Development of Crowd Counting And Density Estimation Model Using CNN

Alubankudi Olaoluwaseni T .
Department of Electrical and Electronics Engineering
Federal University of Technology
Akure , Nigeria
amopsy4real@gmail.com

Ogunti Erastus O.
Department of Computer Engineering
Federal University of Technology,
Akure, Nigeria
eooguntі@futa.edu.ng, oguntig@gmail.com

***Abstract*—-Crowd counting and density estimation is very essential and challenging in the current time. Different tactics have been used to tackle the trouble of crowd counting. In this research, we recommend to enhance a crowd counting model on the use of a convolutional neural system with the elimination of its definitely linked layers and evaluating the model performance. We operate on a tasking crowd counting dataset and few photos introduced to the dataset which achieves the revolutionary modern-day effects and show the efficacy of our new technique**

Keywords—component; Dataset; Crowd; counting; Density Estimation

Introduction

The estimation of numbers of human being in an image is known as crowd counting. Crowd estimation has been an energetic lookup vicinity on account that the creation of computer vision, and the improvement of a range of kingdom of the art methods has progress from being an not possible venture to one of pastime from researchers. People collect together in a specific place for distinct functions such as political rallies, pilgrimages, church crusades, marketplaces, schools, and sports events. These gatherings by and large are overcrowded which may also motive site visitors' delays, stampedes and accidents. Therefore counting the quantity of humans in crowded vicinity is very important. The direct approach is counting every character in the crowd, which is slow and unreliable. Crowd counting is a major challenge that types a simple building block for a number real-world purposes such as density estimation which maps a crowd picture fed into a density estimation algorithm to determine the range of people per pixel in the image, crowd conduct analysis, reconnaissance in current warfare, anomaly detection and congestion analysis.

Fig 1a crowd image

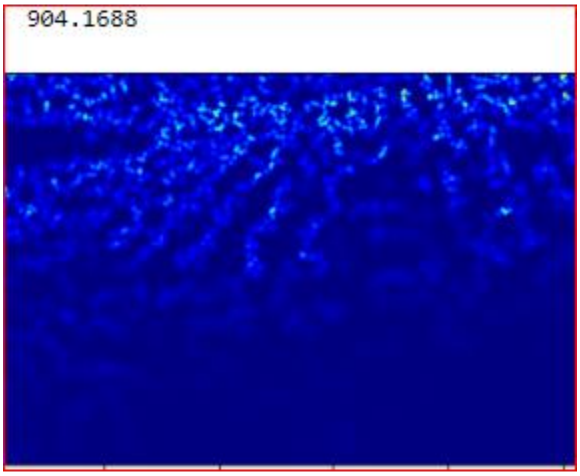


Fig1b: Density estimation

I. Related works

Computer imaginative have solved using wide variety of procedures in tackling crowd counting.[1] grouped the counting method into detection, regression and density estimated approaches. For the detection perspective, counting of peoples' head or shoulder in the crowd is done using a detector. [2] utilize haar element for head like delineations and applied SVM categorizer to group these aspects as the contour of head or not.[3]also presented that the input picture used to be paramount segmented into foreground – background sections and a HOG characteristics based head-shoulder feeler was once used to realized every individual in the pack. but places where it is overcrowded the method could not give better results due to heavy occlusion so regression method was adopted in which low features like textures edges[4] and gradient features[5]are extracted to pair in a regression function by matching the features got into counts[6].For the density estimated method, the number of humans in an image is represented by a density map in which [7]proposed how to match the density map object and the image crowd linearly.[8] categorize a picture into five major crowd concentration classes and utilized a flow of combined CNNs in improving the approach everywhere the additional CNN used to be trained on the photographs mis -categorized by way of the first CNN. These techniques likewise left out the benefits supplied by means of the multitude concentration mark maps

A. CNN based approach

CNNs achievement in numerous computer visualization obligations has motivated scholars to continuously use their capabilities in getting to know nonlinear features in crowd pictures respect to its corresponding density maps or counts. A range of CNN-based methods can be classified into these strategies based totally on networks property and training approach [9]

i. Network property: This grouping is achieved based on the properties possessed by the ConvNet. These are Basics, context-aware models, scale aware models and multi-task models.

ii. Training process: Classification the usage of the method used in training data. These are Patch-based coaching and end-to-end training.

