A-CCNN: ADAPTIVE CCNN FOR DENSITY ESTIMATION AND CROWD COUNTING

Saeed Amirgholipour[1], *Xiangjian He*[1,*], *Wenjing Jia*[1], *Dadong Wang*[2], *Michelle Zeibots*[3]

[1] Global Big Data Technologies Centre, University of Technology Sydney, Australia
[2] Quantitative Imaging, CSIRO Data61, Australia
[3]Institute for Sustainable Futures, University of Technology Sydney, Australia

ABSTRACT

Crowd counting, for estimating the number of people in a crowd using vision-based computer techniques, has attracted much interest in the research community. Although many attempts have been reported, real-world problems, such as huge variation in subjects' sizes in images and serious occlusion among people, make it still a challenging problem. In this paper, we propose an Adaptive Counting Convolutional Neural Network (A-CCNN) and consider the scale variation of objects in a frame adaptively so as to improve the accuracy of counting. Our method takes advantages of contextual information to provide more accurate and adaptive density maps and crowd counting in a scene. Extensively experimental evaluation is conducted using different benchmark datasets for object-counting and shows that the proposed approach is effective and outperforms state-of-the-art approaches.

Index Terms— Crowd counting, Scale Variation, Adaptive Counting CNN

1. INTRODUCTION

Nowadays, density estimation and counting the number of people in a crowded scene is a desirable application especially in restricted, public event places such as train stations. Incidents, traffic delay, and even terrible stampedes may be caused by overcrowding in such a scene. Generally, there is an urgent need for real-time decision making corresponding to crowd changes. To deal with this situation, there exist various challenges caused by occlusions, size and shape variations of people, perspective distortion, etc. Thus, correctly counting in crowded areas is very necessary for many real-world applications including visual surveillance, traffic monitoring, and crowd analysis.

The existing approaches for crowd density estimation can be divided into two main groups, i.e., detection based methods, and feature regression-based methods [1]. Detection based methods (also called direct methods) segment and detect every individual people or object in a scene with pre-trained classifiers and then simply count them. However, in complex scenes with severe occlusions and extremely crowded scenes, these approaches often fail to detect individuals and therefore produce inaccurate countings. In the feature regression-based approaches (also called indirect approaches), learning algorithms or statistical methods are utilized to analyze the image appearance features of a crowded scene, and then estimate the number of people or objects based on image appearance. Thus, these methods are more suitable for dealing with highly crowded scenes where detecting individuals often fail.

In this paper, based on the recent advance of Counting Convolutional Neural Network (CCNN) [2], we propose a new adaptive CCNN architecture, abbreviated as A-CCNN, which processes each part of an input image using an optimally trained CCNN model to estimate the corresponding density map accurately. As illustrated in Fig. 1, to tackle the counting problem, our A-CCNN model is able to regress the density function corresponding to a specified section. This allows our model to accurately localize density maps for unseen images.

The most remarkable properties that make the proposed model outstanding for crowd analysis are: (1) the ability to handle large-scale variations in people's sizes when appearing in images; and (2) the facility to generate local density maps within a crowd scene. Therefore, the proposed model can give a complete view about the scattering of a crowd. Compared to the prior works, our approach does not use different CCNN architectures and only tries to select the most effective Hyper Parameters (HPs) for generating a CCNN model. Thus, it can learn to address scale variations in an image with a simple and effective way.

2. RELATED WORKS

In recent years, many researchers [3, 4, 5] have developed deep learning models for image segmentation, classification, and recognition, and achieved excellent results. Inspired by these, Convolutional Neural Network (CNN) models have been proposed to learn to count people and produce density maps in images simultaneously, and they have worked well for objects of approximately the same size in an image or a video. Sindagi and Patel [6] proposed an end-to-end cascaded network of CNNs that can learn globally relevant and dis-

[*] Corresponding author.

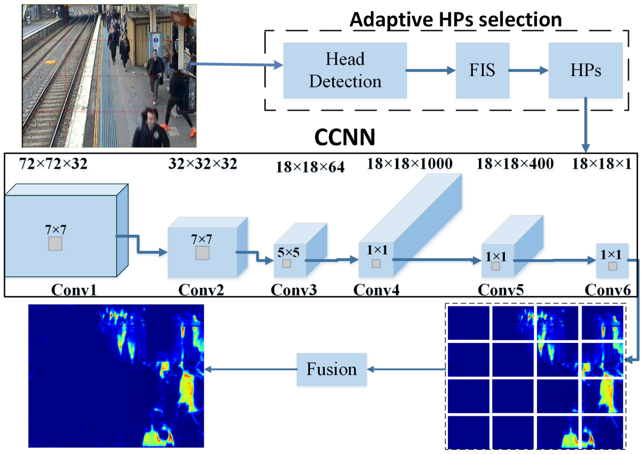

Fig. 1. The overview of our proposed A-CCNN crowd counting method. For an input image, our A-CCNN first estimates head size and corresponding position, and then utilizes a fuzzy engine to determine HP of the CCNN models for generating the density map.

criminative features to estimate highly refined density maps with low count errors. Onoro-Rubio and Lopez-Sastre proposed a regression model called Counting CNN (CCNN) [2], which can map the appearances of input image patches to corresponding density maps. They also introduced a Hydra CNN based on the idea of multi-scaling crowd counting and achieved a sufficient advantage in comparison with the previous models.

