DFE-Crowd: Dense Feature Extraction for Single Image Crowd Counting

Naveed Ilyas, Andres Caceres Najarro, and Kiseon Kim
Gwangju Institute of Science and Technology

Abstract

In this paper, we address the effective utilization of features from lower and lower-middle layers to enhance crowd counting performance. Generally, the features extracted at lower layers are utilized locally and are not passes to higher layers, which result in low counting accuracy. To tackle this issue, we proposed a dense feature extraction model. Combination of task independent general features and task specific features are very useful to obtain high counting accuracy. We have tested the proposed technique on Venice dataset. Results justify its relative effectiveness in terms of selected performance.

I. Introduction

The number of people and their spatial distribution are two significant measurements for understanding crowded scenes. With the boost of convolutional neural networks (CNNs), various CNN-based crowd counting algorithms have mushroomed for addressing the difficulties of crowd counting [1]. One of the most significant advantages of CNN-based crowd counting is its ability to learn powerful features.

Existing CNN-based crowd counting techniques enhance the counting accuracy by using well known networks such as multi-column, multi-tasking, dilated, and de-convolutional [2]. These networks have been widely used individually or in combination with each other to increase the performance at the cost of major shortcomings such as large amount of training time, ineffective branch structure, sparse pixel sampling rate, information loss, and extraction of irrelevant information. Authors in [3] and [4] used a multi-column architecture for density estimation by taking advantages of different receptive fields. Besides multicolumn, kernel size in each column is fixed which means that column can only handle a specific set of density scenes. Further, multi-columns in [1] and [4] learned similar type of features irrespective of different filter size [5].

Based on these observations, we proposed a dense feature extraction model (DFE-Crowd). Our model comprises of two main sub-modules, (i) general feature extraction module (GFEM), and (ii) dense feature extraction module (DFEM). The proposed model is capable to aggregate the task independent and task specific features at higher layers from lower and middle-lower layers to enhance the crowd counting accuracy while obtaining the relevant contextual information.

II. The Proposed Model

The architecture of the proposed approach is shown in Fig. 1. Firstly, DFE-Crowd counting starts from ground truth density estimation. Secondly, our proposed model employs GFEM (inspired from VGG-16) which is used to obtain simple to complex deep features. Thirdly, our proposed network utilizes DFEM which enables the network to extract deep and relevant features. Fourthly, multiple deep stacked convolution (DSC) blocks in DFEM are densely connected with each other, thus enhancing the ability of the network to handle perspective distortion while propagating the information to higher layers. Lastly, the output from subsequent DSCs (densely connected with each other) are aggregated by using fusion module (FM). Thus, output of one DSC has direct access to each layer of the subsequent DSCs, resulting in a contiguous information pass collected through FM.

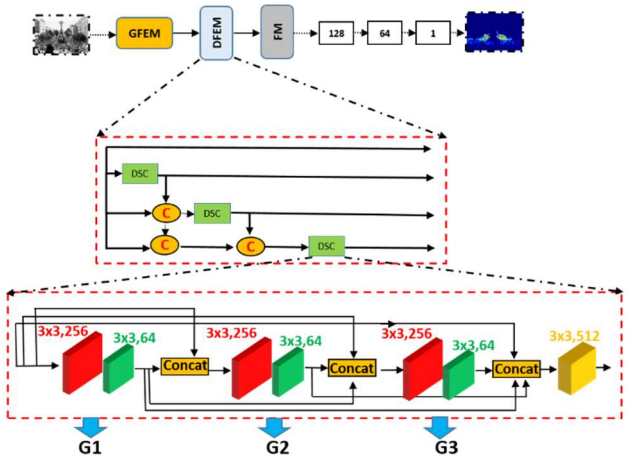

Fig.1 DFE-Crowd: A dense feature extraction network for single image crowd counting

A. Dense Feature Extraction Module (DFEM)

Pedestrians in a crowded scene usually suffer from perspective distortion. To address this challenge for enhance counting accuracy, the estimated feature map must have comprised of low to complex, dense features. Therefore, we built a DFEM by cascading three deep stacked convolution (DSC) blocks as shown in Fig. 1 (middle). We include three DSCs in DFEM module that are densely connected with each other. The upper DSC accepts output from lower ones, which result in aggregation of information from lower and lower-middle layers to upper layers. Further, task independent general features extracted at lower layers are propagated to higher layers thus combined with task specific features extracted at higher layers. In addition, each DSC block is further divided into three groups (G1, G2, and G3) each of these has two

convolutional layers with varying channel sizes. The G3 accepts output from G2, and G2 from G1, such that the upper group accepts output from lower ones as shown in Fig. 1 (bottom). In this way all the groups are densely connected with each other, which results in extraction of deep features.

