SURVEY OF CROWD COUNT &DENSITY ESTIMATION METHODS: CONTEXTUAL, MULTI TASK/STAGE AND COLUMN-BASED CNN

Pankaj Sharma,

Student, Department of Computer Science and Engineering,

Manipal University Jaipur,

Jaipur, INDIA.

Mail Id: pankajsharma0071uk@gmail.com

Harish Sharma,

Department of Computer Science and Engineering,

Manipal University Jaipur,

Jaipur, INDIA.

Mail Id: harish.sharma@jaipur.manipal.edu

Sunita Singhal,

Department of Computer Science and Engineering,

Manipal University Jaipur,

Jaipur, INDIA.

Mail Id: sunita.singhal@jaipur.manipal.edu

***Abstract*:** Crowd Analysis is an interesting topic in Digital Image Processing field. Crowd analysis is based on video imagery. In crowd analysis usually crowd density and crowd behaviour is estimated, crowd tracking, anomaly detection in crowds etc. Since the security of people is main concern in heavily crowd places like concerts, malls etc. Researchers are continuously working on this for the better results. Several approaches are followed earlier in this domain such as crowd detection based, and crowd clustering based. Presently, regression-based approach is being followed. Regression based approach is quite good for crowd analysis of dense crowd. In this theory we have provided our point of view in evolution of contextual, column and multi task-based CNN crowd estimation models. We will discuss pros and cons in all our surveyed algorithms.

Fig 1. Dense crowd difficult to count

Keywords: Convolution Neural Network, crowd counting, Deep Convolution Neural Network

1. Introduction

Crowd Estimation or counting have been in researchers to do list with crowd behaviour analysis[2, 3, 4, 5], tracking and monitoring[1] for anomaly detection [6, 7]. Occlusions, non-uniform distribution, uneven illumination, scale and perspective are the problems that still challenges researchers to tackle in the field of Digital Image Processing. But these problems also boost researchers to make a model which is more efficient than previous models.

The number of pedestrians/person/individuals per unit area is called crowd density. Crowd density estimation targets to map an image of crowd to its corresponding crowd density map which shows the number of persons per pixel. This helps in preventing the crowd disasters such as stampede and life of people.

Crowd density five different levels (people/m^2) as:

Level	People/m^2
Jammed flow	X >2.0 (very high)
Very dense flow	2.0>X>1.27 (high)
Dense flow	1.26>X>0.81 (moderate)
Restricted flow	0.8>X>0.5 (low)
Free flow	0.5>X (very low)

Table 1: Different levels for crowd density.

Above dense flow it is very difficult to determine the desired object. This is the challenge that researchers are being put through.

People Safety: Monitoring people through surveillance is prime concern for preventing any threat. Applications used military equipment, MNC assets, international airports, mall and marts. Older ways in nowadays are not effective no new robust models are needed to handle all challenges on the table.

***Disaster management*:** In many cases of crowd accumulation such as religious events, sports events, music concert events, public and political rallies faces the life-threatening risks due to stampede occurred in crowd.

***Information* :** Crowd analysis, satellite imagery helps in weather information and surveillance. Determining Disease from images is also in practice. Planning any new infrastructure, building, equipment in some examples of image processing for getting precise information.

***Virtual environments*:** Crowd analysis methods are also used in virtual reality application that can provide accurate simulations.

Many surveys of crowd counting, and analysis have been produced but they are carried traditional methods. At present, CNN-based approaches are influencing the crowd analysis due to its capacity to handle large crowd datasets. CNN-based models are having lower error rates. As we all know more the information/data we have then more knowledge we have which results in precise results. Multi task and stage-based CNN models have been in top tier nowadays in image processing approaches. We are going to discuss these state of art models here on.

Crowd Estimation Approaches

2.1 Detection-based approaches

Detecting the object was our prime stage in this evolution of image processing, where earlier we do this with specimen template. This detection can be performed either on parts-based or monolithic detectors. Traditional detector for crowd count are haar [8, 10, 11] and Histogram Oriented gradients [9] with Support Vector Machine, random forest[13], boosting[12] etc have been shown better results. The major drawbacks of these methods are having small or low-density crowd images. Only the highly-dense crowd image is enough for shaking these techniques. In detection patch-based learning for classification of desired object from the image is used to count that. This approach came under supervised learning.

