International Journal of Advanced Smart Convergence Vol.12 No.4 147-153 (2023)
http://dx.doi.org/10.7236/IJASC.2023.12.4.147

IJASC 23-4-18

Density Change Adaptive Congestive Scene Recognition Network

Jun-Hee Kim†, Dae-Seok Lee††, Suk-Ho Lee†††

†*Bachelor Degree Candidate, Dept. of Electronic Engineering, Dongseo University, Korea*
††*Team Manager, Buil Planning Co., Korea*
†††*Professor, Dept. Artificial Intelligence Appliance, Dongseo University, Korea*

E-mail petrasuk@gmail.com

Abstract

In recent times, an absence of effective crowd management has led to numerous stampede incidents in crowded places. A crucial component for enhancing on-site crowd management effectiveness is the utilization of crowd counting technology. Current approaches to analyzing congested scenes have evolved beyond simple crowd counting, which outputs the number of people in the targeted image to a density map. This development aligns with the demands of real-life applications, as the same number of people can exhibit vastly different crowd distributions. Therefore, solely counting the number of crowds is no longer sufficient. CSRNet stands out as one representative method within this advanced category of approaches. In this paper, we propose a crowd counting network which is adaptive to the change in the density of people in the scene, addressing the performance degradation issue observed in the existing CSRNet(Congested Scene Recognition Network) when there are changes in density. To overcome the weakness of the CSRNet, we introduce a system that takes input from the image's information and adjusts the output of CSRNet based on the features extracted from the image. This aims to improve the algorithm's adaptability to changes in density, supplementing the shortcomings identified in the original CSRNet.

Keywords: *People Counting, Congestive Scene, Deep Learning, CCTV*

1. Introduction

Recently, due to a lack of management of the crowd in crowded areas, there have been numerous incidents of stampedes since the occurrence of the crowded accidents. In response to this, each local government is constructing a system to measure the real-time flow of people to manage the crowd. To achieve this, there is a growing interest in researches that can measure the flow of people in real-time based on the CCTV system. In other words, algorithms that count the number of people based on the images acquired through CCTV are gaining attention. When these algorithms are combined with CCTV systems, they can be widely used in various fields such as security, public safety, retail, and more.

Artificial neural networks performing crowd counting can be broadly categorized into object detection-

Manuscript Received: october. 14, 2023 / Revised: october. 20, 2023 / Accepted: october. 25, 2023
Corresponding Author: petrasuk@gmail.com
Tel: +82-51-320-1744, Fax: +82-51-327-8955
Professor, Department of Computer Engineering, General graduate school, Dongseo University, Korea

based networks, regression-based networks, and density measurement-based networks[1][2][3][4][5]. Among these, the object detection-based method is an early approach where human objects are directly detected, and the count of detected people is then calculated. However, this method shows vulnerabilities for images with high density, as densely packed objects may partially obscure each other, making it challenging to accurately detect people. To address this, regression-based artificial neural networks were developed to learn the relationship between features extracted within local patches and the number of people. Subsequently, density measurement-based methods combined the information of saliency maps with the regression-based approach, resulting in more accurate crowd counting. Many CNN models have integrated multi-scale information to enhance the accuracy of these models[6][7][8][9]. However, one of the significant contributions of the paper proposing CSRNet is experimental evidence that using multi-scale does not necessarily lead to improved performance[10]. This highlights that, even without complex calculations and intricate structures, satisfactory crowd counting performance can be achieved. The CSRNet has gained significant influence in the relevant research field by using a simple structure that does not use multi-scale but still shows good counting performance.

However, the CSRNet exhibits a weakness in performance when applied to test data that shows different densities than the densely populated image data it was trained on. In this paper, we propose a method to improve the performance of crowd counting by combining the self-attention mechanism and adding a neural network that reflects the change in the density. Experimental results demonstrate that a higher accuracy compared to the existing CSRNet can be achieved with the proposed modification.

2. Congestive Scene Recognition Network(CSRNet)

The original CSRNet focuses on encoding deep features of crowded scenes and generating high-quality density maps. Additionally, it proposes the use of dilated convolution to mitigate the loss of spatial resolution information caused by pooling layers. The original CSRNet is trained by a loss function that calculates the mean squared error between the predicted density map $Z(X_i;\ \theta)$ and the ground truth density Z_i^{GT} for each input image X_i in the dataset:

$$L_1(\theta) = \frac{1}{2N}\sum_{i=1}^{N} ||Z(X_i;\ \theta) - Z_i^{GT}||_2^2 \tag{1}$$

