·图形图像处理·　　文章编号：1000-3428(2020)09-0254-07　　文献标志码：A　　中图分类号：TP391

基于注意力机制的狭小空间人群拥挤度分析

张　菁，陈庆奎

（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘　要：人群拥挤度的分析对维护公共安全极为重要，在空间狭窄的环境下，由于视角受到局限，人与人、人与物品的遮挡十分严重，并且人的尺度不一，密度不均匀，使得传统人群拥挤度监控方法较难直接统计出具体人数。为此，提出一种基于注意力机制的狭小空间人群拥挤度分析方法，旨在量化人群，通过卷积神经网络回归拥挤率分析当前空间内的人群拥挤程度。设计一个注意力模块作为网络的前端，通过生成对应尺度的注意力图区分背景和人群，保留精确的像素点位置信息，以减轻输入图像中各种噪声的影响。在此基础上，将注意图和原始图片通过对应像素点相乘，注入到微调的残差网络中训练得到人群拥挤率。实验结果表明，该方法能够预测出拥挤率，准确反映当前人群拥挤程度，实现人群的流量控制。

关键词：人群拥挤度；狭小空间；注意力机制；卷积神经网络；残差网络

开放科学（资源服务）标志码（OSID）：

中文引用格式：张菁，陈庆奎. 基于注意力机制的狭小空间人群拥挤度分析[J]. 计算机工程，2020，46(9)：254-260，267.

英文引用格式：ZHANG Jing，CHEN Qingkui. Analysis of crowd congestion degree in narrow space based on attention mechanism[J]. Computer Engineering，2020，46(9)：254-260，267.

Analysis of Crowd Congestion Degree in Narrow Space Based on Attention Mechanism

ZHANG Jing，CHEN Qingkui

（School of Optical-Electrical and Computer Engineering，University of Shanghai for Science and Technology，Shanghai 200093，China）

【Abstract】The analysis of crowd congestion degree is very important to maintain public safety. Generally，in a narrow space the perspective is limited，and the human-human occlusion and human-item occlusion are serious. In addition，because of the different scales of people and uneven density，the traditional methods often fail to directly get the specific number of people in a narrow space. To address the problem，this paper proposes an analysis method of crowd congestion degree in a narrow space based on the attention mechanism in order to quantify the crowd. The method analyzes the congestion degree in the current space through the regression congestion rate of the Convolutional Neural Network (CNN). It designs an attention module as the front end of the network，distinguishes the background and the crowd by generating attention maps of corresponding scales，and retains accurate pixel position information to reduce the impact of various noises in the input image. The attention graph and the original image are multiplied by corresponding pixels and injected into the fine-tuned ResNet to train the crowd congestion rate. Experimental results show that the proposed method can predict the congestion rate，accurately reflect the current crowd congestion degree，and realize crowd flow control.

【Key words】crowd congestion degree；narrow space；attention mechanism；Convolutional Neural Network (CNN)；Residual Network(ResNet)

DOI：10.19678/j.issn.1000-3428.0055701

0　概述

随着城市人口的增多，在狭小空间的场景下，人群密度高，很容易造成拥挤，从而引发安全隐患。为维护公共安全，需要实时获取人群的拥挤程度，根据不同的拥挤等级实施不同的调度策略合理分配资

基金项目：国家自然科学基金(61572325，60970012)；高等学校博士学科点专项科研博导基金(20113120110008)；上海重点科技攻关项目(14511107902，16DZ1203603)；上海市工程中心建设项目(GCZX14014)；上海智能家居大规模物联共性技术工程中心项目(GCZX14014)；上海市一流学科建设项目(XTKX2012)；沪江基金研究基地专项(C14001)。

作者简介：张　菁(1994—)，女，硕士，主研方向为图像处理、人工智能、GPU 并行计算；陈庆奎，教授、博士、博士生导师。

收稿日期：2019-08-09　　**修回日期：**2019-09-23　**E-mail：**851964502@qq.com

源[1],实现公共场所下人群的流量控制,因此,空间内人群拥挤度的分析极为重要。传统人群拥挤度的监控主要靠人力监控,但很难同时且长时间监控多个场景,并且人力监控主观性较强,不同的人给出的拥挤等级的标准都不尽相同。然而,通过深度学习来自动获取人群拥挤率可以节约人力物力,提高工作效率和准确率。