2.2 Regression-based approaches

Recently, ages of regression-based approaches are body parts-based. Its was still difficult for them to detect people when it is for very dense crowd from the images which also have other defects like perspective, occlusion etc. Regression based learning models were started using patch based technique for count person patch by patch by comparing the feature with patch. Background-foreground segmentation also helped in getting results. Then area, perimeter and ration feature extractors also came into picture which have shown good results.

Model trained is solely dependent on the corresponding mapping they are done but if model is used for new scene then the result became degraded.

2.3 Neural Network-based approaches

Neural Networks have also huge help from past decade in more efficiently than others. Due to neural network we can easily achieve the better result as its neural network natural tendency to discover insights which we are unable to see or formulate, which allow digital image to be better represented as not in the above approaches. The trained model can be easily transferred to different perspective scene or crowd scale and will give the better result from that scene also. Just last layers tuning will provide more accurate result.

2.4 ***Density based approaches:*** In this approach we use continuous density maps instead of counting the pedestrians in the crowd. This approach is quite in handy as it estimates the crowd by mapping the input image and maps it to draw density map and we also use this in CNN crowd counting models also. It presents a data driven model to count crowd in unseen scenarios.

2. CNN-based methods

Earlier CNN based methods are used to predict the number of objects instead of crowd density map. Zang et al is used one approach in which he

introduced the CNN with ability to address the complexity of background and density map with the integrated crowd count denoted by the sum of pixel values. After that CNN role in Digital Image Processing has increased. ConvNet (CNN) mostly used to recognise patterns in an image. When these patterns are further processed in the ConvNet model then the model starts recognizing more complex features and eventually accuracy also improves. In paper [16] author is categorized these methods which are based on property of the networks and training approach as shown in Fig. 2.

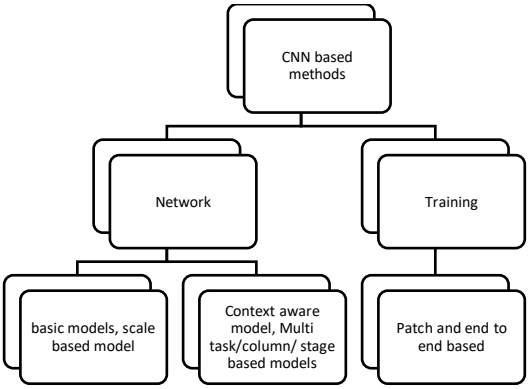

Fig. 2 Categorization of existing CNN-based approaches[16].

Based on the property of the networks, they have classified the approaches into the following categories:

- Basic models: In this approach the ConvNet model involve basic CNN layers. Such models were amongst the earlier ones of CNN used for crowd counting and density estimation.
- Scale-aware models: In this approach of CNN model, the variations in scale of the image is considered like in face detection, in which problem there is face scale for face of each individual is different and to optimize this problem we have CNN models such as multi-column or multi-resolution architectures.
- Context-aware models: By accumulating the local and global information present in an image with a CNN model just to achieve lower error rate.
- Multi-task models: Multi Task Models in Machine learning are used in various fields of data science such as speech recognition, Natural Language Processing (NLP) and Computer vision also. Actually, these models are here to provide generalisation in any field. Example as combining crowd monitoring and crowd counting and these tasks are performing in only one model simultaneously.

Another categorization of CNN approach are on the basis of Training process, here are two categories:

- Patch-based inference: In this approach, the image is divided into equal patches and then each patch is computed in the model and then further send to the pooling process or anything else. This operation is then accumulating the result from that convolution layer and then feedforward it to next layer for further operation. Here patch sizes can be different in any convolution layer. This model is computationally complex though.
- End to end model: In this model whole image is taken as input and operation in model happen simultaneously on whole image. Here the number of computational steps is reduced as we are not taking patches of image which increases the complexity of model.

Apart from all mentioned uses of CNN can also be used in Denoising the data and reducing the dimensionality of the data.