[8]employs CNN for crowd density estimation, the place they removed specific network links according to the commentary of the existence of related feature maps, which notably accelerated the estimation procedure. Secondly, they proposed a two ConvNet classifier cascade, which expanded on accuracy and velocity of today's works. The algorithm was once tested on three datasets (PETS_2009, a subway photo sequence and a ground fact picture sequence). The proposed approach uses education boosting technique where samples are categorized to challenging and normal samples. [10]instigated an end-to-end CNN structure which accept input as a total photo and gives an output of the counting effects directly. A pre-trained CNN was used and inputs the photograph into it to acquire a set of complex structures, these elements are then paired to counting numbers, the usage of recurrent community layers with reminiscence cells. A contemporary result was once achieved, which proved the efficacy of this method.[9] presented some other end-to-end method, but this time, cascaded. This approach examine both crowd depend classification and density map estimation. Network layers learn distinct international facets which aids in gathering fantastically refined density estimation and limit mistakes related with crowd count. [11]developed a convolutional neural network named CrowdNet in which a deep gaining knowledge of framework for estimating crowd density in nevertheless photographs of giant crowds are reflected. An addition of deep and shallow absolutely convolutional network efficaciously captures each excessive degree and low-level features which was tested using the UCF_CC_50 dataset and it performs very well when compared to models of different state of the artwork.

III METHOD

Shanghai Tech Dataset is used for this research study. The first step is the pre processing stage, which secure all usable pictures in the dataset are ready for forming the model. Following is main model stage, in which the model is developed, lastly is the inference stage where the mean absolute error and mean squared error for the model established.

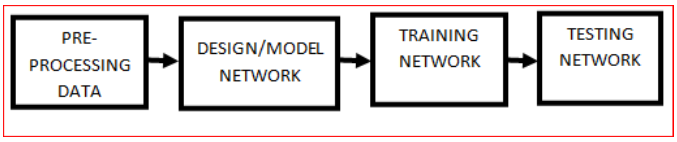


Fig 2: The network model flow chart

Data Pre-Processing

Collecting Data is a great manner of any lookup study. Imprecise or inadequate facts can have an influence on the effects on the results about and sooner or later lead to invalid or skewed outcomes [12]. In this research, the shanghai dataset used has two phase in the training and testing, part A for dense crowd and phase B for sparse crowd. Both section of the dataset had been divided into coaching and testing in the same ratio below. The dataset was once subdivided into two parts:

i. Training Data 70% for training

ii. Testing Data 30% for testing

Density Map Generation

The ground fact (ground truth) was created by using really blurring every head annotation,the usage of aGaussian kernel normalized to summing to one. Following the method of generating density maps in [13], in overfilled sections kernels which are geometrically adaptive are utilized to confront the incredibly crowd. Ground fact used to be generated through thinking about the spatial circulation of the snap shots in every dataset. The geometry-adaptive kernel is described in equation 2 as:

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

Where $\sigma_i = \beta \bar{d}_i$

For each marked object x_i in the ground truth, we use d_i to designate the average space of k adjoining neighbors. To produce the map density, δ(x-x_i) was convolved with Gaussian kernel and standard deviation σ_i where x is the location of pixel in the image. The arrangement utilize by [14] was adopted where β known to be 0.3 and k which is 3 was adopted. In areas of sparse crowd input the Gaussian kernel was adapted to the average head dimensions to blur all the annotations

Density Map Counting

Density map is responsible for the non linear dispersal of objects in an image, relative to the total quantity of objects. Given the density map, D_ID_I the total count N_I is got by integrating pixel values in density map D_I over the entire image

$$N_I = \sum_{\rho\epsilon I} D_I(p)$$

B Data Processing and Cleaning

In this step, ground truth had been transformed into density maps and ground fact were additionally generated for datasets except ground truths. The ground reality is a sparse matrix consisting of head annotations in a particular image. All snap shots in the dataset are anticipated to have a corresponding ground truth. A KD tree affords area partitioning facts structure for organizing factors in a K-dimensional space. The KD tree is used to compute the K-nearest neighbor of the sparse matrix. Average cost got from the distances between head annotations used to be elevated by means of 0.3 as advised by using [15]. The value acquired was equated to sigma, and then the sparse matrix was once exceeded through a Gaussian filter to form a density map. Adding collectively all the cells in the density map gives total quantity of human beings in the image.

