DISAM: DENSITY INDEPENDENT AND SCALE AWARE MODEL FOR CROWD COUNTING AND LOCALIZATION

Sultan Daud Khan[1], Habib Ullah[1], Mohammad Uzair[2], Mohib Ullah[3], Rehan Ullah[4], Faouzi Alaya Cheikh[3]

[1]College of Computer Science and Engineering, University of Ha'il, Saudi Arabia
[2]Defence and Systems Institute at University of South Australia, Australia
[3]Department of Computer Science, Norwegian University of Science and Technology, Norway
[4] College of Computer, Department of IT, Qassim University, Saudi Arabia

ABSTRACT

People counting in high density crowds is emerging as a new frontier in crowd video surveillance. Crowd counting in high density crowds encounters many challenges, such as severe occlusions, few pixels per head, and large variations in person's head sizes. In this paper, we propose a novel Density Independent and Scale Aware model (DISAM), which works as a head detector and takes into account the scale variations of heads in images. Our model is based on the intuition that head is the only visible part in high density crowds. In order to deal with different scales, unlike off-the-shelf Convolutional Neural Network (CNN) based object detectors which use general object proposals as inputs to CNN, we generate scale aware head proposals based on scale map. Scale aware proposals are then fed to the CNN and it renders a response matrix consisting of probabilities of heads. We then explore non-maximal suppression to get the accurate head positions. We conduct comprehensive experiments on two benchmark datasets and compare the performance with other state-of-the-art methods. Our experiments show that the proposed DISAM outperforms the compared methods in both frame-level and pixel-level comparisons.

Index Terms— Crowd counting, Convolution networks, Head detection, Classification

1. INTRODUCTION

With increase in population and rapid urbanization, crowd occurrences are regularly observed during concert, political and religious gatherings. Although these gatherings serve peaceful purposes, yet crowd disasters still occur. To ensure public safety, it is critical to understand crowd dynamics and congestion circumstances at crowded scenes [1][2][3]. 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.

Crowd counting can provide useful piece of information for future event planning and public space design. 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 advantages and has become the prime focus of many researchers, localization in high density images has received least attentation from the research community. Localization provide 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. Localization information can be used to detect and track a person in dense crowds [4]. Localization can also be used to rectify counting errors from automated counting algorithms. Localization provides the estimated locations (bounding boxes or dots) of the individuals in the image and the analyst can easily find and remove false positives. This process can also help the analyst annotating high density images efficiently and effectively.

In this paper, we propose a novel model for crowd counting and localization of people in a crowd image. Our goal is to detect and estimate the location of human heads as head is the only visible feature in high density crowd images. In order to achieve this goal, we propose a model that is composed of three components. First component is a Convolution Neural Network that acts as a head detector. In the second component, we generate scalemap and obtain scale aware head proposals by using scalemap. We then feed each proposal to the CNN and obtain classification score for each proposal. After processing all proposals, a response matrix is obtained, where higher responses indicate high probabilities of heads. Finally, non-maximal suppression is applied to the response matrix and final detection results are produced at the original resolution.

The proposed model has following contributions: (1) the ability to count and localize human head in high density crowd images; (2) handles scale variations in head sizes appearing in image; and (3) generates density 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.

2. 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 [5], which utilizes three columns with filter size of different receptive field to compensate for perspective distortion. The CNN regression model with two configurations [6] estimates the number of people in a single image. Switch-CNN [7] uses multiple CNN based crowd counting architectures and proposes switching strategy to select one network based on the performance. Contextual Pyramid CNN [8] 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 [9]. Crowd density is estimated in [10] 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. 2) These methods can not 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* [11, 12, 13], 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 [14] 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 [12] in a way that we also train a head detector. Unlike feeding general object proposal to the network as proposed in [12], 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.

3. PROPOSED METHODOLOGY

In this work, to count and localize the people in images with large scale variations, we propose a new Density Independent and Scale Aware Model (DISAM). The pipeline of our proposed model is shown in Figure 1. It comprises of three main components which are described in the following sections.

