

# Content-aware Density Map for Crowd Counting and Density Estimation

M. M. Oghaz<sup>1</sup>   A. Khadka<sup>1</sup>   V. Argyriou<sup>1</sup>   P. Remagnino<sup>1</sup>

<sup>1</sup>Kingston University, London, United Kingdom

{m.maktabdaroghaz, a.khadka, vasileios.argyriou, p.remagnino}@kingston.ac.uk

## Abstract

Precise knowledge about the size of a crowd, its density and flow can provide valuable information for safety and security applications, event planning, architectural design and to analyze consumer behavior. Creating a powerful machine learning model, to employ for such applications requires a large and highly accurate and reliable dataset. Unfortunately the existing crowd counting and density estimation benchmark datasets are not only limited in terms of their size, but also lack annotation, in general too time consuming to implement. This paper attempts to address this very issue through a content aware technique, uses combinations of Chan-Vese segmentation algorithm, two-dimensional Gaussian filter and brute-force nearest neighbor search. The results shows that by simply replacing the commonly used density map generators with the proposed method, higher level of accuracy can be achieved using the existing state of the art models.

**Keywords:** deep learning, crowd counting, crowd analysis, convolutional neural networks, computer vision, segmentation, Chan-Vese

## 1 Introduction

The study of human behavior is a subject of great scientific interest and probably an inexhaustible source of research. One of the most cited and popular research topic in human behavior analysis is study of crowd features and characteristics. In recent years, crowd analysis has gained a lot of interest mainly due to its wide range of applications such as safety monitor-

ing, disaster management, public spaces design, and intelligence gathering, especially in the congested scenes like arenas, shopping malls, and airports [1, 2].

Crowd counting, localization and density estimation are crucial objectives of an automated crowd analysis system. Accurate knowledge of the crowd size, location and density in a public space can provide valuable insight for tasks such as city planning, analyzing consumer shopping patterns as well as maintaining general crowd safety. Several studies attempt to produce an accurate estimation of the true number of people present in a crowded scene through density estimation.

Deep learning has proven superior to classic computer vision and machine learning techniques that tend to struggle with the complexity of crowd counting and behavior analysis models. [3].

Generally, crowd counting and density estimation approaches can be divided in two categories: detection-based methods (specific) and regression-based methods (holistic). Detection-based methods generally assume each person on the crowd can be detected and located individually based on its individual features and characteristics. These approaches are preferable in sparse crowd analysis where crowd occlusion is negligible. Holistic crowd counting and behavior analysis approaches utilize global crowd features and characteristics to estimate crowd size, flow and density. These approaches are preferable in dense crowd analysis, where crowd occlusion is significant. Due to high amount of occlusions these approaches only utilize heads as

deterministic feature [4].

However, crowd counting and density estimation is not a trivial task. Several key challenges such as severe occlusions, poor illumination, camera perspective and highly dynamic environments further complicate crowd analysis. Moreover, poor quality of annotated data increases to complexity of crowd counting and behavior analysis in crowded environments. The existing crowd counting and density estimation benchmark datasets are not only limited in terms of the quantity, but also lack in terms of annotation strategy.

In regression-based crowd counting and density estimation approaches, people heads are the only visible body part in an image. Thus, these approaches use heads as the only discriminant feature. Meanwhile, the existing benchmark datasets such as UCF-CC-50 and ShanghaiTech only provide people heads centroid pixel instead of masking the entire head region. Hence, the recreation of the ground truth head masks is accomplished through a static two-dimensional Gaussian filter or a dynamic two-dimensional Gaussian based on the  $K$  nearest neighbors. However, the dynamic Gaussian approach based on proximity of the nearest neighbors mitigates the issue to some extent, but this technique is not content aware and incorporates significant amount of noise into ground truth data [5, 6].

