IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 18, 2019, accepted May 13, 2019, date of publication May 23, 2019, date of current version June 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918650

Scale Driven Convolutional Neural Network Model for People Counting and Localization in Crowd Scenes

SALEH BASALAMAH[1], SULTAN DAUD KHAN[2], AND HABIB ULLAH[2]
[1]College of Computers and Information Systems, Umm Al-Qura University, Makkah 24211, Saudi Arabia
[2]College of Computer Science and Software Engineering, University of Hail, Hail 55211, Saudi Arabia

Corresponding author: Saleh Basalamah (smbasalamah@uqu.edu.sa)

This work was supported in part by the University of Hail and in part by Umm Al-Qura University.

ABSTRACT Counting and localization of people in videos consisting of low density to high density crowds encounter many key challenges including complex backgrounds, scale variations, nonuniform distributions, and occlusions. For this purpose, we propose a scale driven convolutional neural network (SD-CNN) model, which is based on the assumption that heads are the dominant and visible features regardless of the density of crowds. To deal with the problem of different scales of heads in different regions of the videos, we annotate a set of heads in random locations of the videos to develop a scale map representing the mapping of head sizes. We then extract scale aware proposals based on the scale map which are fed to the SD-CNN model acting as a head detector. Our model provides a response matrix rendering accurate head positions via nonmaximal suppression. For experimental evaluations, we consider three standard datasets presenting low density to high density crowd scenes. Our proposed SD-CNN model outperforms the state-of-the-art methods in terms of both frame-level and pixel-level analyses.

INDEX TERMS Convolutional neural networks, non-maximal suppression, head detection, crowd counting, motion analysis.

I. INTRODUCTION

With increase in population and rapid urbanization, crowd occurrences are regularly observed in the form of concert, political and religious gatherings. Although these gatherings serve peaceful purposes, yet present a lot of problems to security agencies and management. To ensure public safety, it is critical to understand crowd dynamics and congestion circumstances at crowded scenes [16], [39]. Crowd analysis can be used in numerous applications, for example, in detecting critical crowd levels, detecting anomalies, and tracking individuals or group of individuals. Among them, the most important and emerging application is to count the number of people in the scene.

The problem of crowd counting is to estimate the number of people attending the event or participating in political or religious gathering. This type of information is also very important for both political and safety point of view. Crowd counting can provide useful piece of information that could provide support in future event planning and public space design. Moreover, crowd counting can substantially reduce the cost by deploying exact number of security personnel required for public safety and security. Though crowd counting has numerous applications and has become the prime focus of many researchers, acquiring information about the localization of people in high density images has received least attention from the research community. The problem of localization is to find the exact location of the people in the scene. With the localization information, one can find out the distribution of people in the environment which is very crucial for crowd managers. Moreover, localization information can be used to detect and track [38] a person in dense crowds. Localization can provide an aide in generating the ground truth data that can be used to rectify counting errors generated by automated counting algorithms. Localization information provides the estimated locations (bounding boxes or dots) of the individuals in the image and the analyst can easily find and rectify the errors by removing false positives. This process can provide huge support to the coders for annotating high density images efficiently and effectively which is very tedious and hectic job.

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

Accurate crowd counting and localization are essential to handle large crowds for public safety. Most of the exiting crowd counting methods [2], [33], [37], [42], [43] are regression based that estimate the crowd count via regression of density maps. However, these methods only estimate count, they cannot localize individual pedestrian and therefore cannot produce the distribution of pedestrians in the environment. On the other hand, traditional methods [3], [14], [41] estimate the count via detecting individual pedestrians in the scene. These methods perform well in low density situations, where all parts of the pedestrian are fully visible. However, the performance of these method degrade when applied in high density situations. This attribute to the severe occlusion and clutter in the scene due to which most parts of the human body are not visible. In high density situations, where the people stand very close to each other and due to the half body occlusions, head is the only visible part. Although several strides have been made in human head detection [20], [30], [32] during the recent years, head detection in the images is still a challenging task. Due to variation in scales and appearances of heads, it remains a big problem to precisely distinguish human heads from the background. Moreover, the smaller sizes of the human heads make the problem even worse.

Generally, most of the state-of-the-art methods treat head detection as object detection problem. Object proposal generation is a pre-processing step and has been widely used in modern object detection pipelines. Object proposals are used to guide the search of objects and avoid exhaustive search across all the image locations. Recent methods use low-level image cues, such as saliency, gradient and edge information [8], [36], [46] to hypothesize objects in images. Later on, DeepBox [18] improved the proposals by re-ranking the object proposals generated by EdgeBox [46]. In DeepProposal [11] method, object proposals are generated by an inverse cascade from the final to the initial layer. MultiBox [23] extracts object regions by bounding box regression based on CNN features maps. However, person scenes and images are usually complex and have large variations in scales, appearances, and human poses. Consequently, the current state-of-the-art region proposal methods are less effective and usually results in low recall rates when applied to complex scenes. To address this problem, we propose a different strategy for generating object proposals to detect human heads in multiple scales. Our framework consists of the following three major components:

1) We generate scale-aware object proposals by generating a scale map. Scale map is generated by first sampling random person positions and then compute perspective values for each sampled position based on their relation to person's head size and then a linear regression is applied to fit the sampled values in each image based on the perspective geometry.
2) The second part is an object proposal classification network, which classifies each proposal into two classes (head/background).
3) Non-maximal suppression is applied to the response matrix and final detection results are produced at the original resolution. The response map is a matrix with resolution equal to the size of input image and obtained after processing all the proposals. The values of response map represent the classification score of all input proposals.

