IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 6, 2022, accepted May 2, 2022, date of publication June 8, 2022, date of current version June 20, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3180738

Single Convolutional Neural Network With Three Layers Model for Crowd Density Estimation

ADAL ALASHBAN, (Student Member, IEEE), ALHANOUF ALSADAN, NORAH F. ALHUSSAINAN, AND RIDHA OUNI

Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: Adal Alashban (439204375@student.ksu.edu.sa)

This work was supported by the Deanship of Scientific Research, King Saud University, KSA.

ABSTRACT Crowd density estimation is an important topic in computer vision due to its widespread applications in surveillance, urban planning, and intelligence gathering. Resulting from extensive analysis, crowd density estimation reflects many aspects such as similarity of appearance between people, background components, and inter-blocking in intense crowds. In this paper, we are interested to apply machine learning for crowd management in order to monitor populated area and prevent congestion situations. We propose a Single-Convolutional Neural Network with Three Layers (S-CNN3) model to count the number of people in a scene and conclude about the crowd estimation. Then, a comparative study for density counting establishes the performance of the proposed model against the convolutional neural networks with four layers (single-CNN4) and Switched Convolutional neural networks (SCNN). ShanghaiTech dataset, considered as the largest data base for crowd counting, is used in this work. The proposed model proves high effectiveness and efficiency for crowd density estimation with 99.88% of average test accuracy and 0.02 of average validation loss. These results achieve better performance than the existing state-of-the-art models.

INDEX TERMS Crowd counting, density estimation, GPU, switching convolutional neural network (SCNN).

I. INTRODUCTION

More people have chosen to live in the city in recent years, where the advantages of this phenomenon are enriching cultural life and at the same time making good use of accessible urban infrastructure, attracting a wide range of people to coordinate various activities. Global and national events attract large crowds, whether indoors or outdoors. Typically, these events involve at least one activity necessitating simultaneous participation of attendees, such as viewing a display, watching an open-air show, passing through checkpoints, or entering areas that restrict the number of people present. However, such events are prone to overcrowding, where there were no rapid and efficient way to obtain an overview of the whole venue and effectively communicate with those in other areas, often resulting in collisions and overcrowding. In contrast, a centralized monitoring system capable of estimating crowd density in various locations at the same time constitutes a preferable option for effective and reliable decision making to ensure the people safety while allowing them to continue enjoying the event.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

In the Kingdom of Saudi Arabia (KSA), there are several seasons for Hajj, Umrah, and Ramadhan (especially the 27th of Ramadan), when people from different countries of the world overpopulate Mecca to perform religious ceremonies in a limited period of the year. In various countries and during the celebration of their founding, national or independence days, the security authorities are mobilized in different places to prevent negative impact of crowding. In this context, the development of an Artificial Intelligence (AI)-based method helps to count the crowds present in different places, provides the ability to control them from accidents and prevent the spread of contagious disease such as Covid-19. Overcrowding problems, wherever they are, can be counted, monitored and controlled using centralized and automated system integrating crowd density estimation and machine learning.

The research on automatic detection, counting and density estimation in a large-scale crowd is playing a significant security and management role for controlling huge crowd in different places. Monitoring large events including huge crowd requires great efforts to ensure the attendees' safety and the delivery of adequate services. Many studies have been conducted to come up with a good monitoring system.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Estimating crowd density will support management authorities in coordinating and scheduling them between different areas during their transitions. With the recent advances in machine learning, various CNN models have been proposed to better resolve some estimation issues and leverage both crowd density estimation and classification accuracy [1]. However, those models used crowd size as a discriminator for crowd density.

Deep Learning (DL), being a specialized form of machine learning, can handle a large amount of data. It is considered as the most technical in machine learning that gives high efficiency in many applications like face recognition and crowd counting. There is a subset of machine learning methods where multilayered neural networks can learn from vast amount of data. Deep learning, such as CNN, LSTM, RNN, GAN, RBFN, MLP, SOM, DBN, RBM, and Autoencoders, supports this mission due to its architecture including more than three layers.

Many studies have discussed ways to automate the monitoring and organizational processes to obtain a macro-scale view of the entire event and conduct central decisions. Several technological solutions have been suggested, some of them rely purely on sensors, while others employ image processing and computer vision. However, the most recent solutions focus on neural network approaches. In this work, we are interested in developing a Convolutional Neural Network (CNN) model with three layers for crowd density estimation. We switched the density estimation challenge from a regression problem to a classification problem by defining specific classes estimating various density levels and providing different crowd indicators. So, this model is able to classify a scene into a crowd density class, which allows enabling appropriate and timely actions to mitigate congestion and reduce the risk of hazardous incidents. For credible evaluation, we employed the most known ShanghaiTech dataset for performance measurements.

This paper is organized as follows. Section II introduces a background in terms of crowd density estimation and convolutional neural network. Section III reviews the relevant state of the art studies covering convolutional neural networks for crowd counting. Section IV presents a detailed description of our proposed CNN model. Section V describes the simulation environment and the experimentation details. Section VI illustrates a comparison of results between the proposed model and some existing models. Section VII shows the work contributions. Finally, Section VIII concludes the paper and outlines some future works.

II. BACKGROUND

Event organizers monitor crowds with the help of volunteers who are distributed around the venue to direct and guide the attendees. However, while the decisions on when and where to direct the crowd are all individual decisions, without having a global and updated overview of the event. Random and unsupervised crowd movement can cause congestion, overcrowding, and in more serious cases injuries due to stampedes. However, each individual responsible for the event only monitors and organizes the movement of one small group in one area. That's why many studies have investigated crowd density classification and estimation as elements of crowd management solutions. Deep learning techniques have been used for crowd density counting, which is a field of Artificial Intelligence (AI) using algorithms to provide computers with the ability to identify patterns from mass data to make predictions. This learning method allows computers to perform specific tasks autonomously. Some of the crowd density estimation solutions are based on different technologies: sensor technology approaches, computer vision approaches, and neural network approaches.

A. GENERAL PROCEDURE FOR CROWD DENSITY ESTIMATION

Crowd counting consists of crowd density estimation or counting the number of persons in a certain scene (image). Previously, the crowd was manually counted by crowd scientist, responsible of tallying the number of people in certain areas of an image and then extrapolates them for estimation. This method is characterized by waste of time and effort as well as of high error possibility. Because they attract and gather a huge number of people in a confined place, crowds are common in sports, festivals, political, and religious activities [2].

The Jacobs' Method is the most widely used method for counting crowds during religious ceremonial meetings, protests, and rallies. Jacobs' approach is dividing a crowd's area into pieces, calculating the average number of people in each sector, and multiplying by the total number of sections filled [3]. Crowd density estimation helps in the development of management techniques such as the design of safe public spaces and an emergency evacuation plan. The procedure for crowd density estimation is to establish relationships between image parameters from various image processing techniques and actual crowd densities at an investigation site [4]. There are two main different approaches: direct and indirect approaches. The direct approach tracks and counts people simultaneously, as long as people are correctly segmented. The indirect approach relates between a set of measurement features and learning algorithms of the whole crowd to carry out counting and estimating process by pixel-based analysis, texture-based analysis, and corner point-based analysis [5].