Model Development

The model was developed from VGG-16 model and our dilated convolution approach. The important idea is to set up a deep Neural Network for extracting high stage points to produce striking density maps devoid of significantly growing the complexity of the neural network.[16] ,[17] , [18], VGG-16 was first selected as the front-end of the model because it has the capacity to be used for switch learning-a approach of the usage of weights of pre-trained models to create new models, considerably reducing the training time and stops us from re-inventing the wheel. The VGG-16 also has an easy-to-use structure that enabled us concatenate the back-end of the system for density map generation.

In CrowdNet [16], 13 layers were without delay took out from the VGG-16 and a more 1 x 1 convolutional layer used to be brought to serve as the output layer. The absenteeism of alterations resulted in a feeble performance. Other structural design such as VGG-16 used by[14] has the density stage classifier for labeling entry photos earlier transferring them to the utmost suitable column of the MCNN, while the CP-CNN through[17] accommodates the result of organization with the points from density map generator. This study, fully-connected layers of the VGG-16 was once removed and the model is built on convolutional layers in the VGG-16. VGG-16 community (except fully-connected layers) used to be adapted and only use 3x3 kernels. According to [18], using small filtersof larger convolutional layers is greater environment friendly than the use of big kernels of small numbers of layers when focused on identical size of field receptive. With wide variety of layers wished to use from VGG-16 used is chosen, fully connected layer is disposed.

Model Training

The major 10 convolutional layers are modified from a fine-proficient VGG-16. For the other strata, the initial 6 value is derived from a Gaussian initialization with a 0.01 preferred deviation. Stochastic gradient descent (SGD) is utilized stably getting to know charge at 0.000001 in the course of training. Also, Euclidean distance was chosen to quantify the difference between the ground fact and the generated density map estimation which is comparable to the works of [14], [19] and [16]. Below is given as loss function

$L(\theta) = \frac{1}{2N}\sum_{i-1}^{N} ||Z(X_i;\theta) - Z_i^{GT}||_2^2$ Where N is the size of training batch and $Z(X_i;\theta)$ is the output by the model with parameters shown as ϴ. X_i signifies the input appearance of picture while Z_i^{GT} is the ground fact of the input picturesX_i..The whole be counted of the human beings in the image can be bought by means of sum up over the anticipated density map. The network is skilled by back-propagating the usage of 12 loss computed with appreciate to ground-truth.

Following previous work of [9] and [19].MAE and RMSE was once employed as the evaluation or constrast metrics. Let N denotes number of tested images, ⟦Count⟧_gt^((n)) be the original counts, ⟦Count⟧^((n)) is the expected number of count for the n-th check image. The contrast metrics are known as follows :

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\left|Count^{(n)} - Count_{gt}^{(n)}\right|$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left|Count^{(n)} - Count_{gt}^{(n)}\right|^{2}} \qquad 8$$

where n is the number of images tested, C_i is the original crowd count for the n-th image sample and $\hat{C}_i$ is the corresponding predicted count. C_i and $\hat{C}_i$ are given by the integrating over the ground truth density map.

The assessment measure was successfully applied in preceding research [20], [21] to calculate the overall performance of predicted distinct models. [22], [23] indicated that they are reliable. The differences between the valves from a model and the actual received values are measured by the root-mean-square error (RMSE) that is commonly used. RMSE is a measure of exactness, and it is used to scrutinize predicting errors of quite a few developed models for a particular dataset. RMSE is continually a high-quality number, and a cost of 0 would point out a perfect in shape to the data .When the value of RMSE decreases, the overall performance of the model is high. RMSE is the square root of the common of squared errors. The magnitude of the squared error is relational to the effect of every error on RMSE, thus, RMSE is greater punishing of errors. [24 mention that the estimated count of image 〖Count〗^((n)) used to be generated in the equation

