Instance Segmentation on Real time Object Detection using Mask R-CNN

Ravalisri.Vasam, Padmalaya Nayak

***Abstract:** In the ever-advancing field of computer vision, image processing plays a prominent role. We can extend the applications of Image processing into solving real-world problems like substantially decreasing Human interaction over the art of driving. In the process of achieving this task, we face several challenges like Segmentation and Detection of objects. The proposed thesis overcomes the challenges effectively by introducing Instance segmentation and Binary masks along with Keras and Tensorflow. Instance segmentation is used to delineate and detect every unique object of interest according to their pixel characteristics in an image. Mask RCNN is the superior model over the existing CNN models and yields accurate detection of objects more efficiently. Unlike conventional Neural Networks which employs selective search algorithm to identify object of interest, Mask RCNN employs Regional Proposal Networks(RPN) to identify object of interest. For better results Image pre-processing techniques and morphological transformations are employed to reduce the noise and increase pixel clarity.*

***Keywords:** Computer vision, Object detection, Instance segmentation, RCNN, Regional Proposal Network.*

I. INTRODUCTION

In modern technology, Image segmentation contributes a major role in Computer Vision. Image segmentation is described as, segmenting into set of pixels or multiple significant regions as per specific application. The major intention of segmentation is for easy analysis by reducing information complexity and it is additionally useful in compressing the images. The segmentation is performed using techniques of Deep learning. It is an advanced branch of machine learning algorithms which parse data, and make use of it to learn and apply that structured/unstructured data in informed decisions from what we have learned. In Deep learning, it creates structured algorithms in layers known as “artificial neural network” which extrapolates an optimal decision on its own from data that can learn. Image segmentation is applied in different fields such as autonomous driving [1], medical imaging [2], satellite imaging, human machine interaction, industrial inspection, military, biometrics image retrieval, extrapolating the features and identifying the objects of interest from the image [3].

Classification and detection are the main image level tasks. Classification is described as categorizing each image to be identical whereas detection is refereed to localizing and recognizing an object. Segmentation and Detection are combinedly implemented in instance segmentation. In this segmentation object of interests are identified and segmented for every known object within an image are segmentations are instance-aware[4]. In CNN, multilayer perceptrons usually refer to fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The "fully-connectedness" of these networks makes them prone to overfitting data. CNNs take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. Therefore, on the scale of connectedness and complexity, CNNs are on the lower extreme. Deep learning techniques have achieved state-of-the-art results for object detection, such as on standard benchmark datasets[13] and in computer vision competitions. Most notably is the R-CNN,(Region-Based Convolutional Neural Networks), as along with the proposal definition , Fast Region with CNN (Fast R-CNN), Faster Region with CNN (Faster R-CNN)[11] and Mask R-CNN have been proposed. Mask R-CNN is a conceptually simple, flexible, and general framework for object instance segmentation. Our approach efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. The most recent technique called Mask R-CNN that is capable of achieving state-of-the-art results on a range of object detection tasks.

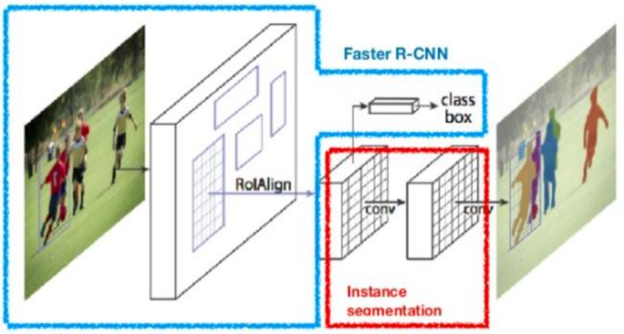


Fig A. Framework of Mask R-CNN for Instance Segmentation.

II. RELATED WORK

In recent years, there has been a continuous research going on CNN. Mostly, the R-CNN, and the extensions of it namely Fast R-CNN and Faster R-CNN and overcomes the issues of the previous method.

M Loknath eta.al proposed an algorithm that uses Fast R-CNN & RPN [5] for detection. ROI is given as input to RCNN network where the regional proposals are provided by RPN,further combined to form a single network[14] which detects a specific object

Revised Manuscript Received on October 15, 2019

* Correspondence Author
Ravalisri Vasam, CSE Department, GRIET, Hyderabad, India. Email: ravalisri.v@gmail.com
Padmalaya Nayak, CSE Department, GRIET, Hyderabad, India. Email:padma_nayak@yahoo.com

Retrieval Number: A1265109119/2019©BEIESP
DOI: 10.35940/ijeat.A1265.109119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

on sharing Convolutional features. No need of getting an ROI, as we use a Unified network that makes process cost free. With increasing number of regional proposals we compared the accuracies of trained VGGNet with two different datasets. MS COCO[11] and PASCAL VOC 2012 on a low cost GPU. At saturation point, the increment in mAP value upto 2000 proposals is viewed with increment in number of proposals. Our outcomes are contrasted and the condition of the algorithm calculation with an expansion of 1.2% as far as mAP for 1800 proposals.