Here, N represents the batch size, and θ is the parameter of CSRNet. $Z(X_i;\ \theta)$ denotes the output of CSRNet, which is the generated output, and Z_i^{GT} is the ground truth density map. At the test time, the people count value C_i for the input image X_i is computed using the following formula:

$$C_i = \sum_{l=1}^{L}\sum_{w=1}^{w} Z_{l,w} \tag{3}$$

Here, L and W represent the length and width of the density map, and $Z_{l,w}$ is the brightness value of Z at position (l, w) in the density map. One of the weaknesses that the CSRNet exhibits is that it cannot adaptively address changes in the density. This weakness stems from the fact that for different densities, the density maps show different characteristics as the overlapping of targets show different characteristics for different densities. This results in a drop of accuracy in the counting, when applied to a scene with different

density than the data the CSRNet has been trained on.

3. Proposed Method

In this paper, we propose a model that combines a self-attention module to address the existing weaknesses in CSRNet, along with a network $\alpha(\cdot)$ for adjusting coefficient values. Figure 1 illustrates the diagram of the proposed system. The self-attention module [11] has been a successful module used in Vision Transformer (ViT). This module treats the pixels or patches as the elements of the sequence and divide the input sequence into the query, the key, and the value parts. The attention scores are then computed by the dot product of the query and the key vectores, and then, the relationships between different pixels or patches are captured to attend to relevant parts of the input when making predictions. The reason for adding the self-attention module to the existing CSRNet is to improve the accuracy by paying attention to surrounding patterns in the image. Crowd images often have complex structures and patterns because many people are close to each other. Through the self-attention, the model is better able to detect various poses and directions of people, especially in densely populated areas where individuals may be obscured by others or objects.

The additional network, i.e., the density change adaptive network $\alpha(\cdot)$ is implemented as a CNN and serves to adjust the predicted people count value based on the image features. In other words, when the i-th input image is X_i, instead of using the predicted coefficient value C_i from the original CSRNet, the final coefficient value is corrected by multiplying it with $\alpha(X_i; \varphi)$ denoted as $\alpha(X_i; \varphi) * C_i$. The role of $\alpha(\cdot)$ is to adjust the value of C_i by reflecting the features of the image X_i as $\alpha(\cdot)$ takes as the input the image X_i.

In the proposed model, the CSRNet is trained using the original loss function proposed in [4]. The loss function calculates the mean squared error between the predicted density map $Z(X_i; \theta)$ and the ground truth density Z_i^{GT} for each input image X_i in the dataset:

$$L_1(\theta) = \frac{1}{2N}\sum_{i=1}^{N} ||Z(X_i; \theta) - Z_i^{GT}||_2^2 \tag{3}$$

Here, N represents the batch size, and θ is the parameter of the CSRNet. $Z(X_i; \theta)$ denotes the output of CSRNet, which is the generated output, and Z_i^{GT} is the ground truth density map. The network $\alpha(\cdot)$ is trained using the mean squared error as the loss function, calculated between the product of the output result of $\alpha(\cdot)$ and the existing crowd prediction, and the actual crowd count:

$$L_2(\varphi) = \frac{1}{N}\sum_{i=1}^{N} |\alpha(X_i; \varphi) * C_i - C_i^{GT}| \tag{4}$$

Here, C_i^{GT} represents the actual number of people in the image X_i, and C_i is the number of people calculated by CSRNet's predicted density map $Z(X_i; \theta)$ for the image X_i, computed using the following formula:

$$C_i = \sum_{l=1}^{L}\sum_{w=1}^{w} Z_{l,w} \tag{5}$$

Here, L and W represent the length and width of the density map, and $Z_{l,w}$ is the brightness value of Z at position (l, w) in the density map. In the proposed model the final predicted number of people is not C_i as used in [10], but the modified number $\alpha(X_i; \varphi) * C_i$.

The design of the structure of $\alpha(\cdot)$ needs a lot of care, since the input is an image and the output of $\alpha(\cdot)$ is a single value which has to compensate for the differences in the densities in the images. In other words, the single output value has to reflect the characteristics of the whole image region. However, this is not easy because the size of the image has to be reduced to a single value, which will produce loss in the information.

In our experiments, using a mere max pooling layers results in a loss of information. Therefore, we use the max pooling layers only at the first and the second layers of the network, and use the average pooling before the last layer of the network.

For the density adaption network $\alpha(\cdot)$, we use a convolutional neural network with 5 layers with max pooling layers in the first and the second layers. Before the last layer the average pooling shrinks the spatial size of the to 8×8. Then, the last layer flattens the feature map and outputs a single value. Figure 1 shows the overall diagram of the proposed method.