The most gigantic and famous crowd in the world occurs in the sacred sites of Mecca. It is considered as the world's largest human gathering as the Muslim holy pilgrimage, attracts millions of humans to Mecca every year for AlHajj. According to statists, the pilgrimage has attracted almost 2.5 million pilgrims in 2019, and over 3 million pilgrims at its height in 2012 [6]. CNN can be one of the best solution for managing Al Hajj crowd, which provides a framework for solving the problem of estimating the level of crowding in order to avoid accidents [7].

B. CONVOLUTIONAL NEURAL NETWORK (CNN)

A Convolutional Neural Network (ConvNets or CNN) is a type of artificial neural network [8] that uses image pixels as input to perform tasks such as image identification, classification, object detection, and face recognition. The process of CNN image classifications is taking image pixels as input and outputting a class such as: ''cat,'' ''dog'' and ''bird,'' and it can count the number of objects in the image.

The CNN consists of an input layer, multiple hidden layers and an output layer. The hidden layers usually consist of a series of convolutional layers, ReLU layers, pooling layers and a fully connected layer. The convolutional layer has a set of filters to filter parameters which need to be learned. The ReLU layers are used to improve neural networks by speeding up the training process. In the neural network, the pooling layer is utilized to minimize the amount of parameters and processing. The Fully Connected (FC) layer makes the final classification decision [9].

Convolution and pooling are two essential procedures that are always included in CNN. The convolution process with several filters is capable of extracting features (feature map) from the data set while preserving their spatial information. Pooling, also known as subsampling, is a technique for reducing the dimensionality of feature maps created by the convolution procedure. The most frequent pooling techniques in CNN are maximum pooling and average pooling [10]. Figure 1 describes the CNN's structure.

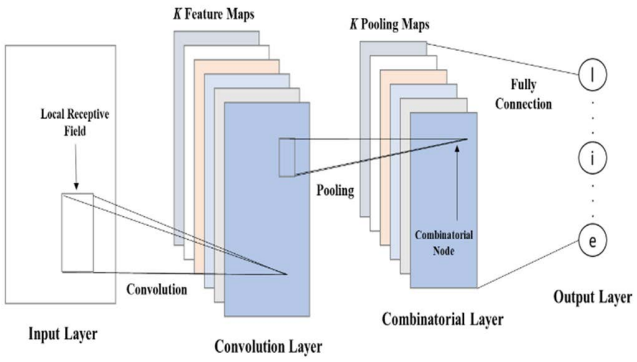

FIGURE 1. Structure of the Convolutional neural network (CNN).

C. OTHER APPROACHES USED FOR CROWD COUNTING

1) SENSOR-BASED SOLUTIONS

Several studies have addressed the problem of crowd estimation using sensor technologies such as Wireless Sensor Networks (WSN) and Radio Frequency Identification (RFID) [11] and [12]. Sensor technology approaches rely on the assumption that crowd members possess network devices that can transmit wireless signals (e.g., RFID wristbands, sensor tags, or smart phones). Certain sensor-based technologies estimate crowd density by counting the Unique Identifiers (UIDs) of the signals sensed or read in the area of interest. Some other solutions relate crowdedness to the Received Signal Strength Indicator (RSSI), Channel State Information (CSI), and Link Quality Indicator (LQI). The latter approaches are less complex in terms of processing, whereas they suffer from serious deployment problem in terms of number of devices and well as their signal communication reliability. Furthermore, their success is strongly determined by the cooperation of crowd members. Finally, these approaches constitute an exhausting and expensive choice for the targeted crowd size.

2) COMPUTER VISION SOLUTIONS1

In contrast, computer vision solutions require fewer devices and do not depend on crowd members' cooperation. It only needs a camera installed at the location to capture images of the crowded scene. Those images are then processed to extract useful information pertaining to the crowd density level [13] and [14], and statistical analysis is performed to find similarities in crowd texture between different images.

However in [15], the authors used mapping based on crowd count and foreground pixels obtained from binary and infrared images. However, statistical analysis and supervised feature extraction are computationally complex and time consuming, as well as not producing universal solutions because of the application-orientated process. Overall, while computer vision solutions require fewer devices for a particular area compared with sensor-based solutions, it is not generalizable to any crowd scene and cannot adapt to new features.

III. LITERATURE REVIEW

Since the estimation of the crowd density is important to manage and control the crowding, crowd management researchers are studying the effectiveness of neural networks for crowd density estimation [17] and [18] leveraging both crowd density estimation and classification accuracy.

Fu *et al.* [19], suggested an optimized convolutional neural network which overcomes the accuracy and speed requirements of engineering applications problems in present methods. This method constitutes the first research work using the convolutional neural network for crowd density estimation. It optimizes the multi-stage convolutional network structure by removing some network connections that have similar feature maps. This contribution increases the speed of estimation and reduces the computation cost for both training and testing phases. In addition, the authors designed a cascade of two convolutional neural network classifiers to improve the accuracy. The first classifier identifies samples that are clearly misclassified, whereas the second one reclassifies those samples. They applied the method on three datasets: PETS_2009, a Subway video and a Chunxi_Road video. They defined five classes; each class in each dataset has different range of people number. Finally, they established a comparative study in terms of results where their method outperforms other related work.

Oñoro & López-Sastre [20], proposed two deep learning approaches to count objects in images. The first approach, named Counting CNN (CCNN), is formulated as a regression model where the network learns how to map the appearance of the image patches to their corresponding object density

maps. The second approach, called Hydra CNN, is a scale-aware counting model able to estimate object densities in different very crowded scenarios without the scene's information. Hydra CNN learns a multiscale non-linear regression model, which uses a pyramid of image patches extracted at multiple scales to perform the final density prediction. Each approach is evaluated on three datasets; the UCSD pedestrian, the UCF CC 50, and the TRANCOS datasets. Sam *et al.* [21], used Switch Convolutional Neural Network (Switch-CNN) model for crowd counting. This model is considered as multi-column CNN (MCNN) based on three different architectures CNN regressors. It consists of a classifier (switch) employed to select the best regressor for an input crowd scene patch. Then, it divides the input image into nine non-overlapping patches. This model assumes that the characteristics of the crowd, such as density and appearance, can be consistent in a given patch for a crowd scene. It has a similar architecture for each CNN regressor; four convolutional layers with two pooling layers. The authors used three different datasets including ShanghaiTech. The classification accuracy reaches 64.39% for CNN-small and 73.75% for VGG-16. Their model learns to group crowd places based on latent factors correlated with crowd density.

