Benchmarking Detection and Regression based Crowd-Counting Techniques

Marko Brodarič
Faculty of Electrical engineering University of Ljubljana
Tržaška cesta 25, 1000 Ljubljana, Slovenija
mb4646@student.uni-lj.si

Abstract

Crowd counting is a computer vision problem, where algorithms are designed to estimate the count of people in crowded scenes, for purposes of the crowd monitoring in public gatherings, for statistics or to maintain adequate crowd density on public places during health crisis. Many approaches were proposed in past decades, addressing these problems, starting from traditional detection methods, and developing to modern regression methods. However, it is not known, how these methods perform on images with different characteristics and what are their advantages and disadvantages. To address this gap, a study of one detection and one regression based model, in different crowd densities, in presence of distractors and face visibilities is done in this paper. It is shown, that modern regression model generally outperforms traditional detection method, but in some scenarios like smaller crowd counts, presence of distractors or crowd with visible faces, detection method outperforms modern regression model.

1. Introduction

Crowd counting is a computer vision problem, where algorithms are designed to estimate the count of people in crowded scenes from images. Many methods were developed in recent years, addressing this problem, due to important applications in crowd monitoring on social events, such as musical concerts and protests. Using this, trampling and suffocation in public gatherings can be prevented, and it can also simplify ensuring adequate crowd density on public places during a pandemic [17].

Methods, developed in this field, can be divided into two groups: detection and regression. Detection approaches are based on object detection in images, where objects of interest are often the faces, the whole body or combination of different body parts. The information about detections is then used to estimate the crowd count, typically by counting detections or by averaging detection numbers of subregions. Regression methods, on the other hand, perform the mapping from input features of the whole picture, to estimate the crowd count, where input in regressor can be a set of low level features [15] or a whole image, where a density map is first calculated and regression is then performed to estimate the number of people in the crowd [12].

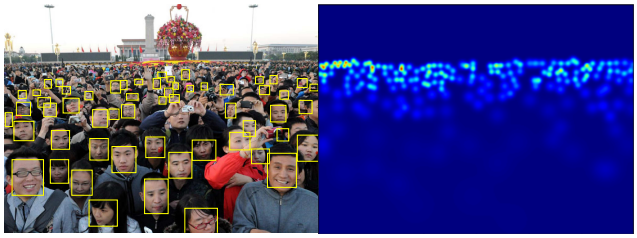

Figure 1. Performance evaluation of detection and regression based crowd-counting techniques. One face detection (as shown on the left image) and one density map regression model (as displayed on the right image) were evaluated on publicly available JHU-CROWD++ dataset. The paper analyses the impact of the presence of distractors in images, different face visibilities and evaluates performance of models on different crowd densities.

Even though a lot of research had been done in this field, it is not known how detection and regression methods perform on different crowd densities, under circumstances where faces are visible and on images, where distractors (objects, that could mislead the model into counting it as a person) are present. In this paper, we try to address this gap by evaluating the performance of one detection and one regression based crowd counting model on challenging dataset. This study analyses the overall performance of models, their robustness to presence of distractors in images, performance on images with visible faces and on different crowd densities. The results show, performance of detection model significantly improves in case of images with well visible faces. Even though regression model highly outperforms detection mode on whole dataset, the latter yields better results on lower crowd densities. In case of distractors, both models perform approximately the same.

The main contributions of the paper are:

- An evaluation of one detection and one regression based crowd counting model on a challenging dataset.
- An analysis of the impact of presence of distractors in images on crowd counting performance.
- A study of influence of face visibility on performance and discussion on reasons for performance differences.
- A performance evaluation of models on different crowd densities.

The rest of the paper is structured as follows. In Section 2, an overview of the existing crowd counting methods is made, in Section 3, dataset and methodology for models evaluation is presented. Section 4 provides experimental results and presents main findings. Finally, in Section 5, the conclusions of this study are discussed.

2. Related Work

In the past decades, a significant amount of work has been published, proposing various approaches to cope with the problem of the crowd counting. In this section, a brief overview of this field is given, firstly, focusing on the detection based methods, then, on the regression based methods.

