A Review on Techniques for Multiple Object Recognition from a Scene

Swetha O[1] C. Ramachandran[2]
[1]M. Tech Student [2]Associate Professor
[1,2]Department of Electronics & Communication Engineering
[1,2]College of Engineering Thalassery, India

Abstract— Object recognition is a computer vision technique for identifying objects in an image or video. Object recognition is a key output of deep learning and machine learning algorithm. Humans can recognize any object in the real world easily without any effort and hence are able to do their work efficiently. Similarly if robots too become efficient in understanding their environment it will be able to perform many complex tasks better. Some of those tasks include object tracking, industrial inspection, bio imaging, robotic vision etc. One of the trending applications of them all is in the intelligent vehicle system. Object recognition helps in recognizing objects or obstacles in the path of a robot, sign boards and people in case of a driverless car, abnormalities during any kind of inspection etc. Histograms of Oriented Gradients (HOG), CNN, RCNN, Fast R-CNN, Faster R-CNN, FCN, Mask R-CNN, YOLO, and YOLO9000 are some of the techniques discussed here.

Keywords: Object Recognition, HOG, CNN

I. INTRODUCTION

Object recognition is one of the most popular research topics in computer vision field. There are various applications object recognition projects, such as robotic navigation, video analysis for security purpose, autonomous driving, and structural visual inspection. Object recognition is able to provide valuable information for semantic understanding of images and videos and so also helps in image classification, human identification, face recognition. It's an efficient method by which machines can understand their surroundings which may include vehicles, people, traffic lights or other obstacles in the path etc. Recently based on the developments and implementation of deep learning technique, object recognition related frameworks have achieved significant improvements.

Just a CNN (convolution neural network) based frameworks for solving physical problems has been going down gradually. Several improvements made to CNN based object detector are Region based Convolution Neural Network (R-CNN) [3], Fast R-CNN [4], Faster R-CNN [5], Fully Convolution Network (FCN) [9], Mask R-CNN [10], You Only Look Once (YOLO) [6], YOLOv2 [7].Semantic segmentation is one of the high-level tasks that help in complete scene understanding. One most successful state-of-the-art deep learning techniques for semantic segmentation stem from a common forerunner: the Fully Convolution Network (FCN) by Long et al. [9].Mask RCNN [10] performs instance segmentation which requires detecting all objects in an image and segmenting each instance.

II. DIFFERENT METHODOLOGIES

There had been a great deal of improvements in the field of object recognition. Various methodologies which improved the efficiency over time are explained below.

Feature extraction methods are used in the object recognition and classification procedures. In [1], improved Histograms of Oriented Gradients features are used to represent the edge information of images. HOG is adopted as the basic feature descriptor. It has high discriminative power. In video surveillance system moving objects can be segmented from the background. For real time human tracking applications, extract the front-image from each frame of video, and employ HOG descriptor to mark the human in the sub image.

Convolution Layers:- It is the first layer. Characteristic maps are formed here. CNN manipulates complex components to convolve the complete data as well as the fundamental component maps.

Pooling Layers:- Spatial resolution of the data is reduced, for the subsequent convolution layer so that amount of parameters and computation can be reduced. This operation does not affect the intensity dimension of the input. Also called Sub sampling

Fully Connected Layers:- These are the last few layers which compiles the data extracted by previous layers to form the final output. This layer provides us with a category prediction.

In [2] by S. Hayat, CNN uses a stochastic gradient decent to update weighting filter and coupling coefficient. Pooling and convolution operations are done. An activation function, Rectified Linear Unit (ReLU) is used for category recognition. Comparison of the performance of the proposed model with BOW (Bag of Words) and HOG-BOW approaches based on linear L2-SVM classifier is done. CNN method is statistically evaluated in terms of accuracy, loss and time efficiency.