Comparing to other state-of-the-art methods, our framework has the following contributions:

- Ability to count and localize human heads in both low density and high density crowd images.
- Handles scale variations by generating scale-aware proposals.
- Generates density maps (response maps) which give the distribution of humans in the scene.
- Unlike previous crowd counting models that only estimate the crowd count, our method handles counting and localization problems simultaneously.

We perform extensive experiments on standard benchmarks datasets, i.e, UCSD dataset [4], and WorldExpo'10 [42] and UCF-CC-50 [13] to show the superiority of our approach over state-of-the-art methods.

II. RELATED WORKS

Deep learning has achieved tremendous success in the recent years. In the literature, various deep learning models are proposed for image segmentation, object classification and detection with excellent results. Inspired by the success of deep learning, the CNN models have been proposed in literature to estimate the count of people from the image. Generally deep learning models for crowd counting can be classified into two major categories, 1) *Regression based methods*, 2) *Detection based methods*. Regression based methods estimate the crowd count by performing regression between the image features and crowd size. In CNN based methods, density maps are generated from the image and count is obtained by performing integration over the density map. A Multi-column Convolutional Neural Network (MCNN) is proposed in [44], which utilizes three columns with filter size of different receptive field to compensate for perspective distortion. The CNN regression model with two configurations [42] estimates the number of people in a single image. Switch-CNN [29] uses multiple CNN based crowd counting architectures and proposes switching strategy to select one network based on the performance. Contextual Pyramid CNN [35] estimates the count by generating high-quality crowd density by incorporating global and local contextual information of crowd images. Different density estimation methods are compared in [15]. Crowd density is estimated in [45] by using different regression networks. Although the Regression based methods work well in high density situations as they capture generalized density information from the crowd image yet they suffer from the following limitations. 1) The performance of these methods degrade when applied to low density situations due to overestimating the count.

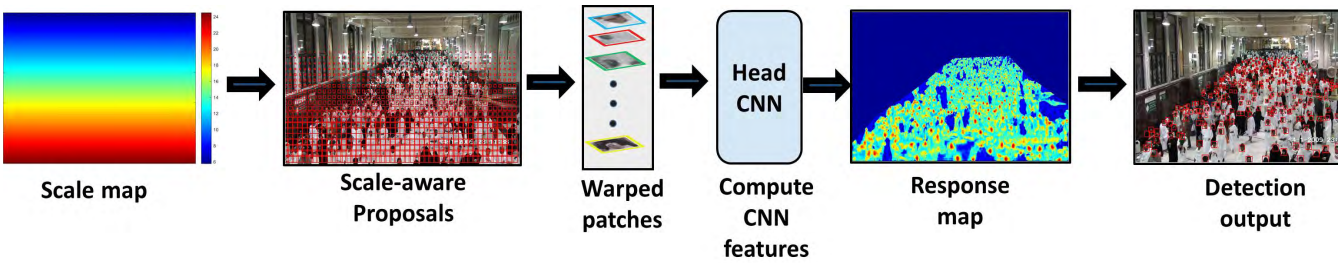

FIGURE 1. SD-CNN Model. We generate scale aware proposals based on the scale map which are fed to the SD-CNN model for rendering the response map. We then apply non-maximal suppression to detect and localize heads in the scene.

2) These methods cannot localize pedestrian in the scene and thus provide no information about the distribution of pedestrians in the environment which is very crucial for the crowd managers and security personnel.

On the other hand, *detection based methods* [30]–[32], train object detectors to localize the position of each person, where crowd count is the number of detections in the scene. A hybrid method is proposed in [21] that incorporates both regression and detection based counting and adaptively decide the appropriate counting mode for different image locations. Our proposed model is similar to [32] in a way that we also train a head detector. Unlike feeding general object proposal to the network as proposed in [32], we generate scale-aware proposals by using a *scale map*. Scale map estimates the object scales and use them to guide proposals rather than exhaustive searching on all scales. From our experiments, we observed that generating scale-aware proposals are very effective and can reduce the search space and ignores false positives at improper scales.

III. SCALE DRIVEN MODEL FOR COUNTING AND LOCALIZATION

In this work, to count and localize the people in images with large scale variations, we propose a new scale driven convolutional network (SD-CNN) model. The pipeline of our proposed model is shown in Figure 1. It comprises of three main components. Firstly, we annotate the sizes of the heads in random locations of the image to generate the scale map. Secondly, the scale map is subsequently used to produce scale aware proposals. This procedure is illustrated in details in Fig. 2. Finally, the scale aware proposals are fed to the SD-CNN model to detect and localize heads.

