The Comparison of Crowd Counting Algorithms based on Computer Vision

IOP Conf. Series: Journal of Physics: Conf. Series **1187** (2019) 042012 doi:10.1088/1742-6596/1187/4/042012

The Comparison of Crowd Counting Algorithms based on Computer Vision

Zhaoqing Wang[1, *], Qishu Deng[2, a], Yusheng Zhao[2, b]

[1]School of Information & Communication Engineering, Beijing Information Science& Technology University, Beijing, China
[2]International School, Beijing University of Posts and Telecommunications, Beijing, China

[*]Corresponding Author E-mail: mj741561@163.com

[a]dengqishu@bupt.edu.cn; [b]zhaoyusheng@bupt.edu.cn

Abstract: This paper aims to compare the three mainstream solutions for today's crowd counting and analyze the highlights of each model. In MCNN, they proposed a multi-column parallel convolutional neural network structure that generates population density maps by adapting crowd changes caused by camera view-points and resolution using filters with different size receptive fields. In Switch-CNN, they added a density classifier to the MCNN to enable the use of local density changes in the crowd. In CSRNet, they abandoned the structure of a multi-column convolutional neural network, using the first ten layers of VGG-16 as the front part and the convolutional neural network as the latter part. From the analysis results, CSRNet shows advanced performance. In addition, we analyzed the comparison results of three convolutional neural networks, and derived the trend of convolutional neural network structure.

1. Introduction

Computer vision began in the late 1960s. Firstly, it was designed to imitate the human visual system as the basic of robots' intelligent behavior [1]. In 1966, to realize the computer vision and let computer present "what it saw", people connect computer to the camera [2][3]. Today, many computer vision algorithms including extracting edges from images, non-polyhedral and polyhedral modeling, labeling lines, representing objects as the interconnection of smaller structures, optical flow and motion estimation are based on the early researches [1]. Then, people use variations of graph cut to solve image segmentation. In a word, computer vision covers several aspects. One of the important applications of computer vision is to accurately estimate the number of people in an image or video.

Crowd counting is a key technology to control crowd and ensure public safety. Monitoring crowds in video or images has important market applications. In addition, information such as the number, density or distribution of participants can be obtained through crowd counting method, which provides effective safety guidance of public places such as stations, shopping malls and plazas. It can also achieve greater economic benefits by improving service quality, analyzing customer behavior, optimizing advertising and resource allocation. Moreover, mature crowd counting technology can be extended to other fields, such as the estimation of vehicle density on traffic roads, microbes counting in microscopic images, and species protection in ecological tribes.

In recent years, with the gradual development of computer vision technology, a large number of crowd counting methods have been proposed. Among them, the core algorithms of crowd counting in the field of deep learning are Multi-CNN, Switch-CNN and CSRNet [4][5][6]. This paper would

mainly compare the similarities and differences of those algorithms as well as discuss their advantages and disadvantages. Summarize the trend of those core algorithms of the crowd counting. In all, this paper aims to obtain a clearer understanding on crowd counting algorithm and figure out the future development trends, in order to achieve greater breakthroughs.

2. Literature Review

2.1 Tracked visual features trajectories clustering approaches

Tracked visual features trajectories clustering have been widely used in crowd counting. By using the established pedestrian database to train the classifier, the KLT tracker is used to track the features detected in the pedestrian videos [7] [8]. In addition, the module based on feature tracking and clustering of KLT feature tracker and spatiotemporal clustering module can filter and delete invalid trajectories, calculate the motion features of the target and automatically assign trajectories to each object [7]. However, this is a very computationally intensive method and can only be applied to the crowd counting in the video.

2.2 Feature-based regression approaches

Feature-based regression can be applied to crowd counting of still images. When extracting low-level features, the background can be removed, and various features in the foreground image such as textures can be extracted. Linear, piece-wise linear or neural network regression functions can be established to estimate the number of people [9]. Additionally, feature-based regression approaches can be used to extract and analyze features in image blocks to search objects with specific characteristics. For the crowd with high density and high occlusion, it is more effective to use a regression function that has been trained by low-level features. This approach may be easily affected by sparse and unbalanced situations, causing large changes in feature parameters [10]. There is currently no good solution.

