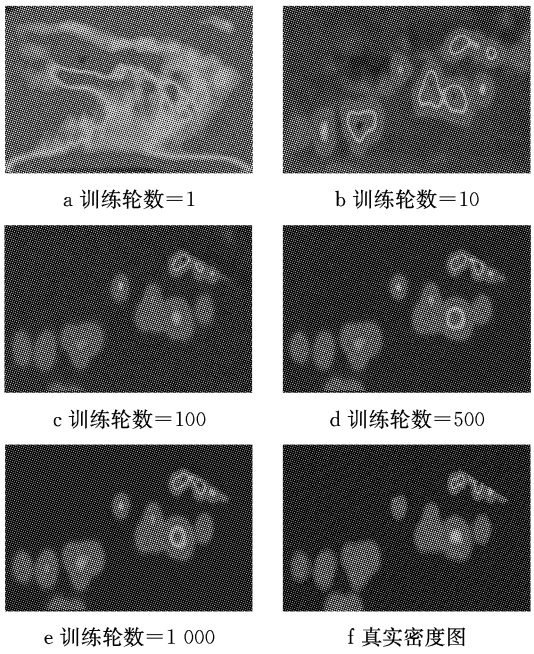


图 6 不同训练阶段生成的密度图的对比

Fig.6 Density maps generated in different stages

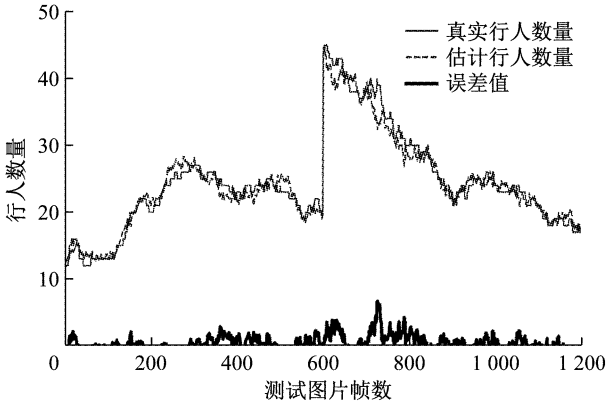


图 7 UCSD 数据集中估计行人数量与真实行人数量随测试图片帧数变化的曲线

Fig.7 Estimated and actual number of pedestrians in UCSD dataset

第 2 列学习第 1 列与真实密度图之间的残差,如图 8 所示,残差不均匀地分布在每一个行人处,主要存在于人群密集区. 加入第 2 列并训练后,模型的计数性能得到提升,图 9 所示是 MAE 和 MSE 在 UCSD 数据集中随测试视频图片帧数的变化曲线,相比于只使用第 1 列,加入第 2 列后模型的 MSE 和 MAE 均有下降. 从图 9 可以看出,MAE 和 MSE 的数值都比较小,并且图中两条曲线的变化趋势平缓,说明 GB-MSCNet 的计数准确性和稳定性都达到了一个较高的水平.

