计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：基于空间维度循环感知网络的密集人群计数模型
作者：付倩慧，李庆奎，傅景楠，王羽
收稿日期：2020-05-12
网络首发日期：2020-10-12
引用格式：付倩慧，李庆奎，傅景楠，王羽．基于空间维度循环感知网络的密集人群计数模型[J/OL]．计算机应用.
https://kns.cnki.net/kcms/detail/51.1307.TP.20201010.1550.007.html

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(****)**-0000-00 doi:10.11772/j.issn.1001-9081.2020050623

基于空间维度循环感知网络的密集人群计数模型

付倩慧，李庆奎*，傅景楠，王羽

北京信息科技大学 自动化学院，北京 100192

(*sdlqk01@126.com)

摘 要: 考虑目前对具有透视畸变的高密度人群图像特征提取的局限性，提出了一种融合全局特征感知网络（GFPNet）和局部关联性特征感知网络（LAFPNet）的人群计数模型（LMCNN）。GFPNet 为 LMCNN 主干网络，其输出的特征图进一步序列化并作为 LAFPNet 的输入，利用循环神经网络在时序维度上对局部关联性特征感知的特点，将单一的空间静态特征映射到具有局部序列关联性特征的特征空间，从而有效的削减了透视畸变对人群密度估计造成的影响。为了验证所提模型的有效性，该模型在 Shanghaitech 和 UCF_CC_50 数据集上与算法原子卷积空间金字塔网络（ACSPNet）相比，在 Shanghaitech Part_A 和 UCF_CC_50 数据集上平均绝对误差分别至少减小了 15.9 和 56，均方误差分别至少减小了 30.7 和 86.6。LMCNN 注重空间维度上前后特征的相关性，通过对空间维度特征与单图像内序列特征的充分融合，减少由透视畸变引起的人群计数误差，更加准确的预测密集区域人数，从而提高人群密度回归精度。

关键词: 人群计数；人群密度估计；卷积神经网络；多列卷积神经网络；长短时记忆神经网络

中图分类号:TP391.4 **文献标志码: A**

Dense crowd counting model based on spatial dimensional circular perception network

FU Qianhui, LI Qingkui, FU Jingnan, WANG Yu

(School of Automation, Beijing Information Science and Technology University, Beijing 100192, China)

Abstract: Considering the limitations of the feature extraction of high-density crowd image features with perspective distortion, a crowd counting model that combines global feature perception networks（GFPNet）and local association feature perception networks（LAFPNet）, named LMCNN was proposed. GFPNet was the backbone network of LMCNN, and serializes its output feature map as an input to the LAFPNet. Using the recurrent neural network to sense the local association feature in the time-series dimension, a single spatial static feature was mapped to a feature space with local sequence association features, thus effectively reducing the impact of perspective distortion on crowd density estimation. To verify the effectiveness of the proposed model, experiments were conducted on Shanghaitech and UCF_CC_50 datasets. Compared to the algorithm: Atrous convolutions spatial pyramid network（ACSPNet）, the mean absolute error of LMCNN respectively decreases 15.9 and 56 at least on Shanghaitech Part_A and UCF_CC_50 dataset, and the mean square error respectively decreased by 30.7 and 86.6 at least. LMCNN pays attention to the association between the front and back features in the spatial dimension. By fully integrating the spatial dimension features and the sequence features in a single image, it can reduce the crowd counting error caused by perspective distortion, and more accurately predict the number of people in dense areas, thereby improving crowd density regression accuracy.

Keywords: crowd counting; crowd density estimation; convolutional neural network; multi-column convolutional neural network; long short-term menory neural network

0 引言

人群计数是智能视频监控系统的重要任务之一，目前对高密度人流的公共场合缺乏有效预警措施，易造成人群踩踏等隐患性事件。因此，分析人群行为趋向且对其安全隐患提供有效预警信息的人群计数系统对于建设智慧安全城市具有重要意义[1]。人群计数面临许多挑战，例如遮挡、高度混乱、人员分布不均匀、透视畸变、视角失真等问题。

通常人群计数方法是基于检测或回归。检测方法是通过对个体定位以给出计数[2-3]；回归方法是将人群作为整体研究对象，分析人群分布以建立特征与人数的映射关系，即通过人群特征与人数的映射关系给出计数。基于检测方法，He

收稿日期: 2020-05-12; **修回日期**: 2020-09-18; **录用日期**: 2020-09-19。

基金项目:促进高校内涵发展-研究生科技创新项目(5121911048)

