The Multi-channel and Multi-scale Network for Crowd Counting

Pengze Wang*, Wei Wu, Yang Su, Xin Li and Yongsheng Duan

School Of Information Engineering, Wuhan University of Technology, Wuhan 430070, China

*Corresponding author e-mail: 634839284@qq.com

Abstract. In this paper, we propose a novel multi-channel and multi-scale network for processing crowds in crowded scenarios and improving counting accuracy. In the estimated crowd count study, different distribution groups have different contributions to the total number of crowd, and the more crowded people have stricter requirements on details. Therefore, we designed two branches in the crowd counting network: the backbone network performs feature extraction operations on the original image, which mainly obtains effective information from the global, and our branch network focuses on the crowd gathering area, which better focuses on the details of the crowd distribution. Finally, the global information is complemented with local details to obtain high-quality feature expressions. To deal with scale changes, Inspired by atrous spatial pyramid pooling structures, we introduce dilated convolution with different sampling rates in the network to expand the receptive field. We carried out a large number of experimental verifications on popular data sets, and the proposed method is superior to existing methods.

1. Introduction

With the continuous increase of the urban population, the crowding situation has gradually increased, which has also caused a series of security issues. In order to deal with these problems, people have researched and implemented calculation methods for the number of people in different scenarios. Traditional crowd counting mainly includes detection-based methods and regression-based methods. The detection method uses a sliding window to detect the scene crowd and count the number of people. There are two methods: overall detection [1] (Such as Haar wavelets) and partial body detection [2]. it is difficult to handle dense situations; the idea of the regression method is to learn the mapping relationship between the characteristics of the crowd [3], which is mainly divided into two steps, the first step is to extract low-level features, and the second step is to learn a regression model, This method has shortcomings in sparse scenes and also ignores local information, Lempitsky et al. [4] Proposed a method to solve this problem by learning the linear mapping between local features and their density maps, Pham et al. [5] Using random forest regression to learn non-linear mapping instead of linear mapping.

In recent years, deep learning has been paid attention to because of its outstanding ability to be used in crowd counting. Predicted density maps can get better results, and due to the great success of visual classification and recognition, researchers began to use CNN to predict density maps. Boominathan et
al. [6] Use deep and shallow networks to improve feature expression. In order to extract head features at different scales, [7] uses multiple columns of different convolution kernel size networks to perform count prediction. Sindagi et al. Added global and local information to improve network performance. [8] is to pre-process the original image, and select the most suitable channel for each block to generate a high-quality density map. And Cao et al. network design borrows the inception structure for scale aggregation. Liu et al. Combining regression and detection methods to improve accuracy. Ranjan et al. proposed the ic-CNN structure, using a combination of high and low resolution methods to improve the density map quality. A model of CSRNet [9] is proposed, which uses dilated convolution to extract deeper features. Liu et al. [10] Let the network generate high-quality density maps by detecting the population distribution.

Crowds are often unevenly distributed in the scene, and different distributions have different effects on the total population (as shown in Fig.1). Existing network frameworks will lose some detailed information during feature extraction, and the details will have a greater impact on densely populated areas, and often small-scale effective information is easily lost, which results in relatively low feature information. In order to deal with these problems, in this article, we have added a branch network to the existing framework with better performance to supplement the crowd details. The original network obtains useful feature information from the global, and the branch network complements the local details with the global information by operating the crowd area from which background information is removed, which improves the feature map's relatively concentrated crowd feature expression, which contains more detailed information to improve the quality of the density map. In order to cope with the acquisition of multi-scale information of pictures, we introduce dilated convolution structures with different sampling rates. Dilated convolution can expand the receptive field without losing resolution.

![Figure 1. The picture (a) shows the distribution information of the crowd in the scene. Take the same size of space (b) in the picture. The population of different densities contains different numbers of people.](image)

2. Proposed Solution

As shown in Fig.2(a), this is our network framework. It consists of two parts. The first part mainly implements the complementation of global and local effective information. The second part uses atrous spatial pyramid pooling to handle crowd multi-scale and occlusion problems. In the following, we will introduce the details of the network.

2.1. Multi-path Learning

We use VGG16 as the feature extraction part because it has excellent learning ability and a lightweight network framework. We removed its fully connected layers and kept 10 convolutional layers. Different population densities have different effects on the total number of people, and the requirements for detailed information are different. However, the existing network has the loss of crowd details. We have added a branch network, which is used to learn the details of the crowd. Global information is
complementary. According to [10], we preprocess the background information of the original image (as shown in Fig.2(b)). Considering that the background will also contain a small number of people and the direct fusion of the crowd distribution will not improve the details, so picture background information is removed as the input of the branch network, the features extracted by this network provide more detailed information for the crowd, while the main network extracts features from the original image. It contains more comprehensive information than the branch network and compares it with the feature information extracted by the branch Complementary to improve the detailed information of the crowd, for generating a better density map.

![Figure 2. Our network framework (a) and preprocessing module (b).](image)

2.2. Multi-scale fusion
There is a pooling operation in the network, which will change the size of the output, reduce the resolution and lose details. Dilated convolution is introduced in [9]. This operation can also expand the receptive field without losing resolution, and it obtains more detailed information than extracting features after pooling and then upsampling. Considering the different population distribution in different scenarios and the existence of multi-scale problems in the crowd, the atrous spatial pyramid pooling structure [11] (as shown in Fig.3) with different sampling rates is introduced in the second part. Dilated convolution can effectively capture multi-scale information. This sampling scheme is more suitable for people in different scenarios. There is a pooling operation in the previous part, and the size of its output will also be reduced. We finally upsample it (bilinear) to maintain the true density map size.

