International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Locate, Size and Count: Accurately Resolving People in Dense Crowds via Detection

Sanyam Swami (Student)[1], Prof. Sonal Fatangare (Guide)[2], Saisagar Singh(Student)[3], Nandakumar Swami(Student)[4], Pranay Sankatala(Student)[5]

[1,3,4,5] Student, Dept of Computer Engineering, RMD Sinhgad School of Engineering, India
[2]Professor, Dept of Computer Engineering, RMD Sinhgad School of Engineering, India

---

Abstract - *This way of life makes life easier for people and increases the use of public services in metropolises. We present a CNN-MRF- grounded system for counting people in still images from colourful scenes. Crowd viscosity is well represented by the features deduced from the CNN model trained for other computer vision tasks. The neighbouring original counts are explosively identified when using the lapping patches separated strategies. The MRF may use this connection to smooth conterminous original counts for a more accurate overall count. We divide the thick crowd visible image into lapping patches, also prize features from each patch image using a deep convolutional neural network, followed by a fully connected neural network to regress the original patch crowd count. Since the original patches lap, there's a strong connection between the crowd counts of neighbouring patches. We smooth the counting goods of the original patches using this connection and the Markov random field.*

Key Words: **Convolutional Neural Network, Sign Language, Machine Learning, Image Processing, Feature Extraction**

1. INTRODUCTION

There are two major groups of being models for estimating crowd density and counting the crowd direct and circular approaches. The direct approach (also known as object discovery grounded) is grounded on detecting and segmenting each person in a crowd scene to get a total count, while the circular approach (also known as point grounded) takes a picture as a whole and excerpts some features before getting the final count. Due to variations in perspective and scene, the distribution of crowd density in crowded crowd images is infrequently harmonious. As a result, counting the crowd by looking at the entire picture is illogical. As a result, the divide-count-sum approach was acclimated in our system. After dividing the images into patches, a regression model is used to collude the image patch to the original count. Eventually, the accretive number of these patches is used to calculate the global image count. There are two benefits of image segmentation: To begin with, the crowd density in the small picture patches has a fairly invariant distribution. Second, image segmentation improves the quantum of training data available to the regression model. Because of the benefits mentioned over, we can train a more robust regression model.

2. LITERATURE SURVEY

Crowd safety in public places has always been a serious but delicate issue, especially in high-density gathering areas. The higher the crowd level, the easier it is to lose control, which can affect in severe casualties. In order to prop in mitigation and decision-making, it is important to search out an intelligent form of crowd analysis in public areas. Crowd counting and density estimation are precious factors of crowd analysis, since they can help measure the significance of conditioning and give applicable staff with information to prop decision-making. As a result, crowd counting and density estimation have become hot motifs in the security sector, with operations ranging from videotape surveillance to traffic control to public safety and civic planning. A crowd monitoring system is in veritably high demand these days. Still, current crowd monitoring system products have a number of excrescencies, similar as being constrained by operation scenes or having low perfection. In particular, there is a lack of exploration on tracking the number of pedestrians in a large-scale crowded area (see Figure 1). The detection-based methods and the regression-based methods are the two types of crowd counting styles. Detection-based crowd counting styles generally employ a sliding window to descry each pedestrian in the scene, calculate the pedestrian's approximate position, and also count the number of pedestrians . For low-density crowd scenes, detection-based methods may produce decent results, but they are oppressively confined for high-density crowd scenes. The early regression based styles attempt to learn a direct mapping between low-level features deduced from original image blocks and head count. Direct regression-based approaches like these only count the number of pedestrians while missing essential spatial information. Learning the linear or non-linear mapping between original block features and their matching target density maps, as indicated by references, may integrate spatial information into the literacy process. Experimenters were inspired by the Convolutional Neural Network's (CNN) performance in numerous computers vision tasks to use CNN to learn nonlinear functions from crowd images to density maps or counts. In 20205, Wang et al used the Alexnet network structure to