$$Count^{(n)} = \sum_{l=1}^{L}\sum_{w=1}^{W} z_{l,w}$$

L is the length of density map

W is the width of density map

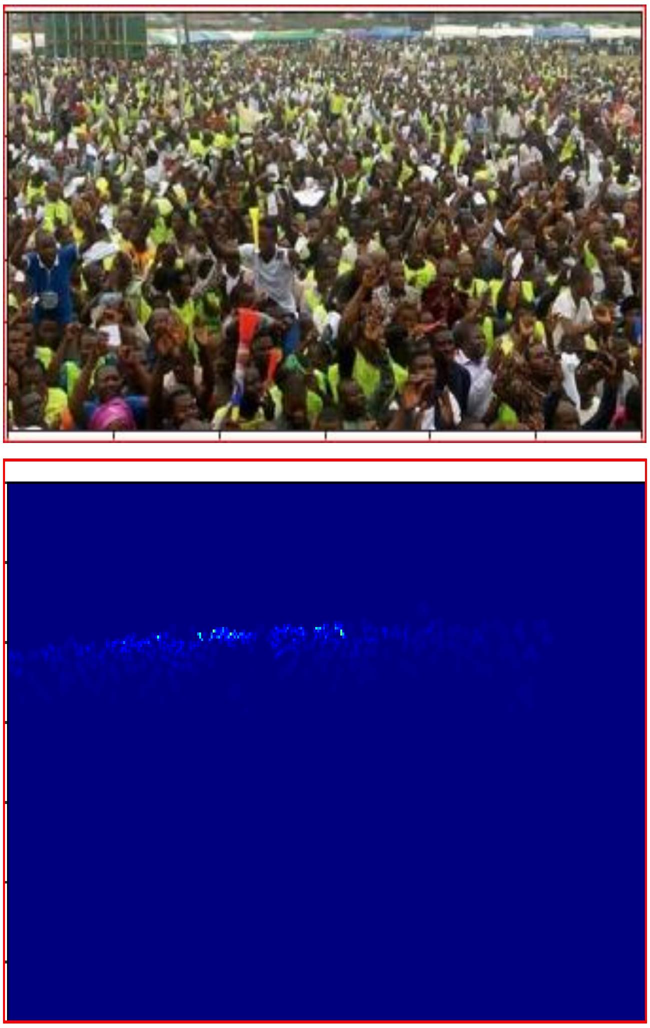
$z_{l,w} z_{l,w}$is the pixel at point (l,w) of a particular density.

$Count^{(n)}$ is the estimated crowd count for a particular image.

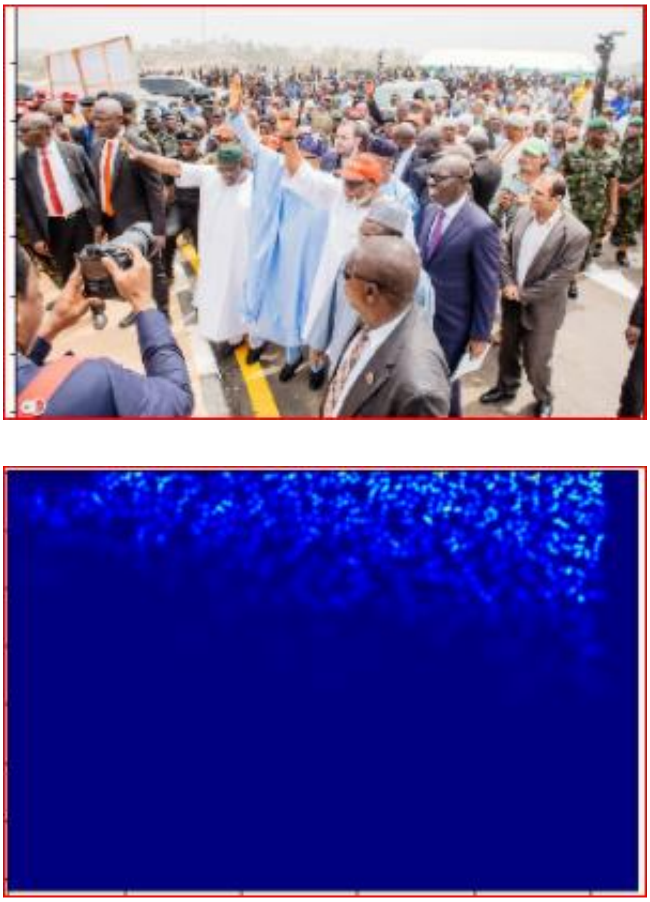
The ShanghaiTech crowd counting dataset was first presented by [13]. The dataset has 1198 clear photos with a whole of 330,165 individuals. This dataset was presented as two aspects: Section A incorporates 482 images and Section B incorporates 716 images. Section A consists of primarily incredibly congested scenes. Section B consists of pix captured from avenue and shoppingwith rather sparse crowd scenes. 300 pictures for training inPart A; four hundred images for training in Part B, the ones remaining for each are for testing.