As higher layers feature always contain more semantic information, and low-layer features contain more detail information. Combination of the features extracted from low-level layers and high-level layers plays an important role for obtaining relevant contextual information. An effective combination of GFEM and DFEM plays a vital role to increase the counting accuracy by aggregating detailed and semantic features. Especially, the DFEM with densely oriented structure is useful to extract and propagate the features to subsequent layers in a dense fashion.

III. Simulation Setup and Results

In this section, we evaluate the crowd counting accuracy on Venice dataset. The proposed technique is compared and evaluated by using mean absolute error (MAE) and mean square error (MSE) which is defined as:

$$MAE = \frac{1}{N}\sum_{i}^{N} |y_i - y_i'|, \qquad (1)$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(y_i - y_i')^2}. \qquad (2)$$

Where N is the number of images in one test sequence, y_i is the estimated count, and y_i' is the corresponding ground truth count. The first 10 layers are fine-tuned from pre-trained VGG-16 architecture. The DFE-Crowd is trained by Adam optimizer with learning rate 1e-6 and momentum 0.9. We use PyTorch plateform with NVIDIA GeForce GTX 1070 with 8 GB memory.

The Venice dataset contains 167 images with fixed resolution of 1280×720. It is collected from Venice city with varying perspective. Sparse and non-uniform density levels make it a low density dataset. We compare DFE-Crowd with existing state-of-the-art counterpart [3,4,5,6] as shown in Table I.

Table. I Crowd Counting Accuracy on Venice Dataset

Technique	Venice Dataset	
	MAE	MSE
MCNN [3]	145.4	147.3
Switch-CNN [4]	52.8	59.5
CSRNet [5]	35.8	50.0
HydraCNN [6]	35.0	**29.9**
DFE-Crowd	**23.83**	34.59

Table I depicts the performance comparison between the proposed DFE-Crowd and state of the art techniques based on the MAE and MSE. It is observed that the DFE-Crowd provides the lowest and second lowest counting error based on the MAE and MSE, respectively. The reason is extraction of low to complex, deeper, and relevant features from lower, middle and higher layers, provides better information. Dense connections among DSCs are useful to propagate the information to higher layers. Further, aggregation of task independent and task specific features extracted from lower and lower-middle layers enhance the counting accuracy. The combination of local and global features further enhanced the counting accuracy by incorporating salient features for final density map.

IV. Conclusion

In this work, we proposed a DFE-Crowd using dense feature extraction for single image crowd counting that is trained in an end-to-end manner. Due to strong relevant feature aggregation property from lower and lower middle layer to higher layer, the performance is enhanced in terms of counting accuracy. The combination of local and global features has been shown to be useful for improving the crowd counting accuracy. In future, we intend to use un-supervised learning with manifold approach to further enhance the counting performance.

ACKNOWLEDGMENT

This research was a part of project titled Development of Ocean Acoustic Echo Sounders and Hydro-Physical Properties Monitoring System, funded by Ministry of Oceans and Fisheries, Korea.

Reference

[1] N. Ilyas, A. Ahmad, and K. Kim. "CASA-Crowd: a context-aware scale aggregation CNN-based crowd counting technique," *IEEE Access*, vol.7, pp. 182050-182059, Dec. 2019.

[2] N. Ilyas, A. Shahzad, and K. Kim. "Convolutional-neural network-based image crowd counting: review, categorization, analysis, and performance evaluation," *Sensors*, vol. 20, Jan. 2020.

[3] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "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, June 2016.

[4] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*., vol. 1, pp.4031-4039, July 2017.

[5] Y. Li, X. Zhang, and D. Chen, "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, June 2018.

[6] D. Onoro-Rubio and R. J. L´opez-Sastre, "Towards perspective-free object counting with deep learning," *In European Conference on Computer Vision.*, pp. 615–629, Mar. 2016.