Inspired by the Hydra CNN method, some researchers have tried to utilize more complex deep models to solve the problem caused by the significant variance of crowd's appearance in a captured image/video. Deepak et al. [5] proposed a switching CNN to select the best CNN regressor for each of different receptive fields and achieved better results than the state-of-the-art for crowd counting. Kumagai1 et al. [7] proposed a mixture of CCNNs and adaptively selected multiple CNNs according to the appearance of a test image for predicting the number of people. Zhang et al. proposed a multi-column network and three independent CNN architectures and then used the combined features of these three networks to get a density map [3].

Our work presented in this paper is based on the CCNN architecture [2]. The CCNN approach takes a small patch of the input image as input and generates the corresponding density map for the image patch. By utilizing the sliding window technique, it extracts patches and applies a CNN model to regress the density function. Therefore, CCNN is formulated as a regression model that generates object density maps based on the corresponding appearances of image patches.

Formally, in the original CCNN model, the ground truth density map D_I is defined as,

$$D_I(p) = \sum_{\mu \epsilon A_I} N(p, \mu, \Sigma), \qquad (1)$$

where A_I represents the number of annotated points in the image I, and N $(p; \mu; \Sigma$) represents a normalized 2D Gaussian function with a mean of μ and a covariance of Σ, evaluated at each pixel position p.

The CCNN utilizes two crucial HPs for generating models, i.e., the patch size and the value of Σ in the Gaussian function. Through careful analysis of CCNN, we have noticed that it has a significant problem in producing a correct density map because CCNN treats the whole parts of the input image in the same way. Therefore, CCNN cannot achieve an acceptable accuracy in density estimation when a scene has a large-scale variation in the sizes of objects. We have observed that more accurate density maps can be produced when the values of the above mentioned two HPs are optimally and properly chosen.

3. ADAPTIVE CCNN

In our work, to handle crowd images with large varieties in targets' appearances, we propose a new A-CCNN model for crowd counting. As shown in Fig. 1, our A-CCNN architecture takes an image as input and equally divides the image to 16 parts and then determines the average of heads' sizes and position in different parts of an image. Then, by utilizing a Fuzzy Inference System (FIS), it feeds each image section with the same FIS linguistic output value to an appropriate CCNN model with a proper HP to obtain the corresponding density map for each section. In the end, it merges the output of different parts to achieve the final density map output.

In reality, the sizes of people who are closer to the camera appear to be bigger than those of the people who are further from the camera. Based on our observation from experiments, we find that there is a relationship between HPs and the scales of people. Thus, we use smaller (and larger) Σ's and patch sizes for the areas containing smaller (and larger) targets. Then, we train CCNN models to create density maps for different sizes of patches. For each image in the testing stage, our A-CCNN model extracts image patches from it and generates their corresponding object density maps by utilizing the relative CCNN models according to their sizes. Then, the density maps of these patches are assembled into the density map of the testing image.

Compared with CCNN, we have made the following improvement in the proposed A-CCNN. Firstly, we use different patch sizes in A-CCNN according to the sizes of people in the patches, different from using the same patch size for all patches in the original CCNN. Secondly, we have utilized various Σ values to generate the training patches. The Σ of the Gaussian function in Eq. 1 is changed to adapt the size of a patch. In comparison with the Switch-CNN, our proposed

A-CCNN uses only one well-known CCNN model with adaptive HPs, so it has less complexity than the Switch-CNN with different CNN architectures.

The process of the proposed A-CCNN is summarized as follows and detailed in the following subsections. First, we perform tiny-face detection [8] to estimate the sizes of heads in each patch of an image. Then, by feeding the head sizes and the corresponding head positions to a fuzzy inference system (FIS), we generate the appropriate HPs corresponding to the patches. Finally, these HPs are used to train CCNNs that can adaptively generate the density maps for various patches.

3.1. Head Detection

To obtain the most suitable values of HPs, we need to know the sizes of people or objects in different parts of an image. Therefore, the tiny-face detection approach [8] is used to detect faces in each part of the input image. It creates a coarse image pyramid of the input image and then feeds the scaled inputs into a CNN to get the template responses. Finally, the final detection results are produced by applying the non-maximum suppression (NMS) at the original resolution.

