Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 2, 2022

Crowd Counting Based on Context-Aware and Multi Scale Feature Fusion

Guoli Zhang

Tiangong University, Tianjin 300387, China

Abstract: Crowd counting plays an important role in public security. Estimating the number of people in an image with congested crowd accurately is a challenging task. The crowd counting method based on fully convolutional network can perform well in crowd image with complex scene. In this paper, to address the counting problems of occlusion， background clutter and perspective effect, we proposed a simple but effective method called Context-aware Multi scale Fusion Network(CMF Net).The CMF Net applied VGG network as backbone to extract coarse features. Then, three context-aware multi-scale fusion modules (CMFM) are adopted. Each CMFM consist of multi-scale feature extraction module (MEM) and context-aware feature extraction module (CEM). In addition, we propose adaptively dense connection to promoted information transmission in the counting network. Experiments on four datasets demonstrate that our network achieves competitive and effective results.

Keywords: Crowd Counting; Deep Learning; Computer Vision; Attention Mechanism; Feature Fusion.

1. Introduction

With the development of urbanization, a large number of congested crowd scenes have been appeared in large city. And it hides many safety hazards. In this year,29nd October 2022, over 150 people dead after stampede at Halloween event in Seoul, South Korea. Therefore, researching of crowd counting is extremely important in intelligent surveillance, disaster management and social distance monitoring. In Figure 1, we shown the crowd scenes with different density.

Figure 1. Crowd scenes with different density level

Traditional crowd counting method based on Object Detection [1] is not suitable for highly congested scene. Inspired by the success of deep convolutional networks, varieties of crowd counting method based on CNN have been proposed. These networks usually adopt an encoder-decoder structure, and generated the predicted density map. Due to perspective effect of different camera viewpoint, the scale of people in the image varies dramatically. And counting in an image also has other problems, such as occlusion, illumination changes, background noises and clutter. To solve these difficulties, MCNN [2] adopt a multi-column CNN network to capture scale-aware features with different receptive fields. CP-CNN [3] designed a context-aware pyramid network to combine the local and global features. CSR-Net[4] employed a deeper single-column network, it adopted dilated six convolutional layers to further enlarge receptive fields.[5] analyzed the impact of the wrong predictions on background regions in crowd counting.

In summary, we analyzed that the context information, enlarging the receptive fields and multi-scale feature extraction are the keys to solve the problems of crowd counting. In this paper, the contributions of our work are as follows:

We proposed a single-column and multi-branch structure based on dilated convolutional layers with different rates to extract multi-scale features. It can solve the problem of perspective effect. And we proposed adaptively dense connection to promote the information transmission.

We applied attention mechanism to capture context-aware features. This context information can effectively handle the large scenes variations among the crowd images.

We introduced a novel crowd counting network (CMF Net) by end-to-end structure. Extensive experiments on four datasets demonstrate that our network is effective and competitive to handle the problem of perspective effect.

2. Proposed method

In this section, we firstly described the overall network of our CMF Net in Figure 2. And then we have to introduce the detailed information about multi-scale Feature extraction module (MSM), context feature extraction module (CFM) and adaptively dense connection.

2.1. Overview

Crowd counting task is regarded as a density map regression problem from an image. We adopted the first 10 layers of VGG16[6] networks as the backbone of the CMF Net because of it can balance feature extraction efficiency and computational complexity. Due to avoid overfitting and the number of dataset imagers is not sufficient, we adopted transfer learning on Imagenet dataset of classification task to pre-train VGG16 network. The backbone's output resolution is 1/8 size of original input image. The output feature of image I are generated by:

$$f_v = \mathcal{F}_{vgg}(I) \quad (1)$$

Then, we applied context-aware multi-scale fusion

modules (CMFM) to capture multi-scale features and context information. Subsequently, we enhanced the resolution of the feature map by up sampling with the factor of 2. We repeated this procedure for 3 times and use adaptively dense connections to increase information transmission. Finally, two dilated and one kernel size of 1*1 convolutional layer as back end to generate a density map of input image.

2.2. Context-aware Multi-scale Fusion module

Due to perspective effect, the people scale in one image change dramatically. And the scenes in training set differ from one another. [7] proved that context information is helpful for crowd counting task. Therefore, fusing the multi scale

features and context information is the main solution of the problem. We proposed Context-aware Multi-scale Fusion module (CMFM), it consist of multi-scale feature extraction module (MEM) and context-aware feature extraction module (CEM). At last, we fused these features with residual connection, element-wise product and addition. The description of CMFM in Figure 3.

2.2.1. Multi-scale feature extraction module (MEM)

Inspired by ASPP [8], we designed a multi branch pyramid structure with aurous convolution layers to capture multi scale features. Each module has four parallel dilated convolutional layers with different dilation rates. The output feature of MEM contains multi scale information of crowd image. Then, to balance the counting efficiency and the number of computational parameters, we adopted the way of cascade to fuse pyramid features.

