Analysis of Vehicle Detection using Region-based Convolutional Neural Networks (RCNN)

KMN Syed Ali Fathima[1],
[1]*Research scholar, Register Number: 18131172282016,*
Sarah Tucker College, Manonmaniam Sundaranar University

DR.K. Merriliance[2]
[2]*Assistant Professor, Department of Computer Application (MCA),*
Sarah Tucker College, Manonmaniam Sundaranar University

Abstract-Vehicle detection and its tracking can help us to avoid accidents to a greater extent. Vehicle detection will locate the presence of objects in the frame along with its position and also with the classification of the located vehicles. Researchers have shown that region-based works will give good performance for object detection. In this paper, we propose a region-based deep learning network to detect the vehicles. Region-based learning is implemented by combining Faster RCNN and Grid RCNN. **The proposed work can detect multiple vehicles in an image. The evaluation of the proposed work gives better results in terms of mAp, Ap, precision and recall when it is applied for CBCL dataset.**

Keywords: Faster RCNN, Grid RCNN, Vehicle detection

1. INTRODUCTION

Computer visualization inconvenience is an entomb remedial methodical field that manages with how computers can be finished to improve significant level perceiving from futuristic images. Computer visualization undertakings consist of schemes for gathering, progressing, evaluating, developing and deciding digital images, and extraction of high-dimensional data from the real world to produce numerical or symbolic information. Vehicle detection is a challenging and important research area of image processing. It is broadly used in computer vision.

The complete understanding of an image should not simply contemplate on categorizing dissimilar images, but also on specifical approximation the positions of substances enclosed in every representation. Object detection, one of the essential computer visualization troubles, is capable to make available important in sequence in linguistic discerning of images.

Many kinds of research were done to detect the objects. Deep learning-based object detection gives good results in detecting vehicles on real road images. But, the vehicle detection in real-time images fails in larger variations of luminosity, intense occlusion, and bulky disparity of object levels. This paper proposes a region-based deep learning technique because instead of working on the entire image, working in the region will yield a better result.

Section 2 discusses the related works done in vehicle detection. Section 3 explains the proposed work which helps to detect the vehicles. The evaluation details are analyzed in Section 4. Section 5 gives the conclusion.

2. RELATED WORK

Detection of a vehicle is one of the important problems in computer vision research. Recent advances in object detection are driven by region-based methods. Basri and Jacobs [1] have exhibited the benefit of separating nearby locale limits for detection. Edelman et al. conjectured that these perplexing neurons could take into account coordinating and detection of 3D objects from a scope of perspectives.

Ke and Sukthankar[6] applied PCA on the gradient image. The PCA-SIFT used in this work yields a 36-dimensional descriptor which is fast for matching. Reducing false-positive rates by more than an order of magnitude relative to the best Haar wavelet-based detector from [2]. Lowe et al. [4] proposed a scale-invariant feature transform (SIFT) to detect and describe local features in digital images. It locates certain key points and then furnishes them with quantitative information is called descriptors. Gaussian $g(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/2\sigma^2}$ was estimated with adequate precision using a 1D essential part. Here,$\sigma = \sqrt{2}$. Convolution with the information picture can be capably enlisted by applying two goes of the Gaussian function in the parallel and erect directions. HOG [13] is used as a feature descriptor along with linear SVM classifier as a regional image descriptor. The HOG descriptor

technique counts occurrences of gradient orientation in localized portions of an image - detection window, each entity pixel surrounding the pixel shade transforms in both x-axis and y-axis, the gradient of a continuous multi-variable function,$\nabla q(x,y) = \begin{bmatrix} r_x \\ r_y \end{bmatrix} = \begin{bmatrix} \frac{\sigma q}{\sigma x} \\ \frac{\sigma q}{\sigma y} \end{bmatrix} = \begin{bmatrix} q(x+1,y)-q(x-1),y \\ q(x,y+1)-q(x,y-1) \end{bmatrix}$, which gives the vector of limited derived of everyone the inconsistent. Assume q(x, y) traces the redden of the pixel at position (x, y), the slope vector of the pixel (x, y) is identified, $\frac{\frac{\sigma q}{\sigma x}}{\frac{\sigma q}{\sigma y}}$ -the term is the partial derivative on the x-path, which is calculated as the color variation connecting the closest pixels.

