Postprint: A Crowd Counting Model for Public Spaces Based on Image Field-of-View Division

Authors: Yuan Jian, Wang Shanshan, Luo Yingwei

Date: 2020-09-28T00:00:00+00:00

Abstract

To address the problem of uneven crowd distribution and varying target scales in public spaces that affect crowd counting accuracy, a public space crowd counting model based on image field-of-view division is proposed. The model first divides the image scene into two regions: near-field and far-field views. For the near-field region, a YOLO-based network is employed for pedestrian detection, and scene constraints are added to avoid duplicate counting across the near-field and far-field regions; for the far-field region, improved MobileNets are utilized to extract crowd density distribution features, and a super-resolution reconstruction module is introduced to enhance the quality of crowd density maps, ultimately obtaining the total crowd count in the entire image by summing the two components. Tested on the Shanghai Tech and Mall datasets, the results demonstrate that the model achieves significant improvements in accuracy and robustness. Experimental results verify the feasibility of the model.

Full Text

Preamble

Vol. 38 No. 3
Application Research of Computers
ChinaXiv Cooperative Journal

Public Place Crowd Counting Model Based on Image Field Division

Yuan Jian, Wang Shanshan, Luo Yingwei
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: To address the challenges of uneven crowd distribution and varying target scales in public places that affect crowd counting accuracy, this paper proposes a novel crowd counting model based on image field division. The model

first partitions the image scene into near-field and far-field regions. For the near-field region, a YOLO-based network is employed for pedestrian detection, with scene constraints added to avoid duplicate counting across the near and far fields. For the far-field region, an improved MobileNets architecture extracts crowd density distribution features, and a super-resolution reconstruction module is introduced to enhance the quality of the crowd density map. The final crowd count for the entire image is obtained by summing the counts from both regions.

Testing on the Shanghai Tech and Mall datasets demonstrates significant improvements in accuracy and robustness. Experimental results confirm the feasibility of the proposed model.

Keywords: crowd counting; convolutional neural network; lightweight

0 Introduction

With socioeconomic development and continuous population growth, crowd activities have become increasingly diverse, and large-scale social gatherings are on the rise. Scenes of crowded populations in public places such as stations, tourist attractions, and shopping malls are commonplace, posing significant challenges to public management and safety. Crowd density is closely related to public safety—excessively high density can easily trigger panic and even stampede incidents. Traditional video surveillance systems require dedicated personnel for monitoring, consuming substantial human resources. If computers could monitor and analyze the number of people in a scene in real time, automatically issuing alerts when crowding trends emerge to enable timely intervention by relevant authorities, it would be of great significance for ensuring public safety in crowded places. However, accurate crowd estimation remains challenging due to complex and uncontrollable environments, irregular crowd distributions, mutual occlusion, uneven illumination, and camera perspective issues. This paper investigates the crowd counting problem in public places and proposes a public place crowd counting model that offers higher accuracy and faster computation speed for crowd recognition in images.

1 Related Work

An increasing number of researchers have begun to focus on crowd counting problems. Current research can be broadly categorized into three approaches: pedestrian detection-based methods, regression-based methods, and deep learning-based methods.

References [2, 3, 5-7] extract full-body pedestrian features (such as Haar wavelets, HOG, and edge features) to train classifiers for detection. These algorithms achieve good results in low-density crowds with few people. References [4, 9, 11, 12] address the problem through local feature-based methods,

where reference [12] extracts head contour features via Haar wavelet transform and employs perspective transformation techniques for more accurate crowd size estimation. This approach shows some improvement for partially occluded crowds, but as crowd density increases and occlusion becomes more severe, the algorithm becomes more time-consuming and its accuracy remains unsatisfactory.

Regression-based methods [13, 14, 17–19] learn a mapping from low-level features (such as edge features [14, 18] and texture features [18]) to crowd counts, establishing regression models between image features and crowd numbers. These methods treat crowds as a whole, successfully addressing occlusion issues but neglecting the importance of pedestrian spatial distribution. Consequently, researchers [15, 16, 20] incorporated spatial information into the learning process by establishing regression models that learn linear mappings between image features and corresponding object density maps. Reference [15] employs linear regression based on image SIFT features to obtain crowd density distribution maps, then integrates the density maps to derive the final crowd count. This method features fast detection under uncomplicated environments while avoiding the difficulties of learning to detect and locate individual object instances. Although density regression-based methods improve counting accuracy to some extent, they still rely on hand-crafted crowd features.