2.2.2. context-aware feature extraction module (CEM)

We utilized attention mechanism to extract the context-aware information. This pixel-wise context information based on attention can highlight the foreground information of people's head and suppress the noise of background region. Making the network pay more attention to the crowd region. The CEM consist of channel attention branch and space attention branch. We devided the CBAM

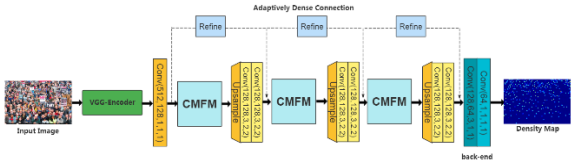

Figure 2. The architecture of our proposed CMF Net.

The Convolution layers are designed as Conv (input channels, output channels, kernel size, padding, dilation rate) network [9] to get the two branches. And combined them in parallel way. The attention masks generated by CEM produced the multi-scale feature from the MEM. At last, we applied residual connection and element-wise addition to fuse context-aware features and original feature map. The procedure can be formulated as follows:

$$M_c(F_{in}) = \sigma\left(MLP\left(AvgPool(F_{in})\right) + MLP\left(MaxPool(F_{in})\right)\right) \quad (2)$$

$$M_s(F_{in}) = \sigma(f^{7\times7}([AvgPool(F_{in}); MaxPool(F_{in})])) \quad (3)$$

$$\mathrm{F_{out}} = \mathrm{M_c} \times \mathrm{F_{ms}} + \mathrm{M_s} \times \mathrm{F_{ms}} + \mathrm{F_{in}} \quad (4)$$

where $M_c(F_{in})$ represents the channel-wise attention weights generated by channel attention branch of CEM and $M_s(F_{in})$ represents the spatial attention weights. MLP is the abbreviation of multi-layer perceptron. σ denotes sigmoid function. $f^{7\times7}$ represent the convolutional layer with the kernel size of 7×7.F_{in} ,F_{ms},F_{out}denotes the input feature map, multi-scale feature and the output feature map of the whole CMFM respectively.

2.3. Adaptively dense connection

The decoder of CMF Net adopted gradually up sample strategy. Thus, common dense connection [10] is not suitable for our network. We designed a refine block to up sample and refine the low-resolution feature map. Then, we combined the refine block and dense connection as adaptively dense connection. The detailed description of Adaptively dense connection is in Figure 4. It is helpful for the back propagation during the training and enhance the information transmission, thereby we can apply deeper network.

2.4. Loss Function

We use Euclidean distance loss as the loss function of our network to measure the pixel-wise difference between the predicted density map and the ground truth density map, the description of loss function are as follows:

$$\mathrm{L}(\mathrm{X_i}, \theta) = \frac{1}{2\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} ||\mathrm{F}(\mathrm{X_i}, \theta) - \mathrm{D_i}||_2^2 \quad (5)$$

where N represents the number of training images. $\mathrm{X_i}$ is the input image. θ are the parameters of our network? $\mathrm{F}(\mathrm{X_i}, \theta)$ represents the predicted density map and the $\mathrm{D_i}$ represents the ground truth density map.

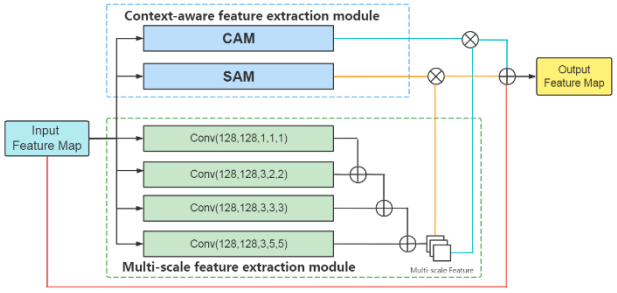

Figure 3. The architecture of CMFM. The CAM and SAM represent the channel attention module (Eq.2.) and the spatial attention module (Eq.3.).

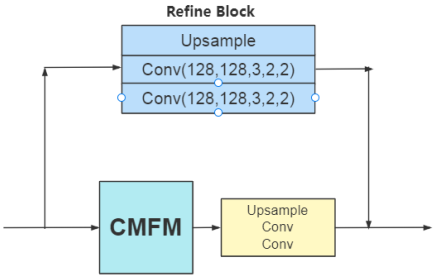

Figure 4. This is the description of our Adaptively dense connection

3. Experiments

In this section, firstly we introduced the evaluation metrics. Then, we described our implementation details. At last, we compared our network to other functions

3.1. Evaluation metrics

Consistent with the mainstream methods, we adopted mean absolute error (MAE) and mean square error (MSE) to evaluate the performance of our network, which are defined as:

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} |\mathrm{C_i^{pre}} - \mathrm{C_i^{gt}}|$$

and

$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i^{pre} - C_i^{gt})^2}$$

where N is the number of images. The C_i^pre represent the predicted number of people in the i-th image. And the C_i^gt is the ground truth number of crowd people.

