Computers, Materials & Continua
DOI:10.32604/cmc.2020.013522
Article

Tech Science Press

A Deep-CNN Crowd Counting Model for Enforcing Social Distancing during COVID19 Pandemic: Application to Saudi Arabia's Public Places

Salma Kammoun Jarraya[1,2,*], Maha Hamdan Alotibi[1,3] and Manar Salamah Ali[1]

[1]Department of Computer Science, FCIT, King Abdulaziz University, Jeddah, Saudi Arabia
[2]MIRACL-Laboratory, Sfax University, Sfax, Tunisia
[3]Department of Computer Science, King Khalid University, Abha, Saudi Arabia
[*]Corresponding Author: Salma Kammoun Jarraya. Email: smohamad1@kau.edu.sa
Received: 10 August 2020; Accepted: 12 September 2020

Abstract: With the emergence of the COVID19 virus in late 2019 and the declaration that the virus is a worldwide pandemic, health organizations and governments have begun to implement severe health precautions to reduce the spread of the virus and preserve human lives. The enforcement of social distancing at work environments and public areas is one of these obligatory precautions. Crowd management is one of the effective measures for social distancing. By reducing the social contacts of individuals, the spread of the disease will be immensely reduced. In this paper, a model for crowd counting in public places of high and low densities is proposed. The model works under various scene conditions and with no prior knowledge. A Deep CNN model (DCNN) is built based on convolutional neural network (CNN) structure with small kernel size and two fronts. To increase the efficiency of the model, a convolutional neural network (CNN) as the front-end and a multi-column layer with Dilated Convolution as the back-end were chosen. Also, the proposed method accepts images of arbitrary sizes/scales as inputs from different cameras. To evaluate the proposed model, a dataset was created from images of Saudi people with traditional and non-traditional Saudi outfits. The model was also trained and tested on some existing datasets. Compared to current counting methods, the results show that the proposed model has significantly improved efficiency and reduced the error rate. We achieve the lowest MAE by 67%, 32% .and 15.63% and lowest MSE by around 47%, 15% and 8.1% than M-CNN, Cascaded-MTL, and CSRNet respectively.

Keywords: CNN; crowd counting; COVID19

1 Introduction

During pandemics such as coronavirus disease (COVID19), social distancing is applied as a preventive precaution to preserve human lives and reduce the severity of disease spread. Globally, 18.5 million confirmed cases and over 700 thousand deaths had been reported until today. Social distancing is an effective non-medical interference measure for preventing the transmission of diseases. Social distancing can be implemented through various rigorous measures such as banning travel, lockdown, and closing

public places. These aggressive measures have made a significant impact on economies around the world. Other less severe measures include warning people to keep a safe distance between individuals and controlling crowds in public places. Enabling technologies and artificial intelligence techniques can play a significant role in implementing and enforcing these measures [1].

Crowd counting is a practical approach for crowd control management. Public places like malls and supermarkets can be monitored seamlessly through surveillance cameras and crowd counting software. Whereby alerts are sent to organizers or visitors if a particular area has reached the maximum allowed capacity according to the social distancing related quota.

While a plethora of research has been conducted on crowd counting technologies, limited research has been undertaken on crowds with unique and uncommon characteristics, such as Saudi people in Saudi public places. In Saudi public places, crowd counting comes with several issues, such as non-uniform illumination, extreme clutter, occlusions, non-uniform spreading of people, perspective, intra-scene\inter-scene differences in appearance, and scale. These characteristics make optimization and unification extremely challenging. The wide range of crowd analysis applications, together with the complexity of the problem, have been guiding researchers in recent years to improve the efficiency of these techniques and come with new and effective solutions [2].

In this paper, we propose a novel method for crowd counting in high and low crowded public places under various scene conditions without prior knowledge. The proposed method is based on CNN model to count people/visitors who appear in video frames in public places. The model accepts arbitrary image sizes and scales as inputs from a diversity of surveillance camera types. The cameras are connected to an IoT architecture to provide pictures from different public places.

The rest of the paper is organized as follows: an overview of crowd counting and density map generation is discussed in Section 2. The proposed model is presented in Section 3. Section 4 discusses the experimental results on several datasets. And finally the conclusions are discussed in Section 5.

2 Background

Crowd counting methods can be grouped into four main classes: Detection-based approaches [3–5], regression-based approaches [6–8], density estimation-based approaches [6,9–11], and CNN-based approaches [12–14].

Studies showed that, regardless of the pros and cons of other methods, CNN-based regressors have considerably outperformed the outdated crowd counting approaches. The representations from the local features of non CNN-based approaches are not sufficient to satisfy a high level of performance and accuracy.

