Crowd Counting and Density Estimation In High Density Crowds Using Convolutional Neural Network

Adwan Alownie Alanazi[1], Sultan Daud Khan[2]

[1]University of Ha'il, Saudi Arabia
[2]National University of Technology, Pakistan

Abstract

Crowd counting is an important task for crowd monitoring in Masjid Al-Haram, where millions of people gather every year to fulfil religious obligation. Several strides have been made to automatically estimate the density and count from images. However, it still remains a challenging task due to variations in view points, scales and illumination. In this paper, we propose a novel approach for the crowd counting based on Convolutional Neural Network (CNN). In this approach, we first divide the input image into non-overlapping blocks and then each block is further sub-divided into cells. For each cell, we extract corresponding patch in the image and then feed to CNN. We then train a binary CNN classifier, which classifies each patch into two classes, i.e, head or background. We evaluate our method on our own dataset which we collected from different location of Masjid Al-Haram. From the experiments, we show to achieve 90% accuracy. We compare our proposed method with other state-of-the-art methods and from the experimental results, we show that our proposed method outperforms other state-of-the-art methods

Key words:

Crowd detection; Fourier analysis; Crowd analysis

1. Introduction

Ensuring crowd safety and security in crowd of pedestrians is an important research topic. Crowd counting and density estimation is receiving much attention from research commu-nity since last few years [30], [33], [27]. Huge mass events are frequently occurring in urban areas, for example, reli-gious and political gathering, marathons, concerts, etc gather large number of people in a limited area. In most of cases, even peaceful gatherings end up with crowd disasters [4]. The responsibility of ensuring public safety and security fall on the shoulder of security personnel and crowd managers. Traditionally, for public surveillance, surveillance cameras are installed in different locations of the scene. Generally, security personnel in control room manually monitor different activities of the crowd. However, this traditional way of surveillance is a hectic job and prone to errors.

With the advancement of computer vision technology, several computed aided tools are developed that automatically analyse the crowd event and report suspicious activities [11]. With the tremendous success of convolutional neural net-work in the task of object detection, classification and image segmentation, researchers are utilizing convolutional neural network to explore the problem of density estimation and crowd counting.

Crowd density estimation and crowd counting has received tremendous importance from the computer science commu-nity. Crowd density estimation provides the distribution of pedestrians in the scene. Moreover, localization of people has numerous applications. (1) Localization provides information about the exact location of people in the scene which is important for crowd managers. (2) It can save the significant cost by deploying required number of security personnel. (3) It provides a semi-automated way of annotating human heads in high density crowds, since it is extremely hard and tedious job to manually annotate humans in high density crowds.

Several methods have been reported in recent years to estimate the crowd count from images. A most recent survey about the existing algorithms and techniques is reported in [23], [31], where the authors classified different algorithms based on their applicability and performance. They evaluated different methods using different challenging datasets. Most of existing methods are based on regression-based models, where the count is computed by just following the regression between the image features and crowd count. However, these regression-based methods have the following limitations. (1) These methods do not localize and detect human heads in crowd. (2) These methods are prone to errors when applied in low density situations.

Generally, regression-based methods [2], [24], [10], [32], [16], [17], [26] are more robust in estimating density in high density crowd. Since there is rich texture in high density crowds and regression-based methods fairly capture the regular patterns in crowd.

High density crowded scenes shows regular and repetitive structures in the form of textures. This notion is exploited by Marana et al. in [14], [15] where crowded scenes is classified into different categories based on the density level. In order to extract rich texture, Gray Level Dependence Matrix (GLDM) is used. After computing GLDM, features like energy, homo-geneity, entropy and

Manuscript received Fabruary 5, 2020
Manuscript revised Fabruary 20, 2020

contrast are computed which are later on used to train Support Vector Machine (SVM). Further-more, in [15], the author proposed a novel feature known as Minkowski fractal dimension for crowd density classification. The input image is first converted to the binary image and then apply different values of dilation in order to extract different morphological structures. These dilated structures are then mapped to the corresponding number of pixels by employing linear regression. After linear regression, the slope of the line is then used to used to calculate Minkowski fractal dimension features. Further studies observe that the values of Minkowski fractal dimension features increases with increase in crowd density. Wang et al. [28] extended the work and proposed an approach that divides the input image into different blocks. From each block both GLDM and LBP are extracted and con-catenated to represent a feature vector. These feature vectors are used to train SVM classifier. Xiaohua et al. [29] extract tex-ture features from the image using discrete wavelet transform. The authors argued that high density crowded scenes have high frequency of repetitive structures in contrast to low density crowds. Rahmalan et al. [20] proposed Translation Invariant Orthonormal Chebyshev Moments to extract texture from the crowded scenes. He argued that different moments represent different densities in the crowd.

