Object Detection and Classification using Deep Learning

Ashitosh Phadtare , Dhiraj Choudhary D , Prathipati Jayanth , Varun Arya
Department of Information Technology
National Institute of Technology, Surathkal, Karnataka, India

***Abstract*—The study aims to tackle the task of developing an efficient Deep Learning model to identify and classify objects as viewed in images. Computer Vision is an important area of research which presents a vast potential yet untapped.**

***Index Terms*—Deep Learning, Object Recognition, Image Analysis**

I. Introduction

Object detection is one of the Major Computer Technology, which is deeply connected to the image processing and computer vision and it interacts with detecting instances of an object such as human faces, vehicles, building, tree. The main aim of face detection algorithm is to determine whether a face is present in the image or not. Many detection problems including object detection, face detection, emotion detection, and face recognition, etc. are resolved successfully by Concept of Neural Networks. The face detection work as to detect multiple faces in an image. Task of counting and locating objects in images have been the attention of several approaches. These methods help to control and count people, vehicles count. As expected, the majority of these methods are based on the well-known object detection task, including the recent methods based on convolutional neural networks, Mask-RCNN, RetinaNet, multi-scale variants, multi-scale deep feature learning network, Gated CNN and ensembles of models.

In this paper, we are proposing a method for counting and locating objects based on (CNN) convolutional neural networks. The method is based on a density estimation map with the confidence that an object occurs in each pixel considered. Density map extraction allows a better refinement of the occurrence of objects in each pixel of the image. Our proposed method uses a feature map enhancement with a Pyramid Pooling Module (PPM) Technique that allows to incorporate global information at different scales. Consequently, the proposed method incorporates sufficient global context information for a good characterization of objects with its hierarchical context module. We hypothesize that this approach is most suitable for situations of high object density, as it detects information in each pixel with the density map and improves this learning with information provided by the Pyramid pooling Module.

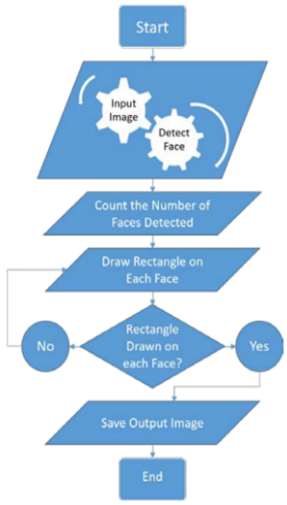


Fig. 1. Basic Flow of Our Method

II. Related Works:

We compare the performance of our model with some basic methods that are present already such as YOLO,YOLO9000,Faster R-CNN,Retina Net,VGG-GAP,VGG-GAP-HR, Deep IoU CNN.

- YOLO is proposed using a multi-scale deep feature learning network (MDFN) which will generate the abstract and semantic features from the concrete and detailed features yield at the top of the base networks. It integrates the contextual information into the learning process through a single-shot framework.The semantic and contextual information of the objects would be activated by multi-scale receptive fields on the deep feature maps. The red, yellow, blue and green components represent four different sizes of filters, which correspond to different object expressions

- VCC-GAP is proposed to provide additional contextual information to simple one-look counting models through dot annotations that takes similar annotation effort as obtaining counts alone. They generated a coarse ground truth Gaussian activation map (GAM) or saliency map from dot annotations available for counting. Next,they have incorporate the idea of back-propagating the differ-

ential error between the predicted class activation map (CAM) and ground-truth GAM alongside the counting errors with the goal to suppress false detections and enhance false negatives.

- LPN(Layout Proposal Networks) and spatial kernels are proposed to simultaneously count and localize target objects (e.g., cars) in videos recorded by the drone. Different from the conventional region proposal methods, they leverage the spatial layout information (e.g., cars often park regularly) and introduced these spatially regularized constraints into our network to improve the localization accuracy. To evaluate counting method, they present a new large-scale car parking lot dataset (CARPK) that contains around 90,000 cars captured from different parking lots.
- Deep IoU CNN, Eran Goldman with his team proposed a method designed to accurately detect objects, even in such densely packed scenes. Our method includes several innovations. We propose learning the Jaccard index with a soft Intersection over Union (Soft-IoU) of network layer. This measure provides valuable information on the quality of detection boxes. They explain how detections can be represented as a Mixture of Gaussians (MoG), reflecting their locations and their Soft-IoU scores. An Expectation-Maximization (EM) based method is then used to cluster all these Gaussians into groups, resolving the detection overlap conflicts.

III. Proposed Methodology

Our method for object detection uses a three-channel image, with w × h pixels, as input, and processes it with a CNN. The object counting and location is modelled after a 2D confidence map estimation.

