ISSN 2616-5775 Vol. 4, Issue 7: 52-58, DOI: 10.25236/AJCIS.2021.040708

MSFANet: Crowd Density Estimation Based on Multi-scale Feature Adaptation

Hui Gao[1,a], Miaolei Deng[2,b], Dexian Zhang[2,c,*], Wenjun Zhao[2,d]

[1]School of Mechanical and Electrical Engineering Henan University of Technology, Zhengzhou, China
[2]School of Information Science and Engineering Henan University of Technology, Zhengzhou, China
[a]ghshow@139.com, [b]dmlei2003@163.com, [c]zdx@haut.edu.cn, [d]841804007@qq.com
**Corresponding Author*

Abstract: *To solve the problem that the number of heads cannot be accurately extracted due to scale changes in crowd counting, a crowd density estimation based on multi-scale feature adaptive network was proposed(MSFANet). Firstly, deep convolution neural network is used to extract semantic features of different scales. Then, the expansion convolution network is introduced, and the expansion convolution in the scale-up unit is combined with the traditional convolution to further increase the receptive field and reduce the information loss caused by channel competition. Finally, channel attention is introduced to enhance the ability of multi-scale feature extraction and ensure the integrity of relevant and important information in the image. Experiments on datasets (ShanghaiTech, UCF_CC_50 and WorldExpo10) show that this algorithm is more accurate and robust than the current mainstream crowd counting algorithms.*

Keywords: *Multi-scale, convolutional neural network, crowd counting, density map*

1. Introduction

With the rapid development of economy and the rapid growth of urbanization in China, the total number of large-scale cities has reached the first place in the world. The research of intelligent video surveillance technology is of great significance to public safety. As one of the important tasks of intelligent video surveillance and analysis, crowd counting has rapidly developed into one of the research hotspots in the field of computer vision. Existing crowd counting algorithms can be divided into two categories: traditional machine learning algorithms and crowd counting algorithms based on convolutional neural networks. Crowd counting based on traditional machine learning is divided into detection-based[1]. The sum of is based on regression[2]. All the methods need complex processes such as data preprocessing, feature extraction and detector design, and the counting error increases with the increase of crowd density.

In recent years, the proposal and development of deep learning[3].Especially the crowd counting algorithm based on convolutional neural network[4]. It has been widely used in the field of computer vision. Zhang et al[5] proposed to use three convolution kernels of different sizes to extract the features of crowd images respectively, and to generate a crowd density map according to the human head markers. Li et al[6] proposed to use a single column structure, in which the front end is used to obtain basic semantic information and the back end is used to generate crowd density information. Kang[7] proposed to use pyramid network to deal with scale change. Jiang[8] proposed adaptive pyramid loss (APLoss) to calculate the estimated loss of sub-regions hierarchically, which reduced the training deviation. Bai[9] and others put forward adaptive expansion convolution and a new supervised learning framework-self-tuning supervision. Yang[10] proposed a reverse perspective network to solve the scale change of input images. Using CNN to estimate the crowd density can not only effectively avoid manual feature extraction in traditional algorithms, but also obtain detailed information of crowd distribution, so this method gradually occupies a dominant position.

In view of the influence of image scale change and background interference in complex scenes, this study proposes a multi-scale feature adaptive crowd density estimation method. By expanding the image receptive field and adaptively selecting channels, we can extract the features of the concerned head position, accurately obtain the number of people in the scene, and monitor the change of crowd number in the scene in real time, which can effectively prevent abnormal situations such as overcrowding and trampling.

ISSN 2616-5775 Vol. 4, Issue 7: 52-58, DOI: 10.25236/AJCIS.2021.040708

2. Density Estimation

2.1. Density Map

Based on CNN's crowd counting method, the image marked with the center position of the head is input, the predicted crowd density map is output, and the total number of people in the image is obtained by integration and summation. For example, if the coordinate of a marked point is xi, the head center point of the point can be expressed as δ(x−xi) and the images of n heads can be expressed as:

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

Then, the center point of human head is normalized by Gaussian kernel filter[11], and the corresponding crowd density map is generated. The specific calculation is as follows:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_\sigma \tag{2}$$

Among them, Represents Gaussian kernel filter, σ represents the standard deviation of Gaussian kernel filter. Because of the sample differences between different data sets, this paper sets different super parameter σ according to different data sets to make corresponding density maps for training. The total number of people in the image can be obtained by integrating the finally generated population density map.

