DEEP LEARNING BASED OBJECT DETECTION-A REVIEW

[1]Aadithiya Mohan K., [2]Ramanand A. C.
[1]M. Tech, Signal Processing and Embedded Systems
Government College of Engineering Kannur
aadithiyamohank@gmail.com
[2]Assistant Professor, Department of Electronics and Communication Engineering
Government College of Engineering Kannur
ramanandac@gcek.ac.in

Abstract—Object detection is an essential but challenging problem in the broad image analysis field. It plays a significant part in many applications and getting distinct consideration in current years. To completely understand an image, we should not specifically focus on categorizing dissimilar images, but similarly attempt to exactly estimate the ideas and positions of objects confined in respective images. This task is mentioned as object recognition, that typically comprises with sub tasks such as face detection, pedestrian recognition, and skeleton recognition. Outdated object recognition approaches are made with handcrafted features with thin trainable architectures. Hence performance effortlessly decays by creating complex ensembles, combine several image features in low levels with the high level framework from object indicators and scene classifiers. By means of the fast expansion in deep learning, precise and powerful methods, that can absorb high-level and deeper features are presented to address that difficulties in outdated designs. These representations perform in a different way in net system style, training approach, and optimization role. Even though there are vast approaches exist, a detailed analysis of the works regarding general detection leftovers. In this study, we provide a detailed review of deep learning-based object detection frameworks.

Index Terms—Object detection, Video object detection, Face detection, Pedestrian recognition, Skeleton recognition, Deep learning.

I. INTRODUCTION

Object detection is a mixture of image sorting with exact object localization, delivering a broad and appropriate understanding of given image. Object detection involves two main stages,that is object localization which regulates the position of an item in an image and object sorting which find out to which group the object fits to. Occasionally localisation in object detection turn into hard due to blockings, substantial differences in vantage point, lightning biasness and scales[2]. Around three main sub steps exist in a detection task. Informative area choice is a method of choosing the objects that seem in an image with the flexible aspect proportions or dimensions. It is extremely computational, for the reason that the complete image is to be scanned.

Feature abstraction is the main step in object recognition that is grounded on pictorial feature mining to characterize the nature of the item. SIFT, HOGG, HAAR are some examples. Sorting is a procedure of categorising target body from all other group. Examples are SVM, ada Boost. Region proposal grounded detectors tails outmoded procedure of object recognition. That is foremost driving the section's proposal generation then categorizing the areas into diverse classes. It initially scans the complete image and afterwards concentrates on ROI. Example for such detectors are R-CNN, SPP, fast and faster RCNN, FPN.

In R-CNN object detection, initially an input picture is used, then we find out region of interest by means of some suggestion method[13],[14]. Selective search is the proposal scheme used here. Areas that are changing colours, textures, scales, as well as enclosure which make an object in an image.

Discriminatory search recognizes these outlines in the image and based on that suggests several areas. Whole these areas are reformed as per the input of the CNN, then separate areas are given to the convolutional neural network, CNN. Then abstract features in every section. Besides that support vector machine is adopted to divide these sections into diverse classes. Then a bounding box regression is adopted to forecast the bounding box of every known area.

Regression classification detection uses regression issue for nearby disconnected bounding boxes. Class possibilities are predicted by solo neural net straight from the entire image in single valuation. Examples are YOLO, SSD, multibox, attention net, Yolo-V2. In YOLO detection system, split the image into a lattice of dimensions s*s. Individual grid then predicts bounding boxes and its confidence score. Every confidence score displays in what way exact it comprises an object. Individual box forecasts the possibility that the box holds an object. This results in to a mixture of bounding boxes from images each grid. Every grid similarly forecasts conditional class probability. At that point take the product of distinct box prediction confidence values with class probabilities which gives the class definite confidence score for every box. Then NMS procedure is used to predict the outcome. Non-maximum suppression algorithm, pick only single box when numerous boxes overlay with each other and distinguish the similar object. Yolo is very fast since it adapts a solo convolutional neural network style. It can be used in real time environment. It treats the sorting as a regression issue. YOLO is more generalized. It outclasses other systems when specifying from normal imageries to additional fields alike artwork[2],[4].