Table 1: Estimated Errors on ShanghaiTech dataset for proposed method

Dataset	Method	MAE	RMSE
ShanghaiTech Part A	CNN-WFC	66.59	95.5
ShanghaiTech Part B	CNN-WFC	11.01	14.2

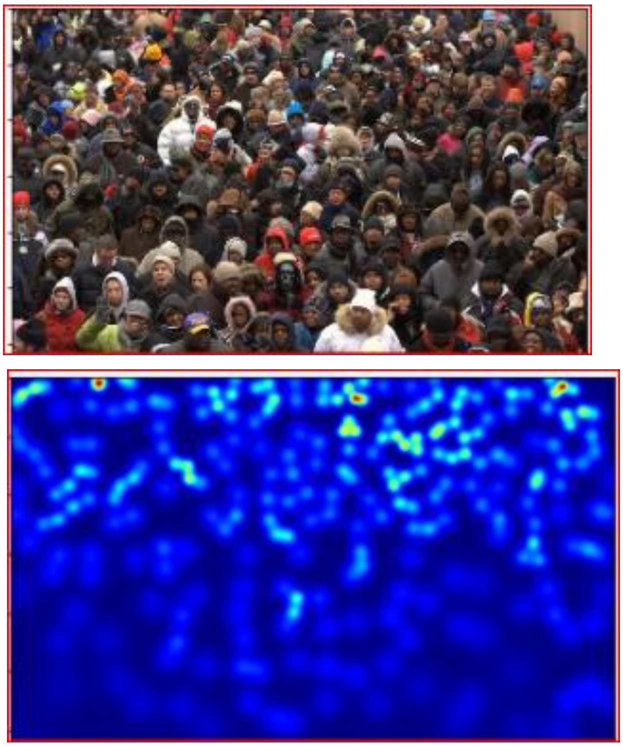


a
Original count 389.9
Predicted count 395



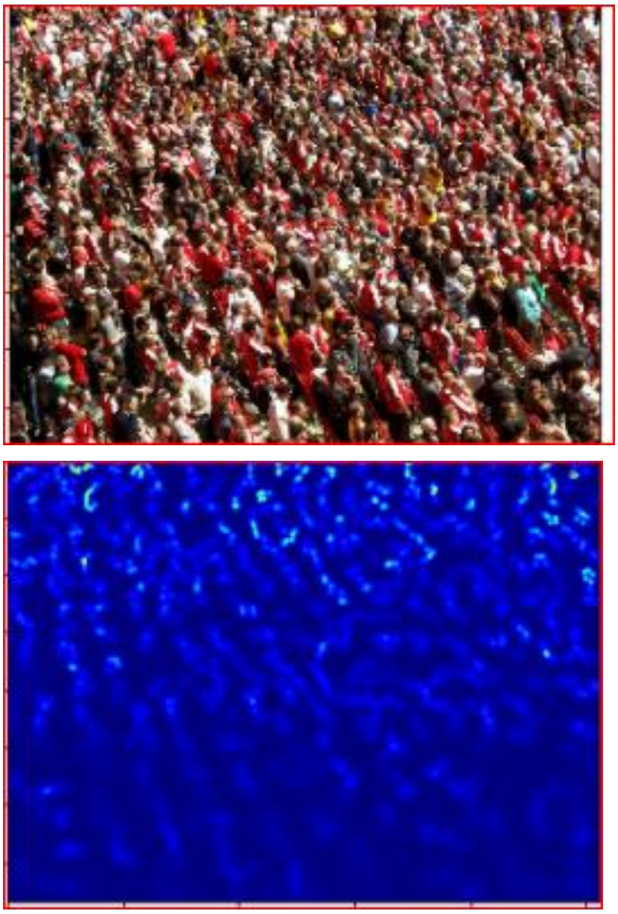
b

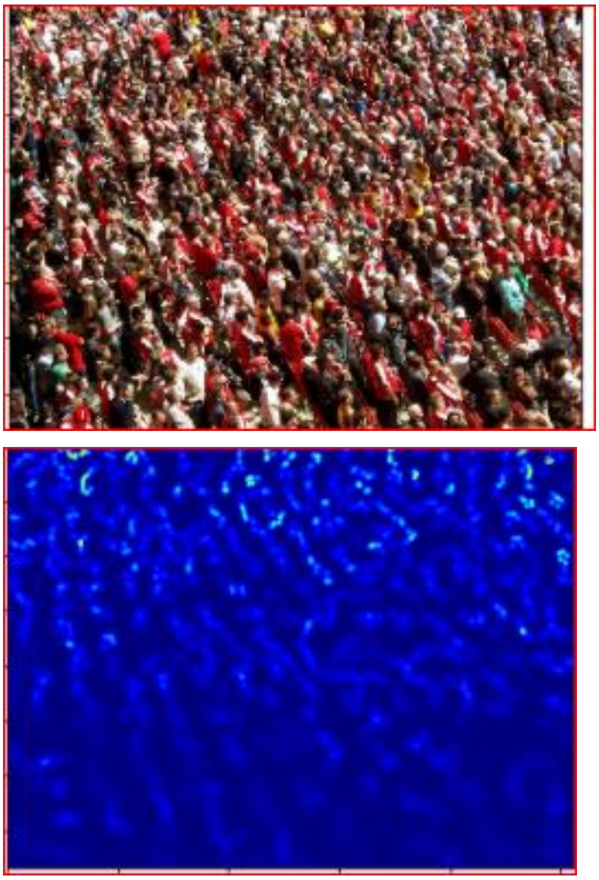
Original count 544
Predicted count 539



c

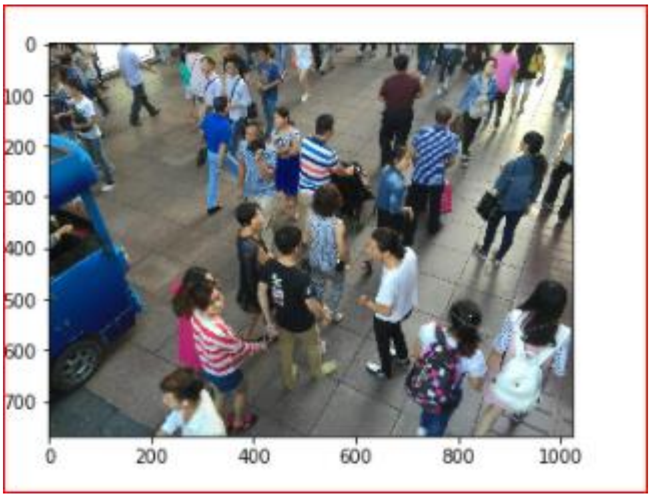
Original count 243
Predicted count 234



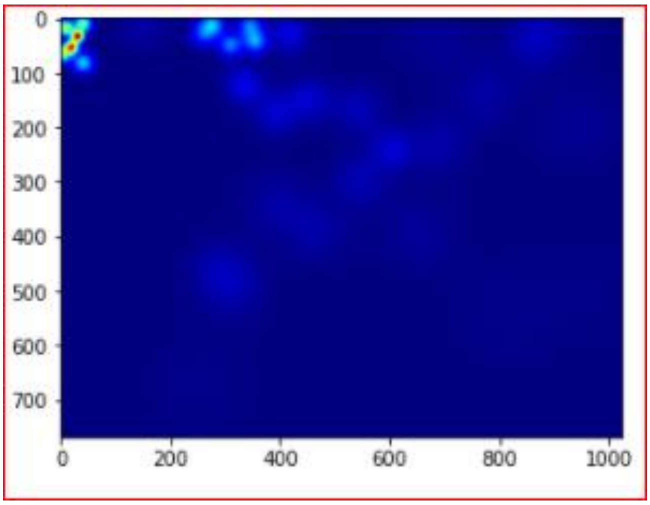


d
Original count 799
Predicted count 629

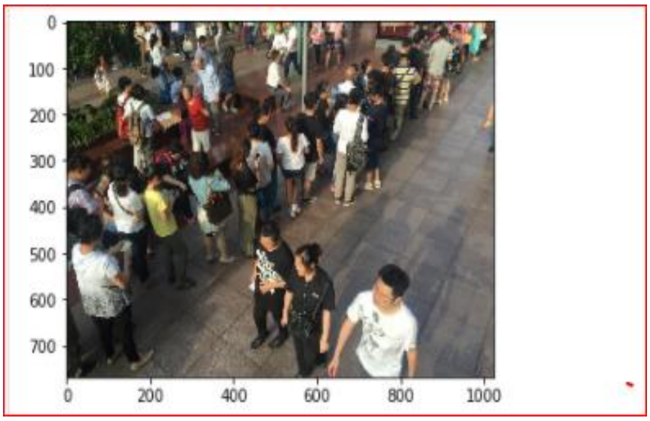
Figure 3: Density Estimation results of shanghai dataset for dense population c and d , Density estimation of pictures added to shanghai data a and b

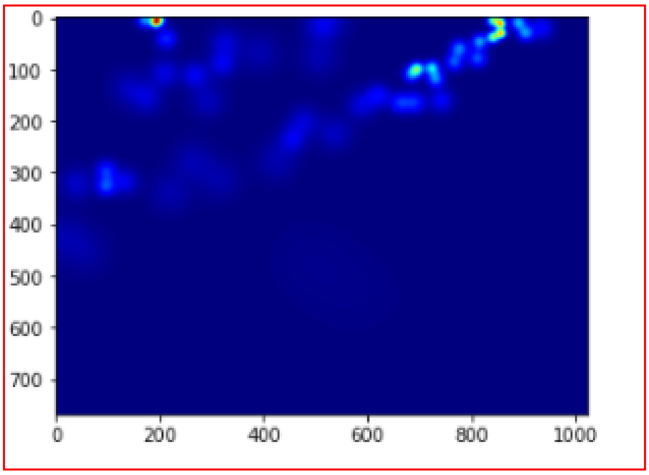


d



e
Original count 30
Predicted count 38





F

Original count 49
Predicted count 44

Figure 4: Density estimation results of shanghai datasets for sparse population e and f

