上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

智能视频分析

人数统计的多种实现

电子系

xx

019xxxxxxxxx

导师：xx

2020 年 6 月

摘 要

本次实验中，我选择了第七次课的大作业，使用所学方法进行人数统计的实验。我主要使用了深度学习的方法，选用了基于目标检测的 YOLO v3 模型，基于多列卷积神经网络的 MCNN 模型，利用检测方法进行人群计数的 LSC-CNN 模型，以及多任务的视觉语言联合预训练的 12-in-1 模型，使用 4 种不同的模型分别对低密度人群和高密度人群两段监测视频进行了人数统计实验。

首先，我在 CDNET 和 Videezy 网站上分别选取了一段低密度人群和一段高密度人群的视频，然后分别使用 YOLO v3 和 MCNN 两种模型对两段视频进行了计数实验，发现 YOLO v3 的低密度人群计数效果更好，而 MCNN 在高密度人群计数中表现更佳。接着，我使用了较新的利用检测方法进行人群计数的 LSC-CNN 模型对两段视频做了实验，发现在高密度人群计数中效果得到了提升。然后我还将人群计数任务联系到了 VQA 任务，并使用了最新提出的视觉语言联合预训练的 12-in-1 模型进行了实验，发现在低密度人群计数中效果理想。最后，我对以上的实现方法进行了比较总结，进一步加深了对人群计数算法的理解。

关键词：人群计数；深度学习；多种计数模型；高低密度人群

1 理论基础

随着人口的稳定增长和世界城市化进程的推进，国内外的大型活动中频发踩踏事件，已经造成了不小的伤亡，如 2015 年上海外滩踩踏事件,已达到了我国规定的重大伤亡事故级别。因此，人群计数问题的研究也越来越火热，若能通过准确估计当前场景的人群密度，并安排相应的安保措施，则可以有效减少或避免此类事件的发生。早期传统的人群计数算法主要分为基于检测的方法、基于轨迹聚类的方法，以及基于回归的方法，近些年随着深度学习的快速发展，基于深度学习的计数算法也因其出色的特征学习的能力被广泛应用于人群计数任务当中，而相应的数据集也接踵而至，如 ShanghaiTect A/B，UCSD，Expo2010，Mall，UCF-CC-50 和 UCF-QNRF 等。

1.1 传统的人群计数算法[1]

1.1.1 基于检测的方法

早期的人群研究主要聚焦于基于检测的方法，比如使用一个滑动窗口检测器来检测场景中人群，并统计相应的人数。基于检测的方法主要分为两大类，一种是基于整体的检测，另一种是基于部分身体的检测。基于整体的检测方法，最直观、最直接的方法就是人体检测，通过一组行人图像训练出一个行人检测器，利用从行人全身提取的 Haar 小波，梯度方向直方图 HOG，边缘等特征去检测行人，行人检测中常用的分类器主要有 SVM, boosting 和 随机森林等方法，基于整体检测的方法主要适用于稀疏的人群计数。随着人群密度的提升，人与人之间的遮挡变得越来越严重，此时就要引入基于部分身体检测的方法，部件检测能在一定程度上解决人群密集情况下的人数统计的问题，当人与人之间存在部分遮挡时，部件模型同样有效，通过检测身体的部分结构，例如头，肩膀等去统计人群的数量，可以有效实现密集人群的人群计数。

1.1.2 基于轨迹聚类的方法

轨迹由随时间连续变化的位置序列构成，特征点是指运动或状态发生明显变化的位置，通过特征点可以将轨迹划分为若干个子轨迹。轨迹数据包含丰富的语义，基于轨迹聚类的方法依赖于假定个体运动场或视觉特征相对一致，因此相干的特征轨迹可以被聚合到一起表示移动的个体。如基于非监督的贝叶斯聚类方法跟踪局部特征，并将其聚合成簇，利用 KLT（Kanade Lucas Tomasi）跟踪器来获取一组丰富的低级跟踪特征，然后通过对轨迹聚类来推断监控区域中的人数。通过轨迹聚类可以保留原有的时空语义特性，能更全面反应人群的运动和行为模式。

1.1.3 基于回归的方法

基于检测的方法难处理人群之间严重的遮挡问题，所以，基于回归的方法逐渐被用来解决人群计数的问题。基于回归的方法，主要思想是通过学习一种特征到人群数量的映射，这类方法步骤主要分为两步，第一步提取低级的特征，第二步是学习一个

