Development of Crowd Management System Using CSRNET

Dr.S.J.Subhashini, C.Jothi Aruna, C.Kaja Priya, S.Sanmugapriya, N.Varshini

Associate Professor, Dept. of C.S.E, Velammal College of Engineering and Technology, Madurai, India

UG Scholar, Dept. of C.S.E, Velammal College of Engineering and Technology, Madurai, India

UG Scholar, Dept. of C.S.E, Velammal College of Engineering and Technology, Madurai, India

UG Scholar, Dept. of C.S.E, Velammal College of Engineering and Technology, Madurai, India

UG Scholar, Dept. of C.S.E, Velammal College of Engineering and Technology, Madurai, India

ABSTRACT: Smart cities aim not solely to form people's lives more pleasant however additionally safer using advanced technology. Being a packed community areas like colleges, stadiums, subway stations or holy spots on pilgrim's journey impacts not solely the amount of human convenience however specifically the threat of human security. An abnormal crowd conduct will cause push, mass panic, stampede, crowd-crush, and inflicting overall management loss. Ancient crowd-counting techniques in an image are remodeled via machine-learning and artificial-intelligence techniques into intelligent crowd-management techniques. This paradigm shift offers several advanced features in terms of accommodative watching and also the governance of dynamic crowd gatherings. Adaptive observing, distinguishing/recognition, and also the management of numerous crowd gatherings will improve several crowd-management-related tasks in terms of potency, capacity, reliableness, and safety. The Congested Scene Recognition Network called CSRNet has two major components: a convolutional neural network (CNN) is the front-end , feature extraction and an expanded CNN for the back-end that uses expanded kernels to deliver larger reception fields and to switch pooling operations and also produce a data-driven and deep learning technique that may recognize extremely congested scenes and perform correct count estimation as well as present high-quality density maps. . CSRNet is an easy-trained model attributable to its pure convolutional structure, which triggers an alarm just in case of accelerating crowd level. The recommended framework provides a good technique to attach and alert all the workers resources forthwith, preventing danger ensuing from crowd surge.

KEYWORDS: convolutional neural network (CNN), Congested Scene Recognition (CSRNET)

I. INTRODUCTION

Crowd counting from unconstrained scene images may be critical task in several real-world applications like urban police work and management; however it's greatly challenged by the camera’s perspective that causes vast look variations in people’s scales and rotations. It may be performed in numerous ways that, like digital-image process, machine learning, and deep learning. Typical ways address such challenges by resorting to mounted multi-scale architectures that area unit usually unable to hide the for the most part varied scales whereas ignoring therotation variations. Growing range of network models are developed to deliver promising solutions for crowd flows monitoring, assembly controlling, and alternative security services. Current ways for congested scenes analysis are developed from simple crowd count (which outputs the amount of individuals within the in the targeted image) to density map presenting (which displays characteristics of crowd distribution).This development follows the demand of real-life applications since identical range of individuals might have fully totally different crowd distribution, so simply count the amount of crowds isn't enough. The distribution map helps us for obtaining additional correct and comprehensive info, which can be important for creating correct selections in risky environments, like stampede and riot. However, it's difficult to get correct distribution patterns. One major drawback comes from the prediction manner: since the generate density values follow the pixel-by-pixel prediction, output density maps should enfold abstraction coherence so they will gift the sleek transition between nearest pixels. Also, the diversified scenes, e.g., irregular crowd clusters and totally different camera views, would create the task particularly for mistreatment ancient ways while not deep neural networks (DNNs).The DNN-based methods attributable to the high accuracy they have achieved in semantic segmentation tasks and also the important progress they need to be created in visual saliency. The additional bonus of using DNNs comes from the dynamic hardware community where DNNs are rapidly examined and executed on GPUs, FPGAs, and ASICs. Among them, the low-power, small-scale schemes are especially suitable for deploying congested scene analysis in surveillance devices.

II. RELATED WORK

The solutions that are capable enough for crowd scenes analysis can be classified into following categories: detection-based methods, regression-based methods, and density estimation-based methods. By combining the deep learning, the CNN-based solutions show even stronger ability during this task and crush the normal ways.

