CROWD_SZ: A Large-scale Multi-view Crowd Counting Semantic Dataset

Jiajia Wu
School of Rail Transportation
Soochow University
Suzhou, Jiangsu, China
20194246023@stu.suda.edu.cn

Yufeng Lin
Public Security Bureau Command Center
Suzhou Industrial Park
Suzhou, Jiangsu, China
18913585632@163.com

Cheng Wu
School of Rail Transportation
Soochow University
Suzhou, Jiangsu, China
cwu@suda.edu.cn

Jin Zhang
School of Rail Transportation
Soochow University
Suzhou, Jiangsu, China
zhangjin1983@suda.edu.cn

Lijun Zhang
School of Rail Transportation
Soochow University
Suzhou, Jiangsu, China
zhanglijun@suda.edu.cn

***Abstract*—In recent years, there have been many large social gatherings and stampedes. High-density crowd counting and density estimation have become a research hotspot in the field of video surveillance. However, traditional datasets expose the limitations of a single perspective and limited crowd size, which cannot meet the research needs of a wide-area place. This paper proposes a new large-scale multi-view dataset, taking the urban life square near Jinji Lake in Suzhou city, Jiangsu Province, China as the research object. A single camera cannot cover the whole place, so we collect surveillance images from multiple perspectives. The low-altitude monitoring image has obvious human characteristics, while the high-altitude monitoring image provides the trend of crowd distribution. Combining these two kinds of information, the trend of crowd change can be predicted more accurately. This dataset is characterized by rapid crowd change in a short time, large aggregation scale and complex illumination conditions, which brings new challenges to crowd counting research.**

***Index Terms*—Crowd counting; Semantic understanding; Data fusion.**

I. Introduction

With the rapid development of urbanization, more large-scale competitions, cultural exchanges, and entertainment activities are held. The actual crowd is often greater than the number of people that can be accommodated in the venue, which has caused a series of unexpected safety problems [1]. To improve event management and safety, related research has shifted focus to the field of crowd technology [2]. Thanks to the widespread use of video surveillance systems, the all-around installation of video capture equipment provides more research data for the field of crowd density estimation, making it possible to accurately count crowds in dense places. Different from other datasets, the pedestrian features in the crowd dataset are small and fuzzy, which makes it more difficult to capture. In addition, changes of perspective, over-dark or over-exposed environmental illumination, and crowd occlusion hurt feature extraction, as shown in Figure 1.

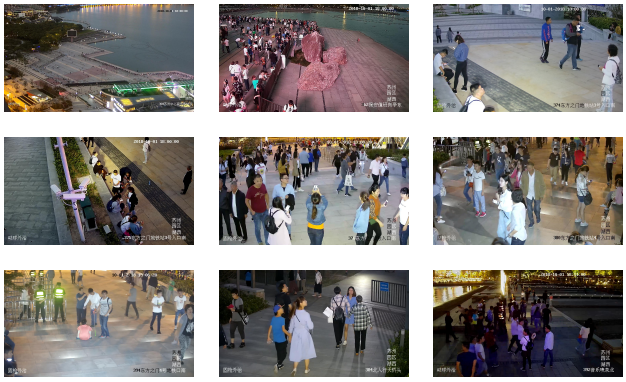


Figure 1. Example from CROWD_SZ.

At present, a lot of research is devoted to crowd counting in natural scenes. Some public places such as squares and stations also have a strong demand for accurate real-time crowd counting [3]. Traditional crowd counting methods are mainly divided into two categories: detection-based methods [4]–[7] and regression-based methods [8]–[10], [21]. Early crowd studies mainly used detection-based methods. First, assuming that the crowd is composed of individuals, use the sliding window detector to detect the crowd and calculate the crowd count. In a dense crowd environment, however, detection-based methods are difficult to solve the problem of serious occlusion inside crowds. Therefore, regression-based methods were introduced. By learning the mapping of a feature to the number of people, these methods extract the foreground features and then use a regression function to estimate the number of people. The mapping process usually adopts the linear [12] or piecewise linear [11] function model. In recent years, more attention is paid to the crowd counting technology based on deep learning. It is different from the traditional crowd counting method, which uses multiple convolutional