In this regard, our study attempts to address the limitation of the existing crowd counting and density estimation benchmark datasets through a content aware annotation technique. It employs combinations of nearest neighbor algorithm and unsupervised segmentation to generate the ground truth head masks. The proposed technique first uses the brute-force nearest neighbor search to localize the nearest neighbor head point, then it identified the head boundaries using Chan-Vese segmentation algorithm and generates a two-dimensional Gaussian filter on that basis. We believe that by simply replacing the  $kNN$ /Gaussian based ground truth density maps in an existing state of the art network with the proposed content aware approach in this study, higher level of accuracy can be achieved.

The rest of this paper is organized as following: section 2 summarizes the related work, section 3 describes the existing datasets and anno-

tation strategies, section 4 presents the proposed methodology, section 5 presents the experimental results and finally section 6 concludes the findings of this research.

## 2 Related Work

Over the last decade, there have been several studies to address the problem of crowd counting and density estimation using deep learning techniques.

Liu *et al.* [7] proposed a universal network for counting people in a crowd with varying density and scale. In this study the proposed network is composed of two components: a detection network (DNet) and an encoder-decoder estimation network (ENet). The input first run through DNet to detect and count individuals who can be segmented clearly. Then, ENet is utilized to estimate the density maps of the remaining areas, where the numbers of individuals cannot be detected. Modified version of Xception used as an encoder for feature extraction and a combination of dilated convolution and transposed convolution used as decoder. Authors attempted to address the variations in crowd density with two literally isolated deep networks which significantly slows down the process lacks novelty.

In another study, Mehta *et al.* [8] proposed independent decoding reinforcement branch as a binary classifier which helps the network converge much earlier and also enables the network to estimate density maps with high Structural Similarity Index (SSIM). A joint loss strategy, the *binary cross entropy* (BCE) loss and *mean squared error* (MSE) loss used to train the network in an end to end fashion. They have used variation of the U-net model to generate the density maps. The proposed model shows notable improvements in recreation of the crowd density maps over the existing models.

A study by Oh *et al.* [9] attempt to address the uncertainty estimation in the domain of crowd counting. This study proposed a scalable neural network framework with quantification of decomposed uncertainty using a bootstrap ensemble. The proposed method incorporates both epistemic uncertainty and aleatoric uncertainty in a neural network for crowd counting. The proposed uncertainty quantification method pro-

vides additional auxiliary insight to the crowd counting model. The proposed technique attempt to address the uncertainty issue in crowd counting. However the use of unsupervised calibration method to re-calibrate the predictions of the pre-trained network is questionable.

In another study Olmschenk *et al.* [10] attempt to address the inefficiency of the existing crowd density map labeling scheme for training deep neural networks. This study proposes a labeling scheme based on inverse k-nearest neighbor (*ikNN*) maps which does not explicitly represents the crowd density. Authors claim a single *ikNN* map provides information similar to the commonly practiced accumulation of many density maps with different Gaussian spreads.

A study by Idrees *et al.* [5] stems from the observation that crowd counting, density map estimation and localization are very interrelated and can be decomposed with respect to each other through composition loss, which can then be used to train a neural network. This study

Several other studies including [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23] attempted to address crowd counting, localization and density estimation issues yet majority of these approaches employed the flawed ground truth density map generation approach.

### 3 Annotation Strategy

In a dense crowd scenario, aside from people heads which are usually fairly visible, the majority of the other body parts are subject to heavy occlusion. This makes heads the only reliable discriminant feature in dense crowd counting and localization. Existing crowd counting and density estimation benchmark datasets such as UCF-CC-50 and ShanghaiTech provide the heads centroid pixel location as labels. Conducting the crowd counting and density estimation as a regression task, seeks for regional isolation of the heads in the form of a binary mask. As the head size is subject to various factors such as camera specifications, point of view, perspective, distance and angle, generation of such mask could be challenging task, given the heads centroid pixel is the only provided form of annotation in existing benchmark datasets.