但是人群拥挤度的估计面临着许多挑战,如背景杂乱、严重遮挡、密度不均匀、场景内和场景间的尺度以及视角变化等。近年来,随着深度学习和大规模人群数据集的发展,多数量化当前场景人群的方法都取得了显著的成果。但是多数模型只适用于室外、视角广阔、人群密度均匀的场景,而在如直升电梯、楼道、隧道、车厢等狭小空间内,视角局限、人群遮挡严重、图像尺度不一致增加了对人群密度分析的难度。

本文针对狭小空间场景下的人群进行分析,提出一种基于注意力机制的人群量化方法。该方法将拥挤率作为网络真实值,通过对其进行回归得到估计的拥挤率,并在 ResNet 提取特征的基础上进行微调,以适应本文数据集的训练,在此基础上引入注意力机制,构建一个新的注意力注入网络及数据集 NS-DATASET。

1　相关工作

人群被视为一个连续密度函数[2],其对任意图像区域的积分得到该区域内行人的数量。近年来,主流的量化人群的方法主要是通过标记的人头点来生成密度图,将密度图作为网络的真实值,对卷积神经网络回归的密度图进行积分得到具体人数。为解决人群多视角、多尺度分布不均匀的问题,文献[3]提出多列卷积神经网络(MCNN)来提取多种图像多尺度特征;文献[4]提出单列卷积神经网络 CSRNet 使用了空洞卷积,在扩大感受野的同时,不损失图像分辨率,以克服多列卷积训练的弊端;文献[5]提出利用单列深层网络 saCNN,通过跳过连接将两层的两个特征映射组合在一起,完成图像多尺度的提取。这些方法都旨在生成更精准的密度图,通过对密度图积分来计算具体人数,从而分析空间内拥挤程度。

但是在狭小空间内视角受到的局限,人与人、人与物品的遮挡十分严重,利用目标检测[6]很难精准地识别出每个人,并且人的尺度不一,人头点的密度不均匀,对于基于密度图[3-5]的方法,本文的实验结果表明,回归的密度图在密度小、空间狭窄的环境下,没有展现出很好的效果,因此只能用于视野广阔、密度均匀和高度拥挤的场景。

目前多数方法利用卷积神经网络将人群划分为不同的拥挤等级,通过不同的等级分类实现对人群密度的监控。文献[7-9]通过分析图像的纹理特征来提取人群密度特征,然后采用 SVM 进行分类,将人群密度分为若干个等级。文献[10-11]通过比较目前流行的深度学习网络框架,选取了 GoogLeNet[12]作为主干网络,对人群密度进行分类。针对人群密度不均匀的情况,文献[13]依据像素稠密度将图像中的人群分割成若干团块,每块分为高低密度两类。上述方法都能得到有效的密度等级分类,但是分化的类别单一,并不能很好地起到量化人群的作用。

为了能更好地量化人群,本文设计了一个能直接反映当前场景拥挤率的模型——设置网络的真实值为拥挤率(当前实际人数除以空间最大容纳数)。这相比于回归密度图的方法,能将数据归一化,减小数据间的差异,并且更注重人群整体密度情况而不是聚焦到单个人数上面,提高了人群量化的准确性。而和直接将人群分为若干个等级相比,本文的模型相当于将拥挤率分为 0~100,共 101 个等级,量化的更加细致,更能反映当前人群密度的真实情况。本文选择采用微调过的 ResNet50[14]作为网络主干,因为 ResNet 有着强大的表征能力,使得目标检测和图像识别的许多计算机视觉应用都得到了性能提升。

但由于网络的真实值是拥挤率,只是一个数值,模型在训练前并没有区分背景与目标,只能依赖于大量的数据不断学习来调整权重,导致了模型训练过程收敛速度慢,且难以收敛,给训练增加了难度。因此,本文引入了注意力机制来增强卷积,为网络拥挤率的回归提供先验知识。

