# 人群密度估计综述

人群计数（Crowd Counting），又称人群密度估计。

人群计数在实际场景中的应用并不广泛。大多数方案还是在公开数据集上做实验，部分公司有商业使用，但是属于商业机密，还无法得知具体情况。

高密度人群聚集容易发生各种意外事件、所以监控与分析高密度人群，防止意外事件发生，具有重要的现实意义，分析高密度人群其中一个最重要的参考就是人群数量、评估聚集人群的数目、分布方式，有利于实时分离与管控，防止意外发生。

## 难点

人与人之间的类内遮挡。

人和背景之间的相似性。

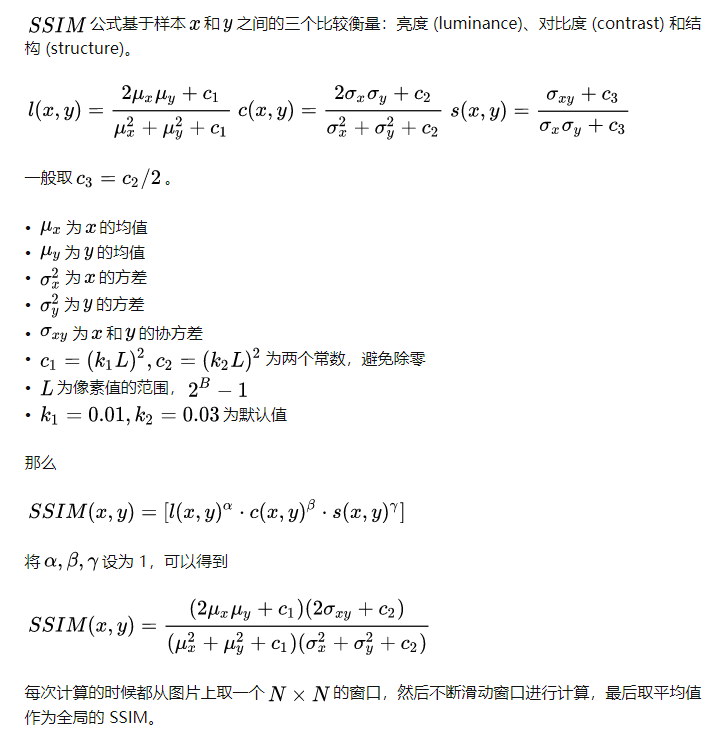
摄像头的广视角。

是二分类问题还是回归问题。

## 评价指标

SSIM

使用SSIM衡量估计密度图和真实值之间的一致性/相似性，计算三个局部统计值，均值方差协方差。SSIM的值从-1到1，两个图片相同时值为1。使用11×11的正则化高斯核函数标准差维1.5来估计局部统计。使用卷积层实现，设定权重不变。这是一个类似于卷积的操作。



回归方法的评价标准

平均绝对误差(MAE)和均方误差(MSE)，前者表征算法估计的准确性，后者表征算法估计的稳定性。

## 数据集

### 生成groundtruth

每一张人群图像中的每个人头所在近似中心位置的坐标作为人工标注，根据已知的每个人头位置，再估计该位置所在人头的大小，这样可以得到该人头的覆盖区域，通过一种方法（MCNN中采用几何自适应高斯核），将该区域转化为该区域内可能为人头的概率，该区域概率和为1（或者表示每个像素可能有多少个人）。最终我们可以得到一张人群密度图。

高斯核来估计图像中每个人头的大小并转换为密度图。一般都是二维高斯核，sigma为高斯核标准差，beta是个设定值取0.3。各个人头区域的概率之和为1，得到对完整人群图像的密度图后，对其进行积分（求和）就是人群数目了。