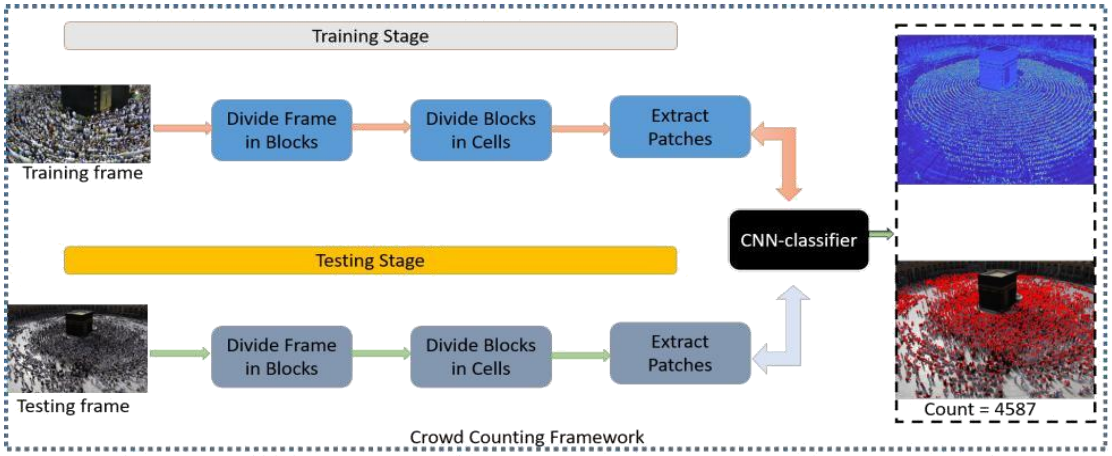

Fig. 1 Shows the pipeline of our proposed framework. Input is the image and output is the density map and count of people in the image.

In contrast to above texture-based features, Local Binary Pattern (LBP) is most commonly used feature and has been extensively employed for texture analysis. Most of existing crowd density estimation methods use histogram of LBP to classify density levels in crowd. Zhe et al. [28] proposed an approach that first compute LBP from input image and then employ GLCM on LBP image to compute contrast, entropy, homogeneity and energy features. Fradi et al. [6], [7] project high dimensional LBP features to low- dimensional feature space by using principle component analysis (PCA). In the same way, Jalali et al. [9] proposed an approach that compute LBP image from the original image and then apply Gabor filter to compute mean and standard deviation of the output.

The above-mentioned methods achieved success in estimating the crowd density, however, these models suffer from common limitations. (1) These methods works well in low density crowds while in high density crowds, these methods fail to provide the desired output. (2) These models are based on hand-crafted features, which were susceptible to significant changes in illumination and lighting conditions. A small change in illumination can significantly change the texture of the image. (3) These models are blind and can not localize pedestrians in the scene, since these methods are regression-based models.

Convolutional Neural Networks (CNNs) achieved significant success in object detection [22], [19], [3], [1], [21], classification [25], [8] and segmentation [13], [12] tasks. With the advent of CNN, research community is considering employing CNN for crowd counting and density estimation tasks. CNN takes a raw image and learn hierarchical features. The initial layers of CNN learn low-level features, like edges, and higher layers learn the context. Therefore, in this paper, we explore Convolutional neural network for crowd counting and density estimation tasks.

In this paper, we proposed Convolutional neural network-based framework for crowd counting and density estimation. For this purpose, our framework has the following pipeline.

- Divide input image into a set of non-overlapping blocks B. Then each block bi ∈ B is further subdivided into cells.

- Extract patch equal to the size of cell from the input image.
- Resize the patch in order to fit the input of convolu-tional neural network.
- We make a batch of 64 patches and feed forward to the CNN.
- The result is the score map, where high response represents the likelihood of head and low response represents the presence of background.
- Non-maximal suppression algorithm is applied on score map to detect human head in crowd. The overall pipeline of our framework is shown in Figure 1.
- Compare to the other state-of-the-art methods, our pro-posed methods have the following contribution
- Our proposed method use detection of human heads to count the number of people in high density crowds.
- Our proposed method localizes all pedestrians in the scene.
- In contrast to other methods, our proposed method solve crowd counting and density estimation problem simultaneously.
- We evaluated our approach on dataset collected from the videos of Masjid-al-Haram.

