Real Time Crowd Counting: A Review

Salwa Thasveen.M
Department of ECE
MES College of engineering
Kuttippuram, India
salwathasveen@gmail.com

Mredhula.L
Department of ECE
MES College of engineering
Kuttippuram, India
mredhulapradeep@mesce.ac.in

***Abstract*—Crowd counting is a process of counting number of people or objects in videos or images .This process has various applications related to our day to day life such as urban planning, health care, disaster management, public safety management, and defense. Thus new researches are going on in this field. The crowd techniques are broadly classified as supervised learning based and unsupervised learning based techniques. Some of traditional and convolutional neural network crowd counting techniques are discussed in this paper. Crowd counting techniques are facing various limitations such as occlusion, distortion in scale and perspective, non uniform distribution. As the crowd density increases, calculation complexity also increases. Most of the crowd counting methods involves density estimation. Density estimation gives an idea about the spatial distribution of people along with the count.**

Keywords—Crowd counting, Convolutional neural network, Density estimation

I. Introduction

World is witnessing exponential growth of population. Currently average increase in population is estimated as 82 million people per year and it is keep growing .Increase in the population will increases gatherings such as Sporting events, Political rallies, and music concerts[12]. In large gatherings of people, better management is essential for the safety and security. Better management can be realized through crowd analysis. Crowd analysis method includes crowd counting and density estimation.

In water festival of Cambodia about 347 people were killed on 22 November 2010. At the funeral of General Qasem Soleimani in Kerman, Iran, 56 were killed and 200 are injured. Many disasters are occurring around the world because of crowd turbulence. Crowd turbulence disaster may occur due to mass-panic, crowd crushes and loss of control. This kind of accidents can be prevented by early detection of the uncontrolled flow of crowd. Crowd behavior analysis is complex because it has both physiological and dynamic characteristics. Crowd scene is generic terms used represent objects with high density. So it is not only used to represent a group of people but also a school of fishes or a herd of sheep. Humans have the ability to extract useful information of behavior patterns in the surveillance area, monitor the scene for abnormal situations in real time, and provide the potential for immediate response [2]. But there are several limitations for the simultaneous observation of various signals in high density crowd. So technologies were introduced in the field of crowd analysis in order to overcome this limitation.

II. APPLICATION AND CALLENGES OF CROWD COUNTING

There are various applications in crowd counting which includes,

(a) For designing public spaces: Design should optimize the safe gatherings and movements of crowd
(b) Attendance Counting : It is helpful for counting attendance in educational institutions or work places
(c) Disaster Management: To prevent stampede in public events.
(d) Intelligence gathering system: Can be used in various sectors to reduce queue length

Challenges in Crowd counting:

(a) Occlusions
(b) High Clutter
(c) Non-Uniform illumination
(d) Scale and Perspective
(e) Non-Uniform distribution of people

III. CROWD COUNTING TECHNIQUES

There are various traditional and modern techniques for crowd counting. These techniques are broadly classified based on supervised learning and non supervised learning. The input data used is known and labeled in supervised learning whereas in unsupervised classification, the used data and labels are unknown.

Supervised classification consists of both traditional and CNN based techniques. The traditional techniques can be broadly classified into three. They are

- Detection based approaches
- Regression based approaches
- Density estimation based approaches

CNN based methods are broadly classified into three based on,

- Property of the network
- Training process
- Image view based

A. *Traditional Approaches*

Various traditional approaches have introduced to resolve crowd counting problems. The broad classification of traditional approaches into Detection based approaches, regression based approaches and density estimation based approaches is done by Loy et al. [8]

a) Detection-based approaches

This is one of the methods on which initial researches are focused.Detection style frame work based on a sliding window detector. The number of people was counts based on the information gained from this. Methods used in detection based approaches are,

- Monolithic style
- Part based detection style

In monolithic approach, a classifier is trained using features extracted from full body. Whereas in part based detection style the classifier is trained for detecting for specific body parts such as head and shoulders to estimate count of people. Min et al. [5] proposed a method for estimating number of people by MID based foreground segmentation and head and shoulder detection.

b)Regression based approaches

The monolithic and part based detection style method will not applicable for dense crowd. Because feature of the full body or parts like head and shoulder cannot be extracted as crowd density increases. To overcome this drawback regression based method is introduced. Regression model consists of two steps, Low level feature extraction and regression modeling. This method is a mapping between low level features extracted from image patches and count. Features that are extracting from image patches are

- Texture features
- Foreground features
- Edge features
- Gradient features

When the feature is extracted various regression techniques such as Gaussian process regression, linear regression, piecewise linear regression etc… can be used for mapping extracted features to crowd count.

c)Density estimation based approaches

Occlusion and clutter are major drawbacks of the regression based approaches. Density estimation can overcome this for an extent. Density based method incorporate spatial information. In density estimation method crowd count of a region can be estimated by integral over that region .