![Figure 3. Structure of ASPP, we use different sampling rates to obtain different scale information.](image)
2.3. Ground truth generation

In the real density map method, we still use the method of [9], using the geometry-adaptive kernels to deal with crowded scenes, and blur each head's annotations by Gaussian kernel. The geometry-adaptive kernel is defined as:

\[ F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x), \text{with } \sigma_i = \beta d_i \]

Where \(d_i\) represents the average distance of \(k\) nearest neighbors, and \(x_i\) represents the target object with ground truth value \(\delta\), and convolves \(\delta(x - x_i)\) with the Gaussian kernel (standard deviation is \(\sigma_i = \beta d_i\)).

3. Evaluation and comparison

In this section, we have conducted a large number of experiments on our network on different datasets, and compared their results with the previous latest methods.

3.1. Datasets

For the datasets section, we used ShanghaiTech dataset, UCF_CC_50 dataset and The UCSD dataset. The ShanghaiTech crowd counting dataset contains 1198 annotated images for a total of 330,165 people. UCF_CC_50 dataset includes 50 images with different perspectives and resolutions. The UCSD dataset has 2000 frames captured by surveillance cameras.

3.2. Training Details

For each image, we crop 9 patches that are 1/4 of the original image, 4 of which are from non-overlapping parts, and the remaining random crops, which double the training set by mirroring the patch. The optimization method uses Adam. The learning rate is 1e-6 during training. The Euclidean distance is used to measure the difference between the generated density map and the ground truth. We trained 2000 epochs on the RTX2080-gpu using pytorch framework.

3.3. Comparative Results

On MAE and MSE, our scheme is compared with the latest CSRNet and ADCrowdNet(AMG-DME) and previous methods on different datasets. The comparison results are summarized in Table 1. For the quality of the density map, we compared with the latest method in ShanghaiTech Part A dataset, and the results are shown in Table 2. We are also showing our test sample (as shown in Fig.4).

| Table 1. Algorithm comparison (MAE/MSE). |
|------------------------------------------|
| Part A                                  | Part B | UCF_CC_50 | UCSD    |
|------------------------------------------|--------|-----------|---------|
| MAE | MSE  | MAE | MSE  | MAE | MSE | MAE | MSE | MAE | MSE |
|-----|------|-----|------|-----|-----|-----|-----|-----|-----|
| MCNN | 110.2 | 173.2 | 26.4 | 41.3 | 377.6 | 509.1 | 1.07 | 1.35 |
| CSRNet | 68.2 | 115.0 | 10.6 | 16.0 | 266.1 | 397.5 | 1.16 | 1.47 |
| ADCrowdNet | 66.1 | 102.1 | 7.6 | 13.9 | 257.9 | 357.7 | 1.10 | 1.42 |
| Ours | **63.0** | **101.6** | **8.5** | **12.6** | **255.6** | **366.4** | **1.02** | **1.28** |

| Table 2. Density map quality comparison (PSNR/SSIM). |
|------------------------------------------|
| Method                                  | CP-CNN | CSRNet | ADCrowdNet | Ours |
|------------------------------------------|--------|--------|-------------|------|
| PSNR                                    | 21.72  | 23.79  | 24.48       | **24.87** |
| SSIM                                     | 0.72   | 0.76   | 0.88        | **0.89** |
4. Conclusion

In this paper, we propose a novel multi-channel and multi-scale network. Multi-channel networks complement local details and global information. Part of the detailed information, the introduction of the dilated convolution structure with different sampling rates improves the network's ability to handle different scenarios. Experiments show that our solution is superior to existing solutions, generates high-quality density maps, and improves counting accuracy.

Acknowledgments

This work was guided by teacher, Wei Wu and helped by my classmates.

References

[1] Paul Viola and Michael J Jones. Robust real-time face detection. International journal of computer vision, 57 (2): 137-154, 2004.
[2] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence, 32 (9): 1627-1645, 2010.
[3] Antoni B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In Computer Vision, 2009 IEEE 12th International Conference on, pages 545-551. IEEE, 2009.
[4] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In Advances in Neural Information Processing Systems, pages 1324-1332, 2010.
[5] Viet-Quoc Pham, Tatsuo Kozakaya, Osamu Yamaguchi, and Ryuzo Okada. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In Computer Vision (ICCV), 2015 IEEE International Conference on, pages 3253-3261. IEEE, 2015.
[6] Lokesh Boominathan, Srinivas S.S. Kruthiventi, and R. Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In Proc. ACM MM, pages 640-644, 2016.
[7] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 589-597, 2016.
[8] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, volume 1, page 6, 2017.
[9] Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proc. IEEE CVPR, pages 1091-1100, 2018.
[10] Ning Liu, Yongchao Long, Changqing Zou, Qun Nium Li Pan and Hefeng Wu. ADCrowdNet: An Attention-Injective Deformable Convolutional Network for Crowd Understanding In Proc. IEEE CVPR, 2019.
[11] LiangChieh Chen, George Papandreou, Florian Schroff and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv preprint arXiv:1706.05587, 2017.