注意力机制多被用于理解上下文语义。近年来,基于注意力机制的图像分割也获得了成功,如文献[15]设计的注意力机制利用全卷积网络(FCN)作为中间层来合并多尺度的特征,文献[16]提出上下文编码模块,结合扩张卷积和多尺度策略提出了语义分割框架 EncNet,用于捕获图像场景的上下文语义,选择性地突出与类别相关的特征图。但是,现有很多成熟的方法训练时往往需要大量的标记图像。对于图像分割而言,要得到大量的完整标记过的图像非常困难。因此,不少基于弱监督定位的 CNN 被提出,如文献[17]设计了类激活映射,能在各种各样的图像识别任务中,使网络直到最后一层都保留其显著的定位能力,通过类激活映射权重,反映不同区域的相对重要性,完成图像区域的划分。文献[18]利用多列空洞卷积完成了弱监督、半监督的学习,大大减轻了图像分割定位的训练难度。

本文的注意力模块旨在通过二分类的网络生成注意力图,将输入图像分为了背景区域和人群区域,能为主干网络预示出人群的区域,更好地学习人群的特征。为了解决标记人群区域轮廓困难的问题,本文参照文献[17-18]中的目标定位方法,通过弱监督学习,自动获取人群区域的位置。将注意特征图连接到卷积特征图,能使神经网络的训练关注人群聚集区域,更好地适应高噪声场景。

2 网络整体结构

本文在 ResNet50[14] 网络提取特征的基础上,添加了注意力机制模块,设计一个新的注意力注入的网络,网络结构如图 1 所示。网络前端的注意力模块将输入图像分为人群和背景,生成更关注人群目标的注意力图(图 1 中注意力图圈出来的部分即为注意力模型得到的网络需要聚焦的区域,为更直观地定位本文的注意力关注区域,将输入图片作为背景,实际训练中的背景区域像素点注意力得分为零),然后将注意力图和原始图像进行点相乘,采用的方式是图像处理中的 pixel-wise 操作[19],即两张图片对应像素点相乘。将处理好的特征图作为 ResNet 的输入,并将损失函数 Loss 直接作为网络的目标函数,对特征图进行回归,直接输出该特征图的人群拥挤率 Crowd。

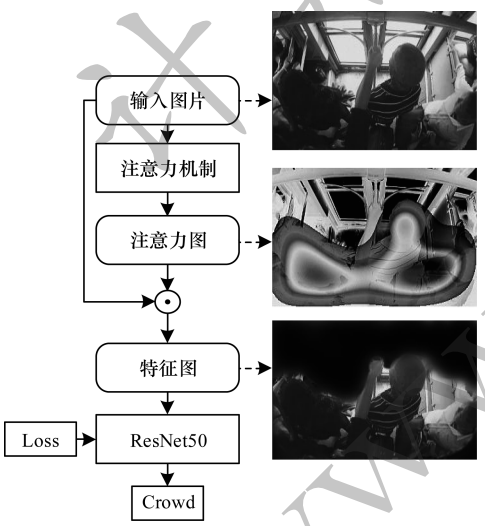


图 1 整体网络结构

Fig. 1 Overall network structure

2.1 注意力机制

注意力机制被广泛应用于计算机视觉领域,其基本原理是利用输入图像的相关信息而不是所有可用信息来计算神经网络,获取长程交互性,能够在纳入相对位置信息的同时维持平移等效性,从而极大地减少网络的计算量。本文设计的注意力模块作为网络的前端,通过生成注意力图来提取精确的像素点定位细节,为主干网络预示出人群的区域以及人群的拥挤等级,将注意特征图连接到卷积特征图,能使神经网络的训练更加关注人群聚集区域,以减轻输入图像中各种噪声的影响。

生成注意力图的工作流程如图 2 所示,网络结构采用的是微调过的 GoogLeNet[12],该网络在图像分类和目标定位上都展现出了很好的性能。将 GoogLeNet 的 Inception4e 后面的层移除,保持图像分辨率为 14 像素×14 像素,为使最后的注意力图和原始输入图片能够融合,保留图片像素点的位置信息,每个卷积层后面都利用 padding 对特征图进行填充,在卷积降低图像分辨率的同时,保持尺度不变。在卷积输出前,参照文献[17]的方法,使用全局平均池化(GAP)和 Softmax 层,将输出的每个类别的权重映射回卷积特征图,从而生成注意力图。