The prosperity of CNNs in most computer vision challenging problems has triggered researchers to deeply investigate and study the nonlinear functions from images of crowds to the identical counts or identical density maps. Most CNN crowd counting techniques still require an input image with a fixed size. This requirement is "artificial" and will probably decrease the image understanding of precision. However, when CNNs have been initially proposed, most related studies and research depended on image patches and required the use of fixed-size images [15,16]. However, since the quality of the density maps is poor, new approaches need to enhance the density maps. Recently, a multi-column style technique has been published, which provides better quality density maps and improved performance [17,18]. However, when the CNN complexity is increased, the model suffers from a non-effective branch structure and prolonged training time.

Crowd counting methods have two main challenges. First, ROI (region-of-interest counting), where the chosen region of study may affect the accuracy and performance of calculation. The second category is LOI (line-of-interest counting), which calculates the number of people crossing a chosen line. Since the proposed model in this study will be used in public places, the region-of-interest counting is adopted, based on a single image to count both high and low crowded places.

3 Related Work

In recent years, crowd analysis techniques have gained significant interest due to the cutting-edge achievements of Convolutional Neural Network (CNN) models. CNN's capabilities are used for investigating the non-linear functions from crowd images to their corresponding density maps.

The majority of CNN-based crowd counting models use fixed-size input images [16,19–20], where the models suffer from low-quality density maps due to the structure of their networks. To solve the low-quality density maps challange, multi-column architectures have been introduced [17,21]. However, when deeper networks are applied, new challenging issues arise, which are: non-effective branch structure and an extended period of time required for training.

CNN's network property approaches are classified into three categories [2]. Early deep learning methods for assessing crowd density and counts used basic CNN models with basic CNN network layers [14,19]. Scale-aware models used advanced CNN models, which are robust to varying scales, such as multi-column architectures [22]. Finally, context-aware models integrated images of regional and global contextual information into CNN framework to reduce estimation errors [22].

A multi-column architecture (MCNN) for images with random crowd densities and various perspectives was proposed in Zhang et al. [18]. To ensure the robustness of the variation in object scales, representation of varied object scales in images has been supported by the construction of large, medium, and small-sized networks.

Training regressors with a multi-column network on every input spot has been proposed as a crowd counting method in Zhang et al. [18]. Babu Sam et al. [17] argued that training specific collection of spots in images with varied crowd densities would significantly improve the performance of the model. Babu Sam et al. [17] proposed a switching CNN that stimulates the multi-column network by using multiple independent regressors with switching classifiers and sensory domains. The proposed model chooses an optimal regressor for a particular input spot.

A Contextual Pyramid model (CP-CNN) was proposed in Sindagi et al. [23]. The model explicitly joins local and global contextual material of crowd images to improve the quality of crowd estimation of crowd densities. The model consists of Local Context Estimator (LCE), Global Context Estimator (GCE), a Fusion-CNN (F-CNN), and Density Map Estimator (DME). However, the proposed model was complex and required long training time.

The CSNet model in Li et al. [24] applied improvement to the quality of density maps with deeper network of single-column structures and introduced a dependency on VGG-16 [25]. The first ten layers of VGG-16 were used without applying a fully connected layer. Dilated convolutional layers have been used as the back-end to extract deeper information of salience without risking the resolution of the output.

4 System Model

The production of the new real-time crowd counting model in this work has gone through several stages; with several steps in each stage. In this section, we discuss the offline work stage that is used to generate a counting model based on deep convolutional neural networks (CNN). The general structure of the offline work consists of the following three steps:

(1) Data acquisition and collection.

(2) Generate a full convolution deep CNN that can deal with the processed data. This is done by following the structure of the first 11 layers of VGG-19 small filter size and changing the last convolution layers with dilation convolution and multi-column for back end.

(3) Evaluate the model with regard to different conditions, including different dilation rates, different crowd levels, and various image environments. The model is evaluated using the most popular crowd counting dataset Shanghai-tech parts A and B, as well as the Saudi dataset. The results are compared with the existing models.

4.1 Data Collection and Preparation

As the first step, we record a new and particular large-scale crowd counting dataset to build a large number of training data required by CNN, and to include the unique and uncommon characteristics of Saudi Arabia public places. This dataset was prepared explicitly for Saudi people in public places. It consists of 673 frames with different head numbers and different crowd levels. The counts range from 1 to 450, with an average of 80 people in view. The pictures were taken from different camera angles, which allowed our system to recognize persons wherever the camera was. The frames are recorded from different videos and from different places, such as malls, restaurants, events, walkways, and airports. In addition, a set of images were gathered from openly available websites and social media, such as Google and Instagram. These images cover different events, including concerts and stadiums as shown in Fig. 1. In Tab. 1, we show more details about the collected dataset. After collecting the videos, we extracted the frames which have a different number of people and have a different distribution. Then we started the process of selecting frames by labeling each person in the image and generating the ground truth file. We label the heads in high crowd images and label the whole body in the less crowded image.