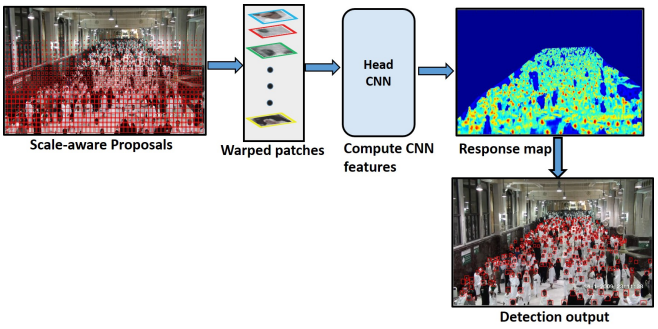

Fig. 1. Pipeline of Density Independent and Scale Aware Model (DISAM).

3.1. Generating Scale Aware Object Proposals

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.

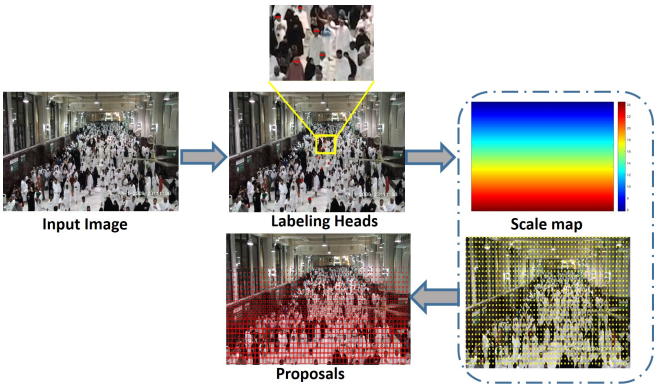

Fig. 2. Density Independent and Scale-Aware proposal generation pipeline.

Several object proposal methods are reported in literature. In DeepProposal [15] method, object proposals are generated by an inverse cascade from the final to the initial layer. MultiBox [16] and SSD [17] extract object regions by bounding box regression based on CNN features maps. However, in

high density crowded scenes, where the people are usually stand very close to each other and due to the high occlusions, head is the only visible part. The small size of the head makes the detection even much worse. Consequently, the current state-of-the-art region proposal methods are less effective and usually results in low Recall rates when applied to high density images. To address this problem, we propose different strategy for generating object proposals to capture range of scales for smaller objects as shown in Figure 2.

To generate object proposals, the first step is to estimate a scale map S for the input image I. In order to generate a scale map, we need to consider the effects of perspective. Inspired by [18], for each scene, we randomly select a group of adults between the two extremes (up and down) of the image and label their heads by drawing a straight line between the two points on the head. The line represents the size of the head as shown in Figure 2 (zoomed view) of portion of an image. We then approximate the scale map by linearly interpolating between two extremes of image. The scale map shown in Figure 2, where the red colors shows bigger size of head while smaller head sizes are highlighted in blue color. 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. For this, we overlaid a gird G of points on the image. Ideally, the resolution of the grid G and scale map S are the same as the resolution of the input image I. Let $S(p_i)$ represents the size of head (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. As we are interested in head detection, we keep the square-like aspect ratios $\Re \in [\frac{2}{3}, \frac{3}{2}]$ for all bounding boxes and refer them as candidates. From the Figure 2, it is obvious that the size of proposals in the down extreme of the image are bigger attributing to the bigger size of the head while the size of proposals becomes smaller as we move up.

3.2. Head Detection

Our head detector follows the model of R-CNN [19] and uses scale-aware proposals instead of selective search proposals to capture head with different scales. For each input proposal, we extend the bounding box with small margin to capture local image context around the head. The corresponding image patch is then re-sized to 224 x 224 pixels to fit the input layer of the CNN. Our R-CNN model is based on the architecture of Oquab et al. [19] extended by one fully-connected layer with 2048 nodes that are initialized randomly and followed by ReLu and DropOut. Oquab et al. [19] is initially trained on ImageNet and then fine tuned on images of human faces extracted from HollywoodHeads dataset [20]. In its original form, however, this head detector is of limited use to us since we are working with very high density images and in high density images human head barely spans a few pixels and face is almost unrecognizable. Therefore we fine-tuned the network on training images from WorldExpo'10 [6] data set.