Early crowd research typically involved extracting global or local features. In recent years, with technological advancements and the widespread application of deep learning in computer vision, numerous deep learning-based algorithms have been proposed. Deep Convolutional Neural Networks (CNNs) have become one of the most successful deep models due to their excellent feature learning capabilities. Researchers have gradually begun to consider using deep learning algorithms, particularly deep convolutional networks, to solve crowd counting problems in complex scenes [21–26]. Reference [22] proposes an algorithm that alternately optimizes density map estimation and crowd count estimation (CrowdCNN), marking the first application of deep convolutional networks to cross-scene crowd density estimation and counting. Reference [24] proposes a single-column counting model based on dilated convolutional neural networks, which significantly reduces network parameters and training difficulty while substantially improving crowd counting accuracy and density map reconstruction quality. Reference [26] employs inception-like modules to extract multi-scale head information, using different convolution kernel sizes simultaneously in each convolutional layer and finally obtaining the density map through deconvolution.

In summary, CNN-based methods greatly simplify complex tasks such as foreground segmentation and target detection/localization. However, for crowd estimation in public places, these methods still have certain limitations: (a) CNN-based algorithms are essentially regression-based and more suitable for scenes with relatively uniform crowd density distribution. In reality, crowd flow in public places exhibits high randomness, often characterized by coexisting

high and low densities. Additionally, cameras in public places are typically positioned above crowds, and various shooting angles produce images with diverse perspectives, resulting in different scale changes across different image field regions. Therefore, these algorithms lack universality for such non-uniform scenes. (b) Existing CNN-based algorithms typically address scale variation and mutual occlusion by setting up multi-column networks with different convolution kernel sizes, which makes networks wider and deeper. Moreover, training requires continuous adjustment of kernel sizes to adapt to crowd scale changes, leading to excessive computational load, poor scene adaptability, and inability to perform real-time crowd counting.

Based on the above analysis, this paper proposes a Public place crowd counting model based on Image Field Division (IFDM). This model achieves accurate crowd estimation in public places with stronger scene adaptability, higher computational accuracy, and smaller network scale. Experimental verification demonstrates that the model possesses good generalization capability and strong robustness.

2 Overall Structure of IFDM Model

The overall structure of the IFDM model is shown in Figure 1 [Figure 1: see original paper]. First, the crowd image is divided into near-field and far-field regions based on its depth information map. The depth information contains relative positional information of objects, reflecting their distance from the camera source. The IFDM model employs the method from reference [27] to obtain single-image depth information, then uses local similarity of depth information colors [28] to partition the image into near-field and far-field regions according to local pixel clustering boundaries. For the near-field region, we propose a pedestrian detection counting algorithm with scene constraints (SCPD) that performs pedestrian detection via a YOLO-based network and adds scene constraints to avoid duplicate counting across near and far fields. For the far-field region, we propose a high-quality density map regression integral counting algorithm (HQDPRI) that designs a lightweight network combined with a super-resolution reconstruction module to extract crowd density distribution features and generate high-quality crowd density maps through mapping. The final crowd count for the entire image is obtained by summing the counts from both field regions.

3 Pedestrian Detection Counting Algorithm with Scene Constraints (SCPD)

After extracting the segmentation boundary from the depth image through clustering, it is mapped to the original crowd image for region segmentation. The region partitioning results are shown in Figure 2 [Figure 2: see original paper]. As seen in Figure 2(d), pedestrians in the near-field region have distinct individual features, rich information, and relatively minor occlusion. We first employ conventional feature learning training with a traditional convolutional

network, where the CNN takes static crowd images as input and uses pre-trained weights to generate visual features, followed by the YOLO architecture [29] as the detection module. However, during experiments, we found that when segmenting images based on depth information, the cutting line often splits people near the boundary in half, potentially causing duplicate counting across near and far fields. To solve this problem, we propose the SCPD algorithm based on YOLO network [29] pedestrian detection. After YOLO detection, the algorithm removes detection boxes whose center coordinates fall within restricted areas (invalid regions), thereby avoiding duplicate counting and reducing false detections.

The YOLO network first divides the input image into S×S grid cells, with each cell assigned B initial candidate boxes of different specifications. Predicted candidate boxes are extracted by the convolutional neural network, with the number of candidate boxes per image being S×S×B. The confidence of whether a target exists in a predicted candidate box is set as:

$$Conf(Object) = Pr(Object) \times Pr(Class|Object) \times IOU_{truth}^{pred}$$

where $Pr(Object)$ indicates whether a target needs to be detected exists in the grid, and IOU_{truth}^{pred} represents the intersection-over-union between the ground truth box and the predicted box.