Object proposal generation is a pre-processing step and has been widely used in modern object detection pipelines. Object proposals are used to guide the search of objects and avoid exhaustive search across all the image locations. To generate object proposals, the first step is to estimate a scale map S. In order to estimate the scale map S, we need to understand the underlying factors that cause scale variations in the image. From empirical evidence, we confirm that drastic perspective distortions in images cause scale variations in the image as illustrated in Figure 3 (left image). The perspective distortion is related to camera calibration

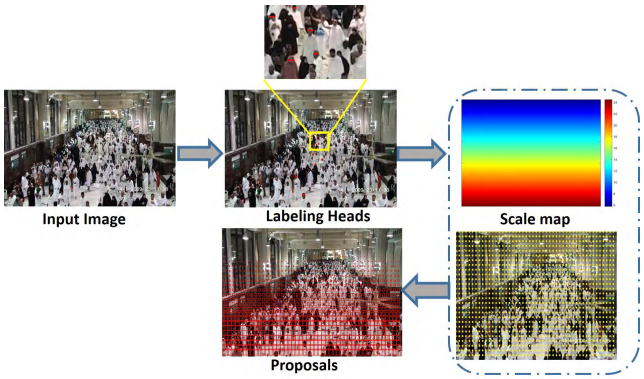

FIGURE 2. Proposal generation. After we annotate the sizes of heads in the image, we produce scale map depicting mapping of head sizes from the original image.

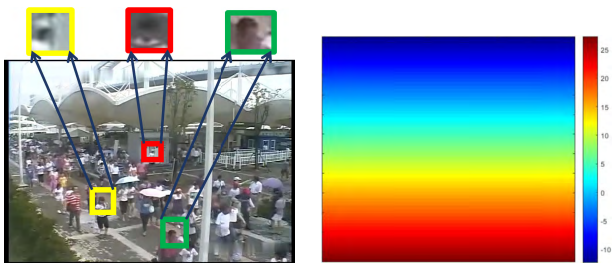

FIGURE 3. The size of the person head drastically changes due to perspective distortions as shown in the image on the left. The size of head at the bottom (in green) is bigger than size of head on the top (yellow) in the image. The estimated scale map on the right captures this perspective distortion at every location in the image.

which estimates 6 degrees-of-freedom (DOF) [10] and indicate the scale change from near to far in an image as shown in Figure 3. Therefore, we exploit perspective information to estimate the scale map S.

The value p_i of any pixel i of the scale map S represents a perspective value and defined as the number of pixels representing one meter at that location in the real scene [42]. Hence, the perspective value is related to the observed size of pedestrian in the image. We estimate the perspective value for each pixel by using perspective geometry of pinnhole camera as shown in Figure 4. In the Figure, a person of height P_H is walking on the ground, shot by the camera located at the

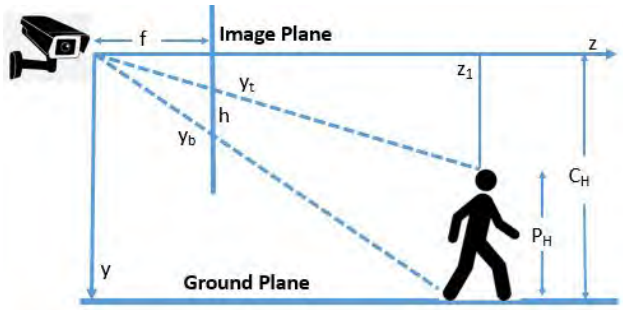

FIGURE 4. Perspective geometry of pinnhole camera.

height C_H from the ground. The head and feet of the person is mapped on the image plane at y_t and y_b, respectively, and f is the focal length. The cartesian coordinate system with y-axis represents the vertical direction, while z-axis represents the depth. From the perspective geometry of pinnhole camera, we can solve the similar triangles as follows

$$y_t = \frac{f(C_H - P_H)}{z_1}, \tag{1}$$

$$y_b = \frac{f(C_H)}{z_1} \tag{2}$$

From the above Equations 1 and 2, we compute the height h of pedestrian as

$$h = y_b - y_t = \frac{f(P_H)}{z_1} \tag{3}$$

we obtain height h by dividing both sides of Equation 3 by y_t as

$$h = \frac{P_H}{C_H - P_H} y_t \tag{4}$$

After obtaining height h of pedestrian, the perspective value p is given by:

$$p = \frac{h}{P_H} = \frac{1}{C_H - P_H} y_t \tag{5}$$

In order to generate scale map for input image, we approximate P_H to be the average size of adults (1.76 m) [42] for every pedestrian. To estimate C_H, we manually labeled height of random adults at different locations. Then we find the perspective value p_i of pixel i as $p_i = \frac{h_i}{1.76}$. We then employ linear regression method on Equation 5 and generate scale map. The scale map shown in Figure 3 (right image), captures the scale variations at every location in the image, with the values decreased from bottom to top indicating the change in person scale from front to remote end of the image and have same values in the same row. The red colors in the scale map represent bigger sizes in the image and blue color represents smaller sizes. The vertical bar shows the range of scales in the input image.