To train the network, we use stochastic gradient descent (SGD) with momentum to optimize the parameters by minimizing the sum of independent log-losses. For training we extracted scale-aware patches that only contain one head and feed them to the network as positive sample. Negative sample were generated from the background and visible human torso, since we want our head detector to give high response only on heads and not on other body parts. Unlike traditional R-CNN that uses SVM for second pass of training, we use the output of CNN to score the proposals. After feeding all the proposal to the network, the output is the response map $M(p_i)$, where $M(p)$ is the score of the proposal and $p_i \in G$ is the location in the image. The higher values of the response map indicate the presence of head while lower values represent the background.

3.3. Localization

For the localization task, to get the precise location of the heads, we post-process the response map by finding local peaks/ maximums base 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. Therefore,to find an optimal strategy for localization is an important research direction.

4. EXPERIMENTS

We evaluate the performance of our proposed DISAM using two challenging crowd counting datasets, i.e, UCSD dataset [18] , and WorldExpo'10 [6] dataset. We use Mean Absolute Error (MAE) as an evaluation measure to compare the counting performance of the DISAM against the state-of-the-art methods and is defined as.

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

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 data set consists of 2000 frames of size 238 x 158 captured from a single camera. We split the dataset into training and testing images in the same way as [18]. On the other hand, WorldExpo'10 is a large scale dataset contains 1132 annotated video sequences captured by 108 surveillance cameras. The test set is 5 hour long video sequence from five different scenes with frame size of 576 x 720 pixels. We perform experiments on all five scenes of WorldExpo'10 data set. The quantitative results for both datasets are reported in Table 1. From the table, it is obvious that our DISAM outperforms other competitive methods with

Table 1. Comparative analysis with other methods in term of Mean Absolute Error (MAE) are presented considering UCSD [18] and WorldExpo'10 [6] datasets.

Methods	WorldExpo'10 [6]	UCSD [18]
Zang et al. [6]	12.9	1.60
M-CNN [5]	11.6	1.07
Kang et al. [9]	13.4	1.12
Switching CNN [7]	9.4	1.62
CP-CNN [8]	8.86	-
DecideNet [14]	9.23	-
Liping et al. [10]	-	1.03
Proposed DISAM	8.65	1.01

Table 2. 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		
	AvP	AvR	AUC	AvP	AvR	AUC
Zang et al. [6]	45.87	39.23	0.45	65.64	59.65	0.64
M-CNN [5]	55.24	52.28	0.51	69.74	65.67	0.71
Kang et al. [9]	42.98	39.27	0.41	67.28	55.32	0.67
Switching CNN [7]	58.29	45.39	0.54	63.87	55.63	0.61
CP-CNN [8]	63.65	58.67	0.61	60.42	49.82	0.58
DecideNet [14]	61.37	53.61	0.57	64.75	59.64	0.63
Liping et al. [10]	65.72	47.91	0.58	71.73	68.68	0.72
Proposed DISAM	69.46	67.65	0.69	73.58	71.68	0.74

lowest MAE of 1.01 and 8.65 for UCSD and WorldExpo'10 datasets, respectively.

We also quantify and compare the localization performance of our method with other state-of-the-art methods. 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 2. 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 WorldExpo'10 dataset as compared to UCSD dataset. This is due to the fact the WorldExpo'10 dataset contains more dense images with heavy occlusions as compared to UCSD. We also show some qualitative results of our proposed method in Fig. 3. The first row depicts the sample images from the UCSD dataset which represents low density scene and the second row depicts the sample images from two different scenes of WorldExpo'10 dataset representing relatively more complex and high density scenes. It is worth noticing that our method performs well in both high and low density scenes and it 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.

Fig. 3. Results of samples frames from UCSD and WorldExpo'10 data sets. 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.