II. REVIEW OF LITERATURE

Most important subsection of general object detection is face detection. In the below sections, different face detection methods are reviewed. In the first section DEFace is reviewed, which uses receptive context module and FPN principle.

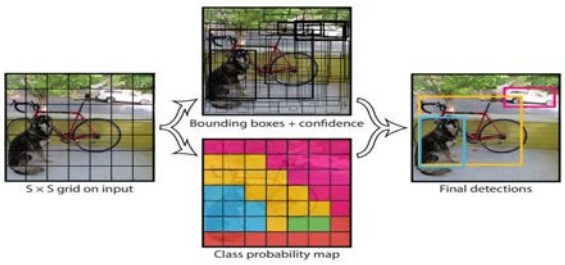

Fig. 1. Object detection using YOLO[2].

In the second section reviewed SFA, which adopts multibranch, multi-scale testing and training face detection architecture. In the third section, salient object detection has reviewed. In the next section reviewed a fast and accurate system for face detection, identification, and verification, which uses deep pyramid single shot detector and a fixed anchor for face. In the last section improvement of NMS algorithm is reviewed detailly.

A. DEFace: Deep Efficient Face Network for Small Scale Variations

Hoang et al.[6] proposed DEFace detector, a method to detect faces having sizes smaller compared to 12 pixels also having obstructions like body parts of human or face masks. Using a feature pyramid network (FPN) principle and a combination of low high resolution, detection of objects pocessing different dimensions are often possible. Small faces are spotted by extending the FPN and the P layer range expansion. For enhancing feature descriminabilty, adding receptive context module (RCM) in every feature head. which is predicted from the top down pathway in FPN architecture. Various datasets like face detection benchmark(FDDB), WIDER face and masked faces dataset (MAFA) such as obstructions due to the face mask, hair, and other body parts were deployed for the evaluation of performance and found better results compared to existing methods. The method also maintains the time for processing while detecting the very small faces. Due to the speed and processing time, single-stage method of detecting face in additional with RCM is used here.

The proposed system uses deep neural networks for the better detection of face motivated by RetinaNet and extend feature pyramid networks (FPN) and also having an extra set of RCM to add to feature maps, which is for the detection challenge of small faces. Context features of multiple scale output from the RCMs are passed into 2 convolution networks. Then it is given into different set of convolutional networks named classifications and regression nets. For avoiding a higher cost in computation and problems associated with the optimization and the supervision, classification nets and regression nets are choosing as shallow but not deep convolution layers each.

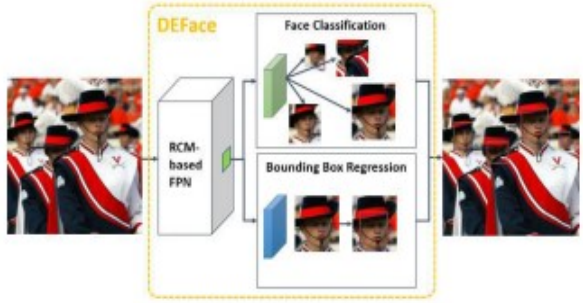

Fig. 2. The deep face detection network with multiple scales for the face[6].

The regression based bounding boxes posess similar structure of initial RetinaNet and 2 sub-nets are used. Fig. 2. indicates the deep face network designed for the detection of variable scaling faces[6]. Any deep CNN can be used for extraction of features like VGG-16, ResNet, or SSD. Here ResNet-50 is deployed in the backbone field, since it has better stability for localization as well for the recognition of objects. The basic extraction module in recent detection networks adopts ResNet and VGGNet. They exhibits identical downside when using only the singleness in receptive fields, that lead to an imbalance among the network's fields and the aspect ratio of faces. This study included an extra module of RCM to avoid this problem[11]. To enhance generalization, this method uses random flip and scaling, and resizing patches to 1280 *720 to get a training set with larger values[6].