After generating the scale map S, the next step is to generate object proposals. We uniformly overlaid a gird G of points on the image and generate bounding boxes with grid points as their centers. Let $S(p_i)$ represents the size of pedestrian (in pixels) at location p_i. For every point $p_i \in G$, we generate bounding box of size $S(p_i)$ with point p_i as its center. Ideally, the resolution of the grid G and scale map S are the same as the resolution of the input image I; nonetheless this would imply huge computational costs. In order to avoid this problem, we define a parameter α, the value which is in the range of $\{0 < \alpha \leq 1\}$, indicating the resolution of the grid. Consider $R_x \times R_y$ is the original resolution of the image. The resulting resolution of the grid G is $G_x \times G_y$, where $G_x = \alpha(R_x)$ and $G_y = \alpha(R_y)$. Generally, the higher the value of α increase the resolution of the grid which results in large number of proposals. In this case, higher number of proposals are concentrated near the areas which likely to contain a pedestrian. However, the downside is that with lower values of α will produce small number of proposals which result in lower recall rates. This issue introduces a tradeoff in selection of parameter α. From the experiments evidences, we found that value of $\alpha = 0.65$ is ideal for most of the cases, so we fix α to 0.65 in the experiments.

The scale map S for the UCSD [4] and WorldExpo'10 [42] datasets were generated by labeling the height of pedestrians. However, for UCF-CC-50 [13] dataset having high dense crowds, the above process of generating the scale map is not applicable. The reason is the pedestrian bodies are not visible for labeling in such dense scenes. In dense crowd scenes, head is the only visible part and we noticed that similarly to the observed pedestrian height, the size of the head also changes due to perspective distortions. Therefore, in this case we interpret perspective value p_i by labeling head size as shown in Figure 2 (zoomed view). After labeling heads, instead of employing conventional linear regression, we adopt a novel non-linear regression to fit the perspective value. After computing mean perspective value at each sampled row y_t, we employ parametric tanh function to fit the average values over the entire row of y_t by

$$p = a \cdot \tanh(b \cdot (y_t + c)) \tag{6}$$

where a, b and c are the parameters.

IV. DETECTION NETWORK

After generating scale-aware proposals, the next step is to classify each proposal into two classes, i.e, head and background. Our detection network follows the classical R-CNN mode [12] and instead of using selective search [40] for proposal generation, we use scale-aware proposals. Before feeding to the network, we extend the bounding box of each proposal by a small margin and then image patch corresponding to each proposal is resized to fit the input layer of the CNN. For the head detection, we keep the square-like aspect ratios $\Re \in [\frac{2}{3}, \frac{3}{2}]$ for all bounding boxes.

The classical R-CNN is based on AlexNet architecture [17] which is pretrained on ImageNet [9] dataset. In addition to AlexNet, we used several other alternatives, for example, VGGS [5], VGG-verydeep-16 [34], and Oquab et al. [25]. From the experiment, we noticed that VGGS slightly outperforms AlexNet but was slower in both training and testing.

TABLE 1. Datasets. Summary of the three datasets including UCSD dataset [4], WorldExpo'10 [42], and UCF-CC-50 [13] is presented in terms of number of frames, total scenes, resolution, frames per second, and crowd size.

Dataset	Number of frames	Total scenes	Resolution	Frames per second	Crowd size
WorldExpo'10 [42]	4.4 million	108	576x238	50	1 - 253
UCSD [4]	2000	1	158x238	10	11-46
UCF-CC-50 [13]	50	50	various	images	94-4543

TABLE 2. Comparative analysis with other techniques on UCSD [4] dataset.

Methods	MAE	MSE
Regression Based Models		
Lempitsky et al. [19]	1.7	-
Kernel Ridge Regression [1]	2.16	7.45
Multi output Ridge Regression [7]	2.25	8.08
Gaussian process Regression [4]	2.24	7.97
Cumulative Attribute Regression [6]	2.07	6.86
CNN Based Models		
Zang et al. [42]	1.60	3.31
Count forest [26]	1.61	4.4
Liping et al. [45]	1.03	1.37
MCNN [44]	1.07	1.35
Faster R-CNN [28]	2.89	9.25
Proposed SD-CNN	1.01	1.28

TABLE 3. Comparative analysis with other techniques on WorldExpo'10 [42] dataset using MAE metric.

Methods	S1	S2	S3	S4	S5	Average
Zhang et al. [42]	9.8	14.1	14.3	22.2	3.7	12.9
MCNN [44]	3.4	20.6	12.9	13.0	8.1	11.6
Lingbo et al. [22]	2.6	11.8	10.3	10.4	3.7	7.76
Faster R-CNN [28]	18.65	37.94	19.84	42.67	15.27	26.87
SD-CNN	2.9	10.8	10.1	9.4	3.9	7.42

TABLE 4. Comparative analysis with other techniques on UCF-CC-50 [13] dataset.