Table 2: Counting Performance of ShanghaiTech dataset between CNN-WFC and other related previous method.

Method	**ShanghaiTech Part A**		**ShanghaiTech Part B**	
	MAE	**RMSE**	**MAE**	**RMSE**
MCNN (Zhang *et al.*, 2016)	110.2	173.2	26.4	41.3
CSRNet (Li *et al.*, 2018)	68.2	115.0	10.6	16.0
CP-CNN (Sindagi and Patel, 2017)	73.6	106.4	20.1	30.1
IC-CNN (Ranjan *et al.*, 2018)	68.5	116.2	10.7	16.2
Switch-CNN (Sam *et al.*, 2017)	90.4	135.0	21.6	33.4
ACM-CNN(Zhikan etal ., 2019)	72.2	103.5	17.5	22.7
Our crowd CNN model	**66.59**	**95.2**	**11.00**	**14.2**

CONCLSION

In This work an advantageous CNN based totally approach was developed and applied to generate density map and estimate crowd rely for both sparse and dense population with the use of the Shanghai datasets. . Dilated convolutional layers was once employed to acquire the background data in the dense scenes making the model to increase the receptive subject besides reduction in picture pleasant with the aid of the dilated convolutional layers

REFERENCES

[1] Loy, C.C., Chen, K., Gong, S., and Xiang, T. (2013). Crowd counting and profiling: Methodology and evaluation, in: Modeling, Simulation and Visual Analysis of Crowds. *Springer*, pp. 347–382.

[2] Lin, S.F., Chen, J.Y., and Chao, H.X. (2001). Estimation of number of people in crowded Scenes using perspective transformation. IEEE Transactions on Systems, Man, and Cybernetics-Part A: *Systems and Humans*, Vol.31, No. 6, pp.645–654

[3] Li, M., Zhang, Z., Huang, K., and Tan, T. (2008). Estimating the number of people in crowded Scenes by mid based foreground segmentation and head-shoulder detection, in: Pattern Recognition,. *ICPR 2008. 19th International Conference on, IEEE*. pp. 1–4.

[4] Mikolajczyk,K.;Zisserman,A.;Schmid,C.shape rEcognitionWith Edge-Based Features.2003.Available online:http://hal.inria.fr/inria-00548226/(accessedon 11 september2020)

[5] Hwang, J.W.; Lee, H.S.Adaptive image interpolation based on local gradient features. IEEE signal Process. Lett. 2004,11,359-362.