作者简介: 付倩慧(1996—)，女，山东聊城人，硕士研究生，主要研究方向：图像处理、供应链系统；李庆奎(1971—)，男，山东临沂人，教授，博士，主要研究方向：切换时滞系统、供应链系统；傅景楠(1993—)，男，福建莆田人，硕士研究生，主要研究方向：图像处理、深度学习；王羽（1996—），女，北京人，硕士研究生，主要研究方向：图像处理、供应链系统。

等[2]利用 KLT（kanade-lucas-tomasi）跟踪器构建跟踪环节，基于多尺度块局部二进制模式（multi-scale block local binary patterns，MBLBP）模型由对点跟踪转化为对人跟踪，实现实时计数。Gao 等[3]在卷积神经网络（convolutional neural network，CNN）的基础上结合级联 Adaboost 算法作为特征提取器学习人群头部特征，通过对头部检测结果进行计数来获得人群数量。检测方法在人群个体独立且分；；布均匀的简单场景下能实现准确计数，其需要目标具有清晰的轮廓特征。受监控设备的视角、距离以及场景中光照影响，如图 1 所示的大规模、高密度、宽视野的复杂场景，设感受野大小相等：$s_1=s_2=s_3=s_4$，设每个感受野中人数密度为 $d_i\ (i=1,2,3,4)$，由于透视畸变影响导致 $d_1<d_2<d_3<d_4$，即相同大小感受野中所捕捉到的人群数量随着景距的加大而增大，该现象对人群计数任务造成极大困难。

图 1　高密度人群

Fig.1 High-density crowd

随着近年来深度学习技术的飞速发展，目前国内外更趋向于采用基于 CNN 密度回归方法实现计数任务。Zhang 等[4]提出一种基于 CNN 的跨场景人群计数框架，通过人群密度和人群计数两个相互关联的学习目标与数据驱动方法，捕捉人群纹理模型，提高未知场景中人群计数准确率。Zhang 等[5]提出多列卷积神经网络（Multi-column CNN，MCNN）以采用不同尺度卷积核的感受野更充分捕捉人群密度分布特征，在无透视矫正变换矩阵下较为准确地生成密度图。在此基础上，Sam 等[6]提出基于图像中人群密度变化的切换卷积神经网络，利用多个 CNN 回归器的固有结构和功能差异预测人群数量，通过执行差分训练机制来应对大规模和视角差异。Zhang 等[7]提出了尺度自适应网络结构，采用 MCNN 提取特征，并在网络间增加自适应调整机制使之对特征的输出能够自主调节到同一尺寸以回归到人群密度图，在密度回归损失函数中加入基于人数的误差计算，密度图的回归精度随即提高。近年来，深度卷积神经网络在多个计算机视觉研究主题中取得了成功，Pu 等[8]提出基于深层深度卷积神经网络的人群密度估计方法，即引入 GoogleNet 和 VGGNet 实现跨场景人群密度估计方法。随着 CNN 深度的不断增加，其参数量与网络结构的增加，网络训练变得困难，准确度未有效提升。为了进一步消除透视畸变、视角失真导致比例变化对计数准确率的影响，郭等[9]提出基于 CNN 与密度分布特征的人数统计方法，将人群图像依据密度进行划分，高密度图像的透视畸变问题使用多核回归函数处理，稀疏图像中个体分布较均匀，即对场景去噪后基于个体位置研判人群数量。Xu 等[10]提出深度信息引导人群计数方法，即基于图像深度信息将其划分为远景与近景，远景区域基于 MCNN 回归人群密度图，近景区域基于 YOLO 框架检测人群数量，并结合空间上下结构消除误差。Ma 等[11]提出原子卷积空间金字塔网络（Atrous convolutions spatial pyramid network，ACSPNet），递增的原子速率排序的原子卷积核增大感受野并保持提取特征的分辨率，空间金字塔中使用无规则卷积进行多尺度感知且连接集成多尺度语义。陆金刚等[12]提出多尺度多列卷积神经网络（Multi-scale MCNN，MsMCNN）的计数模型，引入多尺度连接以结合不同卷积层的特征学习视角和尺度的变化。马皓等[13]提出了特征金字塔网络人群计数算法，结合高级语义特征和低层特征以多尺度特征回归密度图。郭等[14]提出 GB-MSCNet（Gradient Boosting Multi-Scale Counting Net）人群计数网络，该结构增大网络输出层感知野以保存细粒度信息，同时将多尺度特征融合生成更准确密度图。Zhu 等[15]提出层次密度估计器和辅助计数分类器，引入软注意力机制，基于分级策略从粗到精挖掘语义特征以解决缩放比例变化和视角失真的问题。Wang 等[16]提出基于通道、空间注意力机制的人群计数网络，分别从通道和空间维度自适应的选择不同接受场[17]的特征，适当获取空间上下文信息。