The confidence map is a 2D representation of the probability of an object occurring in each pixel. The confidence map estimation was improved by including global and local information through a Pyramid Pooling Module (PPM).

Our approach for Object Detection is divided into four main phases:
(1) feature map generation with a CNN
(2) feature map enhancement with the PPM
(3) multi-sigma refinement of the confidence map
(4) object position obtention by peaks in the confidence map

A. Feature Map using CNN:

The first step is using a convolutional neural network to extract a feature map from a given input image. The feature map is used to characterize the input image and allow the confidence map estimation for the object detection task. This feature map extraction module is based on the VGG19, where the first two convolutional layers have 64 filters of a 3 × 3 size and are followed by a maximum pooling layer with a 2 × 2 window.

The last two convolutional layers have 256 filters with a 3 × 3 size. All convolutional layers use the rectified linear units (ReLU) function, with a stride of 1 and zero-padding, returning an output with the same resolution as the input.

We evaluated two variations of our method for different input images dimensions. The first variation receives an input image with 512 × 512 resolution and produces a feature map in the final layer with 64 × 64 resolution. Proportionally, the second variation receives an input image with 1024 × 1024 pixels, and the output feature map has a resolution of 128 × 128. Despite the low resolution, this map can describe quite some relevant features extracted from the image.

B. Improving feature map with PPM (Pyramid Pooling Module):

Many CNN cannot incorporate sufficient global context information to ensure a good performance in detecting and characterizing high-density objects.

To solve this issue, we adopted a global and subregional context module called PPM. This module allows CNN to be invariant to scale since it associates subregional and global information in the feature map. The PPM combines the features of four pyramid scales, with resolutions of
1 × 1, 2 × 2, 3 × 3 and 6 × 6, respectively.

The highest general level, shown in orange in the figure, applies a global max pooling which creates a 1 × 1 feature map to describe the global image context, like the number of objects detected in the input image. The other levels divide the input map into subregions, forming a grouped representation of the image with their sub-context information, as dense or sparse regions.

The levels of the Pyramid Pooling Method contain feature maps of varying sizes. Because of this, we used a 1 × 1 convolution layer with 512 filters after each level. We up sampled the feature maps to the same size as the input map with bilinear interpolation. At last, the feature maps are combined with the input map to form a much-improved description of the image. This step ensures that some minute object data is not lost in the PPM phase.

C. Multi-sigma refinement:

In this phase, the improved feature map obtained by PPM is used as input for the T stages that estimates the confidence map. The first stage receives the feature map and generates the confidence map C1 by using five convolutional layers:

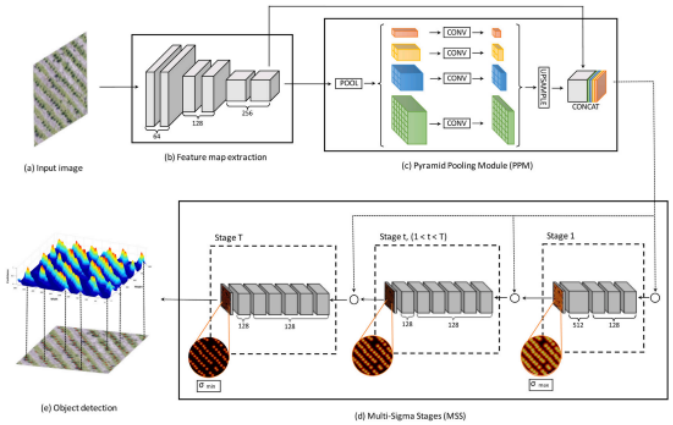
- three layers with 128 filters with a 3 × 3 size
- one layer with 512 filters with a 1 × 1 size
- one final layer with a single filter, corresponding to the confidence map.

At a subsequent stage t, the prediction returned by the previous stage Ct - 1 and the feature map from the PPM process are concatenated. They are used to produce a refined confidence map Ct. The T - 1 final stages consist of six convolutional layers:

- five layers with 128 filters with a 7 × 7 size
- one layer with 128 filters with a 1 × 1 size

The last of the layers have a sigmoid activation function so that each pixel represents the probability of the occurrence of an object (values between [0, 1]). The remaining layers consist of a ReLU activation function.

Through the multiple stages, we proposed hierarchical learning of the center of the object. The first stage roughly predicts the position, while the other stages refine this prediction.



D. Generation of confidence maps:

As mentioned earlier, to train our method, a confidence map C_t is generated as a ground truth for each stage t by using the centre of the objects as annotations in the image. The C_t is generated by placing a 2D Gaussian kernel at each centre of the labelled objects. The Gaussian kernel has a standard deviation (σ_t) that controls the spread of the confidence map peak.