Shaqing Ren et.al proposed an algorithm that uses Fast R-CNN and RPN for detection. Detection networks running time have been decreased by, Advances like SPPnet [6] and Fast R-CNN exposing region proposal calculation as a bottleneck. Introducing a detection network, which offers full image convolution features by presenting an RPN, where the regional proposals are enabled for cost free. Here, Objectness scores and object bounds are predicted by the RPN, a fully convolutional network at each position. Fast R-CNN utilizes the end-to-end trained RPNs to create regional proposals with high-quality for detection.With a basic substituting enhancement, convolutional features can be shared by trained RPN and Fast R-CNN. For the profound VGG-16 model [7], a frame rate of 5fps on a GPU is found on our detection framework while accomplishing accuracy of 73.2% mAP on PASCAL VOC 2007 on state-of-the-art object detection and 2012 (70.4% mAP) utilizing 300 recommendations for every image[4].

Ross Girshick proposed a training algorithm of single stage which mutually figures out how to refine their spatial areas and characterizing object proposals. Object detectors of ConvNet-based state-of-the-art training process is streamlined. A Detection network (VGG16 [20]) 9× faster than R-CNN [8] and 3×faster than SPPnet [9] can be trained by resulting method.At runtime a mAP of 66% of top accuracy on PASCAL VOC 2012 is accomplished while the images in 0.3s of detection network is processed at runtime.

Jaun Du has proposed one of the best representatives of CNN, You only look at once(YOLO) [10], where the custom of CNN family is broken and improved totally a better approach for illuminating the detection of objects in a simple and more efficient way. At the point when CNN arrangement creates Faster R-CNN, the mAP has arrived at 76.4, though, the frame rate (FPS) of it is 5 to 18 that is less than the ongoing impact. Accordingly, the most earnest necessity of item discovery improvement is to quicken the speed. Here, the YOLO had quickest speed and accomplished the energizing unmatched outcome[11] with FPS 155, and its mAP can likewise reach up to 78.6, the two of which have outperformed the exhibition of Faster R-CNN incredibly. YOLO V1 and V2 are the two versions of YOLO and the YOLO's layers, characteristics and algorithms are introduced.YOLOv2 accomplishes a superb the tradeoff among speed and precision just as an object detector with solid speculation capacity to view to the entire image.

III. PROPOSED WORK

In this paper, we propose a technique called Mask R-CNN a state-of-the-art framework for image segmentation that reduces the computational time and segments on variable size images [12]. Mask R-CNN is an extension of Faster R-CNN by including a branch for anticipating a mask of an object[17] in parallel with the current branch for identifying the bounding box[8]. Mask R-CNN is thoughtfully straightforward: Faster R-CNN [16] has two yields for every competitor object, a class name, and an offset of a bounding box; to this, we include a third branch that yields mask of an object [13]. It is an intuitive and natural thought. But the extra mask result is different from the output of boxes and class, which requires a lot of better spatial design extraction of an object. Next, we present the key components of Mask R-CNN, including arrangement in pixel-to-pixel [14], where it is the fundamental missing bit of Fast/Faster R-CNN. It is easy to prepare and adds just a little overhead to faster R-CNN, which runs at 5 fps. Besides, for other tasks Mask R-CNN is generalized easily, to sum up to different undertakings, e.g., enabling us to predict human postures in a similar framework.We show top outcomes in each of the three tracks of the COCO suite [17] of difficulties, including segmentation of instances [15], detecting the bounding box of an object, and individual keypoint identification. Without whistles and bells, Mask R-CNN beats all current, single-model sections

on each task, including the COCO 2016 test victors.

IV. IMPLEMENTATION

The implementation procedure is followed by different steps to get a noise free and pixel clarity image that helps an autonomous car to drive safely.

A. Image Processing:

Initially, an input image given is converted to grayscale[7] form that undergoes morphological transformations.

Here, transformations depends upon the shape of the images,typically performed on binary images[13]. We need two kinds of inputs, an original image and a kernel or organizing component that chooses the idea of activity. Two fundamental operations of morphological are Erosion and Dilation. One of its variation structures named as Opening comes into play.

Erosion: It erodes away the foreground object boundaries. Through the image the kernel slides. In an original image the pixel value (either 1 or 0) will be viewed as 1 in particular if every one of the pixels under the portion is 1, else, it is dissolved i.e., 0. In an image, foreground object thickness diminishes or just white area diminishes. It is valuable for evacuating little repetitive noises, confine two associated objects and so forth.