Methods	MAE	MSE
idrees et al. [13]	419.5	590.3
Zhang et al. [42]	467.0	498.5
Liping et al. [45]	302.3	411.6
MCNN [44]	377.6	509.1
MRA-CNN [43]	240.0	352.6
Faster R-CNN [28]	592.09	672.19
Proposed SD-CNN	235.74	345.6

In the same way, VGG-verydeep-16 performed well but was much slower. Oquab et al. on the other hand performed better and achieved similar speed in both training and testing compared to AlexNet. In this paper, we used Oquab et al. pre-tained on ImageNet. For training the network, we assign each bounding box to one of the two classes,i.e, head and background. We decide this assignment based on intersection-over-union (IoU), which represents the overlap ratio between the candidate bounding box and ground truth bounding box. We fix a threshold value of 0.5 and any bounding box for which IoU $\geq$ 0.5 will be assigned to positive class, while the remaining bounding boxes will be assigned to the negative class. We keep training batch size of 64 proposals. We initialize the parameter of the network using ImageNet pre-trained network of Oquab et al. We minimize the parameter of the network with stochastic gradient descent (SGD) with momentum of 0.9 and weight decay 0.0005. We initialize the learning rate at 0.01, and decrease it by a factor of 10 after the validation error reaches saturation point.

For the localization task, to get the precise location of the heads, we post-process the response map by finding local peaks/ maximums based on fixed threshold. This process is also known as non-maixmal suppression. We use 1-1 matching strategy to compare the predicted locations with the ground truth locations and use Precision and Recall metrics for evaluation. The performance of the localization task is mainly affected by changing the threshold value.

V. EXPERIMENT RESULTS

In this section we discuss both qualitative and quantitative analysis of the results obtained from the experiments. We evaluate our SD-CNN framework using three publicly available datasets, UCSD dataset [4], WorldExpo'10 [42] and UCF-CC-50 [13]. The summary of the datasets is described in Table 1. Generally, these datasets are annotated in a way that can only be useful for evaluating the performance of regression models. Typically, in these datasets, every individual pedestrian is annotated with a dot in the scene. These dot annotations are not suitable for training our SD-CNN model or other detection based methods. Therefore, we annotated each pedestrian with a bounding box that cover whole body of pedestrian. In the same way, we also annotated the head of each pedestrian using the bounding box.

After annotation, we then trained different models discussed in Section IV on Titan Xp with learning rate at 0.01 and decrease it by a factor of 10 after the validation error reaches saturation point.

TABLE 5. Localization performance of different methods in terms of Average Precision (Avg), Average Recall (AvR) and Area Under Curve (AUC). The values of AvP and AvR are represented in percentages.

Methods	WorldExpo'10			UCSD			UCF-CC-50		
	AvP	AvR	AUC	AvP	AvR	AUC	AvP	AvR	AUC
Zang et al. [42]	45.87	39.23	0.45	65.64	59.65	0.64	35.27	29.67	0.29
MCNN [44]	55.24	52.28	0.51	69.74	65.67	0.71	33.27	35.64	0.31
Kang et al. [15]	42.98	39.27	0.41	67.28	55.32	0.67	24.13	30.27	0.27
Liping et al. [45]	65.72	47.91	0.58	71.73	68.68	0.72	34.28	31.19	0.31
Faster R-CNN [28]	25.18	27.53	0.21	33.28	37.62	0.30	14.52	12.69	0.14
Proposed SD-CNN	69.46	67.65	0.69	73.58	71.68	0.74	45.67	40.12	0.45

For the sake of comprehensive evaluation, we divide the experiment setup into two phases. In the first phase, we evaluate and compare the crowd counting performance while in the second phase, we evaluate and compare localization performance of our proposed SD-CNN model with other state-of-the-art methods.

A. COUNTING PERFORMANCE

In this section, we evaluate the performance of different crowd counting methods. We use Mean Absolute Error (MAE) and Mean Square Error (MSE) as evaluation measures to compare the counting performance of the SD-CNN against the state-of-the-art methods and is defined as.

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\mu_t - G_t| \tag{7}$$

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(\mu_t - G_t)^2 \tag{8}$$

where T is the total number of testing frames. While μ_t and G_t are the predicted and ground-truth count of pedestrians respectively at frame t.

The UCSD dataset consists of 2000 frames of size 158 × 238 captured from a single camera at 10 fps. We follow the same settings as in [4] and use frames from 601 to 1400 as training, and the remaining 1200 are used in the testing phase. The dataset captures low density crowds where crowds are sparsely distributed. We evaluate and compare our results with different regression and CNN based methods. The results of SD-CNN and other methods are reported in Table 2. From the table, it is obvious that our SD-CNN outperforms other state-of-the-art methods.

