文章编号：2095-4980(2023)07-0934-05

一种基于卷积神经网络的轨道交通场景人群计数模型

杨路辉[1]，湛忠义[1]，潘尚考[1]，刘光杰[*2]，陆　斌[3]

(1.南京理工大学 自动化学院，江苏 南京 210094；2.南京信息工程大学 电子与信息工程学院，江苏 南京 210044；
3.南京熊猫信息产业有限公司，江苏 南京 210038)

摘　要：现有的人群计数方法不能够完全适用于轨道交通场景中，为此，提出一种基于卷积神经网络的人群计数模型。模型采用VGG16作为前端网络提取浅层特征，提出一种基于Inception结构改进的M-Inception结构，结合空洞卷积构成后端网络，增大感受野，适应多监控角度下不同尺寸的行人目标；并提出一种融合行人总数估计损失和密度图损失的加权损失函数。将本文模型与4种现有模型进行对比实验，结果表明，提出的人群计数算法在地铁场景中的平均绝对误差和均方误差仅为1.46和2.13，优于4种对比模型。考虑到模型的实际应用，将模型部署到海思嵌入式芯片上，实测结果表明，模型可在嵌入式芯片上取得较高的计算速度和准确率，满足实际应用场景的需求。

关键词：人群计数；地铁场景；空洞卷积；嵌入式实现

中图分类号：TP309　　**文献标志码：**A　　**doi：**10.11805/TKYDA2020550

A crowd counting model for rail transit scene based on convolutional neural network

YANG Luhui[1]，ZHAN Zhongyi[1]，PAN Shangkao[1]，LIU Guangjie[*2]，LU Bin[3]

(1.School of Automation，Nanjing University of Science & Technology，Nanjing Jiangsu 210094，China；
2.School of Electronic & Information Engineering，Nanjing University of Information Science & Technology，
Nanjing Jiangsu 210044，China；3.Nanjing Panda Information Industry Co.，Ltd.，Nanjing Jiangsu 210038，China)

Abstract：The existing crowd counting methods are not suitable for the subway scene. Therefore, a crowd counting model based on convolutional neural network is proposed. The model takes the VGG16 as the front-end network to extract the shallow features, and an M-Inception structure is combined with the dilated convolution to form the back-end network, which can increase the receptive field and adapt to different sizes of pedestrian targets at different monitoring angles. And a weighted loss function combining the head count loss and density map loss is proposed. The proposed algorithm is compared with four existing models. The experimental results show that the Mean Absolute Error(MAE) and Mean Square Error(MSE) of the proposed algorithm are 1.46 and 2.13, better than those of the four comparison models. The proposed model is deployed to Hisilicon embedded chip. The test results show that the proposed model can achieve high computing speed and accuracy on the embedded chip, which can meet the requirements of the actual application scenarios.

Keywords：crowd counting；subway scene；dilated convolution；embedded implementation

在城市轨道交通系统日益发达的今天，为实现精细化客流管理，需要高效准确的人群计数技术。近年来，基于深度学习的计算机视觉算法应用在人群计数领域，依据其主要原理大致可分为基于人体目标检测的算法[1-3]和基于密度图回归的算法[4-14]。基于检测的算法一般基于目标检测模型，试图检测出图片中的每个行人，但在人群拥挤严重的场景，检测准确率无法保障，只适用于人群遮挡较少的场景。基于密度图回归的算法通过建立图片特征与人群密度图的回归关系，在拥挤场景下能够实现更好的计数效果。在地铁站内部，高峰时期的人群遮

收稿日期：2020-10-22；**修回日期：**2021-03-24

基金项目：国家自然科学基金资助项目(U1836104)

[*]**通信作者：**刘光杰　email:gjieliu@gmail.com

挡严重，此时基于检测的算法无法正确检测人数。因此基于密度图回归的算法更加适用。前人的工作中使用的数据集大多数来自街道摄像头或商场内部摄像头，其拍摄角度较高，每个行人目标成像大小接近。然而地铁站内监控摄像头安装位置较低，行人目标距离摄像头远近不同，在监控图像中的成像大小也不相同。因此，针对地铁监控场景中角度更加复杂的行人计数问题，需要设计更加有效的数据集和算法。