回归模型。通过回归模型求出人群特征与人数之间的函数或利用分类器将人群特征映射到对应的人群密度等级，如此可以定量或定性地研究人群人数估计。常用的人群特征有前景像素特征、纹理特征和角点特征，分类器有支持向量机 SVM (Support Vector Machine)、反向传播 BP(Back Propagation)神经网络以及自组织映射 SOM (Self Organizing Maps)神经网络等，回归模型有高斯处理回归、线性回归、SVM 回归等。

1.2 基于深度学习的人群计数算法

随着深度学习在计算机视觉领域的发展，人群计数领域也取得了长足的进步。本次实验中我针对基于深度学习的人群计数算法，选取了 4 种不同的模型进行了实验对比，分别是基于目标检测的 YOLO v3 模型，基于多列卷积神经网络的 MCNN 模型，利用检测方法进行人群计数的 LSC-CNN 模型，以及多任务的视觉语言联合预训练的 12-in-1 模型。

1.2.1 YOLO v3 目标检测算法[2]

YOLO v3 是一种 One-Stage 的目标检测算法，这类检测算法不需要 Region Proposal 阶段，可以通过一个 Stage 直接产生物体的类别概率和位置坐标值，One-Stage 的目标检测算法主要包含 YOLO 系列和 SSD 算法。本次实验我采用了 YOLO v3 对高、低密度人群视频进行目标检测，从而完成人群计数，图 1.1 是 YOLO v3 的主要架构。

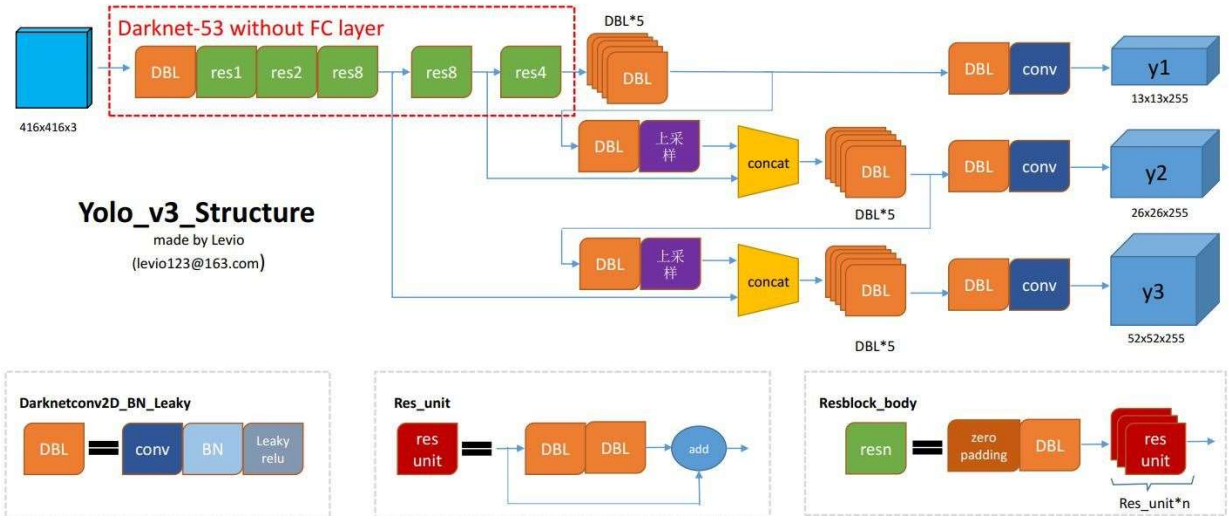

图 1.1. YOLO v3 的主要架构

YOLO v3 使用了 darknet-53 作为特征提取的 backbone，在精度上与 Resnet 相当，在计算速度上却得到了很大的提升。为了加强算法对小目标检测的精确度，YOLO v3 中采用了类似 FPN 的 upsample 和融合做法，在多个 scale 的 feature map 上做检测。此外，在 loss function 种，作者替换了原有的用 softmax 获取类别得分并用最大得分的标签来表示包含再边界框内的目标，而对图像中检测到的对象执行多标签分类，就是对每种类别使用二分类的 logistic 回归。从效果来看，YOLO v3 可以取得和 SSD 同样的精度，速度却提升了 3 倍，是精度与速度兼顾的模型。