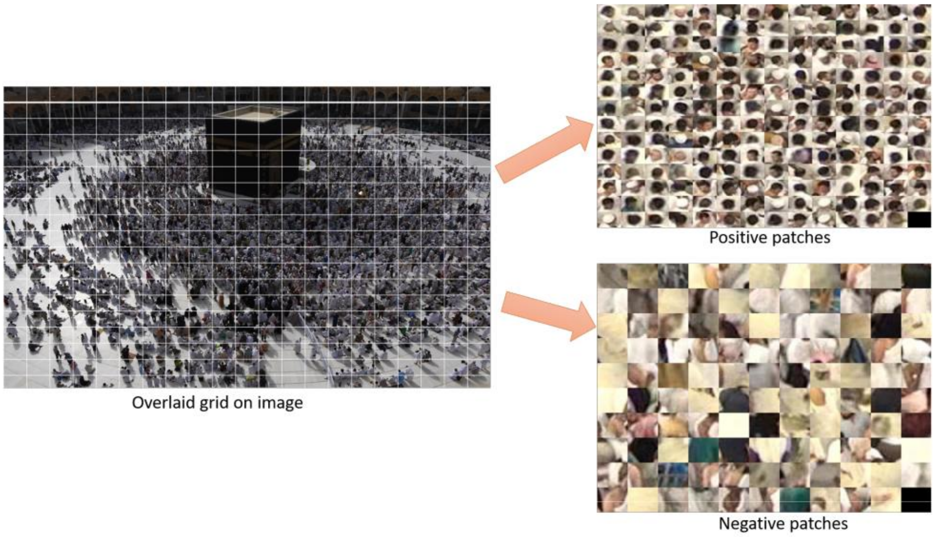

Fig. 2 shows the grid overlaid on image (on left). Right images show the extracted patch corresponding to the grid cells. These patches are used to train the network.

2. Proposed Methodology

The proposed framework takes an input image and estimate density and count as an output as shown in Figure 1. The input image is divided into blocks. Blocks are further subdivided into cells in order to capture the scale of head, since the head lies in multiple scales and the size of head is small. We then extract image patch corresponding to each cell and make a batch of

This batch of patches is then feed to CNN obtain detection map. We give details of each processing step as follows:

A. Extracting patches from image

For each input image, we overlay gird as shown in Figure 2, ideally equal to the size of input image. From the our empirical studies, we observe that higher the resolution of the grid, higher will be the accuracy but increase computational cost. While lower resolution of grid results in lower accuracy. So this creates a trade off between the accuracy and computational cost. In order to address this problem, we define a parameter

Ω which decide the resolution of the grid. In our experimental setup we keep the value of $\Omega = 0.5$. Let F_x and F_y represent the x and y direction of input image. The resolution of the resulting grid will be $G_x = \Omega F_x$ and $G_y = \Omega F_y$.

B. Convolutional Neural Network

Our proposed method use the pre-trained model of Oquab et al. [18] The model is initially trained on Imagenet dataset [5] and then fine-tunned on our own dataset for head detection. For generating object proposals, we follow the same pipeline of RCNN and adopt selective search strategy. We pre-process each object proposal before feeding to the network. In order to capture contextual information, we

extend the bounding box by a small margin. We then extract the patch from the image and resize the image to 224 x 224 to fit the input of CNN. Since we are interested in detecting heads, therefore we keep square like aspect ratios. From empirical studies, we observe that in high density crowds, head only covers few pixels therefore,
we keep the size of bounding box to a minimum value of 10 pixels.

For training, we optimize the network parameters by minimizing the loss function using stochastic gradient descent (SGD) with momentum. In contrast to classical RCNN that used CNN as a feature extract and employ SVM to train the classifier, we directly use CNN to score each object proposal. We observed that this approach works efficiently than traditional RCNN. After the classifier assign score to each object proposal, we then generate a score map. Score map is a heatmap, where hot color (red) represent high score value, where as cold values (blue) represent the background. The score map also contains noise and sparse values which are removed by employing Non-maximal suppression method. It is worth mentioning that the performance of our detector depends on the threshold value. The higher the threshold value, the high will be the localization accuracy but lower will be the counting accuracy. Empirically, we determined the value of threshold and fix its value to 0.5.