3.2. Adaptive HP Selection by FIS

As shown in the Fig. 1, to get the values of the HPs, a FIS is designed to adaptively select the values of HPs according to the sizes and the positions of heads. As shown in Fig. 2, the FIS receives the fuzzy information about head sizes and positions and outputs the fuzzy linguistic variables in the form of fuzzy. We choose the same Gaussian membership function for all input and output variables. Small, Average and Big are the fuzzy linguistic variables according to head sizes, and Up, Middle and Down are the fuzzy linguistic values according to head positions. The output linguistic variables are High-Pred, Mid-Pred, and Low-Pred.

Based on the Gaussian membership function, the input values are converted into the fuzzy linguistic variable in FIS. Then, the fuzzy if-then rules developed based on the Mamdani method [9] are used to map the input variables to appropriate fuzzy output variables. In total, nine fuzzy if-then rules are presented in Table 1. In general, higher (and lower) values of Σ and sliding window (patch size) produce density maps with lower (higher) counts of numbers of people. As an illustration, if an output of FIS is High-Pred (Mid-Pred, Low-Pred), the corresponding CCNN is trained with low (medium, big) HP values.

3.3. Training Parameters

Showing the effectiveness of the proposed A-CCNN, we use the same training parameters as in [2], except for two HPs, which are the patch sizes and Σ's, for people counting and density estimation. These two HPs are empirically determined on the training dataset. Similar to the approach in [2], a stochastic gradient decent algorithm is used during training. The momentum, the learning rate, and the weight decay are set to be 0.9, 0.0001 and 0.001 respectively. After 25 epochs, the model can reach a local optimum.

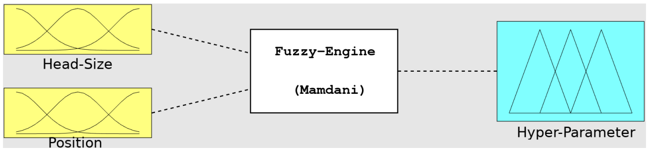

Fig. 2. The fuzzy inference engine, where the head size and corresponding positions are the two inputs and the level of HPs for CCNN is the output.

Table 1. The fuzzy rule table for selecting HPs

Input		Output
Head Size	Position	
Small	Up	High-Pred
Small	Middle	High-Pred
Small	Down	Mid-Pred
Average	Middle	Mid-Pred
Average	Down	Mid-Pred
Average	Up	Low-Pred
Big	Up	Mid-Pred
Big	Down	Low-Pred
Big	Middle	Low-Pred

4. EXPERIMENTAL RESULTS

To evaluate the performance of our A-CCNN algorithm, experiments are conducted on three challenging crowd counting datasets, i.e., the UCSD dataset [10], the UCF-CC dataset [11], and the dataset of Sydney Trains Footage (STF) [12]. Note that the first two are public benchmark datasets.

The Mean Absolute Error (MAE) is used as the evaluation metric for comparing the performance of A-CCNN against the state-of-the-art methods, and it is defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| C_i - C_i^{GT} \right|, \qquad (2)$$

where N is the number of images, C_i is the crowd count predicted by the model being evaluated, and C_i^{GT} is the crowd count from the human annotated one (i.e., ground truth).

4.1. The UCSD Dataset

The UCSD crowd counting dataset consists of 2000 frames of size 238×158 from a single far distance scene. We split the dataset into four subsets of training and testing images in the same way as in [2].

Table 2 presents the MAE results for our proposed A-CCNN and six state-of-the-art methods. As shown in Table 2, our A-CCNN performs competitively against other approaches with the lowest ever MAE of **1.04** and **1.48** for the upscale and minimal subsets respectively. Furthermore, in the other subsets, the results indicate that A-CCNN outperforms the CCNN by more than **8** percent. Overall, proposed model reaches the best average result with MAE of **1.35**.

Table 2. Comparison of the MAE results between A-CCNN and state-of-the-art crowd counting on UCSD crowd-counting dataset [10]

Methods	Max	Down	Up	Min	Avg
Density Learning [13]	1.70	1.28	1.59	2.02	1.64
Count Forest [14]	1.43	1.30	1.59	1.62	1.49
Arteta et al. [15]	**1.24**	1.31	1.69	1.49	1.43
Zhang et al. [16]	1.70	**1.26**	1.59	1.52	1.52
Switch-CNN [5]	-	-	-	1.62	1.62
CCNN [2]	1.65	1.79	1.11	1.50	1.51
A-CCNN	1.51	1.36	**1.04**	**1.48**	**1.35**

4.2. The UCF-CC Dataset

The UCF CC 50 [11] is a small dataset with 50 picture collections of annotated crowd scenes. We have followed the same experimental settings as those of six other state-of-the-art models [5].

