Multimedia Tools and Applications (2020) 79:1057–1073
https://doi.org/10.1007/s11042-019-08208-6

# Multi-scale dilated convolution of convolutional neural network for crowd counting

Yanjie Wang[1] · Shiyu Hu[1] · Guodong Wang[1] · Chenglizhao Chen[1] · Zhenkuan Pan[1]

Received: 21 January 2019 / Revised: 30 July 2019 / Accepted: 12 September 2019 /
Published online: 17 October 2019

## Abstract

Growing numbers of crowd density estimation methods have been developed in scene monitoring, crowd safety and on-site management scheduling. We proposed a method for density estimation of a single static image based on convolutional neural network naming Multi-scale Dilated Convolution of Convolutional Neural Network (Multi-scale-CNN). The proposed method employed the method of density maps regression to learn the mapping relationship between single-image and density maps through convolutional neural network. The adopted network structure is composed of two major components to adapt changes of characters scales in crowd images, a convolutional neural network for the general feature extraction and the other is multi-scale dilated convolution for disposing the scale change problem. It is insufficient for currently study that tackled the multi-column or multi-input convolutional neural networks to solve multi-scale problems. Our method utilizes a single-column network to extract features and combines multi-scale dilated convolution to aggregate multi-scale information to address the shortcomings of two networks. The multi-scale dilated convolution module aggregates multi-scale context information systematically by making use of dilated convolution without reducing the receiving domain, thereby integrate the underlying detail information into the high-level semantic features to promote the perception and counting ability of network for small targets. This paper demonstrates the proposed network structure in ShanghaiTech dataset, UCF_CC_50 dataset and worldexpo'10 dataset, and compares the results with numbers of current mainstream crowd counting algorithms, proves that our method surpasses current state-of-the-art methods and has excellent counting accuracy and robustness. The training and testing codes of our method models can be downloaded at https://github.com/doctorwgd/Multi-scale-CNN.

**Keywords** Image processing · Crowd counting · Deep learning · Dilated convolution

✉ Guodong Wang
doctorwgd@gmail.com

[1] College of Computer Science and Technology, Qingdao University, Qingdao 266071, People's Republic of China

## 1 Introduction

Extensive crowd flows in cities are commonplace at railway station during holidays, shopping malls for holiday parties and supermarkets for promotion. The crowd is prone to disasters and has great security risks of daily life, and if the department concerned has scanty ability to evacuate residents swiftly and diligently at the peak traffic.

The significant point behind estimate crowd density in the public infrastructure accurately is self-evident: control and manage the crowd effectively. The statistical data prevents diversity dangerous accidents, such as stampede and riot, follows the demand of real-life applications. Customers' buying preferences and potential commercial value can be obtained by analyzing the crowd density distribution of shopping malls. Hence, crowd counting has a comprehensive range of application and social significance, especially in the field of intelligent monitoring and security. All above could be critical for making correct decisions in high-risk environments and growing numbers of crowd counting problems in single images and surveillance videos in recent years. The challenging status quo of severe occlusion, changes in illumination, distortion of viewing angles, uneven population distribution and complex scene backgrounds in real-life scenarios has no effective algorithm can be applied at present, which makes the accurate and robust crowd counting algorithm indispensable of the important research directions in the field of computer vision.

### 1.1 Related work

The current crowd counting algorithm is divided into two natural configurations, including the detection theory based methods and the regression theory based methods.

The detection-based algorithm trains a detector to identify individual individuals in the image and receives the sum through the individuals. Early detection-based algorithms introduced Haar wavelets [28] or histogram oriented gradients (HOG) [9] as features detector required the detection of the whole body. With the decrease of resolution, the increase of the crowd and the change of viewing angle limited by the mutual occlusion of the crowd in dense environments, its accuracy rate is gradually reduced and robustness is poor [14, 15]. Following the idea proposed before, the potential solutions for crowd scenes analysis [10, 15] make an attempt to count by detecting a part of the body, and the algorithm of detecting the head or shoulder in the algorithm [14] for detecting a part of body improved the performance of such kind of algorithms effectively. Although the partial detection methods mitigate the impact of crowd overlap and occlusion on the performance of the algorithm, the detection-based algorithm is still inadaptable for dense crowd scenarios. In addition, the detection-based algorithm generates a mass of computational resources and computation time via traverses the trained detection window across the whole image.

Scholars get the mapping relationship between the characteristics of the crowd and the number of people then regress the number of people in the image in order to improve the counting performance in dense crowd scenarios. The algorithm is divided into two procedures: low-level feature extraction and regression framework. The global, local, and texture features of the scene are mediums to generate the low-level information of the image. Linear regression [20], ridge regression [5] and neural networks as a series of regression processes regular learn the mapping between the low-level information and the total number of crowd after low-level information extracted. Regression-based algorithms have succeeded in dealing with overlap and occlusion problems, these algorithms have deficient in returning the amount of crowd which arise certain important loss of spatial information. Lempitsky et al. [13] proposed to achieve crowd counting by estimating the density map as a whole. By capturing density map, the unpitched noise caused by the image could be reduced availably.

At present, Artificial Intelligence (AI) develops rapidly, and it has received great attention both in academia and IT industry. Convolutional neural network is one of the most commonly used models in deep learning. It has been applied in many aspects, such as Document clustering [1, 2], Retrieval [6, 11], Object Recognition [22].