Detection-based methods. Crowd counting methods were primarily based on the object detection and detection counting. Features for classificator training can be derived based on the whole human body (monolithic detection approach), or from the specific body parts, such as the head and the shoulders (part-based detection approach). Papageorgiou and Poggio [14] present a pedestrian detector based on Haar wavelets as descriptors, and Support Vector Machine (SVM) as the baseline classifier. In [6], Depoortere *et al.* introduced an extended and optimized version of that method. Dalal and Triggs [5] suggested a monolithic detection approach using locally normalized Histogram of Oriented Gradient (HOG) descriptors as feature set, which provided an improvement in performance relative to wavelet descriptors feature set. Promising results on challenging datasets were also achieved with covariance matrices used as object descriptors and classification of points lying on Riemannian manifolds, as presented by Tuzel *et al.* [19]. Liebe *et al.* [10] claimed, the pedestrian detection is a too difficult task for a single model or feature. They proposed several stage aggregating principle, where they take into account local and global features.

Although monolithic detection approach methods are successful in cases of low density crowds, major occlusions and higher density crowds pose problems. These are partly addressed with part-based detection approach methods. Lin *et al.* [13] proposed a procedure where they combined Haar wavelet transform to extract the area of the head-like contour, and SVM to classify whether this area is the contour of a head or not. Wu and Nevatia [20] represented a human body as an assembly of body parts. They used AdaBoost algorithm with edgelet features to train different part detectors. The outputs of these detectors are then combined to form a likelihood model. Li *et al.* [11] also proposed part-based detection approach, where they used Mosaic Image Difference (MID) foreground segmentation algorithm to detect active areas, and HOG based head and shoulder detection algorithm.

Regression-based methods. Even though part-based detection approaches reduce some issues detection-based methods have, performance of these approaches falls when faced with extremely dense crowds. To solve this problem, methods, that base on counting by regression, were proposed. Chan *et al.* [1] presented a technique, where a crowd is firstly segmented into smaller components of homogeneous motion, then a set of features is extracted from each component and finally, a crowd count for each segment is estimated using a Gaussian process regression function, which maps feature sets to the crowd count. In [2] the idea was refined, removing the deficiencies of Gaussian process regression, by switching to a Bayesian Poisson regression. A similar nonholistic idea was described by Ryan *et al.* [15]. A similar approach was improved by Kong *et al.* in [9] by using normalized feature histograms to achieve a viewpoint invariant model. In [4], Chen *et al.* proposed a model, that replaces a large number of regressors, needed in the localized density estimation, with a single regression model that automatically learns the mapping between a feature set and a multidimensional output. Further the problem of sparse and imbalanced data, which affects the mapping between a feature vector and a crowd count, was addressed by Chen *et al.* in [3], by introducing a cumulative attribute concept. Though mentioned regression methods improved the week points of detection-based ones, they are not reliable in case of extremely dense crowds, due to low resolution, perspective, foreshortening etc. Idrees *et al.* [8] treated dense crowds as a texture, and therefore they applied the Fourier analysis, a head detection and an interest-point detection (SIFT) in the local neighbourhoods, which are then combined, in the respect to their confidences, in the local crowd counts. The global crowd count is then estimated using the local counts and a multiscale Markov Random Field (MRF).

Although, regression based methods outperform detection based approaches in challenging conditions, both approaches fail in some degree of scene difficulty (e.g., high crowd density, bad lighting, severe occlusions etc.). This paper evaluates the performance of one detection based and one regression based model in different scenarios.

(a) (b) (c) (d)

Figure 2. Example images from JHU-CROWD++. The dataset contains low density images (a) and high crowd count images (b), in different weather conditions (example of low visibility due to fog in (c)), and an instance of distractor image (d)

3. Methodology

In this section we describe models and review the characteristic of datasets used in this paper, we discuss the experimental setup and introduce the performance metrics utilized in the evaluation.

3.1. Crowd Counting Models

Two crowd counting models are evaluated in this paper, representing the detection and the regression based approaches. Details on the selected models are provided below.

- **Haar and HOG Based Head Detection.** The first model utilizes the head detection and estimates the crowd counts by counting the detections, as proposed by Devireddy in [7]. The Head detector is constructed using two classifiers that are trained with two different features, the Haar wavelets and the Histogram of Oriented Gradients. Both are trained by AdaBoost learning algorithm. The head regions detected are then tracked using the Kanade-Lucas-Tomasi feature tracker. Each detected head is identified by a serial number, which increases for each detected head. Hence, the highest number is the total head count, which is the estimation of crowd count.