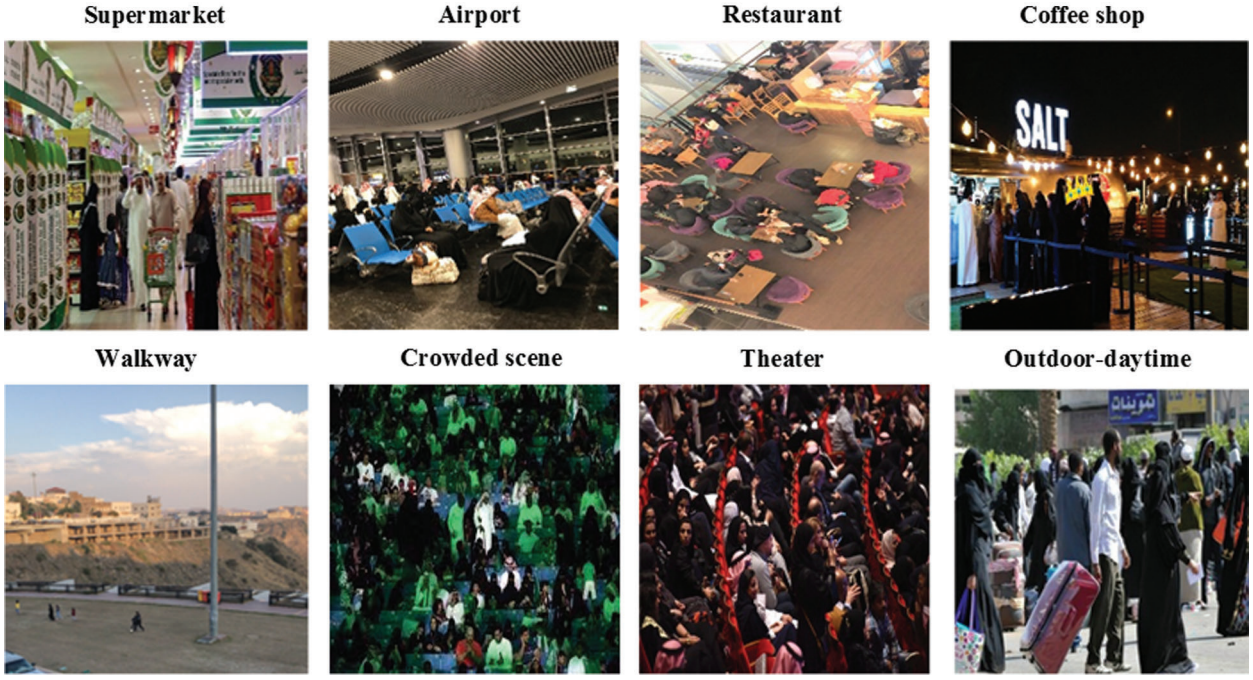

Figure 1: Sample images from the Saudi dataset

In this work, the method of generating density maps used in Zhang et al. [18] and Li et al. [24] is applied. Highly congested crowd scenes are attempted by using the geometry-adaptive kernels. Each head annotation is blurred using a Gaussian kernel normalized to1. The ground truth is generated using the spatial distribution of all the images in each dataset. The geometry-adaptive kernel is defined as follows:

Table 1: Saudi dataset description

Videos	Frames	Locations	Indoor/ Outdoor	Challenges			
				Non-uniform distribution	Non-uniform illumination	Occlusions	High clutter
#1	80	Mall	Indoor		✓	✓	
#2	60	Mall				✓	
#3	70	Mall			✓	✓	✓
#4	40	Mall		✓	✓	✓	✓
#5	50	Mall		✓		✓	
#6	35	Super-market					✓
#7	50	Airport		✓		✓	
#8	70	Coffee shops					✓
#9	75	Restaurants		✓			✓
#10	30	Walkway-daytime	outdoor	✓			
#11	40	Restaurants-night				✓	
From web	70	Daytime and night time		✓	✓	✓	✓

$$F(x) = \sum_{i}^{n} \delta(x - xi) * G\sigma i\,(x);\ with\ \sigma i\ =\ \beta \overline{di} \tag{1}$$

For each targeted object *xi* in the ground truth δ, the average distance of *k* nearest neighbors is indicated by $\overline{di}$. To generate the density map, $\delta(x - xi)$ is convolved with a Gaussian kernel and a parameter σi (standard deviation), where *x* is the position of pixel in the image. In our experimentations, the configuration in Sindagi et al. [12] where β = 0:3 and k = 3 is followed. To blur all the annotations in sparse crowd images, the Gaussian kernel to the average head size is adapted.

4.2 Crowd Counting Model Generation Based on CNN

Our main contribution in this work is to generate accurate people counting model from an arbitrary single image, with any random crowd density and random camera perspective. However, this seems to be a rather challenging task, considering that in the dataset there exists a significant inequality scale of the bodies in different images, which requires the employment of features at different measures and counting people in various images.