Since most candidate boxes do not contain pedestrians or any targets, to reduce network learning difficulty, we set $Pr(Object) = 0$ for candidate boxes without targets. As we only need to perform pedestrian classification for candidate boxes containing targets, we set the target class $Pr(Class|Object) = 1$.

The specific steps of the SCPD algorithm are as follows:

a) Input the crowd image to be detected.

b) Based on prior spatial distribution information of crowds, use equation (2) to partition regions where false detection may occur. The polyline equation varies with different scenes, and the area between the cutting line and polyline is defined as the invalid region. Setting the probability of the target being a pedestrian as $Pr(person|object)$, the confidence that the candidate box contains a pedestrian is:

$$Conf(person) = Pr(object) \times Pr(person|object) \times IOU_{truth}^{pred}$$

For subsequent processing of detection results, assume the top-left and bottom-right coordinates of a detection box are (x_{min}, y_{min}) and (x_{max}, y_{max}), respectively. Then the center coordinates of this bounding box are:

$$c_x = \frac{x_{min} + x_{max}}{2}, \quad c_y = \frac{y_{min} + y_{max}}{2}$$

If $c_x \in [x_{i-1}, x_i)$ and $c_y < y_i$, it indicates the detection box appears in the invalid region. Such boxes are directly removed, and the person is not counted in the final statistics. The comparison of detection results before and after adding constraints is shown in Figure 3 [Figure 3: see original paper].

c) Count the detection results after constraint addition and output the corresponding number of people in the near-field region.

4 High-Quality Density Map Regression Integral Counting Algorithm (HQDPRI)

Significant differences exist in features among crowds of different densities. Far-field region crowds typically have smaller scales and more severe mutual occlusion, making target detection approaches ineffective. Therefore, this part adopts density map regression for calculation. Reference [21] uses a multi-column parallel network to extract multi-scale head features, but this leads to excessive parameters and many inefficient branch structures. Reference [20] first divides images into patches, then uses a classification network to determine which subnetwork each patch should be fed into. Although achieving good detection results, it shares the same problems as reference [21]: high computational cost and simple patch division that affects counting prediction accuracy. Therefore, this paper proposes the HQDPRI algorithm, which uses a lightweight network combined with a super-resolution reconstruction module to extract crowd density distribution features and generate high-quality crowd density maps through mapping. The final count for this part is obtained by integrating the high-quality density map.

4.1 Lightweight Deep Convolutional Network with Super-Resolution Reconstruction Module

Although region partitioning eliminates the need to continuously adjust convolution kernel sizes to adapt to crowd scale changes, the far-field region still suffers from dense crowd distribution and mutual occlusion. Based on the improved lightweight network MobileNets [30] for feature extraction, HQDPRI introduces a super-resolution reconstruction module to design a new convolutional neural network for crowd counting in images.

The backbone network is improved based on MobileNets, replacing standard convolution operations with depthwise convolutions and $1\$\times 1 pointwise convolutions, comprising 27 convolutionall$ and the mean pooling and fully connected layers are removed, ultimately outputting a density feature map at 1/16 of the original image size. Specific parameter changes are shown in Table 1 . The network avoids pooling layers and instead implements downsampling by setting depthwise convolution stride to 2. This combination significantly reduces parameters and computation while maintaining similar accuracy, improving detection speed. Compared with the commonly used VGG16 network model, it achieves similar computational accuracy but reduces computational complexity by 27 times. To obtain more

accurate precision, a super-resolution reconstruction module is introduced in the latter part of the network to improve density map quality.

4.2 Super-Resolution Reconstruction Module

Super-resolution reconstruction technology enables focused analysis of targets to obtain higher spatial resolution images of regions of interest. Current deep learning-based single-image super-resolution reconstruction has achieved great success in reconstruction efficiency and computational cost. Reference [31] proposes directly processing low-resolution images through convolutional networks for super-resolution and introduces an effective sub-pixel convolutional layer that learns a set of upscaling filters to map low-resolution features to high-resolution outputs. This approach eliminates bicubic interpolation and greatly reduces computational cost. Based on reference [31], this paper introduces super-resolution reconstruction technology into the network structure to optimize density map quality and obtain more accurate computational precision.

