改进的基于卷积神经网络的人数估计方法

张红颖*, 王赛男, 胡文博

中国民航大学电子信息与自动化学院, 天津 300300

摘要 估算监控场景中的人数是安防监控的重要任务之一,当人群密集、行人之间存在遮挡时,人数估计较困难。因此,针对密集场景下的人数估计问题,提出了一种改进的基于卷积神经网络的人数估计方法。为了改善摄像透视畸变带来的影响,分别利用深层网络和浅层网络提取人群特征,深层和浅层网络分别设计了不同核大小的卷积层,并将提取到的特征通过一个具备多尺度提取能力的结构进行融合。实验结果表明,改进后的网络模型所获取的人群密度图更加贴近原场景信息,人数估计结果也更加精确。

关键词 机器视觉; 人数估计; 卷积神经网络; 深度学习; 人群密度

中图分类号 TP391 **文献标识码** A **doi**: 10.3788/LOP55.121503

Improved Method for Estimating Number of People Based on Convolution Neural Network

Zhang Hongying*, Wang Sainan, Hu Wenbo

College of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, *China*

Abstract Estimating the number of people in the surveillance scene is one of the important tasks of security monitoring. However it is difficult to estimate the number when the crowd is with clutter and severe occlusion. An improved crowd counting method based on the convolution neural network is proposed as for the number estimation under dense scenes. In order to reduce the effect of camera perspective distortion, the deep network and shallow network are used to extract the crowd characteristics, respectively. The convolution layers with different kernel sizes are also designed. Moreover, the extracted features are fused through a special structure with multi-scale extraction capability. The experimental results show that the crowd density map obtained by the improved network model is closer to the original scene information and the obtained prediction results are more precise.

Key words machine vision; number of people estimation; convolution neural network; deep learning; crowd density

OCIS codes 150.1135; 100.4996

1 引 言

近年来,由于人群聚集而引发的安全问题引起了社会的普遍关注,公共集会场合(如演唱会、大型赛事、宗教集会现场)的人群安全问题使得对人群密度和人数估计的需求越来越多。同时,随着计算机运算速度的提升和视频图像处理系统的发展,机器视觉已逐渐成为人工智能领域最重要的研究课题之一,采用计算机视觉的方法对公众场合进行人数估计成为近年来该领域的热门研究方向[1-3],具有广泛的应用前景。

传统的人数估计算法主要从特征表达入手,基于某种或某几种特征回归得到人群密度或人群数量[4-5]。文献[6]对加速稳健特征(SURF)特征点进行聚类构建人群特征向量,并用支持向量回归机对高密度人群数量进行统计。文献[7]结合了尺度不变特征变换(SIFT)特征点、傅里叶分析、小波变换、灰度共生矩阵(GLCM)特征、头部检测器等多元特征来估计人群人数。这些传统的方法往往依赖于前景提取的准确性和特征设计的有效性,容易受到光照变化、摄像透视以及人群遮挡的影响。随着深度

收稿日期: 2018-05-14; **修回日期**: 2018-06-14; **录用日期**: 2018-07-05

基金项目: 国家自然科学基金民航联合研究基金重点项目(U1533203)、中央高校基本科研业务费项目中国民航大学专项资助(3122017005,3122018C004)

* **E-mail**: carole_zhang0716@163.com

学习技术的发展，一些学者开始将深度学习的理论应用到人数估计中，在场景适应能力方面取得了很好的效果。Zhang 等[8]利用卷积神经网络(CNN)自动提取图像中人群密度特征来估算不同场景的人数，但其在训练和测试中均使用透视矩阵，而很多场景的相机透视关系很难获得。文献[9]利用深层和浅层网络相结合的方式组成 Crowdnet 网络，以深层网络获得高层语义信息，浅层网络获得低层特征，将两个网络结合，在不使用透视矩阵的情况下，能较好地改善摄像透视带来的影响，但其只是将两个网络的输出平均相加，并没有考虑权重问题，特征表达能力不足。

鉴于 CNN 在人数估计中的良好效果，本文提出了一种改进的基于卷积神经网络的人数估计方法。利用深层网络和浅层网络相结合的方式提取图像特征，将提取到的深层特征和浅层特征经过具有多尺度提取能力的结构进行融合。改进算法不需要前景分割，也不需要对图像进行一系列的预处理，通过网络的学习能力自主提取特征，有效实现密集场景人数估计。