3.3 GB-MSCNet 计数性能实例验证

为了能够更客观地说明本文所构建的网络结构

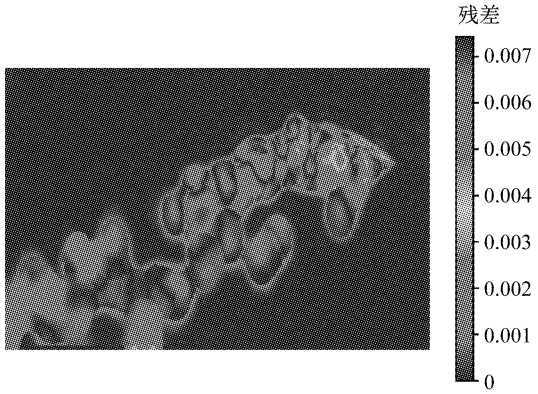


图 8　GB-MSCNet 中第 1 列产生的残差图

Fig. 8　Residual map of first column of the network

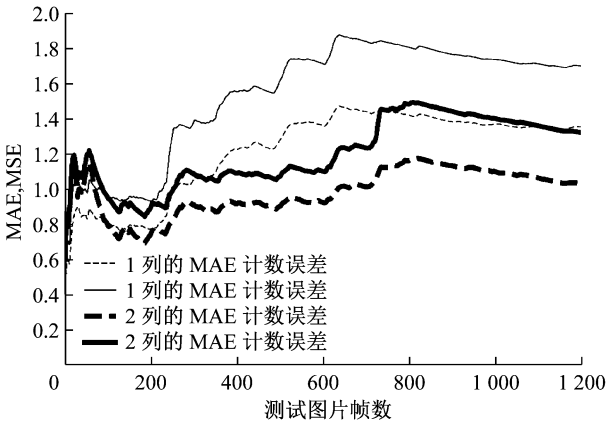


图 9　UCSD 数据集中评价指标 MAE 和 MSE 随测试图片帧数的变化曲线

Fig. 9　MAE and MSE in UCSD dataset

在人群计数问题中的有效性，本实例验证使用 UCSD 数据集作为衡量对于稀疏人群的计数性能的数据集，ShanghaiTech 和 UCF_CC_50 数据集作为衡量对于密集人群的计数性能的数据集，最后使用采集于真实场景的包含光照和视角变化的 TongjiCanteen 数据集测试本文的方法对于这些变化的鲁棒性，同时对本文构建的网络结构的性能进行评价.

3.3.1　稀疏人群实例验证

UCSD 是稀疏人群场景下获得的一组数据集，该数据集中包含 2 000 张分辨率为 238 像素×158 像素的灰度监控照片，平均每张图片有 25.0 个标注行人. 为了与 GB-MSCNet 的感知野大小相匹配，在训练及测试时将 UCSD 数据集中所有的图片放大 7 倍至 1066 像素×1666 像素. 本次试验中第 601 至第 1 400 张图片被用作训练集，余下的图片被用作测试集. 该数据集中人群图片、真实密度图、使用 GB-MSCNet 生成的密度图如图 10 所示. 用不同的方法测试 UCSD 数据集，得到不同网络结构下的计数误差，实验结果如图 11 所示.

分析图 11 中的 MAE 和 MSE 数值可知，GB-MSCNet 在计数准确率与计数稳定性两方面都展示了其良好的性能，证明了该网络在对相对稀疏的人群进行计数时表现良好.

3.3.2　密集人群实验验证

ShanghaiTech数据集是密集人群场景下的一组数据集，该数据集分为 A 和 B 两部分，其中，A 部分中包含 482 张图片，平均每张图片包含 501.4 个标注行人；B 部分中包含 716 张图片，平均每张图片包含 123.6 个标注行人. 根据之前研究的设定[2]，对于 A 部分，300 张图片被用作训练集，余下的 182 张作为测试集；对于 B 部分，400 张图片被用作训练集，余下的 316 张作为测试集. 该数据集中人群图片、真实密度图、使用 GB-MSCNet 生成的密度图如图 12 所示.

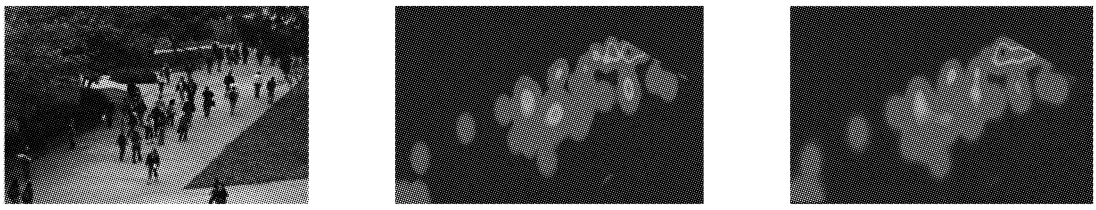


图 10　UCSD 实验结果示例

Fig. 10　Example of experiment results on UCSD dataset

使用不同的方法对 ShanghaiTech 数据集进行测试，得到不同网络结构下的计数误差，实验结果如图 13 所示. 从图 13 中可以看出，GB-MSCNet 在该数据集中的计数准确率和稳定性均较好.