We next evaluate and compare the performance of our framework and other state-of-the-art methods using WorldExpo'10 dataset. This data set was first introduced by Zhang et al. [42] and contains 1132 annotated video sequences which are captured by 108 cameras from different viewpoints. There are total of 199,923 head annotations that span over 3980 frames. In training stage, total 3380 frames are used and for the testing we used five different video scenes. Thanks to the author of [42] for providing the perspective maps. For the fair comparison, we use ROI regions provided by the [42] in each test scene. We use the same evaluation metric (MAE) and the results are presented in Table 3. It can be observed from Table 3 that SD-CNN outperforms existing approaches on an average scale while achieves comparable performance in five different scenes. From the results, we infer that the perspective information generally increases the performance of our proposed SD-CNN considering various scenarios. Liu *et al.* [22] proposed regression based crowd counting approach which works well in extreme dense situations, since they can capture density dependent information. The proposed method by Liu *et al.* [22] is dependent on the density of crowd. Therefore, this is the reason that Liu *et al.* [22] outperforms our method by a small margin in scene S1 and S5, since these scenes contains high density crowd with rich texture information. In most cases, when the density of crowd changes, the performance of Liu *et al.* [22] degrades. Moreover, Liu *et al.* [22] is regression based model and cannot localize persons heads, and thus cannot provide information about the distribution of pedestrians in the environment which is very crucial for the crowd managers and security personnel.

UCF-CC-50 [13] is a challenging dataset which contains 50 annotated images of different resolutions, view points, and with the densities drastically changing from 94 persons/image to 4543 persons/image. We followed the same standard of 5-fold cross-validation proposed by [13] for evaluating and comparing the methods. We evaluate and compare the results of different state-of-the-art methods in Table 4. From Table 4, it is obvious that our proposed SD-CNN outperforms other state-of-the-art methods. This experiment signifies the importance of using perspective information for estimating crowd count in images with widely varying densities.

In Table 2, Table 3 and Table 4, we evaluated and compared the performance of Faster-RCNN [28] on three datasets. The results of our proposed method are significantly better than Faster-RCNN. Faster-RCNN achieve good results only if the size of objects is very large. Faster-RCNN is based on PASCAL VOC dataset for training and testing where the actual size of most objects in the dataset is large. However, in our problem we are interested in detecting heads (size of 10–15 pixels), which are usually small. The detection network in Faster R-CNN has trouble to detect such small objects. The performance of Faster-RCNN becomes worse when applied to high density situations. The reason is that the ROI-pooling layer builds features only from one single high level feature map. For example, the backbone of