CNN-based counting method in different scenarios that pre-trained a part of the picture in the scene to obtain the scene network model was proposed by Zhang [24] et al., only fine-tune the pre-trained network model according to the perspective of the new scene to show good performance on the test set when testing a new scene; but it rely on pre-training and requires perspective in the new scene which greatly limits its practical application.

MCNN [32] is a three pre-trained columns CNN structure. Intuitively, each column of MCNN is dedicated to a certain level of congested scene and different columns extract features of different scales. Sparse kernels of different sizes have employed in each column. Large-scale receptive fields are larger and easier to extract features of large-scale characters while small-scale kernels tackle the characteristics of small-scale characters, the structure of multi-columns handle changes in the crowd image effectively compared with a single-column CNN. However, it is necessary to train each column separately and then take the pre-trained network of columns to concat a multi-column structure and fine-tune the training finally while the complex structure and more parameters of the multi-column network make the training produce become complicated. DRSAN [16] is also used in the same structure, increasing the redundancy and computation of the network. Sam [23] and others who under the influence of Zhang et.al adopted a three-column network structure in response to the phenomenon. The added VGG-based classifier before the three-column network structure makes a difference. The input separated according a nine-square grid and the classifier determines which column network is the option of processing the segmented image. The final effect has a very obvious improvement, but trained classifiers require numbers of complicated training.

Recently, face detection [21] and target recognition [30] achieved excellent results by means of multi-tasking and apply multi-task to the field of crowd counting becomes a phenomenal status quo. Marsden et al. [18] proposed a deep residual network called ResnetCrowd while carried out crowd counting, violent behavior detection and image population density classification. Influenced by the application of cascaded convolutional neural networks in multi-task learning [3, 4, 8, 21], Sindgai et al. [12] taken a cascaded CNN to learn simultaneously. Image density classification and density map regression in another method.

Essentially, crowd density estimation is a perception and detection of pedestrian. This paper adopts the density map regression method based on pedestrian to solve the occlusion problem. Compared with the traditional detection-based methods, the CNN-based crowd counting algorithm improves the counting accuracy greatly and reduces the occlusion effectively. The model has the ability to perceive multi-scale targets to solve the multi-scale target problem caused by the projection effect. The core problem the crowd density estimation in the open scene ought to figure out is the perception of multi-scale targets and small targets. Unavoidably, the current crowd counting algorithm still has the problem of poor computing ability for small targets in high-density crowd images. The research in this paper is focused on solving the problem of scale change and small target counting. The main contributions include the following:

(1) This paper proposes a population density estimation algorithm based on multi-scale dilated convolution feature maps fusion convolutional neural network (Multi-scale-CNN) on the basis of the strengths and weaknesses of vast existing crowd counting algorithms. The algorithm adopts the method of density map regression, which has a capacity to count the crowd of any resolution images without pre-processing. For handling the problem of human scale change and small target counting in crowd images, we segment the convolutional neural network in the algorithm into two sub-networks, a convolutional neural network as the front-end for feature extraction and a multi-scale dilated CNN for convolution feature fusion. Multi-scale dilated convolution settled the multi-scale perception problem and feature fusion is utilized to improve the counting performance of the algorithm for small targets in high-density crowd images.

(2) Since the scale change problem morbidly, the multi-scale dilated convolution adopts the single-column convolutional neural network to append the design of the dilated convolution module with multiple receptive fields. The characters of various scales in the image are performed by setting different dilated factors. Feature extraction. Multi-scale dilated convolution is composed of multiple columns of convolutional layers with different receptive fields stacked side by side and each column is consisted of different dilated convolution, characters of different scales are extracted, feature extraction module is adapted to scale changes in images by integrating features extracted from multi-column convolutional layers. The feature extraction module adapts to scale changes in the image by fusing the features extracted from the multi-column convolutional layers.

(3) In order to figure out the small target counting, the feature fusion part fuses the maps extracted by multi-scale dilated convolution, obtains advanced features and low-level features and uses the feature map to generate the density maps and improves the algorithm to high density. Improve the counting ability of the algorithm for small and medium targets in high-density crowd images.

(4) We conducted a series of experiments on ShanghaiTech [32], UCF_CC_50 [27] and worldexpo’10 [31], three current mainstream crowd counting datasets. The experimental results are compared with some state-of-the-art crowd counting algorithms to prove the accuracy and robustness of the proposed Multi-scale-CNN.

## 2 The proposed method

### 2.1 Multi-scale and dilated convolution neural networks

Our research is based on GoogLeNet [27] which has achieved through multi-scale convolutional neural networks. Employed inception structure, the structure makes good use of the computing resources in the network, and increases the width and depth of the network without increasing the computing load. In order to optimize the network quality, Hebbian principle and multi-scale processing are adopted synchronously. GoogLeNet [27] has achieved good performance in classification and detection. We demonstrate the splitting of the convolutional layer in the segmentation task, its accuracy improved significantly and suitable for the pooling layer, pooling layers are taken to maintain invariance and avoid over-fitting broadly, they reduce the spatial resolution greatly, which means the spatial information of the feature map is missed. Dilated convolution uses a “perforated” filter to alternately convolve and concatenate the layers. This layer expands the receptive field without increasing the number of parameters or the amount of computation. In other words, the same number of layers we used can achieve a larger receptive field without introducing more operations than ordinary convolution. Fisher Yu [29] used the dilated convolution to achieve high accuracy in semantic segmentation tasks.