We use different values of σ_t for each stage t to refine the object centre prediction during each stage. The σ_1 of the first stage is set to a maximum value σ_{max} while the σ_T of the last stage is set to a minimum value σ_{min}. The σ_t for each intermediate stage is equally spaced between [σ_{max}, σ_{min}]. The early stages should return a rough prediction of the centre of the objects, and this prediction is refined in the subsequent stages.

E. Object localization from confidence map:

Object locations are obtained from the confidence map of the last stage (C_t). Then, the estimation of the peaks (local maximum) of the confidence map is carried out by analysing the 4-pixel neighbourhood of each given location of p. Thus, p = (x_p, y_p) is a local maximum if
C_t(p) $>$ C_t(ν) for all the neighbours ν, where ν is given by (x_p ± 1, y_p) or (x_p, y_p ± 1).

IV. Experiments

A. Dataset Exploration

To put the robustness of our method to the test, we pitted it against several challenging dataset of images people in a mall and people at a park. We are using this image dataset because there are different human population densities, ranging from crowded cases to more spread out cases shown in the figure. This variation in density challenges the model for counting and locating objects. The people were also of different age groups and in different positions. This gives us the freedom to evaluate the proposed method with The images which are from The mall dataset which was collected from a publicly accessible webcam for crowd counting and profiling research. Over 60,000 pedestrians were labelled in 2000 video frames. We watermarked the data systematically by labelling the head position of every pedestrian in all frames. With Video length: 2000 frames ,Frame size: 640x480 and Frame rate: $<$ 2 Hz. The images To test the robustness of the proposed approach, we also pitted our method against two well-known image datasets which are USCD pedestrian database and cell counting from [X] by Victor Lempitsky and Andrew Zisserman. We compare the statistics of the prediction metrics against revolutionary image detetction and counting methods such One-Look Regression (Mund-henk et al., 2016), IEP Counting (Stahl et al., 2019), YOLO (Redmon Farhadi, 2017), YOLO9000 (Redmon Farhadi, 2017), Faster R-CNN (Ren et al., 2017), RetinaNet (Hsieh et al.,2017; Lin et al., 2020),LPN (Hsieh et al., 2017), VGG-GAP (Aich Stavness, 2018), VGG-GAP-HR

B. Experimental Analysis

Statistical metrics used are the mean absolute error (MAE) (Chai Draxler, 2014; Wackerly et al., 2014), root mean squared error (RMSE) (Chai Draxler, 2014; Wackerly et al., 2014), the coefficient of determination (R2) (Draper Smith, 1998), the Precision, Recall, and the F-Measure, were used to measure the performance all these are attributes of confusion matrix of the training and testing data. In the model crucial parameters such as Batch size which is 8 and number of epochs are 5 due to technical limitations .Model is Trained and tested on a device with Intel(R) Core(i5) CPU E3-1270@3.80 GHz, 256 GB memory, and a NVIDIA GeoForce GTX 1650 Graphics Card. The methods were implemented using Pytorch on the Windows operating system.

V. Result Disquisition

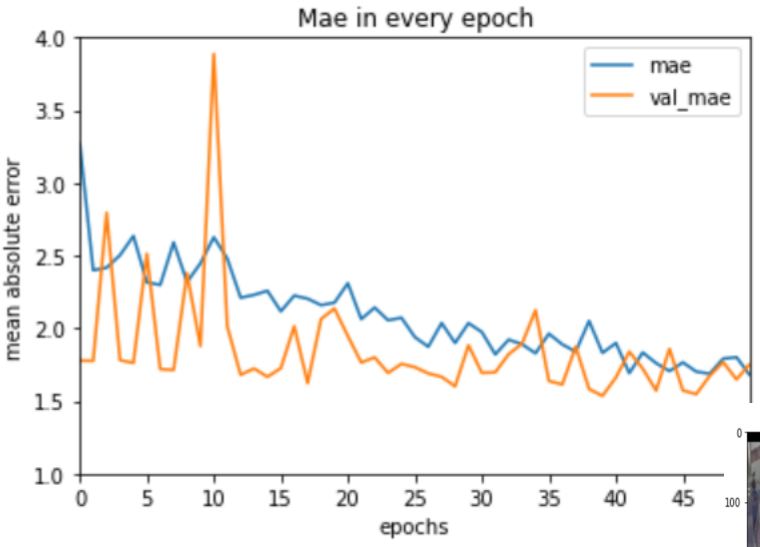
Showcasing the performance of our model against other well known models which have similar application.

we demonstrate the influence of different parameters, which includes number of epochs and finally comparing the results with a baseline of the proposed method. For this, we used the mall dataset and the car counting datasets (CARPK and PUCPR+)

here is a graph that represents MAE vs Number of epochs

Here we compare the MAE of the three data sets which have been used to train our model.