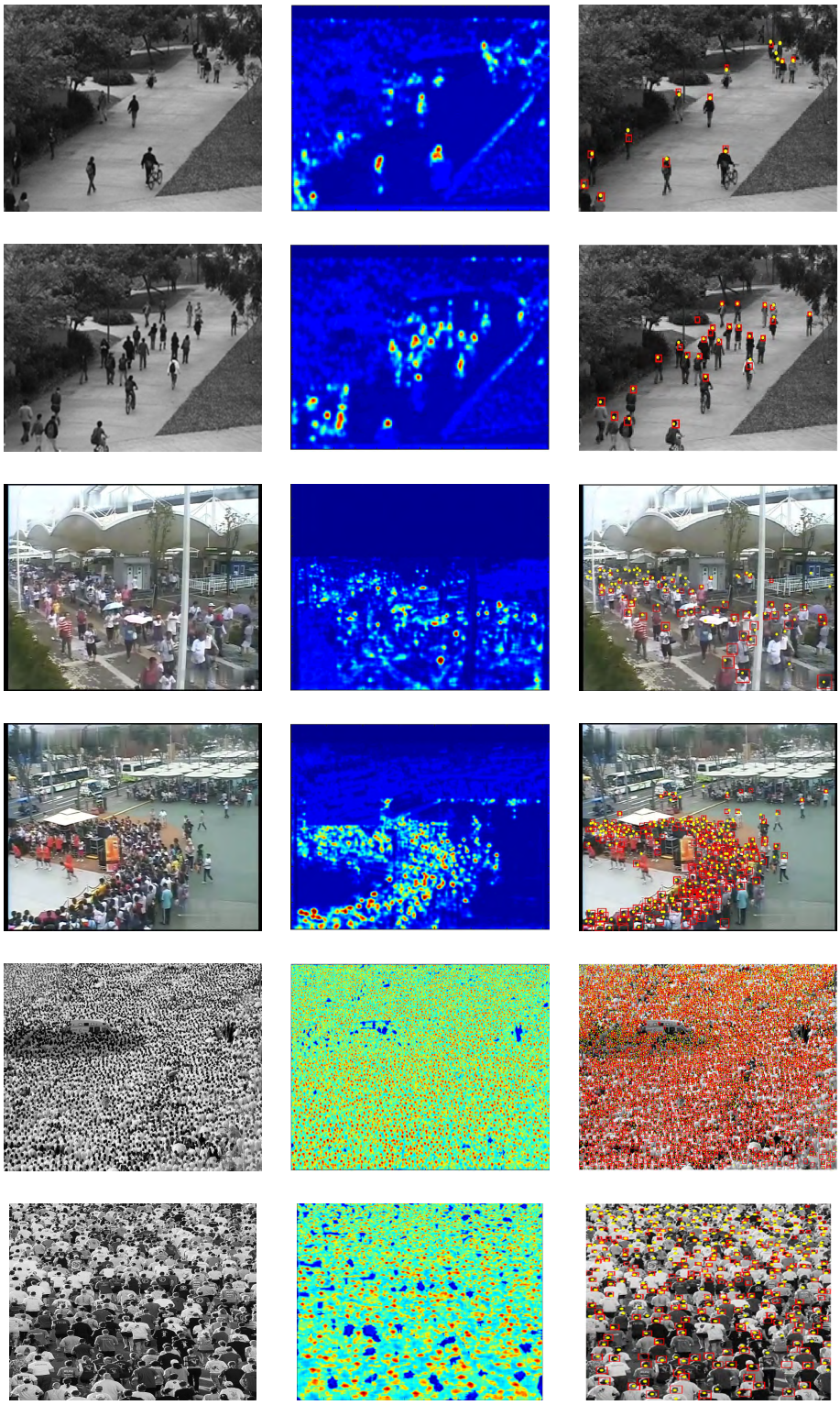

FIGURE 5. Results of samples frames from UCSD (1^{st} and 2^{nd} rows), WorldExpo'10 (3^{rd} and 4^{th} rows) and UCF-CC-50 datasets (5^{th} and 6^{th} rows). The first column represents the input sample images from different datasets. The second column shows the corresponding responses maps (density maps), while the third column shows the final detections. The yellow dot represents the groundtruth while the red bounding box is the predicted location by our approach. The Figure can be best viewed in color.

Faster-RCNN (e.g, VGG-16) model does ROI-pooling from the 'conv5' layer, which has an overall stride of 16. When the object size is less than 16 pixels, the projected ROI pooling region is less than 1 pixel in the 'conv5' layer even if the proposed region is correct. Thus the detector will have much difficulty to predict the object class and bounding box location based on information from only one pixel.

B. LOCALIZATION PERFORMANCE

In this section, We evaluate both qualitatively and quantitatively the localization performance of our framework. In order to quantify the localization error, we associate the center of estimated bounding box with the ground truth location (single dot) through 1-1 matching strategy. We then compute Precision and Recall at various thresholds and report the overall localization performance in terms of area under the curve. In order to estimate the location, we use the same density maps generated by state-of-the-art methods followed by non-maxima suppression algorithm. The results are reported in Table 5. It is obvious that our proposed model presents higher Precision and Recall rates as compared to the state-of-the-art methods. These results attribute to the fact that our model generates scale-aware proposals that capture wide range of head sizes in each image. It can also be observed that all other methods present lower rates for UCF-CC-50 dataset as compared to WorldExpo'10 and UCSD datasets. This is due to the fact the UCF-CC-50 dataset contains more dense images with heavy occlusions as compared to WorldExpo'10 and UCSD datasets. We also show some qualitative results of our proposed method in Figure 5. From the Figure 5, it is obvious that the sample images from the UCSD dataset represent low density scene. The sample images taken from two different scenes of WorldExpo'10 dataset represent medium densities and the images from UCF-CC-50 represent relatively more complex and extreme high density scenes. From our experiments, we find out that our method performs well in both high and low density scenes and is independent of the scene density. As it is clear from the figure, that in most of cases, our proposed method precisely localizes the heads even in the complex scenes.

Our method will incur computation time. We compute the computation time of proposed framework using WorldExpo10 and UCSD dataset. We found that our frame work took 0.87 and 0.34 seconds to process an image from WorldExpo10 and UCSD datasets, respectively. We further investigated time complexity of our proposed framework using UCF-CC-50 dataset and found out that computation complexity of our framework is directly related to the image resolution. UCF-CC-50 dataset contains different images with various resolutions. High image resolution will lead to high computation complexity, since large number of proposals will be generated to estimate the response map. We compute the computation time for each image and found out that average computation time for UCF-CC-50 dataset is 1.78 seconds.

VI. CONCLUSION

This paper presented a novel SD-CNN model to estimate the count by detecting and localizing the humans in dense crowd scenes. To tackle the problem of scale variations, we generated scale-aware head region proposals by exploiting the perspective information. This strategy has significantly reduced the classification time and also resulted in boosting the detection accuracy. We evaluated SD-CNN on three datasets, i.e, UCSD, WorldExpo', and UCF-CC-50 and have achieved noticeable improvements in the results.

In our future work, we would further improve the localization results since the localization accuracy is mainly affected by the post-processing step (non-maxima suppression in our case).

ACKNOWLEDGMENT

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU for this research.

REFERENCES

[1] S. An, W. Liu, and S. Venkatesh, "Face recognition using kernel ridge regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.

[2] C. Arteta, V. Lempitsky, and A. Zisserman, "Counting in the wild," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 483–498.