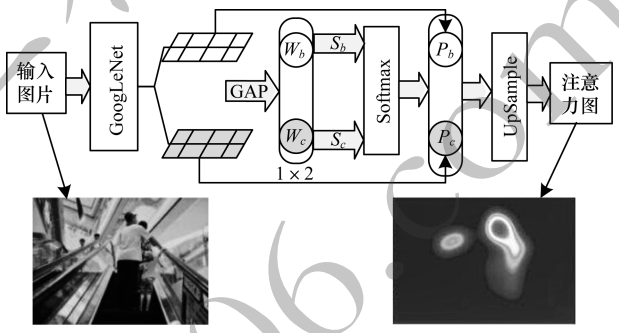


图 2 Attention 模块流程

Fig. 2 Procedure of Attention module

本文将注意力模块设计为一个二分类的网络,并将图片内容分为背景和人群。如图 2 所示,F_b 和 F_c 是最后一层卷积输出的两个通道的特征图,$F_b(x,y)$、$F_c(x,y)$ 分别表示背景和人群在坐标 (x,y) 上像素点的激活值,通过全局平均池化后(图 2 中的 GAP),得到长度为 2 的向量,每个长度对应一个类别权重 W_b、W_c,那么 Softmax 层的输入 S_c、S_b 如式(1)所示,Softmax 层的输出 P_c、P_b 可根据式(2)得到,最后将每个像素点坐标的得分 $P_{c,b}(x,y)$ 与特征图 $F_{c,b}(x,y)$ 根据式(3)进行线性加权融合后,再利用向上采样(UpSample),得到与原始图片尺寸相同的注意力图。

$$S_{c,b}=\sum_{x,y}S_{c,b}(x,y)=\sum_{x,y}W_{c,b}\cdot F_{c,b}(x,y) \tag{1}$$

$$P_{c,b}(x,y)=\frac{\exp(S_{c,b}(x,y))}{\sum_{c,b}\exp(S_{c,b}(x,y))} \tag{2}$$

$$\mathrm{AM}=\sum_{x,y}P_c(x,y)\cdot F_c(x,y)+\sum_{x,y}P_b(x,y)\cdot F_b(x,y) \tag{3}$$

2.2 ResNet 模块

ResNet 是一个由微软开发的深度卷积网络,它主要通过残差连接来工作,ResNet[14] 接受域大于输入图像,使训练数百层甚至数千层成为可能,且

在这种情况下仍能展现出优越的性能。本文参照 VGG 网络[20]的结构,将一个 5×5 卷积层分解成 2 个串联的 3×3 卷积层,并将原始的 ResNet50 的 7×7 卷积层替换为 3 个串联的 3×3 卷积层,在保持接受域的大小不变的同时,减少了网络参数,并且引入更多的非线性:一个大卷积核只有一次激活的过程,而更多串联的小卷积核对应着更多次激活的过程,从而增加了网络的表达能力,可以去拟合更高维的分布。本文还去掉了最后的 Softmax 层,选用全连接层作为网络最后一层,将 1 000 维的向量改为 101 维,用于表示数值在 0~100 的人群拥挤率,将分类问题转化为回归问题。原始的 ResNet50 网络和微调的网络结构如表 1 所示,每层卷积步长设计为 2。

表 1　ResNet50 网络结构

Table 1　Structure of ResNet50 network

卷积层	ResNet50	微调的 ResNet50	输出大小
Conv1	7×7,64,stride 2	[3×3,64]×3 stride 2	112×112
Conv2_x	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	56×56
Conv3_x	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	28×28
Conv4_x	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1\,024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1\,024\end{bmatrix}\times6$	14×14
Conv5_x	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2\,048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2\,048\end{bmatrix}\times3$	7×7
全连接层	1000d_fc Softmax	101d_fc	1×1

2.3　网络损失函数

网络的损失函数 $L(\theta)$ 如式(4)所示:

$$L(\theta)=\sum_{i=1}^{N}\|F_c(X_i,\theta)-\mathrm{Crowd}_i\|^2 \tag{4}$$

其中,N 为训练图片张数,X_i 表示输入的图片,θ 是模型训练得到的参数,则模型预测出的图片中的拥挤率可以表示为 $F_c(X_i,\theta)$,Crowd_i 表示总网络真实值,即数据标记的拥挤率。参照欧氏距离[21]损失函数,并将其作为目标函数,用于连续值训练样本的拟合。