The formation of the ground truth binary head

masks in majority of the existing studies is either accomplished through a static two-dimensional Gaussian filter or a dynamic two-dimensional Gaussian filter paired with  $k$  nearest neighbors approach. The static two-dimensional Gaussian filter assigns a fixed size Gaussian filter to each head regardless of the head size and proximity of the nearest neighbor. This approach does not attempt to compensate for crowd density, distance and camera perspective and incorporates significant amount of noise into ground truth data. The dynamic two-dimensional Gaussian filter approach employs the nearest neighbors search through  $k$ -d tree space partitioning approach, prioritizes the speed over integrity and does not deliver optimal results. In this approach the Gaussian filters are centered to the annotation points and spread based on the average euclidean distance among the three nearest neighbors. In both approaches, the spatial accumulation of all Gaussians creates the global density map for the given image. The following formula shows the commonly used dynamic two-dimensional Gaussian approach:

$$D(x, f) = \sum_{h=1}^T \frac{1}{\sqrt{2\pi}f(\sigma_h)} \exp\left(-\frac{(x-x_h)^2 + (y-y_h)^2}{2f(\sigma_h)^2}\right) \quad (1)$$

where  $T$  is the total number of the heads presents in the given image,  $\sigma_h$  is the sized for each head point positioned at  $(x_h, y_h)$  determined by  $k$ -d tree space partitioning approach based on the the average euclidean distance among the three nearest neighbors and  $f$  is a scaling constant.

The dynamic Gaussian approach based on the  $k$  nearest neighbors attempts to mitigate the crowd density, distance and camera perspective issues to some extent. However, this technique is not content aware and it introduces a significant amount of noise into the ground truth data, which in turn negatively affects the model's accuracy. Figure 1 shows some sample images from the ShanghaiTech dataset along with their respective ground truth density maps. It can be observed that both approaches are fairly unreliable and inconsistent in determining the true head sizes.



Figure 1: From top to bottom: sample images from the ShanghaiTech dataset, density map based on static two-dimensional Gaussian filter and density map based on dynamic two-dimensional Gaussian filter using  $k$ -d tree space partitioning technique.

## 4 Methodology

In order to address the shortcomings of the existing ground truth density maps generation approaches, this study offers a content aware technique using combinations of Chan-Vese segmentation algorithm, two-dimensional Gaussian filter and brute-force nearest neighbor search.

This technique is based on the Mumford-Shah functional for segmentation, and is widely used in the medical imaging field. The Chan-Vese segmentation algorithm is able to segment objects without prominently defined boundaries. This algorithm is based on level sets that are evolved iteratively to minimize an energy, which is defined by weighted values corresponding to the sum of differences intensity from the average value outside the segmented region, the sum of differences from the average value inside the segmented region, and a term which is dependent on the length of the boundary of the segmented region. As the head boundaries in highly dense crowds are not clearly defined, this technique can be used to segment the head regions from the background. Chan-Vese algorithm attempt to minimize the following energy function

in an iterative process [24].

$$\begin{aligned}
 F(c_1, c_2, G) = & \mu \cdot Len(G) + \nu \cdot Area(in(G)) \\
 & + \lambda_1 \int_{in(G)} |u_0(x, y) - c_1|^2 dx dy \\
 & + \lambda_2 \int_{out(G)} |u_0(x, y) - c_2|^2 dx dy
 \end{aligned} \tag{2}$$

where  $G$  denote the initial head which manually set to a  $5 \times 5$  bounding box centered on the labelled head point,  $c_1$  will denote the average pixels' intensity inside the initial head region  $G$ , and  $c_2$  denotes the average intensity of a square box, centered to the annotation head point and its boundary extended to the nearest neighbor head point.  $\lambda_1$ ,  $\lambda_2$  and  $\mu$  are positive scalars, manually set to 1, 1 and 0 respectively. A two-dimensional Gaussian filter which extends to the  $G$  mean and centered to the head point is used to create the ground truth head mask.

Unlike  $k$ -d tree space partitioning technique which does not always delivers the absolute nearest neighbors, brute-force nearest neighbor search technique always guarantees to find the absolute nearest neighbors regardless of the distribution of the points. The brute-force nearest neighbor search technique does take considerably longer time ( $O(n^2)$  vs  $O(n \log n)$ ) to find the nearest neighbors. However, since generating the ground truth density maps is a single-pass preliminary operation in crowd counting and density estimation, speed is a less of a priority. Since, the Chan-Vese segmentation algorithm only uses the very nearest neighbor head point to determine the boundary of the outside region, the brute-force nearest neighbor search only looks for the very nearest head point. To create the global density map, we employed an exclusive cumulative of the Gaussians which addresses the head mask overlap issue. To maintain the count integrity, density map has been normalized at each iteration.