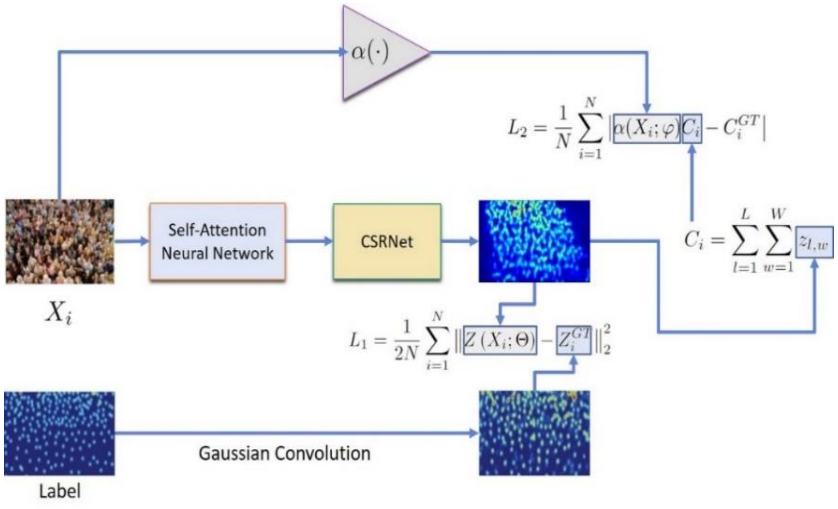

Figure 1. Data Flow Diagram of the Proposed Method

In an alternative version of the proposed method, we suggest incorporating the density adaption network, denoted as $\alpha(\cdot)$, with the density map directly as its input. This suggestion is grounded in the assumption that the density map inherently captures the characteristics of density within the image. Additionally, we advocate for applying self-attention to the density map, as it often delineates relationships between distinct objects more effectively than the original RGB image. Figure 2 illustrates the schematic representation of this variant in our proposed method. We introduce both approaches, recognizing that the first method can also be further evolved in alternative directions, potentially leading to enhanced algorithms.

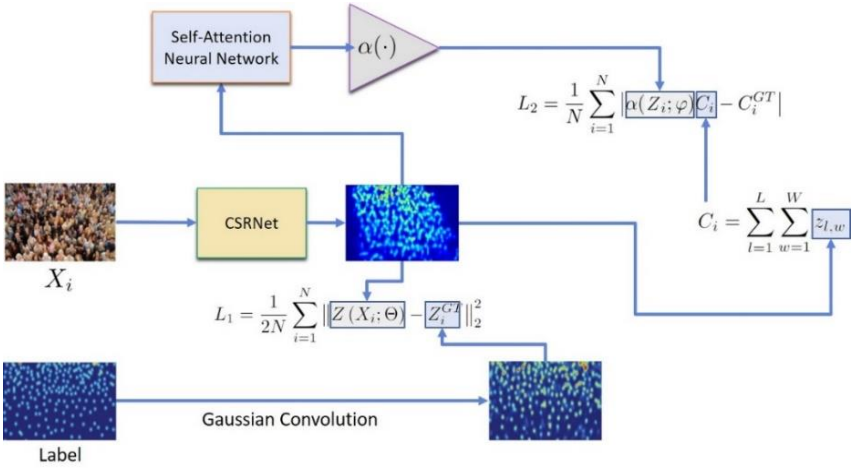

Figure 2. Data Flow Diagram of the Variant of the Proposed Method

4. Experimental Results

The training of the proposed model utilized the Shanghai dataset [3], which consists of 1,198 labeled images containing a total of 330,165 individuals. When evaluating the performance of the proposed model, we employed the Mean Absolute Error (MAE) measure. The conventional CSRNet measures the MAE using the following formula:

$$MAE = \frac{1}{N_{test}}\sum_{i=1}^{N}|C_i - C_i^{GT}| \tag{6}$$

As the proposed model predicts the number of people using $\alpha(X_i;\varphi) * C_i$, the MAE is calculated with the following formula:

$$MAE = \frac{1}{N_{test}}\sum_{i=1}^{N}|\alpha(X_i;\varphi)C_i - C_i^{GT}| \tag{7}$$

Figure 3 shows two experimental results where the proposed method shows a more correct crowd counting result than the original CSRNet. Here, we used the variant of the proposed method. The results show that even though the two scenes have different numbers of crowds, the proposed method can adaptively adapt itself to each situation.