2.3 Deep learning approaches

The CNN-based algorithms are very popular right now, behaving better in accuracy and flexibility in the field of crowd counting. According to the algorithms mentioned above, foreground segmentation is indispensable, but it is difficult to implement. In contrast, the DCNN proposed in [11] does not require foreground segmentation and hand-crafted feature extraction. It can count the total crowd number based on the textures. [4] proposed the MCNN structure. By using different sizes of receptive fields, this algorithm is available to do crowd counting in any static crowd image of any size and resolution. This algorithm can also be transferred to other targets by fine-tuning. [5] Combined with previous studies, Switch-CNN was established, which increased the accuracy of prediction depending on the changes of image density. In addition, researchers are constantly expanding data sets to ensure efficient learning of CNN [12]. Using pure convolutional layer as the core of the structure, a deeper network CSRNet was built, which reduced the computational complexity to some extent [6]. So far, the crowd counting algorithm in deep learning has obtained many breakthroughs and achievements. Therefore, this paper will further compare the three core CNN-based algorithms, discuss their structures and innovation.

3.Comparison

Estimating the number of people in a given population by using the Convolutional Neural Network (CNN) is a relatively engineering study point in the domain of deep learning. There are two main types of solutions. In the first type of network, the input is the image, and the output is the estimated number of people. The input to the second type of network is also an image, but its output is the density map of the crowd, then the number is obtained by integration. However, the output of the second network has more information in the density map than the output of the first one, which gives the distribution of the crowd in space. This paper will use three specific networks as an example to

analyze and compare the network structure using population density counting.

3.1 Multi-CNN

In the case of crowd counting, since the images usually contain heads of very different sizes, it is not possible to detect the characteristics of the crowd density using the filters with reception field of the same size. Hence, they proposed the Multi-Column Convolutional Neural Network (MCNN), which consists of three columns of parallel CNNs with local receptive fields of different sizes. For simplification, in addition to the size and number of filters, each column CNN uses the same network structure. Two 2*2 max pooling layers are used in each column CNN (Because the max pooling layers are used twice, each training sample needs to be down-sampled by 1/4 before the corresponding density map is generated.). At the same time, a rectified linear unit (ReLU) is used as an activation function [13]. To reduce computational complexity, a smaller number of filters are used for CNNs with larger filters. Avoid the distortion of the graphics when stacking the output feature maps of all CNNs and mapping them to the density map. This network uses a 1×1 filter [14] so that the input image can be of any size. Finally, Euclidean distance is used to measure the difference between the estimated density map and the ground truth .The network structure of MCNN is illustrated in Figure 1:

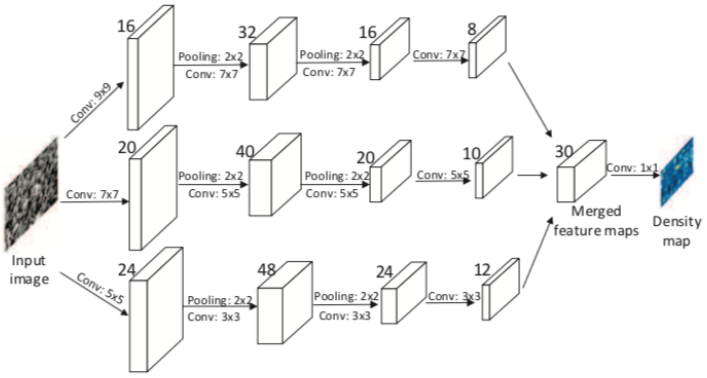

Figure 1: The structure of the proposed multi-column convolutional neural network for crowd density map estimation. [4]

3.2 Switch-CNN

Except for the ability to model large scale variations, the Switch Convolutional Neural Network (Switch-CNN) can also use local density changes in crowd scenarios well. Because the weighted averaging technique used in MCNN will weaken the details in the image, reduce the contrast of the image, and blur the edges in the image to a certain extent. It is difficult to achieve satisfactory fusion effect in most applications, so the ability to leverage local variations in density becomes especially important. The structure of Switch-CNN consists of three CNN regressors with varying receptive fields and the switch that selects the correct regressor for the input patchs. This network uses three CNN regressions introduced in it, R1, R2, R3 [4]. R1 is a 9*9 large-size filter that captures advanced features in the scene. R2 and R3 are 7×7 and 5×5 filters respectively to capture features in low scales. The switch network consists of a switch classifier and a switch layer. The switch classifier infers the label of the CNN regressor suitable for the input image patches. The switch layer receives the label and relays the patches to the correct regressor. The switch uses the adaptation of VGG16 network as the switch classifier for three-way classification [15]. The fully-connected layers in VGG-16 are banned by the global average pool (GAP). GAP is followed by a smaller fully-connected layers and a 3-class softmax classifier, corresponding to three CNN regressors in the Switch-CNN. Switch-CNN first divides the image into 3*3 non-overlapping patches based on a certain crowd characteristics, then uses a switch classifier to classify the patches by density standard, and then relays the patches to the independent CNN regressor with different receptive fields and field-of-view. The network structure of