1.2.2 MCNN 人群计数算法[3]

Multi-Column Convolutional Neural Network (MCNN)是一种简单有效的多列卷积神经网络结构(MCNN)，可以将图像映射到对应的人群密度图上，发表于 CVPR 2016 会议中。该方法允许输入任意尺寸或分辨率的图像，每列 CNN 学习得到的特征可以自适应由于透视或图像分辨率引起的人/头大小的变化，并能在不需要输入图的透视先验情况下，通过几何自适应的核来精确计算人群密度图，图 1.2 是 MCNN 的主要架构。

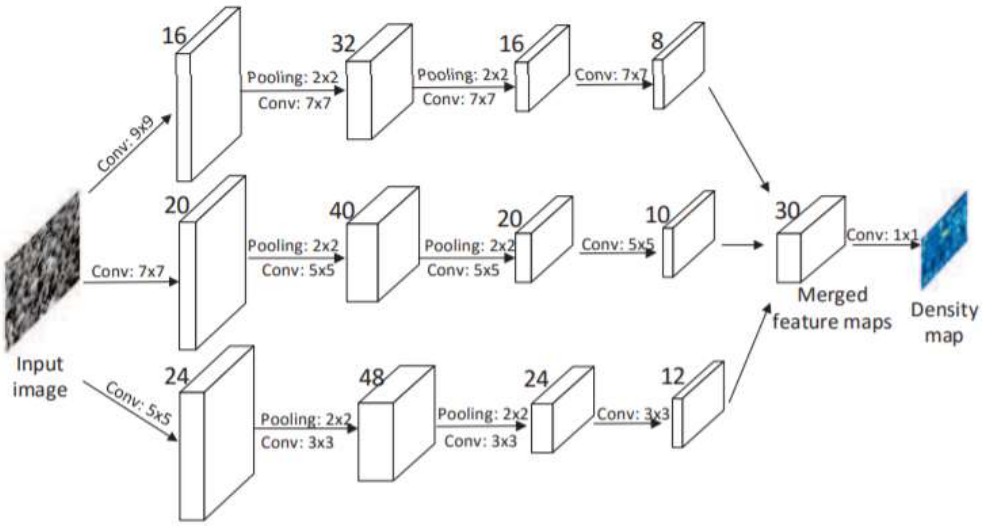

图 1.2. MCNN 的主要架构

由图 2 可以看出，MCNN 网络的每一列并行的子网络深度相同，但是滤波器的大小不同（大，中，小），因此每一列子网络的感受野不同，能够抓住不同大小人头的特征，最后将三列子网络的特征图做线性加权得到该图像的人群密度图，而作者用一个 1*1 滤波器的卷积层代替了完全连接的层，因此模型的输入图像可以是任意大小的，避免了失真。MCNN 的一个优势在于能学习到不同大小人头对应的密度图，因此，如果该模型用一个包含各种大小人头的大数据集来训练，则该模型可以很容易地适应（或迁移）到另一个人头大小是一些特定的尺寸的数据集，如果目标域只包含少量的训练样本，可以简单地将 MCNN 的每一列前几层固定，只微调最后的少量卷积层，这样固定前几层使学习的知识可以被保留，微调后几层很大程度上降低了模型适应目标域的计算复杂度。此外，作者还收集了一个新的数据集 Shanghaitech A/B，用于人群计数方法的评价，该数据集共包含 1198 张图，330,165 个精确标定的人头，比现有的数据集包含更复杂的情况，能更好地测试方法性能。

1.2.3 LSC-CNN 人群计数算法[4]

LSC-CNN是一个利用检测的方法进行人群计数的模型，它是一个端到端的单阶段的方法。LSC-CNN 可以同时处理多个尺度信息并在多个分辨率图像上进行预测，多个分辨率图像上的输出构成最终的预测结果。图 1.3 是 LSC-CNN 的架构，首先，Feature Extractor 在多个分辨率图像上提取特征，然后，多尺度特征图被输入到一系列的 Multi-scale Feedback Reasoning(MFR)单元中，之后经过提取的特征进行融合，并用

于预测 box，最后，Non-Maximum Suppression(NMS)从多个分辨率图像上确定有效的预测结果，并结合生成最终结果。在训练时，LSC-CNN 对 GWTA 训练阶段生成的伪 ground truth 进行像素级分类以完成网络优化，为了训练模型，最后的一个阶段使用了 GWTA 模块。GWTA 模块使用了 Winnners-Take-All（WTA)loss，可以挑选合适的 ground truth box。