[3] G. J. Brostow and R. Cipolla, "Unsupervised Bayesian detection of independent motion in crowds," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 594–601.

[4] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–7.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," May 2014, *arXiv:1405.3531*. [Online]. Available: https://arxiv.org/abs/1405.3531

[6] K. Chen, S. Gong, T. Xiang, and C. Change Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2467–2474.

[7] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. BMVC*, vol. 1, 2012, pp. 1–11.

[8] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3286–3293.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[10] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 930–943, Aug. 2003.

[11] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool, "Deepproposal: Hunting objects by cascading deep convolutional layers," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2578–2586.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[13] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554.

[14] H. Idrees, K. Soomro, and M. Shah, "Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 1986–1998, Oct. 2015.

[15] D. Kang, Z. Ma, and A. B. Chan, "Beyond counting: Comparisons of density maps for crowd analysis tasks—Counting, detection, and tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1408–1422, May 2019.

[16] S. D. Khan, S. Bandini, S. Basalamah, and G. Vizzari, "Analyzing crowd behavior in naturalistic conditions: Identifying sources and sinks and characterizing main flows," *Neurocomputing*, vol. 177, pp. 543–563, Feb. 2016.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[18] W. Kuo, B. Hariharan, and J. Malik, "DeepBox: Learning objectness with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2479–2487.
[19] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1324–1332.
[20] W. Li, H. Li, Q. Wu, F. Meng, L. Xu, and K. N. Ngan, "Headnet: An end-to-end adaptive relational network for head detection," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[21] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "Decidenet: Counting varying density crowds through attention guided detection and density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5197–5206.
[22] L. Liu, H. Wang, G. Li, W. Ouyang, and L. Lin, "Crowd counting using deep recurrent spatial-aware network," Jul. 2018, *arXiv:1807.00601*. [Online]. Available: https://arxiv.org/abs/1807.00601
[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.
[24] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 615–629.
[25] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1717–1724.
[26] 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. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3253–3261.
[27] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 705–711.
[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[29] 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)*, vol. 1, Jul. 2017, pp. 4031–4039.
[30] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein, "Person head detection in multiple scales using deep convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7.
[31] 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.
[32] M. Shami, S. Maqbool, H. Sajid, Y. Ayaz, and S.-C. S. Cheung, "People counting in dense crowd images using sparse head detections," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[33] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5245–5254.
[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[35] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1861–1870.
[36] H. Ullah, A. B. Altamimi, M. Uzair, and M. Ullah, "Anomalous entities detection and localization in pedestrian flows," *Neurocomputing*, vol. 290, pp. 74–86, May 2018.
[37] H. Ullah, M. Uzair, M. Ullah, A. Khan, A. Ahmad, and W. Khan, "Density independent hydrodynamics model for crowd coherency detection," *Neurocomputing*, vol. 242, pp. 28–39, Jun. 2017.
[38] M. Ullah, F. A. Cheikh, and A. S. Imran, "Hog based real-time multi-target tracking in Bayesian framework," in *Proc. 13th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2016, pp. 416–422.
[39] M. Ullah, H. Ullah, N. Conci, and F. G. B. De Natale, "Crowd behavior identification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1195–1199.
[40] K. E. A. Van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation as selective search for object recognition," in *Proc. ICCV*, 2011, pp. 1–8.
[41] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, 2005.
[42] 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)*, Jun. 2015, pp. 833–841.
[43] Y. Zhang, C. Zhou, F. Chang, and A. C. Kot, "Multi-resolution attention convolutional neural network for crowd counting," *Neurocomputing*, vol. 329, pp. 144–152, Feb. 2019.
[44] 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)*, Jun. 2016, pp. 589–597.
[45] L. Zhu, C. Li, Z. Yang, K. Yuan, and S. Wang, "Crowd density estimation based on classification activation map and patch density level," in *Neural Computing & Applications*. Berlin, Germany: Springer, 2019, pp. 1–12.
[46] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Zürich, Switzerland: Springer, 2014, pp. 391–405.

SALEH BASALAMAH received the M.Sc. degree from the University of Bristol, U.K., in 2000, and the Ph.D. degree from Imperial College London, U.K., in 2005. He is currently an Associate Professor with Umm Al-Qura University, Saudi Arabia. His research interests include computer vision and multimedia.

SULTAN DAUD KHAN received the M.Sc. degree in electronics & communication engineering from Hanyang University, South Korea, in 2010, and the Ph.D. degree from the University of Milano-Bicocca, Italy, in 2016. He is currently an Assistant Professor with the University of Hail, Saudi Arabia. His research interests mainly focus on computer vision application to pedestrian and crowd analysis.

HABIB ULLAH received the Ph.D. degree from the University of Trento, Italy, in 2015, and the M.Sc. degree in computer engineering from Hanyang University, Seoul, South Korea, in 2009. He is currently an Assistant Professor with the University of Hail, Saudi Arabia. His research interests include computer vision and machine learning.