随着研究的不断深入，上述文献对于计数的准确率已取得不错的效果。但是局限于密度区域的划分[9-10]，其需要大量的计算成本确定区域边界和削弱前景对象之间的上下语义依存关系，从而降低模型在复杂场景中的性能；或引入多尺度感受野的卷积层[11-14]，增大模型复杂度且未考虑到图像视角旋转变化，一定程度上限制了模型对视角变化的鲁棒性。因此，受文献[18]的启发提出了一种应用于空间维度上循环感知的人群计数模型（LMCNN）。本文的主要工作如下：

1） 引入空间维度上局部特征循环感知网络，基于 LSTM 单元的顺序编码对图像区域之间的全局空间上下依存关系进行存储，进一步在局部区域学习透视畸变、视角失真的高密度人群内部透视变化信息。

2） 设计端到端特征图分割机制，将主干网络的特征输出自发的进行序列化，从而实现不同维度的特征空间的转换。

1　高密度人群计数模型

如图 2 所示，本文提出了一种应用于空间维度上循环感

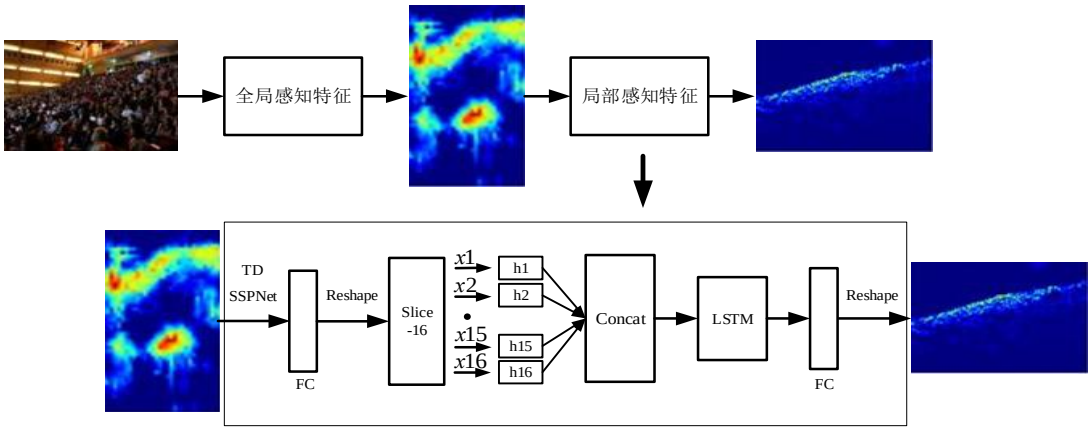


图 2 LMCNN 模型结构

Fig.2 Structure of LMCNN model

知的高密度人群计数模型，由全局特征感知网络（Globalfeature perceptionnetwork，GFPNet）和局部关联性特征感知网络（Local association feature perception network，LAFPNet）两部分组成。为了进一步提高密集人群计数准确性，改善图像透视畸变、视角失真问题，由 GFPNet 有效提取人群图像分布特征，后由 LAFPNet 循环感知特征图序列化的内部空间维度，以进一步在空间维度上迭代提取深层语义特征，最终计算残差值以优化人群图像密度。

1.1 GFPNet

GFPNet 将输入图像经处理后转换为低维特征图，提取深层语义特征。受 zhang 等[5]工作的启发，设计了多尺度工作模式的 GFPNet（如图 3 所示），基于不同尺度的卷积核对图像中不同密度的人群空间分布特征进行提取，即不同尺度的感受野提取不同比例的图像特征。由于透视畸变等问题导致卷积核提取特征的局限性，通过 LAFPNet 基于全局上下相关性以进一步细化密集区域图像特征。如图 3 所示，GFPNet 由三列并行的神经网络组成，每列共有 5 层不同尺寸卷积核和通道数的卷积层，以及两层最大池化层，其中 Conv 层参数 $s*k*k*N$ 中 s 表示卷积核步长，$k*k$ 表示卷积核尺寸，N 表示卷积核通道数；其中 MaxPooling 层参数 $2*2$ 表示为池化区域且池化步长为 2。

