Crowd Counting via Multi-Column Deconvolution Networks

Jibing Jiang1, Mengxuan Bai2, Bo Liu1, Guihong Zhou1*

1 Agricultural University of Hebei, Baoding, Hebei, 071000, China
1* Agricultural University of Hebei, Baoding, Hebei, 071000, China
2 Agricultural University of Hebei, Baoding, Hebei, 071000, China
3 Agricultural University of Hebei, Baoding, Hebei, 071000, China

*Corresponding author’s e-mail: hebau_zgh@163.com

Abstract: The difference in distance between the crowd and the camera results in different scales of the population in the image. In the study of crowd counting, this paper first proposed the method of adding deconvolution to multi-column network for the problem that the population size in the image does not affect the accuracy of counting, namely multi-column deconvolution neural network (Multi-Column Deconvolutinal Neural Network, MCDNN). Compared with the results of the Multi-Column Convolutional Neural Network (MCNN) on the ShanghaiTech dataset, the MSE is reduced by 42.8 and the MAE is reduced by 26.6. On the UCF_CC_50 dataset, the MSE is reduced by 109.4 and the MAE is reduced by 86.3. Compared with other advanced methods, the proposed MCDNN network achieves better performance on the data set, effectively solves the problem of inaccurate population counting accuracy due to scale changes, and can generate a better quality density map.

1. Introduction

With the rapid increase in the crowd and the rapid development of the economy, more and more large-scale collective activities have emerged, and many social problems caused by crowds have become increasingly prominent, especially in some densely populated areas (attractions, stations, shopping malls and other public places, there will be frequent accidents. In the 2014 New Year's Eve, the stampede in the Shanghai Bund Scenic Area was shocked. At the scene, crowds and safety warnings were insufficient, resulting in 36 deaths and 49 injuries. Therefore, crowd counting has become an important research topic in the field of intelligent video surveillance, and it has great value in preventing crowd overcrowding and security early warning. However, due to the complexity of occlusion, high clutter, uneven distribution of people, uneven illumination, changes in appearance, scale and perspective within the scene and between scenes, the accuracy of the results obtained is far from optimal.

In recent years, with the continuous development of computer vision technology, relevant researchers have proposed a lot of people's technical methods. There are three basic methods known from the literature [1,2]: (1) regression-based methods. (2) A method based on density estimation. (3) CNN-based methods. The first two methods are more suitable for sparse crowd scenarios. The regression-based method is a mapping relationship between features extracted from local image blocks and their counts, and can capture the overall density information of the crowded areas to improve the accuracy of the crowd count [3,4,5]. However, when returning to the overall crowd, the key spatial information is ignored based on the regression method.
Based on the method of density estimation, authors of the paper\textsuperscript{[6]} proposed a nonlinear mapping method between local image block features and density maps using a random forest framework. The paper\textsuperscript{[7]} proposed a method based on rich features and random projection forests, using richer and broader features for crowd density estimation, but if using density map estimation based methods in crowd-intensive scenarios, there will be people in the crowd. The density distribution is uneven and the scale is not equal.