We observed that detection obtained using the above model are erroneous and need to be refined for accurate crowd counting. The lower accuracy of the detector attributes to significant variations in scales of human heads which is hard to be captured using selective search strategy. In order to solve the scale problem, we extend the above network to incorporate scale variations in the image. For this purpose, we use information at the image level to estimate the coarse scale of heads in the input image. For this reason, we divide the input image into grid of cells. In our experiments, we use four different resolutions i.e. 16 x 16, 32 x 32, 64 x 64, 128 x 128 with the stride of 50% of the resolution. We extract patches at multiple resolutions and resize those patches to fit the input of CNN. As a result, we obtained four different score maps corresponding that represents the coarse scale and location of head. We trained this network in the same way but with following loss function.

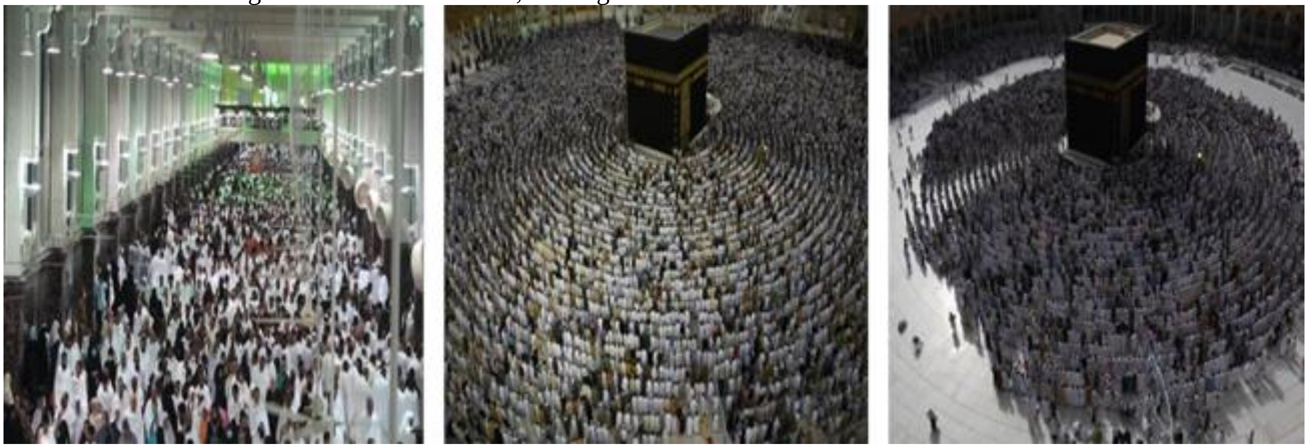

Fig. 3 Sample frames collected from different locations of Masjid-Al-Haram. Images shown in figure have different crowd densities, illuminations and camera viewpoints.

$$L(g_s(i), j_s) = \sum_{j\in\{0,1\}} \log(1 + exp^{((\_1)_{js+j+1} g_{s,}j(i))} \qquad (1)$$

where gs(i) represents the confidence of the grid cell s of input image i, js ∈ {0, 1} represents range of score that indicates the probability of head and background. The classifier will assign class head to grid cell, if the overlap between grid cell and ground truth is larger than 0.5, otherwise the cell is classified as background.

3. Experiments

For the experimental evaluation, we use Titan Xp GPU with 12 GB for training and testing. We implemented our proposed CNN using caffe library. We use images collected from different locations of Masjid-al-Haram. The sample of images are shown in Figure 3. We also use the same dataset for the evaluation and comparison with other state-of-the-art methods. Our dataset is composed of frames with resolution of 1080 x 920 but captured from different viewpoints. The images also depict different crowd densities at different timings.