- **CSRNet.** The second model selected for the study is Congested Scene Recognition Network (CSRNet) proposed by Li *et al.* [12]. It is composed of a front-end and a back-end. The first part deploys the first 10 layers from VGG-16 [16] (fully-connected layers are removed). The output size of the front-end network is 1/8 of the original input size. The back-end also adapts the first 10 layers of VGG-16 network with only three pooling layers instead of five. Using the bilinear interpolation with the scaling factor of 8, the output density maps resize to the same resolution as the input image. To train the model, the ground-truth dot annotations are blurred using a normalized Gaussian kernel, using geometry-adaptive kernels. A Euclidean distance between a ground-truth and an estimated density map is used as a loss function.

3.2. Datasets

Models we evaluate in this paper are pre-trained: Haar and HOG based head detector is trained on the dataset of positives and negatives of human head as described in [7]. Regression based CSRNet model is trained on the Shanghaitech dataset [21].

Evaluation is performed on JHU-CROWD++ [18], which consists of 4372 images, crawled from the internet, and the respective annotations (all three parts of set were used: train, test and validation). The average crowd count is 346, while the biggest count reaches 25791 people. The distribution of the different crowd counts across the dataset is shown in Figure 3. Annotations are divided into the two groups: the image-wise, which provides information about the image crowd count, the scene type, the weather condition and the distractor presence, and the head-wise, which marks the head position, its width and height, the occlusion level and whether it is blurred or not.

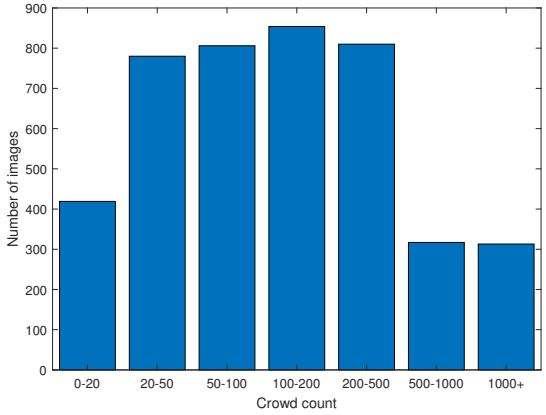

Figure 3. Number of images for different crowd count groups across the dataset

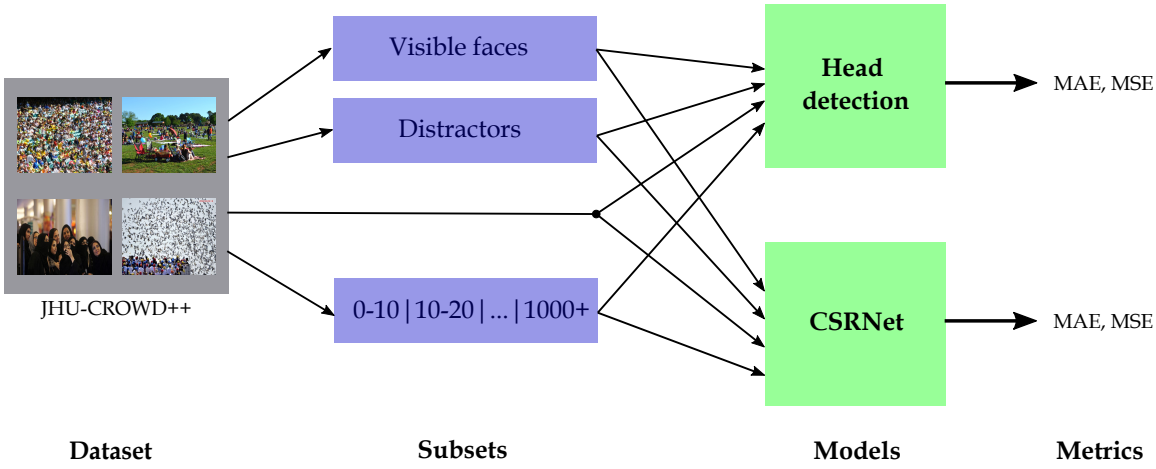

Figure 4. The chosen dataset is divided into the smaller subsets, then, evaluated models are tested. Finally, the performance metrics are compared.