Kumagai *et al.* [22], proposed an architecture of expert and gating CNNs to select the most compelling feature extractor CNN. Using a filter suitable for specific image perspective, each column was trained only on the images of a particular domain. This work used two challenging crowd counting datasets; the UCF CC 50 dataset and the Mall dataset. Later on, Al-Hadhrami *et al.* [23], provided a Single Convolution Neural Network with four convolution layers (Single-CNN4) for crowd counting considered as classification-based problem solving. They divided the input images into nine non-overlapping patches. Then, they examined the model with eight different phases distinguished in terms of the labeling process, training epochs, and dataset splitting.

Hu *et al.* [24], used a single-column network with multiple filter sizes to capture features at various scales by reducing the filter size after alternate sets of convolution layers. Two supervisory signals, crowd count and crowd density, are employed to learn crowd features and estimate the specific counting. They tested the approach on a dataset containing 107 crowd images with 45.000 annotated humans inside, each with headcounts ranging from 58 to 2201.

Dai *et al.* [25], proposed a DSNet for crowd counting, which is simple and easily trained but effective network. The DSNet is a Dense Scale Network used for Crowd Counting and made up of blocks that are densely linked dilated convolutional layers. As result, it can generate features with various receptive fields and capture crowds at various scales. Then, they evaluated their DSNet on four public datasets for crowd counting (ShanghaiTech, UCF-QNRF, UCF-CC 50 and UCSD). DSNet attained the best performance and made substantial gains (20 % on ShanghaiTech and UCSD, and 30 % on the others).

Sindagi and Patel [26], proposed an end-to-end cascaded network of CNNs with two-columns producing a combination of count estimation and count group classification. The first column classifies the crowd group and shares the extracted features to enhance the accuracy of the crowd count mapping of the second column. Extensive experiments on highly challenging publicly available datasets show that this method achieves lower count error and better density maps. Saqib *et al.* [27], proposed a Motion Guided Filter (MGF) based on Deep Convolution Neural Network (DCNN) and Faster-RCNN for crowd counting. They evaluated the performance of their approach on three publicly available datasets (PETS2009, UCSD dataset and Mall dataset). They used MGF to recover misdetections and improve the mean average accuracy of overall detections. It resulted in improving crowd counting and density estimation, as measured by the Mean Absolute Error (MAE) test for this Method (VGG16 + MGF). The results were compared to other techniques, which achieved 1.27 for the UCSD dataset, 1.89 for the Mall dataset and 1.21 for PETS2009 dataset.

Marsden *et al.* [28], proposed a scale-aware model to reduce the computation complexity of the multiple columns, feeding multiple scales of an image into the network, and the crowd counting estimated for each scale. The final count represents the average of all estimates performed with Shanghaitech and UCF CC 50 datasets. Zeng *et al.* [29], proposed a multi-scale convolutional neural network (MSCNN) for single image crowd counting based on scale-relevant features identification process. They evaluated their model for crowd counting on two separate datasets, including the UCF CC 50 and ShanghaiTech datasets. The results achieved by the MSCNN model outperforms other related works in terms of accuracy and robustness, where the obtained MAE is 83.8 on ShanghaiTech dataset and 363.7 on UCF CC 50 dataset.

Zhou *et al.* [30], proposed a Multiscale Generative Adversarial Network (MS-GAN) for generating high-quality crowd density maps of arbitrary crowd density scenes. The MS-GAN combines a multiscale convolutional neural network (generator) and an adversarial network (discriminator). The multiscale generator utilizes the fusion features from multiple hierarchical layers to detect people with a large-scale variation. The resulting density map produced by the multiscale generator was processed by a discriminator network trained to solve a binary classification task between a poor quality density map and real ground-truth ones. The experiments showed that adversarial training improved the performance of density map prediction.

Pu *et al.* [31], used a deep convolutional neural network (ConvNet), in which each layer only received input from the immediate previous layer. The classification of Subway carriage dataset into five density classes was also based on predetermined ranges of crowd count. This deep model provided an accuracy of 82% and 86% on five and three density classes, respectively, attributed to the close boundaries between class ranges and the slight variance in crowd count of the samples in the same class.

Kasmani *et al.* [32], proposed an Adaptive Counting Convolutional Neural Network (A-CCNN) that uses an ideally trained CCNN model. This model is used to analyze each

component of an input image in order to properly estimate the appropriate density map. The most notable features making the proposed model exceptional for crowd analysis resides on the capacity to manage large-scale differences in people's sizes appearing in the image, as well as the ability to create local density maps within a crowd scene. Therefore, the proposed model can give a complete view about the scattering of a crowd. They evaluated their method on different datasets in terms of MAE metric, achieving 367.3 for the UCSD dataset and an average of 1.35 for the UCF-CC dataset. Tripathy & Srivastava [33], proposed a two-input stream multi-column multi-stage convolutional neural network (TIS-MCMS-CNN) model to classify the crowd of PETS2009 and UCSD datasets into five density classes based on crowd count. This model used a shallower but wider model consisting of two-three column networks trained in parallel. This model tried to enhance the accuracy of previous models by capturing spatial and temporal features.

To increase crowd counting performance, Sang *et al.* [34], proposed scale-adaptive convolutional neural network (SaCNN) model. Their experiments were conducted on ShanghaiTech dataset and SaCNN model was compared with MCNN and Switching-CNN models. The experimental results of their improved SaCNN in part A are 75.84 for MAE and 124.88 for MSE. While in part B, the MAE is 11.03 and MSE is 18.55. Liu *et al.* [35], used just the first layer of the CNN pre-trained offline on ImageNet that is utilized in the statistical CNN features and then counted using the SVM (support vector machine). Using UCSD dataset, this approach demonstrated better accuracy and efficiency than other related methods, where MAE and MSE achieved 2.29 and 9.05, respectively.

Although, various network architectures and learning parameters have been proposed to better resolve some estimation issues. However, most existing models in the literature estimate the actual crowd count [36] and [37], while very few studies further classify the overall crowd density.

IV. PROPOSED APPROACH

The main goal of this work consists of using deep learning to develop a Model capable of learning and analyzing crowd features and then predicting the crowd density class. This work allows an appropriate decision-making for efficient management and incident prevention in a crowded environment.

The reason for this research is to count the crowds in an automated way without human intervention in order to deal with the crowds in a professional and faster way, where we used a dataset that contains images with a different number of people. The modal counts the heads in a specific spot where people are crowded. As this helps the security authorities to break up the crowd to prevent accidents, and at the present time, distancing and avoiding crowds is essential to prevent the spread of Covid-19 disease

A. PREPROCESSING

The preprocessing phase consists of increasing the size of the dataset by introducing new images using the available data and a segmentation method. The segmentation divides the image into nine non-overlapping sub-images and then calculates the density maps for each part. In one hand, the sum of the density map gives the number of heads in that map. In the other hand, it is necessary to know how the total count is distributed over the image segments. Thus, the dot-map is first created for the entire image, and then it is segmented. Subsequently, the sum and placement of '1's in a dot-map segment represents the headcount and locations of heads in the equivalent image segment. The full dot-map is also needed to perform labeling. Figure 2 shows the segmentation method.