SURF descriptor [10]for image matching and object recognition, uses a numeral estimate of the determinant of Hessian blob detector, which can be calculated with 3 integer procedures using a precomputed integral image, feature descriptor is based on the sum of the Haar wavelet response around the point of interest, also be computed with the integral image. $\mathcal{H}(x,\sigma) = \begin{bmatrix} \mathcal{L}_{xx(x,\sigma)} \mathcal{L}_{xy(x,\sigma)} \\ \mathcal{L}_{xy(x,\sigma)} \mathcal{L}_{yy(x,\sigma)} \end{bmatrix} \ldots\ldots\ldots \mathcal{L}_{xx(x,\sigma)} = \frac{\sigma^2}{\sigma x^2 g(\sigma)} uses\ \mathcal{L}$-scale-space representation, then detecting scale-space maxima of this operator $\mathcal{H}(x,\sigma)$ and obtains differential blob detector with automatic scale selection. Ross et al. [19] used region-based convolutional neural networks to generate create locale proposition from a system whose last FC layer at the same time predicts different (e.g., 800) boxes, which are utilized for R-CNN, object recognition, For regression, embrace the definitions of the 4 directions, with $u_x = \frac{x-x_a}{w_a}, u_y = \frac{y-y_a}{h_a}, t_w = \log(w/w_a), u_h = \log(h/h_a)$ and $u_x^* = \frac{x^*-x_a}{w_a}, t_y^* = \frac{y^*-y_a}{w_a}, t_w^* = \log(w/w_a), t_h^* = \log(h^*/h_a)$, where x, y, w, and h denote the two coordinates of the box centre, width, and height. Variables x, x_a, and x^* are for the calculated box, anchor box, and ground-truth box respectively, as bounding-box regression from an anchor box to a nearby ground-truth box. Fast R-CNN expands on productively order object proposition utilizing profound convolutional networks, to improve preparing and testing speed while likewise expanding identification precision.

Girshick et al.[22], a fast R-CNN network was proposed which accepts a whole picture as information. The system first procedures the entire picture with a few convolutional (conv) and max-pooling layers to deliver a conv highlight map. At that point, for each article proposition, a locale of intrigue (RoI) pooling layer removes a fixed-length highlight vector from the component map. Each element vector is taken care of into a grouping of completely associated (Fc) layers that at long last branch into two kin yield layers: one that produces softmax likelihood evaluates over K object classes in addition to a catch-all "foundation" class and another layer that yields four genuine esteemed numbers for each of the K object classes. Every set of 4 qualities encodes refined jumping box position for one of the K classes. He et al. [23proposed a technique for object identification called Faster R-CNN which joins Region Proposal Network (RPN) as a Region of Interest (RoI) applicant extractor. Faster R-CNN [23] is progressively vigorous to deal with the huge variety of vehicle different scales.

Grid R-CNN [38] is object detection and identification framework structure, where the customary relapse definition is supplanted by a lattice point guided confinement instrument and the unequivocal spatial portrayals are productively used for excellent limitation. Grid R-CNN isolates the item jumping enclose locale to matrices and utilizes a fully convolutional network.

GoogleNet[28] is a deep learning structure which applied distinctive scale convolution portions (1X1; 3X3 and 5X 5) a similar component map in a given layer. This strategy is caught by utilizing multi-scale includes and summed up these highlights together as a yield highlight map.

VGG Net [27] is collected from five groups of convolutional layers and three FC layers. There are two convolutional layers in the first two groups and three convolutional layers in the next three groups. Between each group, a Max Pooling layer is applied to lessen spatial element. ResNet [29] which condensed optimization difficulty by introducing shortcut relations extract attribute embed of the complete reflection and concatenate it with area characteristic to progress recognition. DenseNet [30] which is retain the deep layer features and better in sequence flow by succession the input with the remaining output, oppressed features from thin layers were reproduction and acquired.

He et al. [31] proposed Mask R-CNN, which anticipated bouncing boxes and division covers in corresponding to create and afterwards RoIPool or RoIAlign are consumed to extort features for these proposals. The extort features are then used for supplementary proposal degeneration and classification, ROI Align layer which tended to the quantization issue by bilinear introduction at partially inspected positions inside every grid, He et al. [31] and Dai et al.[32] be trained integrated occurrence segmentation framework and optimize the detector with pixel-level supervision. Cascade R-CNN [50] prepares multi-stage R-CNNs with growing IOU threshold phase-by-phase and thus the multi-stage R-CNNs are consecutively more prevailing for exact localization.

Ren et al.[40] used to merge the Faster-RCNN reproduction with two diverse convolutional neural networks (VGG-16 and ResNet-50) [41]. RetinaNet [48] which focus to resolve the localization precision of solitary-stage detectors because of the stumpy concentration accuracy of predefined anchor and it has single-stage deterioration. RefineDet[49] is a two-step deterioration to develop the concentration precision for single-stage detectors which recovers solitary-stage detector by two-step pour lose ground.