三列子卷积神经网络中 Conv-5 层输出的特征图经 Concat 相连接，并且经 Conv-6 层中卷积核尺寸为 1*1 且通道数为 1 的卷积层感知后生成特征图。经两层最大池化处理，其生成特征图的分辨率为输入图像的 $1/4$，大大降低网络的参数量和耦合度，以及优化工作量。

1.2 LAFPNet

LAFPNet 优化人群密度特征图，提出基于 LSTM 单元的顺序编码将区域之间的全局空间上下依存关系存储，该网络主要由空间变换网络[19]、LSTM 单元和残差优化所组成。

其中 $\boldsymbol{TD}$ 为空间位置变换，基于其空间不变性特点，从 GFPNet 输出特征图中突出局部高密度区域，公式如（1）所示。为保证 LAFPNet 能输入任意尺寸特征图，引入了空间金字塔池化层[20]（Spatial Pyramid Pooling Network，SPPNet），避免因不满足输入尺寸要求对特征图进行缩放造成失真等问题。根据 SPPNet 理论，基于网络高度对其特征提取，每层金字塔网络的感受野为 $2^m \times 2^m$，$m=(0,\cdots,n-1)$，考虑到 GFPNet 中有五层卷积层，将金字塔层数设为 5。受限于 Caffe 中 LSTM 单元输入要求为一维向量，Slice 层在特征图的垂直维度上进行分割，将特征图按照其高度均等分割成 16 份，分别为 LSTM 单元 16 个时间步输入。

$$\boldsymbol{TD}=\begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \end{bmatrix} \tag{1}$$

其中 q 为将特征图平移、旋转、裁剪等运算的变化参数。

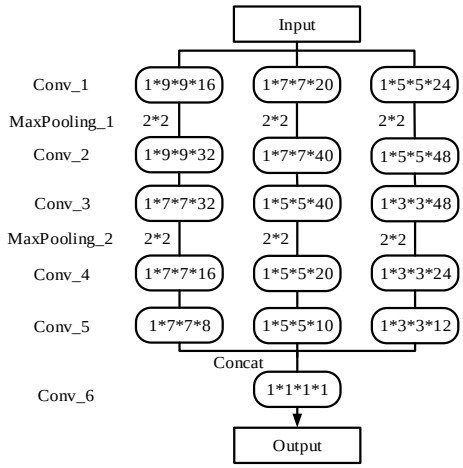


图 3 全局特征感知网络结构

Fig.3 Structure of global feature perception network

1.2.1 LSTM 单元

如图 4 所示，LSTM 单元对一段连续型特征具有较强的记忆性，能够记忆历史特征信息，每个时刻提取的特征都会对下一时刻提取的特征产生影响。因此，LSTM 单元能够感知局部高密度区域空间位置中上下特征的相互依存关系，从而使局部区域人群分布特征的感知更为细化，减弱图像畸变、视角失真对于人群计数影响。该“记忆性”特点由遗忘门、输

入门、输出门三个门控单元控制。遗忘门 $\boldsymbol{f}_t$ 控制之前特征信息对当前的影响，其输入为前一时刻的输出 $\boldsymbol{h}_{t-1}$ 和当前时刻的输入 $\boldsymbol{x}_t$；输入门 $\boldsymbol{i}_t$ 控制信息流的更新；输出门 $\boldsymbol{o}_t$ 控制网络信息流出，分别如下所示：

$$\boldsymbol{f}_t = sig\,\mathrm{mod}(\boldsymbol{W}_{\mathrm{f}} \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{f}}) \tag{2}$$

$$\boldsymbol{i}_t = sig\,\mathrm{mod}(\boldsymbol{W}_{\mathrm{i}} \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{i}}) \tag{3}$$

$$\boldsymbol{o}_t = sig\,\mathrm{mod}(\boldsymbol{W}_{\mathrm{o}} \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{o}}) \tag{4}$$

$\boldsymbol{c}_t$ 表示当前的记忆状态，遗忘门和输入门对其进行更新，其中 $\tilde{\boldsymbol{c}}_t$ 为当前时刻记忆单元的更新值，公式为：

$$\boldsymbol{c}_t = \boldsymbol{f}_t \times \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \times \tilde{\boldsymbol{c}}_t \tag{5}$$

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_{\mathrm{c}} \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{c}}) \tag{6}$$

