Current researches and trends of crowd counting in the field of deep learning

Zhi Li[1] ,Yong Li[1,2] ,Xipeng Wang[1]
1. Postgraduate Brigade, Engineering University of PAP
2. College of Information Engineering, Engineering University of PAP
Xian, China
lee_lifestyle@163.com, lilili819@163.com, wanghp369@sina.com

***Abstract*—Crowd counting is a branch of computer vision, which has important practical significance. This paper summarizes the current situation and development trend of crowd counting based on computer vision. Firstly introducing the methods of crowd counting with shallow learning, and then presenting the research status and trend of deep learning in crowd counting. Then, the performance evaluation of crowd counting method is introduced, and several benchmark datasets for crowd counting test and assessment are provided. Finally, the focus of research at the current stage is affirmed, and the prospect is made.**

Keywords—crowd counting, deep learning, shallow learning

I. INTRODUCTION

It has always been relevant research counting the number of objects (especially people) in images and videos. It is also often used in real-world applications, including crowd management, architectural and spatial design and analysis, and security. In some cases, it is vital to know the number of people, such as at public gatherings, marathons, parks and transport hubs. However, crowd counting in a high dense scene is a challenging task, which needs[1]to be performed by some experienced person. With the development of the computer, people hope that all the work can be done automatically by machine, and has continuously researched and developed the crowd counting based on computer vision. At present, counting people in images or videos by deep learning method has become the mainstream method. Based on the traditional shallow way, this paper supplement the research status and development trend of crowd counting in the field of deep learning.

II. EARLY RESEARCH

In the early stage, researchers mainly extract features from the input image based on the shallow learning method, create the mapping relationship between the number of people and features, to obtain the number of people in the picture.

A. Methods based on digital regression

For example,[2]proposed a privacy-preserving crowd monitoring system, which uses dynamic texture motion model to divide the crowd into components of uniform motion, then extracts a group of simple overall features from each segmented region, and uses Gaussian process regression to learn the corresponding relationship between the elements and the number of people in each segment. A supervised learning framework [3] is proposed. The counting problem is regarded as estimating the density of the image, and the number of objects in the area is given by integrating the image area. Point annotation is introduced into the training image. [4] A single multiple-output regression model is proposed, which can count the local population in space. By connecting multiple local image features as input, the population count for learning spatial localization is used as various outputs, and a single joint regression is used.[5] A semi-supervised learning framework is used to reduce the burden of data annotation by using the unlabeled data given by a large number of underlying population distribution structure patterns. These methods only capture some features in the original image and use machine learning and mathematical method learning model to map the relationship between elements and crowd numbers. However, researchers have to focus on finding more practical features to estimate more accurate results. At the same time, workers marked many annotations, which also increases the workload of data processing. Survivability is a fundamental characteristic of a large-scale network, and would not disappear due to system evolution or external environment changes. In the large-scale network, there exists information exchange between the failed subsystems and other subsystems, and thus may cause cascading failures. The increase of failed subsystems may cause the whole large-scale network failure.

B. Methods based on human detection

Methods based on human detection [6]: a human detector is trained through a group of pedestrian images (as shown in Fig. 1).

Fig. 1 Counting the number of people based on human detection

And the total detection count of pedestrians in the image. For example, using the traditional detector based on hog [7] or the detector based on deep learning (such as Yolo [8] or RCNN [9]), but the effect of these methods[10-11]is straightforward to be severely affected by occlusion in crowd scenes.

III. RESEARCH IN THE FIELD OF DEEP LEARNING

With the increase of population and severe occlusion, low resolution, pixels of each person and perspective, it is difficult to detect the head and face in the images. Researchers start to try other methods for crowd counting. With the development of LeNet-5[12]in 1998 and AlexNet [13], which won the championship in the 2012 visual recognition competition, the application of deep learning method to image problems has aroused researchers' interest. Deep learning models have the function of automatically learning the potential features in the training data. The deep learning model is applied to crowd counting for the first time. [14]proposed a deep convolution neural network for cross-scenes crowd counting task. At the same time, a data-driven method is proposed to select samples from training data and fine-tune CNN model before training, so that it can adapt to an unknown target scene.