3　实验结果与分析

本文实验分为 2 个部分,前半部分以生成注意力图为目的,采用的是弱监督学习方式,即给出的一张图像里面包含哪些类别,而不需要完整标记出目标位置、轮廓等信息。后半部分目标则是通过网络模型的训练,回归密度图人群的拥挤率。输入是经过注意力图处理过的原图,真实值是人工计算给出的人群拥挤率,计算方法将在 3.2 节中给出,网络学习率采用的是 Adam[22],并设初始步长为 1e-7,batch_size 设为 4。Adam 是通过计算梯度的一阶距和二阶距估计而为不同的参数设计独立的自适应学习率,以便模型收敛后自动调节步长,比传统的随机梯度下降法更具高效性。

网络整体代码采用的是 Tensorflow 的框架。本文对 NS 数据集的 3 个场景分别进行训练,利用数据集的 80% 作为训练集,剩下的 20% 作为测试集。采用的 GPU 为 2 个 NVDIA GeForce GTX TITAN X。

为验证注意力模块是否同理论上一样能加快训练收敛速度以及增加精确度,本文将在 3.4.1 节中对有无注意力模块进行对比实验。并且在本文的数据集上,实验了已有的开源方法(3.4.2 节),证明了本文模型在狭小空间的场景中较其他方法各方面的性能都有提升,为防止本文模型在 NS-DATASET(Narrow Space)上产生数据依赖,参考文献[4]中的 6 个公共数据集,在3.4.3 节中展示了模型在公共数据集上的结果。

3.1　数据集的设计

本文构建一个新的数据集 NS-DATASET,该数据集中共计 17 800 张图片,它们都是在狭小空间内,视角受到局限如图 3 所示,图 3(a)为斜上方视角,如楼道、天桥隧道,图 3(b)为正上方视角,如车厢的下车门通道,图 3(c)为正前方视角,如前后车厢。

(a)场景1　(b)场景2　(c)场景3

图 3　NS-DATASET 图片不同视角展示

Fig. 3　NS-DATASET pictures displayed from different perspectives

3.2　网络真实值

本文数据集是全新的,需要分别对每个场景进

行训练,将人工标记的拥挤率作为模型真实值,如表 2 所示。拥挤率计算如下:设该场景最大容纳人数为 N,当前人数为 n_i,则第 i 张图片的拥挤率 Crowd_i 表示为:

$$\text{Crowd}_i = \frac{n_i}{N} \tag{5}$$

表 2 网络模型真实值

Table 2 Real vaule of network model

图片	当前人数	总人数	拥挤率/%
	0	30	0.0
	18	30	60.0
	29	30	96.7

3.3 模型评价指标

参照文献[3-5]的方法,通过 MAE、MSE 评价模型最终的量化人群的准确率。计算公式如式(6)、式(7)所示:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\text{Crowd}_i - C_i| \tag{6}$$

$$\text{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\text{Crowd}_i - C_i|^2} \tag{7}$$

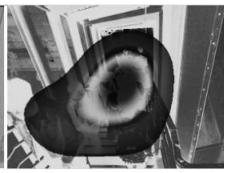
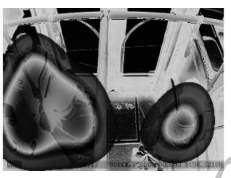
其中,C_i 表示模型预测结果,Crowd_i 表示人数真实值,N 表示统计的图片张数。MAE 强调所有测试集图片的预测拥挤率的平均误差,能更好地反映预测值误差的实际情况,而 MSE 为估计值和真实值差的平方的期望值,可以评价数据的变化程度,MSE 越小,表示数据精确度越高。

3.4 结果对比

表 3 为抽取的实验结果,分别从 NS 数据集的 3 个场景中随机选取一张图片。从表 3 可以看出,本文的注意力模型能准确地关注人群区域(注意力图中的色圈区域),并且模型估计的人群拥挤率大体上和本文给出的真实值一致,能正确地反映当前实时的人群拥挤度。