With the successful application of Convolutional Neural Networks (CNN) in computer vision technology, many CNN-based methods are applied to crowd counting problems. The method of automatically extracting the density estimation of image features based on convolutional neural network has become the mainstream method to solve the crowd counting problem. The data set used by the crowd count is the coordinate position of the image of each person's head (i.e., the marked pixel position), and the coordinate position of the pedestrian's head needs to be converted into a label density map of the supervised data, that is, according to the labeled human head. The position and kernel density estimates generate a crowd density map corresponding to each crowd image, then return the density map to the crowd characteristics, and finally combine the numerical values of the crowd density map as the final predicted number. The scale problem is a key factor that limits the accuracy of crowd counting. Some CNN-based methods are particularly suitable for solving scale changes by multi-column or multi-resolution networks. Automatic selection of appropriate column CNN networks for crowd density estimation. Authors of the paper\textsuperscript{[8]} proposed a spatial pyramid method based on order, which can not only process sub-image blocks of multiple scales, but also speed up the calculation of network models. However, the largest pooling method used in pooling will result in loss of feature information. Authors of the paper\textsuperscript{[9]} proposed to use a combination of shallow and deep networks to perform extensive data enhancement by sampling patches from multi-scale image representations. Although the problem of scale changes is solved, the number of parameters in the network is increased, resulting in network Performance has dropped dramatically. The paper\textsuperscript{[10]} designed a U-net structured generation network, generated a density map from the input patch, and directly used the antagonistic loss to shrink the solution to a real subspace, thus weakening the fuzzy problem of density map estimation. A clearer density map is obtained, but the applicability of the perspective effect is poor. The paper\textsuperscript{[11]} proposed to aggregate multi-scale convolution features extracted from the whole image into a compact single vector representation, which can be efficiently and accurately counted by the vector of local aggregation descriptors (VLAD), but with increased training. The difficulty of the training leads to inefficient training. In the paper\textsuperscript{[12]}, a crowd counting method for deep convolutional neural networks is proposed. The method uses the crowd count and density map to alternately return to two related learning objectives. The two learning objectives can be converted to each other, thereby improving the accuracy of crowd counting. Rate, but the application of this method is limited because it requires a perspective view of the image, which is not readily available during actual training and testing.

2. Proposed method

The crowd density estimation is essentially the detection and perception problem of pedestrian targets. In order to solve the problem of multi-scale, small target perception and single network structure only perceiving targets within a certain scale, the author of literature\textsuperscript{[6]} proposed multi-column network structure (MCNN) for the first time. Each column of susceptors has the ability to perceive crowd targets within a certain scale, and has the ability to perceive different scale ranges. The paper\textsuperscript{[10]} added the understanding of the convolution process in the single-level convolution network, which effectively reduced the error and improved the accuracy of the crowd count.

Inspired by the paper\textsuperscript{[6,10]}, we proposed a multi-column network structure and deconvolution process to design a Mutil-Column Deconvolutinal Neural Network (MCDNN), MCDNN network structure such as Figure 1 shows. The method can effectively solve the problem that the crowd has different scales and the traditional pooling method is easy to lose a large amount of important information, and the accuracy of the crowd counting is improved.
2.1 Network structure

The network input is an image frame, and the head target position is used as the supervised information. The crowd density map is calculated by the convolution-deconvolution process, and the values on the density map are integrated and summed, that is, the number of people in the whole picture is obtained.

The MC-DNN network model has three columns of parallel CNN channels. The first column of the network structure contains 4 convolutional layers, 1 pooling layer. Conv1_1 convolution kernel size is 9×9, Conv1_2, Conv1_3, Conv1_4 convolution kernel size is 7×7, the step size is 1. In order to keep the size of the feature map after convolution pooling, there is a deconvolution operation at a later time; the second column of the network contains 4 convolution layers and 2 pools. The Conv2_1 convolution kernel has a size of 7×7, Conv2_2, Conv2_3, and Conv2_4 convolution kernels are all 5×5, and the step size is 1. There are two deconvolution operations to ensure the feature size and the first column. The second column of the third column network of the network structure includes four convolutional layers and three pooling layers. The conv3_1 convolution kernel has a size of 5×5, and the Conv3_2, Conv3_3, and Conv3_4 convolution kernels have a size of 3×3. The step size is 1, and the deconvolution operation is performed 3 times, which ensures that the size of the feature map is consistent after the final merge. In the network structure, the maximum pooling method is adopted for down-sampling, with a 2×2 size filter and a step size of 2. The width and height of the feature map become the sum of the initial input feature map after one down-sampling. Half the height. The last convolution kernel 3×3 convolution layer maps the feature map to the density map. The output of the network is a crowd density map. After integrating, the number of people in the image can be obtained. In the preprocessing of the data set, there are W×H image frames, 1/2W×1/2H image frames, and 1/4W×1/4H image frames, respectively. The convolution operation part of each column is composed of four convolution layers of different size receptive fields. The large receptive field corresponds to a large-sized convolution kernel, and the small receptive field corresponds to a small-sized convolution kernel. In order to reduce the computational complexity, optimize the number of parameters, use a smaller number of large-scale convolution kernels, the number of times of deconvolution operation, depending on the feature map input after the fourth convolutional layer of the convolution operation size. The function of deconvolution not only ensures that the final merged feature map size is W×H, but also compensates for the loss of detailed information caused by the previous pooling operation.