The first layer of the super-resolution reconstruction module uses two $3\$\times 3 kernels instead of a 5\times\$5$ kernel. This not only maintains the same receptive field while increasing network depth and enhancing nonlinear feature expression but also improves neural network feature learning effectiveness to some extent. The second and third layers use depthwise separable convolutions instead of standard convolutions, and Batch Normalization is omitted to better adapt to image reconstruction tasks. Through the first two convolutional layers, a feature image with r^2 channels (where r is the target upscaling factor) and the same size as the input image is obtained. The third sub-pixel convolutional layer then rearranges each pixel' s r^2 channels into an $r \times r$ region, corresponding to a sub-block of size $r \times r$ in the high-resolution image. Thus, a feature image of size $H \times W \times r^2$ is rearranged into a high-resolution image of size $rH \times rW$, yielding an optimized crowd density map. This process does not actually involve convolution operations but only transforms image size, making it more efficient.

4.3 HQDPRI Algorithm Steps

The HQDPRI algorithm steps are as follows:

a) Feed the crowd image collection into the improved MobileNets backbone network to extract convolutional features.

b) Use a Gaussian kernel function $G_\sigma(x)$ with standard deviation σ for convolution to obtain the crowd density function $F(x)$ with head coordinates x_i. The calculation formula is:

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

where $\delta(x - x_i)$ represents the head annotation point at coordinate x_i, and the k-nearest neighbor distances of annotation points in the image are denoted as $d_1^i, d_2^i, ..., d_m^i$, with $\bar{d}_i$ being the average distance between the i-th head and its k nearest neighbors. Experiments [21] prove that $\beta = 0.3$ yields the best crowd density map results.

The super-resolution reconstruction module rearranges pixels of the density map $F(x)$ through sub-pixel convolution to improve density map quality. The calculation formula is:

$$PS(F(x)) = F\left(\left\lfloor \frac{x}{r} \right\rfloor, \left\lfloor \frac{y}{r} \right\rfloor\right)_{mod(x,r),mod(y,r)}$$

where PS is a periodic shuffling operator that rearranges elements of a $C \times rH \times rW$ tensor. It periodically activates different sub-pixel positions during filter convolution, with x, y being output pixel coordinates in high-resolution space.

c) Calculate the number of people N_p in this part by integrating and summing the high-quality density map:

$$N_p = \sum_{i=1}^{N} \int \int F_{SR}(x)dx$$

where $F_{SR}(x)$ is the super-resolved density map.

5 Experiments and Analysis

To validate the effectiveness of IFDM, we selected the Shanghai Tech [21] and Mall [6] datasets as experimental data sources. The Mall dataset contains surveillance video data from a shopping mall with relatively small scene variations and sparse crowds. The Shanghai Tech dataset, randomly selected from the internet, features higher crowd density and richer scene variations. Training and test set partition details are shown in Table 2 .

The experimental environment is based on Linux 64 Ubuntu 16.04, using the TensorFlow deep learning framework with a GTX-Titan X GPU.

5.1 Model Training

This paper uses Euclidean distance as the loss function to measure the difference between predicted and ground truth crowd density maps:

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

where θ represents model parameters, N is the total number of training images, X_i is the i-th input image, and $F(X_i;\theta)$ and F_i represent the i-th predicted and ground truth crowd density maps, respectively.

To accelerate model convergence, we use the Adam optimization algorithm with adaptive learning rate, setting the initial learning rate to $1e-5$ and batch size to 4. Based on experience, insufficient training data can easily lead to overfitting. To prevent this, we process training images by cropping each image into four non-overlapping blocks of equal size, effectively expanding the training set by 4 times.

5.2 Evaluation Metrics

Model performance is measured using Mean Absolute Error (MAE) and Mean Squared Error (MSE), as shown in equations (9) and (10):

1) **Mean Absolute Error (MAE)**

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|z_i - z_i'|$$

where N represents the number of test images, z_i' is the crowd count obtained from the predicted density map, and z_i is the actual number of people in the image. MAE indicates the accuracy of network predictions; smaller MAE values mean more accurate crowd count estimation.

2) **Mean Squared Error (MSE)**

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(z_i - z_i')^2$$

MSE reflects the difference between estimates and ground truth; smaller MSE values indicate better algorithm robustness.