Copyright (c) IARIA, 2022. ISBN: 978-1-61208-941-6

neural networks to extract head features of different sizes. This method can get better prediction results for densely distributed regions. Zhang et al. [13] proposed a crowd counting model (Crowd CNN) based on a six-layer convolutional neural network. By alternately optimizing the true value of the crowd count and the true density map of the crowd, better robustness was obtained. Bai et al. [28] extracted the global coarse-grained motion of the crowd from the perspective of high altitude. The local fine-grained density characteristics of people with line of sight occlusion are extracted from the low altitude perspective, and a holographic model of the temporal and spatial evolution of crowd situation is established. Fu et al. [14] designed two ConvNet cascaded classifiers to estimate crowd density by optimizing the convolutional neural network. Zhang et al. [15] proposed a Multi-Column Convolutional Neural Network (MCNN) structure to extract multi-scale image features. After image inputting, the crowd density estimation map was obtained. Finally, the density map was integrated to obtain the estimated crowd count value. Hu et al. [16] used the ConvNet structure to extract crowd features and then used two signals of crowd count and crowd density to learn crowd characteristics and estimate the number of people. Jiang et al. [17] focused on the improvement of density maps, using additional multi-scale markers to increase the diversity of deep neural networks, and achieve the high-performance crowd density map estimation. From the traditional detection-based and regression-based methods to the application of deep learning frameworks, most of the current methods count by extracting human head features of different scales. Unfortunately, such methods cannot effectively deal with the problem of target detection in dense crowds in natural scenes. The main problems are:

- The change of perspective and the position of equipment will lead to a great change in the size of the human head in dense crowds, and then the feature extraction method based on a single or finite-size convolution kernel has difficulty extracting the full size of a human head.
- The number of targets contained in a dense crowd varies greatly, and the number changes significantly in a short time. An image can contain several thousands of people. Therefore, the method based on multi-frame fusion detection can not be applied to the estimation of the number of people with obvious differences.
- Research objectives in a dense crowd are often unevenly and irregularly distributed. It is difficult to describe the change in the global situation using methods based on statistical reasoning.

In view of the typical weaknesses of the dense crowd target detection methods, existing studies lack of dense crowd scenarios to meet the research needs. Some existing crowd datasets are mostly ideal, with a single scene and unchanged illumination, which cannot fully reflect the complexity of the problem, such as UCSD [18], Mall [19], WorldExpo'10 [20], UCF_CC_50 [13], Shanghaitech [15] and so on. UCSD [18] is the first dataset in the crowd counting field. It consists of 2,000 frames of images and pedestrian annotations in each frame, and the video frames are extracted from a single scene. The density of images in this dataset is low, with an average of 15 people in each image. Chen et al. [19] collected a new dataset Mall with different illumination conditions and crowd density. The images in this dataset have a higher crowd density. But, like UCSD, they are all part of a single video sequence, so the scene does not change. WorldExpo'10 [20] contains 108 videos in 5 different scenes, with a total of 3,980 frames of images. UCSD and Worldexpo'10 contain only low and medium density scenarios and lack high-density scenarios. The UCF_CC_50 dataset [13] specializes in collecting ultra-highdensity crowd scenes and contains only 50 images. The generalization ability of the training network with a limited number of training samples is reduced, which affects the test results. The Shanghaitech dataset [15] has better diversity than previous datasets in terms of scenes and density levels. It is divided into two parts: part_A (including images of a high-density crowd) and part_B (including images of a lowdensity crowd). It contains 1,198 images with 330,165 annotations. Although these datasets provide some images for counting, they are lacking in sample number, image complexity, and scene diversity. Qi Wang et al. [22] proposed a new dataset Nwpucrowd, with 5,109 images and 2,133,238 annotations, which has been greatly improved in terms of quantity and provides a platform for researchers to compare the calculation results of the test set. Sindagi et al. [23] proposed a dataset called JHU-CROWD++, which was collected under different scenes and environmental conditions, including some images based on severe weather and illumination changes. However, these two datasets did not fully consider background interference factors in the nature scene, and the identification of human body contour is very unsatisfactory in the high-altitude image with an ultra density of the crowd.