UCF_CC_50 数据集中包含 50 张密集人群的图片，总共 63 974 个标注行人，平均每张图片包含 1 279.5 个行人，是目前公开的数据集中人群最密集的数据集. 该数据集中人群图片、真实密度图、使用 GB-MSCNet 生成的密度图如图 14 所示. 根据之前研究的设定[11]，在该数据集中使用五折交叉验证，其实验结果如图15所示. 从图15中可以看出，GB-

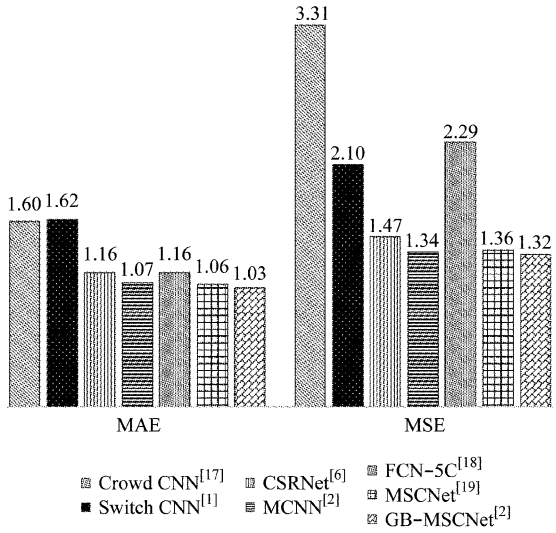


图 11 UCSD 数据集网络结构计数误差值

Fig. 11 Experiment results on UCSD dataset

MSCNet 在非常密集的人群图片中同样获得了较高的准确率与较好的稳定性.

3.3.3 TongjiCanteen 数据集验证

为了验证本文提出方法对于光照和视角变化的鲁棒性,本文采集并标注了 4 段 60 s 的真实场景下的视频,4 段视频拍摄于同一场景,但光照条件、视角都不相同. 图 16 是该数据集中不同条件下的人群图片、真实密度图和使用 GB-MSCNet 生成的密度图.

该数据集中,在每个视频前 50 s 每秒采集 3 帧图像作为训练集,剩余的所有图像作为测试集,在 4 个不同的视频中,GB-MSCNet 均表现良好,具体的 MAE 与 MSE 如表 1 所示.