In the dilated convolution, the kernel with a $k \times k$ filter is amplified to $k + (k\text{-}1)\ (r\text{-}1)$ with the dilation rate r. Therefore, it can be flexible to aggregate multi-scale contextual information while maintaining the same resolution. The normal convolution gets 3 × 3 receptive field, and the dilated convolution of the dilation rate of 2 and 3 provides 5 × 5 and 7 × 7 receptive field, shown in Fig. 1. Obviously, a 3 × 3 convolution kernel with the dilation rate of 2 has the same receptive field as the 5 × 5 convolution kernel, but the number of parameters is only 9 and is 36% of the 5 × 5 convolution parameter number. The same 3 × 3 convolution kernel with the dilation rate of 3, the receptive field is the same as the 7 × 7 convolution kernel, but the number of parameters is only 9, which is 18% of the 7 × 7 convolution parameter number.

## 2.2 Multi-scale-CNN network structure

By the influence of shooting Angle, height, distance and other factors, the size of people in the crowd images have diverse from each other. Undoubtedly, people who closed to the shooting lens occupy more pixels in the picture and presented on a larger scale while those far away from the lens reveal on a smaller one. Moreover, the changes in the scale of people in different crowd pictures which leads to the problem of scale changes of people in crowd images. CNN has invariance to the translation and rotation of the figure in the picture due to the weights sharing and mechanism of local connection and has strong ability of dealing with the change of people's appearance and position. However, CNN does not have the capacity for immutability for human scale changes. Therefore, if the crowd counting algorithm deals with the multi-scale problem, it needs CNN to be able to perceive the scale change of the figure in the images.

Therefore, we propose a convolutional neural network can adapt the image with variation scales. The point of real head has converted into a density map before training, and training is conducted through supervised network learning. As shown in Fig. 2.

Feature extraction: Composed by convolution activation and maximum pooling. Each of which is consisted of three convolution activated connections and a maximum pool. The

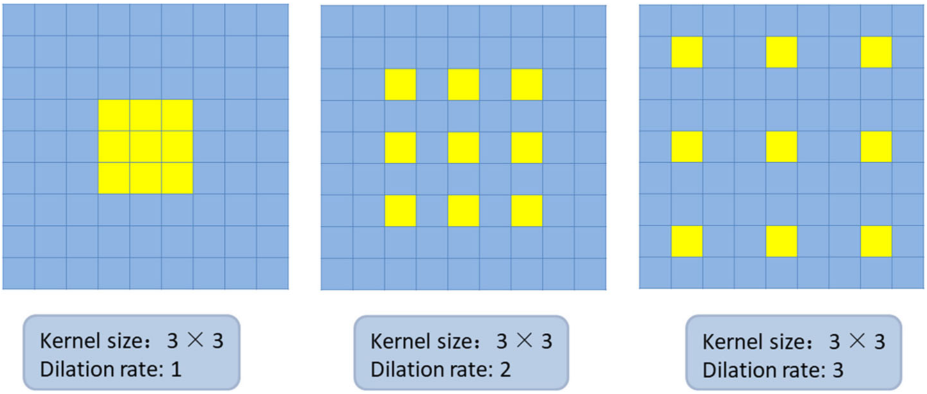


**Fig. 1** 3 × 3 convolution kernels with different dilation rate as 1, 2, and 3

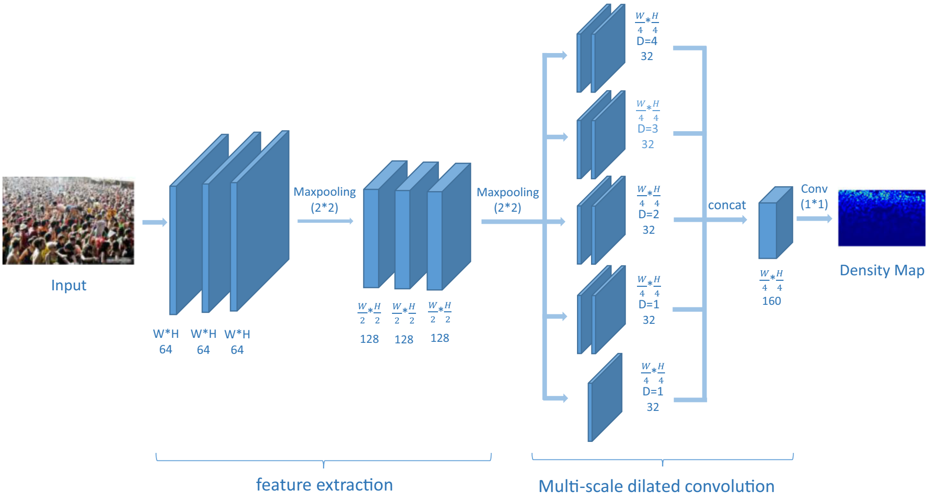

Input
W*H 64
W*H 64
W*H 64
Maxpooling (2*2)
$\frac{W}{2}*\frac{H}{2}$ 128
$\frac{W}{2}*\frac{H}{2}$ 128
$\frac{W}{2}*\frac{H}{2}$ 128
Maxpooling (2*2)
$\frac{W}{4}*\frac{H}{4}$ D=4 32
$\frac{W}{4}*\frac{H}{4}$ D=3 32
$\frac{W}{4}*\frac{H}{4}$ D=2 32
$\frac{W}{4}*\frac{H}{4}$ D=1 32
$\frac{W}{4}*\frac{H}{4}$ D=1 32
concat
$\frac{W}{4}*\frac{H}{4}$ 160
Conv (1*1)
Density Map
feature extraction
Multi-scale dilated convolution

**Fig. 2** The architecture of Multi-scale-CNN network

convolution of the first step generates 64 feature maps while the second part generates 128. More filters used, more features obtained, to make up for the damage of detailed features as the size is reduced by max-pooling.