2 基于卷积神经网络的人数估计方法

卷积神经网络提供了一种端到端的学习模型，通过一系列的卷积运算自动学习图像中的人群特征，并通过反向误差传播不断修正特征参数。基于卷积神经网络的人数估计方法主要分为以下 4 个步骤：1) 准备训练网络所需的人群图像数据集，划分训练集和测试集；2) 构建网络模型；3) 用训练集数据训练网络，调整优化网络参数；4) 测试集测试网络模型，并得到人数估计结果。

2.1 网络结构设计

通常情况下，为了获得更大的监控视野，相机安放角度为倾斜向下，但这样会造成人群图像不同程度的透视畸形，具体表现为“远小近大”。距离相机近的人占据的面积大，有更加精细的特征，而距离相机远的人占据的面积小，只有模糊的头部团块。为了改善透视畸形带来的影响，基于文献[9]，设计了一种深浅结合的卷积神经网络(DASCNN)模型，其中深层次网络用来获取人体面部等高层语义信息，而浅层次网络用来提取模糊头部团块之类的低层特征。与文献[9]不同的是，改进算法将深浅网络各自获得的输出特征通过一个具有多尺度提取能力的结构相融合，避免了直接对特征图进行全局平均而产生的特征损失，使特征能够得到选择和融合。DASCNN 网络结构如图 1 所示。

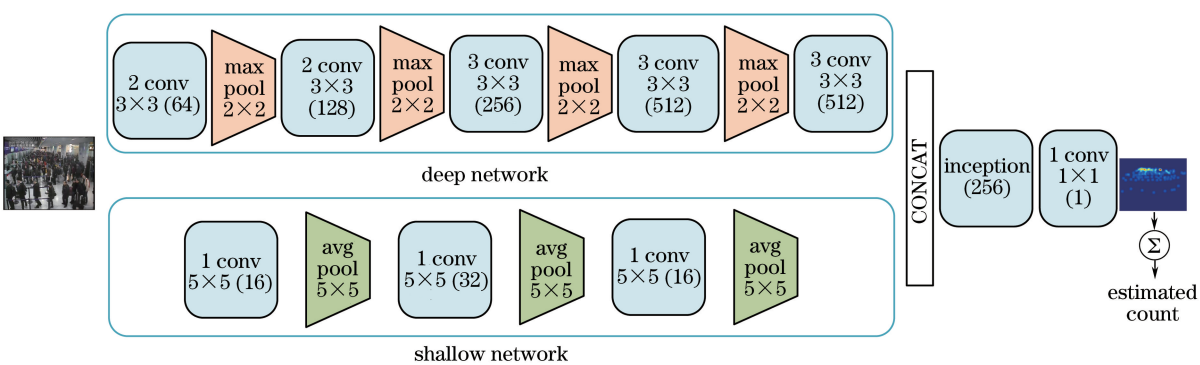

图 1 DASCNN 网络结构

Fig. 1 Architecture of DASCNN

1) 深层网络

视觉几何模型 VGG-16 模型是典型的深层学习模型，其结构简洁，能较好地保存图像的局部位置信息，良好的拓展性使其迁移数据的泛化能力较强，因此深层网络经过调整 VGG-16 模型得到。

VGG-16 结构模型最初训练的目的是图像分类。在图像分类任务中，每个图像有一个分类标签，而要得到人群密度图需要对每一个像素点做预测，因此使用一个全卷积层来替代原来 VGG 网络的全连接层。VGG-16 的网络模型有 5 个最大池化层，每个层的移动步长为 2，输出特性的大小只有输入图像的 1/32。由于目标人群密度图太小会导致分辨率过低，因此移除了 VGG-16 网络的第 5 个最大池化层并将第 4 个最大池化层的池化步长改为 1，这样输出特征图的分辨率是输入图像的 1/8。当池化步长改变时，使用文献[10]中的方法，以保证感受野不变。