3.2. Implementation Details

In the training stage of our experiment, we set the learning rate at 0.00001 and adopt Adam optimizer.

In our approach, the size of input feature is arbitrary. We set the batch size to 1 for the datasets that have arbitrary size images. And we adopted the batch size of 4 for the Shanghai Tech_ part_ B dataset with fixed size images. For each dataset, we applied 300 epochs training our model. We use Pytorch1.7 as deep learning framework to implement our method. Nvidia 3090 GPU is the hardware platform of our experiment.

3.3. Experimental Comparison

We test our network on three publicly dataset: Shanghaies, UCF-CC-50 and UCF-QNRF. The summarization of datasets as shown in Table 1. And the comparison with other experiments are as shown in Table 2. And in table 2, - denotes there is no experimental data in original paper.

Among these datasets, we summarized that Shanghaies and UCF-QNRF have a medium-level density crowd distribution. UCF-CC-50 dataset has the largest average number of people in the images. Shanghaies has a relatively sparse crowd density distribution. Table 1 display the comparison among these networks. In summary, our CMF Net has a good performance for sparse and medium-level density scenes.

3.4. Ablation Study

In this section, we analyzed the effectiveness of each module of our network. We have ablation study on Shanghaies dataset. The experimental comparisons are in Table 3.

Table 1. The datasets summarization

Datasets	Resolution	Train set	Test Set	Max	Min	Avg
SHA	Arbitrary	300	182	3,139	33	501
SHB	1024*768	400	316	578	9	123
UCF50	Arbitrary	50	50	4,543	94	1,279
QNRF	Arbitrary	1201	334	12,865	49	815

Table 2. The datasets summarization

Method	ShanghaiTech_A		ShanghaiTech_B		UCF-CC-50		UCF-QNRF	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN	110.2	173.2	26.4	41.3	377.6	509.1	277	426
MSCNN	83.8	127.4	17.7	30.2	363.7	468.4	-	-
SwichCNN	90.4	135.0	21.6	33.4	318.1	439.2	228.0	445.0
CSRNet	68.2	115.0	10.6	16.0	266.1	397.5	120.3	208.5
LSC-CNN	66.4	117.0	8.1	12.7	225.6	302.7	120.5	218.2
MDCount	84.2	130.7	11.8	19.15	**103.1**	**158.1**	111.3	203
LigMSA	76.6	121.4	10.9	17.5	231.5	339.7	-	-
ours	**65.2**	**110.3**	**8.2**	**11.9**	166.3	285.7	**110.7**	**201.3**

Table 3. The effects of each modules

	ShanghaiTech_A	
	MAE	MSE
backbone	71.1	121.4
Backbone+CMFM(MEM)	70.3	119.0
Backbone+CMFM(MEM+CEM)	69.1	118.7
Backbone+CMFM*2	68.2	116.1
Backbone+CMFM*3	66.9	114.0
Backbone+CMFM*3+adaptively dense connection	**65.2**	**110.3**

The results shown the modules we proposed are effective for crowd counting task.

4. Conclusion

In this paper, we present a crowd counting architecture named CMF Net. To solve the problem of perspective effect, we proposed multi-scale feature extraction module and fused context information. We adopted adaptively dense connect to further promote information transmission. The results of extensive experiments shown our CMF Net achieves competitive performance in accuracy and robustness.

References

[1] Dollar P, Wojek C, Schiele B and Perona P 2012 Pedestrian detection: An evaluation of the state of the art IEEE transactions on pattern analysis and machine intelligence 34(4) pp 743–761.

[2] Zhang Y, Zhou D, Chen S, Gao S and Ma Y 2016 Single-image crowd counting via multi-column convolutional neural network. Proceedings of the IEEE conference on computer vision and pattern recognition pp 589–597.

[3] Sindagi A and Patel M 2017 Generating high-quality crowd density maps using contextual pyramid cnns IEEE International Conference on Computer Vision (ICCV) pp 1879–1888.

[4] [4] Li Y, Zhang X and Chen D 2018 Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 1091-1100.

[5] Davide M, Bing S, Rahul R V, Joseph T 2021 Understanding the Impact of Mistakes on Background Regions in Crowd Counting Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1650-1659.

[6] Karen S and Andrew Z 2014 Very Deep Convolutional Networks for Large-Scale Image Recognition Computer Vision and Pattern Recognition (cs.CV) arXiv:1409.1556.

[7] Weizhe L, Mathieu S and Pascal F 2019 Context-Aware Crowd Counting Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp 5099-5108.

[8] Liang-Chieh C, George P, Iasonas K, Kevin M and Alan L Y 2017 DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[9] Sanghyun W, Jongchan P, Joon-Young L and In S K 2018 CBAM: Convolutional Block Attention Module Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19.

[10] Gao H, Zhuang L, Laurens van der M and Kilian Q. W 2017 Densely Connected Convolutional Networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4700-4708.