Single-Image Crowd Counting using Multi-Column Neural Network

Rinku Mahesh Sharma
PG Student
SSVPS's B.S.Deore College of Engineering
Dhule, 424005, India

ABSTRACT

Crowd scene understanding is an important and challenging problem in computer vision. Most of studies based on tracking individuals, crowd counting, finding region of motion and alarming crowd flaws have came into existence. The task of crowd counting and density map estimation is riddled with many challenges such as occlusions, non-uniform density, intra-scene and inter-scene variations in scale and perspective. Nevertheless, over the last few years, crowd count analysis has evolved from earlier methods that are often limited to small variations in crowd density and scales to the current state-of-the-art methods that have developed the ability to perform successfully on a wide range of scenarios. The success of crowd counting methods in the recent years can be largely attributed to deep learning and publications of challenging datasets. One of the appropriate method that can accurately estimate the crowd count from an image with arbitrary crowd density and arbitrary perspective is using the state-of-the-art i.e. convolution neural network. The technique used for the crowd detection and crowd density estimation is through the Multicolumn Convolution Neural Network architecture. The model allows the input image to be of arbitrary size or resolution with high accuracy and produces a state of art results. The proposed work is implemented with the Shanghaitech dataset, which is among the largest dataset. The model produces highly precise and accurate results with the estimate crowd count and density map.

General Terms

Computer Vision, Neural network, Crowd detection, Crowd density.

Keywords

Deep learning, Convolution neural network, Crowd density, Multicolumn Convolution Neural Network.

1. INTRODUCTION

Human population is one of the most ever growing thing which is happening at a very high phase. As a result, this growth has indirectly increased the incidence of the crowd. The purpose of the gatherings has an important effect on large-scale assets and crowded behavior [1]. Most of the studies in crowd detection are based on tracking individuals, crowd estimation and finding the region of motion. Therefore, the analysis of mobility and behavior of the crowd is one of the greatest interests in many scientific kinds of research in public service, security, and safety and computer vision. There are large crowds of confusion, resulting from pushing, mass panic, and stampede or crowd crushes [5] and causing an overall loss of control. In recent years many massive stampedes have taken place around the world claiming many victims.

To prevent these fatalities, early automatic detection of critical and un-usual situations in the large-scale crowd is necessary. This would certainly help, as result, to make an appropriate decision for emergency control and security Crowd analysis can be used for detecting critical crowd levels, detecting and counting of people and also for detecting anomalies in crowded scenes [7]. Moreover, it can be used for tracking individuals or group of people in crowds. Crowd analysis is one of the most intelligent tasks in such intelligent visual surveillance systems. It can be used for automatic detection of critical crowd level, detection and count people, and also to detect anomalies and alarming faults of the crowd.

During the recent years convolution neural networks have been proving the most promising research directions in computer providing significant results. The CNN have been successfully applied in object detection, image classification, face recognization, image retrieval, digit recognization, pose estimation, pedestrian detection and scene recognization [1]. The major reasons underlying their success lie in the Graphics Processing unit computational power and affordability as well as in the availability of large annotated datasets.

2. RELATED WORK.

The purpose of a crowd counting is to count the number of people in the crowded places where the density estimates represent the number of people per pixel on the map by placing crowded images on your density map. Crowd analysis has a great interest in a large number of critical applications. Most of the previous researches are focused on detection style framework. M.J.Jones and Snow [2] used the method of sliding window detector to detect the people in the given particular scenes of using a scanning window type pedestrian detector using spatiotemporal information. This method with holds the advantage of both the motion and the appearance information to classify the moving object.

B.Leibe [14] used the top-down segmentation for detection of the pedestrian in the crowded scenes. They combined the local and the global features in a potential segmentation. They soon concluded that it was totally [13] system dependent and variations in results during overlapping were also observed. Zhao [11] also got complex occlusion with a huge crowd or overlapping .the next approach was using regression technique for counting using mapping between features extracted from the image patches to their counts. M.Shah [7] and Idrees explained that a single feature and detection method is not capable to provide sufficient information to accurately calculate the presence of high density problem.

