Single-Image Crowd Counting via Multi-Column Convolutional Neural Network

Liangjie Cao

Jun 12, 2018

1. Experiments

They compare their method with the work of Zhang *et al.*, which also uses CNNs for crowd counting and achieved state-of-the-art accuracy at the time. Following the work of [4], they also compare their work with regression based method, which uses Local Binary Pattern (LBP) features extracted from the original image as input and uses ridge regression (RR) to predict the crowd number for each image. To extract LBP features, each image is uniformly divided into 8 x 8 blocks in Part A in Figure 1 and 12 x 16 blocks in Part B in Figure 1, then a 59-dimensional uniform LBP in each block is extracted and all uniform LBP features are concatenated together to represent the image. The ground truth is a 64D or 192D vector where each entry is the total number of persons in corresponding patch. They compare the performances of all the methods on Shanghaitech dataset in Table 1.

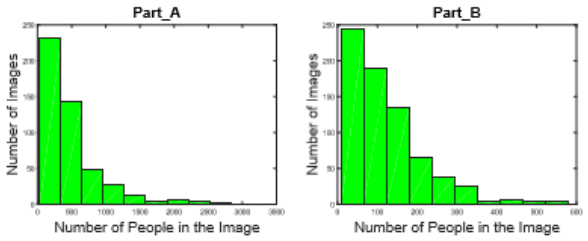

Figure 1. **Histograms of crowd counts of our new dataset**

	Part A		Part B	
Method	MAE	MSE	MAE	MSE
LBP+RR	303.2	371.0	59.1	81.7
Zhang *etal* [4]	181.8	277.7	32.0	49.8
MCNN-CCR	245.0	336.1	70.9	95.9
MCNN	**110.2**	**173.2**	**26.4**	**41.3**

Table 1. **Comparing performances of different methods on Shanghaitech dataset**

2. The effect of pretraining in MCNN

They show the effect of our model without pretraining on Shanghaitech dataset Part A in Figure 2. They see that pretrained network outperforms the network without pretraining. The result verifies the necessity of pretraining for MCNN as optimization starting from random initialization tends to fall into local minima.

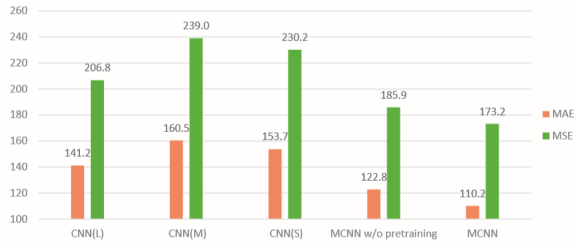

Figure 2. **Comparing single column CNNs with MCNN and M-CNN w/o pretraining on Part A. L, M, S stand for large kernel, medium kernel, small kernel respectively**

3. Single column CNNs vs MCNN

Figure 2 shows the comparison of single column CNNs with MCNN on Shanghaitech dataset Part A. It can be seen that MCNNs significantly outperforms each single column CNN for both MAE and MSE. This verifies the effectiveness of the MCNN architecture.

4. Comparison of different loss functions

They evaluate the performance of our framework with different loss functions. Other than mapping the images to their density maps, they can also map the images to the total head counts in the image directly. For the input image X_i (i = 1,...,N), its total head count is z_i, and $F(X_i;\theta)$ stands for the estimated density map and θ is the parameters of MCNN. Then they arrive the following objective function:

$$L(\theta) = \frac{1}{2N}\sum_{i=1}^{N} \| \iint_S F(X_i;\theta)dxdy - z_i \|^2 \qquad (1)$$

Here S stands for the spatial region of estimated density map, and ground truth of the density map is not used. For this loss, they also pretrain CNNs in each column separately. They call such a baseline as MCNN based crowd count regression (MCNN-CCR). Performance based on such loss function is listed in Table 1, which is also compared with two existing methods as well as the method based on density map estimation (simply labeled as MCNN). They see that the results based on crowd count regression is rather poor. In a way, learning density map manages to preserve more information of the image, and subsequently helps improve the count accuracy.

5. The UCF CC 50 dataset

The UCF_CC_50 dataset is firstly introduced by H. Idrees *et al.* [1]. This dataset contains 50 images from the Internet. It is a very challenging dataset, because of not only limited number of images, but also the crowd count of the image changes dramatically. The head counts range between 94 and 4543 with an average of 1280 individuals per image. The authors provided 63974 annotations in total for these fifty images. They perform 5-fold cross-validation by following the standard setting in [1]. The same data augmentation approach as in that in Shanghaitech dataset. They compare their method with four existing methods on UCF_CC_50 dataset in Table 2. Rodriguez *et al.* [3] employs density map estimation to obtain better head detection results in crowd scenes. Lempitsky *et al.* [2] adopts dense SIFT features on randomly selected patches and the MESA distance to learn a density regression model. The method presented in [1] gets the crowd count estimation by using multi-source features. The work of Zhang *et al.* [4] is based on crowd CNN model to estimate the crowd count of an image. Their method achieves the best MAE, and comparable MSE with existing methods.

Method	MAE	MSE
Rodriguez *et al.* [3]	655.7	697.8
Lempitsky *et al.* [2]	493.4	487.1
Idrees *et al.* [1]	419.5	541.6
Zhang *et al.* [4]	467.0	498.5
MCNN	377.6	509.1

Table 2. **Comparing results of different methods on the UCF CC 50 dataset**

References

[1] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, 2013.

[2] Victor S. Lempitsky and Andrew Zisserman. Learning to count objects in images. In *NIPS*, 2010.

[3] Mikel Rodriguez, Ivan Laptev, Josef Sivic, and Jean Yves Audibert. Density-aware person detection and tracking in crowds. In *ICCV*, 2011.

[4] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, 2015.