3.3.4 实时性分析

本次实验所用的显卡型号为 GTX1070,处理器为 Intel i7 处理器. 在 TongjiCanteen 数据集中,测试

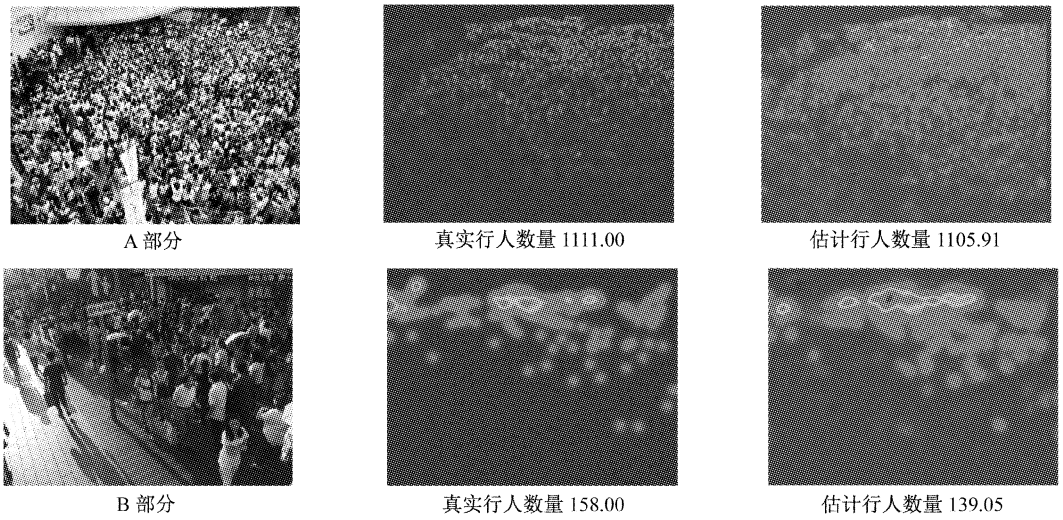
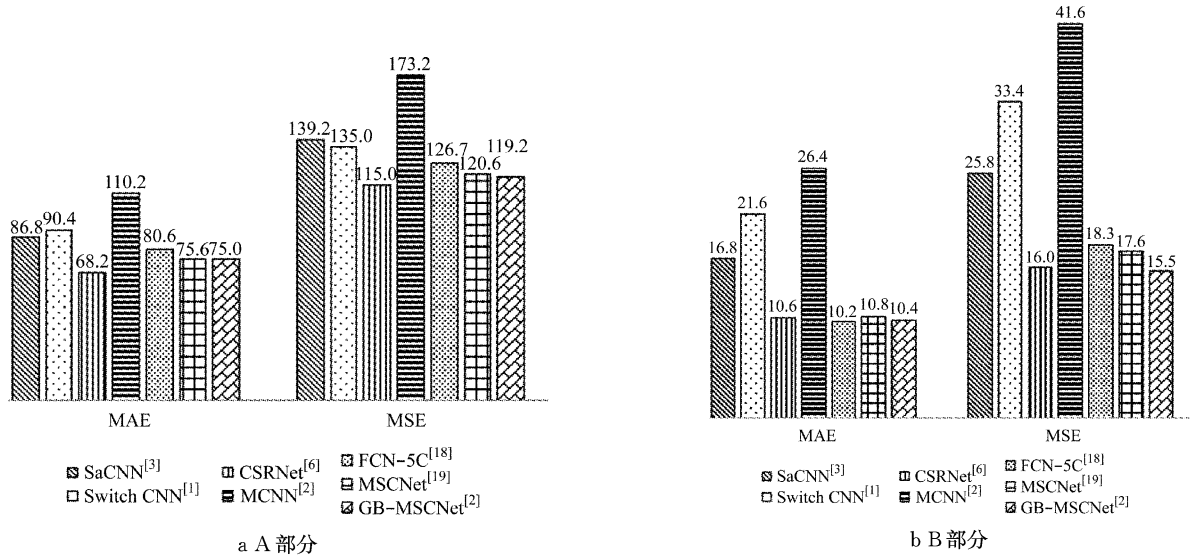


图 12 ShanghaiTech 实验结果示例

Fig. 12 Example of experiment results on ShanghaiTech dataset



a A 部分 b B 部分

图 13 ShanghaiTech 数据集计数误差值

Fig. 13 Experiment results on ShanghaiTech dataset

图 14　UCF_CC_50 实验结果示例

Fig. 14　Example of experiment results on UCF_CC_50 dataset

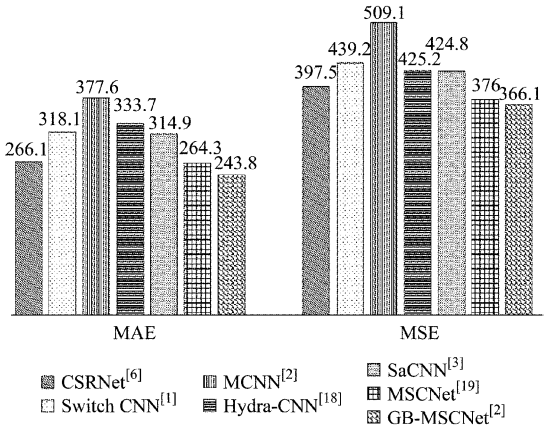


图 15　UCF_CC_50 数据集计数误差值

Fig. 15　Experiment results on UCF_CC_50 dataset

表 1　真实场景下计数误差值

Tab. 1　Experiment results in real scenario

场景	MAE	MSE	平均人数
视角一、光照条件一	0.66	0.92	8.33
视角一、光照条件二	1.22	1.64	20.58
视角二、光照条件一	0.55	0.84	9.28
视角二、光照条件二	0.95	1.31	23.03

的图片分辨率为(307 像素×425 像素)，本文提出的方法在该条件下的图像处理速度如图 17 所示. 从图 17 中可以看出，GB-MSCNet 平均每秒能够处理 38.9 帧图片，达到了在实际应用中的实时性要求.