The success of CNN in computer vision tasks has inspired researchers to exploit their abilities for learning non-linear functions from crowd images to their corresponding density maps or corresponding counts. C.Wang and M.fu [22] were the very first researchers to apply the technique of CNNs for the task of crowd density estimation. Wang [24] proposed end-to-end deep CNN regression model for counting people from images in extremely dense crowd. Y.Zhang and Y.Ma [25] proposed a simple but effective algorithm using Multi-column Convolutional Neural Network (MCCNN) architecture. It allows an image to be arbitrary size or resolution. Here the model uses different filters for each column of different sizes to model the density maps corresponding to heads of different scales. The issue of occlusion while counting people in a clustered environment or in a very dense crowd may happen.

L.Zeng and T.Zhang [26] used an multi scale CNN for single image crowd counting. It extracts scale relevant features from crowd images using a single column network based on the multi-scale blob. A.Bansal [28] proposed the model which used multiple features such as Scale-Invariant Feature Transform (SIFT), Fourier analysis, wavelet decomposition, Gray-Level Co-occurrence Matrix (GLCM) features and low confidence head detection to estimate the counts.

3. MCNN MODEL FOR CROWD DETECTION AND CROWD ANALYSIS.

CNN has become a most powerful subject for researchers and engineers in the area of Computer Vision and Pattern Recognition [1]. It is due to the ability of CNN to learn features from data, and use them for tasks such as image classification, object recognition, and object detection. Here we propose a Multi-column convolution neural network (MCNN) for crowd detection and the crowd density estimation. The proposed method ensures robustness to large variation in object scales. The proposed system uses density map based technique to have more accuracy on crowded images. The crowd analysis is estimated using two methods either by network or through a density map. The first method of using network gives output of the estimated head count with the provided input image. The second method is by using the density map of the crowd. By using the second technique we can estimate the number of people per square meter and then obtain the head count by integration.

3.1 MCNN Architecture

The proposed architecture utilizes the basic idea of convolution neural network (CNN) to implement multicolumn convolution neural network (MCNN). As the fact is clear that the images usually contain heads of very different sizes, so the filters used are of different sizes receptive with to the field to learn the map from the raw pixels to the density maps .The Multi-column CNN (MCNN) is implemented to learn the target density maps. In the architecture of MCNN, the filters of different sizes are used to model the density maps corresponding to heads of different scales for each column. The following figure shows the actual architecture of the model and it is explained further.

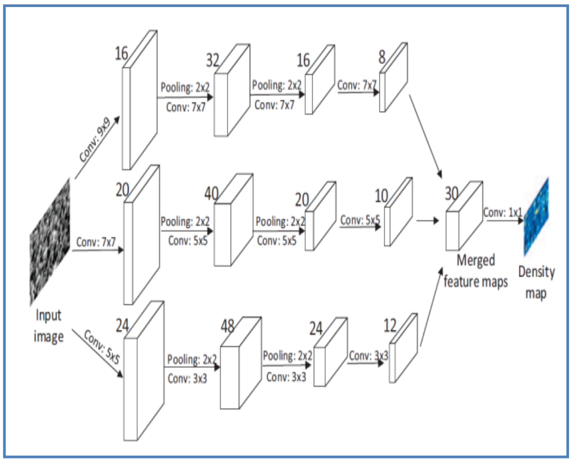

Figure 1 : Proposed MCNN Architecture

The architecture of MCNN contains three columns of CNN whose filters have different sizes. The overall structure of MCNN is illustrated in Figure1.1.It contains three parallel CNNs whose filters are with receptive fields of different sizes. We use less number of filters for CNNs with larger filters to reduce the computational complexity. The same network structure is utilized for all columns. The structure is convolution–pooling–convolution–pooling except for the sizes and numbers of filters. Max pooling is applied for each regions, and Rectified linear unit (ReLU) is adopted as the activation function. The output feature maps of all CNNs are stacked together and finally mapped to a density map. The filters are used to map the features maps to the density map.