And lastly our model performance on other datasets which consists of images of cars which are PUCPR+ (Hsieh et al., 2017) is a subset of the PUCPR dataset 1u (de Almeida et al., 2015), and it is composed of 100 training images and the CARPK dataset (Hsieh et al., 2017) is composed of about 1000 training images (42,274 cars) and about 450 test images (47,500) cars.



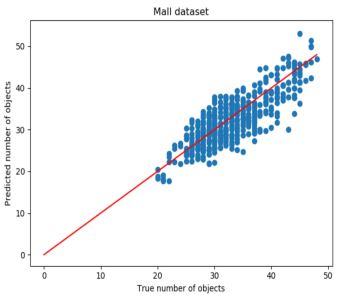


Fig. 2. MALL Dataset

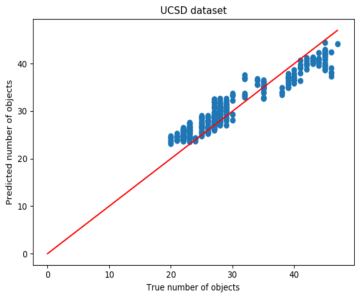


Fig. 3. UCSD Dataset

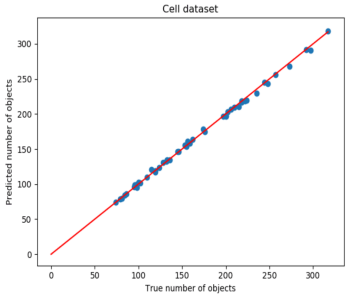


Fig. 4. CELL Dataset

Fig. 5. Side by side comparision of the data sets

Dataset	Min. #objects	Max. #objects	MAE
Fluorescent cells	74	317	1.89
UCSD	20	47	2.27
Mall	20	48	2.86

Fig. 6. Comparitively lower Density

Fig. 7. Comparitively Higher Density

PUCPR+ comparative results.

Method	MAE	RMSE	R^2	Precision	Recall	F-Measure
YOLO	156.00	200.42	–	–	–	–
YOLO9000	130.40	172.46	–	–	–	–
Faster R-CNN	39.88	47.67	–	–	–	–
RetinaNet	24.58	33.12	–	–	–	–
One-Look Regression	21.88	36.73	–	–	–	–
IEP Counting	15.17	–	–	–	–	–
VGG-GAP	8.24	11.38	–	–	–	–
LPN	8.04	12.06	–	–	–	–
Deep IoU CNN	7.16	12.00	–	–	–	–
VGG-GAP-HR	5.24	6.67	–	–	–	–
GAnet	3.28	4.96	–	–	–	–
Crowd-SDNet	3.20	4.83	–	–	–	–
Proposed Method	3.16	4.39	0.999	0.832	0.829	0.830

CARPK comparative results.

Method	MAE	RMSE	R^2	Precision	Recall	F-Measure
One-Look Regression	59.46	66.84	–	–	–	–
IEP Counting	51.83	–	–	–	–	–
YOLO	48.89	57.55	–	–	–	–
YOLO9000	45.36	52.02	–	–	–	–
Faster R-CNN	24.32	37.62	–	–	–	–
RetinaNet	16.62	22.30	–	–	–	–
LPN	13.72	21.77	–	–	–	–
VGG-GAP	10.33	12.89	–	–	–	–
VGG-GAP-HR	7.88	9.30	–	–	–	–
Deep IoU CNN	6.77	8.52	–	–	–	–
GSP	5.46	8.09	–	–	–	–
Crowd-SDNet	4.95	7.09	–	–	–	–
GAnet	4.61	6.55	–	–	–	–
Proposed Method	4.45	6.18	0.975	0.767	0.765	0.763

VI. Applications

Object detection is a computer vision technique for locating instances of semantic objects in digital images or videos. Applications of object detection include:

- Surveillance systems: Object detection can be used to monitor and analyse live video feeds from surveillance cameras.
- Self-driving cars: Object detection is used to identify and track objects such as other vehicles, pedestrians, and traffic signs.
- Robotics: Object detection can be used to enable robots to navigate and interact with their environment.
- Image retrieval: Object detection can be used to retrieve images from a large database that contain a specific object or attribute.
- Augmented reality: Object detection can be used to overlay digital information on top of real-world images.
- Industrial automation: Object detection can be used to automate inspection and quality control processes in manufacturing and assembly line.
- Medical imaging: Object detection can be used to assist in the diagnosis and treatment of medical conditions by identifying specific structures in medical images like X-Rays, MRIs, CT scans, etc.
- Agriculture: Object detection can be used to monitor crop growth and detect pests or diseases.
- Retail and marketing: Object detection can be used to track customer behaviour and improve in-store marketing strategies.