This method effectively detect faces in different lighting and pose situations. Although most face detectors existing have difficulty in locating large faces, DEFace shows large variations. Although there are no basic-truth bounding box, the DEFace method finds a much smaller face. This method solves the problem with face spotting in situations such as obstructed by facemasks, hair, body-parts and faces under a dimension of 12 pixels. The experiments and results on various datasets WIDER Face, FDDB and MAFA dataset showed the method proposed performs well compared to other existing methods in the case of the smaller face spotting, and efficiency[6].

B. SFA: Small Faces Attention Face Detector

Zhang et al.[1] proposed SFA. It is a novel face detector which is scale invariant, referred to as Small Faces Attention (SFA) face detector, meant for improved spotting of very small faces. They use a multiple branch of face detection design that pays a lot of consideration to faces having small scale. Feature maps in near branches are fused. To end, concurrently implement multi scale training, testing to form the model strong to varied scale. The breakthrough effort of ”Viola and Jones” used Haar features to train;that attained a decent precision. Afterwards that, several methods are projected, based on that obtained progressive in face recognition. LBP in addition to its extension strategies presented native features for face detection, remained robust to brightness dissimilarities.

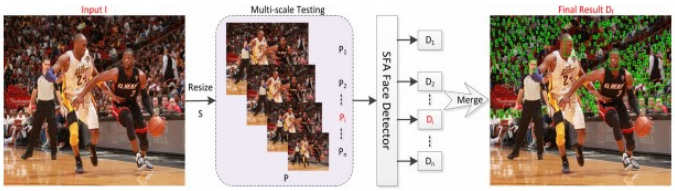

Fig. 3. SFA architecture[1].

General architecture: The input image(I) is resized for making a picture group P = P1, P2, ..., Pn and scale S = S1, S2, ..., Sn for testing in multiple scales. Every image Pi will use SFA for creating detection result as Di. Final results combines this and get Df. VGG-16 extract features of resized Pi image. Detection module M0 to M3 perceive faces with small and large scale respectively. NMS is used to create detection result Di.

Multi-branch detection architecture: it is essential to spot small faces, from initial detection layers wherever it keeps a lot of low-level features. This innovative scale-invariant face recognition design, termed multi-branch detection architecture distinguish faces by means of detection modules M0 to M3, VGG-16. That modules have 4 to 32 strides and they detect small faces, medium faces, and large ones. Figure 3 shows how an SFA works.

Small faces sensitive anchor design: Anchor-based detection of face approaches can be observed as a classification problem, that verify if the anchor is a face or not. For the improved spotting of small faces, small faces sensitive (SFS), anchor design is adopted. Anchors of size varied from 4, 64, 128 to 512, which promises that several scales of faces with abundant features for fine detection. 4 is exact small sized anchor in this technique. And anchors of 4 to 32 are small scale detection values. Advantage from the ”multi-branch detection” design, SFA sensibly organizes ”small faces sensitive anchors”, that advances the face scale. AR*BS gives size of anchor, SFA will balance the positive, negative anchors minibatch ratio 1:3.

Identifying hard samples is essential to support the detector power in training[1].

Feature map fusion: A distinctive pose, serious obstruction, extreme brightness, little resolution make CNN-based-featureextraction very hard to attain necessary and whole features. Consequently, utmost of small faces appear as hard faces. Detect small scale or hard face, using the Feature Map Fusion (FMF) strategy. FMF strategy significantly improves the detection degree on the set of WIDER FACE dataset, containing small faces.

Multi-scale training and testing: As an alternative of using static scale in training, testing stage, here accomplish MultiScale Training" (MS-Training) and "Multi-scale testing" (MSTesting) approach to acquire additional features across an extensive variety of measure, that makes this method extra strong towards diverse scale and meaningfully advances the recognition performance. In the training stage, fore mostly resize the smallest sideways of the image I up to Si. Minor faces sensitive anchor strategy is critical for perceiving slight faces. Feature map synthesis approach is capable for distinguishing tough faces. Several tactics are arranged in SFA for the sake of improved sensing of slight faces, like multi-branch recognition style, minor faces searching anchors design-feature map blend strategy- multiple scale training, testing strategy. These approaches make SFA fast, effective, and strong to sense faces in unconstrained situations, particularly for minor faces.[1].