The density of the crowd, as well as its distribution, are very significant in the selected datasets. Typically, there is substantial occlusion for most bodies in images. Hence, traditional methods such as detection-based methods do not perform well in such settings. Since there might be a notable difference in the scale of the object such as people or heads in the images, we need to use features at various scales collectively to estimate the crowd counts correctly in different images.

To overcome the above mentioned challenges, we proposed a novel framework based on deep-convolutional neural network (CNN) and density map for crowd counting in an unfixed-resolution single image. The basic idea of the proposed network design is to deploy a deeper CNN for catching high-level features with larger receptive fields and produce high-quality density maps without brutally increasing network complexity.

To count the number of people in an input image via CNNs, we generate density maps of the crowds (from the input images) to estimate how many people exist per square meter. The rationale behind using density maps is that the method preserves more information compared to the total head number of the crowd. Also, the density map will provide the spatial distribution pattern of the crowd in the given image. In addition, in training the density map through a CNN, the learned filters will be more adaptive to heads of various sizes. Therefore it will be more suitable for arbitrary inputs whose viewpoint effect varies notably. However, the filters are more semantic meaningful, and as a result, they improve the exactitude of counting.

We choose the structure of the first 11 layers of VGG-19 [25] as the front-end of deep-CNN because of its powerful transfer learning capability. And it has an adaptable design for smoothly concatenating the back-end, which is made for high-quality density map production.

The original VGG-19 is built using a single column that consists of Convolutions layers (used only 3 × 3 size), maximum pooling layers (used only 2 × 2 size), and fully connected layers, resulting in 19 layers.

However, the lack of modifications results in poor performance. In CrowdNet [26], the authors directly carve the first 13 layers from VGG-16 and add a (1 × 1) convolutional layer as an output layer instead of a fully connected layer. As observed from the literature, some architectures use VGG-16, such as MCNN [18], which uses VGG-16 as a classifier of the density level for labeling input pictures before forwarding them to the most suitable column of the multi-column network.

While in CP-CNN [4], the VGG-16 acts as an ancillary without boosting the utmost accuracy since it combines the effect of classification with the features from the density map generator column. In the proposed model, we first remove the fully-connected layers of VGG-19 and consider it as the classification part. The DCNN is a fully convolutional network; different image sizes can be used both in prediction and in training modes. Then the proposed deep-CNN is built with convolutional layers and three pooling layers in VGG-19.

The front-end network output is 1/8 the size of the original input. Continuation of stacking the basic components in VGG-19 (more convolutional layers and pooling layers), will lead to further downsizing of the output and limits the production of high-quality density maps.

Alternatively, we deploy dilated convolutional layers inspired from Yu et al. [27], as the back-end of our deep CNN, to maintain the output resolution and extract deeper information of saliency.

One of the key aspects and significant components of our back-end deep-CNN is the 2-D dilated convolutional layer which can be described as follow:

$$f(n,m) = \sum_{j=1}^{n} \sum_{i=1}^{m} \mathrm{y(n\ +\ D\ -i;\ m+\ D\ -j)w(i;\ j)} \tag{2}$$

where f(n; m) refers to the production of dilated convolution from image y(n,m) and a filter w(i; j) with n length and m width, while D refers to the dilation rate. However, in a normal convolution, the dilation rate r = 1. Dilated convolutional layers are an excellent alternative of pooling layer and have shown notable enhancement of accuracy in segmentation [28–31].

4.2.1 Network Configuration

According to the improvement applied on the original VGG-19, we suggest three network configurations of deep-CNN. The back-end in Fig. 2 has a similar front-end organization but with different dilation rates in the back-end.