Multi-scale dilated convolution: Five channels with different receptive fields to obtain the characteristics of different crowd densities. The first channel uses the convolution of a 64 filters, the second channel has two convolution operations convolved in series with 64 kernels, the third channel to the fifth employ the convolution of two dilated factors of 2, 3, 4, 64 kernels respectively. Taken the operations convolved with 1 × 1 to get a feature map which is the final cowed density feature map ultimately.

In conclusion, our crowd density model of Multi-scale-CNN has three main characteristics. Using a uniform feature extract parts, the computation is economized and the redundancy of computation is reduced. Multi-scale convolution is introduced to get the feature maps from different receptive fields to achieve good results for different crowd densities. The dilated convolutional neural network is used to make the difference in the receptive fields of different channels larger without increasing the amount of calculation to adapt the change of population density well.

Euclidean distance is measured the distance between the labeled density map and the predicted density map to carry out the supervised training by the network. The loss function of the network is defined as follows:

$$L(\theta)=\frac{1}{2N}\sum_{i=1}^{N}||E_i(I_i;\theta)-D_i||_2^2 \tag{1}$$

where $\theta$ is a parameter to be learned in a fully convolutional neural network. $I_i$ represent the $ith$ training image on the training set and $N$ is the number of images in the training set. $D_i$ and $E_i$ represent the labeled density map and the predicted density map by the full convolutional neural network respectively. $L(\theta)$ make the loss as a result of calculating the Euclidean distance between the plotted density map and the predicted density map.

Two pooling layers employed on account of the size of a human head usually bulky and the distance between people and the camera is closed in the real scene, result in the fact when designing the network the receptive fields cannot be diminutive. Although the utilize of pooling layer reduce the resolution of the final predicted density map, it increases the sensing field of the network which is conducive to the response of the network to large heads and the learning of more meaningful feature representations.

## 3 Experimental results

### 3.1 Multi-scale-CNN network structure

Mean absolute error (MAE) and mean square root error (MSE) take as evaluation criteria to evaluate proposed Multi-scale-CNN and the state-of-the-art crowd counting methods. MAE and MSE of the test dataset with N images are defined as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{2}$$

$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{3}$$

where $y_i$ and $\hat{y}_i$ represent the number of people in the i-th test image, the number of people from human labeled annotations and the number of people predicted. MAE calculates the average absolute error and represents the average difference between predicted results and the actual. RMSE is used to calculate the mean square error and indicated the degree of dispersion between the differences. MAE is the indicator to test the accuracy of predicted crowd counting in the test set and MSE is used to evaluate the robustness of the algorithm.

### 3.2 Experimental environment and settings

The implementation of the proposed method's training and testing are based on the Pytorch [7], a framework of deep learning. Generating the label density map and setting the width parameter of the two-dimensional Gaussian convolution kernel to 4. Symmetric left and right operations are performed on the training set in order to prevent the network from over-fitting. Specifically, nine quarter-sized image blocks are cropped from different regions of the original image. The size of the final output density map is one quarter of the original because the network uses two filters with a kernel size of 2 × 2 and a step size of 2 results in the label density used when the network calculates the loss is one quarter of the original size. In the training phase, the initial learning rate of the network is set to 0.0001and reduced by 10 times while the training loss is no longer reduced. The network optimization algorithm is a batch random gradient descent method (SGD with momentum) with a batch size of 32 and an impulse of 0.9.

Multi-scale-CNN requires a crowd image and its corresponding density map at the training phase. The existed crowd counting dataset only provides the coordinate values of the head annotations in the image as data labels, but not directly used for network training make it necessary to utilize these datasets before converting the coordinate values of the human head annotations in the image to a data tag for crowd counting.

For Multi-scale-CNN, the quality of the density map generation of the training data affects the performance of the algorithm intuitively. The density map generation used in the proposed method will be described below. Assuming there is a personal head annotation at the coordinate $x_i$ in the image, the point can be represented as a δ function $\delta(x-x_i)$. Thus, for a image with N personal head points, it can be expressed as:

$$H(x)=\sum_{i=1}^{N}\delta(x-x_i) \tag{3.1}$$

However, the density equation is discrete. A Gaussian filter $G_\sigma$ can be convolved with the density equation to obtain a density equation as shown in Eq. (3.2) to convert it into continuous,.

$$F(x)=H(x)*G_\sigma(x) \tag{3.2}$$

$$F(x)=H(x)*G_{\sigma_i}(x),\quad \sigma_i=\beta\overline{d_i} \tag{3.3}$$

Two test images from the ShanghaiTech dataset have been selected in the article to generate respective density maps. The two images and respective corresponding density maps are shown in Fig. 3.