本文提出一种适用于地铁站拥挤场景的人群计数模型：MPCNN(Metro Passenger flow Counting Neural Network)。该模型基于卷积神经网络设计，使用VGG16[15]作为前端网络提取浅层特征，然后基于Inception结构设计了一个可更好适应目标尺寸变化的M–Inception结构，同时加入空洞卷积组成后端网络。本文还提出一种融合行人总数估计损失和密度图损失的加权损失函数，可提高模型训练结果的准确率。实验结果表明，本文提出的算法在地铁场景中的性能提升明显，显著优于现有模型。考虑到模型的实际应用，将模型移植到海思嵌入式人工智能芯片上，模型移植后的精确度平均损失仅为3.32%，且单张图片推理平均时间仅为0.875 s，能够满足实际应用需求。

1　MPCNN 模型

1.1　MPCNN 模型结构

网络结构的前端网络采用预训练好的VGG16模型的前10层，后端网络采用M–Inception网络和空洞卷积构成。Inception网络结构由文献[16]提出，能够使用不同尺寸的卷积核增大感受野。对于地铁场景下的监控图像，人体与头部在图片中显示尺寸比例经常不同(近处人体尺寸大，远处人体尺寸小)，因此Inception网络的多尺寸卷积核可提取不同尺寸大小的人体特征。本文基于Inception网络结构，设计了适合本场景下的M–Inception模块，如图1所示。模块包含4个卷积分支。第1个分支为1个1×1卷积层，用于提取全局特征；第2个分支为1个1×1卷积层连接1个3×3卷积层，用于提取卷积特征；第3个分支为1个池化层连接1个1×1卷积层，用于提取筛选过的全局特征；第4个分支为1个1×1卷积层连接2层3×3卷积层，用于提取更深层的卷积特征。最后将4个卷积分支的特征图进行拼接，能够有效获取全局与局部的卷积特征。

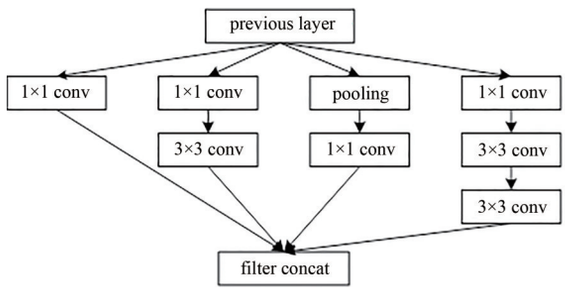


Fig.1 Structure of M–Inception
图1　M–Inception模块结构图

考虑到人体尺寸和视角变化是影响人群密度估计的主要因素，输入图像数据在下采样与上采样变换过程中信息会丢失，为此在M–Inception模块之后引入空洞卷积，空洞卷积在保持图像分辨力的同时，实现交替卷积，增大了感受野，更适合人群密度估计。后端网络由一个M–Inception结构、5层空洞卷积层和1个普通卷积层构成，完整的网络结构如图2所示。