常用的数据集生成脚本：

https://github.com/gaoguangshuai/survey-for-crowd-counting

https://github.com/leeyeehoo/CSRNet-pytorch/blob/master/make\_dataset.ipynb

### 常用的数据集

UCF CC 50人头标注数据集

UCSD 行人标注数据集

GCC Dataset

Fudan-ShanghaiTech Dataset

Venice Dataset

UCF-QNRF Dataset

ShanghaiTech Dataset

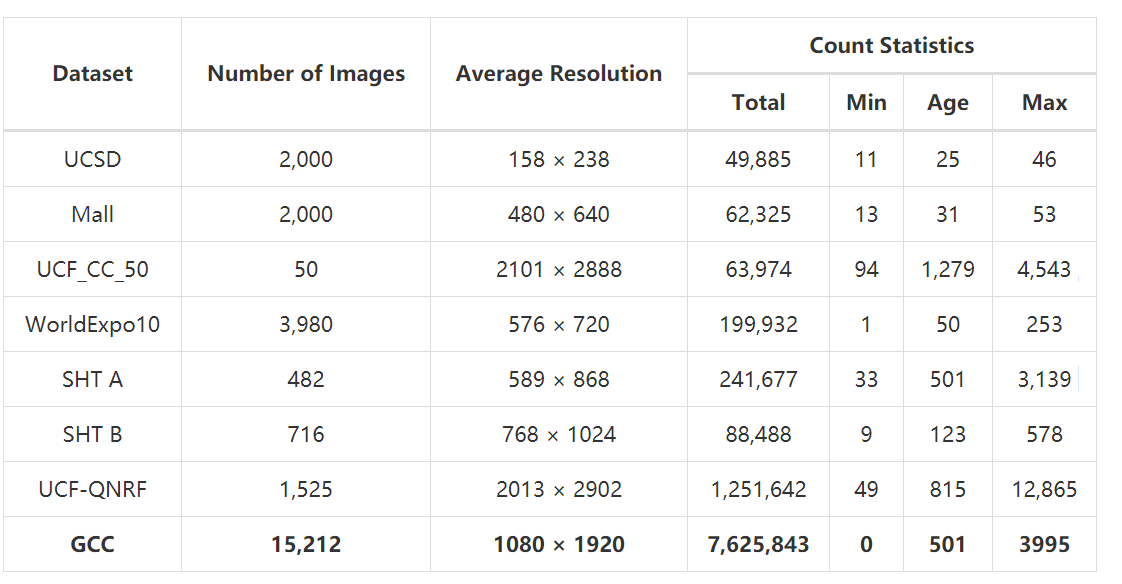
WorldExpo'10 Dataset

Mall Dataset

SmartCity Dataset

AHU-Crowd Dataset

JHU-CROWD Dataset（2020年最新的数据集）



## 主流算法

### 传统特征提取方法VS深度学习的方法

传统特征提取方法

早期的密集人群技术主要是基于人头计数与行人计数、人脸、身体等部位、这些方法都基于SIFT、HAAR、HOG等特征传统的图像特征提取技术、这些方法在面对遮挡、密集人群的时候常常失灵、无法较准确的统计评估出密集人群，特别是对大规模的密集人群，如超过2000+以上的拥挤人群，很难正确计数。

深度学习的方法

另外一种现在比较流行的方法是生成人群heat map（热力图），然后人群计数就变成对热力图的积分计算。还可以计算每平方米的行人密度与聚集程度，这种方法是伴随着深度学习到来逐渐流行起来，使用卷积神经网络的方法基本上都是在这个思路上面进行各种拓展。

流程：图片-》backbone-》回归（密度估计）-》双线性上采样（尺寸相同）。

### 四种主流思路

人群计数的通常的方法大致可以分为四种

detection-based methods

regression-based methods

tracking-based methods

density estimation-based methods

1. 基于目标检测的方法

这类方法就是通过对图像上每个行人或这人头进行定位与识别，再根绝结果统计人数。优点在于可以做到准确的行人或者人头位置(人头检测)，但缺点在于对高密度的人群图像来说，其检测效果差。

这种方法比较直接，在人群较稀疏的场景中，通过检测视频中的每一个行人，进而得到人群计数的结果，一般是用基于外观和运动特征的boosting，贝叶斯模型为基础的分割，或集成的自顶向下和自底向上的处理，这种方法在人群拥挤情况下不大奏效，需要运用到基于部件模型（如DPM）的检测器来克服人群拥挤遮挡的问题。

1. 基于回归的方法

这类方法可以叫做人群数目估计，没有精确定位行人位置，而是对大概的人群数目给出个估计值；优点在于对高密度人群图像来说，其效果是比基于目标检测方法的好，但缺点没有精确的定位。

1. 视觉特征轨迹聚类

对于视频监控，一般用KLT跟踪器（即Lucas光流法）和聚类的方法，通过轨迹聚类得到的数目来估计人数。

1. 密度图回归

每一张人群图像中的每个人头所在近似中心位置的坐标作为人工标注，根据已知的每个人头位置，再估计该位置所在人头的大小，这样可以得到该人头的覆盖区域，通过一种方法（MCNN中采用几何自适应高斯核），将该区域转化为该区域内可能为人头的概率，该区域概率和为1（或者表示每个像素可能有多少个人）。最终我们可以得到一张人群密度图。

基于CNN的密度图估计多数采用了 multi-column based architecture (MCNN) 的架构，这种结构存在几个问题：

* Multi-column CNNs 比较难训练
* Multi-column CNNs 引入了冗余的网络结构
* 需要 density level classifier ，这样计算量比较大
* 这些网络用了很大一部分参数用于 density level classification，用于密度图估计的参数占小部分。

## 一些trick

数据增广

先在原始图像上随机找100个位置，每个位置裁剪出1/4尺寸的局部图像。然后用水平翻转和加入噪声的方法创建另外200个局部图像。

根据经验，如果是from scratch training，对于这几个数据集建议采用多batch size训练或者采用GCC-SFCN中加padding的方案，对于有预训练参数的模型（AlexNet，VGG，ResNet等），建议采用单一batch size进行训练。

Q：语义分割和人群计数非常类似，能不能直接用一些分割网络呢？

A：二者同属于逐像素任务，前者为逐像素分类，后者为逐像素回归。根据我的实验，某些分割网络直接修改最后一层为回归层后，其效果与backbone相比，提升非常有限。甚至性能会有所下降。深层问题暂时还没有仔细思考。不过据我所知，有人对此问题已经做了研究，大家耐心等待即可。"语义分割和人群计数的任务十分接近，人群计数中引入FPN这种特征金字塔将1/8的密度图输出转换到和输入相同的大小，是否会比直接上采样更好？

有个问题想请教一下，为什么resnet上效果很好，但现在大多数计数模型都是以vgg为backbone呢？

VGG模型简单好用，ResNet一般用作刷榜使用。

在人群计数领域，回归问题在深层网络往往会出现梯度消失的问题，如果有新增的很多模块（需要随机初始化的层），模型将很难训练。我们的res50、101模型仅仅新增两层，受这种影响不大。如果设计过深的模块，模型将会很难训练。一般我们写论文会将重心放在验证自己的insight上，不会过分关注于调参。因此大家使用VGG居多。这在其他很多领域也是如此。

LabelNormalize，即放大density-map上的所有像素值，这能提升精度。