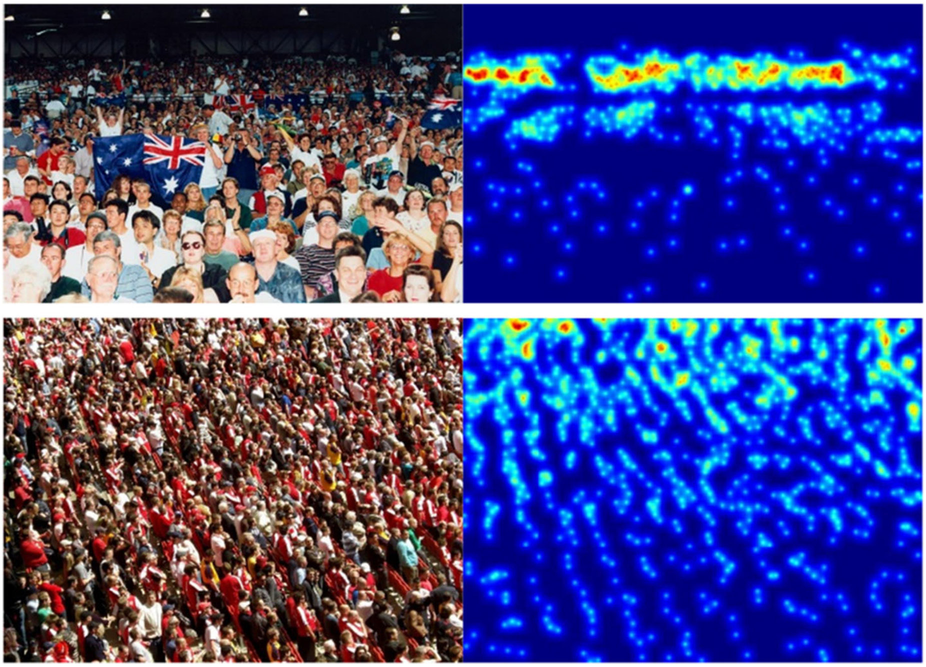


**Fig. 3** Original images and corresponding crowd density maps obtained by convolving geometry-adaptive Gaussian kernels

### 3.3 Evaluation and comparisons

The experiment is conducted on three public standard datasets, the most widely applicable ShangHaiTech dataset, the UCF_CC_50 dataset and worldexpo'10 dataset. Table 1 lists the relevant information for the three public datasets. In Table 1, Number is the total number of pictures in the dataset. Max is the maximum number of single images while Min represents the minimum number of single images, Average is the average number of single images in the dataset and Total express the total number of people in the dataset, Resolution is the resolution of the image in the dataset.

#### 3.3.1 ShanghaiTech dataset

The ShanghaiTech dataset, a standard large dataset, has the largest number of labeled populations and divided into two elements, Part A and Part B.

Part A is image selected from the Internet containing a total of 482 annotated images with a different resolution. In Table 1, the total number of people in each image is 501.4, the minimum is 33, and the maximum is 3139. The total number of people reaches 241,677, and the population density is high. The dataset has been split by the publisher into a training dataset include 300 images as a test dataset remain 182 verification images.

Part B, relatively sparse images with a total of 482 annotations. The resolution of each image is 768 × 1024. As can be seen from Table 1, the average number of people per image is 123.6, at least 9 people, and the maximum is 578 people. The total number of people reaches 88,488, and the crowd density is low. The data set contains 400 training pictures and 316 verification pictures.

For the ShanghaiTech dataset, the classical symmetry enhancement method has been used to obtain more data without over-fitting, which only performs left-right symmetry and doubles the number of images. Then per picture of module is cut into 9 pieces. The size of each picture is 1/4 of the original and the interval is 1/6 of the picture, so the original picture is enhanced to 18 pieces. Thus, Part A's training set has 5400 images while Part B has 7200. The test set is not processed.

We compared the performance of Multi-scale-CNN with the current five crowd counting algorithm on the ShanghaiTech dataset. Zhang et al. [17] designed a single-column structure has capable of cross-scenario counting and learned the number and density map from the picture simultaneously; Marsden et al. [25] proposed a simple single-column CNN structure inputs results of different scale pictures to the total number of people by converting the pictures into different proportions; MCNN [32] is a three-column CNN structure. Different columns of

**Table 1** Data information of two existing datasets

| Dataset | ShanghaiTech | | UCF_CC_50 | worldexpo'10 |
|---|---|---|---|---|
| | Part A | Part B | | |
| Number | 482 | 716 | 50 | 3980 |
| Max | 3139 | 578 | 4543 | 253 |
| Min | 33 | 9 | 94 | 1 |
| Average | 501.4 | 123.6 | 1297.5 | 50.2 |
| Total | 241,677 | 88,488 | 63,974 | 199,923 |
| Resolution | different | 768 × 1024 | different | 576 × 720 |

networks have various sizes of convolution kernels, and the scale change problem is solved by the multi-scale feature extracted by the three-column network. This method learns the mapping relationship between pictures and density maps; Cascaded. MTL [26] belongs to two columns of CNN structure. This network performs simultaneous picture density classification and density map regression based on cascaded CNN structure and multi-task learning. Switching-CNN [23] is connected by a picture density classifier and multi-column CNN. Composition, after the picture is input into the network, the classifier determines the density level first, and selects a column network from the multi-column CNN for density map regression according to the classification result of the classifier.

The comparison results of Multi-scale-CNN with five algorithms are shown in Table 2. Our method has a certain gap compared to CP-CNN [26] on Part A. This is because CP-CNN has a deeper network as a context predictor, although this increases the complexity of the network. But Table 2 indicates that Multi-scale-CNN achieves the best MAE on Part B. MSE is also excellent compared with Switching-CNN indicates that the crowd counting algorithm proposed in this paper has a strong ability on counting accuracy and robustness. Figure 4 shows examples of ground truth density maps and estimated density maps of images in Part A and B of ShanghaiTech.

