Journal of Robotics, Networking and Artificial Life
Vol. 7(4); March (2021), *pp.* 279–283
DOI: https://doi.org/10.2991/jrnal.k.201215.015; ISSN 2405-9021; eISSN 2352-6386
https://www.atlantis-press.com/journals/jrnal

Research Article

Crowd Density Estimation based on Global Reasoning

Li Wang, Fangbo Zhou, Huailin Zhao*

School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai, China

ARTICLE INFO

Article History

Received 10 September 2019
Accepted 04 December 2020

Keywords

Global reasoning unit
graph convolutional network
crowd density estimation

ABSTRACT

The problem of crowd counting in single images and videos has attracted more and more attention in recent years. The crowd counting task has made massive progress by now due to the Convolutional Neural Network (CNN). However, filters in the shallow convolutional layer of the CNN only model the local region rather than the global region, which cannot capture context information from the crowd scene efficiently. In this paper, we propose a Graph-based Global Reasoning (GGR) network for crowd counting to solve this problem. Each input image is processed by the VGG-16 network for feature extracting, and then the GGR Unit reasons the context information from the extracted feature. Especially, the extracted feature firstly is transformed from the feature space to the interaction space for global context reasoning with the Graph Convolutional Network (GCN). Then, the output of the GCN projects the context information from the interaction space to the feature space. The experiments on the UCF-QNRF dataset demonstrate the effectiveness of the proposed method.

1. INTRODUCTION

In the current social and economic activities, large-scale crowd gathering has become more and more common, and the effective supervision of crowd gathering has become increasingly important. Accurate crowd counting and density estimation can help people effectively avoid stampede and riots caused by high crowds and help control crowd flow during epidemic prevention and control. Counting the number of people in a high-density population is difficult because of occluding each other partly and non-uniform distribution in complex scenarios.

The current mainstream crowd counting algorithm is a crowd density estimation algorithm based on deep learning. It mainly uses the powerful feature expression ability of Convolutional Neural Networks (CNNs) to learn the non-linear mapping relationship between input images and output density maps. Compared with the traditional crowd counting algorithms, CNNs have made significant progress in crowd counting. Such as, Zhang et al. [1] proposed MCNN, a typical multi-column architecture, in which convolution kernels of different sizes are used in each column to obtain multi-scale information. Li et al. [2] proposed CSRNet, which uses the first ten layers of VGG-16 as the front end of the network, and the back end uses hollow convolution to increase the receptive field of the network. However, the above method only models the local region, and the effect of the global region is slightly poor. This leads to difficulty in extracting context information and poses a certain challenge for accurate crowd density estimation.

In order to overcome the inherent limitations of the convolution operation, a global reasoning network based on the Graph Convolution Network (GCN) is proposed for crowd counting, and the global relation of the input crowd is modeled. The extracted crowd features are transformed from the feature space to the interaction space for global context inference, the context information is inferred and interpreted in the interaction space, and then the context-aware features are returned to the original coordinate space and combined with the features that were originally extracted. The use of global inference networks increases the relationship between different regions of the image and improves the accuracy of crowd density estimation.

2. RELATED WORK

Traditional machine learning methods are difficult to design crowd features, and it is difficult to obtain one or a set of features to meet the challenges of crowd density, crowd occlusion, different scenarios and perspectives encountered in crowd counting problems. In recent years, due to the improvement of computer computing power, especially the significant enhancement of graphics card performance, deep learning technology has promoted many computer vision problems to achieve very great breakthroughs. Massive data training enables deep learning to obtain powerful feature representation capabilities, which can well cope with the challenges and difficulties commonly found in crowd counting, and obtain more accurate counting results.

Wang et al. [3] proposed a deep network based on the Alexnet architecture for high-density crowd counting, using multi-layer convolutional layers to extract features. Using a fully connected layer predicts the number of extremely dense crowds. The MCNN [1] network uses convolution kernels of different sizes to adapt to the head sizes of different scales, and then combines the three

*Corresponding author. Email: zhao_huailin@yahoo.com

columns of CNNs to obtain the final crowd density map. Sam et al. [4] proposed Switching CNN, which trains several independent CNN crowd density regressors on the image patches. The regressor also uses a multi-column convolution structure, but these multi-column networks are low in efficiency and computationally costly. Wang et al. [5] added a density pre-classification network to the multi-column network. The problem of scale change is dealt with by pre-classification of density. Since then, Li et al. [2] proposed a dilated CNN (CSRNet). Using a convolution kernel with dilated instead of the pooling layer and the convolution layer reduces the amount of data and expands the receptive field range without losing the spatial resolution. In order to regulate the multi-scale density map, Varior et al. [6] also proposed a new scale-perceived loss function to guide the network to specialize in specific head sizes.