This is the motive for our work. In response to the existing problems of the above datasets, this article introduces a large-scale multi-view crowd counting dataset. Table 1 describes the parameter comparison between our dataset and other typical crowd datasets. We collected 5,610 images and 1,738 video files from the monitoring equipment. The images come from different heights and angles (including two high perspectives and nine low perspectives), and the illumination of the scene changes significantly at night. In terms of perspective selection, to facilitate the research of multi-perspective fusion algorithm, the dataset fully considers the basic principle that a high-altitude perspective must include a low-altitude perspective coverage area, which mainly reflects the interrelation

TABLE I. COMPARISON BETWEEN CROWD_SZ AND OTHER CROWD DATASETS.

Dataset		Resolution	Images	Min	Max	Multi-view	Density change
Shanghaitech	Part_A	different	482	33	3,139	✓	✓
	Part_B	768*1,024	716	9	578		
UCF_CC_50		different	50	94	4,543	✓	✗
UCSD		158*238	2,000	11	46	✗	✗
Mall		480*640	2,000	13	53	✗	✓
WorldExpo_10		576*720	3,980	1	253	✓	✓
CROWD_SZ		different	5,610	1	673	✓	✓

(a) Bounding box annotation (b) Human head annotation

Figure 2. Annotation.

and spatial complementarity of different perspectives. At the same time, the images selected from the dataset also highlight the different crowd density levels, illumination conditions, perspective distortion, and other conditions.

In summary, the main contributions of this article are as follows:

- We propose a new large-scale multi-view dataset for crowd counting and density estimation. The dataset includes 5,610 images and 1,738 videos, which makes up for the lack of diversity in traditional datasets;
- For the stock data of the dataset, a large number of complete head annotation files are carefully prepared;
- For the key elements of dense crowd vision, different scene video frames with good statistical dispersion are provided;
- Fully considering the basic principle of "the high-altitude perspective must include the low-altitude perspective coverage area", our dataset provides global and local images, and allows to calculate the number of people in wide-area places according to the time correlation and spatial complementarity between them.

The rest of the paper is structured as follows. Section II describes the annotation method and the classification of the dataset. Section III gives the specific nature of the dataset and some statistical information of the data. Section IV provides the experimental procedures and data of the two methods, and performs crossdataset verification to verify the generalization ability of the dataset. Section V concludes the paper.

II. Annotation

A. Dataset

In the crowd dataset, image acquisition equipment and acquisition scenarios are the main reasons for the deviation of the dataset. To eliminate the deviation, we collected videos and images in the CROWD_SZ dataset from monitoring equipment at different locations in Suzhou Jinji Lake Fountain Square. We recorded the specific location of each perspective and the specific time of each image to ensure that the image time between different perspectives is consistent with the subject of observation. For video files with different perspectives, we save a one-minute video as one file, which also guarantees consistency between different perspectives.

B. Classification

CROWD_SZ is divided into image sets and video sets. The image set contains high-altitude images and low-altitude images, with high-altitude images having two perspectives. The low-altitude images are divided into nine perspectives. Each folder contains 510 images and the corresponding pedestrian header annotation file. Video sets and image sets follow the same classification criteria; Each subclass contains 158 video files and each subclass contains a 1-minute video. We divided the annotation files into training set and test set according to a 3:1 ratio.

C. Annotation method

In the direction of pedestrian detection, many datasets use bounding box annotation [24] and pedestrian torso line annotation [25]. In our dataset, there is a large number of images with pedestrian occlusion and overlap. If the above two annotation methods are still used, multiple boundary boxes and pedestrian trunk lines will overlap extensively, and the pedestrian in the image cannot be accurately detected, resulting in error counting results. Considering the counting requirements, we annotate the head of the pedestrian in the image. As shown in Figure 2, the head position of each pedestrian is marked with a red cross. The marking process is mainly divided into two parts: crowd image labeling and transforming crowd image labeling into a crowd density map.

The label density map generation process is as follows: First, x_i represents the center coordinates of the human head. If there is a human head at a specific position x_a, it can be expressed as $\delta(x - x_a)$, which means that there is a actual person at the x_a coordinate position. If there are N heads in a picture, then this picture can be represented as follows:

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

The density map of the image is generated by the function and gaussian kernel convolution. Due to perspective distortion in the scene, the size of each human head needs to be considered to determine the diffusion function before generating the density map. The size of the head is usually the distance between the centers of two adjacent people. Therefore, we adopt an adaptive method to determine the parameters of each person. The formula for generating the final density map is as shown in (2), where G represents the Gaussian kernel, σ is the standard deviation of the Gaussian kernel, and β is a set value, usually 0.3. Suppose there are k people around this person. d represents the average of the sum of the Euclidean distances of the head from its k neighboring heads.