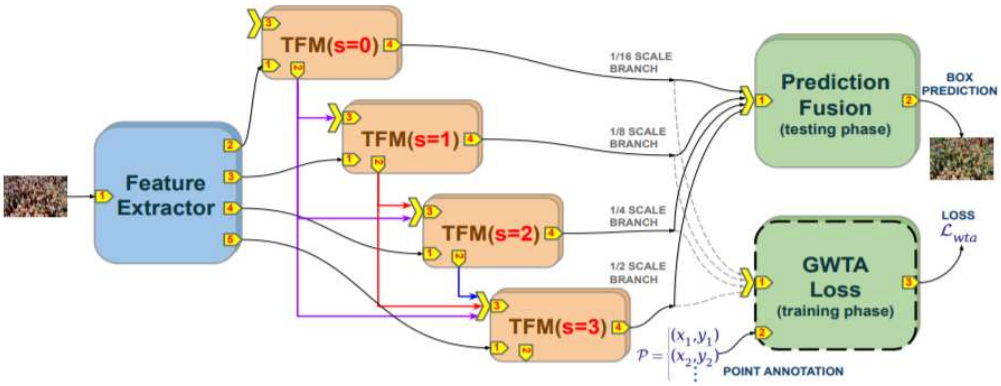

图 1.3. LSC-CNN 的主要架构

总的来说，LSC-CNN 使用了检测的方法进行人群计数，设计了一个新颖的 CNN 框架，可以在高分辨率图像上精确定位人头，此外，作者还设计了一个与从上到下反馈结构相融合的方案，使得网络可以联合处理多尺度信息，方便网络更好地定位人头。在仅有点标注信息的情况下，可以预测每个人头的 bounding box，并且在 GWTA 模块使用了新设计的 winner-take-all 的 loss，有利于在高分辨率的图像上进行训练。

1.2.4 视觉语言联合预训练的 12-in-1 模型[5]

12-in-1 模型是通过多任务训练来学习视觉语言联合表示的一种跨模态模型，该模型涉及了四类任务，视觉问题回答 (Visual Question Answering)，基于图像描述的图像检索 (Caption-based Image Retrieval)，看图识物 (Grounding Referring Expressions) 和多模态验证 (Multi-modal Verification)，并在 12 个不同的数据集上进行联合训练，图 1.4 是 12-in-1 模型的主要架构。

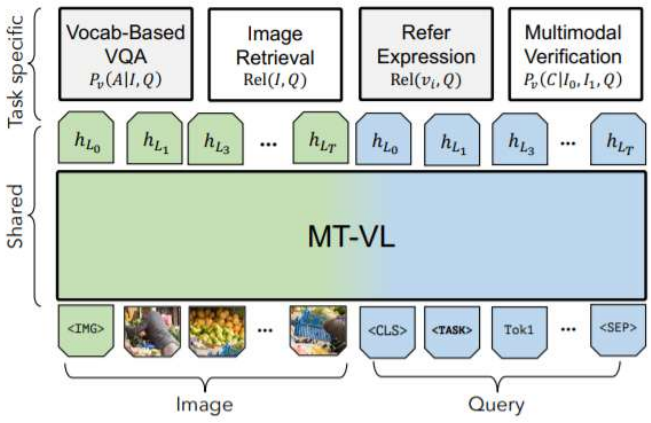

图 1.4. 12-in-1 的主要架构

12-in-1 的使用了视觉语言联合预训练模型 ViLBERT 作为 backbone，并在此基础上做了两个改进：首先，在 mask visual region 时，作者还遮住了具有 IoU > 0.4 的其他区域，以避免泄漏视觉信息，迫使模型更加依赖于语言来预测图像内容。其次，在对不匹配语言进行多模态对齐预测时，作者不强制掩蔽多模态建模损失，有效地消除由负样本引入的噪声。此外，为了同时训练 12 个难度、大小都不同的数据集，作者还引入了动态训练调度器 (Dynamic Stop-and-Go training scheduler)、基于任务的输入标记 (Task-dependent Input Tokens) 和简单的启发式超参 (simple Hyper-parameter Heuristics)。