表 3 实验结果对比

Table 3 Comparison of experimental results %

输入图片	注意力图	真实值	估计值
		78.7	71.3
		56.0	60.4
		28.5	36.4

NS 数据集上模型的测试结果如表 4 所示,将数据集 20% 模型未见过的图片作为测试集,由 MAE、MSE 这两项指标可以看出,大体上 3 个场景预测的人群拥挤率误差都在正常范围内,训练图片最多的场景 3 效果最佳。

表 4 NS 数据集模型上性能评价

Table 4 Performance evaluation on NS dataset model

场景	训练图片数量	测试图片数量	MAE/%	MSE/%
场景 1	4 000	1 000	8.3	17.9
场景 2	2 240	560	11.7	21.6
场景 3	8 000	2 000	6.2	9.8

3.4.1 有无注意力机制实验对比分析

由图 1 网络整体结构所示,网络前端是利用 attention 模块将输入图像分为人群和背景,生成更关注人群目标的注意力图,然后将结合了注意力图和原始图像的特征图作为主干网络的输入。去掉注意力机制,直接将输入图片作为 ResNet 网络的输入,同样能得到一个预测的人群拥挤度。本节主要对去掉注意力模块后的网络进行实验,并与原模型结果进行对比来验证网络添加了注意力机制后,是否同理论上一样去除背景噪声,聚焦人群区域,得到更为精准的人群拥挤率。

图 4 灰色曲线是添加了注意力模块训练的损失值迭代曲线,黑色是去除注意力后,只通过 Resnet50 网络训练得到的迭代曲线,由图 4 可知,注入了注意力机制的网络,从 84 轮开始就逐步收敛,而代表未添加 attention 的黑色曲线,到 170 轮 loss 值才逐渐平稳,并且后期 loss 曲线相比灰色下降的很慢,一直处于振荡的状态。2 种方法的性能指标如表 5 所示,

可以看到 MAE 降低了 40%,MSE 降低了 34%,其中 MAE、MSE 为 3 个场景的平均值。

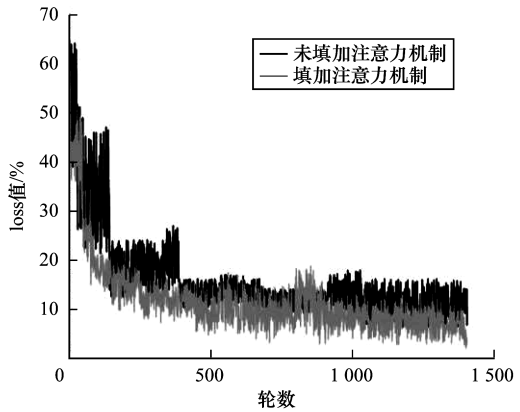


图 4　两种方法 loss 迭代曲线对比

Fig. 4　Comparison of loss iteration curves of two methods

表 5　注意力模块性能对比

Table 5　Performance comparison of attention module　%

方法	MAE	MSE
添加注意力	8.7	16.4
未添加注意力	14.6	25.1

综上可以得出,注入了注意力机制的网络能区分背景和人群,为网络拥挤率的回归提供先验知识,得到更为精确的预测结果。

3.4.2　NS-DATASET 结果对比与分析

本文选取了文献[4]的 CSRNet、文献[3]的 MCNN 以及文献[11]改进的 GoogLeNet 方法作为本文的对比方法。数据集采用的是 NS-ATASET,其中 3 个场景如图 3 所示,本文设计的空间容纳总人数分别为 20 人、30 人和 8 人。MCNN 和 CSRNet 方法产生的是具体人数,而文献[11]的方法产生的是 5 个拥挤等级。为统一评价指标,将 CSRNet、MCNN 的结果分别除以 3 个场景的总人数作为新的结果,文献[11]的方法则通过设置新的真实值重新训练得到拥挤率结果,最终的模型结果评价指标如图 5 所示。