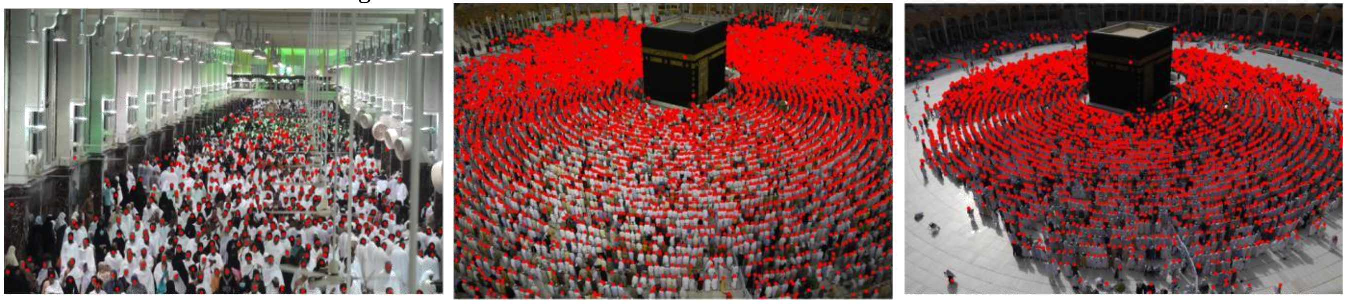

Fig. 4 Shows detection in different frames. The red point show the predicted detection. In first image (from left), ground truth=400 and predicted=375. In second image, ground truth=5590 and predicted==5572. In third image, ground truth=3300 and predicted=3252

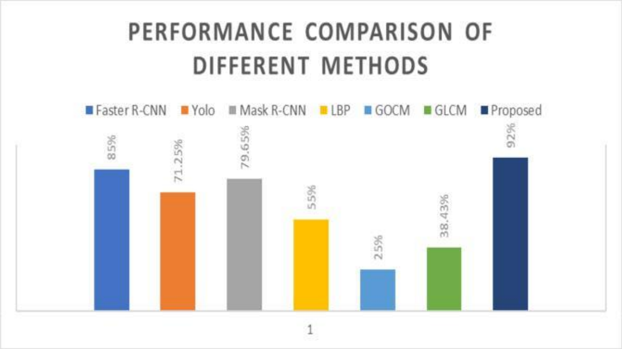

Fig. 5 Comparison of different state of the art methods Our propose method outperforms all existing methods

For counting heads in the input image, we first divide the input image into finite number of cells. Let frame I is divided into finite number of cells. Let C = {c1, c2.., cn} represent the set of cells. We then resize each cell bi ∈ B to fit the input of CNN. We then extract patches corresponding to cells and feed to the network to generate score map. We then employ Non-maxima suppression method to localize and detect heads. Figure 4 depicts the qualitative results. As obvious from Figure, our model precisely detect human head in high density crowds.

In order to demonstrate the effectiveness of proposed method, we compare our results with other existing methods in Figure 5. We use directly the pre-trained models of these methods in our evaluation. From the Figure, it is obvious that our proposed framework outperforms all state of-the-art methods.

The lower performance of other detector attribute to the smaller size of the human head that becomes challenging problem for other detector to detect human head smaller than 23 pixels. Perspective distortions naturally causes scale problem. Due to perspective distortions, the heads near to camera appear large while the head away from the camera appear small. Other detectors like R-CNN uses weak strategy for object proposal generation and do not take into account scale variations in the image. This strategy generally relies on hand-craft features, i.e, edges and can not generalize to different scenes. On the other hand, our proposed method handles these problems in an efficient and effective way. We argue that our method perform best than other state-of-the-art methods due to the adoption of scale-aware strategy.

Acknowledgment

This work is supported by Deanship of Research, University of Ha'il under the grant no: BA-1910. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU for this research.

References

[1] P. Ammirato and A. C. Berg. A mask-rcnn baseline for probabilistic object detection. arXiv preprint arXiv:1908.03621, 2019.

[2] L. Boominathan, S. S. Kruthiventi, and R. V. Babu. Crowdnet: A deep convolutional network for dense crowd counting. In Proceedings of the 2016 ACM on Multimedia Conference, pages 640–644. ACM, 2016.

[3] M. Cheng, J. Bai, L. Li, Q. Chen, X. Zhou, H. Zhang, and P. Zhang. Tiny-retinanet: a one-stage detector for real-time object detection. In Eleventh International Conference on Graphics and Image Processing (ICGIP 2019), volume 11373, page 113730R. International Society for Optics and Photonics, 2020.

[4] M. M. de Almeida and J. von Schreeb. Human stampedes: an updated review of current literature. Prehospital and disaster medicine, 34(1):82–88, 2019.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255. IEEE, 2009.

[6] H. Fradi and J.-L. Dugelay. A new multiclass svm algorithm and its application to crowd density analysis using lbp

features. In 2013 IEEE International Conference on Image Processing, pages 4554–4558. IEEE, 2013.