综合以上数据集来看，GB-MSCNet 在对稀疏人群与密集人群进行计数时，能够准确给出复杂场景中行人的数目，并且对于相机视角和光照变化的鲁

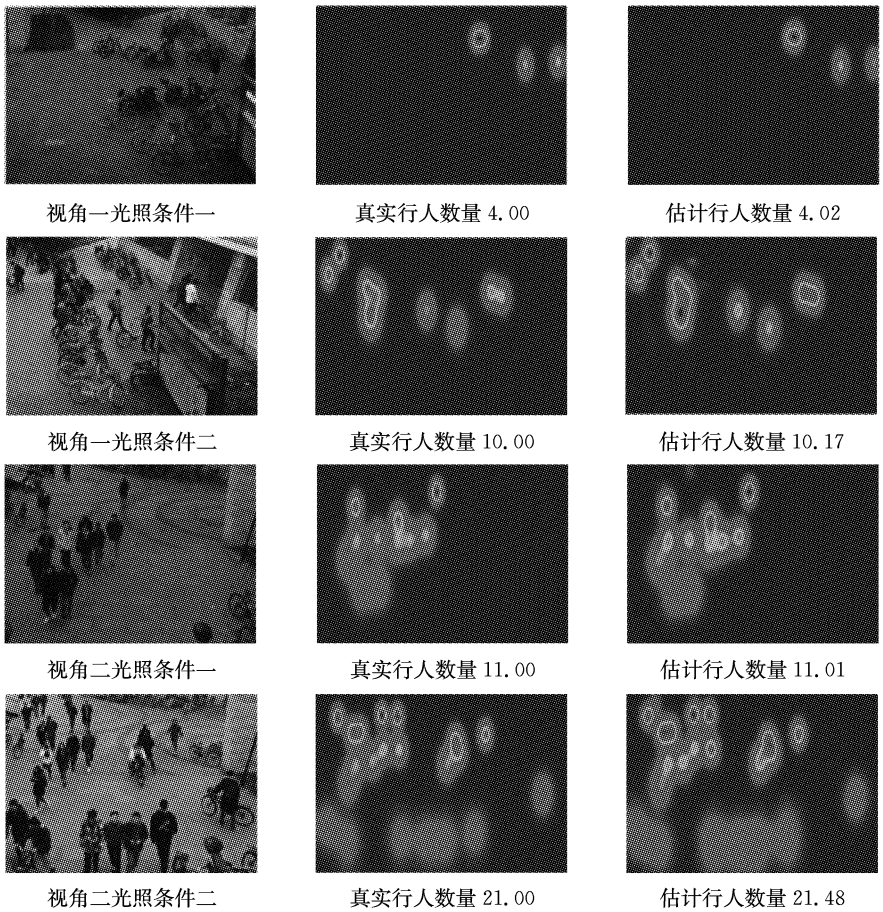


图 16　TongjiCanteen 数据集计数误差值

Fig. 16　Experiment results on TongjiCanteen dataset

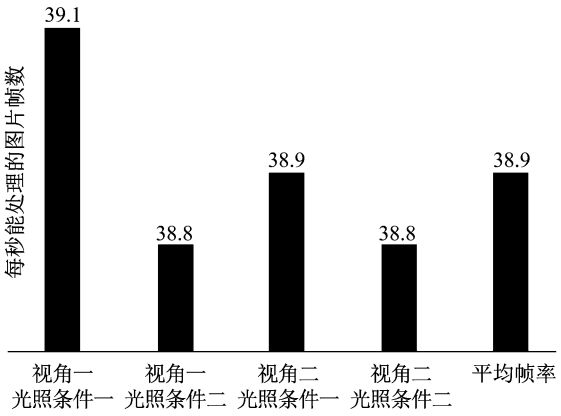


图17 GB-MSCNet在TongjiCanteen数据集中的运行速度

Fig.17 FPS of GB-MSCNet on TongjiCanteen dataset

棒性较强,较之于之前的方法在准确率和稳定性两方面都有所提高.