Switch-CNN is illustrated in Figure 2:

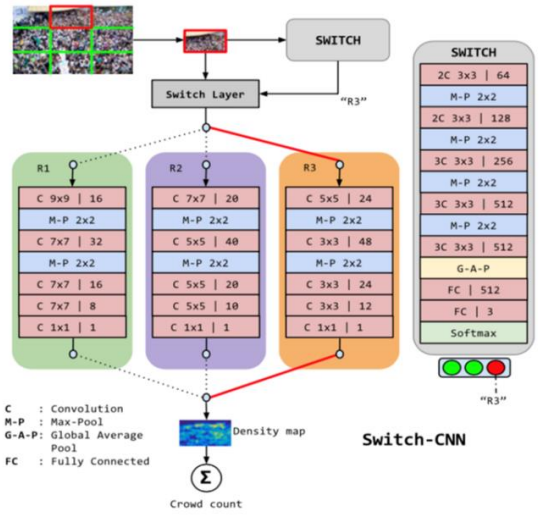

Figure 2. The structure of the proposed switch convolutional neural network for crowd density map estimation. [5]

Configurations of CSRNet			
A	B	C	D
input(unfixed-resolution color image)			
front-end (fine-tuned from VGG-16)			
conv3-64-1 conv3-64-1			
max-pooling			
conv3-128-1 conv3-128-1			
max-pooling			
conv3-256-1 conv3-256-1 conv3-256-1			
max-pooling			
conv3-512-1 conv3-512-1 conv3-512-1			
back-end (four different configurations)			
conv3-512-1	conv3-512-2	conv3-512-2	conv3-512-4
conv3-512-1	conv3-512-2	conv3-512-2	conv3-512-4
conv3-512-1	conv3-512-2	conv3-512-2	conv3-512-4
conv3-256-1	conv3-256-2	conv3-256-4	conv3-256-4
conv3-128-1	conv3-128-2	conv3-128-4	conv3-128-4
conv3-64-1	conv3-64-2	conv3-64-4	conv3-64-4
conv1-1-1			

Figure3 The Structure of Dilated Convolutional Neural Networks for crowd density map estimation.[6]

3.3 CSRNet

CSRNet is a network model for high-density population monitoring. The main idea of this network is to deploy deeper CNN. The benefit of this is that you can capture advanced features with larger reception fields while also reducing network complexity. Dilated Convolutional Neural Networks (DCNN) consists primarily of the front-end VGG-16 convolutional layer portion and the back-end dilated convolutional layers. According to similar ideas in [16][5][17], they chose VGG-16 as the front end of CSRNet [6], and then deleted the classification part of VGG-16 fully connected layer, using the convolution layer of VGG-16. Built CSRNet. After making trade-offs between accuracy and the resource overhead (including training time, memory consumption, and the number of parameters), they retained the top ten layers of VGG-16, with three pooling layers. In order to avoid loss of resolution and to extract deeper information of saliency, they proposed dilated convolutional layers. It uses sparse kernels to alternate the pooling layers and convolutional layers. The character expands the receptive field without increasing the number of parameters or the amount of calculation, reducing the computational complexity. In the dilated convolutional layers, the small-size kernel with the a*a filter is expanded to a +(a-1)(b-1) with an expansion step b. Since CSRNet's output density map is only 1/8 of the input size, they choose the bilinear interpolation with a factor of 8 to scale, which ensures that the output has the same resolution as the input image. The dilated convolutional network shows obvious advantages. The first advantage is that the output shares the same dimensions as the input (no pooling and deconvolution). The second advantage is that the output of the classified convolutional contains more detailed information. The network structure of CSRNet is illustrated in Figure 3.

4. Discussion of Application and Function

The ShanghaiTech dataset contains more viewpoints and a larger density of people than most existing datasets. The comparison results for these three models are showed in Table 1. We can clearly see that CSRNet performs better than Switch-CNN and MCNN, achieving the lowest MAE and MSE compared to other methods. This shows that CSRNet perform noticeably well in high-density scenarios and is capable of high-density population density detection. At the same time, the test results

of the UCSD dataset characterized by low-density scenes are shown in Table 2. The results show that MCNN performs best and achieves the lowest MAE and MSE. This shows that MCNN is qualified for low-density population testing. Although CSRNet is not the best performer, the effect is considerable. This shows that in addition to being able to perform high-density crowd counting, CSRNet is also suitable for low-density crowd counting and is very versatile.