### 3.3.2 UCF_CC_50 dataset

UCF_CC_50 [55], the most challenging crowd counting dataset currently, consists of 50 images taken from the network with a different resolution and angle of view. The scenes of these images are abundant, including common crowd-intensive scenes such as stadiums, marathons and concerts. A large imparity has been created to the number of people in different pictures. With an average of 1280, the minimum number is 94 and the maximum number occurs 4543. Although UCF_CC_50 has more high-density crowd pictures and the crowd density of datasets is larger than other existed datasets, the limited number of pictures in this database becomes the obvious flaw.

For the UCF_CC_50 dataset, we take the 5-fold cross-validation of Zhang et al. [23] to divide the data into 5 portions to test the effect of the model respectively and calculate the average of MAE and RMSE. The image is segmented to 1/4 of the original size and the cutting pitch is 1/12 in view of the limited number of images, so the picture has been enhanced to 36. The enhancement of the test set is merely employed for model training, and the actual effect detection process remain the original picture without change.

**Table 2** Comparing results of different methods on the ShanghaiTech dataset

| Method | Part_A | | Part_B | |
|---|---|---|---|---|
| | MAE | MSE | MAE | MSE |
| Zhang et al. [31] | 181.8 | 277.7 | 32.0 | 49.8 |
| Marsden et al. [17] | 126.5 | 173.5 | 23.8 | 33.1 |
| MCNN [32] | 110.2 | 173.2 | 26.4 | 41.3 |
| Cascaded-MTL [25] | 101.3 | 152.4 | 20.0 | 31.1 |
| Switching-CNN [23] | 90.4 | 135.0 | 21.6 | 33.4 |
| CP-CNN [26] | **73.6** | **106.4** | 20.1 | **30.1** |
| Multi-scale-CNN (ours) | 83.7 | 124.5 | **17.9** | 32.4 |

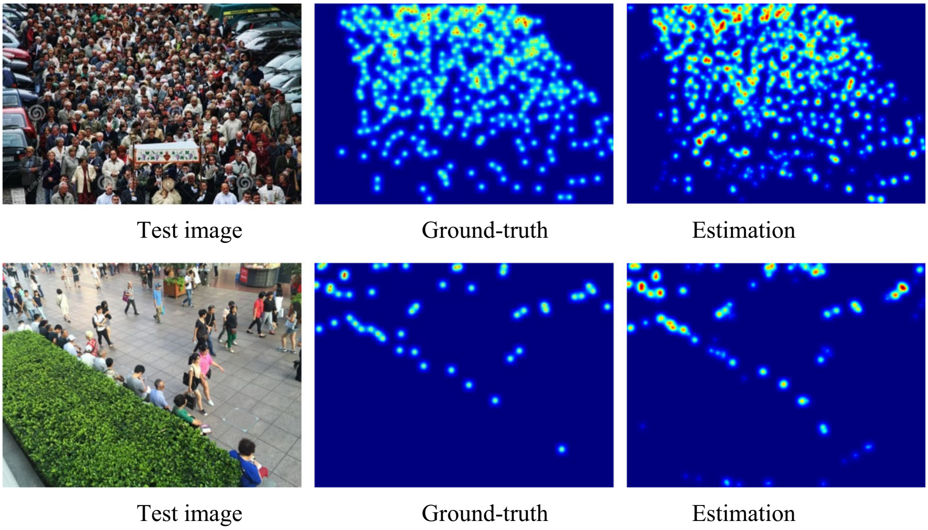


**Fig. 4** The ground truth density map and estimated density map of our method of two test images in part A and B of ShanghaiTech

The proposed Multi-scale-CNN, the same as the current crowd counting algorithm, introduce the five-fold cross-validation method on UCF_CC_50, train on five training sets and test on the corresponding set. The results of Multi-scale-CNN on the test sets are shown in Table 3. The average of results is taken as the counting performance of the algorithm on UCF_CC_50, and the MAE and MSE are 264.93 and 382.14, respectively.

We compared Multi-scale-CNN's test results on UCF_CC_50 with the existed seven mainstream crowd counting methods. The shallow CNN extracts the low-level features in the image as the deep CNN extracts the high-level semantic features and two features merge to generate the density map ultimately; MCNN [32] processes the image scale changes through the three-column CNN, and each column CNN contains different size convolution kernels, different columns extract features of different scales, and features extracted by three columns of CNNs are combined to obtain density maps; Marsden et al. [19] propose a simple single-column CNN structure, which is input into the network by changing the pictures into different proportions. The results obtained by different scale pictures are averaged as the final total number; Hydra-CNN [29] is a multi-input CNN structure, and the image construct an image pyramid structure with three pictures before inputting the network, then different resolutions is changed to the same input into different columns of the network, and the density map is generated by merging the feature extractions extracted by different columns; Cascaded-MTL

**Table 3** The result of 5-fold cross-validation on the ShanghaiTech dataset

| Test set | MAE | MSE |
|---|---|---|
| Fold-1 | 311.25 | 507.77 |
| Fold-2 | 267.73 | 381.61 |
| Fold-3 | 281.94 | 384.48 |
| Fold-4 | 214.28 | 303.89 |
| Fold-5 | 249.47 | 332.96 |
| Avg | 264.93 | 382.14 |