Chan et al.[1] proposed a regression method using Bayesian regression. Pedestrians moving in different directions are segmented based on homogeneous direction as shown in fig.1. Segment feature (Area, perimeter), Internal edge features (Edge length, Edge orientation), Texture features were extracted and mapped to crowd count using two regression based methods, Gaussian process regression and Bayesian Poisson regression. Gaussian process regression gives more accurate result at low density where as Bayesian Poisson regression gives better result at high density.

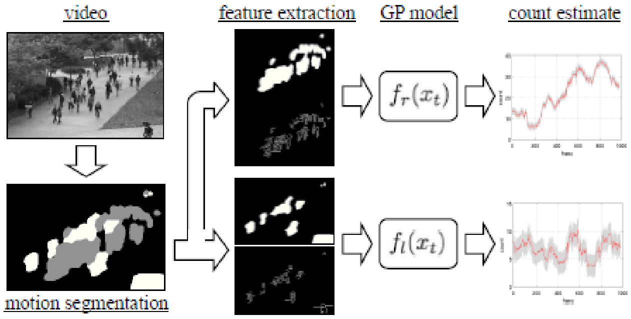

Fig. 1. Bayesian Regression method

Lempitsky et al. [4] proposed a flexible framework based on density estimation. Two types of images, Bacterial cells in fluorescence-light and pedestrians in surveillance video are examined. A density function F is recovered from the images and the count is calculated by integrating the density at region of interest.

B. *CNN based methods*

The CNN based approaches are broadly classified into three based on network property, training process and image view.

a) Property of network

Based on network property CNN based approaches can be categorizes as,

a. Basic CNNs: Involves only basic CNN layers.
b. Scale aware models: Architecture, such as multi-column and multi-resolution, provides robustness to scale.
c. Context aware model: CNN framework involves both global and local contextual information of an image
d. Multi task frameworks: Combining crowd velocity estimation and foreground-back ground subtraction along crowd counting and density estimation.

b) Training process

Based on training process CNN based approaches can be categorized as

a. Patch based training: Input images will convert into patches
b. Whole image based training: Input image use as such.

c) Image View

Based on the perspective of the image, we can classify image view based methods into two.

a. Arial view based : Object and camera are perpendicular to each other
b. Perspective view based: Object and camera are parallel to each other.

Generalized flow diagram [3] of CNN based crowd counting is given in the fig. 2. and the description is given below.

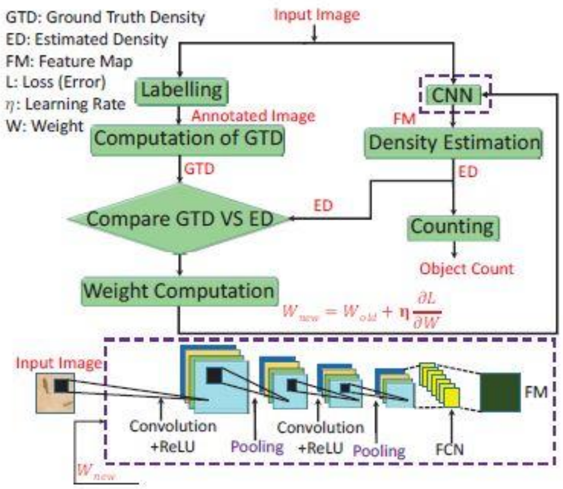

Fig. 2. Generalized flow diagram of CNN for crowd counting

Labeling: In crowd counting, labeling is done by dot annotation. It can be done various tools using LabelMe, Labelbox and RectLabel .

GTD Computation: Ground truth is the information obtained by direct observation. GTD computation can done using geometry-adaptive kernel and Gaussian Kernel.

GTD and ED Comparison: Comparison between GTD and ED provides loss between ground truth and estimated output.

Weight Computation: To minimize loss, update the network weight. W_{new}=$W_{ol\ d}$+ η$\frac{\partial L}{\partial W}$ computes the updated weight.