3. REGION-BASED VEHICLE DETECTION METHODS

This section describes Faster RCNN [23] and Grid RCNN [38] which is used for object detection in the proposed work. Region Proposal Network acquired raw image information as input creates a region of interest which represents the prospect of object subsistence using a process called Selective Search. R-CNN [19] and uses Support Vector Machine (SVM) that classifies the object.

3.1 Faster RCNN

Faster RCNN [23] uses Region Proposal Network (RPN) which gets representation feature maps as an effort and produces a set of object suggestions, each with an objectless gain as output. Faster R-CNN joins the problem of selective search by replacing it with Region Proposal Network (RPN). Firstly, remove include maps from the information picture utilizing ConvNet and afterwards go those maps through an RPN which returns object an RPN which returns object proposals. Finally, these maps are grouped and the bounding boxes are anticipated. Bounding-box regression offset is given in Eqn. (1).

$$u^k = u_x^k, u_y^k, u_w^k u_h^k \qquad \ldots (1)$$

The bounding box regression gives the parameterizations of the four coordinates, where x, y, w, and h denotes the box's interior coordinate and its width and height. Variables x, x_a, and $x^*$are remove include maps from the information picture utilizing ConvNet and afterwards go those maps through an RPN which returns object recommendations. At last, these maps are ordered and the jumping boxes are anticipated. The feature extraction process of Faster R-CNN [23] is called as region proposal network (RPN) which shares convolutional layers with region-based detectors; Region proposals proposition is gotten from RPN and used for ROI element extortion from the output attribute maps of a CNN. To extort characteristics for region proposals, RPN is designed to efficiently predict region proposals with a wide range of scales and aspect ratios.

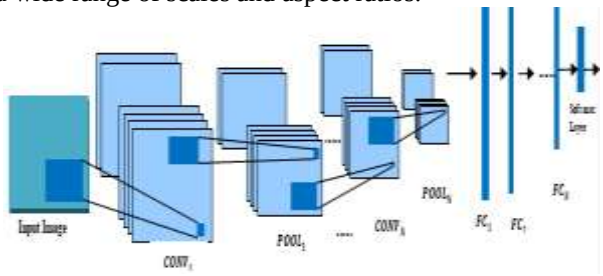

Figure 1 Architecture of Faster RCNN Detector

Figure 1. (a) Illustrates the Faster R-CNN architecture. A Faster R-CNN [23] accepts a whole representation as contribution with a lot of suggestions for their objects. . The system first methods the whole image with several convolutional (conv) and max-pooling layers to produce a conv feature a map. Then, for every point application a region of interest (ROI) collecting layer extorts a locale of intrigue (ROI) pooling layer extricates a predetermined-length include vector from the component map.

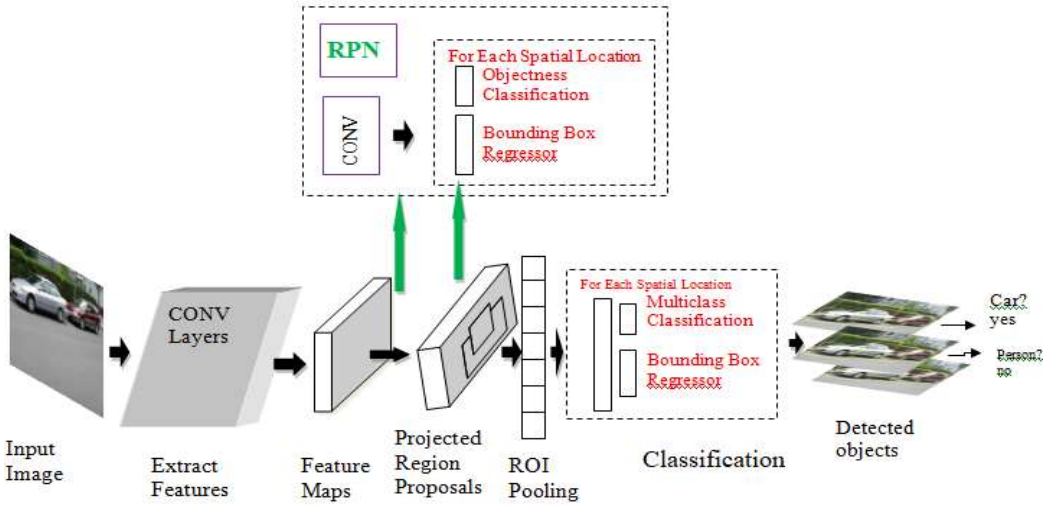

Figure 1.(b) Architecture of Faster RCNN Vehicle (car) Detector

Figure 1 (b) illustrates each element vector is taken care of into an arrangement of completely associated (fc) layers that at long last gives two yield layers: one that produces softmax likelihood appraises over K object classes in addition to a catch-all "foundation" class and another layer that yields four genuine esteemed numbers for every one of the K object classes. Each arrangement of 4 qualities encodes refined jumping box positions for one of the K classes.