A. Detection-based approaches

Most of the early researches concentrate on detection-based approaches employing a moving-window-like detector to detect individuals and count their range [1]. These methods need well-trained classifiers to extract low-level features from the full physical structure (like Haar wavelets [21] and HOG (histogram oriented gradients) [3]). However, they perform poorly on extremely congested scenes since most of the targeted objects square measure obscured. To tackle this problem, researchers observe particular body parts rather than the full body to complete crowd scenes analysis.

B. Regression-based approaches

Since detection-based approaches cannot be adapt to highly congested scenes, researchers attempt to deploy regression-based approaches to seek out the relations among extracted features from cropped image patches, and so calculate the amount of explicit objects. More features, such as foreground and texture features, have been used for generating low-level info[4]. Following similar approaches, Idrees et al. propose a model to extract features by employing Fourier analysis and SIFT (Scale invariant feature transform) interest-point primarily counting.

C. Density estimation-based approaches

During the execution of the regression-based solution, one evaluative feature, called saliency, is overlooked this causes inaccurate leads to local regions. Lempitsky et al. [5] propose a method to solve this problem by learning a linear mapping between features in the local region and its object density maps. It incorporates the particulars of saliency during the learning process. Since the ideal linear mapping is hard to obtain, Phametal. [6] use random forest regression to learn a non-linear mapping instead of the linear one.

D. Switch-CNN

The Switch Convolutional Neural Network (Switch-CNN) can be employed with local density changes in crowd scenarios well with the exception of its ability to model large scale variations. The details in the image will be weakened along with the reduction in contrast of the image and the edges in the image being blurred to a certain extent which is due to the usage of the weighted averaging technique in MCNN. It is challenging to attain satisfactory fusion effect in most implementations, so that it becomes especially influential to leverage local variations in density. The Switch-CNN structure has about three CNN regressors with varying receptive fields leading architectures and the switch that selects the correct regressor for the input patches. The usage of three CNN regressions introduced in it, R1, R2, R3 [7] has been done in this network. R1 is a 9*9 considerable-sized filter that apprehends progressive features in the scene. R2 and R3 are 7×7 and 5×5 filters that are used to detain features in low scales respectively. The switch classifier and a switch layer are the composition of the switch network. The switch classifier deduces the label of the CNN regressor appropriate for the input image patches. The switch layer has been provided with the label and transmits the patches to the correct regressor. The adaptation of the VGG-16 network is used in the switch as the switch classifier for a three-way classification [8]. The global average pool (GAP) has banned the fully-connected layers in VGG-16 and is followed by smaller fully-connected layers and a3-class soft-max classifier, corresponding to three CNN regressors in the Switch-CNN. Initially, the Switch-CNN splits the image into 3*3 non-overlapping patches based on certain crowd characteristics, then uses a switch classifier to classify the patches by density standard, and then relays the patches to the independent CNN regressor with different receptive fields and field-of-view.

E. Contextual Pyramid-CNN:

The CP-CNN method consists of a pyramid of context estimators and a Fusion-CNN. It consists of 4 modules: GCE, LCE, DME, and F-CNN. The former two types are CNN-based networks that encode global and local context present in the input image respectively. DME is nothing but a multi-column CNN that performs the initial task of remodeling the input image to high-dimensional feature maps. Ultimately, F-CNN merges dependent information from GCE and LCE with multi-dimensional feature maps from DME to provide an apparent resolution and finest density maps.

III. PROPOSED ALGORITHM

In this paper, the deeper network called CSRNet is used for counting crowd and generating high-quality density maps, from which the occurrence of overcrowd is detected. Unlike the latest works such as Switching convolutional neural network for crowd counting and contextual pyramid CNNs, which use the deep CNN for ancillary, we focus on designing a CNN-based density map generator. Our model uses convolutional layers as the backbone to support input images with flexible resolutions. To limit the network complexity, we use the small size of convolution filters in all layers. We deploy the first 10 layers from VGG-16 implemented as front-end and dilated convolution layers implemented as a back-end to enlarge receptive fields and extract deeper features without losing resolutions (since pooling layers are not used).