2.2. Loss Function

The difference between the crowd density map and the target density map can be expressed by calculating the Euclidean distance between the pixels corresponding to the points of interest in the two maps, so the loss function can be defined by formula (3).

$$L(\theta) = \frac{1}{2N} \sum_{i=1}^{N} \| G(X_i, \theta) - D_i^{GT} \|_2^2 \tag{3}$$

Where N is the number of images trained in a batch; Is the estimated density value of the ith training sample, and the parameter is θ; D_i^{GT} is the true density value.

3. Methodology

3.1. Network Model

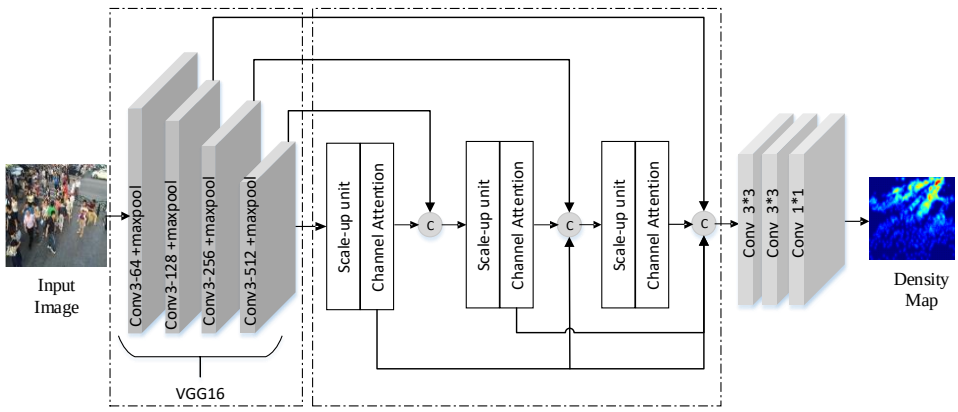

Figure 1: An illustration of the Multi-scale feature adaptive network.

The multi-scale feature adaptive network model proposed in this paper is shown in Figure 1. Because VGG16 network has excellent feature extraction ability[12], our method uses the first 10 layers of VGG16 as the deep convolution network, and the back end is a multi-scale feature adaptive network. The scale expansion unit is added before each channel's attention. When learning different

ISSN 2616-5775 Vol. 4, Issue 7: 52-58, DOI: 10.25236/AJCIS.2021.040708

scale feature maps, different channels are assigned weights to reduce the loss of feature information caused by channel competition. After selecting the channel adaptively, dense connection is adopted to enhance the ability of multi-scale feature extraction and ensure the integrity of relevant and important information in the image.

3.2. Scale-up Unit

Scale-up unit is composed of traditional convolution and expansion convolution. Expansion convolution not only increases the receptive field, but also reduces the information loss caused by channel competition. Through experiments, it is found that when the expansion rate is 2, the head information of different scales can be well extracted. The scale expansion unit is shown in Figure 2.

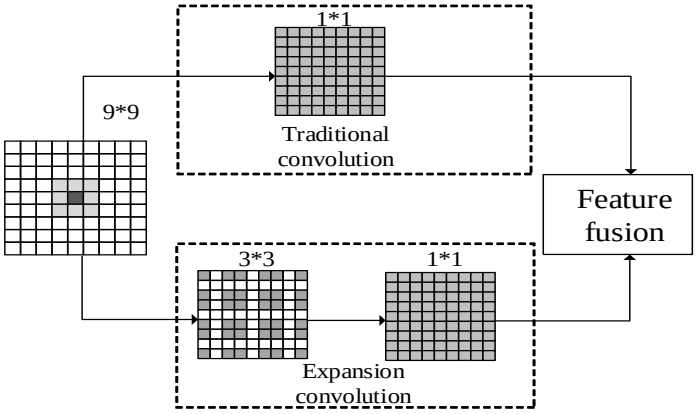

Figure 2: Structure diagram of scale-up unit.