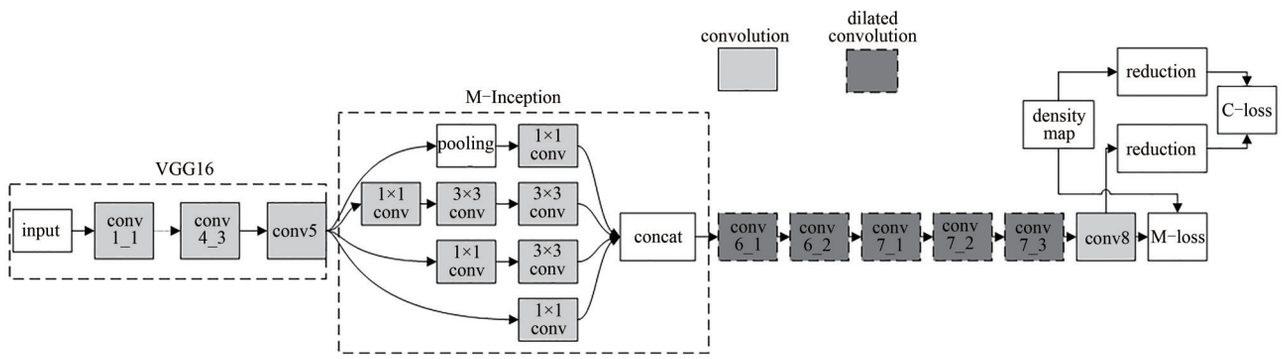


Fig.2 Architecture of MPCNN model
图2　MPCNN模型结构

1.2　加权损失函数

密度图损失见式(1)，采用欧式距离的平均值计算模型密度图误差。行人总数损失见式(2)，采用归一化的欧式距离取平均值计算模型预测行人总数误差。为了提高模型训练效果，提出一种结合密度图损失(M–loss)和行人总数损失(C–loss)的加权损失函数，见式(3)。

$$L_{\mathrm{M}}(\theta)=\frac{1}{2N}\sum_{i=1}^{N_1}\left\| F_{\mathrm{M}}(X_i;\theta)-M_i \right\|^2 \tag{1}$$

式中：θ为网络参数；N_1为训练图片数；X_i为输入图片；F_{M}为输入图像经网络模型所估计的密度图；M_i为输入图像的真实密度图。

$$L_{\mathrm{C}}(\theta)=\frac{1}{2N}\sum_{i=1}^{N_1}\left\| \frac{F_{\mathrm{C}}(X_i;\theta)-C_i}{C_i+1} \right\|^2 \tag{2}$$

式中：F_{C}为输入图像经网络模型预估的人数；C_i为输入图像的实际人数。

$$L_{\mathrm{MC}}=L_{\mathrm{M}}+\lambda L_{\mathrm{C}} \tag{3}$$

式中λ为行人总数估计损失函数的权重。

2 实验与结果分析

2.1 数据集构建

为构建地铁站人群计数数据集，本文从地铁站台监控摄像头采集数据。在不同的时间段截取视频，以获取不同密度人群数据集。本文参考公开数据集数据标注方式，将图片所有人头中心区域位置坐标保存成矩阵形式，使用高斯核卷积的方式生成对应的人群密度图，经过实验后选取高斯核大小为25，方差取值为1.5，转化后的密度图如图3所示。

Fig.3 Convert label image to density map
图3 标注图转成密度图

实验采用了2个数据集：一个是本文构建的地铁行人数据集(Metro-Dataset)，共1 148张，样本内人群人数分布范围为1~88；另一个是公开的ShanghaiTech-PartB数据集，共716张人群图像，样本内人群人数分布范围为9~578，平均123人。由于ShanghaiTech-PartB数据集大小仅为Metro-Dataset的60%，本文对其进行了数据增广，将每张图片裁剪为9块，每块大小为原图的1/4，且其中4块图像无重叠，其余5块随机裁剪。

2.2 评价标准

实验结果评价指标采用平均绝对误差(MAE)和平均方差误差(MSE)，见式(4)。MAE和MSE值越低，表示模型效果越好。

$$\begin{cases} E_{\mathrm{MA}}=\dfrac{1}{N}\displaystyle\sum_{i=1}^{N_2}\left| t_i-\overline{t}_i \right| \\ E_{\mathrm{MS}}=\sqrt{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N_2}\left(t_i-\overline{t}_i \right)^2} \end{cases} \tag{4}$$

式中：N_2为测试图像总数；$t_i-\overline{t}_i$为实际人数与预测人数误差。

2.3 模型结果分析

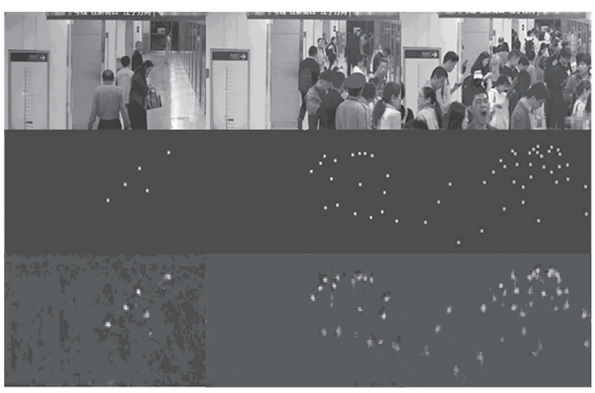
为验证本文提出算法的效果，选取当前表现较好的4种人群计数算法SaCNN[10]、CSRNet[11]、CACC[12]和ADCrowdNet[13]进行对比，且都在Metro-Dataset和ShanghaiTech-PartB数据集上训练、测试。对比测试结果如表1所示。