2) 浅层网络

Chu 等[11]通过实验发现，金字塔架构(该网络结构的特征面数目按倍数增加)特征面的 CNN 模

型比每层特征面数目均相同的 CNN 结构更能有效地利用计算资源。Krizhevsky 等[12]采用相邻的池化窗口间有重叠区域的重叠池化框架，与无重叠池化框架相比，其泛化能力更强，更不易产生过拟合。因此浅层网络的设计使用 16-32-16 的金字塔架构和重叠池化的方式，为了和深层网络的输出维度一致，平均池化层的步长选为 2。

3) 深浅结合的网络

深层网络和浅层网络得到的特征图大小一致，可以直接通过 concat 层将特征响应图组合在一起。为了使得到的特征充分融合，使用一个 Inception 结构[13]，得到 256 维的特征图。同时为了与标准人群密度图的维度保持一致，还需要一个 1×1 的卷积层，得到网络最终的输出。其中 Inception 结构如图 2 所示。

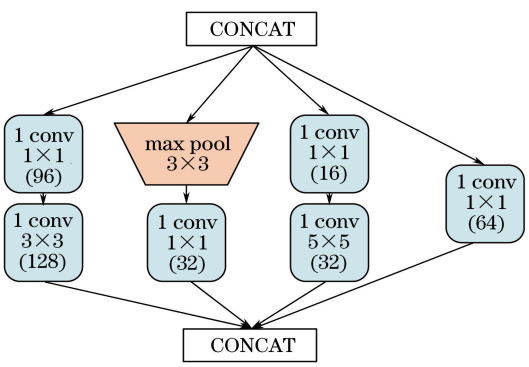

图 2 Inception 结构

Fig. 2 Structure of inception

Inception 结构能够在多尺度上对特征进行提取，这与深层网络和浅层网络的卷积核大小不同相对应。

2.2 数据准备

卷积神经网络是典型的有监督学习，数据标注的好坏直接影响模型的训练结果，因此训练数据中密度图的制作质量对整个网络模型的性能有着重要的影响。采用人工标注的方式，在每一张图像中标注出人头的中心位置，以此得到人头位置的散点图。如果在像素 x_i 上有一个标注，则将其表示为一个脉冲函数 $\delta(x,x_i)$。因此一个带有 N 个标注信息的图像可以表示为

$$\boldsymbol{H}(x)=\sum_{i=1}^{N}\delta(x,x_i)。\tag{1}$$

为了将标注的散点图像转化为连续的密度函数，需要用一个高斯核函数 $\boldsymbol{G}_\sigma$ 与 $\boldsymbol{H}(x)$ 卷积，

$$\boldsymbol{F}(x)=\boldsymbol{H}(x)*\boldsymbol{G}_\sigma(x)。\tag{2}$$

通常给定的图像或视频中并没有给出关于相机的确切信息，也很难获得确切的透视矩阵。借鉴文献[7]的做法，利用相邻标注点的欧氏距离作为判断头部大小的标准。假定在一定的范围内，人群分布的密度是均匀的，那么利用与 x_i 相近的 k 个标注之间距离的平均值可以近似地估计出头部区域的大小，利用这个距离去自适应高斯核。

假设给定的图像中标注点为 x_i，与其最近的 k 个标注点的距离表示为 $\{d_1^i,d_2^i,\cdots,d_m^i\}$，平均距离表示为 $\bar{d}_i=\frac{1}{m}\sum_{j=1}^{m}d_j^i$，于是，与 x_i 相关的像素对应场景中头部区域，这个区域的半径和 $\bar{d}_i$ 成正比。这样高斯核的参数 σ 与 $\bar{d}_i$ 成比例。采用 k-dimensional 树的方法来计算 $\bar{d}_i$。则密度函数表示为

$$\boldsymbol{F}(x)=\sum_{i=1}^{N}\delta(x,x_i)*\boldsymbol{G}_{\sigma_i}(x),\tag{3}$$