2.2 Calculation of crowd density map

Since the dataset images are characterized by the head of the crowd as the supervised signal of the network, the live crowd density map on the ground is generated by the Gaussian nuclear convolution annotation points. According to the literature [6], the normal head size is known. It is related to distance, so 2D Gaussian kernel is used to represent each head target in the density map. The center position of the head target is the center of the 2D Gaussian kernel to calculate the ground truth density map of the
crowd, if the crowd target on an image frame The set is H={h1,h2,...,hN}. If the center coordinate of the head of the target hi is (xi, yi), it can be represented as, and the function represented on a sample picture is:

$$U(x, y) = \sum_{i=1}^{N} \delta(x-x_i, y-y_i)$$

(1)

The Gaussian kernel function of the head target of the crowd is:

$$G_{\beta}(h_i) = \frac{1}{2\pi\beta^2} e^{-\frac{(x-x_i)^2+(y-y_i)^2}{2\beta^2}}$$

(2)

Among them is the Gaussian kernel parameter corresponding to the target. We know U(x, y) is a discrete function. To convert the formula (1) into a continuous function, it needs to be convolved with the Gaussian kernel function so that the crowd density map Q is:

$$Q(H) = U(x, y) G_{\beta}(h_i)$$

(3)

2.3 Loss function

In this paper, the Euclidean distance is used to estimate the difference between the density map and the label map. The purpose is to obtain a high-quality crowd density distribution, and to obtain the optimal crowd count result, the loss function definition of the density estimate. for:

$$L_d = \frac{1}{N} \sum_{i=1}^{N} \| F_d(X_i, \varphi) - D_i \|_2^2$$

(4)

N is the number of sample images, \( \varphi \) is the learning parameter of the proposed network, \( F_d(X_i, \varphi) \) is the estimated density map, \( X_i \) is the input image, \( D_i \) is the true value of the density map. Using the L2 norm paradigm is used to estimate the density map to approximate the true distribution of the crowd.

3. Experiment and result analysis

3.1 Experimental conditions and parameter settings

In this section, we will introduce the experimental data set, experimental evaluation indicators and experimental results of different methods. The experiment in this paper is based on PyTorch open source deep learning architecture, training and evaluation on NVIDIA GTX1080Ti GPU. During the training, all parameters were optimized by Batch Gradient Descent (B GD) and Batch Gradient Descent (BGD). The initial learning rate is 0.00001, the momentum is 0.9, and the weight decay is set to 0.0001 to avoid over-fitting problems in the training phase. Finally, in order to improve the training speed and reduce the memory occupancy rate, each training set of the original data set is randomly cropped during each iteration of the training phase, so as to use the obtained image sub-block for training.

3.2 Evaluation indicators

In this experiment, the experimental details and experimental results were presented by using two major public data sets, ShanghaiTech dataset and The UCF_CC_50 data set, using the crowd count evaluation criteria as Mean Absolute Error (MAE) and Mean Square Error. MSE (Mean Square Error), where MAE represents the accuracy of the experimental method estimation, and the MSE experimental method estimates the robustness. The indicators are defined as follows:
\[ MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \]  
\[ MSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|^2} \]  

(5)  

(6)

N is the number of test sample pictures, \( y_i \) is the actual number of people, and \( \hat{y}_i \) is the estimated number of people corresponding to the i-th sample picture.