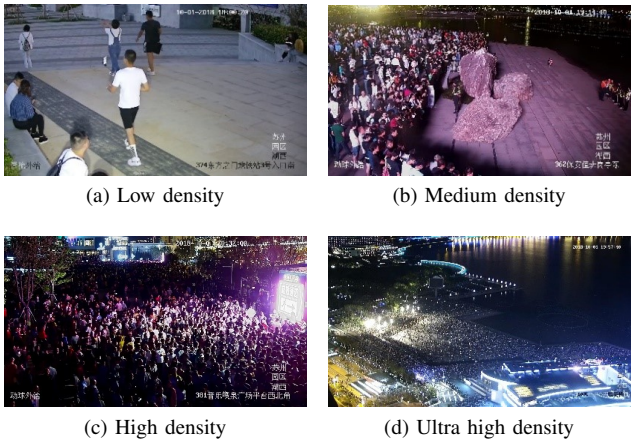
$$F(x) = \sum_{i=1}^{N} \delta(x - x_i\ ) * G_{\sigma_i}\ (x), with\ \sigma_i = \beta \overline{d_i} \tag{2}$$

III. The nature of CROWD_SZ

A. Image collection location and pixel size

Compared with other datasets, the original sizes of the images in our dataset are $1,920 * 1,080$, $2,560 * 1,440$, etc., while the image sizes of other existing datasets mostly do not exceed $1,000 * 1,000$. For example, the image size in UCSD is $158 * 238$. The picture size in Mall is $480 * 640$.

Copyright (c) IARIA, 2022. ISBN: 978-1-61208-941-6



(a) Low density (b) Medium density

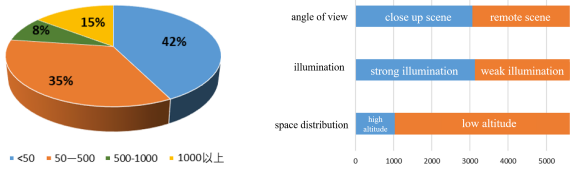
(c) High density (d) Ultra high density

Figure 3. Examples of crowd images with different density levels.

(a) High altitude image (b) Strong illumination image (c) Close up scene

(d) Low altitude image (e) Weak illumination image (f) Remote scene

Figure 4. Examples of different types of images in CROWD_SZ.



(a) (b)

Figure 5. The characteristics of CROWD_SZ.

B. Density distribution

In terms of density (here, we define density as the number of people contained in a single image.), as shown in Figure 3, we divide it into four density levels:

- Low density: the count value is between zero and fifty.
- Medium density: the count value is between fifty and five hundred.
- High density: the count value is between five hundred and one thousand.
- Ultra high density: count value above one thousand.

Compared with other datasets with a single density distribution, our dataset has made progress in the diversity of density. We counted all the pictures included in the dataset, and finally got the distribution of people as shown in Table II. In this dataset, low-density images account for a relatively high proportion of 42%, medium-density images account for 35%, high-density images account for 15%, and ultra high-density images account for 8%, which is enough to meet the research of different density images demand.

C. Diversity

Scene diversity is an important attribute of the dataset. Our dataset contains images taken by cameras at different heights and angles. The scenes, illumination, and pedestrian forms are diverse. As shown in Figure 4, the images in the dataset can be divided into a high-altitude image and low-altitude image according to spatial distribution, strong illumination image, and weak illumination image according to illumination conditions, and close-up scene image and remote scene image according to perspective. Figure 5 depicts the statistical data of the above-mentioned scene diversity. In terms of spatial distribution, high-altitude image data accounts for 1/5 of all data, while low-altitude image data accounts for 4/5 of all data. In the illumination condition category, the number of images with strong illumination accounts for 1/2 of the total, while the other 1/2 are images with weak illumination. Similarly, under the perspective position category, the number of close-up images and remote images each account for 1/2 of the total.

TABLE II. DIFFERENT DENSITY DISTRIBUTION IN CROWD_SZ.