apply CNN to the crowd counting charge. To count the number of pedestrians in the crowd picture, the fully connected layer with 4096 neurons was replaced by a layer with only one neuron. In the same year, Zhang et al discovered that when existing approaches were applied to new scenes that varied from the training dataset, their output was significantly reduced. To address this problem, a data-driven approach was proposed for fine-tuning the pre-trained CNN model with training samples that were close to the density level in the new script, allowing it to acclimate to unknown operation scenes. This approach eliminates the need for retraining when the model is converted to a new script, but it still necessitates a large quantum of training data, and it is delicate to prognosticate the density level of the new scene in practice. In 20206, Zhang et al proposed a multi-column convolutional neural network-based architecture (MCNN) based on the success of multi-column networks in image recognition by constructing a network conforming of three columns of filters corresponding to the receptive fields with different sizes (large, medium, small) to acclimatize to changes in head size due to perspective goods or ima. Of column of the MCNN pre-trains all image blocks during training, also the three networks are combined for fine-tuning training. The training process is complicated, because there is a lot of redundancy in the structure. Sam et al proposed in 20207 that the convolutional neural network for crowd counting (Switching CNN) be used to train regressions using a specific collection of training data patches based on different crowd densities in the picture. The network is made up of multiple independent CNN regressions, analogous to a multi-column network, with the addition of a Switch classifier based on the VGG-16 architecture to pick the best regression for each input block. Alternatively, the Switch classifier and the independent regression are trained. Switching CNN, on the other hand, switches between regressions using the Switch classifier, which is veritably expensive and frequently unreliable. Analogous to Refs, Kumaga et al suggested a hybrid neural network Mixture of CNNs in 20207, believing that a single predictor in colorful scene surroundings is inadequate to directly prognosticate the number of pedestrians (MoCNN). A combination of expert CNNs and a gated CNN makes up the model framework. On the base of the environment of the input picture, the applicable expert CNN is adaptively named. Expert CNNs estimate the image's head count in vaticination, while gated CNN estimates each expert CNN's respectable liability. These odds are also used as weighting factors in calculating a weighted average of all expert CNNs' head counts. Via gated CNN preparation, MoCNN not only trains multitudinous expert CNNs, but also learns the liability of each expert CNN's approximate head count. Still, it can only be used for crowd counting estimation and does not have information on crowd density distribution. Tang et al proposed a low-rank and sparse-based deep-fusion convolutional neural network for crowd counting (LFCNN) that bettered the delicacy of the projection from the density map to global counting by using a regression approach based on low-rank and sparse penalty. By rooting point charts from different layers and conforming them to have the same output size, Zhang et al proposed scale-adaptive CNN (SaCNN) to estimate the crowd density map and incorporate the density map to get a more accurate estimated head count. To achieve the head count in static images, Han et al combined convolutional neural network and Markov Random Field (CNN-MRF), which comported of three corridor: a pre-trained deep residual network 152 to prize features, a fully connected neural network for count regress, and an MRF to smooth the counting goods of the original patches. High correlation of near patches was used to increase count delicacy in this way. In this paper, a feature fusion-based deep convolutional neural network system, FF-CNN (Feature Fusion of Convolutional Neural Network), was proposed to achieve more accurate crowd counting output in high-density and complex surroundings. The aim of FF-CNN was to collude the crowd picture to its crowd density map, and then use integration to get the head count. The geometryadaptive kernels were used to induce high-quality density maps that were used as training ground trueness, as described by MCNN . To gain richer functionality the VGG network was used as the FF-CNN box network. The fusion of high-level and low-level features was achieved using the deconvolution technique . Two loss functions, density map loss and absolute count loss, were combined to optimize for a more precise density map and a more precise crowd count. For each replication, the original images were cropped to 256 256 images using an arbitrary cropping process to maximize sample diversity.

3. PROPOSED SYSTEM

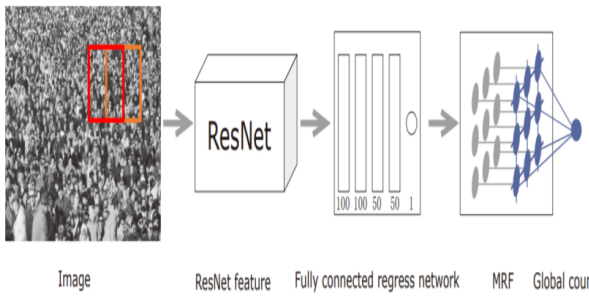
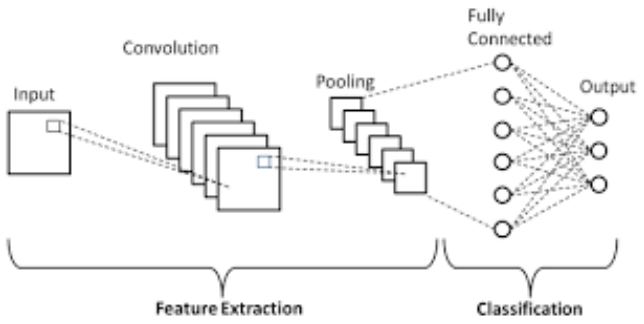