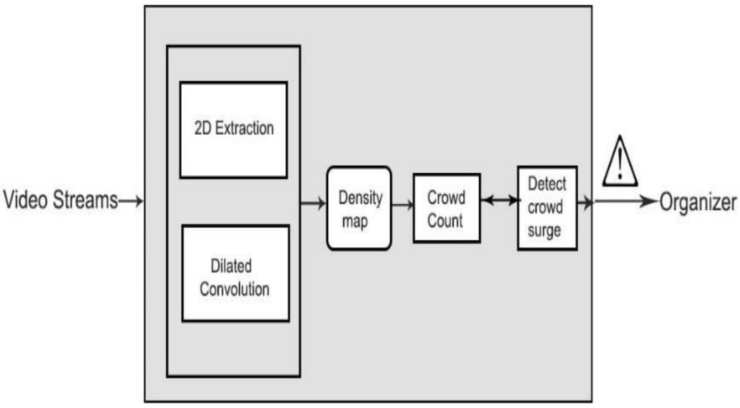


Fig 1. System Architecture

CSRNET ALGORITHM:

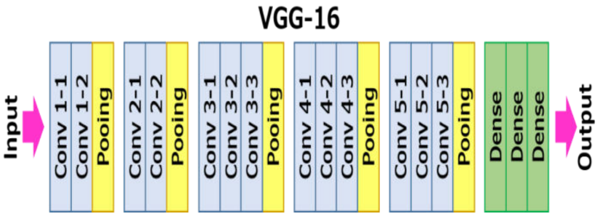
The proposed CSRNet is consist of two major components:

1.A convolutional neural network (CNN) is utilized as the front-end for 2D feature extraction.

2. A dilated CNN delivers the reception fields to replace pooling using dilated kernels.

2-D Feature Extraction:

In this feature extraction we use the front-end CNN which is similar to first ten layers of VGG-16 with three pooling layers, considering the tradeoff between accuracy and the resource overhead. VGG-16 is nothing but a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford. The aim of the model in ImageNet to achieve 92.7% which is top-5 test precision.



The primary thing is that there are 13 convolutional layers, 5 Max Pooling layers and three dense layers which encapsulate to 21 layers but only 16 weight layers. The detailed view of 13 convolutional layer is as follows: Conv 1 has 64 number of filters while Conv 2 has 128 number of filters, Conv 3 has 256 number of filters while Conv 4 and Conv 5 has number of 512 filters. The image is proceeded through a set of convolutional layers, where the filters were used with small receptive field: 3×3. Pooling layer is usually between consecutive convolutional layers. It is used to reduce the spatial size of the representation to scale back the number of parameters and also the computation of the network. The pooling layer is applied independently to each depth slice of the input and it reduces the spatial dimensions of the input. It is mostly used to reduce over-fitting. If an input is applied a MAX POOLING with a filter size and stride of 2*2 then both width and height will be sampled by the input size a factor of 2 keeping the depth unaffected which suggests it discarded 75% of the activation

Segmentation by dilated CNN:

The back-end CNN is a series of dilated convolutional layers producing density maps. The objective of semantic segmentation is categorizing each pixel of the input image into a given set of classes. The main challenge of this is often to mix pixel-level accuracy with multi-scale contextual information. The previous state of the art models is based on the adaptations of convolutional neural networks designed for image classification. The idea of Dilated Convolution was motivated as it enlarges the receptive field while maintaining resolution and also it comes from the wavelet decomposition. This module is an altered version of the adapted VGG-16 network for semantic segmentation by removing the last two pooling and striding layers. Dilated convolutions use specific kernels with dispersed aligned weights. Expansion of both the size of kernel and the sparse weights interval exponentially with dilation factor. By increasing dilation factor, receptive field is additionally expanded exponentially by large kernel. Dilated Convolution (Basic or Large) can always improve the results and doesn't overlap with the other post-processing steps. A dilated convolution is fundamentally an abstraction of the traditional 2D convolution that allows the operation to skip some inputs. This enables an increase in the size of the receptive filter without losing resolution.

IV. PROPOSED MODULES

In the previous section, we have already shown system architecture and its performance. In this section describes module names and its functions. The modules are following.