Density	Low	Medium	High	Ultra high	Total
No. of images	2,389	1,941	460	820	5,610

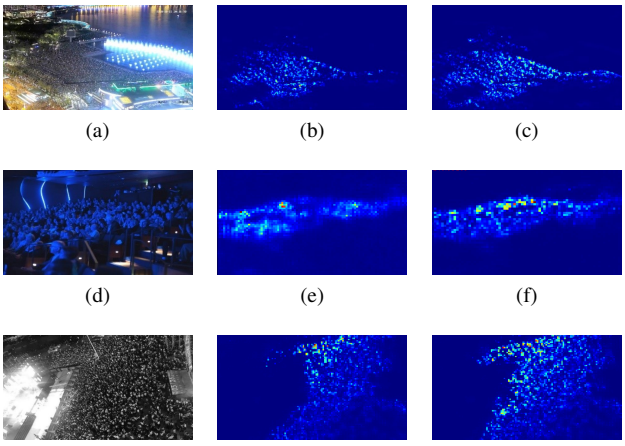
IV. EVALUATION

In this section, we evaluate some typical crowd counting algorithms on the CROWD_SZ dataset. We choose MCNN [15] and CSRNet [26] as the benchmark algorithms for low-altitude image processing in the dataset. Here, MCNN uses three columns of different scale convolution kernels to adapt to different scales of human head sizes, and finally combines the three columns of neural networks to obtain a density map. The CSRNet network model is divided into a front-end network and a back-end network. The front-end network will use VGG [27] (Visual Geometry Group Network) with the fully connected layer removed, and the back-end network will use a hollow convolutional neural network. Its purpose is to generate high-quality crowd density maps while maintaining the resolution while expanding the perceptual field.

A. Image preprocessing

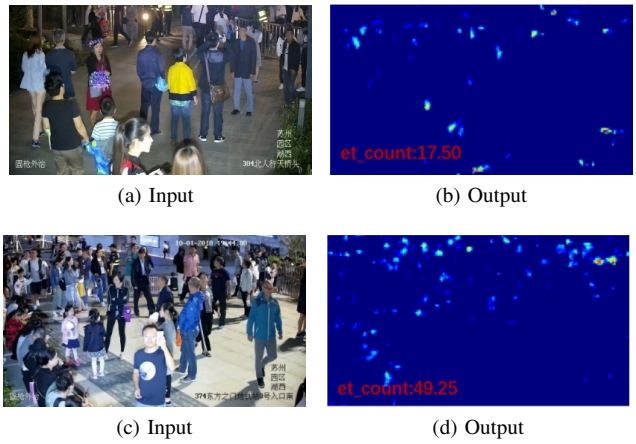
One of the advantages of our dataset is that it contains many night scenes. Take the challenging scenes shown in Figure 6 as an example. To recognize the pedestrian characteristics better, we preprocessed the images, the image needs to be grayed first to obtain the density map. This preprocessing

Copyright (c) IARIA, 2022. ISBN: 978-1-61208-941-6



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 6. Image preprocessing: (a)(d)(g) is the original image, (b)(e)(h) is the density map obtained from the input original image, and (c)(f)(I) is the density map obtained by graying the original image.



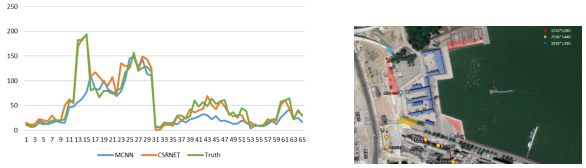
(a) Input (b) Output (c) Input (d) Output

Figure 7. Input and output.

step helps to identify the crowd more comprehensively and carefully, thus making the numerical calculation more accurate. We also selected images from other datasets at night or in week-illumination for verification. Experimental results show that grayscale pretreatment can improve the accuracy of recognition.

B. Experimental analysis

To evaluate two crowd counting algorithms based on the CROWD_SZ dataset, we randomly selected 65 images in the dataset, including 31 images with relatively sparse scenes and 34 images with crowded scenes. Two algorithms, MCNN and CSRNet, are used to estimate the crowd density. The experimental results are shown in Figure 7. We select a low-altitude image as the input data to get the specific number of people and the output density map.