CNN: Extract image features and map the information into a density map.

Zhang et al. [14] presents a deep convolution neural network, multi-task, patch based method for crowd counting. This method even work when a new scene is introduced as input image which is not in dataset. This network is alternatively trained with crowd count and density estimation fig. 3.When a new scene is introduced at the input, the trained network fine-tunes the image to predict the crowd count. Thus the network is adapted to more scenes.

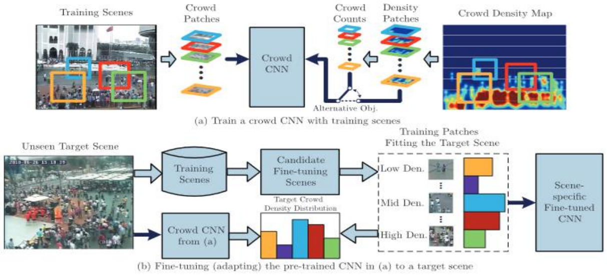
Training Scenes
Crowd Patches
Crowd CNN
Crowd Counts
Density Patches
Crowd Density Map
Alternative Obj.
(a) Train a crowd CNN with training scenes
Unseen Target Scene
Training Scenes
Candidate Fine-tuning Scenes
Training Patches Fitting the Target Scene
Low Den.
Mid Den.
High Den.
Crowd CNN from (a)
Target Crowd Density Distribution
Scene-specific Fine-tuned CNN
(b) Fine-tuning (adapting) the pre-trained CNN in (a) to a target scene

Fig. 3. Cross-scene crowd counting method

Zhang et al. [15] proposed a Multi-column Convolutional neural network (fig.4.) which is adapted to variations in size of heads due to perspective distortion. They collected 1198 images with 330,000 heads. The multi-column architecture has 3 columns with filter of receptive fields in three sizes, large, medium and small. Thus errors due to perspective distortion can be reduced.

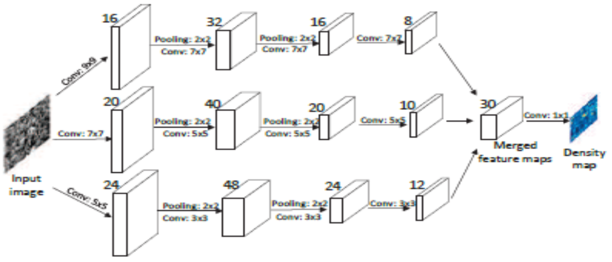

Fig. 4.Multi-column CNN method

Sindagi et al [10] proposed an end to end multi task Cascaded CNN architecture for crowd counting and density estimation. Obtain density map is highly refined and only subjected to lower count error. Proposed method has two parallel networks, high level prior stage and a density estimation stage. (Fig.5). These layers shares an initial set of convolutional layers. It consists of two layers each of which are followed by PReLU activation layer. Both parallel networks consist of 4 convolutional layers with PReLU activation layer after each.

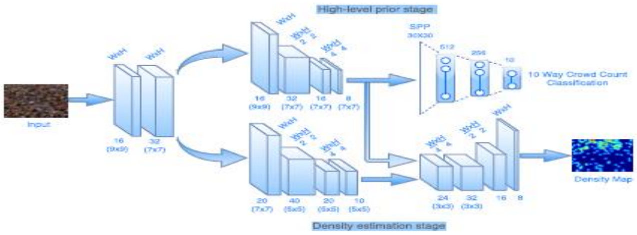

Fig. 5. Multi-task Cascaded CNN method

TABLE I. COMPARISON TABLE

Reference	Approach type	Method	Category		Dataset	MAE	MSE
			Network Property	Training process			
Chan et al.[1]	Traditional	Bayesian regression	-	-	UCSD	2.07	6.86
					Mall	3.43	17.07
Lempitsky et al. [4]	Traditional	Density estimation	-	-	UCF CC 50	493.4	487.1
Zhang et al.[14]	CNN based	Cross scene	Multi-task	Patch-based	UCSD	1.60	3.31
					UCF CC 50	467.0	498.5
Zhang et al. [15]	CNN based	Multi-column CNN	Scale aware	Whole image based	UCSD	1.07	1.35
					UCF CC 50	377.6	509.1
Sindagi et al. [10]	CNN based	Multi task Cascaded CNN	Multi-task	Whole image based	UCF CC 50	322.8	341.4
DecideNet [6]	CNN based	Multi-column CNN	-	Patch-based	Mall	1.52	1.90
					ShanghaiTech PartB	21.53	31.98
ADCrowdnet [7]	CNN based	Single column	-	Whole image-based	UCF CC 50	257.9	357.7
					UCSD	1.10	1.42
CP-CNN [11]	CNN based	Multi-column	-	Whole image-based	ShanghaiTech PartA	73.6	106.4
					ShanghaiTech PartB	20.1	30.1