In Table 3, the MAE performance of our A-CCNN compared with other methods is shown. As shown in Table 3, proposed approach outperforms four out of six methods and improves the MAE score by more than **24** percentage compared to the original CCNN. Considering its simplicity, A-CCNN's performance is comparable to that of Switch-CNN and Hydra-CCNN.

Table 3. Comparison of the MAE results between A-CCNN and state-of-the-art crowd counting on UCF CC dataset [11].

Methods	MAE
Density learning [13]	493.4
Idrees et al. [11]	419.5
Zhang et al. [16]	467.0
MCNN [17]	377.6
Hydra-CCNN [2]	333.73
Switch-CNN [5]	**318.1**
CCNN [2]	488.67
A-CCNN	367.3

4.3. The Sydney Train Footage

To evaluate the robustness of our model on real-world problems with heavy occlusions, low resolution and large variance in people's sizes, we have utilized CCTV footages of a train station in Sydney and created annotated data for training and testing with our proposed approach. An example is shown in Fig. 1. This dataset has two separate scenes, taken by cameras C5 and C9 with 788 and 600 frames, respectively, with crowd varying between 3 to 65. The sizes of the input frames are 576×704, and the mask and annotation are provided. The huge variation in people's sizes and heavy extreme occlusions make it a very challenging task. Generally, in this dataset, the sizes of people who are in front of the cameras are three to four times larger than the sizes of people in further areas.

Table 4 reports the MAE performance on this dataset. The crowd count of A-CCNN is significantly higher than the original CCNN. This reinforces the fact that utilizing our approach can efficiently manage both the difference in appearances and sizes of people. Thus, the various trained CCNNs employed by A-CCNN can provide precise density maps, independent of the datasets.

Table 4. Comparison of the MAE results between A-CCNN and state-of-the-art crowd counting on STF [12].

Methods	C5	C9
Farhood et al. [12]	2.28	2.67
CCNN [2]	3.90	4.23
A-CCNN	**1.69**	**1.87**

5. CONCLUSION

Aiming to tackle the difficult problem of crowd counting such as scale variance and extreme collusion, we have presented an Adaptive CCNN architecture that takes a whole image as input and directly outputs its density map. The proposed method has made full use of contextual information to generate an accurate density map. To leverage the local information, we have utilized the combination of CNN-based head detection and the fuzzy inference engine to choose an optimal CCNN model adaptively to each patch of the input image. We have achieved noticeable improvements on three challenging datasets, i.e., the UCSD, UCF-CC and the crowd dataset collected by ourselves from a train station in Sydney, and have demonstrated the effectiveness of the proposed approach.

6. ACKNOWLEDGMENT

This work was partly supported by Rail Manufacturing CRC and Sydney Trains with UTS project ID PRO17-3968.

7. REFERENCES

[1] 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*, July 2017.

[2] D. Onoro-Rubio and R.J. Lpez-Sastre, "Towards perspective-free object counting with deep learning," in *Proceedings of the ECCV*. Springer, 2016, pp. 615–629.

[3] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang, "Multi-scale convolutional neural networks for crowd counting," *arXiv preprint arXiv:1702.02359*, February 2017.

[4] V.A. Sindagi and V.M. Patel, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, vol. 14, pp. 1–6.

[5] D.B. Sam, S. Surya, and R.V. Babu, "Switching convolutional neural network for crowd counting," in *CVPR*, 2017, vol. 1/3, p. 6.

[6] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in *Proceedings of the ICIP*. IEEE, 2016, pp. 1215–1219.

[7] S. Kumagai, K. Hotta, and T. Kurita, "Mixture of counting cnns: Adaptive integration of cnns specialized to specific appearance for crowd counting," *arXiv preprint arXiv:1703.09393*, March 2017.

[8] P. Hu and D. Ramanan, "Finding tiny faces," in *Proceedings of the CVPR*. IEEE, 2017, pp. 1522–1530.

[9] E.H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," in *Proceedings of the sixth international symposium on Multiple-valued logic*. IEEE Computer Society Press, 1976, pp. 196–202.

[10] A.B. Chan, Z.S.J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proceedings of the CVPR*. IEEE, 2008, pp. 1–7.

[11] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the CVPR*, 2013, pp. 2547–2554.

[12] H. Farhood, X.S. He, W. Jia, M. Blumenstein, and H. Li, "Counting people based on linear, weighted and local random forest," in *The International Conference on Digital Image Computing: Techniques and Applications*, 2017.

[13] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proceedings of the NIPS*, 2010, pp. 1324–1332.

[14] V.Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in *Proceedings of the ICCV*, 2015, pp. 3253–3261.

[15] C. Arteta, V. Lempitsky, J.A. Noble, and A. Zisserman, "Interactive object counting," in *Proceedings of the ECCV*. Springer, 2014, pp. 504–518.

[16] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the CVPR*, 2015, pp. 833–841.

[17] 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 CVPR*, 2016, pp. 589–597.