FIGURE 2. Example of image segmentation.

B. FEATURES EXTRACTION AND LABELING

In this phase, the main feature of an image is extracted, which is the number of heads. Then, all images are labeled based on the number of heads in the image. Then, these images are divided into 20 and 33 labels (classes), used in two different experiments respectively.

C. PROPOSED MODEL ARCHITECTURE DETAILS

In this phase, we propose a Single Convolution Neural Network with three convolution layers (Single-CNN3) to count the crowd in a scene. First, to turn the problem solving from counting to density estimation, several classes are created including different ranges of crowd counts. These ranges are selected to define a specific density level. These levels designate various indicators of s where high levels indicate risk of congestion situations.

Second, we built the Single-CNN3 model with a simple structure to solve the crowd density estimation problem as a classification rather than a regression problem. It consists

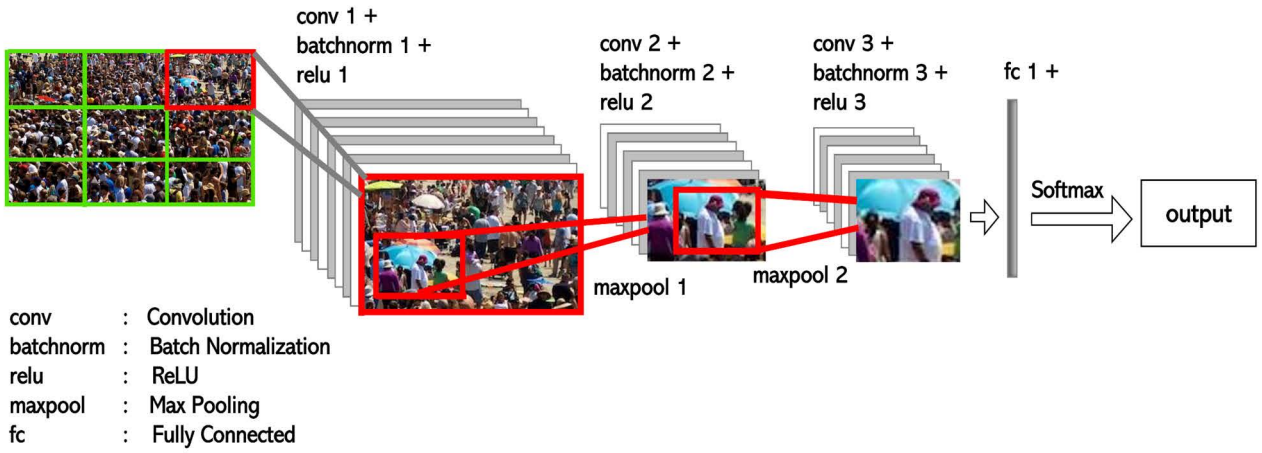

FIGURE 3. Single Convolutional neural network with three layers (Single-CNN3) model.

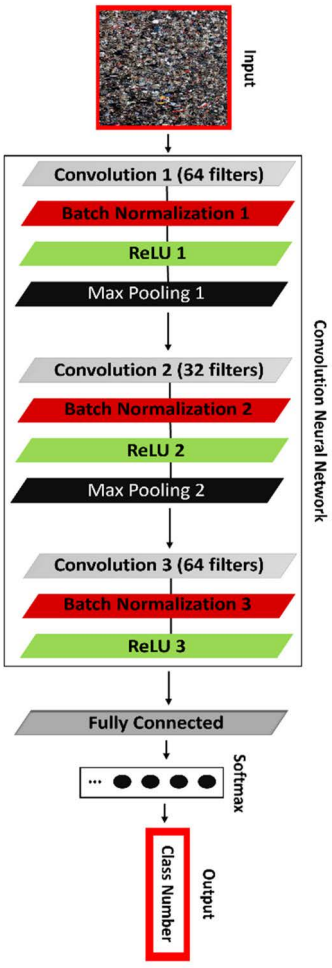

FIGURE 4. Architecture of the proposed model.

of a single CNN regressor with three convolution layers which has [1 1] filter size. The first and third layers uses 64 filters, whereas the second layer uses 32 filters. In addition, the proposed architecture includes two max-pooling layer, three batch normalization layers, three relu layers, one fully connected layer, one softmax layer beside the input and output layers, and finally, an output layer of n neurons representing the n density classes. The model is initialized by a Gaussian distribution with a learning rate of 0.0001.

FIGURE 5. Samples of images from the ShanghaiTech dataset.

The max-pooling layers help in maintaining the dominant features and reduce the size of the input. Earlier, the convolution layers extract low-level features such as edges, while high-level features are extracted in the later convolution layers. Those features are combined in a fully connected layer called softmax before the output layer.

The softmax layer must have the same number of nodes (n) in the output layer to be classified into the n density classes in the final output density layer. Since the size of the dataset is small whereas the number of labels (classes) is huge, the images are divided into nine non-overlapping images (patches) to make the dataset nine times more important. This process needs creating a density map for each patch.

Figure 3 shows the architecture of our proposed model (Single-CNN3) whereas figure 4 provides more details in terms of layers and mechanisms.

V. IMPLEMENTATION

A. TOOLS AND EQUIPMENTS

This proposed model was run on a MATLAB program with specific parameters in training options. Table 1 lists our

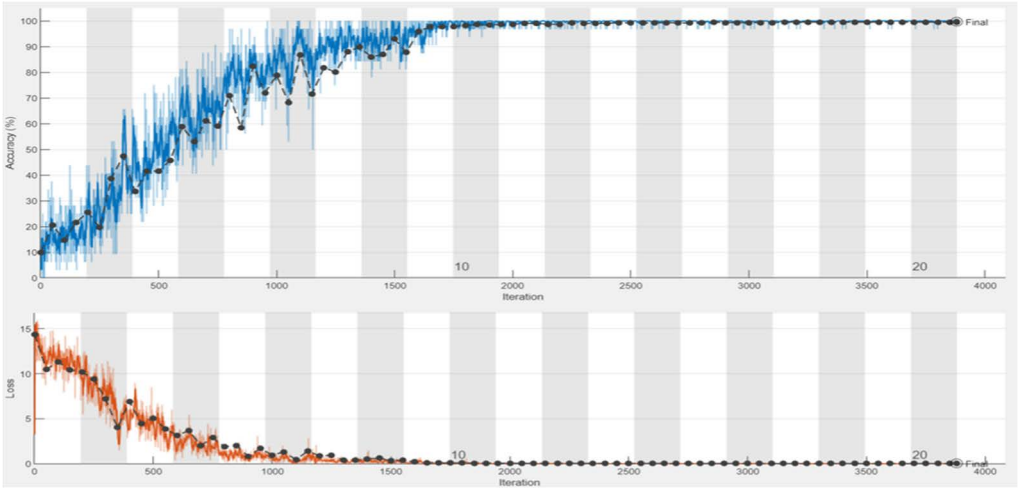

FIGURE 6. Accuracy and loss progress during the training for the first phase.