3.1.1 ROI (Region of Interest) Voting

Each training with ROI is marked with a ground-truth class u and a ground-truth bounding box relapse target v, utilize a perform multiple tasks trouble L on each named ROI to together train for characterization and bounding box relapse:

$$L(p, u, t^u, v) = L_{cls}(p, u) + \lambda[u \geq 1]L_{loc}(t^u, v) \quad \ldots (2)$$

Eqn. (2), denotes the L_{loc} defined over a tuple of relapse focuses for class true bounding box regression targets for class $u, v = (v_x;\ v_y;\ v_w; v_h)$, and a predicted tuple$t_u = \left(t_x^u; t_y^u ; t_w^u;\ t_h^u\right)$, again for class u.

3.2 Grid RCNN

Another object detection system named Grid R-CNN [38] receives a lattice guided restriction component for precise item identification. Grid R-CNN [38] utilizes explicit spatial representation to estimate the location of the grid points, thus reformulating the localization as a classification problem. Grid R-CNN [38] is the first recognition structure that finds the item by anticipating network focuses on the pixel level. Matrix focuses can be gotten with top-notch confinement results. return for capital invested highlights is then used to perform arrangement and restriction. As opposed to past works with a container balance relapse branch, we receive a framework guided instrument for excellent restriction. The framework expectation branch embraces an FCN[20] to yield a likelihood heat map from which we can find the matrix focuses in the bouncing box lined up with the article. With the lattice focuses, we at long last decide the exact item bouncing box by an element map level data combination approach. For the multi-stage identifiers, the proposition is right off the bat produced and afterwards, RoIPool or RoIAlign[8] are used to extricate highlights into the objects.

The architecture of Grid RCNN object detector is given in Fig.2.

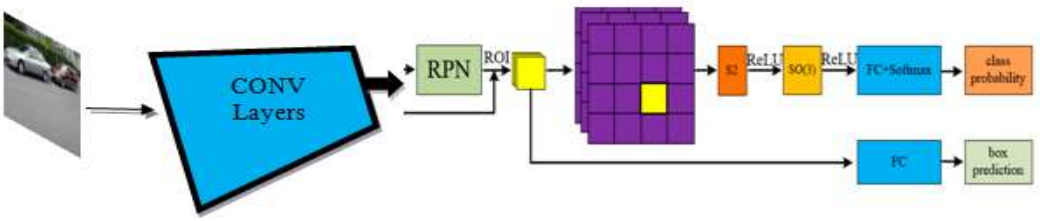

Figure 2. The architecture of Grid RCNN

In this, convolution Layers are fed into RPN to extract features. The RPN is used to provide bounding box proposals. The candidate objects in each bounding box are projected onto a grid map.

3.3 Trouncing Function for erudition province suggestion using Detectors

RPNs appoint a double class mark (of being an article or not) to each stay, dole out a positive name to two sorts of grapples: (I) the stay/stays with the most elevated Intersection over-Union (IoU) cover with a ground-truth box, or (ii) a grapple that has an IoU cover higher than 0.7 with any ground-truth box. Misfortune work for a picture in quicker rcnn indicator [23]is characterized as:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \quad \dots (3)$$

Eqn. (3) is used to reduce the Loss function for an image in Faster RCNN detector [23]. It is implemented by using the cls layer as a two-class softmax layer, i as the file of a stay in a smaller than common batch,p_i as the anticipated likelihood of grapple, ground-truth name p_i^*,t_i as a vector speaking to the 4 defined directions of the anticipated bounding box,t_i^* −as that of the ground-truth box related with a positive anchor. L_{cls} is the chronicle loss and L_{reg} *is the* regression loss.

$$\left.\begin{aligned} x_l = \frac{1}{N}\sum_{j\in E_1} x_j p_j,\ y_u = \frac{1}{N}\sum_{j\in E_2} y_j p_j \\ x_r = \frac{1}{N}\sum_{j\in E_3} x_j p_j, y_b = \frac{1}{N}\sum_{j\in E_4} y_j p_j \end{aligned}\right\} \quad \dots (4)$$

Eqn. (4) is used to reduce the Loss function for an image in Grid RCNN detector [38] by determining the four limits of the case of an item with the anticipated framework focuses. In particular, we signify the four limits arrange as, B =x_l, y_u, x_r, y_b(left, higher, right and base) frame correspondingly, predicted probability p_j, E_i as the position of directories of grid positions that are established on the i^{th} boundary.