3.3. Experimental setup

All images from the database were firstly joined together in one set, on which the evaluation is performed. Because of hardware limitations, images with width or height above 2000 pixels are not used, leaving 3532 images in the evaluation dataset. Some pictures are annotated as the distractors. These are images, with the count close or equal to 0, but their content resemble the crowd-like pattern which can be easily distinguished by the human eye, but it could pose a problem to computer algorithms. For our experiments, images are also divided into different subsets: a set of 64 images containing the distractors extracted using the dataset annotations, subsets of images with different crowd count also extracted using ground-truth annotation and a subset containing 200 images with well visible faces which we hand-crafted for this study.

Our goal is to observe chosen performance metrics of two crowd counting methods on mentioned image database. Using annotations, overall performance is measured as well as performance in different conditions, such as various crowd counts and several distractor images.

3.4. Performance metrics

Crowd counting methods are evaluated utilizing two widely used performance measures: Mean Absolute Error (MAE) and Mean Squared Error (MSE). These are defined as

$$MAE = \frac{1}{K}\sum_{k=1}^{K}|N_k - C_k| \quad (1)$$

$$MSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(N_k - C_k)^2} \quad (2)$$

Table 1. MAE and MSE values for both models evaluated on the whole dataset, on the subset with visible faces and on subset with the distractor images.

Image set	Overall		Faces		Distractors	
Model	MAE	MSE	MAE	MSE	MAE	MSE
Detection	334,3	1004,9	**42,4**	**84,3**	**141,0**	**278,5**
CSRNet	**95,2**	**236,26**	46,3	70,8	176,7	224,7

where K is the number of evaluation images, N_k is the ground-truth count of the k-th image and C_k is the crowd count of the k-th image, predicted by the model. Both metrics produce smaller values for better estimates, but MSE penalize bigger errors more than MAE.

4. Experimental results

In the following section, the analysis of the experimental results is presented.

Overall results. MAE and MSE values for both models can be seen in Table 1. The experiment shows that the CSRNet outperforms the detection model tested on the whole database. The reason for the relatively poor performance of the detection model is the difficulty of the dataset. JHU-CROWD++ is a challenging dataset with the high crowd densities where faces are blurred and occluded. Detection method, which counts detected human heads, fail in such circumstances since it does not detect any human head. CSRNet on the other hand performs regression from the input image to the crowd count through the density map. This kind of mapping of the input features to the crowd count estimates reduces the impact of higher densities and occlusions to some degree.

(a) $y_{GT} = 2$, $y_{det} = 2$, $y_{CSRNet} = 611$

(b) $y_{GT} = 254$, $y_{det} = 3$, $y_{CSRNet} = 255$

(c) $y_{GT} = 0$, $y_{det} = 0$, $y_{CSRNet} = 255$

(d) $y_{GT} = 38$, $y_{det} = 6$, $y_{CSRNet} = 607$

Figure 5. Example of images with the distractors. In (a), the detection model detected two heads, and therefore it estimated the crowd count to 2. CSRNet counted all the objects around as people as well, and so it estimated the crowd count at 611. On the other hand, in the image (b), objects in the sky did not distract the regression model, and it estimated the count to 255, while detection method fails to detect faces, because of the high level of blur and occlusion. In (c) the detection model did not detect faces and so it correctly estimated 0 crowd count, while CSRNet did fail to recognize the distractor image. Last image (d) shows a very difficult scene, and so both, the detection and the regression models, fail to correctly estimate the crowd count.

(a) $y_{GT} = 2$, $y_{det} = 2$, $y_{CSRNet} = 0$

(b) $y_{GT} = 15$, $y_{det} = 13$, $y_{CSRNet} = 222$

(c) $y_{GT} = 323$, $y_{det} = 46$, $y_{CSRNet} = 336$

(d) $y_{GT} = 173$, $y_{det} = 134$, $y_{CSRNet} = 174$

Figure 6. Example of images with well visible faces. In (a), the detection model detects and counts two people, while the regression model counts zero, since the density map does not represent the two people in the picture. In (b), the detection method does not have major problems, since faces are big and visible. On the other hand, regression model maps the input picture to relatively big crowd count, because density map, represented single head as high density and therefore estimated count is too high. Picture (c) represents a crowd with higher density, but visible faces. The regression model estimates the crowd count very close to the ground truth, while the detection model does struggle with the smaller faces further in the background of the crowd. In the last image (d), both methods perform successfully, since faces are relatively big and visible and the pattern still resembles a crowd, so the density map properly represents the crowd density.