DecideNet [6] is an end-to-end framework for crowd counting which make use of both detection and regression. Regression based method is mostly chosen in congested areas whereas detection based methods is more accurate in lessdense crowd scene. *DecideNet* separately generates detection based and regression based density maps. The proposed architecture is given the fig.6. The architecture consists of three blocks,*RegNet, DetNet, QualityNet*. *RegNet* and *DetNet* provide 2 types of density maps D_i^{reg} and D_i^{det} respectively .D_i^{reg} and D_i^{det} upsamples to the size of the input image I_i . D_i^{reg}, D_i^{det} and I_i are then stacked and given as input to *QualityNet.* Final density map D_i is obtained as the output of *QualityNet*.

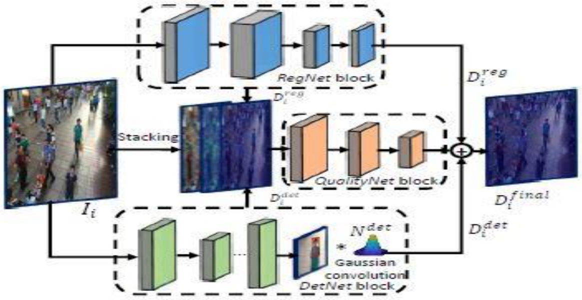

Fig. 6. Architecture of DecideNet

ADCrowdNet [7] consists of two concatenated networks. First one is an attention-aware network called Attention Map Generator (AMG). It detects crowd regions and computes the degree of congestion. Second is multi-scale deformable network called Density Map Estimator (DME). DME generates high quality density map. Architecture of the proposed method is shown is the fig.7. ADCrowdNet is much resistant to various noises and captures crowd features more effectively. AMG classifies an image into background image and crowd image. And obtain an attention map which has higher values for crowded region .It indicates degree of congestion. The pixel wise product of input image and attention map is used as input to DME network. To extract low level features front end uses 10 layers of trained VGG-16 model [9] . Back end uses a structure related to inception module [13] in which multi-scale deformable convolution layers that helps to overcome occlusion, distortion due to perspective view, variations in crowd distribution.

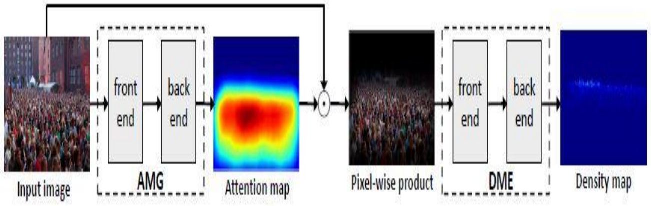

Fig. 7. Architecture of ADCrowdNet

CP-CNN [11] method explicitly incorporates both global and local contextual information of crowd images for crowd estimation and generate high quality crowd density map. The architecture of the proposed method is given in the fig. 8. The

CP-CNN is build up of four modules. They are, Global Context Estimator (GCE), Local Context Estimator (LCE),Density Map Estimator (DME) and a Fusion-CNN(FCNN). The network incorporates global context using GCE and local context using LCE. They are CNN based networks.GCE and LCE classify the input images into five categories. They are, extremely high density (ex-hi), high density (hi), medium density (med), low-density (ex-lo), extremely low-density (ex-lo). DME is multi-column CNN. It transforms input images to high dimensional feature maps. Finally, contextual information gets from GCE and LCE combines with high dimensional feature map obtained from DME. This produces a high quality and high resolution density maps

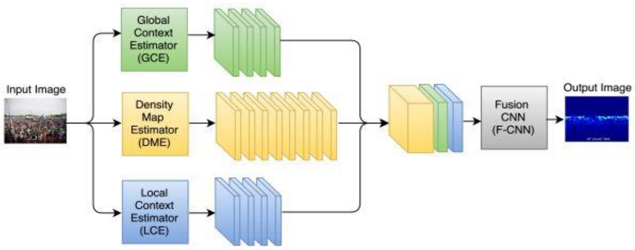
Input Image
Global Context Estimator (GCE)
Density Map Estimator (DME)
Local Context Estimator (LCE)
Fusion CNN (F-CNN)
Output Image