The function is: **cv.erode()**

Dilation: Opposite to Erosion. If atleast one of the elements under the kernel is ‘1’ ,then the value of pixel is ‘1’ which increases the white region.

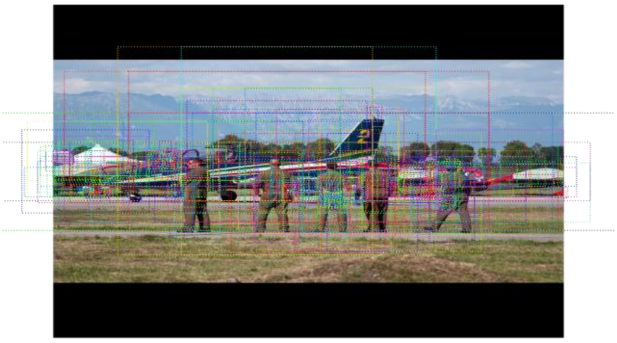
The function is: **cv.dilate()**

Opening: A form of erosion followed by dilation for noise removal. The function is: **cv.morphologyEx()**

Like the ConvNet which we use in Faster R-CNN for feature map extraction from image. In Mask R-CNN it utilizes the ResNet 101 design for features extraction from the images and that features act as next layer input.

B.Anchor Sorting and Filtering:

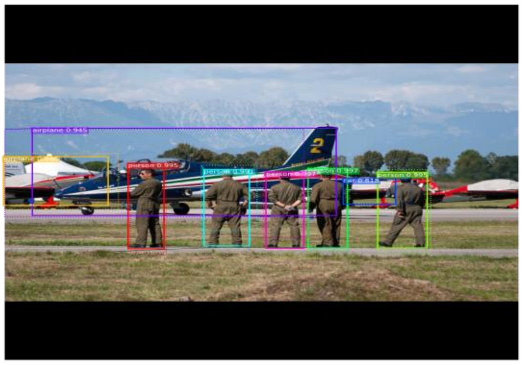
The feature maps collected from the earlier step are taken and a Regional Proposal Network applied(RPN)[16]. This fundamentally predicts the availability of an object in the region (or not). In this progression, we use feature maps or regions which predict the presence of some object by the model. The acquired regions from the RPN may be of various shapes, isn't that so? Subsequently, a pooling layer Region of Interest(RoI) is applied and every one of the areas is converted to a similar shape.



C.Bounding Box Refinement:

The obtained similar shaped regions are passed through a fully connected network such that the class labels and bounding boxes are predicted[18]. A case of conclusive detection boxes (dabbed lines) and the refinement concerned them (strong lines) in the subsequent stage.

Detections after NMS



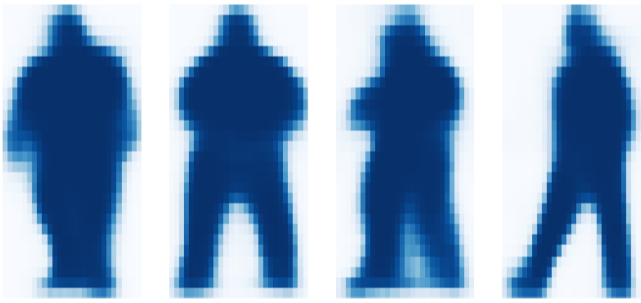
We first figure the RoI[11] so that the calculation time can be decreased. For all the anticipated areas, we figure the Intersection over Union (IoU) with the ground truth boxes. We can PC IoU like this.

IoU = Area of the intersection / Area of the union

Presently, just if the IoU is more than or equivalent to 0.5, we think about that as RoI. Else, we disregard that specific region. We compute this for every one of the regions and afterward select just a lot of regions of value more than 0.5.

D.Segmentation Mask:

When we have the RoIs dependent on the IoU values, the mask branch can be added to the current design[14]. Thus, a segmentation mask is returned for each object contained in a region. It restores a cover of size 28 X 28 for every region that scaled up for interface. The example of generated masks.



V.RESULTS

There are the results of Instance Segmentation on real time object detection of Self Driving cars using Mask R-CNN.

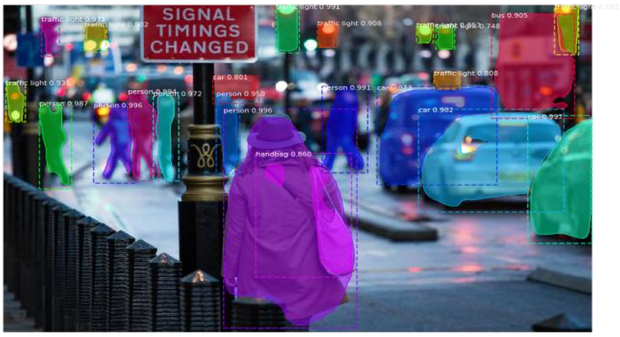


Fig B. Application Of Mask R-CNN for self driving cars.