## 5 Experimental Results

In order to measure the effectiveness of our content-aware crowd density map generator, we have re-trained some of the notable state of the art deep models including Sindagi *et al.* [25]

, Shi *et al.* [22], Li *et al.* [26] and Zhang *et al.* [27] using the density maps generated by the proposed crowd density map generator. We have used the original implementation of these algorithms provided by authors in Github. All algorithms were trained and tested across both UCF-CC-50 and ShanghaiTech datasets using the proposed content-aware crowd density map generator as well as the commonly used existing ground truth density map generator. In some cases we were unable to reproduce the reported performance in the original manuscripts. However, as we were consistent with the experiments across both density map generators, validity and integrity of the comparison is not compromised.

Table 1 shows the mean square error (MSE) comparison between the proposed and existing density map generator across ShanghaiTech dataset part A and B. It can be observed that using the proposed content-aware density map generator, MSE has been consistently decreased across relatively all investigated models. The improvements is more pronounced in ShanghaiTech part A dataset. ShanghaiTech part A dataset exhibits more challenging and dynamic crowd scenarios. The results convey the proposed method could deliver better depiction of the ground truth density maps. Table 2 compares the MSE and mean absolute error (MAE) between the proposed and existing density map generator using extremely challenging UCF-CC-50 dataset. Similar to the results in ShanghaiTech dataset, there is a notable improvement in both MSE and MAE metrics.

Figure 2 compares the density maps generated using the existing approach based on  $k$ -d tree space partitioning technique and the proposed content-aware crowd density map generator. It can be observed that in highly dense crowds, the proposed method generates more granular density maps with lesser overlaps between neighbor Gaussians. The proposed method uses combination of pixels intensity and nearest neighbors to adjust the size of the Gaussians per head. Figure 2 shows this technique significantly improves the integrity of the density map relative to the input image.

Table 1: MSE comparison between the proposed and existing density map generator across ShanghaiTech dataset

Method	Existing Density map Generator		Proposed Density map Generator	
	ShTech-A	ShTech-B	ShTechA	ShTechB
Sindagi <i>et al.</i>	152	31	149	28
Shi <i>et al.</i>	112	26	110	26
Li <i>et al.</i>	115	16	113	16
Zhang <i>et al.</i>	197	66	191	57

Table 2: MSE and MAE comparison between the proposed and existing density map generator across UCF-CC-50 dataset

Method	Existing Density map Generator		Proposed Density map Generator	
	UCF-CC-50 MSE	UCF-CC-50 MAE	UCF-CC-50 MSE	UCF-CC-50 MAE
Sindagi <i>et al.</i>	397	322	397	320
Shi <i>et al.</i>	415	293	414	286
Li <i>et al.</i>	397	266	396	264
Zhang <i>et al.</i>	498	467	483	459



Figure 2: From top to bottom: sample images from ShanghaiTech dataset, density map generated using the existing method and density map generated using the proposed method

## 6 Conclusion

Creating an accurate model for crowd counting and density estimation demands for a large and highly reliable ground truth data in the first place. However, the existing crowd counting and density estimation benchmark datasets are not only limited in terms of size, but also lack in terms of annotation methodology. This study attempted to address this issue through a content-aware technique which employed combinations of Chan-Vese segmentation algorithm, two-dimensional Gaussian filter and brute-force nearest neighbor search to generate the ground truth density maps. Experiment results shows by replacing the commonly practiced ground truth density map generators with the proposed content-aware method, the existing state of the art crowd counting models can achieve higher level of count and localization accuracy.

## Acknowledgments

This work is co-funded by the EU-H2020 within the MONICA project under grant agreement number 732350. The Titan X Pascal used for this research was donated by NVIDIA.