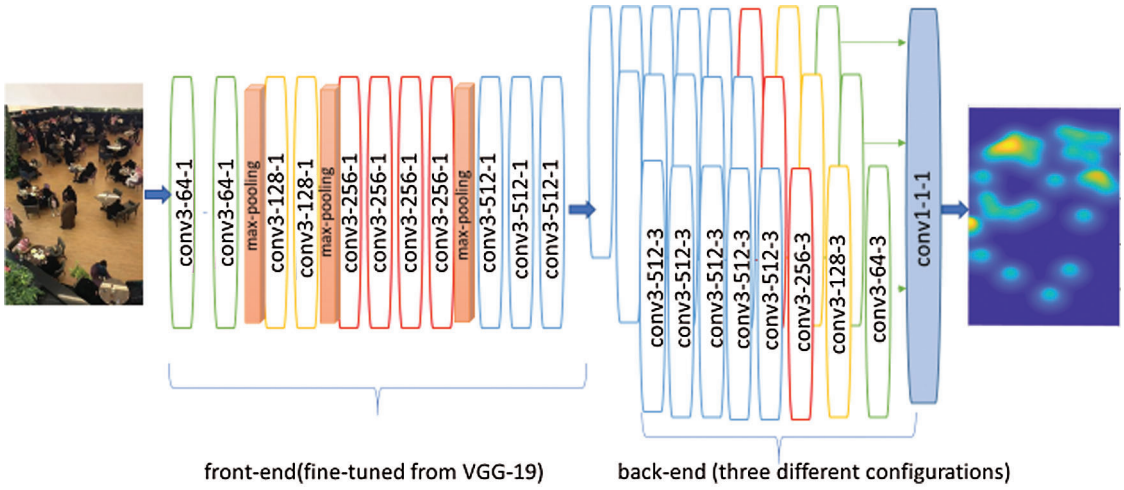

Figure 2: The structure of the proposed counting network DNCC

There are two primary parts in the proposed CNN model. The first part, which is the front-end, comprises the adapted VGG-19 CNN (except fully-connected layers) [32,33] with 3 × 3 kernels for 2D feature extraction. The second part is the back-end containing dilated CNN (DCNN) with more layers, where dilated kernels are utilized for delivering larger reception fields and replacing pooling operations. It is recommended to use DCNN with tree network configurations that have many dilation rates at the back-end, but with using the same front-end arrangement. According to Zhang et al. [22], it is more efficient to utilize smaller kernels with a higher number of conventional layers rather than bigger kernels with a lower number of layers for receptive fields of the same size. The primary consideration is to balance the need for accuracy against the resources involved, such as the number of parameters, the training time, and memory consumption.

Based on our intensive experiments, it was observed that the optimal arrangement involves the use of the first eleven layers of VGG-19 with three rather than five pooling layers so that the adverse impacts of pooling operations on output accuracy could be reduced [22]. The same front-end structure has been maintained while the dataset has been trained starting from the eighth layer until the end of the network. Padding has been employed to keep all convolutional layers at the prior size. The parameters of the convolutional layers have been represented by (conv-kernel-size), (dilation-rate), and (number of filters) where the max-pooling layers have been performed on a 2 × 2-pixel window with stride 2. Fig. 2 shows a visual overview of the proposed counting network DNCC.

4.2.2 Training Details

A direct approach is used for training the DCNN as an end-to-end structure. For the first 11 convolutional layers, initially from fine-tuned, well-trained VGG-19 [25]. A 5-fold cross-validation is performed following the standard setting in Li et al. [24]. The initial values of the remaining layers are derived from a Gaussian initialization with a standard deviation of 0.01. During the training session, stochastic gradient descent (SGD) is used at a constant learning rate of 1e-6. Consistent with [3,19,22],

the Euclidean distance is used for calculating the difference between the ground truth and the estimated generated density map. The equation for the loss function is given below:

$$L(\Theta)\frac{1}{2N}\sum_{i=1}^{N}||\mathrm{Z(Xi;\ \Theta)}\ -\ \mathrm{Z}_i^{GT}||_2^2 \tag{3}$$

where N is the training batch size while $\mathrm{Z(Xi;\ \Theta)}$ represents the output generated by DCNN with a parameter of Θ. Meanwhile, Xi is the input image with Z_i^{GT} as the ground truth result of the input image Xi.

Subsequently, in the data augmentation step (see Fig. 3), the images have been cropped at various places producing nine patches at a quarter of the initial image size. The patches of the first four encompass four quarters of the input image that do not overlap, and the remaining five patches are cropped indiscriminately from the image. Next, the patches are reflected, thus doubling the training set.

Figure 3: Sample from data augmentation step

4.3 Model Evaluation

The evaluation of the proposed model is based Mean Squared Error (MSE) and Mean Absolute Error (MAE) which are defined by Eqs. (4) and (5).

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|G_i - G_i^{GT}\right|^2} \tag{4}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|G_i - G_i^{GT}\right| \tag{5}$$

where N is the number of images in one test sequence and G_i^{GT} is the ground truth of counting. C_i represents the estimated count which is defined as follows:

$$C_i = \sum_{l=1}^{L}\sum_{w=1}^{W}\mathrm{zl;w} \tag{6}$$

L and W show the length and width of the density map, respectively, while zl;w is the pixel at (l;w) of the generated density map. C_i is the estimated counting number for image Xi. Roughly, MAE indicates the accuracy of the estimates, and MSE indicates the robustness of the estimates.

In the next subsections, we present the results of the proposed model on the Saudi dataset and the most popular and challenging crowd dataset, shanghai_tech A&B.