In the last 2 years, attention models have been widely used in various types of deep learning tasks. Zhu et al. [7] proposed a new two-way multi-scale fusion network structure SFANet. This method generates the final high-quality and high-resolution density map through attention mechanism and multi-scale fusion. Wang et al. [8] Introduce the attention mechanism to the crowd counting problem, guide the network to pay more attention to the crowd's head position through the attention module, and suppress background noise by giving less weight to other unimportant regional features.

The above research shows that the problem of crowd picture scale changes has always been the focus and difficulty of crowd counting tasks. In a picture, due to the change in the distance of the camera, two similar people show significant differences when they are at different distances from the camera. For this reason, we can consider increasing the connection between contexts to improve the accuracy of crowd density estimation. Based on the above research, this paper proposes a Graph Convolution-based Global Reasoning Network (GGRNet). Global reasoning units are used to model long-range regional relationships of feature pictures, and context information is added for the final population density estimation.

3. PROPOSED METHOD

In this section mainly introduces the overview of GGRNet for crowd counting. Then presents the backbone network and upsampling. Finally, we introduced the Global Reasoning Unit (GReU) in detail.

3.1. Overview

Convolutional neural network has made great progress in crowd counting task in recent years. Its biggest characteristic is local connection and weight sharing. Compared with other neural networks, it enhances the ability of feature learning and greatly reduces the number of parameters. However, it is precisely because of the local convolution mode that the features extracted by the convolution neural network are limited by the size of the convolution kernel and lack of a large enough receiving field. Only local regions are modeled, not global regions, resulting in the absence of some global features. In the crowd density estimation algorithm based on convolution neural network, context information is essential for the crowd counting network to accurately estimate the number of people, especially in crowded scenes. The limitation of convolution operation will lead to the network cannot effectively capture context information from crowd scene.

Graph convolution-based reasoning methods have been widely used in deep neural networks in recent years, and have greatly improved network performance in many tasks. For example, the GCN network proposed in Kipf and Welling [9] is used for semi-supervised classification. Sarlin et al. [10] used graph neural networks to learn feature matching. This paper proposes a graph-based global reasoning network for crowd counting. By adding the global reasoning unit to increase the global information of the crowd picture, the problem of information loss caused by convolution operation is compensated.

The overall flowchart of the algorithm is shown in Figure 1.

3.2. Backbone and Upsampling

In this paper, we use the first 10 layers of the VGG-16 as our backbone network because of its strong learning ability and flexible architecture, which is convenient for connecting density map generation at the back end. Input the picture and extract the shallow feature information of the crowd image through the convolution operation of the first 10 convolution layers of the VGG-16 network. Its model structure is shown in Figure 2. In the middle, it experienced three maximum pooling processes, and the output feature map size became 1/64 of the original input picture.

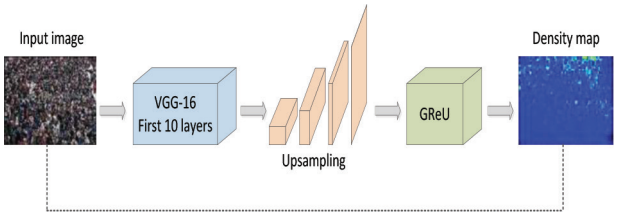

Figure 1 | Crowd counting algorithm based on global reasoning.

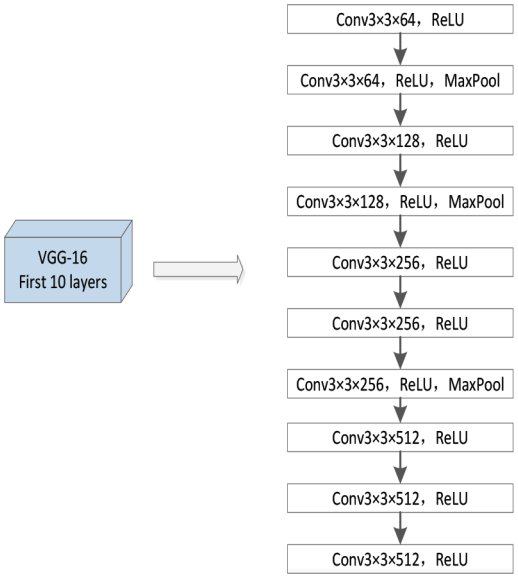

Figure 2 | Structure of the first 10 layers of VGG-16.