式中 $\sigma_i=\beta\bar{d}_i$。

图 3 为某机场候检区的人群图像通过标注和自适应高斯核的方式生成的可视化人群密度图，其中图 3(e)为人群密度值和颜色的对应关系示例。

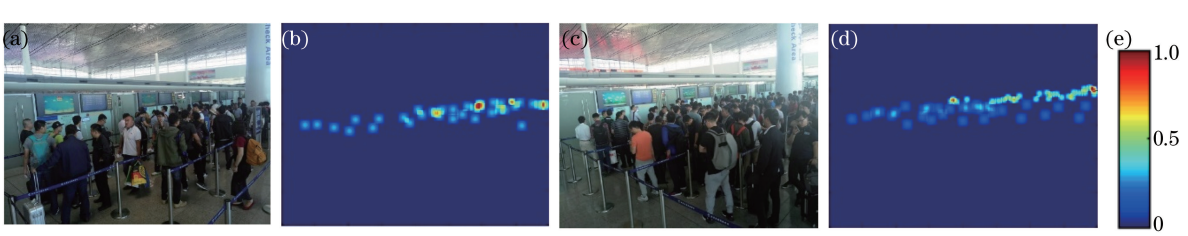

图 3 原图和可视化密度图。(a)图像 1；(b)图像 1 密度图；(c)图像 2；(d)图像 2 的密度图；(e)颜色-密度标尺

Fig. 3 Original images and visual density maps. (a) Image 1; (b) density map of image 1; (c) image 2; (d) density map of image 2; (e) color-density scale

根据一个人头与周围人头之间的平均距离，自适应地调整密度图生成时所用归一化高斯分布内核的大小，对于平均距离较小的人头采用较小的高斯分布内核，而对于平均距离较大的人头采用较大的

高斯分布内核,这种方法适用于人群密集、场景透视矩阵不易获得的情况。

为防止过拟合,采用数据增强的方式来解决训练样本不足的问题。为了提高网络模型对于尺度的适应能力,在数据增强过程中采用构建图像金字塔的方法。将正常大小的图片放在金字塔结构的中间,那么金字塔的上部就相当于图片的收缩,金字塔的底部就相当于对图片的放大。然后进一步对图像金字塔处理后的数据集裁切为 225×225 的图像块并上下左右翻转以得到更多的数据。

2.3 训练网络模型

本文利用卷积神经网络得到人群密度图,通过网络模型输出的预测值与真实的人群密度图之间的差值损失进行网络反向传播更新权值,不断迭代最终完成网络模型训练。计算预测人群密度与真实人群密度之间的差值损失的函数被称为损失函数。损失函数用欧氏距离来计算,对于每一幅输入的图像 $\boldsymbol{X}_i(i=1,\cdots,N)$,总标记人数为 Z_i,损失函数定义为

$$\boldsymbol{L}(\theta)=\frac{1}{2N}\sum_{1}^{N}\left\|\boldsymbol{F}(\boldsymbol{X}_i;\theta)-\boldsymbol{F}_i\right\|^2, \tag{4}$$

式中 θ 为一组从卷积神经网络中学习到的参数,N 为训练图像的总数量,$\boldsymbol{X}_i$ 为输入的第 i 幅图像,$\boldsymbol{F}(\boldsymbol{X}_i;\theta)$为预测得到的密度图,$\boldsymbol{F}_i$ 为标注得到的密度图。

网络的训练中,通过随机梯度下降算法和反向传播算法进行网络优化来训练神经网络,得到最优的网络参数。然而在实际中,由于训练样本的数量非常有限,考虑到梯度消失对深层神经网络的影响,要同时学习所有的参数是不容易的。所以每列卷积都需要先经过预训练,然后再对后面卷积层进行网络微调,避免了训练整个大型网络的计算开销。

网络训练结束后,输入人群图像,经过一系列的卷积和池化运算,最终得到人群密度图,然后对得到的人群密度图进行积分,进而得到图像中估计的人数数值。

$$C_{\text{count}}=\sum_{i=1,j=1}^{i=m,j=n}P_{\text{pixel}}(x_i,y_j), \tag{5}$$

式中 m 和 n 分别为图像的长和宽,$P_{\text{pixel}}(x_i,y_j)$为人群密度图点坐标$(x_i,y_j)$处的像素值大小,$P_{\text{pixel}}(x_i,y_j)\in[0,1]$。

3 实验和分析

实验网络训练的硬件配置情况为 Intel i7、GTX 960。选择人数估计数据集为 CAUC-CROWD、UCF-CROWD[14]、AHU-CROWD[15]。CAUC-CROWD 是国内某机场候检区的旅客图像,UCF-CROWD、AHU-CROWD 是包含各种场景的大规模人群公开数据集。3 种数据集的详细比较如表 1 所示。