C. What is a Salient Object? A Dataset and a Baseline Model for Salient Object Detection

Ali et al.[9] proposed a salient object detection and segmentation, which newly pays a boundless deal and attention in computer vision. It has numerous applications inhumanrobot communication, video summarization, television retargeting, photograph combination, image subdivision, image superiority valuation, image assortment surfing, content-based image reclamation and graphic stalking, image editing and manipulating, object detection, image and video compression. A salient object recognition prototypical encompasses dual phases. 1) choosing objects to process or defining saliency direction of the objects, and 2) segmentation of object zone or dividing the object and its border. Resolutions for the glitches like benchmark for datasets for mounting up models and also model expansion and an extensively approved objective description of the utmost salient object is provided in this method. Most primitive models, Itti et al. exposed instances somewhere their prototypical remained to detect spatial breaks in scenes. Liu et al.[10] investigated saliency recognition like double division problem. 1 for a forefront pixel and null aimed at a pixel in background area.

Some lessons have considered the connection among saliency fixations and decisions. For illustration, Xu et all. inspected the part of complex semantic information like object operability, watch capability and object data like object centrebias for fixation calculation in unrestricted observation of the normal scenes. They raised a huge dataset termed "Object Semantic Images and Eye-tracking (OSIE)". It is detected that eye travels are pointers of noticeable objects. Saliency is wherever a person look or which item stand out. Initial models intention was to forecast the points that individuals looks in an unrestricted observation. They requested two viewers to physically sketch substances by means of the LabelMe open annotation tool. Viewers were trained to precisely slice various things as possible by rejecting reflection of things in glasses, slicing things that are not divisible as unique. This method contains the subsequent two phases: Step 1: An input photograph from which saliency map is computed and an over segmented section plot. For the previous, they practice a complex estimate model to figure out latitudinal outlines in scenes which appeal persons eye activities and graphic consideration.

Here two models are adopted for these purpose such as AWS andHouNIPS, which are computationally efficient. General objectness degree by Alexe et all is used. The crucial feature of

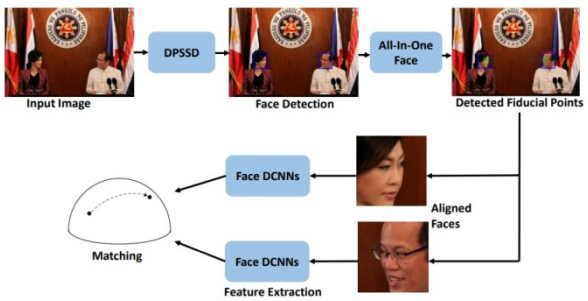

Fig. 4. Deep Pyramid Single Shot Face Detector[7].

this modest model is separating saliency from the segmentation task, means that nowadays it is possible to determine the reason of faults and low performance of methods or Spotting the

incorrect object or defective segmentation. SVO and CBsal create visually more pleasing plots. Goferman highlights object borders more than article peripheries and PCA creates centre influenced maps. Earlier modelling systems has been principally concerted on biased datasets of pictures comprising things at the centre. Here, removing this deficiency and defined in what way unbiased salient recognition datasets could be made. There is no dataset exists that has complete object remarks, clear saliency judgments and eye actions[12]. They projected an identical modest yet influential prototypical founded on super pixels which can be used as a reference line for classical valuation and assessment. Visual saliency recognition method simulates the person's visual arrangement to distinguish the scene and it is broadly used in vision tasks[9][12].