4. SYSTEM SCENARIO.

The model built is fine tuned before its actual execution which means that the images are preprocessed. As the MCNN is trained to estimate the crowd density map from an input image, the quality of density given in the training data is very much important as it determine the performance of system. The crowd is estimated by estimating the head count of each person in the given image. An simple method of converting the annotated image of labeled people heads to a map of crowd density[1] is adopted. Let's assume that, if there is a head at pixel $\boldsymbol{x}_i$, then it is represented as a delta function $\delta(\boldsymbol{x} - \boldsymbol{x}_i)$. Hence an image with N heads labeled can be represented as a function,

$$H(x) = \sum_{i=1}^{N} \delta(x - x_i)$$

The above function is converted into a continuous density function using a convolution of the function with a kernel. Hence the density obtained is as follows,

$$F(\boldsymbol{x}) = H(\boldsymbol{x}) * G_\sigma(\boldsymbol{x}).$$

Usually a density function assumes that the pixels are independent samples in the image plane [17]. For dense crowded scene, the density map is determined by the spread parameter for each person based on its average distance to its neighbors. Convolve $\delta(x - x_\mathrm{i})$ with a Gaussian kernel to estimate the crowd density around the pixel $\boldsymbol{x}_\mathrm{i}$. The density $\boldsymbol{F}$ is, therefore estimated and is defined as

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x)$$

Where, $\sigma_i = \beta d$ ($\beta$is some parameter and d is the average distance estimated through *k* nearest neighbor).The difference between the estimated density map and ground truth is measured by the Euclidean distance[32]. The loss function used to calculate the loss is defined as follows,

$$L(\Theta) = \frac{1}{2N}\sum_{i=1}^{N}\|F(X_i;\ \Theta) - F_i\|_2^2$$

where ,L is loss between estimated density map and the ground truth density map, Θ Is the a set of learnable parameters in the MCNN. N stands for the number of training image, X_i is the input image and F_i is the ground truth density map of image X_i. Here F (X_i;Θ) is the estimated density map generated by MCNN which is parameterized with Θ for sample X_i.

4.1 Training and Testing Of Dataset

The input given to the system is in the form of an image or video. Shanghaitech dataset [33] is used which is considered as the largest one in terms of the number of annotated people. It contains 1198 annotated images, with a total of 330,165 people with centers of their heads annotated. This dataset is divided in two parts Part_A and Part_B. The Part_A dataset contains 482 images which are randomly crawled from the internet and Part_B dataset contains 716 images in are taken from the busy streets of metropolitan areas in Shanghai. The crowd density between the two dataset are totally different. The subsets of Shanghaitech dataset, Part A and Part B are divided into training and testing. Almost 300images of Part A are used for training and the remaining 182 images for testing. Similarly, 400 images of Part B are used for training and 316 for testing. Here we first pre-train each column of MCNN independently.

4.2 Implementation of the system

For implementation of the Multicolumn CNN model for detection of crowd and density estimation training algorithm is used. The preprocessing of the dataset is done through the inbuilt functions which include scaling, labeling and resize. The proposed algorithm has five main steps, model creation, Training model, loss calculation, crowd detection and density analysis and finally the metric evaluation. Different parameters such as batch size etc. are set to create a model .The model is trained and tested on a huge dataset. The losses are calculated and the crowd count and the density map is provided for each of the input image. MAE and MSE are the metrics used for evaluation of the performance of the model. The proposed algorithm is implemented using python, TensorFlow and various other packages provided by the python .Here standard algorithms for crowd detection and estimation are used for implementation.

