ISBN: 978-1-948012-17-1

Ethics and Information Technology (ETIT)

DOI: <http://doi.org/10.26480/etit.02.2020.134.137>

CrossMark

A LOOK AT ADVANCES IN CNN-BASED CROWD COUNTING APPROACHES

Deepesh Deshmukh^a, Chaitanya Deo^a, Ishwar Nigam^a, Mr. Sourabh Dave^{a,b}^a CSE Department, Medi-Caps University, Indore, India^b CSE Department, Medi-Caps University, Indore, India*Corresponding Author Email: kumardeepesh02@gmail.com

This is an open access article distributed under the Creative Commons Attribution License CC BY 4.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ARTICLE DETAILS

Article History:

Received 25 October 2020

Accepted 26 November 2020

Available online 03 December 2020

ABSTRACT

Crowd counting is one of the booming topics and has a large number of applications such as urban planning, public safety, video surveillance etc. Our aim is to compare and analyze and highlight architectural specifications of three state-of-the-art convolutional neural network (CNN) based crowd counting algorithms. These algorithms were developed to perform successfully in a wide range of scenarios, the occurrence of many challenges such as uneven density in the crowd scene, variations in perspective and scale of the image. Spatial Divide and Conquer Network (S-DCNet) makes use of divide and conquer based strategy in the spatial domain of image. In Congested Scene Recognition Network (CSR-Net), it uses dilated kernels and a simple CNN front-end. Spatial/Channel-wise Attention Regression Network (SCAR) instead of just focusing on local appearance it mainly focuses on pixel-wise context of the image and tries to reduce the mistakes on background estimations. From the study, SCAR shows the best performance among these.

KEYWORDS

CNN, Computer Vision, Crowd Counting, CSRNet, SCAR, SDCNet.

1. INTRODUCTION

Computer Vision is to see beyond what humans cannot. There are numerous applications of computer vision like image segmentation, object/motion detection, image similarity, and video analysis. Crowd counting is one such application that has immense application in the area of computer vision which can be used to accurately approximate the count of people present in an image or video. It can help in estimating the density of the crowd and thus provides a lot of data for the analytics; this data can be useful for further research.

The problem of crowd counting and density estimation is of paramount importance and it is essential for building higher level cognitive abilities in crowded scenarios such as crowd monitoring (Chan et al., 2008) and scene understanding (Shao et al., 2015; Zhou et al., 2012). Due to an increase in population globally and the rise of urbanization has led to many rallies, public gatherings, sports events, concerts etc.

Applications of crowd counting and analysis: -

- Safety Monitoring: Places such as stadiums, stations, airports etc. are prone places where public safety is indispensable, footage from video surveillance cameras can be helpful. By using these algorithms various analyses can be achieved such as behavior and congestion analysis.
- Infrastructure Planning: Analyzing crowds in public places such as train stations, bus stops, airports etc. can help in realizing shortcomings and help in optimizing the design of such places.
- Wildlife Monitoring: Crowd counting and analysis can help in getting accurate, detailed information about the animals living in the wild, which in itself is a huge task.

In such cases, crowd counting can help in analysing crowd behaviour and helps in better management.

The problem arises in computer vision is when the image has low resolution or having very high resolution, the image does not have enough luminance, the crowd in the image is non-uniform, scale and perspective of an image also influences the results majorly.

In the past few years many researches have been conducted for computer vision in crowd counting and density estimation, various methods have been proposed such as Convolutional Neural Network based approach, Regression based methods, Density based methods and Density estimation-based methods.

In this work, we mainly compare the analogy and divergence of different networks used in Convolutional Neural Network i.e. SCAR (Spatial/Channel-wise Attention Regression Network) (Gao et al., 2019), SDC-Net (Spatial Divide and Conquer) and CSR-Net (Xiong et al., 2020; Li et al., 2018).

We chose these architectures for our comparison paper because these architectures capture high level features and produce high-quality density maps. These architectures are used greatly because of their effective results.