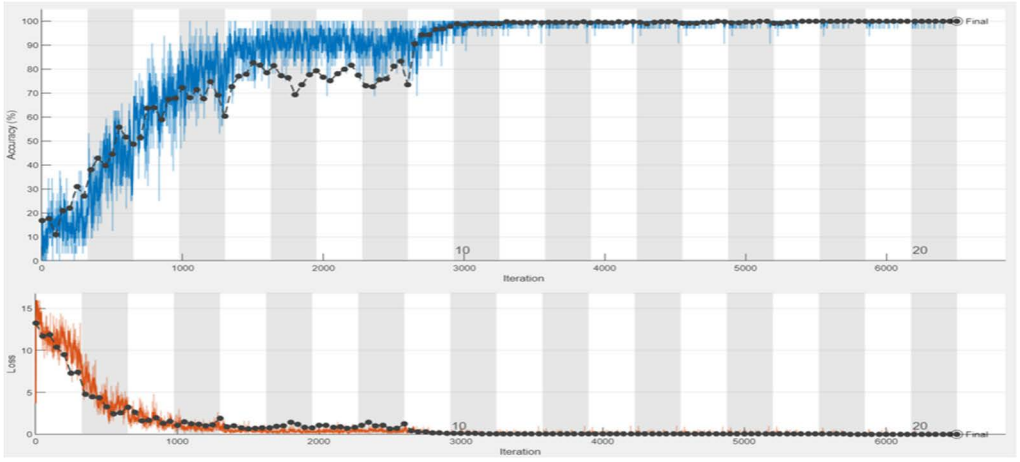

FIGURE 7. Accuracy and loss progress during the training for the second phase.

TABLE 1. Training parameters.

Initial Learn Rate	0.001
Validation Frequency	50
Learn Rate Drop Factor	0.1
Learn Rate Drop Period	8
L2Regularization	0.004
Max Epochs	20
Mini Batch Size	32

training parameters. The simulation equipment is an MSI laptop with Core (TM) i7-9750H CPU @ 2.60 GHz, 16 GB RAM, and NVIDIA GeForce RTX 2060.

B. THE SHANGHAITECH DATASET

ShanghaiTech is a new large-scale crowd density estimation dataset including 1198 annotated images, characterized by a total of 330 165 persons with centers of their heads annotated. This dataset is the largest one in terms of number of annotated people [38]. The dataset consists of two parts. The first part contains 482 images, which are randomly crawled from the Internet as well as from Arabian countries and Makkah. The second part contains 716 images taken from the busy streets of metropolitan areas in Shanghai. Both parts are divided into training and testing groups.

The crowd density varies significantly between the two subsets, making accurate estimation of the crowd more challenging than most existing datasets [39]. Figure 5 shows samples of the dataset.

C. PERFORMANCE METRICS

To evaluate the performance of our proposed model, multiple metrics are used and calculated over several experimentations. Three performance metrics are used in this work.

1) CLASSIFICATION ACCURACY

The accuracy is defined as the percentage of the number of correct predictions to the total number of input [40].

It is given by equation 1.

$$\text{Accuracy} = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions} \quad (1)$$

2) LOSS RATE

The loss rate metric measures how bad the model's prediction was on a single example. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater. Thus, the goal of training a model is to find a set of weights and biases that have low loss, on average, across all examples. The loss is calculated on training and validation; it's how well the model is doing for these two sets, unlike accuracy. Thus, the loss is not a percentage. Instead, it is a summation of the errors made for each example in training or validation sets [41].

3) CONFUSION MATRIX

We used confusion matrix to evaluate, visualize and summarize the performance of a classification model [42]. also knowing the actual and predicted classifications done by our model [43].

VI. EXPERIMENTS AND PERFORMANCE EVALUATION

A. EXPERIMENT AND RESULTS

The Single-CNN3 is applied on the ShanghaiTech dataset [44]. This data set is considered as crowd density estimation dataset including labels used by the neural network to train a model and provide context. It contains a large number of images with various crowd levels, which allows efficient learning on the congestion situation variety.

The model is trained using multiple GPU in parallel to speed up the training process. Part A and part B of the dataset are merged and randomly split into train and test subsets given by 80% and 20%, respectively. In this work, a certain number of classes including a specific range of head cunt are created. Each class represents a specific level of crowdness. Several experimentations are performed for a total number of classes of 20 and 33. Obviously, the classification problem is significantly increased as much as the number of classes is high. Each class includes a certain range of head count, which represents a specific crowd level. Therefore, this step turns our problem solving from regression to classification with large number of classes.

Experimentations are achieved into two phases. Within the first phase, the dataset is relabeled into 20 labels (classes) given by the following samples: 0, 1, 2, 3, 4, 5, 6-10, 11-15, 16-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-100, 101-150, 151-200, 201-250 and 251-300. In the second phase, the dataset is relabeled into 33 labels (classes) which are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-100, 151-200, 201-250, 251-300, 301-350 and 351-600. In the first phase, the model is trained with 20 epochs and executed 5 times. This resulted to 99.77% as average validation accuracy, 0.02 for loss rate, and 99.72% for the test accuracy.

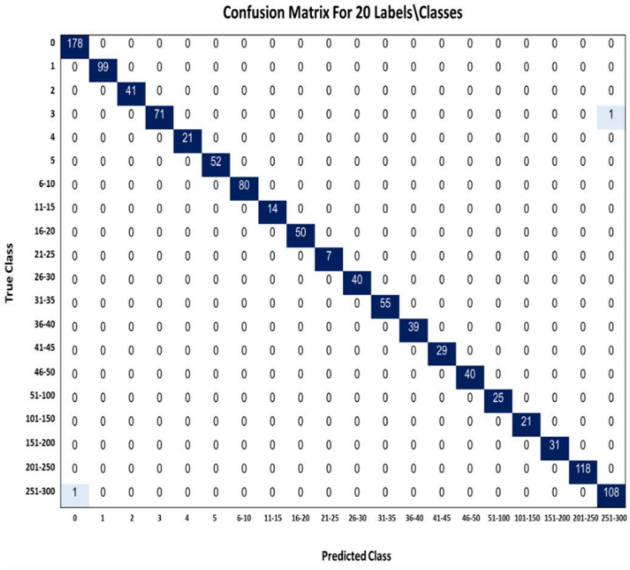

FIGURE 8. Confusion matrix of run #1 in the first phase.

TABLE 2. Results of single-cnn3 in the first phase (20 labels).