从表1结果可知，本文模型在Metro-Dataset上的MAE和MSE仅为1.46和2.13，显著优于4种对比算法；在公开的ShanghaiTech-PartB数据集上，结果略好于几种对比算法。实测结果如图4~图5所示，图中从上到下依次为实际场景图、实际人群密度图、模型预测密度图。结果显示行人数量预测误差仅为1%左右，且算法对不同行人密度场景有较强的适应性。

表1 对比实验结果

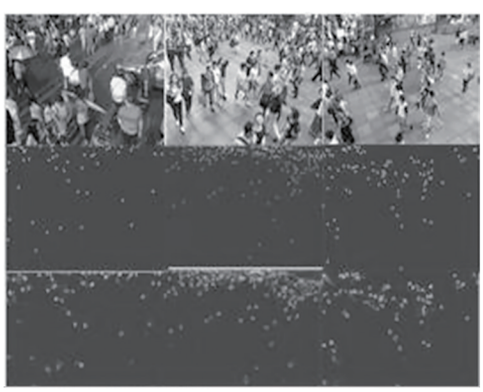
Table1 Comparative experimental results

model	dataset			
	Metro–Dataset		ShanghaiTech–PartB Dataset	
	MAE	MSE	MAE	MSE
SaCNN	2.223 9	3.054 5	16.473 8	24.326 7
CSRNet	1.800 2	2.351 9	11.773 8	18.591 0
CACC	1.702 5	2.221 4	11.621 4	18.482 3
ADCrowNet	1.623 1	2.164 1	11.142 1	17.753 2
MPCNN	1.464 9	2.132 5	11.087 3	17.748 7



ground truth: 5.0 ground truth: 18.0 ground truth: 37.0
detected: 5.48 detected: 17.94 detected: 39.2

Fig.4 Test results on subway dataset
图4 地铁站台数据测试结果



ground truth: 59.0 ground truth: 217.0 ground truth: 89.0
detected: 58.49 detected: 214.76 detected: 90.42

Fig.5 Test results on ShanghaiTech–PartB dataset
图5 ShanghaiTech–PartB数据测试结果

2.4 嵌入式设备移植测试

为测试算法在真实应用环境中的性能表现，本文在海思Hi3559CV100嵌入式人工智能芯片上部署了该算法。Hi3559系列可将Caffe模型转为INT8型WK模型，经过INT8量化后可加快推理速度。实验评估了经过INT8转换的模型精确度损失以及模型计算速度。实验中采集样本80张，将测试样本平均分为8组。取每组数据平均值作为该组测试最终准确率，通过对比转换后的WK模型与未转换的Caffe模型精确度，测试其精确度损失以及在嵌入式芯片上的推理时间结果如表2所示。由表2可知，模型转化后平均损失精确度为3.32%，总体上在可接受的范围内，且嵌入式模型平均测试准确率为92.0%，单张图片平均推理时间仅为0.875 s。本文所提出的人群计数算法移植到嵌入式设备时，能够取得不错的准确率，同时计算速度能够满足实际应用的需求。

表2 本文模型计数准确率测试结果

Table2 Counting accuracy results of the proposed model

No.	Caffe model accuracy/%	embedded transplant accuracy/%	inference time/s
1	96.862	91.179	0.850
2	95.776	90.394	0.880
3	93.746	89.837	0.900
4	95.293	93.792	0.860
5	94.525	89.155	0.880
6	95.748	92.891	0.870
7	96.479	95.987	0.880
8	94.126	92.749	0.880
Avg.	95.319	91.998	0.875

3 结论

针对轨道交通场景中的人群计数问题，本文提出一种基于卷积神经网络的人群计数模型MPCNN。模型采用预训练的VGG16模型中的卷积部分作为前端网络提取图片特征，使用基于M–Inception结构感知不同尺寸的人体头部区域，并结合空洞卷积层扩大感受野。本文模型在一个公共数据集和本文采集的地铁站台数据集上进行了实验，同时将实验结果与当前4种同类型优秀算法进行了对比，实验结果表明，MPCNN的结果优于对比算法。考虑到模型的实际应用场景，将模型移植到海思Hi3559CV100芯片中，实验表明，本文模型可在嵌入式芯片中实现准确快速的人群计数。

参考文献：

[1] SONG Diping, QIAO Yu, CORBETTA A. Depth driven people counting using deep region proposal network[C]// 2017 IEEE International Conference on Information and Automation(ICIA). Macao,China:IEEE, 2017:416-421. doi:10.1109/ICInfA.2017.8078944.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,NV,USA:IEEE, 2016:779-788.