5. RESULT ANALYSIS AND DISCUSSION.

The MCNN model uses Shanghaitech (Part A and Part B) datasets for training and testing the system .It is a large-scale crowd dataset divided in two parts, Part A and Part B. Part A has considerably larger density images as compared to Part B. Both the parts are further divided into training and evaluation sets. The Part A contains images which are randomly crawled from the Internet. Most of them have a large number of people. The training is carried on 300 images and the testing is done on nearly 182 images. The Part B contains the images from busy streets of metropolitan areas in Shanghai. It contains the approximately 400 images for training and 316 images for testing.

5.1 Evaluation metric

The model is evaluated mainly on the two metrics, MAE and MSE [28]. In more simplified words,the MAE mainly indicates the accuracy of the estimates and the MSE indicates the robustness of the estimates. The more the results are accurate, the more the model is precise. Both the absolute error (MAE) and the mean squared error (MSE) are defined as follows:

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad MAE = \frac{1}{N}\sum_{i=1}^{N}|(y_i - \hat{y}_i)|$$

Where N is the number of test images, y_i is the actual number of people in the i^{th} image, and $\hat{y}_i$ as the estimated number of people in the i^{th} image.

5.2 Methods of comparison

Initially the method is compared with the work stated in the literature and the details obtained from the existing system. The following table shows the comparison of the Shanghaitech dataset with the existing dataset. The table 1 shows the comparison of the same.

Table 1 : Comparison of Shanghaitech dataset with existing dataset.

Dataset	Resolution	Num	Max	Min	Average	Total
UCSD	158 x 238	2000	46	11	24.9	49,885
UCF_CC_50	different	50	4543	94	1279.5	63,974
WorldExpo	576 x 720	3980	253	1	50.2	1,99,923
Shanghaitech Part A	different	482	3139	33	501.4	2,41,677
Shanghaitech Part B	768 x 1024	716	578	9	123.6	88,488

The above table gives the brief discussion of five different datasets with their respective resolution, Num as number of images, Max as maximum crowd count in the image, Min as the minimum crowd count in the particular image., Average as the average of the crowd count and Total as the total number of labeled people or the total number of annotated peoples in the given dataset.

It is clearly observed that the Shanghaitech dataset is the largest among the others with a huge number of images. It is also divided into two different parts Part A and Part B. The MCNN model uses the Shanghaitech dataset for training as well as testing the system.

5.3. Results Analysis

The results show that the MCNN model gives precise and remarkable output as compared to the existing methods. The training algorithm of the MCNN model provides more accurate results and hence improves the performance. The proposed model uses the Shanghaitech dataset to conduct the experiments .The Table 2 shows the experiment results of various methods on the Shanghaitech dataset .Hence one can give a statement that the model achieves the state-of-the-art performance.

Table 2. Comparing performance of different methods on Shanghaitech dataset

Shanghaitech Dataset	Part _ A		Part_B	
Method	**MAE**	**MSE**	**MAE**	**MSE**
LBP+RR	303.2	371.0	59.1	81.7
Zhang	181.8	277.7	32.0	49.8
MSNN-CCR	245.0	336.1	70.9	95.9
MCNN	110.2	173.2	26.4	41.3
PROPOSED MCNN	**66.46**	**117.00**	**8.10**	**12.77**

The performance of the MCNN model is evaluated with the evaluation metrics. The above table concludes that the proposed system utilizes the given state-of-the-art performance.The method is more accurate and robust to the large variations in the crowd number or in density. The following figures shows the comparison between the proposed MCNN model with the other methods discussed in the literature.It is observed that minimum values of the MAE and MSE are obtained as compared with the other method