2. LITERATURE REVIEW

2.1 Detection Based methods

In detection-based methods the detection of crowd scenes is done by a sliding window-like detector (Han et al., 2004). It creates bounding boxes around the objects according to the object's scale. Detection usually is done in the part-based or either by monolithic detection. These are trained classifiers that can detect low-level features. Monolithic detection uses the features that are extracted from the whole body. The success of results can

Quick Response Code	This paper was presented at	Access this article online	
	International Conference on Contemporary Issues in Computing (ICIC-2020) - Virtual IETE Sector V, Salt Lake, Kolkata From 25th-26th July 2020	Website: www.intelcomp-design.com	DOI: 10.26480/etit.02.2020.134.137

be affected and can inflate by using Support-Vector machines and Random Forest. Whereas part-based uses, boosted classifiers such as boosted trees for specific body parts. Though these methods work well for detecting faces, it is inefficient on crowded scenes because of motion scenes and some images do not have clarity.

2.2 Regression-based approach

The features extracted from local image patches are mapped to counts (Chen et al., 2012). It does not involve tracking or segmentation or any kind of detection. Some of the features used for encoding low-level information are edge features, texture features, foreground feature and gradient features. To improve the results further both global properties and local features are used together. Various regression techniques used are linear regression (Paragios et al., 2001), Gaussian process regression, neural network (Marana et al., 1998), ridge regression (Chen et al., 2012) and piecewise linear regression to learn low-level feature mapping to crowd count (Chan et al., 2008).

2.3 Deep Learning Methods

In the above approaches hand-crafted features were used which have low representation capability. CNN based approaches are one of the most reliable approaches for crowd counting.

These methods are accurate and flexible which permits ado of crowd counting with static images whilst images can be of arbitrary size and resolution. Deep convolutional neural network proposed by researchers constructed the framework which does not rely on foreground segmentation but the texture of the crowd (Zhang et al., 2015). In Multi-column convolutional network proposed by a research they used convolutional layer of 1x1 filter size with three columns of filters and receptive field (Zhang et al., 2016).

Deep-learning approaches include convolutional neural networks as one of the many types of algorithms, as CNN-based algorithms are considered best for computer vision tasks, here we did comparison of CNN based algorithms.

3. COMPARISON

3.1 Spatial Divide and Conquer Network

The nature of the problem of counting people in image is not a closed-set problem as the range of count of values can vary from 0 to infinity. Thus the problem is countered using a regression model. Since convolutional neural networks have shown promising results with multi-branch architectures in various vision tasks, modern CNNs for deep counting can adopt this idea to enhance the performance on dense regions (Deepak et al., 2017; Xinkun et al., 2018; Zenglin et al., 2018). But the patterns observed in datasets are limited, resulting in network learning from closed sets.

Figure 1: Illustration of Spatial divisions. Assume that closed set count is $[0, 20]$. The instance indicates how the image needs to be divided. Just a single division isn't sufficient. The sub regions need further division.

Counting has a special property of spatial decomposition. Thus the problem of crowd counting can be solved using the idea of Spatial Divide and Conquer (S-DC) as illustrated in [Figure 1].

In this way regardless of whether the model is prepared on a fixed range, 0~20, by isolating the picture into sub pictures until sub-area tallies are under 20, at that point the system can precisely tally these sub images and entirety over every neighbourhood check to acquire the worldwide picture tally (Xiong et al., 2020). The [Figure 2] illustrates the concept of S-DC.

Figure 2: Spatial Division on input image (left) and Feature map (right). To estimate local count the input image is partitioned spatially and up-sampled using the same network. The repeated up-sampling and computation, decoding and dividing feature map of hi resolution is avoided on feature map S-DC.

The network is inspired from RoI pooling (Girshick et al., 2015), it is shown that it is feasible to apply S-DC on a feature map, as conceptually demonstrated in [Figure 3]. The feature map is then decoded and up-sampled, the following prediction layers focus on features of local areas and predict sub region counts. The S-DC network is based upon an encoder built on VGG16 (Simonyan et al., 2014) and an UNet (Ma et al., 2019) like decoder to produce feature maps which are multi-resolution and all share the same counting predictor. Rather than density map regression, continuous count values are discretised into sets of intervals. Also a division decoder is used for deciding which sub region to divide and merge at different levels of sub-region count.