[3] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: Single Shot multibox Detector[C]// European Conference on Computer Vision. Amsterdam:Springer,Cham, 2016:21-37. doi:10.1007/978-3-319-46448-0_2.

[4] ZHANG Yingying, ZHOU Desen, CHEN Siqin, et al. Single-image crowd counting via multi-column convolutional neural network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,NV,USA:IEEE, 2016:589-597.

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

[6] CHAN A B, VASCONCELOS N. Bayesian Poisson regression for crowd counting[C]// 2009 IEEE the 12th International Conference on Computer Vision. Kyoto,Japan:IEEE, 2009:545-551. doi:10.1109/ICCV.2009.5459191.

[7] ZHANG Cong, LI Hongsheng, WANG Xiaogang, et al. Cross-scene crowd counting via deep convolutional neural networks[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Boston,MA:IEEE, 2015:833-841. doi:10.1109/CVPR.2015.7298684.

[8] SINDAGI V A, PATEL V M. Generating high-quality crowd density maps using contextual pyramid CNNs[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice,Italy:IEEE, 2017:1861-1870. doi:10.1109/iccv.2017.206.

[9] CAO Xinkun, WANG Zhipeng, ZHAO Yanyun, et al. Scale aggregation network for accurate and efficient crowd counting[C]// Proceedings of the European Conference on Computer Vision(ECCV). Berlin, Germany: Springer, 2018: 734-750. doi: 10.1007/978-3-030-01228-1_45.

[10] ZHANG Lu, SHI Miaojing, CHEN Qiaobo. Crowd counting via scale-adaptive convolutional neural network[C]// 2018 IEEE Winter Conference on Applications of Computer Vision(WACV). Lake Tahoe, NV, USA: IEEE, 2018: 1113-1121. doi: 10.1109/WACV.2018.00127.

[11] LI Yuhong, ZHANG Xiaofan, CHEN Deming. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA:IEEE, 2018:1091-1100. doi:10.1109/CVPR.2018.00120.

[12] LIU Weizhe, SALZMANN Mathieu, FUA Pascal. Context-aware crowd counting[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach,CA,USA:IEEE, 2019:5094-5103. doi:10.1109/CVPR.2019.00524.

[13] LIU Ning, LONG Yongchao, ZOU Changqing, et al. ADCrowdNet: an attention-injective deformable convolutional network for crowd understanding[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA:IEEE, 2019:3220-3229. doi:10.1109/CVPR.2019.00334.

[14] 向东,卿粼波,何小海,等. 基于深度学习的视频人群计数系统[J]. 太赫兹科学与电子信息学报, 2020,18(3):515-519. (XIANG Dong, QING Linbo, HE Xiaohai, et al. Video crowd counting system based on deep learning[J]. Journal of Terahertz Science and Electronic Information Technology, 2020,18(3):515-519.) doi:10.11805/TKYDA2019234.

[15] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J/OL]. arXiv: 1409.1556, 2014.

[16] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston,MA,USA:IEEE, 2015:1-9. doi:10.1109/CVPR.2015.7298594.

作者简介：

杨路辉(1992-)，男，博士，高级工程师，主要研究方向为深度学习与网络安全. email: yangluhui005@foxmail.com.

湛忠义(1995-)，男，硕士，主要研究方向为深度学习与图像处理.

潘尚考(1996-)，男，硕士，主要研究方向为深度学习与图像处理.

刘光杰(1980-)，男，博士，教授，博士生导师，主要研究方向为网络与多媒体安全、深度学习与图像处理.

陆　斌(1970-)，男，硕士，高级工程师，主要研究方向为城市轨道交通 AFC 系统.