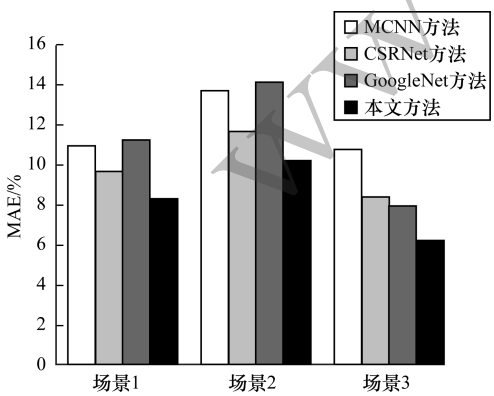


图 5　不同方法在 NS 数据集上 MAE 值对比

Fig. 5　Comparison of MAE values by different methods on NS datasets

在 NS 数据集的 3 个场景上实验了 MCNN、CSRNet、GoogLeNet 3 种方法,并与本文的方法结果进行对比。由图 5 可知,本文方法比以上 3 种方法的 MAE 分别降低了 21.7%、5.23%、30.3%,说明了本文模型在对噪声大、多尺度、人群区域闭塞、分布不均匀的图片处理上相比于其他量化人群方法更有效。

3.4.3　公共数据集上实验对比分析

为研究本文模型在各个公共数据集上预测的人群拥挤度的质量,本文参照文献[11]中的公共数据集进行实验,4 个数据集包含的图片样本数和图片包含的人数如表 6 所示。

表 6　公共数据集组成结构

Table 6　Composition structure of public dataset

数据集	图片数量	平均人数
ShanghaiTech PartA	482	501
ShanghaiTech PartB	716	123
UCF_CC_50	50	1 280
The UCSD	2 000	11 ~ 46

选取 4 个数据集的 80% 作为训练集,20% 作为测试集,并标记了每张图片样本的拥挤率作为网络真实值。参照文献[3-4]中的 MCNN 和 CSRNet 方法在这些数据集上的已有的结果,将其预测的人数误差换算为拥挤率误差来作为本文模型结果的对比。已知其他方法的预测的人数为 n,样本图片中真实人数为 Count,标记的拥挤率真实值为 Gt,则需要的拥挤率 Crowd 可以表示为:

$$\text{Crowd} = \frac{n}{\text{Count}} \cdot Gt \tag{8}$$

在得到每张图片预测的拥挤率后,再利用式(6)计算得到 MAE 对预测结果进行评价,本文的实验结果如图 6 所示。

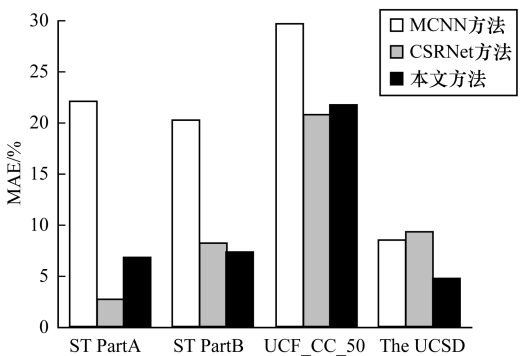


图 6　不同方法在公共数据集上 MAE 值对比

Fig. 6　Comparison of MAE values by different methods on public datasets

从图 6 可以看出,在各个公共数据集上,本文模型比 MCNN 都有更好的效果,但在 ShanghaiTech PartA、UCF_CC_50 平均人数超过 500 人的情况下,模

型效果略微弱于适用于人群高度密集场景的 CRSNet 方法。但在平均人数较少的场景下,ShanghaiTech PartB、The UCSD 方法都比 CRSNet 有着更好的效果。这说明本文方法在其他场景下同样能准确地预测出人群的拥挤程度,并且更加适合狭小的空间内的场景。

4 结束语

本文提出一种注入注意力机制的网络用于分析狭小空间内的人群拥挤度。通过引入注意力模型并结合上下文完成图像特征的提取,获得精准的像素级密集特征,去除不相关背景,完成对场景的感知。实验结果表明,该方法在 NS 数据集下能预测给出图片的人群拥挤率,并且加入注意力模型后,提高了网络的收敛速度。此外,在数据的标记上采用弱监督学习,大幅降低了标记难度和工作量。为提高本文模型的场景泛化能力,同时扩大数据集并增加学习样本,下一步将研究如何提高网络的泛化能力与验证注意力模块的可迁移性。

参考文献

[1] WEI Meng. Crowd density analysis based on convolutional neural network [D]. Hefei: University of Science and Technology of China, 2018. (in Chinese)
魏梦. 基于卷积神经网络的人群密度分析[D]. 合肥: 中国科学技术大学, 2018.