特征信息经提取后，最终输出状态值 $\boldsymbol{h}_t$ 表示为：

$$\boldsymbol{h}_t = \boldsymbol{o}_t \times \tanh(\boldsymbol{c}_t) \tag{7}$$

其中 $\boldsymbol{W}_{\mathrm{f}}$、$\boldsymbol{W}_{\mathrm{i}}$、$\boldsymbol{W}_{\mathrm{o}}$、$\boldsymbol{W}_{\mathrm{c}}$ 为权重矩阵，$\boldsymbol{b}_{\mathrm{f}}$、$\boldsymbol{b}_{\mathrm{i}}$、$\boldsymbol{b}_{\mathrm{o}}$、$\boldsymbol{b}_{\mathrm{c}}$ 为偏置项。

上述公式详细介绍了 LSTM 单元工作机理。单元细胞基于时间顺序连接，每个时刻提取的特征会对后续提取特征提供信息，不同时刻的特征相互影响，增加后续特征提取的可靠性。

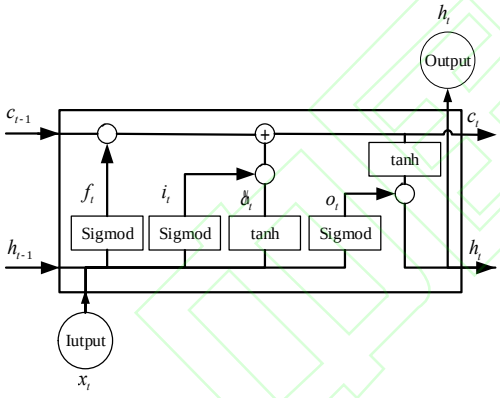


图 4 t 时刻 LSTM 单元结构

Fig.4 Structure of LSTM cell at time t

1.2.2 残差优化

基于密度图回归的方法通常采用随机梯度下降法进行网络优化，以欧氏距离函数进行损失计算，进一步优化结构参数，损失函数定义如下：

$$L(\boldsymbol{\Theta}) = \frac{1}{2N}\sum_{i=1}^{N} \| \boldsymbol{I}(\boldsymbol{X}_i;\boldsymbol{\Theta}) - \boldsymbol{I}_i \|^2 \tag{8}$$

其中 N 代表训练样本的总数，$\boldsymbol{I}_i$ 代表第 i 张训练图片 $\boldsymbol{X}_i$ 对应的真实密度图，$\boldsymbol{I}(\boldsymbol{X}_i;\boldsymbol{\Theta})$ 代表训练输出的密度图，$\boldsymbol{\Theta}$ 代表需要学习的参数，$L(\boldsymbol{\Theta})$ 是训练输出的密度图与真实密度图之间的差距。

1.3 人群密度图

当图像中人群稀疏并且图像中人头尺寸具有一致性时，采用传统高斯核函数方法生成密度图。即假设某个人头中心点在像素 x_i 上，该密度用狄拉克函数来近似表示为 $\delta(x - x_i)$。高斯核滤波器为 $G_\sigma(x)$，人群密度函数为：

$$F(x) = \sum_{i=1}^{n} \delta(x - x_i) \times G_\sigma(x) \tag{9}$$

在高密度人群图像中，人群个体间并非相互独立，由于透视畸变问题，人头在图像中不同区域的尺度不同，则采用自适应高斯滤波器 $G_{\sigma_i}(x)$[12]计算人群密度函数：

$$P(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x) \tag{10}$$

式中 $\sigma_i = \beta \bar{d}^i$ 为高斯核的带宽，$\bar{d}^i$ 为给定图像中人头中心所在的像素点 x_i 与最近的 k 个人头的平均距离，β 为 $\bar{d}^i$ 相对于 σ_i 的权重。经实验，本文 β 设为 0.55 效果较好。

2 实验与分析

为更加充分的验证本模型的有效性，本文实验环境基于 Linux 系统，使用 Caffe 作为训练框架，程序语言为 python 3.5。基于 Shanghaitech 数据集和 UCF_CC_50 数据集进行训练与测试，并与经典、主流的人群计数算法 MCNN[5]、ACSPNet[11]、GB-MSCNet[14]相比较。

2.1 数据集

人群计数任务对数据集的要求一般包括两个方面：不同疏密程度的人群图像以及相对应的人头中心点像素坐标的标签文件。目前人群计数研究所使用的主流开源数据集有 Shanghaitech、WorldExpo'10、UCF_CC_50、UCSD 等，本文实验结果主要基于 Shanghaitech、UCF_CC_50 数据集，其详细信息见表 1 所示：