4.3.1 Results on Saudi Dataset

The dataset consists of 673 images customized and personalized for Saudi people in public places, and with different head numbers and different crowding levels. We use 450 images for training and 217 for testing. The evaluation of the dataset is as follows:

Before choosing the final structure of DCNN, different arrangements for the front end and the back end of the improved VGG has been tested. Tab. 2 shows the evaluation of the different structures. The most critical part depends on the tradeoff between accuracy and the resource overhead (including training time, memory consumption, and the number of parameters). The result of the experiment shows the best tradeoff can be achieved when keeping the first eleven layers of VGG-19 with only three pooling layers instead of five to suppress the detrimental effects on output accuracy caused by the pooling operation. We also test the effect of dilation on our model by using different dilation rate from 1 to 4, and we found that the best dilation rate is 2 on VGG-19 case F.

Table 2: Evaluation of different DCNN structures

Case	Layers	Front-end layers	Back-end Layers	MAE	MSE
A	16	10	6	14.7	26
B	16	11	5	16.03	27.8
C	16	9	7	14.5	25.8
D	19	10	9	13.9	25.7
E	19	9	10	14.2	25.78
F	19	11	8	13.5	25.6

Discussion: We evaluate our dataset based on the different category on the DCNN network; the result was as shown in Tab. 2. Five random frames are selected from each category, and we found that outdoor night images get the highest error rate by 1.5 MSE. The indoor low crowded achieve the lowest error rate of 0.3. However, this was due to several reasons. First, we have more frames from the indoor category compared to the outdoor category during the network training. Also, the outdoor frames are more challenging than indoor frames. In Tab. 3, we present the experimental result and comparisons of three states of art open source models and our DCNN on Saudi dataset. We found that our proposed model achieves the lowest MAE by 67%, 32% and 15.63% and lowest MSE by around 47%, 15% and 8.1% than M-CNN, Cascaded-MTL, and CSRNet respectively. Due to the uncommon and special nature of Saudi society, the state of art models may require special features to be added to the global features. In fact, we have trained our data for crowd counting by accepting the unfixed-resolution color of the images. Thus M-CNN training code has been used as a base, where the findings further support the idea of non-effective branch structure and Cascaded-MTL that have a complicated structure. Fig. 4 presents the produced density map of our method. The 1st column displays testing samples of the set in the Saudi Dataset. While the 2nd column displays the ground truth for samples and the 3rd column shows the produced density map.

Table 3: Estimation errors on the Saudi dataset

Year	Method	MAE	MSE
2016	M-CNN [18]	43.2	49
2017	Cascaded-MTL [12]	20.9	30
2018	CSRNet [24]	18	28.2
2020	DCNN	13.5	25.6