5.3 Comparative Experimental Results Analysis

The Shanghai Tech dataset consists of two parts: part_A and part_B, with significant density differences between them. Part_A contains 482 images randomly crawled from the internet with high crowd density, while part_B contains 716 images captured from busy Shanghai streets with medium density but larger crowd distribution variations. The dataset comprises 1,198 annotated images with a total of 330,165 labeled heads. Table 3 shows experimental comparison results on the Shanghai Tech dataset, where references [20–24] are CNN-based methods. The results show that in part_B test set, the proposed model' s MAE decreases by 35.94% and MSE by 34.53% compared with reference [22]; in part_A test set, MAE decreases by 39.38% and MSE by 38.06%. Compared

with reference [21], MAE in part_A is comparable; compared with reference [24], MAE in part_A increases by 8.52%, because part_A contains extremely dense crowds where distinct near/far field regions are difficult to identify. However, MSE still performs excellently compared with other results. These comparisons demonstrate that the proposed model outperforms classical CNN-based algorithms overall, particularly showing better performance under large crowd distribution variations. Figure 4 [Figure 4: see original paper] illustrates experimental results on the Shanghai Tech dataset.

The Mall dataset consists of 2,000 frames of size 640$×$480, with over 60,000 labeled pedestrians. In addition to varying illumination conditions and crowd densities, the dataset suffers from severe perspective distortion, large object scale and appearance variations, and frequent occlusion. Table 4 shows experimental results on the Mall dataset, where references [13, 17] are traditional methods and [20, 32] are CNN-based methods. Compared with the traditional method [17], IFDM reduces MAE by 40.00% with more significant MSE improvement; compared with the CNN-based method [20], MAE decreases by 16.73% with notable MSE improvement. The Mall dataset has relatively fixed scene variations and sparse crowds per image, and experimental results demonstrate that the model can also achieve accurate estimation and higher robustness for relatively sparse crowd images. Figure 5 [Figure 5: see original paper] illustrates experimental results on the Mall dataset.

5.4 Validation Experimental Analysis

To verify the impact of the super-resolution reconstruction module on model performance, this section analyzes both running speed and performance metrics after removing the module. Table 5 compares performance metrics of the proposed algorithm with and without the super-resolution reconstruction module on Shanghai Tech dataset part_B. The results show that without the super-resolution reconstruction module, MAE increases by 39.5% and MSE by 53.3%.

Table 6 compares total parameters, computation, and running speed with and without the super-resolution reconstruction module for 224$×$224 input images. The data show that adding the super-resolution reconstruction module does not significantly increase parameters or computation. This is because the backbone network is lightweight with inherently fewer parameters and computations than conventional networks, and our improved sub-pixel convolutional layer also greatly reduces parameters and computational complexity. Therefore, the model with the super-resolution reconstruction module maintains fast running speed.

In summary, the model with the super-resolution reconstruction module, although increasing some computational load and reducing running speed compared with the module-free version, effectively improves predicted density map quality and significantly enhances model performance metrics, yielding more accurate predictions.

6 Conclusion

Public place crowd counting is a challenging topic in crowd behavior research and a focus of public safety studies. Public places often contain multiple moving objects of small size with similar appearance, along with mutual occlusion, uneven illumination, and camera distortion, making crowd counting analysis extremely difficult. To better address this problem, we propose using different counting algorithms for different field regions, obtaining the final prediction by summing counts from near and far fields. Experimental analysis shows that although the proposed model achieves significant improvement over existing methods, it still faces challenges for some extremely dense scenes, particularly where near/far field regions are difficult to partition. Future work will focus on improving network structure to adapt to extremely dense crowds and enabling real-time video image analysis. By automatically and reliably obtaining crowd counts or densities from surveillance footage, we aim to provide comprehensive dynamic estimates of crowd flow status, direction, and duration to help staff optimize management.

References

[1] Zhang Junjun, Shi Zhiguang, Li Jicheng. Current researches and future perspectives of crowd counting and crowd density estimation technology [J]. Computer Engineering and Science, 2018, 40(2): 282-291.

[2] Wojek C, Dollar P, Schiele B, et al. Pedestrian Detection: An Evaluation of the State of the Art [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2012, 34(4): 743-761.

[3] Enzweiler M, Gavrila D M. Monocular pedestrian detection: survey and experiments [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31: 2179-2195.

[4] Li Min, Zhang Zhaoxiang, Huang Kaiqi, et al. Estimating the Number of People in Crowded Scenes by MID Based Foreground Segmentation and Head-shoulder Detection [C]// International Conference on Pattern Recognition. IEEE, 2009: 1998-2001.

[5] Leibe B, Seemann E, Schiele B. Pedestrian Detection in Crowded Scenes [C]// Proc of Computer Vision and Pattern Recognition, 2005: 878-885.

[6] Chen Ke, Loy C, Gong Shaogang, et al. Feature Mining for Localised Crowd Counting [C]// British Machine Vision Conference, 2012: 3.