Runs	Val. Loss	Val. Acc.	Test Acc.
#1	0.017	100%	99.88%
#2	0.021	99.68%	99.72%
#3	0.05	99.36%	99.71%
#4	0.041	99.84%	99.42%
#5	0.017	100%	99.88%
Avg	0.029	99.77%	99.72%
Max	0.05	100%	99.88%
Min	0.017	99.36%	99.429%

The best result of the validation accuracy was 100%, whereas the worst result was 99.36%. In terms of loss, the results ranged from 0.017 to 0.05. Finally, the test accuracy reached significant results given between 99.42%. and 99.88%.

In the second phase, the same simulation scenario is performed, except the dataset labeling procedure including 33 classes. Actually, the results are slightly higher than those obtained in the first phase, where the average validation accuracy is 99.86%, the loss is 0.01748, and the test accuracy is 99.8794%. The best result of the validation accuracy was 100%, whereas the worst was 99.61%. The achieved loss is given between 0.011 and 0.027. Finally, the best-performed test accuracy was 100%, and the worst was 99.78%. Figures 6 and 7 show the evolution of the accuracy and loss metrics during the training for both first and second phases, respectively. Tables 2 and 3 show the results for the two phases.

Figure 8 shows the confusion matrix result of run #1 in the first phase of Table 2. Figure 9 shows the confusion matrix result of run #1 in the second phase of Table 3.

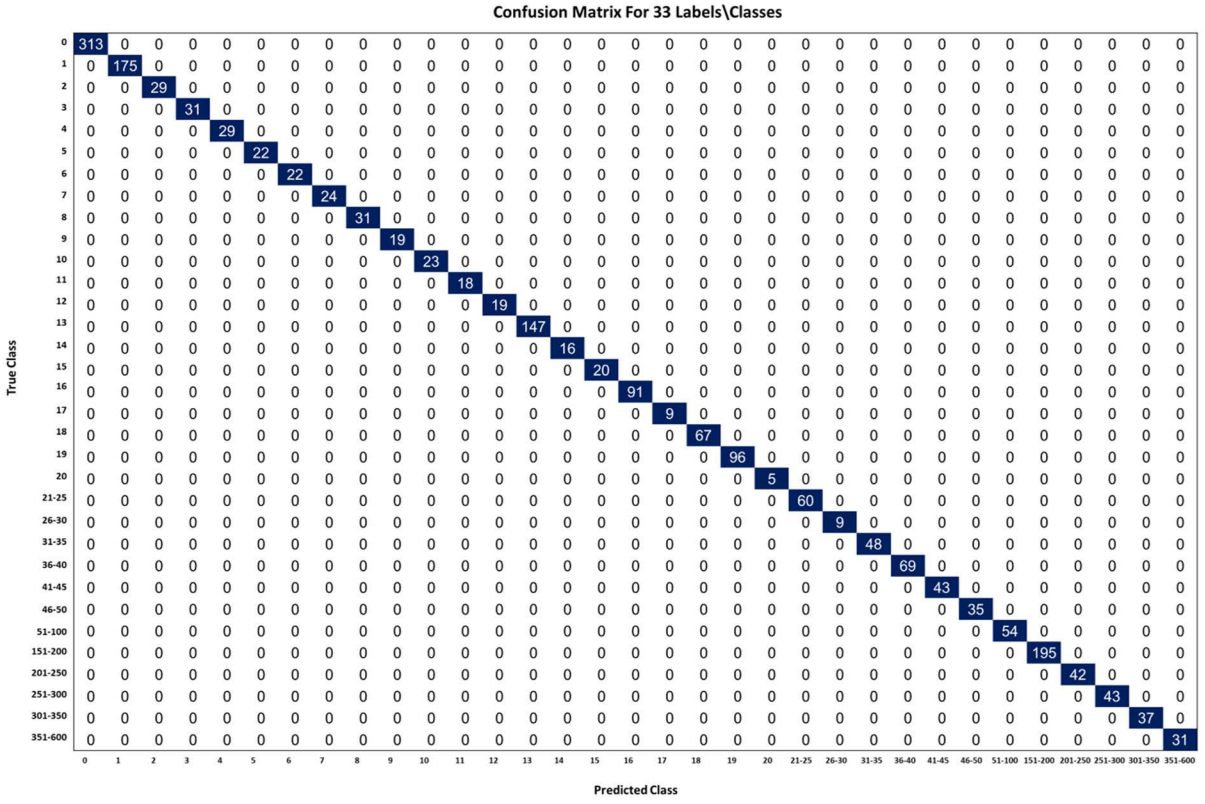

FIGURE 9. Confusion matrix of run #1 in the second phase.

TABLE 3. Results of single-cnn3 in the second phase (33 labels).

Runs	Val. Loss	Val. Acc.	Test Acc.
#1	0.024	99.61%	100%
#2	0.011	100%	99.95%
#3	0.012	100%	99.78%
#4	0.011	100%	99.78%
#5	0.027	99.71%	99.87%
Avg	0.017	99.86%	99.87%
Max	0.027	100%	100%
Min	0.011	99.61%	99.78%

TABLE 4. Comparative summary of switch-cnn, single-cnn4 and single-cnn3 results.

Model	Labels Nb	Val. Loss	Val. Acc.	Test Acc.
Switch-CNN	Ground-truth	NA	NA	76.3
S-CNN4 (5th phase)	20	0.29	99.5%	99.5%
S-CNN4 (8th phase)	33	0.51	99.88%	93.4%
S-CNN3 (avg - 1st phase)	20	0.03	99.78%	99.73%
S-CNN3 (avg –2nd phase)	33	0.02	99.86%	99.88%

B. PERFORMANCE EVALUATION

The performance evaluation establishes a comparison on the results and other characteristics between our proposed Single-CNN3 model against both Single-CNN4 and Switch-CNN models using ShanghaiTech dataset. Table 4 shows the results of the three models in terms of number of classes, validation loss, validation accuracy and test accuracy. Single-CNN3 model achieved the highest result for both validation accuracy and test accuracy using 20 labels. With 33 labels, the Single-CNN4 model performed the highest result in terms of validation accuracy, whereas the Single-CNN3 model achieved the highest result in terms of test accuracy.

From Table 4, Switch-CNN achieved the lowest testing accuracy (76.3%), whereas Single-CNN3 reached the highest testing accuracy given by 99.88% obtained within the second phase. Single-CNN3 also performed a high validation accuracy of 99.86 almost similar to 99.88% got by the Single-CNN4 (8th phase). Moreover, our proposed model succeeds to get the minimum validation loss of 0.02 accomplished as average result during the second phase. As result, the Single-CNN3 model has better performance than Single-CNN4 and Switch-CNN models in terms of accuracy and loss. Although both single-CNN3 and Single CNN4 have simple structures, use single-column CNN and solve the problem as a classification problem.