表 1 CAUC-CROWD、UCF-CROWD、AHU-CROWD 数据集信息

Table 1 Dataset information of CAUC-CROWD、UCF-CROWD、AHU-CROWD

Dataset	Number of images	Resolution /(pixel×pixel)	Number of people
CAUC-CROWD	225	800×602	7-204
UCF-CROWD	50	Different	94-4543
AHU-CROWD	107	Different	58-2201

为了定量分析改进网络的人群计数能力,采用平均绝对误差(MAE)和平均相对误差(MSE)指标对预测的结果进行评价。MAE 反映了预测的精确度,MSE 反映了预测的稳健性,它们的定义分别为

$$E_{\text{MAE}}=\frac{1}{N}\sum_{1}^{N}\left|G(i)-T(i)\right|, \tag{6}$$

$$E_{\text{MSE}}=\sqrt{\frac{1}{N}\sum_{1}^{N}[G(i)-T(i)]^2}, \tag{7}$$

式中 $G(i)$为图像中的实际人数,$T(i)$为预测得到的人数,N 为总测试图像的数量。

3.1 CAUC-CROWD 数据集

CAUC-CROWD 数据集是在国内某机场各个角度拍摄的机场候检通道人群图像,分辨率为 800 pixel×602 pixel,包含人数为 7~204。利用数据增强后的数据集对网络进行预训练和微调。图 4 为测试集中人群图像真实值和预测结果的比较。图 5 分别为单列网络和结合网络得到的人群密度图的可视化结果。

通过对人群密度图的分析可知,DASCNN 最终得到的人群密度图与原图像中人群所在位置更匹配,且密度表示更相符合,在机场场景的人数估计中效果更好。此外,还将本文方法与其他方法进行对比。其中文献[4]方法为基于特征回归的方法,文献[9]方法为 Crowdnet 卷积神经网络的方法。

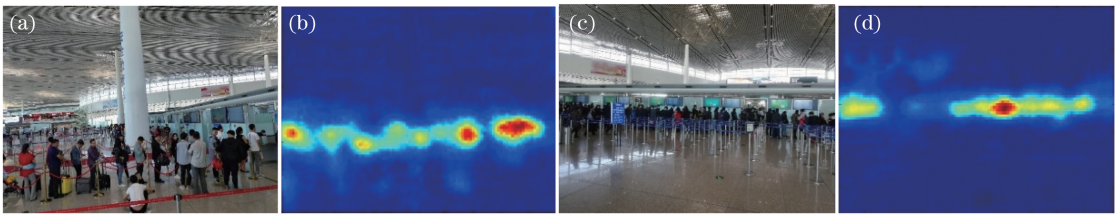

图 4 原图和估计得到的人群密度图。
(a)图像 1 真实值 36 人;(b)图像 1 估计值 31.5 人;(c)图像 2 真实值 22 人;(d)图像 2 估计值 21.7 人
Fig. 4 Original images and estimated crowd density maps. (a) Image 1 with truth value of 36; (b) image 1 with estimation value of 31.5; (c) image 2 with truth value of 22; (d) image 2 with estimation value of 21.7

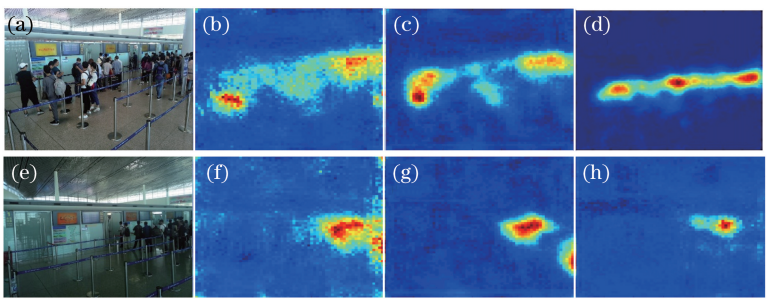

图 5 单列网络和结合网络得到的密度图的对比。(a)图像 1;(b)图像 1 的浅层网络密度图;(c)图像 1 的深层网络密度图;(d)图像 1 的 DASCNN 密度图;(e)图像 2;(f)图像 2 的浅层网络密度图;(g)图像 2 的深层网络密度图;(h)图像 2 的 DASCNN 密度图
Fig. 5 Comparison of density maps obtained by single-row network and combined network. (a) Image 1; (b) density map of image 1 by shallow network; (c) density map of image 1 by deep network; (d) density map of image 1 by DASCNN; (e) image 2; (f) density map of image 2 by shallow network; (g) density map of image 2 by deep network; (h) density map of image 2 by DASCNN