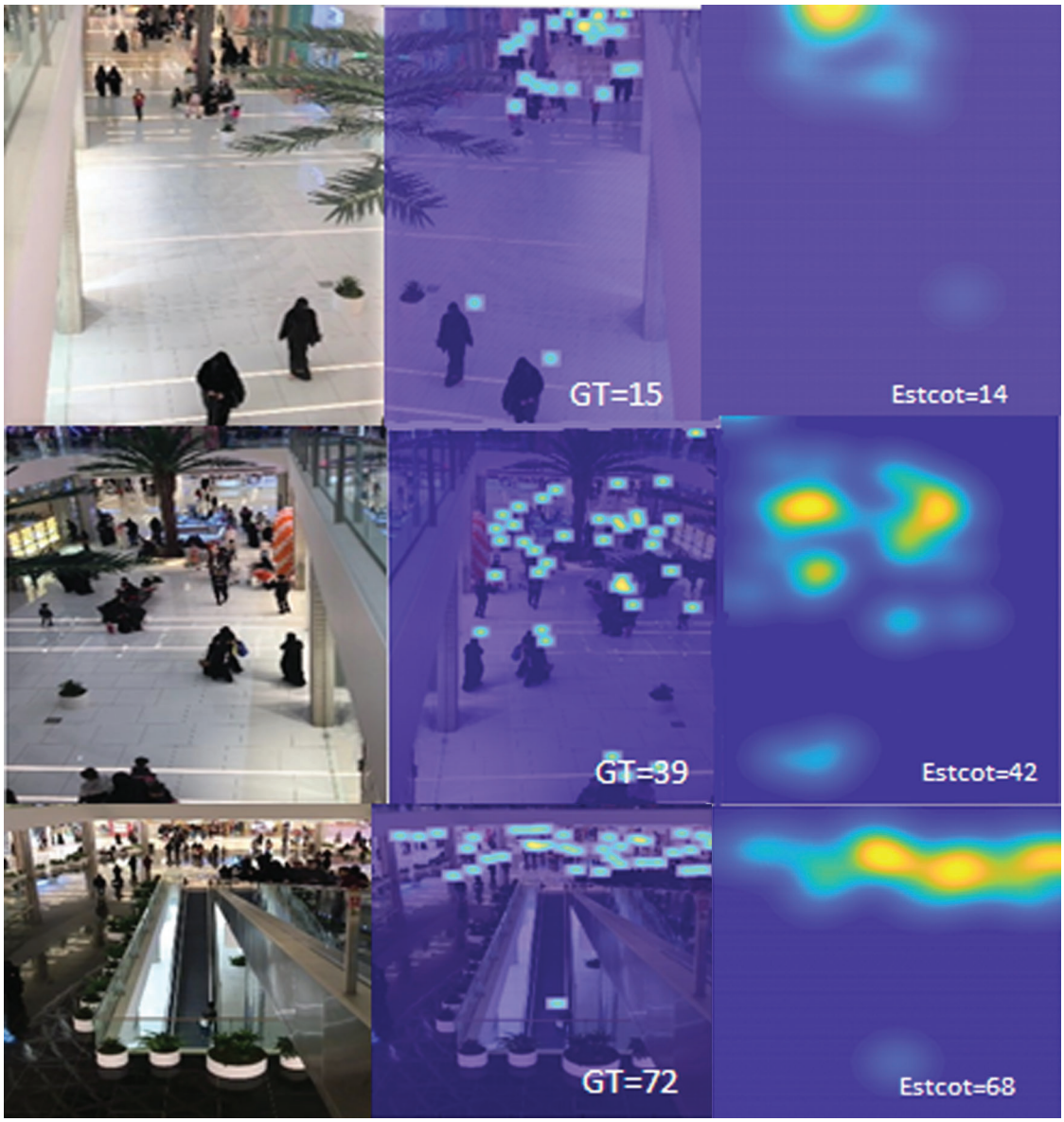

Figure 4: Density maps

4.3.2 Results on Shanghaies Dataset

The Shanghai Tech dataset consists of a total number of 330,165 persons within 1198 annotated images [23]. This public dataset is based on two parts, A and B, as shown in Tabs. 4 and 5, respectively.

Table 4: Approximation errors on the Shanghai Tech dataset part A

Year	Method	Shanghai Tech dataset	
		Part A	
		MAE	MSE
2015	Cross-scene [3]	181.8	277.7
2016	M-CNN [18]	110.2	173.2
2017	FCN [10]	126.5	173.5
2017	Cascaded-MTL [12]	101.3	152.4
2017	Switching-CNN [17]	90.4	135.0
2018	CSRNet [24]	68.2	115
2020	DCNN	66.3	98.4

Table 5: Approximation errors on the Shanghai Tech dataset part B

Year	Method	Shanghai Tech dataset	
		Part B	
		MAE	MAE
2015	Cross-scene [3]	32.0	32.0
2016	M-CNN [18]	26.4	26.4
2017	FCN [10]	23.76	23.76
2017	Cascaded-MTL [12]	20.0	20.0
2017	Switching-CNN [17]	21.6	21.6
2018	CSRNet [24]	10.6	10.6
2020	DCNN	9.8	11.1

Discussion: Our DCNN model has been evaluated and compared to seven other existing related works. The results indicate that our model has achieved the following:

- 63% lower MAE in Part A compared to the Cross-scene.
- 2.5 % Lower MAE than the recent existing work of CSRNet.
- Lowest MAE (the highest accuracy) in Part A compared to other models.
- 54% lower MAE has been achieved in part B.

The results indicate that our method has outperformed not only the counting tasks for extremely dense crowds but also tasks for relatively sparse scenes.

5 Conclusion

Social distancing has proved to be a significant measure in reducing infection rates in COVID19 pandemic. Artificial intelligence technology plays a vital role in encouraging or even enforcing social distancing practice. One of the practices in social distancing is controlling crowded gatherings in public places. For this reason, there is a need for systems with high accuracy and remarkable performance to detect dense crowds.

In this paper, we provide accurate people counting model from an arbitrary single image, with any random crowd density and random camera perspective. A novel deep CNN called DCNN, had been proposed. The proposed model follows the structure of VGG-19 small conve size, that takes any image resolution as an input. Also, a new large-scale crowd counting dataset for the Saudi public area has been created and used to train the model. The results indicate that our proposed model achieves a lower MAE combed to state-of-art methods. More specifically, we achieve the lowest MAE by 67%, 32% and 15.63% and lowest MSE by around 47%, 15% and 8.1% than M-CNN, Cascaded-MTL, and CSRNet respectively.

In this paper, we provide accurate people counting model from an arbitrary single image, with any random crowd density and random camera perspective. A novel deep CNN called DCNN, had been proposed. The proposed model follows the structure of VGG-19 small conve size, that takes any image resolution as an input. Also, a new large-scale crowd counting dataset for the Saudi public area has been created and used to train the model. The results indicate that our proposed model achieves a lower MAE combed to state-of-art methods. More specifically, we achieve the lowest MAE by 67%, 32% and 15.63% and lowest MSE by around 47%, 15% and 8.1% than M-CNN, Cascaded-MTL, and CSRNet respectively.