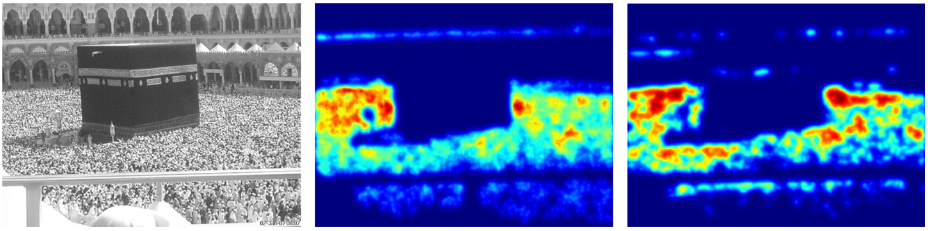
**Table 4** Comparing results of different methods on the UCF_CC_50 dataset

| Method | MAE | MSE |
|---|---|---|
| Zhang et al. [31] | 467.0 | 498.5 |
| MCNN [32] | 377.6 | 509.1 |
| Marsden et al. [17] | 338.6 | 424.5 |
| Hydra-CNN [19] | 333.7 | 425.2 |
| Cascaded-MTL [25] | 322.8 | 397.9 |
| Switching-CNN [23] | 318.1 | 439.2 |
| CP-CNN [26] | 295.8 | **320.9** |
| Multi-scale-CNN (ours) | **264.9** | 382.1 |

[26] Cascading CNN structure and multi-task learning simultaneously perform picture density level classification and density map regression; Switching-CNN [23] consists of a picture density classifier and a multi-column CNN series. After the picture is input into the network, the classifier determines the density level. According to the classification result of the classifier, a list of networks is selected from the multi-column CNN for density map regression. The performance comparison between Multi-scale-CNN and other seven algorithms on UCF_CC_50 shown in Table 4 indicates that the algorithm goes beyond other algorithms in the MAE index, and the optimal CP-CNN [11] among the seven algorithms has improved by 30.9, far exceeding Multi-scale-CNN on ShanghaiTech compared to Switch-CNN. On the MSE indicator, Multi-scale-CNN is second to all algorithms to obtain the optimal MSE Cascaded-MTL proves that Multi-scale-CNN has excellent counting accuracy and good robustness in high-density graph population pictures. Figure 5 shows examples of ground truth density maps and estimated density maps of images in UCF_CC_50.

### 3.3.3 worldexpo'10 dataset

Since some of the earlier methods and dataset focused primarily on single-scene counting, Zhang et al. [31] introduced a data set for crowd counting. The authors attempted a data-driven cross-scenario population count. They collected a new large data set, including 1132 annotated video sequences captured by 108 surveillance cameras, all from the 2010 Shanghai World Expo. The diversity of scene types is ensured by collecting video from cameras with disjoint aerial views. The data set includes a total of 3980 frames of size 576 × 720, of which 199,923 are marked as pedestrians. The dataset is divided into two parts: The training set contains 3380 frames and the test set contains 600 frames from five different scenes Each test set consists of



Test image Ground-truth Estimation

**Fig. 5** The ground truth density map and estimated density map of our method of the test image in UCF_CC_50

**Table 5** Comparing results of different methods on the worldexpo'10 dataset

| Method | Scene 1 | Scene 2 | Scene 3 | Scene 4 | Scene 5 | Avgerage |
|---|---|---|---|---|---|---|
| Zhang et al. [31] | 9.8 | 14.1 | 14.3 | 22.2 | **3.7** | 12.9 |
| MCNN [32] | 3.4 | 20.6 | 12.9 | 13.0 | 8.1 | 11.6 |
| Switching-CNN [23] | 4.4 | 15.7 | 10.0 | 11.0 | 5.9 | 9.4 |
| CP-CNN [26] | **2.9** | **14.7** | 10.5 | **10.4** | 5.8 | 8.86 |
| Multi-scale-CNN (ours) | 3.3 | 15.2 | **10.2** | **9.8** | 4.3 | **8.56** |

120 marked frames, with populations ranging from 1 to 220. For a fair comparison, generated the density map is used to generate a ground truth map of work similar to [31].

At the same time, we consider using ROI mapping to process the output density map generated by the network. We use the same evaluation metric (MAE) suggested by [31]. Table 5 reports the results of different methods on the worldexpo'10 dataset. Our method achieves state-of-the-art performance in terms of average MAE on five scenes.

## 4 Conclusion

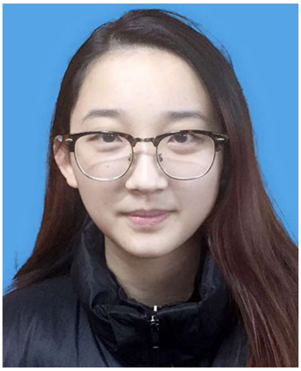
In this paper, a crowd density estimation algorithm based on multi-scale cavity convolution feature fusion convolutional neural network is proposed. The few parameters of the network reduce the complexity of the model and reduces over-fitting. Two pooling layers greatly increase the network's receptive fields, enable the network responds to a large-scale human heads. The crowd density map is predicted first through the network and then the number of people is further obtained, which acquire the distribution of the population in space while make the number of people obtained more accurate finally. The experiments and comparisons on the ShanghaiTech, UCF_CC_50 and worldexpo'10 standard datasets demonstrate the high accuracy and effectiveness of the proposed Multi-scale-CNN for crowd density estimation.