Impact of the presence of the distractors. Experiments are also conducted to evaluate the performance of the two crowd counting models in the presence of the distractors in images. The results for this part are also shown in Table 1. The detection method shows improved performance compared to overall results, and it even scores slightly better MAE score than the CSRNet method, which scored higher MAE scores than on the overall performance. The detection method improved its scores because the crowd like patterns, that are represented in the distractor images, does not resemble human faces and so, the detector does not falsely detect these objects as human heads. Though the detection method does not falsely detect distractor object, it still struggles to score better MAE and MSE scores. The reason for this is the size of true positive human heads in the distractor images. In many of these pictures, the human faces and heads are still very small, and therefore blurred or occluded, which results in the detector failing to detect them. Overall, the presence of distractors does not worsen the performance of the detection method. On the other hand, CSRNet performs worse in the presence of distractors, according to MAE score in comparison to the overall performance. This is the result of training on a dataset with not a sufficient amount of distractors, and therefore the model fails when faced with patterns on images that resembles crowds. Examples of the distractor images, where models succeed and fail are shown in Figure 5.

Impact of faces visibility. The base of evaluated detection method is counting the detected faces. The experi-

ment is designed to test both methods on a subset of images, where faces are well visible in various crowd densities and scene types. The MAE and MSE values are shown in Table 1. The results show drastic improvement in the performance of detection model. The outcome is rather expected, since the method detects faces and counts them. The problem poses smaller, blurred faces, sometimes occluded with headgear, shadows or masks, which are often overlooked by the face detector and therefore the crowd count estimate deviate from the ground truth value. Overall, results are improved in comparison to performance on the whole dataset, and it slightly outperforms the CSRNet, which also reported improved results. Main problems posed to the regression model are images, that does not represent a crowd per se, but rather a smaller group of individuals. In these cases, density maps often represent the high density on a person's head, which results in a higher crowd count on one person, or does not even represent a human head and therefore crowd count on that head is zero. Examples of images with visible faces, where models succeed and fail are shown in Figure 6.

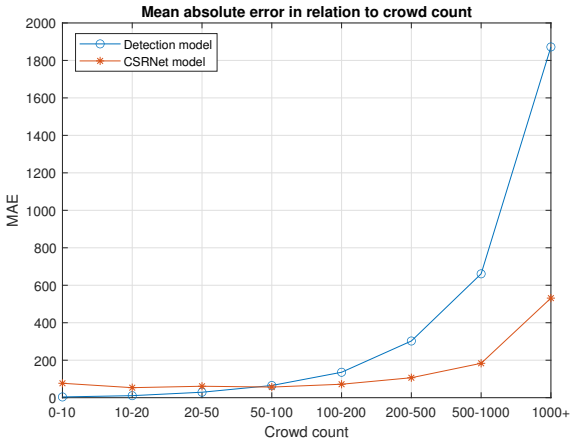

Figure 7. Mean absolute error in relation to the crowd count of both models

Impact of the crowd count. Figures 7 and 8 show the relation between the mean absolute error and the mean square error to the crowd count. Results show, that the detection method outperforms the regression model in the crowd counts under 100. In very low densities, faces are big and visible, so the face detector counts people with ease, while on the other hand, the regression model faces problems with density maps, since in the lower crowd counts, maps can show zero or too many people. As the crow count increases, faces become smaller, pictures become occluded and the advantage of the regression comes to the fore. The estimation error of the detection model shows exponential growth, and it outgrows the error of CSRNet at around 100 people. Even