3.3.1 Learning rate

The learning rate is a tuning boundary in an enhancement calculation, for example, SGD, that decides the progression size at every emphasis while pushing toward at least a least amount of a loss function. Learning rate and linearly (or exponentially) increase it every iteration. Training should be stopped when loss function starts to drastically increase. Record the learning rate and loss (or accuracy) at each iteration.

$$\eta_n = \eta_0 d^{floor}\left(\frac{1+n}{r}\right) \quad \dots (5)$$

Eqn. (5) gives the η as learning rate at iteration n,$\eta_0$as starting learning rate, d as to how much the learning rate should change at each drop, and r compares to the drop rate. Floor work here drops the estimation of its contribution to 0 for all qualities modest than 1.

3.3.2 Optimization SGD Momentum- Minibatch SGD:

Optimization algorithms (such as SGD, RMSprop, Adam) require setting the learning rate — the most important hyper-parameter for training deep neural networks. Be train by reducing a loss:

$$\mathrm{L(w)} = \frac{1}{|\mathrm{X}|}\sum_{\mathrm{x}\in\mathrm{X}} \mathrm{l(x,w)} \quad \dots (6)$$

In Eqn. (6), w is loads of a system, X is a marked preparing set, and l(x,w)- the misfortune processed from tests, l- the sum of a classification loss and a regularization loss on w.

$$w_{t+1} = w_t - \eta\frac{1}{n}\sum_{x\epsilon B}\nabla l(x, w_t) \quad \dots (7)$$

Eqn. (7) is a minibatch inspected from X and n=|B| is the minibatch size,η-the learning rate, t is the cycle list. Force is constrained by a hyperparameter comparable to a ball's mass which must be picked physically—excessively high and the ball will turn over minima which wish to discover

$$v_{t+1} = mv_t + \eta\frac{1}{n}\sum_{x\epsilon B}\nabla l(x, w_t) \quad \dots (8)$$

$$w_{t+1} = w_t - v_{t+1} \quad \dots (9)$$

In Eqn. (8)&(9), m force rot factor and u is the update tensor, η is the learning rate recommendations incorporate the energy strategy, which showed up in Rumelhart, Hinton and Williams' original paper on back engendering learning [16], Stochastic angle plunge with energy recalls the update Δ w at every emphasis, and decides the following update as a straight blend of the inclination. The RPN, which is normally executed as a completely convolutional arrange [20], can be prepared start to finish by back-spread back-propagation and stochastic gradient descent (SGD) [47].

4. PROPOSED WORK

The vehicle detection has mainly two steps: Localization and Classification. The image is given as the input to the framework. The object detection is done as the first step and its location is identified. The identified object is given to the classification step which identifies the vehicle.

$$\mathcal{L}(p, u, t^u, v) = \underbrace{\mathcal{L}_c(p, u)}_{\textit{Classification}} + \underbrace{\lambda[u \geq 1]\mathcal{L}_l(t^u, v)}_{\textit{Localization}} \quad \ldots (10)$$

Eqn. (10) gives the outcome after localization and classification. In the proposed work, Faster R-CNN and Grid CNN are used. Faster R-CNN is increasingly powerful to deal with the enormous assortment of vehicle levels. Faster R-CNN shows an extraordinary presentation for the Pascal VOC [24] and COCO [25] 2D object finding benchmarks. In every case, it has not executed fine and just achieves 56.39% represent normal accuracy on the KITTI [26] vehicle recognition benchmark.

Faster R-CNN fixes this by utilizing another convolutional network (RPN) to produce the locale recommendations. The area proposition time from 2s to 10ms per picture yet also permits the district proposition stage to share layers. Faster R-CNN replaces specific tracking down with a different sub-neural system to produce ROIs, making another 10x accelerate and along these lines testing at a pace of around 7–18 fps. Network CNN models the issue of article recognition as finding a way from a fixed lattice to boxes firmly encompassing the items.

Grid-CNN with around 180 boxes in a multi-scale lattice performs all around contrasted with Fast R-CNN which utilizes around 2K bounding boxes created with a proposition strategy. This procedure makes recognition quicker by evacuating the item proposition stage just as diminishing the number of boxes to be handled, train a regressor to move and scale elements of the grid towards objects iteratively. Grid R-CNN leads to high-quality object localization. Grid R-CNN embraces a lattice guided limitation instrument for exact item identification with the customary relapse based strategies. Grid R-CNN catches the spatial data unequivocally and empowers the position-touchy property of completely convolutional architecture. CBCL train dataset, run SGD for 30k mini-batch cycles, and at that point bring down the learning rate to 0.0001 and prepare for another10k cycles. Learning stops after 40,000 cycles and the boundaries of layers conv1-1 to conv2-2 are fixed during learning for quicker preparing. CBCL vehicle dataset run SGD for small scale clump emphases, and afterwards, bring down the learning rate and train for different cycles. A force and boundary rot (on loads and predispositions) are utilized.