3.3. Channel Attention

Convolution can fuse the spatial information of local receptive fields, but ignores the information sharing between channels. Using channel attention mechanism can enhance the feature extraction ability of convolutional neural network [13]. The channel attention mechanism is shown in Figure 3.

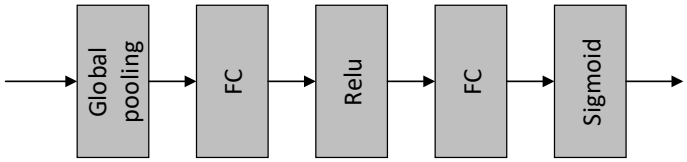

Figure 3: Channel attention mechanism.

Channel attention mechanism pays attention to the effective input of images. Firstly, the spatial dimension of feature map is compressed, that is, the global average pooling and global maximum pooling are used to aggregate the spatial information, and the obtained feature vector is used as the input of the full connection layer, and ReLU is used as the activation function. After passing through the next full connection layer, sigmoid function is used as the nonlinear activation function to enhance the mutual learning between channels and obtain the output vector. Finally, the input feature is multiplied by the output vector to obtain the channel attention feature map F.

4. Experiments and Analysis

4.1. Experimental Environment and Dataset

The algorithm in this paper is trained and tested on Intel 2.4GHz processor under Ubuntu16.04 operating system, and accelerated by GPU (Tesla V100). The experimental framework adopts PyTorch. In the process of network model training, the Batch Size is set to 256, the Learning Rate of the network is 10-5, and the variance is set to 0.01.

The algorithm in this paper is compared with other mainstream algorithms on ShanghaiTech and UCF_CC_50 datasets.

ISSN 2616-5775 Vol. 4, Issue 7: 52-58, DOI: 10.25236/AJCIS.2021.040708

4.1.1. ShanghaiTech

ShanghaiTech consists of 1198 images and 330,165 annotations. According to different density distributions, the data set is divided into two parts: Part_A and Part_B .. Part_A is a picture randomly selected from the internet, and Part_B contains a picture taken from a busy street in Shanghai. Part_A is much denser than part _ b.

4.1.2. UCF_CC_50

UCF_CC_50 is the first truly challenging data set, which was created by public network images. It includes distortion of different perspectives in various densities and different scenes, such as concerts, protests, stadiums and marathons. Considering that there are only 50 images in the data set, five-fold cross-validation protocols are applied to them. Because of the small amount of data, even the most advanced CNN-based method is far from the best result.

4.1.3. WorldExpo10

WorldExpo10 is a large data-driven cross-scene crowd-count dataset collected from the 2010 Shanghai WorldExpo, consisting of 1132 annotated video sequences captured by 108 surveillance cameras. It contains 3920 frames in total, with a size of 576×720 and 199, 923 comments.

4.2. Evaluation Indicators

Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are usually used as evaluation criteria in population counting research. MAE reflects the accuracy of the model, while RMSE reflects the robustness of the model. The specific calculation method is as follows:

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} | \hat{\mathrm{Y}}_{\mathrm{i}} - \mathrm{Y}_{\mathrm{i}} | \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} (\hat{\mathrm{Y}}_{\mathrm{i}} - \mathrm{Y}_{\mathrm{i}})^{2}} \tag{5}$$

Where N is the number of test set images; they are the predicted density map and the real density map of the crowd counting scene, Defined as follows:

$$\hat{\mathrm{Y}}_{\mathrm{i}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \mathrm{P}_{\mathrm{i}}, _{j} \tag{6}$$

And Pi,j is the pixel at (I, J) of the generated density map. Represents the estimate of image Xi.