In [8] Inception v3 of ImageNet is used as image recognition model. For each filter in CNN, Rectified Linear Unit (ReLU), Max Pooling and fully connected layers procedures will be implemented on the input image. The gradient descent procedure renews the values of the weight for improving accuracy after the fully connected layer in the network. Back-propagation of CNN is applied.A more significant gain is obtained with the introduction off regions with convolution neural network (R-CNN) [3].It is a deep neural network (DNN) technique. This method improves thee quality of bounding boxes and extracts high-level features.

R-CNN was proposed by Girshick et al. [3] and obtained a mean average precision (mAP) of 53.3% with more than 30% improvement over the previous best result on PASCAL VOC 2012. Region Proposal Generation means R-CNN adopts selective search to generate about 2000 region proposals for each image, then label categories and bounding box for the image. Then CNN module is utilized to extract a 4096-dimensional feature as the final representation. A pre-trained category- specific linear SVMs are used for object classification. Different region proposals are scored on a set of positive regions and background (negative) regions.

After proposal of R-CNN, an improved model called Fast R-CNN [4] is introduced that jointly optimizes classification and bounding box regression tasks. A fixed length feature vector is extracted from each region proposal with an ROI. Each feature vector is then fed into a sequence of fully convoluted layers and then divides into two output layers. One output layer is for softmax probabilities and the other output layer is for bounding-box positions. In the Fast R-CNN [4], regardless of region proposal generation, the training of all network layers can be processed in a single stage with a multitask loss. So it requires less storage space and improves both accuracy and efficiency. It is repeated multiple times for each region of interest. This method is faster but requires a set of candidate regions to be proposed along with each input image.

Faster RCNN is the modified version of Fast RCNN. The major difference between them is that Fast RCNN uses selective search for generating Regions of Interest, while Faster RCNN uses Region Proposal Network (RPN). RPN takes image feature maps as an input and generates a set of anchor boxes, each with an object score as output.

ROI pooling, in Faster R-CNN [5], causes misalignment between the ROI and the features. It affects classification little because it withstands small translations. However, it has a large negative effect on pixel to-pixel mask prediction. To solve this problem, Mask R-CNN [10] adopts a simple and quantization-free layer, namely, ROI Align, to preserve the explicit per-pixel spatial correspondence. ROI Align is obtained by replacing the quantization of ROI pooling with bilinear interpolation, computing the exact values of the input features at four regularly sampled locations in each ROI bin. This minor change improves mask accuracy greatly, especially under strict localization metrics. Mask R-CNN performs instance segmentation and object detection results. Mask R-CNN is a flexible and efficient framework for instance-level recognition, which can be easily generalized to other tasks with little changes.

A state-of-the-art deep learning techniques for semantic segmentation originates from a Fully Convolution Network (FCN) by Long et al. [9].Mask R-CNN [10] is performed by combining Faster R-CNN[5] and FCN[9]. In FCN[9], the well-known classification models – AlexNet , VGG (16-layer net) , GoogLeNet and ResNet are transformed into fully convolutional ones by replacing the fully connected layers with convolutional ones to output spatial maps instead of classification scores. Next stage pixel wise prediction and up sampling is done to maintain dimension. Spacial details are improved by fusing information from layers with different strides. In the staged version of training, there is a single-stream FCN-32s, which gets upgraded to the two-stream FCN-16s and continue learning, and then finally gets upgraded to the three-stream FCN-8s and finish learning. The learning rate is dropped 100 times from FCN-32s to FCN-16s and 100 times more from FCN-16s to FCN-8s, which is very essential.