表 2 本文方法与其他算法结果对比
Table 2 Comparison of results by proposed method and other algorithms

Method	MAE	MSE
Method in Ref. [9]	6.53	9.43
Method in Ref. [4]	10.75	15.89
Deep network	6.82	10.71
Shallow network	8.34	11.83
DASCNN	4.49	5.65

从表 2 可以看出,基于卷积神经网络的人数估计方法优于传统的特征提取和回归的方法,尤其是对于机场候检区这种人群密集、人群遮挡严重、存在相机透视畸形干扰的场景,本文改进算法有效降低了误差。另外,本文方法所得人数统计的精确度和稳健性也都优于 Crowdnet 网络模型。

3.2 UCF_CC_50 数据集

UCF_CC_50 数据集中包括 50 张灰度图,包含人数为 94~4543,平均每张图像有 1280 人,包括集会、音乐会、球赛现场等多种场景,场景中存在大量透视变换的现象。为了定性地分析本文方法对人群密度的估计效果,将本文方法得到的密度图可视化结果和文献[9]方法得到的密度图可视化结果进行对比,如图 6 所示。

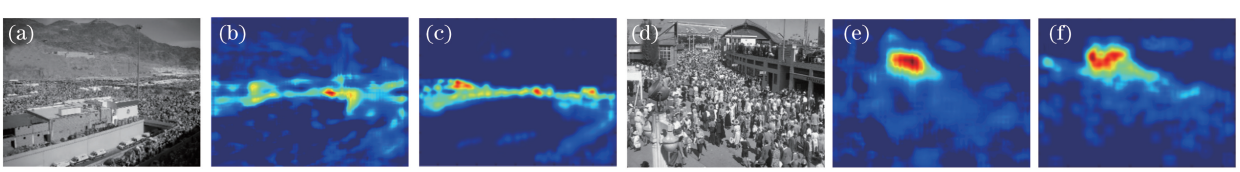

图 6 人群密度估计对比。(a)图像 1;(b)文献[9]预测的密度图;(c)本文预测的密度图;(d)图像 2;(e)文献[9]预测的密度图;(f)本文预测的密度图
Fig. 6 Contrast of crowd density estimations. (a) Image 1; (b) density map predicted in Ref. [9]; (c) density map predicted by proposed method; (d) image 2; (e) density map predicted in Ref. [9]; (f) density map predicted by proposed method

图像1中标注人数为1115,文献[9]方法预测为1143,本文方法预测人数为1132;图像2中标注人数为440,文献[9]方法预测为433,本文方法预测人数为443。相较于文献[9]方法,本文方法预测的人数更加准确,另外从生成的密度图可以看出,本文方法得到的人群密度图与原场景中人群的分布更相近。

为了定量分析本文算法,采用5折交叉验证的方法,与4种经典算法进行比较。其中,文献[16]方法为基于SIFT特征和特定的距离度量的回归方法;文献[14]方法为基于多特征融合的方法;文献[7]方法为经典的CNN方法;文献[9]方法为Crowdnet网络。

表3 本文方法与其他算法结果对比

Table 3 Comparison of results by proposed method and other algorithms

Method	MAE	MSE
Method in Ref. [17]	493.4	487.1
Method in Ref. [14]	419.5	541.6
Method in Ref. [8]	467.0	498.5
Method in Ref. [9]	452.5	—
DASCNN	412.5	523.5

从表3可以看出,本文方法相对于Crowdnet网络,误差降低了8.9%。与其他经典人群计数算法相比,本文方法效果也更优。

3.3 AHU-CROWD数据集

AHU-CROWD数据集中包括107张图像,包含人数为58~2201,属于大规模人群,多俯拍角度,而俯拍角度在机场数据集中比较少,因此本文选择AHU-CROWD数据集中的图像对网络进行微调。采用文献[4,17-18]中统一使用的相对偏差(RD)$R=\frac{1}{N}\sum_{i=1}^{N}\frac{|G(i)-T(i)|}{G(i)}$作为稳健性评价标准。表4为本文算法与其他方法的对比结果。其中文献[4]方法为基于岭回归的方法,文献[17]方法为Haar-like特征结合和Adaboost分类的方法,文献[18]方法为可变形模板方法。