5.Performance Analysis

Evaluation of our advanced vehicle detector method which is formed by combining Faster RCNN and Grid RCNN is given in this section. This is done by using the average precision (AP), confusion matrix, precision, recall, and IoU. Assess the performance of vehicle class because the CBCL object identification benchmark takes the exhibition of the vehicle classification as the assessment average. Several images have been used 1000 images in CBCL dataset, 85% of the dataset is used as training images and 15% of it as test images. Dataset Link: http://cbcl.mit.edu/software-datasets/streetscenes/

i) Perplexity Matrix

The object detection task is utilized to make predicts about item classifications and the directions of existing items. For a solitary ground-truth object, make "valid/invalid" expectation to mean whether ready to prevail with regards to distinguishing the item detecting the object. Table 1 gives the confusion matrix.

Table 1: Confusion matrix

Total Condition	True	False
Predicted Affirmative	True Affirmative(TA)	False Affirmative (FA)
Predicted Pessimistic	False Pessimistic(FP)	True Pessimistic (TP)

ii) Accuracy:

Accuracy gauges how precise your expectations are, demonstrating the level of right positive calculates which is given in Eqn. (11).

$$\text{Precision} = \frac{\text{TA}}{\text{TA+FA}} \quad \dots (11)$$

iii) Recall:

The review quantifies how great our indicator is at discovering all the positives, demonstrating the level of the positive ground-truth protests that locator (faster RCNN and Grid RCNN) finds which is given in Eqn. (12).

$$\text{Recall} = \frac{\text{TA}}{\text{TA+FP}} \quad \dots (12)$$

iv) Intersection over Union (IoU):

The Intersection over Union (IoU) is an assessment metric in the article location territory. It gauges the covering rate between two zones, which is to quantify how great our locator is as for the ground-truth which is given in Eqn. (13).

$$\text{IoU} = \text{Areaof}\frac{\text{Overlap}}{\text{union}}\dots \quad (13)$$

v) Standard accuracy (SA):

Standard accuracy (SA) is a significant level assessment metric in object discovery, which comprises of an accuracy, review, and IoU. For each ground-truth object, we use our identifier to make expectations and afterwards judge the right forecasts by contrasting IoU and the IoU limit. We at that point process the disarray network, exactness, and review. Then, we likewise utilize a certainty score to quantify how much certainty we have in our identifier. We set the scored edge from 0 to 1 and step size to 0.05, and afterwards, we process the accuracy and review sets which are given in Eqn. (14).

$$AP = \int_0^1 p(r)dr \quad \dots (14)$$

Table 2 represents the confusion matrix of the proposed system. In this table, the row indicates the predicted class and the column indicates the actual class.

Table 2.Performance Analysis of Proposed Work

Detection Methods	Accuracy	Recall	F1 Score	Standard accuracy (SA)
Faster RCNN	**0.6544**	**0.6933**	**0.66**	**0.6252**
Grid RCNN	**0.6953**	**0.733**	**0.77**	**0.6752**

Analysis of Vehicle Detection methods shown in below Figure 3.