(a) (b)

Figure 8. (a) is the comparison of prediction results of MCNN and CSRNet. (b) is the installation location distribution and corresponding resolution of monitoring devices.

As shown in Figure 8, we list the estimated number of people in the same image obtained by two algorithms, both of which predict the image with higher accuracy. The detailed description is shown in Table III, we use MAE (Mean Absolute Error) as evaluation criteria, the smaller the value, the higher the estimation accuracy. In crowded scenarios, CSRNet has a higher estimation accuracy. On the contrary, MCNN has a higher estimation accuracy in sparse scenes. Select the high-altitude density map and the low-altitude density map of the same time. Finally, as shown in Figure 9, comparing these two density maps, we can find that the local density is consistent with the high-altitude density, which also shows that our count is more accurate and reflects the current crowd density.

At the same tine, we choose the JHUCROWD++ for cross-dataset induction. Compared with other crowd datasets, JHU-CROWD++ has a large number of images collected under different environmental conditions and scenes, including severe weather changes and illumination changes, which makes it very challenging. We randomly select ten images and evaluate the accuracy of their count. The results are shown in Table IV. We used MCNN and CSRNet to estimate the number of people in the image. The estimated results are the same, which also shows that our dataset has good applicability. However, there are also significant differences between the estimates of MCNN and CSRNet. This shows that CROWD_SZ is more challenging.

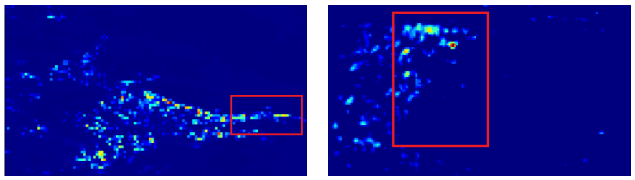
TABLE III. COMPARISON OF MAE VALUES BETWEEN MCNN AND CSRNET IN DIFFERENT SCENARIOS.

Scene \ MAE	MCNN	CSRNet
Sparse scene	0.02839	0.1429
Crowded scene	0.2979	0.1415

TABLE IV. COMPARISON OF MCNN AND CSRNET'S ESTIMATED COUNT OF IMAGES IN TWO DATASETS.

	CROWD_SZ			JHU-CROWD++		
Number	Truth	MCNN	CSRNet	Truth	MCNN	CSRNet
1	7	9.48	11.42	84	82.93	98.65
2	21	17.47	30.39	41	55.22	53.26
3	408	91.08	246.75	13	11.19	14.04
4	39	39.02	32.87	68	28.17	79.23
5	44	39.6	39.71	69	50.32	120.05
6	65	71.63	67.98	42	37.56	58.18
7	17	15.02	10.59	7	5.91	13.02
8	53	46.24	62.52	188	153.5	272.98
9	13	11.48	16.9	65	66.5	81.98
10	23	15.32	24.27	50	47.62	81.52

Copyright (c) IARIA, 2022. ISBN: 978-1-61208-941-6



(a) High altitude image (b) Low altitude image

Figure 9. Density contrast map of the same area in high altitude and low altitude.

V. Conclusion and Future Work

In this work, we introduce a new challenging crowd dataset for large crowd gatherings. The dataset consists of 5,610 images and 1,738 videos, including a large number of night scenes. This dataset contains more images than the existing crowd dataset and has good temporal and spatial correlations between images under different spatial positions, illumination changes, and perspective changes. In addition, we provide a complete header annotation markup file. In this paper, by using a typical crowd counting target detection algorithm and cross-dataset validation process, it is proved that the dataset is larger, more diverse, and more challenging, which is suitable for practical application and can be used as the basis of dense crowd counting research dataset. In the future, we will use spatial complementarity between high altitude images and low altitude images to count vast areas.

VI. Aknowledgement

This work is a part of the project ”Design and consulting service scheme for safety control system of crowd flow in large urban complex and crowd concentrated place”. Thanks to Cheng wu, the corresponding author, for giving me the opportunity to participate in the meeting.

References

[1] L. Soomaroo and V. Murray, ”Disasters at Mass Gatherings: Lessons from History,” Plos Currents, 2012, pp. 1-10.

[2] J. Shao, K. Kang, C. C. Loy and X. Wang, ”Deeply learned attributes for crowded scene understanding,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4657-4666.