[6] Chan,A.B.;Liang,Z.S.J.;Vasconcelos,N.Privacy preserving crowd monitoring:Counting people without people modelsor tracking.In Proceedingsof theIEEE Conference on Computer Vision and Pattern Recognition(CVPR2008),Anchorage,AK,USA,24 june2008;pp.1-7.

[7] Lempitsky, V., and Zisserman, A. (2010). Learning to Count Objects in Images . *Advances in Neural Information Processing Systems*, pp. 1324-1332

[8] Pham, V.Q., Kozakaya, T., Yamaguchi, O., and Okada, R. (2015). Countforest: Co-voting

uncertain number of targets using random forest for crowd density estimation, in: Proceedings of the *IEEE International Conference on Computer Vision*, pp. 3253–3261.

[9] Wang, C., Zhang, H., Yang, L., Liu, S., and Cao, X. (2015). Deep people counting in extremely dense crowds. In *Proceedings of the 23rd ACM international conference on Multimedia*pp. 1299-1302.

[10] Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., and Zhu, C. (2015). Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence* Vol.43, pp.81–88.

[11] Sindagi, V. A. and Patel, V. M. (2017). A survey of recent advances in CNN-based single image crowd counting and density estimation. *Pattern Recognition Letters*, pp.2-17.

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

[13] Boominathan, L., Kruthiventi, S. S., and Babu, R. V. (2016). Crowdnet: A deep convolutional network for dense crowd counting. *In Proceedings of the 24th ACM international conference on Multimedia* pp. 640-644.

[14] Lendasse, A., Bodt, E. D., Wertz, V. and Verleysen, M. (2000). Non-Linear Financial Time

Series Forecasting: Application to the Bel 20 Stock Market Index. *European Journal of Economic and Social Systems,* Vol. 14, pp. 81-91.

[15] Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016). Single image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEEconference on computer vision and pattern recognition, pp. 589–597.

[16] Sam, D. B., Surya, S., and Babu, R. V. (2017). Switching convolutional neural network

for crowd counting. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4031-4039.

[17] Yuhong, L., Xiaofan, Z., and Deming, C. (2018). CSRNet: Dilated Convolutional Neural

Networks for Understanding the Highly Congested Scenes, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recogmition* , pp. 1-16.

[18] Lokesh, B., Srinivas, K., and Venkatesh, B. (2016). CrowdNet: A Deep Convolutional Network for Dense Crowd Counting. *Computer Vision and Pattern Recognition* , pp. 1-5.

[19] Vishwanath, S. A., and Vishal, P. M. (2017). CNN-based Cascaded Multi-task Learning of High level Prior and Density Estimation for Crowd Counting. International Conference on Advanced Video and Signal Based Surveillance pp. 1-6.

[20] Karen, S., and Andrew, Z. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd International Conference on Learning Representations (ICLR 2015) pp. 1-15.

[21] Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016). 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.

[22] Khashei, M. and Bijari, M. (2010). An artificial neural network (p,d,q) model for time-series forecasting. Expert Systems with Applications, Vol. 37, No. 1, pp.479-489

[23] Arunraj, N. S. and Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. International Journal of Production Economics, Vol. 170, pp.321-335.

[24] Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., and Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental pollution*, Vol. 231, pp.997-1004.

[25] Cortez, B., Carrera, B., Kim, Y. J. and Jung, J. Y. (2018). An architecture for emergency event prediction using LSTM recurrent neural networks. Expert Systems with Application, Vol. 97, pp.315-324.

[26] Li, Y., Zhang, X., and Chen, D. (2018). 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.

[27] Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., and Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental pollution*, Vol. 231, pp.997-1004.

[28] Zhao, Z., Li, H., Wang, X., and Zhao, R. (2016). Crossing-line Crowd Counting with two-phase Deep Neural Networks. *European Conference on Computer Vision,*Springer, pp. 712-726.

[29] Aziz, M., Naeem, F., Alizai, M., and Khattak, Dr. (2017). Automated Solutions for Crowd Size Estimation. *Social Science Computer Review*, Vol. 36, pp. 1-22.

[30] Arunraj, N. S. and Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics*, Vol. 170, pp.321-335.

[31] Ranjan, V., Le, H., and Hoai, M. (2018). Iterative crowd counting. *In Proceedings of the* European Conference on Computer Vision (ECCV), pp. 270–285.