Since the feature image output by the VGG-16 network is only 1/64 of the original image size, and the GReU is used to model the long-range area of the crowd image, a design is designed between the VGG-16 network and the GReU upsampling operation to enlarge the image. In this paper, a bilinear interpolation upsampling method is used, and a convolution process is added before and after it to help correct the feature deviation due to upsampling. The model structure is shown in Figure 3.

3.3. Global Reasoning Unit

Referring to the literature [11], a GReU is added to the network. Since a single convolutional layer can only capture the local area relationships covered by the convolution kernel, and the relationship between long-distance areas needs to be achieved by superimposing multiple convolutional layers, this is not only cumbersome and inefficient, but also for network training added difficulties.

In order to overcome this limitation and better obtain the global context feature information, a GReU is selected to be added to the network to project the crowd characteristics on the coordinate space to the interactive space of global context reasoning. As shown in Figure 4, the GReU consists of five convolutions. Two are used for size reduction and expansion (leftmost and rightmost) on the input feature X and output Y. One is used to generate a double projection B between the coordinates and the potential interaction space (Top).

For the interactive space (middle), its implementation of graph convolution using two-direction 1D convolutions, and can be formulated as:

$$Z = ((I - A_g)V)W_g \tag{1}$$

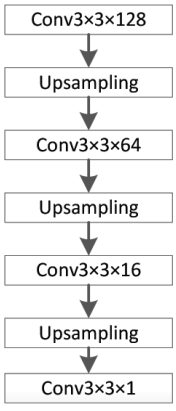

Figure 3 | Design of upsampling network.

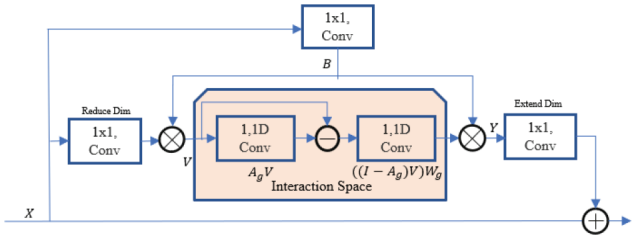

Figure 4 | Architecture of the proposed global reasoning unit.

where A_g and W_g represents the parameters of graph convolution. Here, V encodes the regional features as graph nodes, Z represents the feature map of interactive space output.

In this interaction space, regions with similar semantic features are represented by the same feature node, and their features are stored to form a fully connected graph. The interaction relationship between each node in the fully connected graph is modeled to obtain global context information through the convolution operation. The network then maps the resulting global feature map back from the interaction space to the feature space, combines it with the crowd features that were originally extracted, and outputs the final crowd density estimate map. Sum all pixels in the density map to calculate the total number of people in the crowd picture.

4. EXPERIMENTS

4.1. Dataset

In this paper, the UCF-QNRF [12] dataset is used to train and test the network. The UCF-QNRF dataset contains buildings, vegetation, sky, and roads in real scenes captured in the wild. And it including a variety of scenes with the most diverse perspectives, density, and lighting changes, making the dataset more realistic and representative. The UCF-QNRF dataset contains a total of 1535 pictures, among them, 1201 pictures are used as the training set and 334 pictures are used as the verification set. And the number of labeled heads reached 1,125,642, with an average of 815 people per picture.

The crowd picture resolution in the dataset is relatively large, with an average resolution of 2013 × 2902. The head size of the crowd also varies greatly, and the number of people per picture ranges from 49 to 12,865, which is especially suitable for training deep CNNs.

