# A system for crowd counting in highly congested scenes geared towards smart transportation systems

**Cecilia Veronica Macias Hernandez** [1] **Lina Maria Aguilar Lobo** [1] **Gilberto Ochoa-Ruiz** [2]

## Abstract

In this work we present a novel framework for people counting on crowded scenes, specifically tailored for analyzing the influx of people in transportation systems (i.e. bus and metro stations). This information is to be exploited by a larger physical management system by deploying a predictive analytics framework for optimizing the service in countries like Mexico, where such systems are rather scant. For analyzing the scene, we make use of a computer vision data-driven approach based on fully convolutional neural networks to be implemented on smart cameras, which are capable of performing accurate count estimation through density maps, thus avoiding privacy issues. Herein, we present the proposed approach, the utilized neural net architecture and results which improve upon those in the state of the art.

## 1. Introduction

Counting crowd pedestrians in video has drawn a lot of attention in recent years, as it is especially important for metropolis security and management. Crowd counting is a challenging task due to the presence of severe occlusions, perspective distortions and diverse crowd distributions, in which object detectors cannot work reliably. Therefore, most of the state of the approaches in the literature tackle crowd modeling as an instance of the more general density estimation problem, which has been successfully used in domains such as medicine and biology. As with previous works in the literature, we tackle this problem by deploying deep learning architectures able to learn the regression function that projects the image appearance into an object density map, allowing the derivation of an estimated object density map for unseen images.

*Equal contribution [1]Universidad Autónoma de Guadalajara, Mexico [2]Tecnológico de Monterrey,Mexico. Correspondence to: G. Ochoa-Ruiz <gilberto.ochoa@tec.mx>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Vienna, Austria, PMLR 108, 2020. 

The person count can be obtained from the density map through integration of various image patches in a multi-scale pyramid. The process is shown schematically in Figure 1.

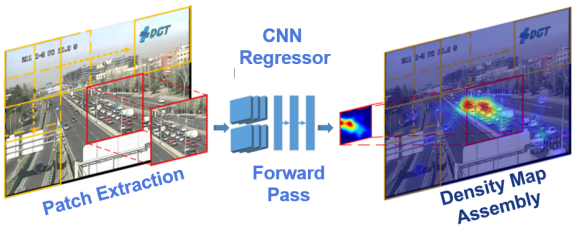


*Figure 1.* Deep learning crowd counting via density estimation

### 1.1. Motivation

In this work, we make use of crowd counting algorithms for implementing a solution geared around the "smart city" paradigm. This work forms part of a larger endeavor in which smart cameras are to be used as sensors in specific bus stops or metro/BRT stations for acquiring information about mobility patterns in public transportation system, helping the city administrators or private companies to improve the service (i.e. better schedules or possibly real-time management of the transportation network).

The main rationale for using smart cameras is that they extract specific metrics from raw video, which can be stored inexpensively in the cloud, reducing the required communication, computational and storage burden typically associated with these technologies.

The proposed solution can be easily complemented with other algorithms running on the camera, making it a viable solution for in-development countries, in which there is an increased interest for this type end-to-end solutions (i.e. for action recognition to detect anomalous behaviours).

### 1.2. Related Works

Although crowd counting has a relatively long history (Lempitsky & Zisserman, 2010), recent successes of CNN-based methods in classification tasks have inspired researchers to employ them for the purpose of crowd counting and density estimation (Li et al., 2015; Cong Zhang et al., 2015).

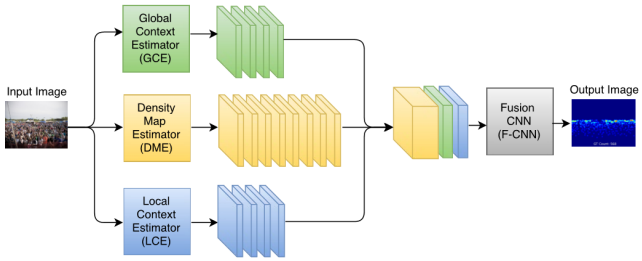


*Figure 2.* Overview of the used context pyramid CNN

Several approaches have been proposed to tackle the huge variations in perspective and crowd distributions (Oñoro-Rubio & López-Sastre, 2016); such approaches usually employ a "multi-branch" structure (Zhang et al., 2016) approach for extracting features at several scales or by mixing shallow and deep networks to the same effect. The main reason for using multi-column NNs (MCNNs) is the flexible receptive fields provided by the convolutional filters with different sizes across the columns (Sindagi & Patel, 2017b)].

## 2. Proposed approach

Recent works (Cao et al., 2018) have shown that the various branches are ineffective and costly (i.e., in terms of complexity and time) and thus, we decided to explore methods based on contextual pyramids CNNs for tackling the variation in crowd distribution due to perspective distortions (Li et al., 2018). The main idea is to deploy a deeper FCN design, using VGG-16 as a front-end with a 1x1 conv layer for density map generation, as depicted in Figure 2, but using dilated conv layers for extracting deeper saliency information. In this manner, the image resolution is maintained, yielding better crowd counts due to the improved density maps.