D. A Fast and Accurate System for Face Detection, Identification, and Verification

Ranjan et al.[7] proposed a new face detector, Deep Pyramid Single Shot Face Detector (DPSSD), that is fast and senses faces with great scale differences, particularly minute faces. Moreover, a novel loss function, termed the Crystal Loss is introduced for verification,identification of face. The main three contributions in this paper are,

1) An innovative face indicator(DPSSD)
2) A modest, new active Loss function, to train face verification net.
3) A precise end to end scheme for programmed face recognition.

This scheme comprises three key segments: face recognition, "face alignment-verification-identification". This system is fast and able to notice faces at huge diversity of scales. It is specifically accurate at spotting minute faces whose size is smaller than 5 percentage of image dimensions. Face recognition is a distinct case of general object detection. For the object recognition task, a SSD trained on VGG-16 is adopted. SSD takes a speed gain over different detectors like Faster R-CNN, because it is solo phase detector and doesn't use proposals. The SSD method is completely convolutional. For the better detection of objects at multiple scales here an extra layer is added. Figure 4 shows the DPSSD With models of convolution for every layer, the objects are spotted from several feature layers. Modified the SSD method in such a means that it can notice minute faces efficiently.

"Anchor pyramid with fixed aspect-ratio" for detecting several scales of faces, feature pyramid assembly inherent within the DCNN and resized input image is deployed here. To deliver contextual info, added bilinear up sampling, convolution layer near the SSD net end. The features developed are rich in localization, because the six carefully chosen layers are added elementwise to these up sampled layers. Consequently, the last detection are created from these up sampled layers by means of the anchor box matching method. WIDER Face dataset is used to train this face detector. The net is prepared by the SSD detection. On test period, the image input is resized with the smallest side having 512 pixels dimension.

The projected structure(face identification, verification) adopts the "All-in-One" face outline for key point localization, a new technique that concurrently accomplishes the responsibilities of face recognition, head-pose approximations, smile sorting, and oldness approximation. The model finely trained together for all responsibilities in a multitask learning framework, that shapes up benefits in distinct tasks, with dual nets for feature depiction and accomplishment of fusion. The dual networks are built on Inception ResNet-v2, ResNet-101, correspondingly.

This paper also debated training and datasets specifics for the structure and in what way it relays to current works on face detection. Dataset bias and domain adaptation is a difficulty for present face recognition arrangements. These schemes are generally trained on one dataset. Though, nets trained on those area may not perform fine for others. CNNs training presently takes numerous hours. Hence there arises the necessity for advancing additional effective designs[7].

E. Improvement of Non-Maximum Suppression in RGB-D Object Detection

Wang et al.[3] proposed "RGB-D object detection" strategies based on CNNs, typically analyses fusion of RGB, depth features, network structure. In 2018, Qiu et al.[4] model determined the NMS algorithm performance is considerably affected by vastly overlying objects, also its localization precision solely depends on maximum score detection. Consequently, they proposed a precise NMS technique, that progressively combines extremely overlapping recognition boxes,

captivating benefit of Regression-Soft NMS, whereas removing their shortcomings. Expand the ”NMS algorithm” to dual RGB-D CNN by means of object depth features, reducing the mislaid detection degree of extremely overlying concentrated objects, thus, enlightening the accurateness of the recognition model.

Traditional NMS algorithm: The NMS algorithm is commonly used to find out the specific prediction box which has the maximum score in the object detection task. The detection method includes mining the feature, once the classifier identifies the classification, every detection box obtains

Fig. 5. NMS algorithm[3].

a score, however the ”sliding window” can create several recognition boxes which are interconnecting other windows. NMS is required to extract the correct prediction box with the maximum score in the neighbourhood and neglet the prediction boxes that produce other lesser scores. The principle of NMS is not that much complex, and it principally includes identifying the IoU of every overlying recognition box to find out the final recognition box. Here IoU denotes a ratio, between intersection and union of dual recognition boxes zones. The detailed stages are defined as: 1) Categorise the scores of whole recognition boxes, at that point select the maximum score and matching box, 2) Examine the residual recognition boxes, if present maximum score is greater than edge T, then remove the equivalent box, 3) Continue to choose the recognition box having the uppermost score and just repeat the whole process mentioned above[3].