[7] H. Fradi, X. Zhao, and J.-L. Dugelay. Crowd density analysis using subspace learning on local binary pattern. In 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pages 1–6. IEEE, 2013.

[8] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger. Multi-scale dense networks for resource efficient image classification. arXiv preprint arXiv:1703.09844, 2017.

[9] M. Jalali Moghaddam, E. Shaabani, and R. Safabakhsh. Crowd density estimation for outdoor environments. In Proceedings of the 8th Inter-national Conference on Bioinspired Information and Communications Technologies, pages 306–310. ICST (Institute for Computer Sciences, Social-Informatics and . . . , 2014.

[10] D. Kang, D. Dhar, and A. B. Chan. Crowd counting by adapting convolutional neural networks with side information. arXiv preprint arXiv:1611.06748, 2016.

[11] S. D. Khan, M. Tayyab, M. K. Amin, A. Nour, A. Basalamah, S. Basalamah, and S. A. Khan Towards a crowd analytic frame arXiv:1709.05952, 2017

[12] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng. H-denseunet hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging, 37(12):2663–2674 2018

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3431–3440, 2015

[14] A. Marana, L. d. F. Costa, R. Lotufo, and S. Velastin. On the efficacy of texture analysis for crowd monitoring. In Proceedings SIBGRAPI'98 International Symposium on Computer Graphics, Image Processing and Vision (Cat. No. 98EX237), pages 354–361. IEEE, 1998.

[15] A. N. Marana, L. D. F. Costa, R. Lotufo, and S. A. Velastin. Estimating crowd density with minkowski fractal dimension In 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings. ICASSP99 (Cat. No. 99CH36258), volume 6, pages 3521–3524. IEEE, 1999

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

[17] D. Onoro-Rubio and R. J. Lopez´-Sastre. Towards perspective-free object counting with deep learning. In European Conference on Computer Vision, pages 615–629. Springer, 2016

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1717–1724, 2014

[19] M. Price, J. Hoover, G. Dagley, S. Wylie, and Q. Tang. Object detection using image classification models, Mar. 5 2019. US Patent 10,223,611

[20] H. Rahmalan, M. S. Nixon, and J. N. Carter. On crowd density estimation for surveillance. 2006

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788

[22] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015

[23] S. A. M. Saleh, S. A. Suandi, and H. Ibrahim. Recent survey on crowd density estimation and counting for visual surveillance. Engineering information processing systems, pages 91–99, 2015

[24] D. B. Sam, S. Surya, and R. V. Babu. Switching convolutional neuralnetwork for crowd counting. In Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition, volume 1, page 6, 2017.

[25] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng.Convolutional-recursive deep learning for 3d object classification. InAdvances in neural information processing systems, pages 656–664,2012.

[26] E. Walach and L. Wolf. Learning to count with cnnboosting. In European Conference on Computer Vision, pages 660–676. Springer,2016.

[27] J. Wan, W. Luo, B. Wu, A. B. Chan, and W. Liu. Residual regressionwith semantic prior for crowd counting. In Proceedings of the IEEEConference on Computer Vision and Pattern Recognition, pages 4036–4045, 2019.

[28] Z. Wang, H. Liu, Y. Qian, and T. Xu. Crowd density estimation based onlocal binary pattern co-occurrence matrix. In 2012 IEEE InternationalConference on Multimedia and Expo Workshops, pages 372–377. IEEE,2012.

[29] L. Xiaohua, S. Lansun, and L. Huanqin. Estimation of crowd densitybased on wavelet and support vector machine.Transactions of the Institute of Measurement and Control, 28(3):299–308, 2006.

[30] X. Zeng, Y. Wu, S. Hu, R. Wang, and Y. Ye. Dspnet: Deep scale purifiernetwork for dense crowd counting. Expert Systems with Applications,141:112977, 2020.

[31] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q.Xu. Crowd analysis: a survey. Machine Vision and Applications, 19(5-

[32] 6):345–357, 2008. S. Chen, S. Gao, and Y. Ma.Single-imageY. Zhang, D. Zhou, crowd counting viamulti-column convolutional neural network. InProceedings of the IEEE conference on computer vision and patternrecognition, pages 589–597, 2016.

[33] Y. Zhou, J. Yang, H. Li, T. Cao, and S. Kung. Adversarial learning for