5. CONCLUSION

This paper presented a novel DISAM to estimate the count by detecting and localizing the human heads in dense crowd scenes. To tackle the problem of scale variations, we generated scale-aware head region proposals by exploiting the perspective effects. This strategy has significantly reduced the classification time and also resulted in boosting the detection accuracy. We evaluated SADM on two datasets, i.e, UCSD, and WorldExpo'10 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).

6. ACKNOWLEDGMENT

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

7. REFERENCES

[1] Mohib Ullah, Habib Ullah, Nicola Conci, and Francesco GB De Natale, "Crowd behavior identification," in *International conference on image processing, IEEE ICIP*, 2016, pp. 1195–1199.

[2] Paolo Rota, Habib Ullah, Nicola Conci, Nicu Sebe, and Francesco GB De Natale, "Particles cross-influence for entity grouping," in *21st European signal processing conference, IEEE EUSIPCO*, 2013, pp. 1–5.

[3] Sultan Daud Khan and Habib Ullah, "A survey of advances in vision-based vehicle re-identification," *Journal of computer vision and image understanding, Elsevier CVIU*, 2019.

[4] Sultan D Khan, Stefania Bandini, Saleh Basalamah, and Giuseppe Vizzari, "Analyzing crowd behavior in naturalistic conditions: Identifying sources and sinks and characterizing main flows," *Journal of neurocomputing, Elsevier NC*, vol. 177, pp. 543–563, 2016.

[5] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Conference on computer vision and pattern recognition, IEEE CVPR*, 2016, pp. 589–597.

[6] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Conference on computer vision and pattern recognition, IEEE CVPR*, 2015, pp. 833–841.

[7] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu, "Switching convolutional neural network for crowd counting," in *Conference on computer vision and pattern recognition, IEEE CVPR*, 2017, vol. 1, p. 6.

[8] Vishwanath A Sindagi and Vishal M Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *International conference on computer vision, IEEE ICCV*, 2017, pp. 1879–1888.

[9] Di Kang, Zheng Ma, and Antoni B Chan, "Beyond counting: comparisons of density maps for crowd analysis tasks-counting, detection, and tracking," *Transactions on circuits and systems for video technology, IEEE TCSVT*, 2018.

[10] Liping Zhu, Chengyang Li, Zhongguo Yang, Kun Yuan, and Shang Wang, "Crowd density estimation based on classification activation map and patch density level," *Journal of neural computing and applications, Springer NCA*, pp. 1–12, 2019.

[11] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Conference on computer vision and pattern recognition, IEEE CVPR*, 2005, vol. 1, pp. 886–893.

[12] Mamoona Shami, Salman Maqbool, Hasan Sajid, Yasar Ayaz, and Sen-Ching Samson Cheung, "People counting in dense crowd images using sparse head detections," *Transactions on circuits and systems for video technology, IEEE TCSVT*, 2018.

[13] Muhammad Saqib, Sultan Daud Khan, Nabin Sharma, and Michael Blumenstein, "Person head detection in multiple scales using deep convolutional neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[14] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann, "Decidenet: Counting varying density crowds through attention guided detection and density estimation," in *Conference on computer vision and pattern recognition, IEEE CVPR*, 2018, pp. 5197–5206.

[15] Amir Ghodrati, Ali Diba, Marco Pedersoli, Tinne Tuytelaars, and Luc Van Gool, "Deepproposal: Hunting objects by cascading deep convolutional layers," in *International conference on computer vision, IEEE ICCV*, 2015, pp. 2578–2586.

[16] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov, "Scalable object detection using deep neural networks," in *Conference on computer vision and pattern recognition, IEEE CVPR*, 2014, pp. 2147–2154.

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision, Springer ECCV*, 2016, pp. 21–37.

[18] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Conference on computer vision and pattern recognition, IEEE CVPR 2008*, 2008, pp. 1–7.

[19] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Conference on computer vision and pattern recognition, IEEE CVPR*, 2014, pp. 1717–1724.

[20] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *Conference on computer vision and pattern recognition, IEEE CVPR*, 2008, pp. 1–8.