3.3 Data sets and experimental results

3.3.1 ShanghaiTech Dataset

The ShanghaiTech dataset contains 1,198 tagged images for a total of 33,165 people. The data set is divided into two parts, PartA and PartB. PartA is a dense group and the data is from the Internet. PartB is a sparse crowd of surveillance video frames from the streets of Shanghai. There are 482 sample images in A, with 300 for the training set and 182 for the test set. There are 716 pictures in B, with 4090 sheets for the training set and 316 for the test set.

In Fig. 2, the crowd density estimation result obtained by the experiment is obtained by taking one frame image on the PartA data set and the PartB data set by using the method. Where (1) represents the original image. (2) The true value of the ground crowd. (3) The estimated results of the MCNN method. (4) The estimated results of our method.

The method proposed in this paper was compared with several popular methods. The experimental results are shown in Table 1.

| methods               | PartA | PartB |
|-----------------------|-------|-------|
| MCNN[6] (2016)        | 110.2 | 41.3  |
| Switch-CNN[7] (2017)  | 90.4  | 33.4  |
| SC_CNN[16] (2018)     | 88.6  | 27.2  |
| SaCNN[17] (2018)      | 86.3  | 25.8  |
| Our method            | 83.6  | 24.6  |
3.3.2 UCF_CC_50 data set
The UCF_CC_50 data set is proposed by the author of the paper[13]. There are 50 sample images, each of which contains between 94 and 4543 people. The images are mostly dense and have great challenges for crowd counting.

![Image](image_url)

Fig.3 running results

In Fig. 3, the crowd density estimation result obtained by the experiment is obtained by taking one frame image on the UCF_CC_50 data set by using the method. Where (1) represents the original image. (2) The true value of the ground crowd. (3) The estimated results of the MCNN method. (4) The estimated results of our method.

The method proposed in this paper was compared with several popular methods. The experimental results are shown in Table 2.

| methods                  | MAE   | MSE   |
|--------------------------|-------|-------|
| MCNN[6] (2016)           | 377.6 | 509.1 |
| Switch-CNN[7] (2017)     | 318.1 | 439.2 |
| SC_CNN[16] (2018)        | 309.7 | 400.0 |
| SaCNN[17] (2018)         | 314.9 | 424.8 |
| Our method               | 291.3 | 399.7 |

3.4 Results analysis
The experimental results obtained by extracting 100 data in the test set in the PartA data set (Fig. 4) and the PartB data set (Fig. 5), respectively.

(1) The estimated value and true value distribution of the number of people per image in the PartA data set.

![Image](image_url)

Fig. 4 PartA estimates and true values
(2) The crowd density estimation results obtained by the method on the UCF_CC_50 data set are shown in Fig.6.

It can be seen from Table 1 that the method in this paper is superior to other methods in the ShanghaiTech dataset. Compared with the latest SC_CNN method, the MAE of the Part_A verification set has decreased by 5.0, the MSE has decreased by 18.0. The MAE of the Part_B verification set has decreased by 0.3, and the MSE dropped by 2.6. It can be seen from Table 2 that on the UCF_CC_50 data set, compared with SC_CNN, the MAE of this method is reduced by 18.4 and the MSE is decreased by 0.3. As shown in Fig. 4 and Fig. 5, the red dotted line represents the estimated person value, and the blue dotted line represents the actual number of people. By observing the trend, it is easier to estimate the value in the high-dense crowd dataset such as the PartA data set and the UCF_CC_50 data set. Deviation from the true value. Figures 3 and 6 also reflect the "overestimation" results of the estimated values above the true values. The main reason for the high estimate is that the interference objects such as buildings and trees in the image are regarded as the head of the human body. On the sparse dataset of PartB crowd, the estimated value is basically consistent with the reality, and the accuracy rate is 99.7%, which indicates that the method exhibits high precision and robustness in sparse crowd conditions. At the same time, in the course of the experiment, the method of this paper is compared with the MCNN method, which not only improves the crowd counting accuracy, but also obtains a clearer crowd density map.