Fig .8. Architecture of CP-CNN

IV. EVALUATION METRIC

A comparison table of crowd counting techniques is given in the TABLE I. Different methods are evaluated using both mean absolute error (MAE) and mean square error (MSE) . MAE and MSE are given by equation,

$$\text{MEA} = \frac{1}{N} \times \sum_1^N |x-y| \qquad (1)$$

$$\text{MSE} = \sqrt{\frac{1}{N} \times \sum_1^N |x-y|^2} \qquad (2)$$

V. CONCLUSION

With the growing population, the necessity of the crowd analysis is increasing because of the more frequent crowd gatherings such as Political rallies, Sporting events, Music concerts. The crowd analysis includes crowd counting and density estimation. This can be done either by traditional approaches that uses hand crafted features or by CNN-based approaches. The CNN-based approaches are further classified based on the training process, network property and image view. The crowd count analysis is riddled with many challenges such as occlusion, non uniform density, variations of scales and perspective. The traditional approaches are limited to low density crowd. By estimating the error of various traditional and CNN-based approaches, CNN-based approaches are more adapted for handling large density crowd with variations in object scale and scene perspective.

References

[1] Chan, Antoni B and Vasconcelos, Nuno, “Counting people with low-level features and Bayesian regression,” *IEEE Trans. Image Process.,* Vol.21, No.4, pp.2160-2177, 2011.

[2] Hospedales, Timothy and Gong, Shaogang and Xiang, Tao,” Video behaviour mining using a dynamic topic model,” *Int. J., of Comput. Vision.,* Vol.98,No.3,pp.303-323,2012.

[3] Ilyas, Naveed and Shahzad, Ahsan and Kim, Kiseon,” Convolutional-neural network-based image crowd counting: review, categorization, analysis, and performance evaluation,” *Sensors,* Vol.20,No.1,pp.43,2020.

[4] Lempitsky, Victor and Zisserman, Andrew,”Learning to count objects in images,”in *Advances in neural information processing systems*,2010,pp.1324-1332.

[5] Li, Min and Zhang, Zhaoxiang and Huang, Kaiqi and Tan, Tieniu,” Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection,”in*2008 19th international conference on pattern recognition*,2008,pp.1-4.

[6] Liu, Jiang and Gao, Chenqiang and Meng, Deyu and Hauptmann, Alexander G,” Decidenet: Counting varying density crowds through attention guided detection and density estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,*2018,pp.5197-5206.

[7] Liu, Ning and Long, Yongchao and Zou, Changqing and Niu, Qun and Pan, Li and Wu, Hefeng,” Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,*2019,3225-3234.

[8] Loy, Chen Change and Chen, Ke and Gong, Shaogang and Xiang, Tao,” Crowd counting and profiling: Methodology and evaluation,” in *Modeling, simulation and visual analysis of crowds*,2013,pp.347-382.

[9] Simonyan, Karen and Zisserman, Andrew,” Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*,2014.

[10] Sindagi, Vishwanath A and Patel, Vishal M,” Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting,” in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS),*2017,pp.1-6,

[11] Sindagi, Vishwanath A and Patel, Vishal M,” Generating high-quality crowd density maps using contextual pyramid cnns,” in *Proceedings of the IEEE International Conference on Computer Vision*,2017,pp.1861-1870.

[12] Sindagi, Vishwanath A and Patel, Vishal M,” A survey of recent advances in cnn-based single image crowd counting and density estimation,” *Pattern Recognition Letters*,Vol.107,2018,pp.3-16.

[13] Szegedy, Christian and Liu, Wei and Jia, Yangqing and Sermanet, Pierre and Reed, Scott and Anguelov, Dragomir and Erhan, Dumitru and Vanhoucke, Vincent and Rabinovich, Andrew,” Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*,2015,pp.1-9.

[14] Zhang, Cong and Li, Hongsheng and Wang, Xiaogang and Yang, Xiaokang,” Cross-scene crowd counting via deep convolutional neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition,*2015,pp.833-841.

[15] Zhang, Yingying and Zhou, Desen and Chen, Siqin and Gao, Shenghua and Ma, Yi,” Single-image crowd counting via multi-column convolutional neural network,” *in Proceedings of the IEEE conference on computer vision and pattern recognition*,2016,pp.589-597.