Figure 3: The architecture of S-DCNet (left) and a two-stage S-DC process (right).

The architecture take on the convolutional layers of VGG16 (Han et al., 2004). The decoder for up-sampling and dividing feature map is inspired from UNet (Ma et al., 2019). To generate the local count C_i and division mask W_i the partitioned feature map is passed classifier and division decoder. The local counts and division masks are merged to find division count DIV_i . Before merging these counts the low resolution count are averaged 2×2 areas of high resolution. "0" represent the Hadamard product.

Rather than regressing count value in image S-DCNet follow (Li et al., 2018) to break local counts into intervals and classify image counts. Specifically, S-DCNet defines and interval partition of $[0, +\infty)$ as $\{0\}, [0, C_1], [C_2, C_3], \dots, [C_{M-1}, C_M]$ and $[C_M, +\infty)$. These $M+1$ sub-intervals are labelled to the 0-th to N-th classes, respectively (Gao et al., 2019). The value of C_M is set such that it is not greater than max local count of the training dataset. For the value count of sub-interval median of that sub interval is used, in case if a region is classified as last sub-interval $[C_M, +\infty)$, C_M is taken as the count value for that region. Though using C_M as the count value may result in systematic error, but this error is diminished using S-DC. [Figure 3] shows the architecture and two stage process of S-DCNet.

3.2 Congested Scene Recognition

It is a data-driven and deep learning method which is used in crowd estimation more accurately in a highly congested scene. The front-end of the network uses VGG16 (Simonyan et al., 2014); having strong ability of transfer learning and its architecture being flexible allows easy addition of the back-end for density map generation. In case of CSR-Net the Classification part of VGG-16 (fully connected layers) is removed and

proposed CSR-Net with convolutional layers in VGG-16 is built causing front end network output 1/8 of input size.

The back end of CSR-Net is built by adding dilated convolutional layers for extracting unique deep information and maintaining output resolution.

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			

Figure 4: The configuration of Congested Scene Recognition (CSR-Net).

A 2-D dilated convolution is calculated using:

$$y(a, b) = \sum_{i=1}^A \sum_{j=1}^B x(a + n \times i, b + n \times j) w(i, j) \quad (1)$$

Applying dilated convolution input $x(a, b)$ and a filter $w(i, j)$ with the length and the width of A and B respectively gives output $y(a, b)$; where n is the dilation rate. For $n=1$, a dilated convolution is equivalent to normal convolution. So, Eq.(1) explains how a 2-D dilated convolution can be calculated.

In dilated convolution, a small size kernel with $(a \times a)$ filter is enlarged to $(a+(a-1)(n-1))$ with dilated strider. For the purpose of maintaining the resolution of the feature map, the processing of the input image is done by two approaches to generate separate same size outputs.

Figure 5: 3 x 3 convolutional kernels with different dilation rate

By the means of the Sobel kernel of 3x3 matrix, the down sampled image obtained by max pooling by factor 2, features the first approach. In the other approach, dilated convolution is used with the same 3 x 3 kernel with a factor of stride 2 [Figure 5.]. The output shares the same dimension as the input. So, the output from dilated convolution contains detailed information.

3.3 Spatial-/Channel-wise Attention Regression

SCAR is a CNN based counting algorithm which focuses on wide-range contextual information present pixel-wise, attention information and local appearance features. SCAR is used to estimate the density map. Since most of the CNN based algorithms mainly focus on local features this does not allow effective encoding of wide-range contextual information. Many

traditional Fully Convolutional Networks cannot encode relations of channels in a layer and fail to predict background as a potential crowd present region.