[2] SINDAGI V A, PATEL V M. A survey of recent advances in cnn-based single image crowd counting and density estimation [J]. Pattern Recognition Letters, 2018, 107: 3-16.

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

[4] LI Yuhong, ZHANG Xiaofan, CHEN Deming. Csrnet: dilated convolutional neural networks for understanding the highly congested scenes [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 1091-1100.

[5] ZHANG Lu, SHI Miaojing, CHEN Qiaobo. Crowd counting via scale-adaptive convolutional neural network[C]//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. Washington D. C., USA: IEEE Press, 2018: 1113-1121.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 779-788.

[7] XUE Cuihong, YU Yang. Fusing LBP and GLCM for crowd density classification algorithm [J]. Video Engineering, 2015, 39(24): 7-10. (in Chinese)
薛翠红, 于洋. 融合 LBP 与 GLCM 的人群密度分类算法[J]. 电视技术, 2015, 39(24): 7-10.

[8] WANG Yalin. Research on crowd density estimation algorithm based on gray level co-occurrence matrix [D]. Xi'an: Xi'an University of Science and Technology, 2013. (in Chinese)
王雅琳. 基于灰度共生矩阵的人群密度估计算法研究[D]. 西安: 西安科技大学, 2013.

[9] ZHANG Zhe, SUN Jin, YANG Liutao. Dual-views hand gesture recognition using fusion features of multi-convolution layers [J]. Journal of Chinese Computer Systems, 2019, 40(3): 646-650. (in Chinese)
张哲, 孙瑾, 杨刘涛. 融合多层卷积特征的双视点手势识别技术研究[J]. 小型微型计算机系统, 2019, 40(3): 646-650.

[10] HAN Xinyi. Research on crowd density estimation algorithm based on deep learning [D]. Xi'an: Xi'an University of Science and Technology, 2018. (in Chinese)
韩新怡. 基于深度学习的人群密度估计算法研究[D]. 西安: 西安科技大学, 2018.

[11] LI Baiping, HAN Xinyi, WU Dongmei. Real-time crowd density estimation based on convolutional neural networks [J]. Journal of Graphics, 2018, 39(4): 728-734. (in Chinese)
李白萍, 韩新怡, 吴冬梅. 基于卷积神经网络的实时人群密度估计[J]. 图学学报, 2018, 39(4): 728-734.

[12] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 14-48.

[13] GUO Jichang, LI Xiangpeng. A crowd counting method based on convolutional neural networks and density distribution features [J]. Journal of University of Electronic Science and Technology of China, 2018, 47(6): 8-15. (in Chinese)
郭继昌, 李翔鹏. 基于卷积神经网络和密度分布特征的人数统计方法[J]. 电子科技大学学报, 2018, 47(6): 8-15.

[14] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 770-778.

[15] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 3431-3440.

[16] ZHANG H, DANA K, SHI J, et al. Context encoding for semantic segmentation [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 101-109.

[17] ZHOU B, KHOSLA A, LABEDRIZA A, et al. Learning deep features for discrimi native localization [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 2921-2929.

[18] WEI Yunchao, XIAO Huaxin, SHI Honghui et al. Revisiting dilated convolution: a simple approach for weakly-and semi-supervised semantic segmentation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 7268-7277.

(下转第 267 页)

(上接第 260 页)

[19] CHEN L C, YANG Y, WANG J, et al. Attention to scale: scale-aware semantic image segmentation[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Washington D. C., USA: IEEE Press, 2016: 3640-3649.

[20] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 1409-1456.

[21] KODIROV E, TAO T, GONG S. Semantic autoencoder for zero-shot learning[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 4447-4456.

[22] MATTHEWS A L, NOY P J, REYAT J S, et al. Regulation of a disintegrin and metalloproteinase family sheddases ADAM10 and ADAM17: the emerging role of tetraspanins and rhomboids[J]. Platelets, 2017, 28(4): 333-341.

编辑 索书志