A. Methods based on density map estimation

The crowd counting task develops a situation in which there are over 100 people in the scene to the case that there are more than 1000 people in the scene to count. The crowd counting task gradually develops into a mainstream method based on density map estimation (as shown in Fig. 2).

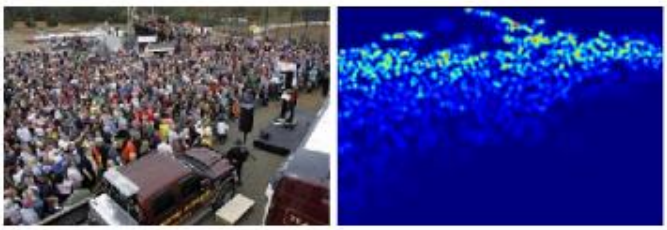

Fig. 2 Crowd counting based on a density map

Compared with the method based on detector and regression, density map can not only provide information about the number of pedestrians but also reflect the distribution of pedestrians. The primary way of density map is to use a Gaussian kernel to simulate a head in the corresponding position of the original image. After performing this operation on all heads in image, normalization operation is shown in the matrix composed of all these Gaussian kernels.

B. Focus on solving scale transformation

Counting people in the image with deep convolution network in 2015 was first proposed. The deep learning algorithm applied to crowd counting began to develop better in accuracy, fewer parameters and performance for large-scale population counting. The proposed algorithm needs to solve the problem of scale transformation caused by the increasing number of the crowd and the different distance between the pedestrian and the camera. [15] proposes a simple and effective multi-column convolutional neural network structure to map the image to the crowd density map. The features learned by CNN in each column can adapt to the size change of the human body/head caused by perspective or image resolution. Besides, once the model is trained on one dataset, it can be easily transferred to another new dataset.

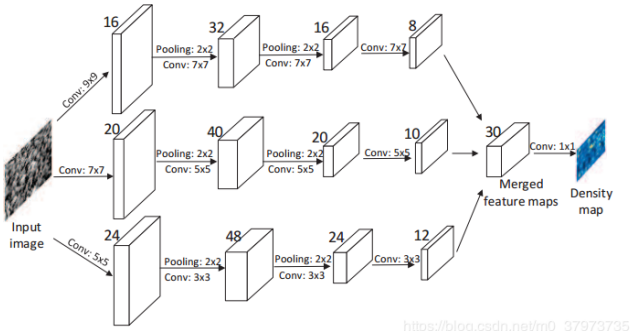

Fig. 3 Multi-column convolutional neural network model

It is a challenging problem due to perspective distortion caused by the change of scale and the distance between the crowd and the camera. Although the multi-column or multi-network CNN model is used to extract scale related features, it becomes more involved in optimization and calculation. Therefore, a new multi-scale convolutional neural network is proposed in [16] to count the crowd in a single image. The network can generate scale related features in a single column structure by using the multi-scale blobs based on the Inception structure, thus making higher counting performance, and the number of parameters is much less than the method in [15]. [17] proposes a recognition network based on CSRNet structure for high congestion scenes. This network is composed of a convolutional neural network as front-end and dilated CNN as a back-end, which extracts two-dimensional features. Using data-driven, it can not only understand congestion scenes, make an accurate estimation, but also generate high-quality density map. So far, the researchers hope that the designed algorithm can learn and adapt to the scale transformation in the image adaptively. [18] focus on density map refinement. A thinning framework is proposed to improve the performance of crowd counting by using thinning density map to train counters. At the same time, an adaptive density map generator is proposed, which directly uses the labelled dot map as input to generate a density map for training counter. This end-to-end framework trains density graph generator and counters jointly, without manually specifying density map an intermediate representation. [19] proposes a learning scale model (L2SM) to solve the problem of massive density changes in crowd count. Density concentration is achieved by using multipole centre loss. L2SM can effectively learn to extend multiple dense regions to similar density levels, which makes the density estimation of dense regions more robust. It has good generalization ability for unseen datasets with different density distribution.