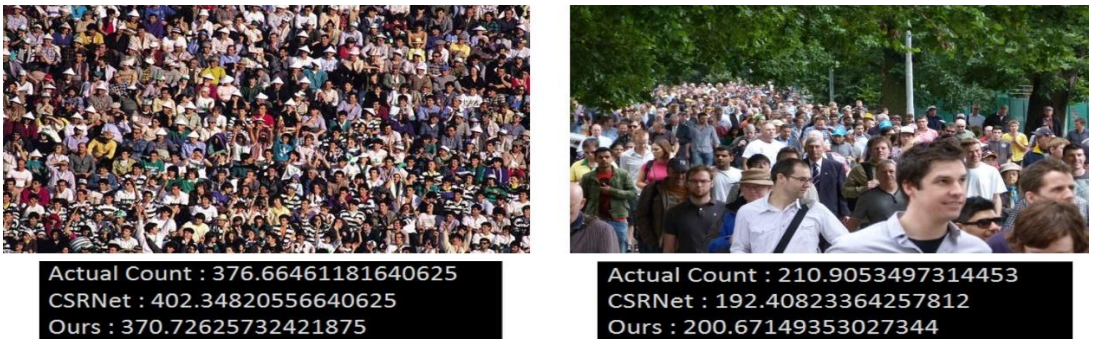

Figure 3. Selected crowd counting results on two different scenes

Table 2. Comparison of Estimation errors on ShanghaiTech Part A dataset between the various crowd counting methods. Results for CP-CNN, Switching CNN, MCNN are from [10].

Method	MAE
MCNN [6]	110.2
CP-CNN [7]	73.6
Switching CNN [8]	90.4
CSRNet [10]	68.4
Proposed	67.7

When experimenting with the Part A test dataset of the Shanghai dataset, the conventional CSRNet yielded an MAE of 68.42, while the proposed model exhibited an MAE of 67.7. In other words, the proposed model showed an improvement in performance by approximately 0.64. Table 2 compares the MAE between the different methods on the ShanghaiTech Part A dataset. The reason that the proposed method shows a lower MAE value than other methods is that the density adaptive network in the proposed method adapts to the change in the density of the crowd and regulates the predicted count by $\alpha(X_i;\varphi) * C_i$.

Conclusion

In this paper, we proposed a new model for counting the crowd number by modifying the original CSRNet to adapt to images with different crowd density. We proposed two different types of models, where each model apply a density adaptive network into it. The proposed model demonstrates improved performance compared to the existing CSRNet. In future research, we aim to explore more varying modules to the proposed method to further enhance performance.

Acknowledgement

This research was supported by 「Fostering Project of R&D promotion complex in Busan」 funded by the Korea government(MSIT) and a local government(Busan)(2023-Development of R&D promotion complex-BUSAN-Segment1-Scaleup1)

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, Vol. 32, No. 9, pp.1627–1645, DOI:https://doi.org/2010.10.1109/TPAMI.2009.167

[2] H. Idrees, I. Saleemi, C. Seibert, M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," In *Proc. of International Conference on Computer Vision and Pattern Recognition,* pp. 2547–2554, June 25-26, 2013. DOI:https://doi.org/10.1109/CVPR.2013.329

[3] A. B. Chan, N. Vasconcelos, "Bayesian poisson regression for crowd counting," In *Proc. of International Conference on Computer Vision*, pp. 545–551, Sep. 29-Oct. 2, 2009. DOI:https://doi.org/10.1109/ICCV.2009.5459191

[4] V. Lempitsky, A. Zisserman, "Learning to count objects in images," In *Advances in Neural Information Processing Systems*, pp. 1324–1332, Dec. 6-11, 2010.

[5] V.Q. Pham, T. Kozakaya, O. Yamaguchi, R. Okada, "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation," In Proc. *International Conference on Computer Vision*, pp. 3253–3261, Dec. 11-18, 2015. DOI:https://doi.org/10.1109/ICCV.2015.372

[6] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 589-597, June 27-30, 2016. DOI:https://doi.org/10.1109/CVPR.2016.70

[7] V. A. Sindagi, V. M. Patel, "Generating high quality crowd density maps using contextual pyramid CNNs," in Proc. of *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1861–1870, Oct. 22-29, 2017. DOI:https://doi.org/10.1109/ICCV.2017.206

[8] D. B. Sam, S. Surya, R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. of Conference on Computer Vision and Pattern Recognition*, pp. 5744-5752, July 21-26, 2017. DOI:https://doi.org/10.1109/CVPR.2017.429

[9] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. of Conference on Computer Vision and Pattern Recognition*,, pp. 589–597, June 27-30, 2016. DOI:https://doi.org/10.1109/CVPR.2016.70

[10] L. Yuhong, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in Proc. of the IEEE conference on computer vision and pattern recognition, pp. 1091-1100, June 18-23, 2018. DOI:https://doi.org/10.1109/CVPR.2018.00120

[11] V. Ashish, et al. "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998-6008, Dec. 4-7, 2017.