VII. WORK CONTRIBUTIONS

The most significant contributions of this research are summarized as follows:

1. It contributes to establishing a system capable of estimating the crowds automatically in record time during specific events such as the pilgrimages in religious ceremonies. It also allows saving time, effort, and cost compared to traditional counting and controlling crowded places.
2. The density estimation challenge has been switched from a regression problem to a classification problem by defining specific classes estimating various density levels.
3. Performance evaluation has been achieved using the most known ShanghaiTech dataset that reflects crowded events

The targeted applications of this system reside in helping the security authorities to control and organize crowds in touristic and recreational places as well as scientific edifices such as (universities, sea beaches, stadiums, and commercial centers). The system can help for Secure Covid-19 vaccine supply supervision in the available centers.

VIII. CONCLUSION

In this paper, we studied the convolutional neural network-based approaches, which are designed to accurately estimate the crowd density level in different environments. Recently, deep learning has attracted the interest of the research community and industry in varying applications of image classification and speech recognition. Our proposed approach CNN with three layers is applied in the ShanghaiTech dataset, which is a large dataset in terms of the annotated heads for crowd counting. The results of the approach have proven high accuracy up to 100% and a low loss rate. Then, a comparative study has been established between our proposed model and the switched convolutional neural networks in terms of accuracy and loss metrics. This model is developed on three layers CNN, used very large and recognized crowd counting dataset and evaluated compared to the existing state of the art models.

In conclusion, our approach contributes to establishing a system capable of estimating the crowds automatically in record time during the pilgrimages in religious ceremonies that count crowds faster and efficiently where helps the security disperse and dismantle crowds for the safety of all. Furthermore, the system contributes to helping for Secure Covid-19 Vaccine Supply Supervision in the available centers.

As future work, our approach can be applied in a more general context, including various databases. Furthermore, the model can be used by the concerned authorities in the governmental and non-governmental sectors for crowd density estimation and control.

DECLARATION OF INTEREST

The authors have no relevant conflicts of interest to disclose.

ACKNOWLEDGMENT

The authors would like to thank the Deanship of Scientific Research in King Saud University for funding and supporting this research through the initiative of DSR Graduate Students Research Support (GSR).

REFERENCES

[1] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 3642–3649, doi: 10.1109/CVPR.2012.6248110.

[2] M. Aziz, F. Naeem, M. Alizai, and D. Khattak, "Automated solutions for crowd size estimation," *Social Sci. Comput. Rev.*, vol. 36, no. 5, Sep. 2017, Art. no. 089443931772651, doi: 10.1177/0894439317726510.

[3] A. Choi-Fitzpatrick and T. Juskauskas, "Up in the air: Applying the Jacobs crowd formula to drone imagery," *Proc. Eng.*, vol. 107, pp. 273–281, Jan. 2015, doi: 10.1016/j.proeng.2015.06.082.

[4] J. H. Yin, S. A. Velastin, and A. C. Davies, "Measurement of crowd density using image processing," in *Proc. 7th Eur. Signal Process. Conf. (EUSIPCO)*, vol. 3. Edinburgh, U.K., 1994, pp. 1397–1400.

[5] S. A. M. Saleh, S. A. Suandi, and H. Ibrahim, "Recent survey on crowd density estimation and counting for visual surveillance," *Eng. Appl. Artif. Intell.*, vol. 41, pp. 103–114, May 2015, doi: 10.1016/j.engappai.2015.01.007.

[6] W. Albattah, M. H. K. Khel, S. Habib, M. Islam, S. Khan, and K. A. Kadir. (Feb. 2020). *Hajj Crowd Management Using CNN-Based Approach*. Accessed: Jun. 20, 2021. [Online]. Available: http://localhost/jspui/handle/123456789/24873

[7] S. D. Khan, Y. Salih, B. Zafar, and A. Noorwali, "A deep-fusion network for crowd counting in high-density crowded scenes," *Int. J. Comput. Intell. Syst.*, vol. 14, no. 1, pp. 1–12, Dec. 2021, doi: 10.1007/s44196-021-00016-x.

[8] Y. Liu, S. Liu, and X. Zhao, "Intrusion detection algorithm based on convolutional neural network," *DEStech Trans. Eng. Technol. Res.*, 2017, doi: 10.12783/dtetr/iceta2017/19916.

[9] M.-H. Oh, P. A. Olsen, and K. N. Ramamurthy, "Crowd counting with decomposed uncertainty," 2019, *arXiv:1903.07427*.

[10] W. Zhu, Y. Ma, Y. Zhou, M. Benton, and J. Romagnoli, "Deep learning based soft sensor and its application on a pyrolysis reactor for compositions predictions of gas phase components," in *Computer Aided Chemical Engineering*, vol. 44. Amsterdam, The Netherlands: Elsevier, 2018, pp. 2245–2250, doi: 10.1016/B978-0-444-64241-7.50369-4.

[11] M. Mohandes, "An RFID-based pilgrim identification system (a pilot study)," in *Proc. 11th Int. Conf. Optim. Electr. Electron. Equip.*, May 2008, pp. 107–112, doi: 10.1109/OPTIM.2008.4602508.

[12] S. Hashish and M. Ahmed, "Efficient wireless sensor network rings overlay for crowd management in Arafat area of Makkah," in *Proc. IEEE Int. Conf. Signal Process., Informat., Commun. Energy Syst. (SPICES)*, Feb. 2015, pp. 1–6, doi: 10.1109/SPICES.2015.7091437.

[13] G. Sen, W. Liu, and H. Yan, "Counting people in crowd open scene based on grey level dependence matrix," in *Proc. Int. Conf. Inf. Autom.*, Jun. 2009, pp. 228–231, doi: 10.1109/ICINFA.2009.5204926.

[14] W.-L. Hsu and T.-H. Chen, "People gathering recognition based on dynamic texture detection," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, Jul. 2015, pp. 334–339, doi: 10.1109/ICMLC.2015.7340944.

[15] N. Hussain, H. S. M. Yatim, N. L. Hussain, J. L. S. Yan, and F. Haron, "CDES: A pixel-based crowd density estimation system for Masjid al-Haram," *Saf. Sci.*, vol. 49, no. 6, pp. 824–833, Jul. 2011, doi: 10.1016/j.ssci.2011.01.005.

[16] R. Ma, L. Li, W. Huang, and Q. Tian, "On pixel count based crowd density estimation for visual surveillance," in *Proc. IEEE Conf. Cybern. Intell. Syst.*, Dec. 2004, pp. 170–173, doi: 10.1109/ICCIS.2004.1460406.

[17] S. D. Khan and S. Basalamah, "Sparse to dense scale prediction for crowd couting in high density crowds," *Arabian J. Sci. Eng.*, vol. 46, pp. 3051–3065, Oct. 2020, doi: 10.1007/s13369-020-04990-w.

[18] S. Basalamah, S. D. Khan, and H. Ullah, "Scale driven convolutional neural network model for people counting and localization in crowd scenes," *IEEE Access*, vol. 7, pp. 71576–71584, 2019, doi: 10.1109/ACCESS.2019.2918650.