[7] Chen Rui, Peng Qimin. Pedestrian detection method based on gradient direction histogram of stable region [J]. Journal of Computer-Aided Design and Computer Graphics, 2012, 24(3): 372-377.

[8] Fan Chunnian, Du Weiping, Liu Yanrong. Pedestrian detection based on HOG features and AdaBoost algorithm [J]. Automation Technology and Application, 2018, 37(07): 89-91.

[9] Wu Bo, Nevatia R. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors [C]// the 10th IEEE International Conference on Computer Vision. ICCV, 2005: 1270-1277.

[10] Viola P, Jones M J. Robust Real-Time Face Detection [J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[11] Felzenszwalb P F, Girshick R B, Mcallester D, et al. Object Detection with Discriminatively Trained Part-Based Models [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[12] Lin Shengfu, Chen J, Chao Hungxin. Estimation of number of people in crowded scenes using perspective transformation [J]. IEEE Trans on Systems Man & Cybernetics Part A-Systems and Humans, 2001, 31(6): 645-654.

[13] Chan A B, Vasconcelos N. Counting People With Low-Level Features and Bayesian Regression [J]. IEEE Trans on Image Processing, 2012, 21(4): 2160-2177.

[14] Ryan D, Denman S, Fookes C, et al. Crowd Counting Using Multiple Local Features [C]// Digital Image Computing: Techniques and Applications. IEEE, 2009: 81-88.

[15] Lempitsky V S, Zisserman A. Learning To Count Objects in Images [C]// Neural Information Processing Systems, 2010: 1324-1332.

[16] Fiaschi L, Nair R, Koethe U, et al. Learning to count with regression forest and structured labels [C]// International Conference on Pattern Recognition, 2012: 2685-2688.

[17] Chen Ke, Gong Shaogang, Xiang Tao, et al. Cumulative attribute space for age and crowd density estimation [C]// Proc of Computer Vision and Pattern Recognition, 2013: 2467–2474.

[18] Li Haifeng, Jiang Zizheng, Fan Longfei, et al. Population estimation algorithm based on density classification and combination characteristics [J]. Computer Application Research, 2018, 35(06): 1891-1895.

[19] Li Jun, Tao Dacheng. A Bayesian Hierarchical Factorization Model for Vector Fields [J]. IEEE Trans on Image Processing, 2013, 22(11): 4510-4521.

[20] Sheng Biyun, Shen Chunhua, Lin Guosheng, et al. Crowd counting via weighted VLAD on dense attribute feature maps [J]. IEEE Trans on Circuits and Systems for Video Technology, 2018: 1788-1797.

[21] Zhang Yingying, Zhou Desen, Chen Siqi, et al. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 589-597.

[22] Zhang Cong, Li Hongsheng, Wang Xiaogang, et al. Cross-scene crowd counting via deep convolutional neural networks [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 833-841.

[23] Marsden M, Mcguinness K, Little S, et al. Fully Convolutional Crowd Counting on Highly Congested Scenes [C]// the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), 2016: 27-33.

[24] Li Yuhong, Zhang Xiaofan, Chen Deming. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1091-1100.

[25] Fan Lvyuan, Tong Minglei, Li Min, et al. Population density estimation using mixed convolution structure in still images [J/OL]. Computer Application Research: 1-6 [2020-03-25]. https://doi.org/10.19734/j.issn.1001-3695.2018.06.0661.

[26] Cao Xinkun, Wang Zhipeng, Zhao Yanyun, et al. Scale aggregation network for accurate and efficient crowd counting [C]// Computer Vision. 15th European Conference (ECCV 2018), 2018: 757-763.

[27] Eigen D, Puhrsch C, Fergus R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network [C]// the 28th Conference on Neural Information Processing Systems (NIPS), 2014, 27: 2366–2374.

[28] Achanta R, Shaji K, Smith, et al. SLIC superpixels. EPFL Technical Report 149300, 2010.

[29] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 779–788.

[30] Howard A G, Zhu Menglong, Chen Bo, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [C]// Computer Vision and Pattern Recognition. 2017: 1-9.

[31] Shi Wenshi, Caballero J, Huszár F, et al. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 1874-1883.

[32] Kumagai S, Hotta K, Kurita T. Mixture of counting CNNs: adaptive integration of CNNs specialized to specific appearance for crowd counting [C]// Proc of Computer Vision and Pattern Recognition, arXiv:1703.09393, 2017.

Note: Figure translations are in progress. See original paper for figures.