A sequential pipeline is followed in SCAR architecture as in [Figure 6] which defines the architecture of SCAR, in which the image data is processed consecutively. The first 10 convolutional layers of VGG-16 Net (Simonyan et al., 2014) are used to extract local features, and the other two convolutional layers with 1×1 kernel size are present in Map Regressor.

Figure 6: The flow chart of SCAR.

The input is first fed into a module consisting of VGG-16 Backbone [17] and dilation module, responsible for local feature extraction.

The output feature maps are then fed to two modules. SCAR consists of two modules named Spatial-wise Attention Model (SAM) and Channel-wise Attention Model (CAM). Spatial-wise Attention Model (SAM) plays an important role in capturing the change in density distribution across regions and it also models the wide-range contextual information, whereas Channel-wise Attention Model (CAM) used to learn relation across channel dimensions of data. The outputs are then concatenated and the density map is generated by the convolutional layer and up-sample operation.

SCAR	
VGG-16 backbone	
conv1:	$[k(3,3)-c64-s1-R] \times 2$
	Max polling
conv2:	$[k(3,3)-c128-s1-R] \times 2$
	Max polling
conv3:	$[k(3,3)-c128-s1-R] \times 3$
	Max polling
conv4:	$[k(3,3)-c512-s1-R] \times 3$
Dilation Module	
	$k(3,3)-c512-s1-d2-R$
	$k(3,3)-c512-s1-d2-R$
	$k(3,3)-c512-s1-d2-R$
	$k(3,3)-c256-s1-d2-R$
	$k(3,3)-c128-s1-d2-R$
	$k(3,3)-c64-s1-d2-R$
Attention Model	
SAM	CAM
Concat	
Regression Layer	
	$k(1,1)-c1-s1-R$
	Up-sample: $\times 8$

Figure 7: The network architecture of SCAR.

4. DISCUSSION AND ANALYSIS OF PERFORMANCE

The Mean Absolute Error and Mean Squared Error are used to compare and evaluate different methods.

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^N |y_i - y'_i|^2} \quad (2)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - y'_i| \quad (3)$$

Where MSE is mean squared error, MAE is mean absolute error, N is total samples, y_i is ground truth, y'_i is estimated count of i_{th} sample. NWPU dataset consists of various scenes and has a density range that goes upto 20k, which is quite high as compared to other existing datasets.

This dataset in particular consists of high resolution images, and average is as high as 2311 x 3383, with large appearance variations.

The ShanghaiTech Dataset (Li et al., 2018) is present in two parts namely Part_A and Part_B, Part_A mainly consists of highly congested scenes and other one contains relatively sparse scenes. This dataset is highly imbalanced. In Part_A more than 80% images consist of crowd count less than 900, and in Part_B more than 80% images are with crowd count less than 240.

Channel wise Attention (CAM) and Spatial Attention (SAM) module in SCAR helps it to leverage not only local information but pixel-wise information and better background density estimation as well. SCAR hugely benefits from its approach of encoding channel relations.

Convolutional layers with dilated kernels in CSR-Net; helps replace pooling layers and maintain the spatial resolution of image and to expand the receptive field which in turn reduces model complexity and the number of parameters.

SDC-Net here is a model with a different and unique approach to solve crowd counting problems, the model can answer the problem of dataset shift. During test time if models are given an image with a number of people which haven't been encountered during training, the S-DC-Net can perform exceptionally well in such cases also, since it uses divide-and-conquer approach. The other two models might perform drastically bad in such situations.

On NWPU dataset, the performance of SCAR is better than S-DC-Net and CSR-Net, MSE and MAE scores of SCAR were better or lower as compared to the other two. This shows that SCAR and its modifications can perform very well in above average-density scenarios and are able to outperform in highly-dense population scenes. Also CSR-Net might not perform best in high density scenes; it performs reasonably well in low density scenes and is a very versatile model. Also from Table.1 it can be seen that S-DC-Net might be a better player when it comes to the huge difference in joint distribution of inputs and outputs during the training and testing phase, since its MAE and MSE are quite low on ShanghaiTech Dataset.