4.3. Results and Analysis

Compared with other mainstream methods in ShanghaiTech data set, the experimental results of this algorithm show better performance on both Part_A and Part_B. Specific comparison results are shown in Table 1.

Table 1: Comparison of different algorithms on ShanghaiTech dataset.

Method	Part_A		Part_B	
	MAE	RMSE	MAE	RMSE
MCNN	110.2	173.2	26.4	41.3
Switch-CNN[14]	90.4	135	21.6	33.4
ACSCP[15]	75.7	102.7	17.2	27.4
PCC Net[16]	73.5	124.0	11.0	19.0
MSFANet(OURS)	77.5	113.2	10.7	16.9

Compared with other mainstream crowd counting methods, the experimental results on UCF_CC_50 data set show that our method has better crowd counting performance. The specific comparison results are shown in Table 2.

ISSN 2616-5775 Vol. 4, Issue 7: 52-58, DOI: 10.25236/AJCIS.2021.040708

Table 2: Comparison of different algorithms on UCF_CC_50 dataset.

Method	MAE	RMSE
MCNN	377.6	509.1
Switch-CNN	318.1	439.2
CSRNet	266.1	397.5
PCC Net	240.0	315.5
MSFANet(OURS)	241.2	313.6

Compared with other mainstream crowd counting methods, the experimental results on WorldExpo'10 data set show that our method has better crowd counting performance. The specific comparison results are shown in Table 3.

Table 3: Comparison of different algorithms on WorldExpo10 dataset.

Method	World-Expo'10					
	Sce1	Sce2	Sce3	Sce4	Sce5	Avg
MCNN	3. 4	20. 6	12. 9	13. 0	8. 1	11. 60
Switch-CNN	4. 4	15. 7	10. 0	11. 0	5. 9	9. 40
CSRNet	2. 9	11. 5	8. 6	16. 6	3. 4	8. 6
ACSCP	2. 8	14. 5	9. 6	8. 1	2. 9	7. 58
MSFANet(OURS)	2. 6	11. 9	9. 2	9. 7	3. 1	7. 30

4.4. Visualization

The method in this paper has superior performance in solving the problem of different head sizes and keeping details in the process of crowd counting. As shown in Figure.4, in two very dense scenes, this method can generate density maps almost the same as the real density maps, and estimate numbers closer to the real population. Thereinto, figure (a) is an original image extracted from PartA of ShanghaiTech, figure (b) is the corresponding real density map, and figure (c) is the predicted density map generated by MSFANet. The method in this paper predicts that the details of density map (c) are more detailed, and the contour of background and foreground distinction is more obvious.

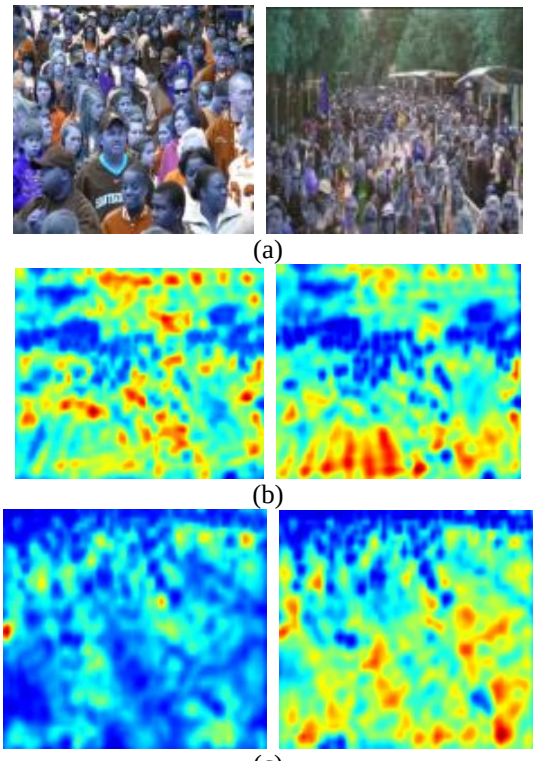

Figure 4 Visualization of feature maps: (a) Input image, (b) Before segmentation infusion, (c) After segmentation infusion. By infusion segmentation information into the counting network, we are able suppress background regions. Note that in the density maps, red color indicates high density and blue color indicates low density.