通过多任务的学习可以获得更广泛的视觉语言联合表示，并用于不同的下游任务中，在本次人群计数的实验中，我将 12-in-1 模型应用到 VQA 这个下游任务中，输入视频的帧图像，以及问题“How many people are there in the picture?”，从而获得画面中人物的数量。

2 实验结果及分析

为了巩固对人群计数算法的理解，我使用了 4 种模型分别对低密度人群和高密度人群视频做了实验。首先，我在网上找了两段包含行人的短视频，一个是在车站月台的监测视频，是一段低密度人群的视频，另一个是商场大厅的监测视频，是一段高密度人群的视频。图 2.1 是两段视频中选取的某些帧图像。

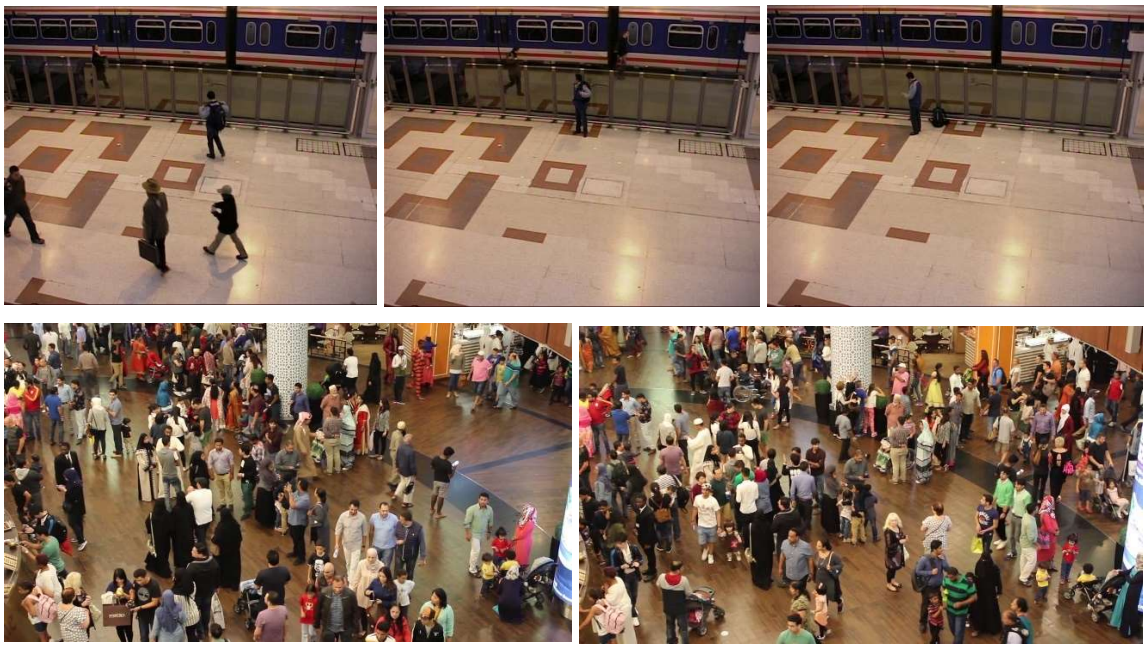

图 2.1. 低密度（上）视频中的 3 张帧图像
高密度（下）视频中的 2 张帧图像

2.1 YOLO v3 目标检测算法

2.1.1 低密度人群

YOLO v3 是目标检测的算法，自然也可以通过计数检测出的人数来完成人群计数人物。图 2.2 是我使用 YOLO v3 对低密度人群视频进行处理获得帧图像，完整的检测视频见附件中的 PPT。从效果上来看，当视频中人物较少时，YOLO v3 可以很好地检测出所有的目标人物，准确地完成人群计数，不过当目标离镜头较远时，YOLO v3 有时未能检测出这种小目标，造成不稳定。

图 2.2. YOLO v3 对低密度人群的计数

2.1.2 高密度人群

接着，我又将 YOLO v3 算法用于对高密度人群的计数中，图 2.3 是检测后的某些帧图像，从结果发现，基于目标检测的 YOLO v3 在人数过多的视频中，表现不理想，即使我将 bounding box 的最大数目设到了 300，但是还是智能检测出 30-50 个人，那些过于紧密的小目标未能成成功检测出，所以，YOLO v3 对高密度人群的计数效果很不理想。