4.2. Density Map Generation

Crowd density estimation during network training, the head position coordinates given in the data set first be converted into corresponding crowd density pictures. In this paper, a method for generating a population density map based on an adaptive Gaussian kernel is used. Its formula is as follows:

$$F(x) = \sum_i^N \delta(x - x_i) * G_{\sigma_i}(x),\ \sigma_i = \beta \bar{d}^l \tag{2}$$

where $G_{\sigma_i}(x)$ represents the Gaussian convolution kernel, x_i is the position coordinates of the human head in the image, $\delta(x - x_i)$ is the Dirac function of the human head, N is the total number of people included in the image, and $\bar{d}^l = \frac{1}{m}\sum_{j=1}^m d^i{}_j$ represents the average distance of the m heads closest to the head. In denser cases it is approximately equal to the head size. β is a hyperparameter, here it takes 0.3. The loss function formula to be optimized by the network is as follows:

$$L(\theta) = \frac{1}{2N}\sum_{i=1}^N \left\| \hat{y}(x_i;\theta) - y_i \right\|_2^2 \tag{3}$$

where θ represents the parameters to be optimized by the network, N is the number of pictures in the training set, x_i represents the input picture, $\hat{y}(x_i;\theta)$ represents the crowd density map estimated

by the network, and y_i represents the crowd density truth value picture corresponding to the input picture.

For network training and implementation, we chose the Pytorch framework. When training the network, the Adam algorithm was used to optimize the network, and the initial learning rate was set to 0.00001.

4.3. Evaluation Metric

At present, the evaluation of the pros and cons of the crowd counting algorithm mainly includes the following two indicators, Mean Absolute Error (MAE) and Mean Squared Error (MSE).

$$\mathrm{MAE} = \frac{1}{N}\sum_{1}^{N}|z_i - \hat{z}_i| \tag{4}$$

$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum_{1}^{N}(z_i - \hat{z}_i)^2} \tag{5}$$

where N represents the total number of pictures in the test sequence, z_i represents the actual number of people in the picture, and $\hat{z}_i$ represents the estimated number of people in the picture.

4.4. Experimental Results

The graph-based global inference network is used for crowd counting. The experimental results on the UCF-QNRF dataset are shown in Table 1.

As can be seen from the table above, the method proposed in this paper is tested on the UCF-QNRF dataset, with MAE of 111.2 and MSE of 189.4. Compared with other more advanced crowd density estimation networks, this method has the smallest test error, and the network performance improves significantly. In terms of MAE and MSE, our method is 7.6% and 9.2% higher than the second-placed CSRNet, respectively.

In order to more clearly see the effect of this method on crowd density estimation, the density map output by the network was visualized and compared with the true value density map. Five images with different density levels were randomly selected for display, as shown in Figure 5.

In Figure 5, the first column is the original crowd image, the second column is ground truth of image, and the third column is the network-estimated crowd density map. In addition, the sum of all pixels of the density graph is given. It can be seen from the figure that the population density map estimated by the network can vividly reflect the population distribution.

Table 1 | The experimental results on the UCF-QNRF dataset

Method	MAE	MSE
Idrees 2013 [13]	315	508
MCNN [1]	277	426
CMTL [14]	252	514
Switching CNN [4]	228	445
CL [12]	132	191
CSRNet [2]	120.3	208.5
Our proposed	**111.2**	**189.4**

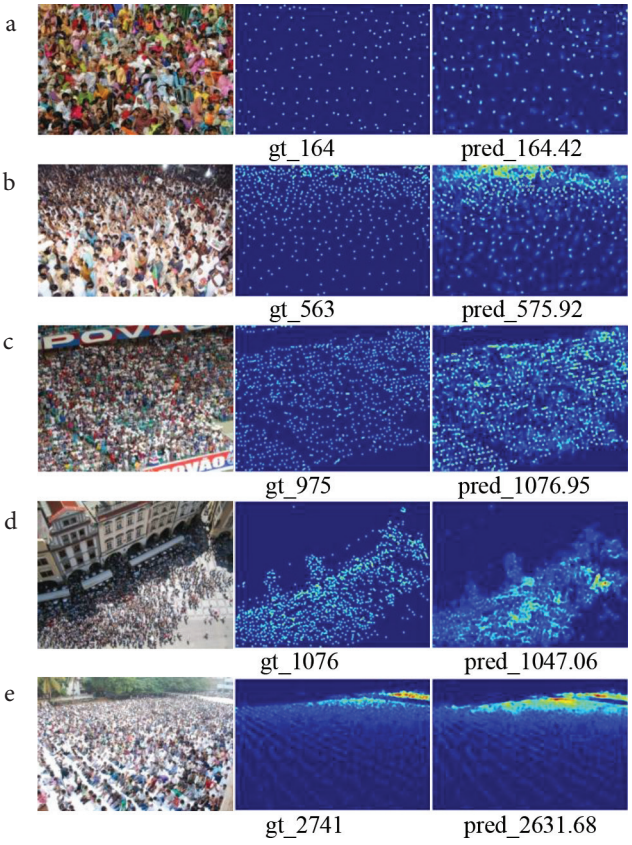

Figure 5 | Visualization of density map.