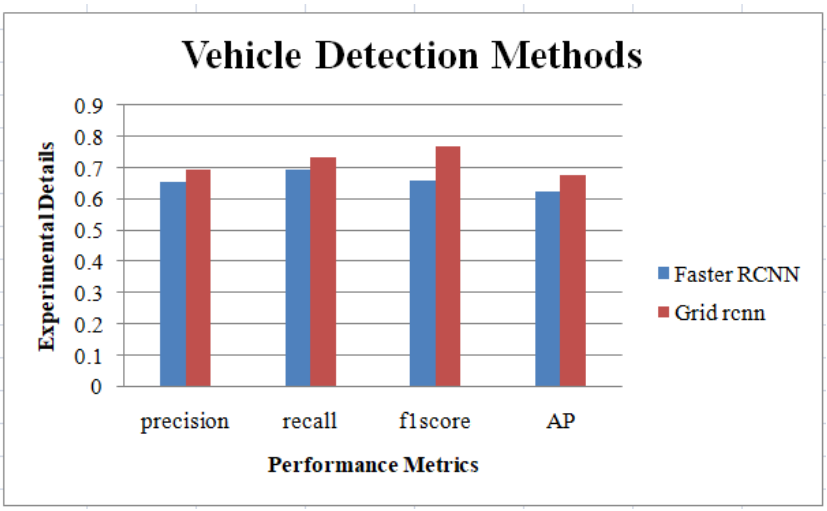

Figure 3.Analysis on Vehicle Detection Methods

6. CONCLUSION

This paper proposes a vehicle detection algorithm which combines Faster RCNN and grid RCNN. The proposed work is an optimizing algorithm which gives highly improved accuracy results on the loss function. It is done for the vehicle images based on the loss function factor in Faster RCNN detector. The framework branch finds the article by anticipating lattice focuses with the position touchy, district mapping instrument is proposed to assist ROIs with getting a bigger speaking to the territory to cover however many matrices focuses as could be expected under the circumstances, which altogether gives a superior exhibition improves the misfortune work in an exactness level. The experimental results show that the performance of the detection is better when detecting the vehicle (car) images and the corresponding detection trained end-to-end with the proposed Grid RCNN. It yields a better result than the faster RCNN method.

References

[1] Basri, R., and Jacobs, D.W. 1997, "Recognition using region correspondences", International Journal of Computer Vision, 25(2):145-166.

[2] A. Mohan, C. Papageorgiou, and T. Poggio, ”Example-based object detection in images by components. *PAMI*, 23(4):349– 361, April 2001.

[3] Paul Viola and Michael Jones, "Robust real-time object detection”, International Journal of Computer Vision, 4, no. 34-47, 2001,P.Viola, M. J. Jones, and D. Snow. “Detecting pedestrians using patterns of motion and appearance”, The 9th ICCV, Nice, France, volume 1, pages 734–741, 2003.

[4] D.Lowe, ”Local feature view clustering for 3D object recognition”, In Proceedings of the IEEE Conference on Computer Vision and pattern recognition, Kauai, Hawaii, pages 682–688. Springer, December 2001.

[5] J.Sivic and A..Zisserman. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the International Conference on Computer Vision, pages II:1470–1477, October 2003.

[6] Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: CVPR (2).(2004) 506 – 513.

[7] D.Lowe, "Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91–110, 2004.

[8] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection", In International Conference on Computer Vision & Pattern Recognition (CVPR'05), vol. 1, pp. 886-893. IEEE Computer Society, 2005.

[9] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, ”Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories”,2006.

[10] Herbert Bay, TinneTuytelaars, and Luc Van Gool, ”SURF: Speeded Up Robust Features”, 2006.

[11] OncelTuzel, FatihPorikli, and Peter Meer, "Region Covariance: A Fast Descriptor for Detection and Classification”,2006

[12] P. Felzenszwalb, R.Girshick, D. McAllester, D. Ramanan, Discriminatively trained mixtures of deformable part models, in PASCAL VOC Challenge, 2008.

[13] Xiaoyu Wang∗ Tony X. Han, Shuicheng Yan, ”A HOG-LBP Human Detector with Partial Occlusion Handling",2009.

[14] Christoph H. Lampert, Matthew B. Blaschko, and Thomas Hofmann, "EfficientSubwindow Search: A Branch and Bound Framework for Object Localization”, Ieee Transactions On Pattern Analysis And Machine Intelligence, Vol. 31, No. 12, December 2009.

[15] FlorentPerronnin, Jorge Sanchez and Thomas Mensink, ”Improving the Fisher Kernel for Large-Scale Image Classification", research gate.2010.

[16] Koen E. A. van de Sande, Jasper R. R. Uijlings Theo Gevers Arnold W. M. Smeulders “Segmentation as Selective Search for Object Recognition”,2011.

[17] Alex Krizhevsky, IlyaSutskever and Geoffrey E. Hinton,” ImageNet Classification with Deep Convolutional Neural Networks”2012.