图 2.3. YOLO v3 对高密度人群的计数

2.2 MCNN 人群计数算法

2.2.1 低密度人群

MCNN 是较早的一种基于深度学习的人群计数算法，它的主要结构是三列卷积神经网络。首先我先将其用于低密度人群的计数任务中，图 2.4 是计数结果的几张帧图

像。从结果来看，在低密度人群视频中，MCNN 有时可以准确计数，有时却计数错误，整体效果不如 YOLO v3 的计数效果。

图 2.4. MCNN 对低密度人群的计数

2.2.2 高密度人群

而在高密度人群的计数中，MCNN 的计数效果要明显优于 YOLO v3 的效果，图 2.5 是 MCNN 对高密度人群的计数结果。但是，MCNN 存在多计数的现象，即计数结果要比图中实际包含的人数要多，所以，在高密度人群的计数任务中，MCNN 的效果明显优于 YOLO v3 但是仍有不足。

图 2.5. MCNN 对高密度人群的计数

2.3 LSC-CNN 人群计数算法

2.3.1 低密度人群

LSC-CNN是一种利用检测方法来进行人群计数的算法，是目前比较新的一种计数模型。首先，我将其用于对低密度人群视频的计数任务中，获得图 2.6 所示的计数结果图。从结果来看，LSC-CNN 在低密度人群的计数中表现同样不好，总是将背景中的车窗也识别为人，从而导致计数出错。我将帧图像放大后观测，并不能判断车窗中是否真的有人，可能由于 LSC-CNN 的训练集中存在这种现象，所以它有时将车窗也识别为人。总之，LSC-CNN 在低密度人群计数中效果不如 YOLO v3 的计数效果。

图 2.6. LSC-CNN 对低密度人群的计数

2.3.2 高密度人群

接着，我将 LSC-CNN 模型用于高密度人群的计数任务中，得到图 2.7 所示的结果图。从结果来看，LSC-CNN 计数的结果相对于 MCNN 要更接近真实值一些，但是图像中仍有很多人未能检测出来，可能是因为我将非极大值抑制 (Non-Maximum Suppression, NMS)的阈值设置太低了，导致一些过于紧密的目标的 bounding box 被删除了。经过实验，将 NMS threshold 提高后，统计的数目的确得到了提升。总之，LSC-CNN 在高密度视频的计数任务中，表现效果要优于 MCNN 算法。

图 2.7. LSC-CNN 对高密度人群的计数

2.4 视觉语言联合预训练的 12-in-1 模型

2.4.1 低密度人群

12-in-1 模型不是一个专门用于人群计数的模型，它是通过多任务训练来学习视觉语言联合表示的一种跨模态模型。在这里我使用它是因为它可以用于 VQA 这个任务当中，而 VQA 任务中就存在回答数目的样例，VQA 的任务是输入一张图片和一个与图像相关的问题，模型会根据图片内容来做出相应的回答，在人群计数中，输入的问题自然就是“How many people are there in the picture?”，模型会根据学习到的视觉语言联合表示特征来进行作答。12-in-1 模型是一种多任务学习的模型，它涉及了 4 类不同的任务，在 12 个不同的数据集上联合训练，具有很强的视觉语言表示能力。

首先，我将 12-in-1 模型用于低密度人群的计数任务中，由于资源的限制，我仅在作者提供的网页版 demo 上选择了几张帧图像进行了测试，从测试结果来看，在低密