5. CONCLUSION

A graph-based global reasoning network for crowd density estimation tasks is proposed in this paper. The crowd feature information extracted by the VGG-16 network within the global area is modeled. The GReU is applied to establish the relationship among long-distance areas in the picture, the context information is added, and the supplementary features for the generation of crowd density maps are provided. Experimental results on the UCF-QNRF dataset show the effectiveness of the method. In the future work, we will continue to optimize the network structure. Under ensuring accuracy, the network will be quantified and tailored, and crowd density will be monitored in real-time.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

REFERENCES

[1] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network,

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Las Vegas, NV, USA, 2016, pp. 589–597.
[2] Y. Li, X. Zhang, D. Chen, CSRNet: dilated convolutional neural networks for understanding the highly congested scenes, Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, USA, 2018, pp. 1091–1100.
[3] C. Wang, H. Zhang, L. Yang, S. Liu, X. Cao, Deep people counting in extremely dense crowds, Proceedings of the 23rd ACM International Conference on Multimedia, Association for Computing Machinery, 2015, pp. 1299–1302.
[4] D.B. Sam, S. Surya, R. Venkatesh Babu, Switching convolutional neural network for crowd counting, Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, USA, 2017, pp. 4031–4039.
[5] S. Wang, H. Zhao, W. Wang, H. Di, X. Shu, Improving deep crowd density estimation via pre-classification of density, Proceedings of the IEEE International Conference on Neural Information Processing, Springer, Cham, 2017, pp. 260–269.
[6] R.R. Varior, B. Shuai, J. Tighe, D. Modolo, Multi-scale attention network for crowd counting, arXiv preprint arXiv:1901.06026, 2019.
[7] L. Zhu, Z. Zhao, C. Lu, Y. Lin, Y. Peng, T. Yao, Dual path multi-scale fusion networks with attention for crowd counting, arXiv preprint arXiv:1902.01115, 2019.
[8] L. Wang, H. Zhao, Y. Li, Research on the multi-scale network crowd density estimation algorithm based on attention mechanism, Proceedings of the 2019 International Conference on Intelligent Informatics and BioMedical Science (ICIIBMS), IEEE, Shanghai, China, 2019, pp. 272–278.
[9] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907, 2016.
[10] P.E. Sarlin, D. DeTone, T. Malisiewicz, A. Rabinovich, SuperGlue: learning feature matching with graph neural networks, arXiv preprint arXiv:1911.11763, 2019.
[11] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, Y. Kalantidis, Graph-based global reasoning networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 433–442.
[12] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, et al., Composition loss for counting, density map estimation and localization in dense crowds, Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 532–546.
[13] H. Idrees, I. Saleemi, C. Seibert, M. Shah, Multi-source multi-scale counting in extremely dense crowd images, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Portland, OR, USA, 2013, pp. 2547–2554.
[14] V.A. Sindagi, V.M. Patel, CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, Lecce, Italy, 2017, pp. 1–6.

AUTHORS INTRODUCTION

Ms. Li Wang

She received the M.S. degree from Shanghai Institute of Technology, Shanghai, China, in 2020. She is a computer teacher in the School of Shanghai Communications Polytechnic. Her main research interests are deep learning and intelligent information processing.

Dr. Huailin Zhao

He received his PhD from Oita University, Japan in 2008. He is a professor in the School of Electrical & Electronic Engineering, Shanghai Institute of Technology, China. His main research interests are robotics, multi-agent system and artificial intelligence. He is the members of both IEEE and Sigma Xi.

Mr. Fangbo Zhou

He received the B.S. degree in automation from Chuzhou University, Chuzhou, China, in 2019. He is currently pursuing the M.S. degree in safety engineering at Shanghai Institute of Technology, Shanghai, China. His research interests include crowd counting and image processing.