C. Focus on other ways

The neural network model used for crowd counting is generally grave, and the accurate labelling of training data is also a labour-intensive work. At present, the method of crowd counting focus on other ways. [20] proved the possibility of the adaptive multi-scale context to coding, and provides an explicit perspective distortion effect model as the input into the deep network, by combining the use of multiple accept domain size characteristics, and learn the importance of each feature in each image position. The method is adaptive to code needed to accurately predict the crowd density of the range of context information, significantly improve the performance of the crowd count. In particular, it produces better density estimates in high-density areas. An unsupervised learning method for counting dense crowd is proposed [21]. Creating broad annotated crowd data is expensive and directly affects the performance of existing CNN-based small dataset counting models. Inspired by these challenges, the Grid winner-take-all (GWTA) auto-encoder was developed to learn several layers of useful filters from unlabeled crowd images. The basic idea is to limit the weight update of the neuron in the convolution output map to the neuron with the maximum activation in a fixed space cell. Almost 99.9% of the network parameters were trained into stacked WTA auto-encoders, using unmarked crowd images, while the rest were updated under supervision. Further analysis shows that this unsupervised method is more effective than fully supervised training when fewer marker data are available. Some researchers learned from Synthetic data for crowd counting in the wild [22]. The crowd counting method based on deep learning starts to seek a breakthrough in ways of training data, marking patterns and multi-view fusion.

IV. PERFORMANCE EVALUATION AND DATASETS

To evaluate the performance of deep learning algorithm for crowd counting, the commonly used evaluation indexes are mean absolute error (MAE), mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM).

MAE is defined as:

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\left|y_n - y_n\right| \tag{1}$$

MSE is defined as:

$$MSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - y_n)^2} \tag{2}$$

N represents the total number of test sets, Yn is the real counting value in the image, and Y'n is the predicted value. The lower the value of MAE is, the higher the accuracy of the algorithm is, while the lower the value of MSE is, the higher the robustness of the algorithm is. PSNR is often used to measure the quality of signal reconstruction in image compression, which is defined as:

$$PSNR = 10 \times \log_{10}(\frac{MAX_I^2}{MSE}) \tag{3}$$

MAX_I represents the maximum value of image point colour. If each sampling point is represented by 8 bits, the value is 255. SSIM is often used to represent the similarity index of two signals, ranging from - 1 to 1. When two signals are identical, the SSIM value is 1. For crowd counting task, MAE and MSE are commonly used to measure its performance.

Datasets for crowd counting in deep learning algorithm are as follows:

(1) UCSD: the dataset is collected from the camera on the sidewalk, including 2000 tagged pedestrian images in the video sequence. It contains 49885 pedestrian instances with a relatively low-density population.

(2)WorldExpo'10: a dataset for cross-scenes crowd counting. It contains 1132 annotated video sequences captured by 108 cameras. There are 3980 frames, of which 199923 are labelled as pedestrians. The crowd count varies from 1 to 220.

(3)Shanghai_Tech: including 1198 images and 330165 annotated headers. Part A of 482 high-density crowd images and part B of 716 different scene types and different density levels. There are different scale crowd and perspective distortion in the dataset.

(4)UCF_CC_50: a challenging crowd counting dataset. A total of 50 images with different resolutions. There are 1280 people in each image on average, and 63075 people are marked in total. Even the most advanced method based on CNN is far from optimal for the dataset.

(5)UCF_QNRF: including 1535 images, with 1251642 people marked, with an average of 815 people per image and a maximum of 12865 people. It is very suitable for training deep convolution neural network.

TABLE I. COMPARISON OF DATASET STATISTICS

Dataset	Number of images	Number of annotation	Average count	Max count	Average resolution	Average density
UCF_CC_50	50	63974	1279	4633	2101×2888	2.02×10^-4
WorldExpo'10	3980	22516	56	334	576×720	1.36´10^-4
ShanghaiTech PartA	482	241677	501	3139	589×868	9.33´10^-4
UCF_QNRF	1535	1251642	815	12865	2013×2902	1.12´10^-4

In the early stage of the medium-density crowd counting test, for example, the density of the USCD dataset is approximately 11-46 people per frame, the density of the mall dataset is 13-53 people per frame, and the USCD_CC_50 dataset has 50 images with an average of each Containing 1280 people, such a high density means that an individual may occupy too few pixels to detect it and verify its existence based on location, which is a challenge to existing technologies.