Table 1.The comparison of three models on ShanghaiTech dataset

	Part A		Part B	
Method	MAE	MSE	MAE	MSE
MCNN	110.2	173.2	26.4	41.3
Switch-CNN	90.4	135.0	21.6	33.4
CSRNet	**68.2**	**115.0**	**10.6**	**16.0**

Table 2.The comparison of three models on UCSD dataset

	Part A		Part B	
Method	MAE	MSE	MAE	MSE
MCNN	110.2	173.2	26.4	41.3
Switch-CNN	90.4	135.0	21.6	33.4
CSRNet	**68.2**	**115.0**	**10.6**	**16.0**

The three columns of parallel CNNs in the MCNN have filters of different size local reception fields, which are significantly modified, compared to the traditional CNN. This can better adapt to the size change of the human head in different scenes. In addition, this network uses the weighted average of the filter of 1*1 to fuse the feature maps from each column CNN, so the density map of any input size image can be generated without distortion. A major change in Switch-CNN is to first divide the input image into 9 patches and then use the switch to take advantage of changes in local population density within the scene. In the switch classifier, the fully connected layer of VGG-16 is replaced by the global average pool (GAP). The advantage of GAP is to regularize the entire network structure to suppress overfitting (random noise of the model overfitting the data set). The highlight of CSRNet is the dilated convolutional layers, which use a sparse kernel to replace pooling layers and convolutional layers to expand the receptive field. This approach brings three distinct benefits: (1) maintaining spatial resolution. (2) Simplified network structure. (3) Reduced computational complexity and number of parameters.

5. Conclusion

In this paper, we compared MCNN, Switch-CNN, and CSRNet from the perspective of network structure and experimental performance. MCNN uses three columns of CNNs with different size receptive fields to achieve adaptation to different sizes of heads. In addition, MCNN also has good adaptability. By using the switch, Switch-CNN uses the factor of local density variation to classify crowd image patches. By using dilated convolutional layers, CSRNet expands the receptive field without reducing spatial resolution, while also enabling density detection for high-density populations. From the overall results, CSRNet is the most condensed and best performing CNN. From MCNN to CSRNet, we have found that the structure of convolutional neural networks has evolved from multiple columns CNN to single columns CNN, and the depth of the network has gradually deepened from shallow to shallow. This is also the future development trend of convolutional neural networks.

Reference

[1] Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

[2] Papert, S. A. (1966). The summer vision project.
[3] Margaret, A. (2006). Mind as machine: a history of cognitive science.
[4] Zhang, Y., Zhou, D., Chen, S., Gao, S., & Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 589-597).
[5] Sam, D. B., Surya, S., & Babu, R. V. (2017, July). Switching convolutional neural network for crowd counting. In*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 1, No. 3, p. 6).
[6] Li, Y., Zhang, X., & Chen, D. (2018, February). CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1091-1100).
[7] Rabaud, V., & Belongie, S. (2006, June). Counting crowded moving objects. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (Vol. 1, pp. 705-711). IEEE.
[8] Yang, T., Zhang, Y., Shao, D., & Li, Y. (2010). Clustering method for counting passengers getting in a bus with single camera. *Optical Engineering*, *49*(3), 037203.
[9] Chan, A. B., Liang, Z. S. J., & Vasconcelos, N. (2008, June). Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1-7). IEEE.
[10] Chen, K., Gong, S., Xiang, T., & Change Loy, C. (2013). Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2467-2474).
[11] Zhang, C., Li, H., Wang, X., & Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. In*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 833-841).
[12] Liu, X., van de Weijer, J., & Bagdanov, A. D. (2018). Leveraging Unlabeled Data for Crowd Counting by Learning to Rank. *arXiv preprint arXiv:1803.03095*.
[13] Zeiler, M. D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q. V., ... & Hinton, G. E. (2013, May). On rectified linear units for speech processing. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*(pp. 3517-3521). IEEE.
[14] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In*Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).
[15] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
[16] Boominathan, L., Kruthiventi, S. S., & Babu, R. V. (2016, October). Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 2016 ACM on Multimedia Conference* (pp. 640-644). ACM.
[17] Sindagi, V. A., & Patel, V. M. (2017, October). Generating high-quality crowd density maps using contextual pyramid cnns. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 1879-1888). IEEE.