3. Contextual-based and multi stage-based CNN methods

Here, we provide our study on the basis of two cases, first one is contextual based, and another is multi stage based on CNN-based crowd density estimation methods. Zhang et al. [17] is found that in existed methods, performance decreases rapidly when applied to a new scene that is not similar from the training dataset. To overcome this problem, they proposed a method which is adaptive to newer image data and learn feature maps from images for crowd counting. To get this goal they have worked on both crowd count and crowd estimation methodologies.

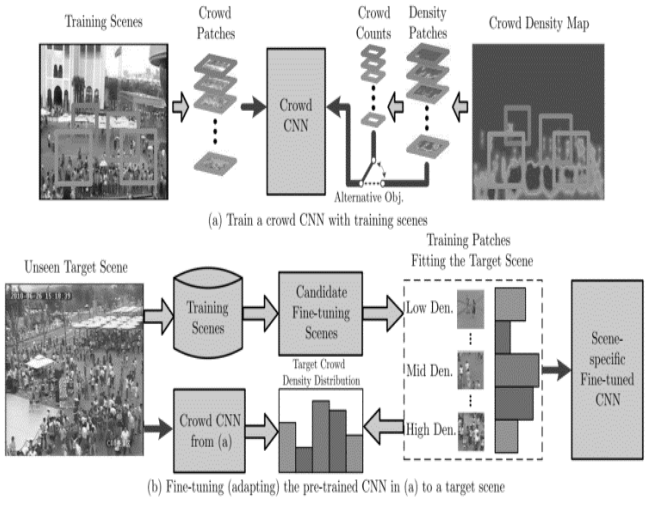

Fig. 3 Cross scene crowd counting proposed by Zhang et al [17]

Attaining the improvement in these two ways of crowd estimation, we are capable to get more accurate results. Adaption nature can only be achieved when there is some similarity between previous target and new samples. As they are adaptive it is not difficult to adapt to new targets. Fig. 3 shows the model diagram. this adaptability brings robustness in the model which means to be able to handle any scale data of different perspective. They had also provided their dataset. They are able to employ on both cross-scene crowd counting and single scene crowd counting with good accuracy.

But Arteta et al. [18] is used a dataset of penguins which is scale variant and have high degree of occlusion. The used dataset is dot annotated as well. Using their multi-task learning algorithm, they have implemented foreground-background segmentation and estimated the count explicitly. Since it is multi-task based so it able to do all that multi classification instead of that single task classification. Since it is contextual based classification then this model also able to train additional regions around the target. A whole contextual patch are trained here for getting good accuracy.

Zhao et al. [19] also proposed a dataset which is surveillance videos trajectorial dataset with maximum number of annotation (5900) at that time. They have only 5 scenes to count the line of Interest pedestrians. It is video based CNN model which is trained pixel-level same as of the single image crowd estimation models. Here they are performing two task-

- counting the people
- monitoring the trajectory(speed)

Both tasks are simultaneously work till the first layer of the CNN model they proposed.

Whereas, in 2017 Sindagi et al. [20] proposed a end to end cascaded CNN model for very dense crowd scenes. They tackled- the scale variant dataset difficulty very well. This model able to keep lower the count error and better density maps with previously proposed models. They did high level density estimation, which results in giving the better performance. their high-level prior stage of CNN provides great insight of the dense crowd image. They are able to learn global features. They are able to generate high resolution feature maps. Fig. 4 represents diagram of their multi task CNN model.

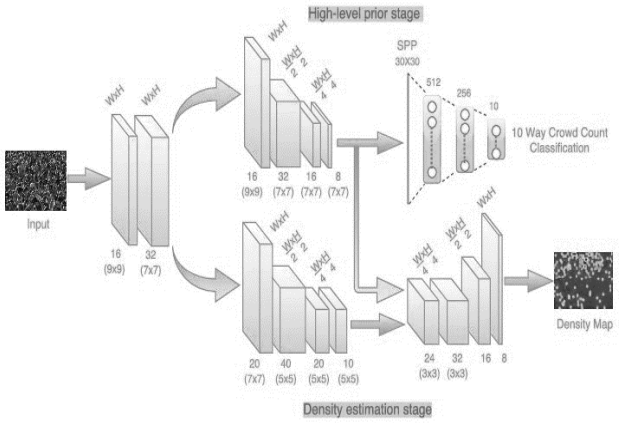

Fig. 4 Overview of Cascaded Multi-task CNN by Sindagi et al [20]