[18] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and YannLeCun. "Overfeat: Integrated recognition, localization and detection using convolutional networks." arXiv preprint arXiv:1312.6229, 2013.
[19] Girshick Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587, 2014.
[20] J. Long, E. Shelhamer, and T. Darrell.Fully convolutional networks for semantic segmentation.In CVPR, 2015.
[21] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1989.
[22] R. Girshick. Fast R-CNN.arXiv:1504.08083, 2015.
[23] ShaoqingRen, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks." In Advances in Neural information processing Systems, pp. 91-99, 2015.
[24] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes (VOC) challenge." International Journal of Computer Vision, 88, no. 2, 2010: 303-338.
[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, PiotrDollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European Conference on Computer Vision, pp. 740-755. Springer, Cham, 2014.
[26] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The KITTI dataset." The International Journal of Robotics Research, 32, no. 11, 2013: 1231-1237.
[27] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large- Scale Image Recognition." *Computing Research Repository (CoRR)*, abs/1409.1556, 2014.
[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in CVPR, 2015.
[29] Kaiming He, Xiangyu Zhang, ShaoqingRen and Jian Sun. "Deep Residual Learning for Image Recognition."IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016: 770-778.
[30] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks., in CVPR, 2017.
[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN", 2017, ICCV , pp. 2980-2988.
[32] J. Dai, K. He, J. Sun, Instance-aware semantic segmentation via multi-task network cascades, in CVPR, 2016.
[33] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in CVPR, 2016.
[34] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, SSD: Single shot multibox detector, in ECCV, 2016.
[35] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, Dssd: Deconvolutional single shot detector, in arXiv preprint arXiv:1701.06659, 2017.
[36] Z. Li, F. Zhou, Fssd: Feature fusion single shot multibox detector, in arXiv preprint arXiv:1712.00960, 2017.
[37] L. Cui, Mdssd: Multi-scale deconvolutional single shot detector for small objects, in arXiv preprint arXiv:1805.07009, 2018.
[38] X. Lu, B. Li, Y. Yue, Q. Li, J. Yan, Grid r-CNN, in CVPR, 2019.
[39] Changqingca, Bo wang, Wenruizhang, Xiaodong Zeng, Xuya, ZhejunfengYutaoliu and Zengyanwu, "An Improved Faster R-CNN for small object detection", IEEE Citation information: DOI 10.1109/ACCESS.2019.2932731, IEEE Access
[40] Ren Y, Zhu CR, Xiao SP, "Small Object Detection in Optical Remote Sensing Image via Modified Faster R-CNN." Journal Citation Reports., vol.8, no.5, 2018.
[41] Huang Zhongjie, Fu Meixia, Ni Kaili, et al. "Recognition of vehicle-logo based on faster RCNN," 4th International Conference on Signal and Information Processing, Networking and Computers. vol.494, pp.75-83,2019.
[42] Huang Jipeng, Shi Yinchuan, Gao Yang. "Multi-Scale Faster-RCNN Algorithm for Small Object Detection."Computer Research and Development.2019, vol.56, pp.319-327.
[43] Liang Zhenwen, Shao Jie, Zhang Dongyanget al. "Small object detection using deep feature pyramid networks."19th Pacific-Rim Conference on Multimedia.2018, 11166 LNCS, pp.554-564.
[44] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In CVPR, pages 779–788, 2016.
[45] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In ECCV, pages 354–370, 2016.
[46] ShengkaiWu ,Xiaoping Li "IoU-balanced Loss Functions for Single-stage Object Detection",
[47] P. Goyal, P. D. A. R, R. B. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour," CoRR, vol. abs/1706.02677, 2017.
[48] T. Lin, P. Goyal, R. Girshick, K. He, and P. Doll A R, "Focal loss for dense object detection," arXiv preprint ar X iv:1708.02002, 2017.
[49] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-Shot Refinement Neural Network for Object Detection",2018.
[50] Z.Cai and N. Vasconcelos, "Cascade R-CNN: Delving Into High-Quality Object Detection", 2018.

AUTHORS

K.M.N.Syed Ali Fathima received the B.sc degree in Computer Science from MS University in 2012, M.sc degree in Computer Science from MS University in 2014 and M.Phil degree in Computer Science from MS University in 2017 under the guidance of
S. ShajunNisha.She is currently pursuing a PhD degree in Computer Science under the guidance of Dr K. Merriliance. Her research interest mainly includes the domain of Deep Learning in Object Detection in Image Processing.

Dr K. Merriliance pursued Bachelor of Science in Computer Science from Manonmaniam Sundaranar University, Tirunelveli in 1996, Master of Science from Madurai Kamaraj University in 1998, Master of Philosophy from Mother Teresa Women's University, Kodaikanal in 2000 and

PhD in Computer Science from Mother Teresa Women's University, Kodaikanal. Currently, she is working as an Associate professor in the Department of Computer Applications, Sarah Tucker College (Autonomous), Tirunelveli. She has published more than 19 papers in international journals and conference proceedings including Elsevier. Her main research work focuses on "An analytical study of various Object Picking Algorithms in Non-Immersive Virtual world". In her research, the performances and the characteristics of various object picking algorithms have been evaluated and analyzed within the non-immersive virtual environment and the results have been produced to show her proposed method is applicable to achieve high efficiency as compared with other traditional algorithms. She has 21 years of teaching experience in the computer science field.