YOLO or You Only Look Once [6] is an object detection algorithm much different from the region based algorithms. In YOLO a single convolution network predicts the bounding boxes and the class probabilities for these boxes. The images are split it into an S x S grid, within each of the grid we take m bounding boxes. For each of the bounding box, the network outputs a class probability and offset values for the bounding box. The bounding boxes having the class probability above a threshold value is selected and used to locate the object within the image. The YOLO consists of 24 convolution layers and 2 FC layers, of which some convolution layers construct ensembles of inception modules with 1×1 reduction layers followed by 3×3 convolution layers. The network can process images in real time at 45 fps, but has some localisation error and is difficult to detect small objects in an image, also the recall is low. An improved version, YOLOv2, was later proposed. YOLO9000[2] is also proposed to detect over 9000 object categories obtained by combining COCO and ImageNet using WordTree. The improved model, YOLOv2, is state-of-the-art on standard detection tasks like PASCALVOC and COCO. With the multi-scale training method the same YOLOv2 model can run at varying sizes, offering an easy tradeoff between speed and accuracy. YOLO9000 simultaneously trained on the COCO detection dataset and the ImageNet classification dataset. Joint training allows YOLO9000 to predict detections for object classes that don't have labelled detection data. Darknet-19 is the backbone for YOLOv2. It has 19 convolution layers and 5 max-pooling layers. It achieves 91.2% top-5 accuracy on ImageNet which is better than VGG (90%) and YOLO network (88%).

III. Results and Comparisons

The accuracy among the discussed technique is measured in terms of Average Precision (AP). AP computes the average precision value for recall value over 0 to 1. Recall measures how good you find all the positives. Comparison of RCNN, Fast RCNN and Faster RCNN based on its speed is done below. The object recognition techniques were tested on popular dataset PASCAL VOC 2007/2012 consisting of 20 categories. For another popular dataset COCO the YOLO v2 architecture with Darknet-19 architecture provides an AP of 21.6.

The segmentation techniques like FCN [9] and MRCNN [10] were also tested on Pascal VOC and COCO dataset respectively. Performance of FCN is measured by mean intersection over union mean IU. Mean IU on Pascal VOC 2011 and 2012 datasets gave 30% improvement, ie 67.5 and 67.2 respectively. Mask RCNN with ResNetXt -101-FPN gave the best AP of 39.8.

Method	Test time per image(sec)	Speed up
RCNN	50	1x
Fast RCNN	2	25x
Faster RCNN	0.2	250x

Table 3.1: Comparison of technique of [3], [4], [5]

MODEL	Architecture	mAP	fps
Fast RCNN	CNN	70	0.5
Faster RCNN	VGG 16	73.2	7
Faster RCNN	ResNet	76.4	5
YOLO (448 x 448)	VGG-16	63.4	45
YOLO v2(544 x 544)	Darknet-19	78.6	40
YOLO v2 (416 x 416)	Darknet-19	76.8	67

Table 3.2: Results when methods were trained on Pascal VOC 2007

Method	mAP
Fast RCNN	68.4

Faster RCNN	70.4
YOLO	57.9
YOLOv2(544x544)	73.4

Table 3.3: Result when methods were trained on Pascal VOC 2012

IV. CONCLUSION

Various object recognition techniques are discussed here, with each technique being an improved version of the previous one. The papers selected for this literature review ranges from local feature extraction to neural network based methods for object recognition. The methods reviewed are Histograms of Oriented Gradients (HOG), CNN, R-CNN, Fast R-CNN, Faster R-CNN, FCN, Mask R-CNN, YOLO, YOLO9000. Comparison of above mentioned techniques are done here based on their accuracy on different datasets. Among these methods, YOLO is a novel framework which is fast and accurate and also supports real time object recognition. Predictions are made from one single network. It can be trained end-to-end to improve accuracy. YOLO is more generalized. It outperforms other methods of object recognition

REFERENCES

[1] S. Zhang and X. Wang, "Human detection and object tracking based on Histograms of Oriented Gradients," 2013 Ninth International Conference on Natural Computation (ICNC), Shenyang, 2013
[2] S. Hayat, S. Kun, Z. Tengtao, Y. Yu, T. Tu and Y. Du, "A Deep Learning Framework Using Convolutional Neural Network for Multi-Class Object Recognition," 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Chongqing, 2018
[3] R.Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014
[4] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015
[5] S. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," in Proc. NIPS, 2015
[6] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016
[7] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017
[8] M. A. Rahman, S. P. Paul, M. Das, M. M. Hossain, R. Haque and M. A. Rahman, "Convolutional Neural Networks based multi-object recognition from a RGB image," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019
[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015
[10] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017