Different from previous models which are using the patch-based level training, Shang et al. [21] proposed a model(Fig. 5) which is an end-to-end convolution neural network(CNN) which takes full image as input and produces the count number. It is contextual based classification of input image. Entire image is fed to the model which is different from patch-based processing. when there is any overlapping region, the model's sharing computation make it easier for getting the local and global features which after processing gives high accuracy for crowd count. This is able to deal with the complexity easily. It is having a faster training process. the contextual insight they get will allow to neglect noise in the image to get high results. The dataset is pretrained using GoogleNet which gives CNN feature maps and then the local count is given by LSTM(Long-short time memory) and CNN layers for final count.

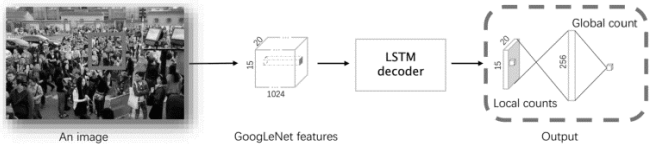

Fig. 5 End-to-end counting method proposed by Shang et al. [25].

Previous models were scale adaptive, while sheng et al. [22] proposed a model which is able to estimate density maps on an image pixel wise. These estimated density maps are obtained on different cases which are local scale attention(gives

three pixel-wise local maps), multi-scale feature extractor and Global scale attention(gives three global maps). the density maps obtained from all these three paradigms will be summed up by fusion network which in result gives the final crowd count. Hence this way are able to predict density map. Multi-scale feature extractor able to get feature maps on three different scales. So, this model is also scale aware model. Due to these three different feature map categories, this model outperforms many models.

4. Discussion

Among all methodologies based on CNN methods Zhang et al. [17] proposed a method which is able to deal problems unlabelled datasets using patch methodology across datasets and able to introduce adaptive models. Due to using unlabelled datasets it is hard to get good feature maps, but fact was that the dataset was easy to produce accurate feature maps due to less crowd density. They have used 72x72 size of patches for training.

In Marsden et al. [23], they are proposed a model which able to deal multi-scalability for predicting the count. But it is not that effective due to its inference stage which is weak to give results. Such models are not able get global feature map which can enhance the count accuracy, but it wasn't in their model so not that robust to predict accurate count. But lately few models [21][22] came which are able to exploit semantic and spatial intelligence present in image dataset which in return gives good local and global feature maps. In some models they have used multi task [20] capability for learning and giving high accurate estimation of count with engaging high-level priors in those. Which in return gives very accurate crowd count. The Dataset also plays an important role in count accuracy. If the image in too blur and have lot of noise, then it is hard to get desired output.

Table 2, is the comparison table of few algorithms we have studied.

5. Conclusion

In this survey, all crowd counting models are not considered, just few state-of-art models in their time are picked. These are able to predict crowd count effectively. CNN-based (contextual-based & multi task-based) approaches to conclude that CNN-based methods are more adept at handling large density crowds with variations in object scales and scene perspective. Additionally, the CNN-based methods drastically improve the estimation error. Apparently, we have discussed and find merits and drawbacks of these models, which in return gives us good view for upcoming challenges that are faced during the prediction of crowd count.

Algorithm	Authors	Dataset Used	Pros	Cons
Cross-scene crowd counting via deep convolutional neural networks	Zhang et al. [17]	WorldExpo'10 crowd counting dataset, the UCSD pedestrian dataset and the UCF CC 50 dataset	Still have errors value greater on current state of art crowd counting model.	Performance decreases rapidly when applied to a new scene that is not similar from the training dataset
Counting in the wild	Arteta et al. [18]	Penguin dataset	Implemented foreground and background segmentation and explicitly count is taken	Model was counting less than the actual count with this much dot annotated dataset
Crossing-line crowd counting with two-phase deep neural networks	Zhao et al. [19]	Surveillance videos trajectorial dataset	Performing two task-counting the people, monitoring the trajectory(speed)	Dataset have good set of imagery so if dataset is too crowded and frames have occlusion it won't performs well

CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting	Sindagi et al. [20]	Shanghai dataset and UFC_CC_50 dataset	High level prior stage and can-do count local and global features simultaneously.	Computation cost may be high, take more time due to slow learning.
End-to-end crowd counting via joint learning local and global count	Shang et al. [21]	UFC_CC and worldexpo'10 crowd counting dataset	End-to-end patch based, faster training, scale adaptive	Still have not able to get deep features for dense crowd.
Crowd counting via weighted vlad on dense attribute feature maps	Sheng et al. [22]	Mall dataset, UCSD dataset and Caltech 10X dataset	Pixel based learning and also able to get local and global count on three different feature maps	If W-VLAD was used instead of VLAD it would have given better results

Table 2: Comparison of algorithms

References

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

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

[3] Zhou, B., Wang, X., Tang, X., 2012. Understanding collective crowd behaviours: Learning a mixture model of dynamic pedestrian-agents, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE. pp. 2871–2878.

[4]Saxena, S., Br´emond, F., Thonnat, M., Ma, R., 2008. Crowd behavior recognition for video surveillance, in: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer. pp. 970–981.

[5]Zhou, B., Tang, X., Wang, X., 2015. Learning collective crowd behaviours with dynamic pedestrian-agents. International Journal of Computer Vision 111, 50–68.

[6]Li, W., Mahadevan, V., Vasconcelos, N., 2014. Anomaly detection and localization in crowded scenes. IEEE transactions on pattern analysis and machine intelligence 36, 18–32.

[7]Benabbas,Y., Ihaddadene,N.,Djeraba,C.,2010. Motion pattern extraction and event detection for automatic visual surveillance. EURASIP Journal on Image and Video Processing 2011, 163682.

[8]Viola, P., Jones, M.J., 2004. Robust real-time face detection. International journal of computer vision 57, 137–154.

[9] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE. pp. 886–893.

[10] Wu, B., Nevatia, R., 2005. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, IEEE. pp. 90–97.

[11]Sabzmeydani, P., Mori, G., 2007. Detecting pedestrians by learning shapelet features, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE. pp. 1–8.

[12] Viola, P., Jones, M.J., Snow, D., 2005. Detecting pedestrians using patterns of motion and appearance. International Journal of Computer Vision 63, 153–161.

[13] Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V., 2011. Hough forests for object detection, tracking, and action recognition. IEEE transactions on pattern analysis and machine intelligence 33, 2188–2202.

[14] Zhao, T., Nevatia, R., Wu, B., 2008. Segmentation and tracking of multiple humans in crowded environments. IEEE transactions on pattern analysis and machine intelligence 30, 1198–1211.

[15] Ge, W., Collins, R.T.,2009. Marked point processes for crowd counting, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 2913–2920.

[16] Vishwanath A. Sindagia, Vishal M. Patel, 2018 A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation, in: Elsevier Ltd website.

[17] Zhang, C., Li, H., Wang, X., Yang, X., 2015. Cross-scene crowd counting via deep convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 833–841.

[18] Arteta, C.,Lempitsky,V.,Zisserman,A.,2016. Counting in the wild, in: European Conference on Computer Vision, Springer. pp. 483–498.

[19] Zhao, Z., Li, H., Zhao, R., Wang, X., 2016. Crossing-line crowd counting with two-phase deep neural networks, in: European Conference on Computer Vision, Springer. pp. 712–726.

[20] Sindagi, V., Patel, V., 2017. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, in: Advanced Video and Signal Based Surveillance (AVSS), 2017 IEEE International Conference on, IEEE.

[21] Shang, C., Ai, H., Bai, B., 2016. End-to-end crowd counting via joint learning local and global count, in: Image Processing(ICIP),2016IEEE International Conference on, IEEE. pp. 1215–1219.

[22] Sheng, B., Shen, C., Lin, G., Li, J., Yang, W., Sun, C., 2016. Crowd counting via weighted vlad on dense attribute feature maps. IEEE Transactions on Circuits and Systems for Video Technology.

[23] Marsden, M., McGuiness, K., Little, S., O'Connor, N.E., 2016. Fully convolutional crowd counting on highly congested scenes. arXiv preprint arXiv:1612.00220.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Google Inc, University of North Carolina, Chapel Hill, University of Michigan, Ann Arbor and Magic Leap Inc., Going Deeper with Convolutions, CVPR 2015