[19] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, and C. Zhu, "Fast crowd density estimation with convolutional neural networks," *Eng. Appl. Artif. Intell.*, vol. 43, pp. 81–88, Aug. 2015, doi: 10.1016/j.engappai.2015.04.006.

[20] D. Onoro and R. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. Eur. Conf. Comput. Vis.*, vol. 9911, 2016, pp. 615–629, doi: 10.1007/978-3-319-46478-7_38.

[21] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4031–4039, doi: 10.1109/CVPR.2017.429.

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

[23] S. Al-Hadhrami, S. Altuwaijri, N. Alkharashi, and R. Ouni, "Deep classification technique for density counting," in *Proc. 2nd Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS)*, Riyadh, Saudi Arabia, May 2019, pp. 1–6, doi: 10.1109/CAIS.2019.8769489.

[24] Y. Hu, H. Chang, F. Nian, Y. Wang, and T. Li, "Dense crowd counting from still images with convolutional neural networks," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 530–539, Jul. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320316300256

[25] F. Dai, H. Liu, Y. Ma, J. Cao, Q. Zhao, and Y. Zhang, "Dense scale network for crowd counting," 2019, *arXiv:1906.09707*.

[26] V. A. Sindagi and V. M. Patel, "CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," 2017, *arXiv:1707.09605*.

[27] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein, "Crowd counting in low-resolution crowded scenes using region-based deep convolutional neural networks," *IEEE Access*, vol. 7, pp. 35317–35329, 2019, doi: 10.1109/ACCESS.2019.2904712.

[28] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Fully convolutional crowd counting on highly congested scenes," 2016, *arXiv:1612.00220*.

[29] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang, "Multi-scale convolutional neural networks for crowd counting," 2017, *arXiv:1702.02359*.

[30] Y. Zhou, J. Yang, H. Li, T. Cao, and S.-Y. Kung, "Adversarial learning for multiscale crowd counting under complex scenes," *IEEE Trans. Cybern.*, vol. 51, no. 11, pp. 5423–5432, Nov. 2021, doi: 10.1109/TCYB.2019.2956091.

[31] S. Pu, T. Song, Y. Zhang, and D. Xie, "Estimation of crowd density in surveillance scenes based on deep convolutional neural network," *Proc. Comput. Sci.*, vol. 111, pp. 154–159, Jan. 2017, doi: 10.1016/j.procs.2017.06.022.

[32] S. A. Kasmani, X. He, W. Jia, D. Wang, and M. Zeibots, "A-CCNN: Adaptive CCNN for density estimation and crowd counting," 2018, *arXiv:1804.06958*.

[33] S. K. Tripathy and R. Srivastava, "A real-time two-input stream multi-column multi-stage convolution neural network (TIS-MCMS-CNN) for efficient crowd congestion-level analysis," *Multimedia Syst.*, vol. 26, no. 5, pp. 585–605, Oct. 2020, doi: 10.1007/s00530-020-00667-4.

[34] J. Sang, W. Wu, H. Luo, H. Xiang, Q. Zhang, H. Hu, and X. Xia, "Improved crowd counting method based on scale-adaptive convolutional neural network," *IEEE Access*, vol. 7, pp. 24411–24419, 2019, doi: 10.1109/ACCESS.2019.2899939.

[35] S. Liu, S. Zhai, C. Li, and J. Tang, "An effective approach to crowd counting with CNN-based statistical features," in *Proc. Int. Smart Cities Conf. (ISC2)*, Sep. 2017, pp. 1–5, doi: 10.1109/ISC2.2017.8090827.

[36] X. Kong, M. Zhao, H. Zhou, and C. Zhang, "Weakly supervised crowd-wise attention for robust crowd counting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2722–2726, doi: 10.1109/ICASSP40776.2020.9054258.

[37] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-Crowd: A large-scale benchmark for crowd counting and localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2141–2149, Jun. 2021, doi: 10.1109/TPAMI.2020.3013269.

[38] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 589–597, doi: 10.1109/CVPR.2016.70.

[39] M. Liu, J. Jiang, Z. Guo, Z. Wang, and Y. Liu, "Crowd counting with fully convolutional neural network," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 953–957, doi: 10.1109/ICIP.2018.8451787.

[40] *Model Accuracy*. Accessed: Dec. 1, 2021. [Online]. Available: https://www.i2tutorials.com/machine-learning-tutorial/model-accuracy/

[41] A. A. Alashban and Y. A. Alotaibi, "Speaker gender classification in mono-language and cross-language using BLSTM network," in *Proc. 44th Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2021, pp. 66–71, doi: 10.1109/TSP52935.2021.9522623.

[42] *Compute Confusion Matrix for Classification Problem—MATLAB Confusionmat—MathWorks Switzerland*. Accessed: Dec. 6, 2021. [Online]. Available: https://ch.mathworks.com/help/stats/confusionmat.html

[43] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating trust prediction and confusion matrix measures for web services ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[44] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 833–841, doi: 10.1109/CVPR.2015.7298684.

ADAL ALASHBAN (Student Member, IEEE) received the B.Sc. degree in networking and telecommunication systems from Princess Nourah Bint Abdulrahman University (PNU), Riyadh, Saudi Arabia, in 2015. She is currently pursuing the master's degree with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University (KSU), Riyadh. She was a Lecturer with the Department of Natural and Engineering Sciences, College of Applied Studies and Community Service, KSU, from 2016 to 2018. Her research interests include artificial intelligence, wireless sensor networks, deep learning, convolutional neural networks, the Internet of Things, computer vision, signals and images processing, and speech and languages processing.

ALHANOUF ALSADAN received the bachelor's degree in computer science from King Saud University (KSU), Riyadh, Saudi Arabia, where she is currently pursuing the master's degree with the Department of Computer Science, College of Computer and Information Sciences. Her research interests include artificial intelligence, machine learning, computer vision, and the IoT.

NORAH F. ALHUSSAINAN received the bachelor's degree in networking and telecommunication systems from Princess Nourah Bint Abdulrahman University (PNU), Riyadh, Saudi Arabia, in 2015. She is currently pursuing the master's degree in computer engineering with the College of Computer and Information Sciences, King Saud University (KSU), Riyadh. Her research interests include machine learning, image processing, signal processing, artificial intelligence, and networking.

RIDHA OUNI received the Ph.D. degree from the University of Monastir in cooperation with the Institut Polytechnique de Grenoble, France, in 2002, and the H.D.R. degree in wireless networking field from the University of Monastir, Tunisia, in 2015.

He was an Assistant Professor with the Institut Préparatoire aux Études d'Ingénieurs de Monastir (IPEIM), Tunisia, from 1999 to 2009. He has been an Associate Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University (KSU), since February 2009. His research interests include wireless sensor networks, the IoT, networks interoperability, and machine learning.