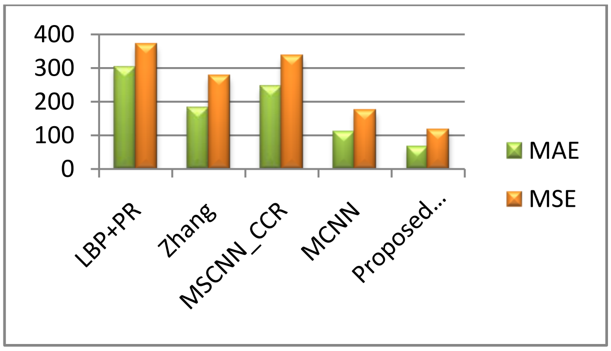

Figure 3. Performance comparison Shanghaitech dataset Part A

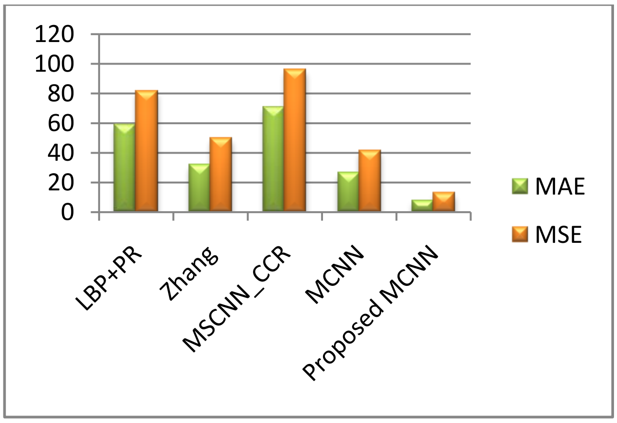

Figure 3. Performance comparison Shanghaitech dataset Part B

6. ADVANTAGES

The advantages of MCNN model for crowd and density estimation can pointed as follows.

- It preserves more information
- It is robust and adaptive to large variations
- As the model is an trained model it can be easily adapted or transferred to another dataset for further estimations.
- The filters are more semantic and meaningful which improves the accuracy of crowd counting and density estimation.
- The results are more precise and accurate as compared to the other methods .

7. CONCLUSION

Crowd detection is one of the challenging problems of computer vision and machine learning. We have proposed a Multi-column Convolution Neural Network which can estimate crowd with higher rate of accuracy in a single image from almost any perspective. CNN is the basic framework to learn efficient features for counting. Hence, the accuracy of the system is increased compared to the other techniques. To get better performance of crowd counting, it requires large label dataset. Here, Shanghaitech Dataset is used which is further divided into two parts, Part A and Part B This is the largest dataset so far in terms of the annotated heads for crowd counting. The model proposed outperforms the state-of-art crowd counting methods on all datasets used for evaluation. The accuracy of the model obtained is also more precise and more accurate as compared to the existing methods. The future work of the proposed system can be extended to, face recognition in crowd for visual surveillance system which can handle challenges such as blurred and overlapped faces in crowded areas.

8. REFERENCES

[1] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-image crowd counting via multi-column convolutional neural network,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 589–597, 2016.

[2] M. J. Jones and D. Snow, “Pedestrian detection using boosted features over many frames,” International Conference on Pattern Recognition, 2008. ICPR 2008. IEEE, pp. 1–4, December 2008.

[3] P. Viola, M. J. Jones, and D. Snow, “Detecting pedestrians using patterns of motion and appearance,” International Journal of Computer Vision, vol. 63, no. 2, pp. 153–161, 2005.

[4] 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, vol. 31, no. 6, pp. 645–654, 2001.

[5] “22 dead, several injured in stampede at mumbai’selphinstone road station, The Hindu,” https://www.thehindu.com/news/cities/mumbai/mumbai-stampede/ article19775073.ece, [Accessed Oct. 22, 2017].

[6] T. Zhao and R. Nevatia, “Tracking multiple humans in complex situations,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 26, no. 9, pp. 1208–1221, 2006.

[7] 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, Elsevier, vol. 107, pp. 1–16, 2017. 53

[8] S. Abdulla, M. Saleh, S. A. Suandi, and H. Ibrahim, "Recent survey on crowd density estimation and counting for visual surveillance," Engineering Application of Artificial Intelligence, ScienceDirect, vol. 41, pp. 103–114, 2015.