Funding Statement: This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, Saudi Arabia, under grant No. (DF-352-165-1441). The authors, therefore, gratefully acknowledge DSR for their technical and financial support.

Conflict of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Trivedi, C. Zakaria, R. Balan and P. Shenoy, "WiFiTrace: Network-based contact tracing for infectious diseases using passive WiFi sensing," *arXiv preprint arXiv: 2005.12045*, 2020.

[2] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, 2018.

[3] C. Zhang, H. Li, X. Wang and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 833–841, 2015.

[4] P. Viola, M. J. Jones and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[5] P. Dollar, C. Wojek, B. Schiele and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[6] V. Lempitsky and A. Zisserman, "Learning to count objects in images," *Advances in Neural Information Processing Systems*, vol. 1, pp. 1324–1332, 2010.

[7] D. Ryan, S. Denman, C. Fookes and S. Sridharan, "Crowd counting using multiple local features," in *2009 Digital Image Computing: Techniques and Applications*, IEEE, Melbourne, VIC, Australia, pp. 81–88, 2009.

[8] A. B. Chan and N. Vasconcelos, "Bayesian Poisson regression for crowd counting," in *Proc. of 12th Int. Conf. on Computer Vision*, *IEEE*, Kyoto, Japan, pp. 545–551, 2009.

[9] B. Xu and G. Qiu, "Crowd density estimation based on rich features and random projection forest," in *2016 IEEE Winter Conf. on Applications of Computer Vision*, Lake Placid, NY, USA, pp. 1–8, 2016.

[10] M. Marsden, K. McGuinness, S. Little and N. E. O'Connor, "Fully convolutional crowd counting on highly congested scenes," in *Int. Conf. on Computer Vision Theory and Applications*, Porto, Portugal, pp. 27–33, 2017.

[11] V. Q. Pham, T. Kozakaya, O. Yamaguchi and R. Okada, "COUNT forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 3253–3261, 2015.

[12] V. A. Sindagi and V. M. Patel, "CNN-Based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *14th IEEE Int. Conf. Adv. Video Signal Based Surveillance*, Lecce, Italy, 2017.

[13] S. Kumagai, K. Hotta and T. Kurita, "Mixture of counting CNNs: Adaptive integration of CNNs specialized to specific appearance for crowd counting," *arXiv preprint arXiv:1703.09393*, 2017.

[14] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5012–5024, 2016.

[15] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu *et al.,* "Crowd counting via adversarial cross-scale consistency pursuit," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 5245–5254, 2018.

[16] C. Wang, H. Zhang, L. Yang, S. Liu and X. Cao, "Deep people counting in extremely dense crowds," in *Proc. of the. 23rd ACM Int. Conf. on Multimedia*, Brisbane, Australia, pp. 1299–1302, 2015.

[17] D. Babu Sam, S. Surya and R. Venkatesh Babu, "Switching convolutional neural network for crowd counting," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 5744–5752, 2017.

[18] Y. Zhang, D. Zhou, S. Chen, S. Gao and Y. Ma, "Single-image crowd counting via multi-column Convolutional Neural Network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 589–597, 2016.

[19] E. Walach and L. Wolf, "Learning to count with CNN boosting," *European Conf. on Computer Vision*, Amsterdam, The Netherlands, pp. 660–676, 2016.

[20] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye *et al.,* "Fast crowd density estimation with convolutional neural networks," *Engineering Applications of Artificial Intelligence*, vol. 43, pp. 81–88, 2015.

[21] L. Zeng, X. Xu, B. Cai, S. Qiu and T. Zhang, "Multi-scale convolutional neural networks for crowd counting," in *IEEE Int. Conf. on Image Processing*, Beijing, China, pp. 465–469, 2018.

[22] L. Zhang, M. Shi and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *IEEE Winter Conf. on Applications of Computer Vision*, Lake Tahoe, NV, USA, pp. 1113–1121, 2018.

[23] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 1879–1888, 2017.

[24] Y. Li, X. Zhang and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 1091–1100, 2018.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] L. Boominathan, S. S. Kruthiventi and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proc. of the 24th ACM Int. Conf. on Multimedia*, Athens, Greece, pp. 640–644, 2016.

[27] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[28] P. Isola, J. Y. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1125–1134, 2017.

[29] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[30] L. C. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv*:1706.05587, 2017.

[31] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv*:1511.07122, 2015.

[32] M. Li, Z. Zhang, K. Huang and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in *19th Int. Conf. on Pattern Recognition*, Tampa, FL, USA, pp. 1–4, 2008.

[33] C. Chung, S. Patel, R. Lee, L. Fu, S. Reilly *et al.,* "Implementation of an integrated computerized prescriber order-entry system for chemotherapy in a multisite safety-net health system," *Bulletin of the American Society of Hospital Pharmacists*, vol. 75, no. 6, pp. 398–406, 2018.