Fig:- System Architecture

We use a fully connected neural network to learn a map from the above features to the original count, and a pre-trained deep residual network to prize features from image patches. Deep convolutional network features have been used in a variety of computer vision tasks, including image recognition, object discovery, and image segmentation. This suggests that the deep convolutional network's learned features are applicable to a wide range of computer vision tasks. The representation capability of the learned features improves as the number of network layers increases. A deeper model, on the other hand, necessitates more data for

preparation. Current datasets for crowd counting are inadequate to train a veritably deep convolutional neural network from scrape. To extract features from an image patch, we use a pre-trained deep residual network. Rather of learning unreferenced functions, their approach resolved the declination issue by reformulating the layers as learning residual functions with reference to the subcaste inputs. To prize the deep features that reflect the density of the crowd, we use the residual network, which was trained on the ImageNet dataset for image bracket. For every three convolution layers, this pre-trained CNN network generated a residual item, bringing the total number of layers in the network to 152. To get the 1000-dimensional features, we resize the image patches to 224 224 pixels as the model's input and prize the fc1000 subcaste's output. Following that, the features are used to train a five-subcaste completely linked neural network. The input to the network is 1000-dimensional, and the network's number of neurons is 100-100-50-50-1. The original crowd count is the network's output. The fully linked neural network's literacy part is to minimize the mean squared error of the training patches. Image Counting Approach Since the group datasets are extremely little, there is a destined number of preparing tests, and when there is a application of deep learning methods, these datasets are worse an acceptable quantum of to prepare a deep system. When we put on deep systems to these datasets then the issues are less feasible but rather more it ought to be. Along these lines, we propose a two-level deep expansion-based methodology for group checking that causes our deep system to handle with the issue.

Algorithm Used CNN



Why CNN?

- CNNs are employed for image classification and recognition of its high perfection.

- The CNN follows a various leveled model which deals with erecting an organization, analogous to a pipe, incipiently gives out a fully associated layer where every one of the neurons are associated with one another and the result is handled.

- Hereafter we are involving Convolutional Neural Network for proposed framework.

4. EXPERIMENTAL AND RESULT

In this paper, we have performed our proposed system on the ShanghaiTech dataset. The testing results are shown in Table 1.

Table 1: Comparison of original count and predicted countfrom various images

Original Count	Predicted Count	Original Count	Predicted Count
1110	897	370	117
296	113	501	460
567	255	1067	904
171	285	320	350
169	86	583	405
816	905	761	440
360	399	340	216
1325	337	415	327

Fig: - Experiment 1

5. CONCLUSIONS

We present a CNN-MRF-based method for counting people in still images from several scenes. Crowd density is well represented by the features deduced from the CNN model trained for other computer vision tasks. The neighboring colorful counts are explosively identified when using the overlapping patches separated strategies. The MRF may use this connection to smooth conterminous original counts for a more accurate overall count. Experimental findings show that the proposed system outperforms other recent affiliated methods.

- The system will give better accuracy for crowd detection from heterogeneous images.

- This approach is suitable to work on image as well as videotape dataset respectively.

- Various feature extraction selection ways provides good detection accuracy.

• System uses RESNET from deep convolutional network that provides up to 152 hidden layers.

REFERENCES

[1] Fruin, J. "Pedestrian planning and design, metropolitan association of urban design and environmental planners." Inc., New York 20.6 (1971).

[2] Zhan, Beibei, et al. "Crowd analysis: a survey." Machine Vision and Applications 19.5 (2008): 345-357.

[3] Zeng, Lingke, et al. "Multi-scale convolutional neural networks for crowd counting." 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017.

[4] Zhang, Cong, et al. "Cross-scene crowd counting via deep convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[5] Leibe, Bastian, Edgar Seemann, and Bernt Schiele. "Pedestrian detection in crowded scenes." 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. IEEE, 2005.

[6] Zhao, Tao, Ram Nevatia, and Bo Wu. "Segmentation and tracking of multiple humans in crowded environments." IEEE transactions on pattern analysis and machine intelligence 30.7 (2008): 1198-1211.

[7] Ge, Weina, and Robert T. Collins. "Marked point processes for crowd counting." 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009.