ISSN 2616-5775 Vol. 4, Issue 7: 52-58, DOI: 10.25236/AJCIS.2021.040708

5. Conclusions

Aiming at the problems of scale change in crowd counting and competition in multi-scale feature fusion, this paper studies and proposes a multi-scale feature adaptive network. The proposed multi-scale feature adaptive network combines expansion convolution with traditional convolution, which effectively solves the problem of information loss caused by channel competition in channel selection of multi-scale features, and improves the robustness of the network to different scales. Finally, experiments show that the algorithm proposed in this paper is suitable for a variety of scenarios, and has achieved good results on a number of data. However, at present, crowd density estimation is mostly based on static images, and online prediction based on video stream is needed in practical applications. Therefore, the next step is to further enhance the robustness of the network to scale changes and improve the real-time performance of the crowd counting network.

Acknowledgments

This work was supported by National Key R&D Program of China 2018******02.

Gao Hui, born in 1983, is a doctoral candidate and mainly researches intelligent information technology. E-mail:ghshow@139.com.

Deng Miaolei, born in 1977, holds a PhD. degree and is an associate professor, mainly engaged in intelligent information technology.

Corresponding author: Professor Zhang Dexian, born in 1961, is a doctoral supervisor with a PhD. degree. He is also the vice-chairman of Information and Automation Sub-association of Chinese Cereals and Oils Association (CCOA), a scientific and technological innovation talent in Henan Province and an academic and technical leader of the Education Department of Henan Province. He mainly studies intelligent information technology, data mining and machine learning.

Zhao Wenjun, born in 1993, is a doctoral candidate and mainly researches intelligent information technology.

References

[1] Sabzmeydani P, Mori G. Detecting Pedestrians by Learning Shapelet Features [C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2007.

[2] Chan A B, Vasconcelos N. Bayesian poisson regression for crowd counting [C]//2009 IEEE 12th International Conference on Computer Vision (ICCV),2009.

[3] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. science, 2006, 313(5786): 504-507.

[4] Lecun Y, Bottou L. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[5] Zhang Y, Zhou D, Chen S, et al. Single-image crowd counting via multi-column convolutional neural network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 589-597.

[6] Li Y, Zhang X, Chen D. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[7] Kang D, Chan A. Crowd Counting by Adaptively Fusing Predictions from an Image Pyramid [J]. arXiv preprint arXiv: 1805.06115, 2018.

[8] X Jiang, Zhang L , Xu M , et al. Attention Scaling for Crowd Counting [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[9] Bai S, He Z, Qiao Y, et al. Adaptive Dilated Network With Self-Correction Supervision for Counting [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[10] Yang Y, Li G, Wu Z, et al. Reverse Perspective Network for Perspective-Aware Object Counting [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[11] Cao, Xinkun, et al. Scale Aggregation Network for Accurate and Efficient Crowd Counting [C]//european conference on computer vision, 2018: 757-773.

[12] Zhang C, Li H,Wang X,et al.Cross-scene crowd counting via deep convolutional neural networks [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015.

ISSN 2616-5775 Vol. 4, Issue 7: 52-58, DOI: 10.25236/AJCIS.2021.040708

[13] Zeiler M D,Fergus R.Visualizing and understanding convolutional networks [M] //Fleet D, Pajdla T, Schiele B, et al. Computer vision ECCV 2014. Lecture notes in computer science. Cham:Springer, 2014,8689:818-833.

[14] Sam D B, Surya S, Babu R V. Switching Convolutional Neural Network for Crowd Counting [C]// Computer Vision & Pattern Recognition. IEEE, 2017.

[15] Zan S, Yi X, Ni B, et al. Crowd Counting via Adversarial Cross-Scale Consistency Pursuit [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.

[16] J. Gao, Q. Wang, and X. Li, "Pcc net: Perspective crowd counting via spatial convolutional network," TCSVT, 2019.