1）Shanghaitech 数据集

Shanghaitech 数据集是由 zhang 等[5]提出的具有多尺度疏密人群分布特点的大型数据集，该数据集包含两部分：Part_A 和 Part_B。Part_A 是从互联网上随机爬取的 482 张不同场景高密度人群图像，包含 300 张训练图片和 182 张测试图片；Part_B 是从上海市繁华区域采集的 716 张中小密度人群图像，包含 400 张训练图片和 316 张测试图片，是目前使用最多的人群计数数据集。

2）UCF_CC_50 数据集

UCF_CC_50 共有 50 张人群图像，每张图片的人群总数在 94 和 4543 之间，其密度分布差异较大。该数据集数据量较少且人群密度较大，对网络计数准确率挑战极大。

本文中所使用的数据集为 Shanghaitech 和 UCF_CC_50 数据集。考虑到人群图像的样本量过少，容易导致过拟合现象发生，在数据预处理阶段进行数据增强，主要通过裁剪操

作以扩充样本，最后将所有图片按照 8:1:1 随机划分成训练集、验证集和测试集。

表 1 Shanghaitech 和 UCF_CC_50 数据集信息

Tab.1 Information of Shanghaitech and UCF_CC_50 dataset

数据集	数量	分辨率	最少人数	平均人数	最多人数	总人数
Shanghaitech Part A	482	不统一	33	501	3139	241677
Shanghaitech Part B	716	768´1024	9	123	578	88488
UCF-CC-50	50	不统一	94	1279	4543	63974

2.2 网络评价指标

网络评价指标采用平均绝对误差(Mean Absolute Error，MAE)与均方误差(Mean Square Error，MSE)，MAE 反映了网络的预测精度，MSE 反映了网络的泛化能力。其公式如下：

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | n_i - \hat{n}_i | \quad (11)$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (n_i - \hat{n}_i)^2} \quad (12)$$

其中 N 为测试集的人群总数，n_i 为第 i 张测试图片的真实人群总数，$\hat{n}_i$ 为第 i 张测试图片的网络预测人群总数。

2.3 LMCNN 计数性能实例验证

将该模型与文献[5]、[11]、[14]进行对比，其中文献[5]注意到 CNN 中卷积核尺寸固定，同一卷积核卷积运算的感受野大小不变，未考虑透视畸变对于人群密度分布的影响，率先提出 MCNN 结构，采用多列卷积与不同尺寸卷积核学习密度分布特征，为人群计数领域经典模型；文献[11]在此基础上进一步采用原子速率排序的原子卷积结构、金字塔结构的无规则卷积块跳过连接方式和权值与功能共享，在保证分辨率的同时扩大感知范围以集成多尺度特征；文献[14]基于优化 Inception-ResNet-A 模块和 Gradient Boosting 集成学习方法设计端到端的网络结构，采用较大的感受野以融合多个尺度的特征，为目前评价指标较好的模型。该模型与文献[5]、[11]、[14]对 Shanghaitech、UCF_CC_50 数据集进行训练与测试，表 2、表 3 分别展示了不同模型在上述数据集下的训练结果。

2.3.1基于 Shanghaitech 数据集实验结果

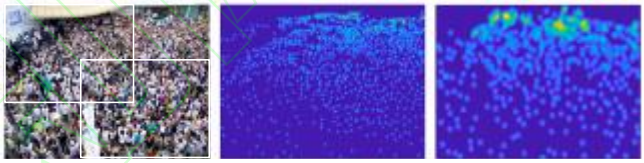
在数据预处理阶段，如图 5、图 6 中(a)所示，将图片随机裁剪相同大小的 4 份，同时保证这 4 份子图相互叠加能够覆盖整张图片。在 Shanghaitech 数据集上，对 Part_A 与 Part_B 分别进行了训练与测试，并与文献[5]、[11]、[14]相对比，结果如表 2 所示：

表 2 基于 Shanghaitech 数据集实验结果

Tab.2 Experimental results based on Shanghaitech dataset

Method	Part_A		Part_B	
	MAE	MSE	MAE	MSE
MCNN[5]	110.2	173.2	26.4	41.3
ACSPNet[11]	85.2	137.1	10.6	16
GB-MSCNet[14]	75.0	119.2	10.4	15.5
LMCNN	69.3	106.4	11.1	14.4