VII. Future Scope

The future scope of object detection is quite broad, as the technology is expected to continue to advance and be applied in new areas. Some potential future developments include:

- Improved accuracy and speed: Object detection algorithms are likely to become even more accurate and faster, making them more useful in real-time applications such as self-driving cars.
- Greater robustness in challenging environments: Object detection systems will become more robust in dealing with challenging environments such as low light, rain, and snow.
- More widespread adoption in industry: Object detection is likely to become more widely adopted in industries such as retail, transportation, and security, as companies look to improve efficiency and automate processes.
- Integration with other technologies: Object detection is likely to be integrated with other technologies such as augmented reality, edge computing, and 5G networks to provide new and more powerful capabilities.
- Increased use in autonomous systems: The use of object detection technology will continue to increase in various autonomous systems such as drones, robots, and self-driving cars.
- More realistic and complex scenes: Object detection models will become more versatile and efficient in dealing with realistic, complex scenes with multiple objects, varying lighting conditions, and cluttered backgrounds.
- Increased use of deep learning: Object detection will continue to leverage deep learning techniques to improve accuracy and efficiency.

VIII. Conclusions

In this study, we proposed a new method based on a CNN which returned good performance for counting and locating objects with a high-density in images. The proposed approach is based on a density estimation map with the confidence that an object occurs in each and every pixel. For this, our approach produces a feature map generated by a CNN, and then apply an enhancement with the PPM module. To improve the predictions of each object, it uses a multi-sigma refinement process, and the object position is calculated from the peaks of the respective refined confidence maps. Experiments were performed in three datasets with images containing eucalyptus trees and cars. Despite the challenges, the proposed method obtained better results than the previous methods. Experimental results on the CARPK and PUCPR+ indicate that the proposed method improves MAE, e.g., from 6.77 to 4.45 on CARPK and 5.24 to 3.16 on database PUCPR+.

Since this is the first object counting and locating CNN method based on a feature map enhancement and a multi-sigma refinement of a confidence map, other types of objects detection approaches. Further research could be focused on investigating the impact on object counting for different choices of distributions (other than Gaussian) used to generate the confidence map. Predictions other than the confidence map can also help in separating objects in high density, such as predicting the boundaries obtained from the Voronoi's diagram.

IX. References

- Aich, S., Stavness, I. (2018). Improving object counting with heatmap regulation. arXiv:1803.05494. Aich, S., Stavness, I. (2019). Global sum pooling: A generalization trick for object counting with small datasets of large images. In Proceedings of the IEEE/CVF conference on

computer vision and pattern recognition (CVPR) workshops

- Hsieh, M., Lin, Y., Hsu, W. H. (2017). Drone-based object counting by Spatially regularized regional proposal network. In 2017 IEEE international conference on computer vision (pp. 4165–4173). http://dx.doi.org/10.1109/ICCV.2017.446.
- Goldman, E., Herzig, R., Eisenschtat, A., Goldberger, J., Hassner, T. (2019). Precise detection in densely packed scenes. In IEEE conf. on computer vision and pattern recognition (pp. 5227–5236). arXiv:1904.00853. He, K., Gkioxari, G., Dollár, P., Girshick, R. (2020). Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 386–397.
- From Semi-Supervised to Transfer Counting of Crowds C. C. Loy, S. Gong, and T. Xiang in Proceedings of IEEE International Conference on Computer Vision, pp. 2256-2263, 2013 (ICCV) PDF Poster Project Page
- Cumulative Attribute Space for Age and Crowd Density Estimation K. Chen, S. Gong, T. Xiang, and C. C. Loy in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2467-2474, 2013 (CVPR, Oral) PDF Poster Project Page
- Crowd Counting and Profiling: Methodology and Evaluation C. C. Loy, K. Chen, S. Gong, T. Xiang in S. Ali, K. Nishino, D. Manocha, and M. Shah (Eds.), Modeling, Simulation and Visual Analysis of Crowds, Springer, vol. 11, pp. 347-382, 2013 DOI PDF
- Feature Mining for Localised Crowd Counting K. Chen, C. C. Loy, S. Gong, and T. Xiang British Machine Vision Conference, 2012 (BMVC) PDF Extended Abstract Poster Project Page