V. CONCLUSIONS

In recent years, the development of crowd counting method based on shallow learning has slowed down, and the most influential research results in this field are focused on deep learning. With the expansion of population scale, scale transformation, perspective distortion and other practical

problems, the deep learning crowd counting method based on computer vision is bound to be the focus of research, and its results are widely used in practical applications. The deep learning method applied to crowd counting is also constantly innovating, and we believe that crowd counting is going to becoming more and more apparent in the future.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation of China under Grant No.6167252 and Research and Innovation Team Project of Armed Police University Grant No.KYTD201807

REFERENCES

[1] R. Melina. How is crowd size estimated? InLife'sLittleMysteries.com, 2010.

[2] Chan, Antoni B., Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking.IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008:1-2.

[3] Lempitsky, Victor, and Andrew Zisserman. Learning to count objects in images.Advances in neural information processing systems. 2010:1-4.

[4] Chen, Ke, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting.In BMVC, vol. 1, no. 2, p. 3. 2012:1-10.

[5] Change Loy, Chen, Shaogang Gong, and Tao Xiang. From semi-supervised to transfer counting of crowds.In Proceedings of the IEEE International Conference on Computer Vision, pp. 2256-2263. 2013:1-8.

[6] Leibe B,Seemann E,Schiele B.Pedestrian detection in crowded scenes[C]//Procof 2005 IEEEComputer Society Conference on Computer Vision and Pattern Recognition (CVPR' 05) 2005:878−885.

[7] DalaiN,Triggs B.Histograms of oriented gradients for human detection[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2005:886−893.

[8] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi.You only look once: Unified, real-time object detection.In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779-788. 2016.

[9] Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik.Rich feature hierarchies for accurate object detection and semantic segmentation.In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587. 2014.

[10] Albiol A,Silla M J,Mossi J M.Video analysis using corner motion statistics[C]//Proc of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance,2009:31-38.

[11] Brostow G J,Cipolla R. Unsupervised Bayesian detection of independent motion in crowds[C] // Proc of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-(CVPR'06),2006:594-601.

[12] LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition.Proceedings of the IEEE 86, no. 11 (1998): 2278-2324.

[13] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks.In Advances in neural information processing systems, pp. 1097-1105. 2012.

[14] Zhang, Cong, 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, pp. 833-841. 2015.

[15] Zhang, Yingying, 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, pp. 589-597. 2016.

[16] Zeng, Lingke, Xiangmin Xu, Bolun Cai, Suo Qiu, and Tong Zhang. "Multi-scale convolutional neural networks for crowd counting." In 2017 IEEE International Conference on Image Processing (ICIP), pp. 465-469. IEEE, 2017.

[17] Li, Yuhong, 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*, pp. 1091-1100. 2018.

[18] Wan, Jia, and Antoni Chan. "Adaptive density map generation for crowd counting." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1130-1139. 2019.

[19] Xu, Chenfeng, Kai Qiu, Jianlong Fu, Song Bai, Yongchao Xu, and Xiang Bai. "Learn to Scale: Generating Multipolar Normalized Density Maps for Crowd Counting." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8382-8390.

[20] Liu, Weizhe, Mathieu Salzmann, and Pascal Fua.Context-Aware Crowd Counting.In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5099-5108. 2019.

[21] Sam, Deepak Babu, Neeraj N. Sajjan, Himanshu Maurya, and R. Venkatesh Babu. "Almost Unsupervised Learning for Dense Crowd Counting." In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, HI, USA*, vol. 27. 2019.

[22] Wang, Qi, Junyu Gao, Wei Lin, and Yuan Yuan. "Learning from synthetic data for crowd counting in the wild." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8198-8207. 2019.