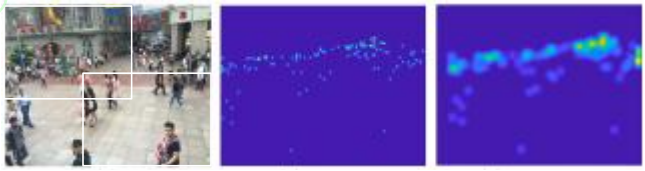
由表 2 来看，LMCNN 模型在稠密度较高的子集 Part_A 上表现优于其它方法，其 MAE 与 MSE 与经典算法 MCNN 相比显著减少，相比于基于 CNN 结构的 ACSPNet 算法分别减少了 15.9 和 30.7，相比于目前较好算法 GB-MSCNet 分别减少了 5.7 和 12.8，说明 LAFPNet 充分学习到了高密度区域特征；但在如图 6(a)所示稠密度较低的子集 Part_B 中，模型 MAE 表现略逊于 GB-MSCNet，Part_B 为低密度图像，背景主要为街道且布局杂乱，LAFPNet 过多提取背景特征，导致 MAE 稍差。图 5 与图 6 分别展示了模型在数据集 Part_A 与 Part_B 随机图像的真实密度图与预测密度图的对比，充分说明网络预测密度图较为准确地反映图片真实密度。



（a）Part_A测试图片 （b）真实密度图 （c）模型预测密度图

图 5 Part_A 测试图片的真实密度图与预测密度图对比

Fig.5 Comparison of real density map and predicted density map of Part_A test picture



（a）Part_B测试图片 （b）真实密度图 （c）模型预测密度图

图 6 Part_B 测试图片的真实密度图与预测密度图对比

Fig.6 Comparison of real density map and predicted density map of Part_B test picture

2.3.2基于 UCF_CC_50 数据集实验结果

在数据预处理阶段，对每张图片都随机裁剪了相同大小的 16 份，并保证这 16 份子图相加能够覆盖整张图片。

在 UCF_CC_50 数据集上进行了训练与测试，并与文献[5]、[11]、[14]相对比，结果如表 3 所示：

表 3 基于 UCF_CC_50 数据集实验结果

Tab.3 Experimental results based on UCF_CC_50 dataset

Method	MAE	MSE
MCNN[5]	377.6	509.1
ACSPNet[11]	275.2	383.7
GB-MSCNet[14]	243.8	366.1
LMCNN	219.2	297.1

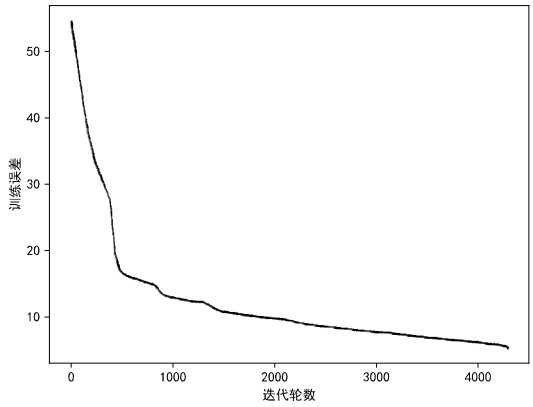


图 7 UCF_CC_50 数据集损失函数曲线

Fig.7 loss of UCF_CC_50 dataset

由表 3 来看，模型在 UCF_CC_50 数据集上的 MAE 和 MSE 较目前较好算法 GB-MSCNet 分别减少了 24.6 和 69，模型性能显著提高，较基于 CNN 结构的 ACSPNet 算法分别减少了 56 和 86.6。该数据集人群密度较大，图像中由于视角失真引起的密度分布变化规律更为明显，充分发挥 LAFPNet 对局部高密度序列特征关联性的感知能力，充分说明 LSTM 单元处理具有明显层次变化信息的图像的学习能力要比仅基于多尺度 CNN 的效果更好。图 7 中展示了在 UCF_CC_50 数据集上训练损失值的变化曲线，从图中可以看出，大约在前 1000 次的迭代中，损失震荡较为强烈，但总体趋势在降低，随着迭代轮数的增加，损失值最终收敛到了较低水平。对于回归任务来说，引进 LSTM 单元在模型准确率上显著提高现有技术水平。

3 结论

本文提出了融合全局特征感知网络与局部关联性特征感知网络的 LMCNN 人群计数模型，该模型以全局感知特征网络为主，基于多模融合策略引入 LSTM 单元，利用 LSTM 单元对序列的关联性特征具有强感知的特点，通过对空间维度特征与单图像内序列特征的充分融合，达到减少由透视畸变和视角失真引起的人群计数误差，较为准确的预测密集区域人数。在 Shanghaitech、UCF_CC_50 数据集上进行实验，结果表明 LMCNN 网络在 UCF_CC_50 数据集上对于高密度人群预测准确率均优于现有算法，说明循环感知单元对于透视畸变、视角失真图像序列特征学习的充分性。但是当人群密度稀疏、背景杂乱时，LMCNN 模型预测能力较弱，主要由于人群个体尺寸变化规律的特征表示程度较小，导致循环感知特征时过多提取背景特征，空间维度上循环感知局部特征的学习能力无法得到很好的提升。在下一步工作中，基于尺度和视角变化特征，进一步提高该模型在低密度人群图像中计数准确性。

1.2 章节中 $\boldsymbol{TD}$ 为矩阵。

参考文献

[1] 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.