## References

1. Abualigah L, Khader A, Hanandeh E (2017) A new feature selection method to improve the document clustering using particle swarm optimization algorithm. J Comput Sci
2. Abualigah L, Khader A, Hanandeh E (2018) A combination of objective functions and hybrid krill herd algorithm for text document clustering analysis. Eng Appl Artif Int
3. Chen J, He L, Yang T (2016) Scale-up purification for rutin hyrdrolysates by high-performance counter-current chromatography coupled with semi-preparative high-performance liquid chromatography. Sep Sci Technol 51(9):152–1530
4. Chen J, Kumar A, Ranjan R, Patel VM, Alavi A, Chellappa R (2016) A cascaded convolutional neural network for age estimation of unconstrained faces, 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS) 1–8
5. Chen K, Loy CC, Gong S, Xiang T (2012) Feature mining for localised crowd counting. Proc Br Mach Vis Conf 21.1–21.11
6. Cheng Z, Chang X, Zhu L (2019) MMALFM: explainable recommendation by leveraging reviews and images, ACM Transactions on Information Systems (TOIS) 37(2)
7. Collobert R, Kavukcuoglu K, Farabet C (2011) Torch7: A matlab-like environment for machine learning
8. Dai J, He K, Sun J (2016) Instance-aware semantic segmentation via multi-task network cascades. IEEE Conf Comput Vis Pattern Recognit (CVPR) 2016:3150–3158
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection, Computer Vision and Pattern Recognition, 2005 IEEE Computer Society Conference on New York IEEE 886–893

10. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645
11. Guo Y, Cheng Z, Nie L et al (2019) Attentive long short-term preference modeling for personalized product search, ACM Transactions on Information Systems (TOIS) 37(2)
12. Idrees H, Saleemi I, Seibert C, Shah M (2013) Multi-source multi-scale counting in extremely dense crowd images. IEEE Conf Comput Vis Pattern Recognit (CVPR) 2013:2547–2554
13. Lempitsky V, Zisserman A (2010) Learning to count objects m images, advances in neural information processing systems 1324–1332
14. Li M, Zhang Z, Huang K, Tan T (2008) Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection, 2008 19th International Conference on Pattern Recognition 1–4
15. Lin SF, Chen JY, Chao HX (2001) Estimation of number of people in crowded scenes using perspective transformation. IEEE Trans Syst Man Cybern Syst Hum 31(6):645–654
16. Liu L, Wang H, Li G et al (2018) Crowd counting using deep recurrent spatial-aware network, Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18) 849–855
17. Marsden M, McGuiness K, Little S et al (2016) Fully convolutional crowd counting on highly congested scenes. ArXiv preprint arXiv: 1612.00220
18. Marsden M, McGuinness K, Little S, O'Connor NE (2017) ResnetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification, 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) 1–7
19. Onoro-Rubio D, Lopez-Sastre RJ (2016) Towards perspective-free object counting with deep learning, european conference on computer vision (ECCV), Springer, Cham
20. Paragios N, Ramesh V (2001) A MRF-based approach for real-time subway monitoring, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), I-I
21. Ranjan R, Patel VM, Chellappa R (2019) HyperFace: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans Pattern Anal Mach Intell 41(1):121–135
22. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
23. Sam DB, Surya S, Babu RV (2017) Switching convolutional neural network for crowd counting. IEEE Conf Comput Vis Pattern Recognit (CVPR) 2017:4031–4039
24. Sharma A, De S, Gupta HM (2014) R Gangopadhyay, multiple description transform coded transmission over OFDM broadcast channels. Phys Commun 12:79–92
25. Sindagi VA, Patel VM (2017) CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) 1–6
26. Sindagi VA, Patel VM (2017) Generating high-quality crowd density maps using contextual pyramid CNNs, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR) 1861–1870
27. Szegedy C et al (2015) Going deeper with convolutions. IEEE Conf Comput Vis Pattern Recognit (CVPR) 2015:1–9
28. Viola P, Jones MJ (2013) Robust real-time face recognition, 2013 Africon, Pointe-Aux-Piments pp 1–5
29. Yu F, Koltun V (2016) Multi-scale context aggregation by dilated convolutions, International Conference on Learning Representations (ICLR)
30. Yu J, Zhang B, Kuang Z, Lin D, Fan J (2017) iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. IEEE Trans Inf For Secur 12(5):1005–1016
31. Zhang C, Li HS, Wang X, Yang XK (2015) Cross-scene crowd counting via deep convolutional neural networks, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 833–841
32. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-image crowd counting via multi-column convolutional neural network, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 589–597

**Yanjie Wang** obtained his B.Sc. degree from Qingdao University in 2017. Now he is a master degree candidate in Qingdao University. His research interests include machine learning and image processing.



**Shiyu Hu** obtained her B.Eng. degree from Qingdao Agricultural University in 2017. Now she is a master degree candidate in Qingdao University. His research interests include machine learning and image processing.

**Guodong Wang** Male, Born in Weifang City, Shandong Province, China, in February, 1980. Now he is an associate professor in College of Computer Science and Technology, Qing University. He received bachelor degree in 2001 and master degree in 2004 in control theory and control engineer, Qingdao University of Science and Technology, and received Ph.D degree in pattern recognition and intelligent system in Huazhong University in 2008. His research Interest include: Variational Image Science, Face recognition, Intelligent video survillance, 3D reconstruction and Medical image processing and Analysis.

**Chenglizhao Chen** received the PhD. degree in Technology of Computer Application from Beihang University, 2017. He is currently an assistant professor at Qingdao University. His research interests include pattern recognition, computer vision, and machine learning.

**Zhenkuan Pan** was born in 1966. He received his Ph.D. Degree in engineering mechanics from Shanghai Jiao Tong University, China in 1992. Currently he is a professor of College of Computer Science and Technology at Qingdao University. His research interests include dynamics and control of multibody systems, computer simulation and variational image processing. He has published numerous papers and conference papers in the area of image processing and object recognition.