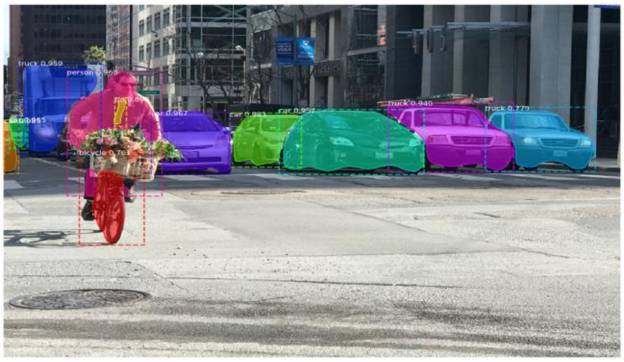


Fig C. Instance segmentation with Mask R-CNN.

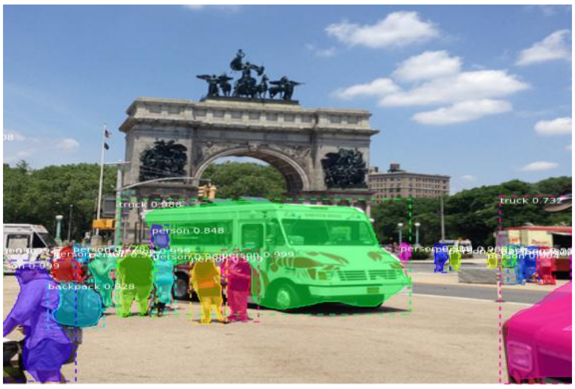


Fig D. Detection of objects using Mask R-CNN

VI. CONCLUSION

In this paper, we described by using the Mask RCNN for Instance segmentation the learning rate is considerately low compared to the semantic segmentation. Also we can resize the image to our ease as we have our own set of convolutional neural network to handle the input images. The instance segmentation can be used in various fields and technologies like Real time application of self driving cars were the cars can identify the presence of objects in front of it and make a safe move so that there will be no damage occurred to the vehicle and for the other people and surrounding objects. This instance segmentation is used in various fields like for face detection , Counting the persons in real time or counting the objects from an image , Identifying features from an image like identifying cancer cells from an image , can be used in traffic footage to identify a required vehicle etc,. The implementation of instance segmentation using Mask R-CNN on real time object detection had acquired a great accuracy compared to previous techniques of R-CNN.

REFERENCES

1. A. Ess, T. Muller, H. Grabner, and L. J. Van Gool, “Segmentation- ¨ based urban traffic scene understanding.” in BMVC, vol. 1, 2009, p.
2. M. Hameed, M. Sharif, M. Raza, S. W. Haider, and M. Iqbal, “Framework for the comparison of classifiers for medical image segmentation with transform and moment based features,” Research Journal of Recent Sciences, vol. 2, no. 6, pp. 1-10, June 2013..
3. M. Yasmin, S. Mohsin, I. Irum, and M. Sharif, “Content based image retrieval by shape, color and relevance feedback,” Life Science Journal, vol. 10, 2013.
4. A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in 2012 IEEE Conference on Computer Vision and Pattern Recognition, June 2012, pp. 3354–3361.
5. Ross Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik, 2016 Region-Based Convolutional Networks for Accurate Object Detection and Segmentation IEEE Transactions On Pattern Analysis And Machine Intelligence 38 1 142-58.
6. D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In CVPR, 2014.
7. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv:1408.5093, 2014.
8. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
9. Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler. segDeepM: Exploiting segmentation and context in deep neural networks for object detection. In CVPR, 2015.
10. Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 806-813).
11. Redmon, J., & Farhadi, A. (2016). YOLO9000: better, faster, stronger. arXiv preprint arXiv:1612.08242.
12. Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In CVPR, 2017.
13. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In CVPR, 2017.
14. C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning hierarchical features for scene labeling,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1915–1929, 2013.
15. J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In CVPR, 2017.
16. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.
17. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ra-manan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Com-mon objects in context. In ECCV, 2014.
18. G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In CVPR, 2017.

AUTHORS PROFILE

First Author Ravalisri Vasam completed her B.Tech in Computer Science and Engineering. She is pursuing M.Tech in Department of Computer Science and Engineering at Gokaraju Rangaraju Institute of Engineering and Technology (GRIET), Hyderabad, Telangana, India.

Second Author Padmalaya Nayak received her B.Tech degree in Electronics and Tele Communication Engineering from IETE, New Delhi, in 1997. She received her M.Tech degree in Computer Science and Engineering from the University of Madras in 2002 and Doctoral degree in Computer Science Engineering from National Institute of Technology, Tiruchirappalli, India in 2010.