## References

- [1] Nuria Pelechano, Kevin O'Brien, Barry Silverman, and Norman Badler. Crowd simulation incorporating agent psychological models, roles and communication. Technical report, Pennsylvania, United States Center for Human Modeling and Simulation, 2005.
- [2] Barry G Silverman, Norman I Badler, Nuria Pelechano, and Kevin O'Brien. Crowd simulation incorporating agent psychological models, roles and communication. 2005.
- [3] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E O'Connor. Fully convolutional crowd counting on highly congested scenes. *arXiv preprint arXiv:1612.00220*, 2016.
- [4] David Ryan, Simon Denman, Sridha Sridharan, and Clinton Fookes. An evaluation of crowd counting methods, features and regression models. *Computer Vision and Image Understanding*, 130:1–17, 2015.
- [5] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–546, 2018.
- [6] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi 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.
- [7] Lei Liu, Jie Jiang, Wenjing Jia, Saeed Amirgholipour, Michelle Zeibots, and Xiangjian He. Denet: A universal network for counting crowd with varying densities and scales. *arXiv preprint arXiv:1904.08056*, 2019.
- [8] Varun Kannadi Valloli and Kinal Mehta. W-net: Reinforced u-net for density map estimation. *arXiv preprint arXiv:1903.11249*, 2019.
- [9] Min-hwan Oh, Peder A Olsen, and Karthikeyan Natesan Ramamurthy. Crowd counting with decomposed uncertainty. *arXiv preprint arXiv:1903.07427*, 2019.
- [10] Greg Olmschenk, Hao Tang, and Zhigang Zhu. Improving dense crowd counting convolutional neural networks using inverse k-nearest neighbor maps and multiscale upsampling. *arXiv preprint arXiv:1902.05379*, 2019.
- [11] Shengqin Jiang, Xiaobo Lu, Yinjie Lei, and Lingqiao Liu. Mask-aware networks for crowd counting. *arXiv preprint arXiv:1901.00039*, 2018.
- [12] Rahul Rama Varior, Bing Shuai, Joe Tighe, and Davide Modolo. Scale-aware atten-

- tion network for crowd counting. *arXiv preprint arXiv:1901.06026*, 2019.
- [13] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7661–7669, 2018.
- [14] Chen Change Loy, Shaogang Gong, and Tao Xiang. From semi-supervised to transfer counting of crowds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2256–2263, 2013.
- [15] Mohammad Hossain, Mehrdad Hosseinzadeh, Omit Chanda, and Yang Wang. Crowd counting using scale-aware attention networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1280–1288. IEEE, 2019.
- [16] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. *arXiv preprint arXiv:1903.03303*, 2019.
- [17] Ze Wang, Zehao Xiao, Kai Xie, Qiang Qiu, Xiantong Zhen, and Xianbin Cao. In defense of single-column networks for crowd counting. *arXiv preprint arXiv:1808.06133*, 2018.
- [18] Di Kang and Antoni Chan. Crowd counting by adaptively fusing predictions from an image pyramid. *arXiv preprint arXiv:1805.06115*, 2018.
- [19] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin. Crowd counting using deep recurrent spatial-aware network. *arXiv preprint arXiv:1807.00601*, 2018.
- [20] Chloe Eunhyang Kim, Mahdi Maktab Dar Oghaz, Jiri Fajtl, Vasileios Argyriou, and Paolo Remagnino. A comparison of embedded deep learning methods for person detection. *arXiv preprint arXiv:1812.03451*, 2018.
- [21] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 270–285, 2018.
- [22] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5382–5390, 2018.
- [23] Deepak Babu Sam, Neeraj N Sajjan, R Venkatesh Babu, and Mukundhan Srinivasan. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3618–3626, 2018.
- [24] Tony F Chan and Luminita A Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001.
- [25] Vishwanath A Sindagi and Vishal M Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.
- [26] Yuhong Li, Xiaofan Zhang, and Deming 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.
- [27] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 833–841, 2015.