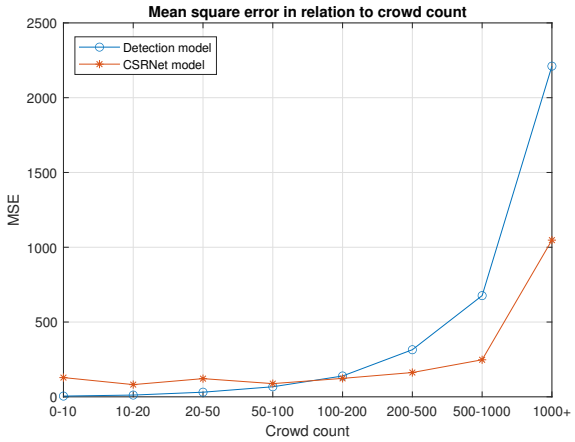

Figure 8. Mean square error in relation to the crowd count of both models

though the regression model performance at lower densities is worse than the detection, it successfully holds relatively low errors until the crowd counts of 1000, before it starts the exponential growth. For instance, at the crowd count 1000, the mean absolute error of the detection model is about three times bigger than the regression absolute count error.

Computational complexity. The evaluation of the CSRNet model was performed on the Google Colab NVIDIA Tesla K80 GPU, which provides 12 GB of video RAM. The evaluation takes approximately an hour on the whole dataset. The detection method was evaluated on the Intel® Core™ i5-7200U CPU, which takes about 45 minutes for the whole dataset.

5. Conclusion

In this paper, the performance of the two crowd counting models is analyzed, using one crowd counting dataset. The results showed that while the regression model outperforms the detection model on the whole dataset, some scenarios still pose problems to the regression method, and therefore the detection model, utilizing a different approach, outperformed the regression model in some scenarios.

It is shown that the detection method is more robust to the presence of the distractors in images, while the regression model should be carefully trained to solve this problem. On the other hand, the detection model shows high dependence on the face visibility. CSRNet does struggle, when it comes to very small groups of people, since generated density maps often falsely estimate a crowd density when it comes to a few people. Finally, we perform the comparison of performances on the different crowd counts. The results shows, that head detection method outperforms the regression model on a lower crowd densities, but the error increases exponentially when the crowd count grows.

The regression method does perform worse at lower counts, but it successfully holds low errors up to a crowd count of 1000. After that, exponential growth of an error can be seen.

Even though the regression model generally outperforms the detection model, traditional method still reports better results in specific scenarios. In the future work, training of the regression model on the carefully composed dataset, to improve the performance on weak spots, exposed in this study, should be made, considering the increasing demand for crowd monitoring during the concerts, protests or health crisis like a pandemic.

References

[1] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008.

[2] A. B. Chan and N. Vasconcelos. Bayesian poisson regression for crowd counting. In *2009 IEEE 12th international conference on computer vision*, pages 545–551. IEEE, 2009.

[3] K. Chen, S. Gong, T. Xiang, and C. Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2467–2474, 2013.

[4] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *Bmvc*, volume 1, page 3, 2012.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[6] V. De Poortere, J. Cant, B. Van den Bosch, J. De Prins, F. Fransens, and L. Van Gool. Efficient pedestrian detection: a test case for svm based categorization. In *Workshop on Cognitive Vision*, volume 1, pages 19–20, 2002.

[7] P. Devireddy. Persons counting by head detection in real time. `https://github.com/Pramod-Devireddy/head_detection`, 2019.

[8] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013.

[9] D. Kong, D. Gray, and H. Tao. A viewpoint invariant approach for crowd counting. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 1187–1190. IEEE, 2006.

[10] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 878–885. IEEE, 2005.

[11] M. Li, Z. Zhang, K. Huang, and T. Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *2008 19th international conference on pattern recognition*, pages 1–4. IEEE, 2008.

[12] Y. Li, X. Zhang, and D. Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

[13] S.-F. Lin, J.-Y. Chen, and H.-X. Chao. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(6):645–654, 2001.

[14] C. Papageorgiou and T. Poggio. A trainable system for object detection. *International journal of computer vision*, 38(1):15–33, 2000.

[15] D. Ryan, S. Denman, C. Fookes, and S. Sridharan. Crowd counting using multiple local features. In *2009 Digital Image Computing: Techniques and Applications*, pages 81–88. IEEE, 2009.

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

[17] V. A. Sindagi and V. M. Patel. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 107:3–16, May 2018.

[18] V. A. Sindagi, R. Yasarla, and V. M. Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *Technical Report*, 2020.

[19] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1713–1727, 2008.

[20] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007.

[21] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016.