4. Conclusion
This paper presents for the first time a multi-column deconvolution CNN network for end-to-end population counting for population density map estimation. The deconvolution is used for regression on the full resolution density map to make up for the loss of detail caused by the largest pooling layer
in the previous period. The whole process is trained in an end-to-end manner. To solve the problem of head size change on the data set without providing perspective information, a geometric adaptive Gaussian kernel method is used to generate the corresponding ground truth density map. Experiments were carried out on the public dataset and compared with the experimental results of the current popular methods. Compared with the MCNN method, a population density map with clearer quality was obtained. It proves the effectiveness of the proposed method.

References
[1] C.C. Loy, K. Chen, S. Gong, and T. Xiang. Crowd counting and profiling: Methodology and evaluation. [J].Modeling, Simulation and Visual Analysis of Crowd(MSVA), 2013, 11(4):347-382
[2] Siqi Tang, Wei Tao, Liangliang Zhang, et al. A deep crowd counting algorithm based on multi-column feature map fusion[J]. Zhengzhou University Journal, Science Edition, 2018, 50(2): 69-74
[3] Ryan, S. Denman, C. Fookes, and S. Sridharan. Crowd counting using multiple local features. [J]. In Digital Image Computing: Techniques and Applications, 2009, 20(14): 81-88
[4] K.Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting[J]. British Machine Vision Conference, 2012, 10(5244):21-26
[5] K.Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. [J]. In European Conference on Computer Vision, 2012, 10(21):1-2
[6] Y.Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single image crowd counting via multi-column convolutional neural network.[J]. In Proceedings of the IEEE CVPR, 2016, 23(3):589-597
[7] DB. Sam, S. Surya, and R. V. Babu. Switching convolutional neural network for crowd counting.[J]. CVPR, 2017., 2017(1):1-7
[8] SHI Z L, YE Y D, WU Y P, et al.Crowd Counting Using Rank-based Spatial Pyramid Pooling Network[J].Acta Automatica Sinica, 2016, 42(6):866-874
[9] Xialei Liu, JWiejier.. Leveraging Unlabeled Data for Crowd Counting by Learning to Rank.[J].CVPR, 2018, 24(4):7661-7669
[10] SINDAGI V A, PATEL V M.Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting[J].Advanced Video and Signal Based Surveillance (AVSS), 2017., 1(6)
[11] ZShi, L. Zhang, Y.Sun, and Y.Ye.. Multiscale Multitask Deep NetVLAD for Crowd Counting [J].In IEEE., 2018., 12(31):4953-4962
[12] C.Zhang, H. Li, X. Wang, and X. Yang.. Cross-scene crowd counting via deep convolutional neural networks. [J].IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2015 , 2(12):833-841
[13] Hldrees, I. Saleemi, C. Seibert, and M. Shah..Multi-source multi-scale counting in extremely dense crowd images.[J]. IEEE CVPR, 2013., 3(43):2547-2554
[14] L. Boominathan, S.S. Kruthiventi, and R.V. Babu. Crowdnet: A deep convolutional network for dense crowd counting. [J].In Proceedings of the 2016 ACM on Multimedia Conference, 2016, 2016(1):640-644
[15] CZhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. [J]. ICCV, 2015., 1(2):1-7
[16] Jun Sang, Weiqun Wu, Hongling Luo et al. A crowding counting method based on single-column convolutional neural network[J]. Acta Automatica Sinica Sinica, 2018, 20(3):65-71
[17] L.Zhang, M. Shi, and Q. Chen. Crowd counting via scale-adaptive convolutional neural network.[J]. In 2018 IEEE Win-ter Conference on Applications of Computer Vision (WACV), 2018, 1(22):1113-1121
[18] Y.Li, X. Zhang and D. Chen. Crsnet: Dilated convolutional neural networks for understanding the highly congested scenes. [J].n Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 5(15):1091-1100
[19] D. Onoro-Rubio and R. J. Lopez-Sastre. Towards perspective-free object counting with deep learning. [J]. In ECCV. Springer, 2016, 3(5):1-7

[20] B. Xu and G. Qiu. Crowd density estimation based on rich features and random projection forest. [J]. IEEE WACV, 2016, 2016(1):1-8