表4 AHU-CROWD人数估计实验结果

Table 4 Experimental results of estimating number of people from AHU-CROWD

Method	MAE	RD
Method in Ref. [4]	207.4	0.578
Method in Ref. [17]	409.0	0.912
Method in Ref. [18]	395.4	0.864
DASCNN	150.3	0.384

从表4可以看出,本文方法依然能够取得较好的结果,表明DASCNN在大规模场景人数估计中具有较强的稳健性。

4 结 论

针对人数估计问题,提出了一种用Inception结构将深层网络和浅层网络结合的卷积神经网络模型,利用深度神经网络提取精细的高层语义特征,浅层网络提取团块等底层语义特征,再用具有多尺度特征提取能力的Inception结构将两个网络提取的特征融合,以此获得更加精准的场景表达。在对比实验中,对深层网络、浅层网络和深浅结合后的网络进行对比,说明了结合后的网络具有更强的特征表达能力;另外在多个数据集中,与其他方法的比较说明本文算法具有较好的稳健性。

参 考 文 献

[1] Yin Z Q, Gu G H, Chen Q, *et al*. Application of three-dimensional depth image in automatic statistic passenger flow system [J]. Chinese Journal of Lasers, 2014, 41(6): 0609003.
尹章芹, 顾国华, 陈钱, 等. 三维深度图像在自动客流计数系统中的应用[J]. 中国激光, 2014, 41(6): 0609003.

[2] Ryan D, Denman S, Sridharan S, *et al*. An evaluation of crowd counting methods, features and regression models [J]. Computer Vision and Image Understanding, 2015, 130: 1-17.

[3] 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.

[4] Chen K, Loy C C, Gong S, *et al*. Feature mining for localised crowd counting[C]// Proceedings of the 23rd British Machine Vision Conference, 2012, 1(2): 3.

[5] Chan A B, Vasconcelos N. Counting people with low-level features and Bayesian regression [J]. IEEE Transactions on Image Processing, 2012, 21(4): 2160-2177.

[6] Liang R H, Liu X D, Ma X Y, *et al*. High-density crowd counting method based on SURF feature [J]. Journal of Computer-Aided Design & Computer Graphics, 2012, 24(12): 1559-1567.
梁荣华, 刘向东, 马祥音, 等. 基于SURF的高密度人群计数方法[J]. 计算机辅助设计与图形学学报, 2012, 24(12): 1559-1567.

[7] 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, 2016: 589-597.

[8] Zhang C, Li H S, Wang X G, *et al*. Cross-scene crowd counting via deep convolutional neural networks[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015: 833-841.

[9] Boominathan L, Kruthiventi S S S, Babu R V. Crowdnet: a deep convolutional network for dense crowd counting[C] // Proceedings of the 2016 ACM on Multimedia Conference, 2016: 640-644.

[10] Chen L C, Papandreou G, Kokkinos I, *et al*. Deeplab semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[11] Chu J L, Krzyżak A. Analysis of feature maps selection in supervised learning using convolutional neural networks [C] // Proceedings of Canadian Conference on Artificial Intelligence, 2014: 59-70.

[12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C] // Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012: 1097-1105.

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv: 1409.1556, 2014.

[14] Idrees H, Saleemi I, Seibert C, *et al*. Multi-source multi-scale counting in extremely dense crowd images [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2547-2554.

[15] Hu Y C, Chang H, Nian F D, *et al*. Dense crowd counting from still images with convolutional neural networks[J]. Journal of Visual Communication and Image Representation, 2016, 38: 530-539.

[16] Lempitsky V, Zisserman A. Learning to count objects in images[C] // Proceedings of International Conference on Neural Information Processing Systems, 2010: 1324-1332.

[17] Jones M J, Snow D. Pedestrian detection using boosted features over many frames[C] // Proceedings of 19th International Conference on Pattern Recognition, 2008: 1-4.

[18] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1-8.