A. DILATED CONVOLUTION:

One of the critical components of our design is that the dilated convolutional layer. A 2-Ddilated convolution is often defined as follow:

y(m, n) = X M i=1 X N j=1 x(m + r × i, n + r × j)w(i, j)

y(m, n) is that the output of dilated convolution from input x(m, n) and a filter w(i, j) with the length and then the width of M and N respectively. The parameter r is the dilation rate. If r = 1, a dilated convolution turns into a traditional convolution. A good alternative of pooling layer is dilated convolutional layers that has a significant accuracy improvement and have been demonstrated in segmentation tasks. Pooling layers has the functionality like maintaining invariance and controlling over fitting, they also progressively reduce the spatial resolution i.e the spatial information of feature map is lost. De-convolutional layers can alleviate the loss of information, but the additional complexity and execution latency may not be suitable for all cases. Adding more convolutional layers can make larger receptive fields but introduce more operations. A small-scale kernel with k × k filter in dilated kernel turns into k + (k − 1)(r − 1) with dilated stride r. Thus, it allows flexible aggregation on of the multi-scale contextual information while keeping an equivalent resolution. Example can be found in Fig where normal convolution gets 3 × 3 receptive field and two dilated convolutions deliver 5 × 5 and 7 × 7 receptive fields respectively.

The input is an image of crowds, and it is processed by two approaches separately for generating output with an equivalent size. In the first approach, pooling layer with factor 2 down sample the input, and then it is passed to a 3×3 Sobel kernel convolution layers. Since the generated feature map is only 1/2 of the original input, it needs to be up sampled by the de-convolutional layer (bilinear interpolation). In the other approach, we try dilated convolution and adapt the equivalent 3 × 3 Sobel kernel to a dilated kernel with a factor which is equal to 2 strides. The output is shared the equivalent dimension as the input. Mostly noted that the output from dilated convolution contains more detailed information.

B. GROUND-TRUTH GENERATION:

The term "ground truth" means accuracy of the training set's classification for supervised learning techniques. Consider an input as x(m,n), a filter as w(i ,j), and the dilation rate r. The output y(m,n) will be:

$$y(m,n) = \sum_{i=1}^{M}\sum_{j=1}^{N} x(m + r \times i, n + r \times j)w(i,j)$$

This equation using a (k*k) kernel generalized with a dilation rate r. The kernel enlarges to:
([k + (k-1)*(r-1)] * [k + (k-1)*(r-1)]

So, then each image has a generated ground truth. Gaussian kernel is used to blur each person's head in a given image. The 4 quarters are divided from the first 4 patches and the remaining patches are randomly cropped. The training set is doubled by using each patch.

C. DENSITY MAP GENERATION:

The ground-truth crowd density map is generated from the ground-truth dot annotations of people. Given an input image, a model is trained to predict the density map, which is summed to obtain the predicted count. Following the method of generating density maps, we use the geometry-adaptive kernels to tackle the highly congested scenes. By blurring each head annotation employing Gaussian kernels, which generates the ground truth considering the spatial distribution of all images from each dataset. The geometry-adaptive kernel is defined as:

$$F(x) = \sum_{i=1}^{n} \delta(x - xi) \times G\sigma i(x), \text{ with } \sigma i = \beta di$$

For each targeted object xi within the ground truth δ, we use di to point out the average distance of k nearest neighbors. To generate the density map, we combine δ(x−xi) with a Gaussian kernel with parameter σi (standard deviation), where exist the position of pixel within the image. In experiment, we follow the configuration in [9] where $\beta = 0.3$ and $k = 3$. For sparse crowd input, we alter to make the Gaussian kernel to the average head size to blur all the annotations.

D.OVERCROWD DETECTION:

The overcrowd detection can be found using the appropriate crowd count in a specific area at any instant and the total number of items in the image using the multiplier we can find the density of people in the image. Provides overcrowding alert if the density of the image crossing a defined threshold value.

V. PERFORMANCE ANALYSIS

The performance metric used in CSRNet is MAE and MSE, i.e., Mean Absolute Error and Mean Square Error.

MAE: The Mean Absolute Error, also known as MAE, is one of the many metrics for summarizing and assessing the quality of a deep learning model.

MSE: MSE is the average of the squared error that is used as the loss function for least squares regression: It is the sum, over all the data points, of the square of the difference between the expected and actual target variables, divided by the number of data points.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | C_i - C_i^{GT} |$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} | C_i - C_i^{GT} |^2}$$

where N is the total number of testing images, Ci and Ci are the ground truth count and estimated count of the i-th image respectively.