4 结语

本文针对基于视觉的人群计数进行了研究,提出了使用集成学习方法Gradient Boosting进行训练的人群计数网络GB-MSCNet,该网络具有较大的感知野,且能够融合各个尺度的特征,从而适应人群计数中行人的尺度变化.实验结果证明,对于不同的人群密度、相机视角与光照变化,该方法都能够较准确地给出图片中的行人数量,证实了该网络对于人群密度、行人大小、光照以及视角的变化具有较强的鲁棒性.与之前的人群计数方法相比较,GB-MSCNet在准确率与稳定性两方面均有较大的提高.

参考文献:

[1] SAM D B, SURYA S, BABU R V. Switching convolutional neural network for crowd counting[C]// IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4031-4039.

[2] ZHANG Y, ZHOU D, CHEN S, *et al*. Single-image crowd counting via multi-column convolutional neural network[C]// IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 589-597.

[3] ZHANG L, SHI M, CHEN Q. Crowd counting via scale-adaptive convolutional neural network[C] // IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe: IEEE, 2018: 1113-1121.

[4] HAN K, WAN W, YAO H, *et al*. Image crowd counting using convolutional neural network and markov random field[J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2017, 21(4): 632.

[5] KUMAGAI S, HOTTA K, KURITA T. Mixture of counting cnns: adaptive integration of cnns specialized to specific appearance for crowd counting [EB/OL].[2017-03-09]http://dx.doi.org/10.1007/s00138-018-0955-6.

[6] LI Y, ZHANG X, CHEN D. Csrnet: dilated covolutional neural networks for understanding the highly congested scenes[C]// IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake: IEEE, 2018: 1091-1100.

[7] CHEN K, LOY C C, GONG S, *et al*. Feature mining for localised crowd counting [C] // Procedings of the British Machine Vision Conference. Surrey: [s.n.], 2012: 1-11.

[8] 狄媛,杨东援. 基于多区域人群的上海公交走廊出行行为[J]. 同济大学学报(自然科学版), 2016, 44(3): 369.
DI Yuan, YANG Dongyuan. Travel behavior analysis of different-regional passengers for public transport corridor in Shanghai[J]. Journal of Tongji University(Natural Science), 2016, 44(3): 369.

[9] 狄媛,杨东援. 基于人群分类的城市公交走廊客流分配模型[J]. 同济大学学报(自然科学版), 2016, 44(2): 235.
DI Yuan, YANG Dongyuan. A passenger classification transportation assignment model for urban public traffic corridor[J]. Journal of Tongji University (Natural Science), 2016, 44(2): 235.

[10] 姬丽娜,陈庆奎,陈圆金等.基于GPU的视频流人群实时计数[J]. 计算机应用,2017, 37(1): 145.
JI Lina, CHEN Qingkui, CHEN Yuanjin, *et al*. Real-time crowd counting method from video stream based on GPU[J]. Journal of Computer Applications, 2017, 37(1): 145.

[11] IDREES H, SALEEMI I, SEIBERT C, *et al*. Multi-source multi-scale counting in extremely dense crowd images[C]// IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013: 2547-2554.

[12] WANG M, WANG X. Automatic adaptation of a generic pedestrian detector to a specific traffic scene[C] // IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs: IEEE, 2011: 3401-3408.

[13] BO W, NEVATIA R. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors[C] // International Conference on Computer Vision. Beijing: IEEE, 2005: 90-97.

[14] CHAN A B, LIANG Z J, VASCONCELOS N. Privacy preserving crowd monitoring: counting people without people models or tracking[C]// IEEE Conference on Computer Vision and Pattern Recognition. Anchorage: IEEE, 2008: 1-7.

[15] OÑORO-RUBIO D, L'OPEZ-SASTRE R J. Towards perspective-free object counting with deep learning[C] // Europe Conference on Computer Vision. Amsterdam: [s.n.], 2016: 615-629.

[16] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL].[2014-09-15].https://arxiv.org/abs/1409.1556.

[17] CHEN X J, GUO R Q, LUO W, *et al*. Visual crowd counting with improved Inception-ResNet-A module [C] // IEEE International Conference on Robotics and Biomimetics. Kuala Lumpur: IEEE, 2018: 112-119.