[2] HE P, MA W H, HUANG L, et al. Real time people counting system [J]. Journal of Image and Graphics, 2011, 16(5): 813-820.

[3] GAO C Q, LI P, ZHANG Y J, et al. People counting based on head detection combining Adaboost and CNN in crowded surveillance environment[J]. Neurocomputing, 2016, 208: 108-116.

[4] ZHANG C, LI H, WANG X, et al. Cross-scene crowd counting via deep convolutional neural networks [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 833-841.

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

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

[7] ZHANG L, SHI M J, CHEN Q B. Crowd counting via scale-adaptive convolutional neural network [C]// Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2018: 1113-1121.

[8] PU S L, SONG T, ZHANG Y, et al. Estimation of crowd density in surveillance scenes based on deep convolutional neural network[C]// Proceedings of the 8th International Conference on Advances in Information Technology. Piscataway: IEEE , 2017: 154-159.

[9] 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): 806-813.(GUO J C, LI X P. 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): 806-813.)

[10] XU M J, GE Z Y, JIANG X H, et al. Depth information guided crowd counting for complex crowd scenes[J]. Pattern Recognition Letters, 2019, 125: 563-569.

[11] MA J J, DAI Y P, TAN Y P. Atrous convolutions spatial pyramid network for crowd counting and density estimation[J]. Neurocomputing, 2019, 350: 91-101.

[12] 陆金刚, 张莉. 基于多尺度多列卷积神经网络的密集人群计数模型[J]. 计算机应用, 2019, 39(12): 3445-3449.(LU J G, ZHANG L. Crowd counting via multi-scale multi-column convolutional neural network[J]. Journal of Computer Applications, 2019, 39(12): 3445-3449.)

[13] 马皓, 殷保群, 彭思凡. 基于特征金字塔网络的人群计数算法[J]. 计算机工程, 2019, 45(7): 203-207.(MA H, YIN B Q, PENG S F. Crowd counting algorithm based on feature pyramid network[J]. Computer Engineering, 2019, 45(7): 203-207.)

[14] 郭瑞琴, 陈雄杰, 骆炜, 等. 基于优化的 Inception ResNet A 模块与 Gradient Boosting 的人群计数方法[J]. 同济大学学报(自然科学版), 2019, 47(8): 1216-1224.(GUO R Q, CHEN X J, LUO W, et al. A Method of Crowd Counting Based on Improved Inception-ResNet-A Module with Gradient Boosting[J]. Journal of TongJi University (Natural Science), 2019, 47(8): 1216-1224.)

[15] ZHU M, WANG X Q, TANG J, et al. Attentive multi-stage convolutional neural network for crowd counting [J]. Pattern Recognition Letters, 2020, 135: 279-285.
[16] WANG S Z, LU Y, ZHOU T F, et al. SCLNet: spatial context learning network for congested crowd counting [J]. Neurocomputing, 2020, 404: 227-239.
[17] LIU W Z, SALZMANN M, FUA P. Context-aware crowd counting [C]// Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 5099-5108.
[18] WANG Z X, CHEN T S, LI G B, et al. Multi-label image recognition by recurrently discovering attentional regions[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 464-472.
[19] MAX J, KAREN S, ANDREW Z. Spatial transformer networks[C]// Advances in Neural Information Processing Systems. Piscataway, NJ: IEEE Press, 2015: 2017-2025.
[20] LI L L, YANG Z Y, JIAO L C, et al. High-Resolution SAR Change Detection Based on ROI and SPP Net[J]. IEEE Access, 2019, 7: 177009-177022.

This work is partially supported by the development of university content--Postgraduate science and technology innovation project (5121911048).

FU Qianhui, born in 1996, M. S. candidate. Her research interests include image processing.
LI Qingkui, born in 1971, Ph. D., professor. His research interests include switching time-delay system.
FU Jingnan, born in 1993, M. S. candidate. His research interests include image processing, deep learning.
WANG Yu, born in 1996, M. S. candidate. Her research interests include image processing.