[9] D. Helbing, D. Brockmann, T. Chadefaux, K. Donnay, U. Blanke, O. Woolley- Meza, M. Moussaid, A. Johansson, J. Krause, and S. Schutte, "Saving human lives: what complexity science and information systems can contribute," Journal of Statistical Physics, vol. 158, no. 3, pp. 1–47, 2014.

[10] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 13, no. 7, pp. 1198–1211, 2008.

[11] Death toll in MP stampede reaches 115; congress wants CM to quit, The Times of India," https://timesofindia.indiatimes.com/india/ Death-toll-in-MP-stampede-reaches-115-Congress-wants-CM-to-quit/ articles how/24151591.cms, [Accessed Oct. 22, 2017].

[12] Kumbh mela chief azam khan resigns over stampede, BBC News," https:// www.bbc.com/news/world-asia-india-21406879, [Accessed Oct. 22, 2017].

[13] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence rated predictions," Machine Learning, vol. 37, no. 3, pp. 297–336, 1999.

[14] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 878–885, June 2005.

[15] S. F. Lin and C. D. Lin, "Estimation of the pedestrians on a crosswalk," International

Joint Conference SICE-ICASE, 2006. IEEE, pp. 4931–4936, October 2006.

[16] V. Rabaud and S. Belongie, "Counting crowded moving objects," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1. IEEE, pp. 705–711, June 2006.

[17] J. Shi and C. Tomasi, "Good features to track," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 593–600, June 1994.

[18] O. Sidla, Y. Lypetskyy, N. Brandle, and S. Seer, "Pedestrian detection and tracking for counting applications in crowded situations," IEEE International Conference on Video and Signal Based Surveillance, November 2006.

[19] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multiscale counting in extremely dense crowd images," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2547–2554, 2013.

[20] C.Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," Proceedings of the 23rd ACM international conference on Multimedia, ACM, pp. 1299–1302, 2015.

[21] P. Viola and M. J. Jones, "Robust real-time face detection," International journal of computer vision, vol. 57, no. 2, pp. 137–154, 2004.

[22] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, and C. Zhu, "Fast crowd density estimation with convolutional neural networks," Engineering Applications of Artificial Intelligence, vol. 43, pp. 81–88, 2015.

[23] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang, "Multi-scale convolutional neural networks for crowd counting," IEEE International Conference on Image Processing (ICIP), pp. 465–469, February 2017.

[24] D. Kang, Z. Ma, and A. B. Chan, "Beyond counting: Comparisons of density maps for crowd analysis tasks-counting, detection, and tracking," Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–14, 2018.

[25] Bansal and K. S. Venkatesh, "People counting in high density crowds from still images," International Journal of Computer and Electrical Engineering, vol. 7, no. 5, pp. 316–324, 2017.

[26] Dertat, "Applied deep learning - part 4: Convolutional neural networks," https://towardsdatascience.com/ applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2, [Accessed May 17, 2018].

[27] "Microsoft - cognitive toolkit - tutorial 2," https://docs.microsoft.com/en-us/ cognitive-toolkit/tutorial2/tutorial2, [Accessed May 23, 2018].

[28] U. Udofia, "Basic overview of convolutional neural network (cnn)," https://medium.com/@udemeudofia01/ basic-overview-of-convolutional-neural-network-cnn-4fcc7dbb4f17, [Accessed May 24, 2018].

[29] "Deep learning for computer vision," https://www.slideshare.net/Tricode/ deep-learning-stm-6, [Accessed May 24, 2018].

[30] S. Setty, M. Husain, P. Beham, J. Gudavalli, M. Kandasamy, R. Vaddi, V. Hemadri, J. C. Karure, R. Raju, Rajan, K. V., and J. C. V., "Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations," National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), pp. 1–5, December 2013.

[31] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 130–136, 1997.

[32] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 833–841, 2015.

[33] M. Mathias, R. Benenson, M. Pedersoli, and V. L. Gool, "Face detection without bells and whistles," European conference on computer vision, pp. 720–735, 2014.