Here, Ci is the estimated count:

$$C_i = \sum_{l=1}^{L}\sum_{w=1}^{W} z_{l,w}$$

L and W are the width of the predicted density map.

Comparison of architectures on ShanghaiTech Part dataset and part B:

METHOD	PART A		PART B	
	MAE	MSE	MAE	MSE
MCNN	110.2	173.2	26.4	41.3
Switch -CNN	90.4	135.0	21.6	33.4
CP-CNN	73.6	106.4	20.1	30.1
Cascaded MTL	101.3	152.4	20.0	31.1
Zhang et al	181.8	277.7	32.0	49.8
CSRNET	68.2	115.0	10.6	16.0

Table 1. Comparative Analysis for Shanghai Tech Data Set

Our model will first find the density map for a given image. The pixel value will be zero if nobody is present. A certain pre-defined value will be assigned if that pixel corresponds to a person. So, calculating the total pixel values corresponding to a person will give us the count of people in that image. It indicates that our method achieves the lowest MAE (the highest accuracy) in Part A compared to other methods and we get 7%lowerMAEthanthestate-of-the-artsolutioncalledCP-CNN. CSRNet also achieves 47.3% lower MAE in Part B compared to the CP-CNN.

VI. CONCLUSION AND FUTURE WORK

In this project, we proposed a novel architecture called CSRNet for crowd counting and high-quality density map generation with an easy-trained end-to-end approach. We used the dilated convolutional layers to aggregate the multiscale contextual information in the congested scenes. By knowing advantage of the dilated convolutional layers, CSRNet can enlarge on the receptive field without losing resolution. We demonstrated our model in four crowd counting datasets with the state-of-the-art performance.

Generally, in this project obtain a higher-quality crowd density map, the geometry adaptive kernels were adopted to generate high-quality ground truth density maps during training and the de-convolution technique were used to combine high-level and low-level features. In our future research, we plan to incorporate our model in other existing crowd flow prediction framework. We will also expand our approach to process consecutive images concurrently and impose temporal consistency, (i.e.) implies correcting ground-truth densities to also account for perspective distortions and be able to properly reason out in terms of ground-plane densities instead of image-plane densities.

REFERENCES

1. Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence, 34(4):743–761, 2012.
2. Paul Viola and Michael J Jones. Robust real-time face detection. International journal of computer vision, 57(2):137– 154, 2004.
3. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR2005. IEEE Computer Society Conference on, volume 1, pages 886–893. IEEE, 2005.
4. Antony B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. InComputerVision,2009IEEE 12th International Conference on, pages 545–551. IEEE, 2009.
5. Victor Lempitsky and Andrew Zisserman Learning to count objects in images. In Advances in Neural Information Processing Systems, pages 1324–1332, 2010.
6. Viet-Quoc Pham, Tatsuo Kozakaya, Osamu Yamaguchi, and Ryuzo Okada. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In Computer Vision (ICCV), 2015 IEEE International Conference on, pages 3253–3261. IEEE, 2015.
7. Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 589–597, 2016.
8. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 2014, arXiv:1409.1556.
9. Yingying Zhang, 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, pages 589–597, 2016.
10. Beibei Zhan, Dorothy N Monekosso, Paolo Remagnino, Sergio A Velastin, and Li-Qun Xu. Crowd analysis: a survey. Machine Vision and Applications, 19(5-6):345–357, 2008.
11. Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. Crowded scene analysis: A survey. IEEE transactions on circuits and systems for video technology, 25(3):367–386, 2015.

BIOGRAPHY

Dr.S.J.Subhashini, working as Assistant Professor of CSE department in Velammal college of Engineering and Technology, Tamil Nadu, India. She received B.E degree in Computer Science and Engineering from Madurai Kamaraj university and M.E degree in Computer Science and Engineering from Anna University, Chennai. She has completed her Ph.D in Anna University Chennai. She has more than 15 years of experience. Her areas of interest include data mining, image processing, wireless networks and cloud computing.

C.Jothi Aruna, C.Kaja Priya, S.Sanmugapriya, N.Varshini UG Scholar, Dept. of C.S.E, studying in Velammal College of Engineering and Technology, Madurai, India