We tested our crowd counting algorithm on the various public datasets in the literature (UCF_CC_50, ShanghaiTech , WorldExpo (Sindagi & Patel, 2017a)) achieving competitive results on the metrics typically reported: the Mean Average and Mean Square Errors (MAE and MSE, respectively). The results of our model in the challenging Shanghai Tech Part A, compared to other methods is show in Table I; the lower the metrics the better results in the regression and thus in the crowd count in average for the dataset.

*Table 1.* Count errors for different methods ShanghaiTech Part A.

| Method | Reference | MAE | MSE |
|---|---|---|---|
| MCNN | (Zhang et al., 2016) | 181.8 | 277.7 |
| SW-CNN | (Sam et al., 2017) | 90.4 | 135.0 |
| C-MTL | (Sindagi, 2017) | 101.3 | 152.4 |
| Ours | | 67.0 | 104.5 |

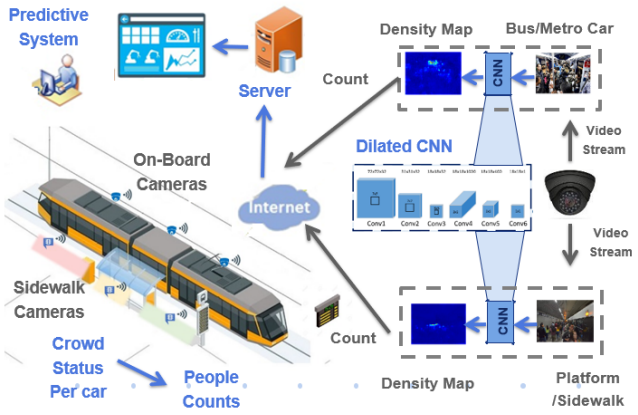


*Figure 3.* Proposed system for people counting based on smart cameras and dilated context-aware CNNs

In Figure 3 we present our approach for crowd counting on public transportation networks. The main idea is to instrument the sidewalks and metro or buses with smart cameras, which incorporate the proposed dilated CNN which automatically yields the person count in a particular car or station This information can be transmitted over a wireless connection and stored in a server for carrying out big data analyses and/or predictive analytics (i.e. for adjusting the QoS in terms of the demand) and be sent back to displays to reduce mobility issues or contingencies at the stations. We haven't yet installed such cameras, but we have obtained permits to gather images from public transportation in Guadalajara to test our algorithms. As it can be observed from the figure, one of the advantages of using density maps over raw video is that the privacy of the users is not compromised, as only people counts are transmitted to the main server.

## 3. Discussion and future work

In this paper we presented a deep learning-based algorithm for crowd counting applications, which achieves competitive results with respect to other methods in the state of the art. We leverage recent advances in segmentation architectures for proposing a novel method makes use of atrous convolution, this obtaining better density maps. The architecture performs especially well in very crowded areas, a factor that we seek to exploit for implementing smart cameras in the analysis and monitoring of transportation systems.

We are currently exploring schemes for efficiently implementing of our algorithms on smart cameras, specifically quantization techniques for embedded systems (Sze et al., 2017); this aspect is of paramount importance for their deployment in low-power devices for smart cities applications.

## References


Cao, X., Wang, Z., Zhao, Y., and Su, F. Scale aggregation network for accurate and efficient crowd counting. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *Computer Vision – ECCV 2018*, pp. 757–773, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01228-1.

Cong Zhang, Hongsheng Li, Wang, X., and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 833–841, 2015.

Lempitsky, V. and Zisserman, A. Learning to count objects in images. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pp. 1324–1332, Red Hook, NY, USA, 2010. Curran Associates Inc.

Li, T., Chang, H., Wang, M., Ni, B., Hong, R., and Yan, S. Crowded scene analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(3): 367–386, 2015.

Li, Y., Zhang, X., and Chen, D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. *CoRR*, abs/1802.10062, 2018. URL http://arxiv.org/abs/1802.10062.

Oñoro-Rubio, D. and López-Sastre, R. J. Towards perspective-free object counting with deep learning. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *Computer Vision – ECCV 2016*, pp. 615–629, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46478-7.

Sam, D. B., Surya, S., and Babu, R. V. Switching convolutional neural network for crowd counting. *CoRR*, abs/1708.00199, 2017. URL http://arxiv.org/abs/1708.00199.

Sindagi, V. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. *CoRR*, abs/1707.09605, 2017. URL http://arxiv.org/abs/1707.09605.

Sindagi, V. and Patel, V. M. A survey of recent advances in cnn-based single image crowd counting and density estimation. *CoRR*, abs/1707.01202, 2017a. URL http://arxiv.org/abs/1707.01202.

Sindagi, V. A. and Patel, V. M. Generating high-quality crowd density maps using contextual pyramid cnns. *CoRR*, abs/1708.00953, 2017b. URL http://arxiv.org/abs/1708.00953.

Sze, V., Chen, Y., Yang, T., and Emer, J. S. Efficient processing of deep neural networks: A tutorial and survey. *CoRR*, abs/1703.09039, 2017. URL http://arxiv.org/abs/1703.09039.

Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. Single-image crowd counting via multi-column convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 589–597, 2016.