度人群的计数中，12-in-1 模型的表现效果很好，通常都能正确地回答图片中人物的数目，图 2.8 就是其中的一个检测结果。

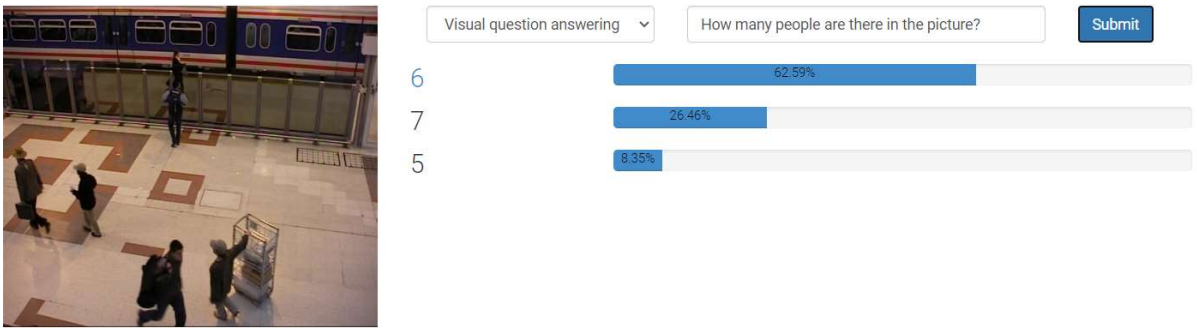

图 2.8. 12-in-1 对低密度人群的计数

2.4.2 高密度人群

但是，在高密度人群的计数中，12-in-1 模型完全不能进行准确的作答，如图 2.9 所示，虽然也能获得一个大致的范围，但是不能用于准确的计数。因为训练集中就没有出现这种有上百人的图像作答样本，而且模型本身也不是专用于人群计数任务，所以 VQA 模型不能处理高密度的人群计数问题。

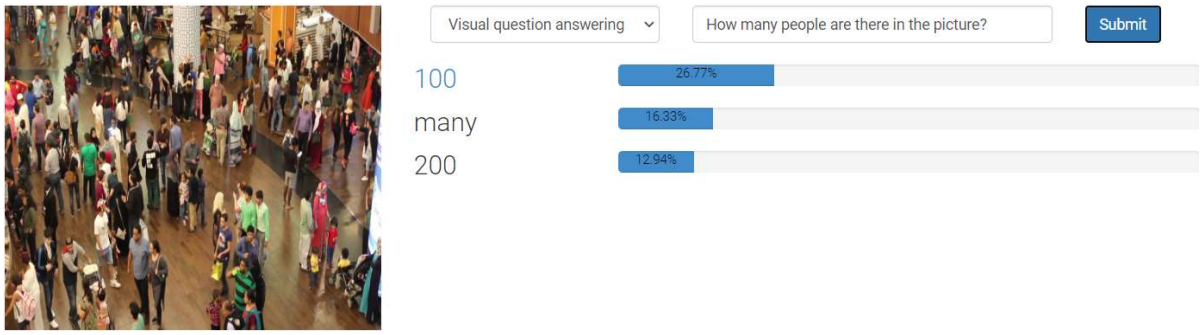

图 2.9. 12-in-1 对高密度人群的计数

3 实验总结

本次实验中，我选用了基于目标检测的 YOLO v3 模型，基于多列卷积神经网络的 MCNN 模型，利用检测方法进行人群计数的 LSC-CNN 模型，以及多任务的视觉语言联合预训练的 12-in-1 模型，分别对低密度人群和高密度人群两段监测视频进行了人数统计实验。

从实现效果来看，YOLO v3 由于是目标检测的算法，所以在目标不是很多的低密度人群的计数任务中表现效果很好，但是在高密度人群计数中表现不佳，仅能检测出 30-50 个目标，和实际相差太多。而 MCNN 模型作为较早的基于深度学习的人群计数

算法，尽管在低密度人群计数中表现不如 YOLO v3，但是在高密度人群计数中效果要明显优于前者，不过却存在计数过多的问题。之后我还使用了较新的 LSC-CNN 模型进行了实验，该模型利用检测方法进行人群计数，在低密度人群计数中表现效果同样不好，但是在高密度人群计数中相较于 MCNN 模型有了一定的提高，不过检测效果受参数 NMS threshold 影响较大。最后，我还使用了一个视觉语言联合预训练的 12-in-1 模型，将人群计数视为视觉问答 (VQA) 任务，输入监测视频的帧图像和问题“How many people are there in the picture?”来获得图像中的人数，该模型在低密度人群计数中表现不错，但是由于训练集和模型本身的限制无法处理高密度人群计数问题。

总的来说，通过这次实验，无论是从理论还是实践上都加深了我对视频人群计数的理解，帮助我对智能视频分析这门课程有了更深一步的了解。

参考文献

[1] 郑世宝．智能视频分析2020春第七章课件．上海．

[2] Joseph Redmon, Ali Farhadi. YOLOv3: An Incremental Improvement. In CVPR, 2018.

[3] Yingying Zhang, Desen Zhou, et al. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In CVPR, 2016.

[4] Sam, Peri, et al. Locate, Size and Count: Accurately Resolving People in Dense Crowds via Detection. In T-PAMI 2020.

[5] Jiasen Lu, Vedanuj Goswami1, et al. 12-in-1: Multi-Task Vision and Language Representation Learning. In CVPR, 2020.