[3] E. Stromatias, ”Developing a supervised training algorithm for limited precision feed-forward spiking neural networks,” Computer Science, 2011, pp. 1-107.

[4] M. Li, Z. Zhang, K. Huang and T. Tan, ”Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection,” 2008 19th International Conference on Pattern Recognition, 2008, pp. 1-4.

[5] L. Wang, L. Xu and M. H. Yang, ”Pedestrian detection in crowded scenes via scale and occlusion analysis,” 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 1210-1214.

[6] M. Enzweiler and D. M. Gavrila, ”Monocular Pedestrian Detection: Survey and Experiments,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 12, 2009, pp. 2179-2195.

[7] O. Tuzel, F. Porikli and P. Meer, ”Human Detection via Classification on Riemannian Manifolds,” 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-8.

[8] U. Khan and R. Klette, ”Logarithmically improved property regression for crowd counting,” 2016, pp. 123–135.

[9] A. B. Chan and N. Vasconcelos, ”Bayesian Poisson regression for crowd counting,” 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 545-551.

[10] K. Chen, C. C. Loy, S. Gong and T. Xiang, ”Feature Mining for Localised Crowd Counting,” British Machine Vision Conference. 2012, pp. 1-12.

[11] N. Paragios and V. Ramesh, ”A MRF-based approach for real-time subway monitoring,” Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, pp. 1034-1040.

[12] V. Rabaud and S. Belongie, ”Counting Crowded Moving Objects,” 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), 2006, pp. 705-711.

[13] C. Zhang, H. S. Li, X. Wang and X. K.Yang, ”Cross-scene crowd counting via deep convolutional neural networks,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 833-841.

[14] F. Min et al., ”Fast crowd density estimation with convolutional neural networks,” Engineering Applications of Artificial Intelligence, 2015, pp. 81-88.

[15] Y. Zhang, D. Zhou, S. Chen, S. Gao and Y. Ma, ”Single-Image Crowd Counting via Multi-Column Convolutional Neural Network,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 589-597.

[16] Y. Hu, H. Chang, F. Nian, Y. Wang and T. Li, ”Dense crowd counting from still images with convolutional neural networks,” Journal of Visual Communication and Image Representation, 2016, 530-539.

[17] H. Y. Jiang and W. Jin, ”Effective use of convolutional neural networks and diverse deep supervision for better crowd counting,” Applied Intelligence , 2019, pp. 2415–2433.

[18] A. B. Chan, Z. S. Liang and N. Vasconcelos, ”Privacy preserving crowd monitoring: Counting people without people models or tracking,” 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-7.

[19] C. C. Loy, S. Gong and T. Xiang, ”From Semi-supervised to Transfer Counting of Crowds,” 2013 IEEE International Conference on Computer Vision, 2013, pp. 2256-2263.

[20] H. Idrees, I. Saleemi, C. Seibert and M. Shah, ”Multi-source Multi-scale Counting in Extremely Dense Crowd Images,” 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2547-2554.

[21] Y. Yuan, Z. Y. Wu and X. Z. Jiang, ”Research on Crowd Counting via Deep Convolution Neural Networks,” Group Technology & Production Modernization, 2017, pp. 49-53.

[22] Q. Wang, J. Gao, W. Lin and X. Li, ”NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 6, 2021, pp. 2141-2149.

[23] V. Sindagi, R. Yasarla and V. M. M. Patel, ”JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, vol. 14, no. 8, pp. 1-17.

[24] W. Lee, J. Na and G. Kim, ”Multi-Task Self-Supervised Object Detection via Recycling of Bounding Box Annotations,” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4979-4988.

[25] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei and Y. Sheikh, ”OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, 2021, pp. 172-186.

[26] Y. Li, X. Zhang and D. Chen, ”CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1091-1100.

[27] Z. Yu et al., ”Convolutional networks with cross-layer neurons for image recognition,” Information Sciences, 2018, vol. 433–434, pp. 241-254.

[28] L. Bai, C. Wu, Y. M. Wang and F. Xie. ”Multiview-Fusion-Based Crowd Density Estimation Method for Dense Crowd,” The Fifteenth International Conference on Systems(ICONS), 2020, pp. 50-55.

Copyright (c) IARIA, 2022. ISBN: 978-1-61208-941-6