Table 1: Comparison of approaches on ShanghaiTech dataset

Method	Part_A		Part_B	
	MAE	MSE	MAE	MSE
CSRNet	68.2	115.0	10.6	16.0
SCAR	66.3	114.1	9.5	15.2
SDCNet	58.3	95.0	6.7	10.7

Table 2: Comparison of approaches on NWPU dataset.

Method	MAE	MSE
CSRNet	121.3	387.8
SCAR	110.0	495.3

5. CONCLUSION

In this paper, we compared differentiated CSR-Net, SCAR and SDC-Net on the basis of their architectures and their performances on NWPU. CSR-Net uses convolutional layers with dilated kernels which helps in maintaining spatial resolution and simply the network. In addition, CSR-Net also has good versatility. By using attention modules on spatial domain and channels, SCAR is able to reduce mistakes in background density estimation thus being good at highly-dense crowd images. By using divide and conquer strategy on spatial domain, SDC-Net is able to generalize its performance on closed-set problems/images. From the overall results, SCAR is the most accurate and best performing CNN among these. From CSR-Net to SCAR, we have found that the strategies in developing the structure of CNNs have evolved, and the depth of architectures has deepened. Also CNN networks that can incorporate additional information such as context and scale may perform immensely well at density estimation and may dominate the Crowd Counting field.

REFERENCES

Cao X, Wang Z, Zhao Y, and Su F, 2018. Scale aggregation network for accurate and efficient crowd counting. In the European Conference on Computer Vision (ECCV), 734–750.

Chan, A.B., Liang, Z.S.J., Vasconcelos, N., 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE. 1–7.

Chen, K., Loy, C.C., Gong, S., Xiang, T., 2012. Feature mining for localised crowd counting., in: European Conference on Computer Vision.

Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu, 2017. Switching convolutional neural network for crowd counting. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5744–5752.

Gao J, Wang Q, Yuan Y. 2019. SCAR: Spatial-/Channel-wise Attention Regression Networks for Crowd Counting, arXiv:1908.03716

Girshick R, 2015. Fast R-CNN. In Proc. IEEE International Conference on Computer Vision (ICCV), 1440–1448.

Han M, Sethi A, Hua W, Gong Y. 2004. A detection-based multiple object tracking method. In 2004 International Conference on Image Processing, 2004. ICIP'04. 5, 3065–3068. IEEE

Li R, Xian K, Shen C, Cao Z, Lu H, and Hang L. 2018. Deep attention-based classification network for robust depth prediction. 2, 3.

Li Y., Zhang X., & Chen D. 2018. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1091–1100.

Ma Z, Wei X, Hong X, and Gong Y. 2019. Bayesian loss for crowd count estimation with point supervision. In Proceedings of the IEEE International Conference on Computer Vision, 142–615.

Marana, A., Costa, L.d.F., Lotufo, R., Velastin, S., 1998. On the ecacy of texture analysis for crowd monitoring, in: Computer Graphics, Image Processing, and Vision, 1998. Proceedings. SIBGRAPI'98. International Symposium on, IEEE. 354–361.

Paragios, N., Ramesh, V., 2001. A mrf-based approach for real-time subway monitoring, in: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, IEEE. 1–1034.

Shao, J., Kang, K., Loy, C.C., Wang, X., 2015. Deeply learned attributes for crowded scene understanding, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. 4657–4666.

Shi Z, Zhang L, Liu Y, Cao X, Ye Y, Cheng MM, and Zheng G, 2018. Crowd counting with deep negative correlation learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5382–5390.

Simonyan K., Zisserman A., 2014. Very deep convolutional networks for largescale image recognition, arXiv preprint arXiv: 1409.1556.

Xiong H, Lu H, Liu C, Liu L, Shen C, Cao Z. 2020. From Open Set to Closed Set: Supervised Spatial Divide-and-Conquer for Object Counting, arXiv:2001.01886

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. 833–841.

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. 589–597.

Zhou, B., Wang, X., Tang, X., 2012. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE. 2871–2878.