Depth fusion NMS algorithm: Old-style NMS algorithm has mainly two defects. Initial one is the choice of the optimum recognition box only depends upon the prediction score, that lacks robustness. Next one is two items that are adjacent together can’t be distinguished at the same time. ”NMS postprocessing” technique built on the depth fusion and also the depth features of RGB-D pictures to form some equivalent enhancements is used here. The aim is to advance the missing recognition rate and localization precision by presenting depth fusion relations. YOLO v3, Darknet-53 is the basic framework, network assembly of CNN, used for object detection in RGBD images. Figure 5 shows the NMS algorithm working.

Impressed by the system of RGB-D i.e. level-by-level feature combination, a dual channel system to extract depth features and RGB in the initial phase, that integrates actual depth features into each branch in RGB to hold out the succeeding forecast classification is designed here. To end, post-processing phase, an enhanced NMS technique is used. As a standard post processing technique, NMS contains difficulties of not sufficiently removing missed and wrong recognitions due to unsuitable IoU threshold choice. Built on the benefits of RGBD depth images object recognition, designed an enhanced NMS procedure which rest on depth fusion, that will increase the discrimination condition. The trial outcomes created on NYU dataset shows that the projected system can significantly advance the discovery of thick objects with higher intersection. It can efficiently decrease the missing object, false recognition percentage, thus improving the accurateness of RGB-D object recognition model. Just like the outdated NMS system, this process likewise faces the difficulty in IoU threshold choice, and it’s hard to prevent the missed recognition of extremely overlying objects with same depths[3][4].

III. CONCLUSION

Presented an overview of ’Deep Learning Based Object Detection’ algorithms. Discussed various steps in an object detection model and identified reggression classification methods and region based methods. And a detailed study of RCNN, YOLO based object detection has done. Mainly five papers are reviewed here. Identified and compared different stages in each model and advantages as well as the disadvantages of each system is mentioned here. Face detection is a main topic area which has distinct progress part of object detection using deep learning methods. Compared to the general object detection algorithm, face detection seems

different because of feature extraction and multiple target detection with large scale variations. Different face detection algorithms to overcome the difficulties with these problems such as DEFace, SFA are reviewed here. Object detection can provide valuable informations regarding semantic understanding in images or videos.That's why it is deployed in many applications; like image classification, human behaviour analysis, face recognition, autonomous driving systems. There are enormous methods available in the field of deep learning based object detection. Many researchers are building more efficient systems by incorporating the benefits of the old-style methods and removing the difficulties in each system from its origin to till now.

REFERENCES

[1] Shi Luo , Xiongfei Li, Rui Zhu and Xiaoli Zhang, "SFA: Small Faces Attention Face Detector", IEEE Transactions 2019

[2] Lubna Aziz , Sah bin Haji Salam , Sara Ayub, "Exploring deep learningbased architecture, strategies, applications and current trends in generic object detection: A comprehensive review", IEEE Transactions 2017

[3] Decheng Wang, Xiangning Chen, Hui Yi, and Feng Zhao,"Improvement of Non-Maximum Suppression in RGB-D Object Detection",IEEE Transactions 2019

[4] Zhong-Qiu Zhao, Peng Zheng, Shou-Tao Xu, and Xindong Wu, "Object Detection With Deep Learning: A Review", IEEE Transactions 2017

[5] Ming-HsuanYang,David J. Kriegman and Narendra Ahuja, "Detecting Faces in Images: A Survey", IEEE Transactions 2012

[6] Toan Minh Hoang , GI Pyo Nam, "DEFace: Deep Efficient Face Network for Small Scale Variations", IEEE Transactions

[7] Rajeev Ranjan, Ankan Bansal, Jingxiao Zheng, Hongyu Xu, Joshua Gleason, Boyu Lu, AnirudhNanduri, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa,"A Fast and Accurate System for Face Detection, Identification, and Verification", IEEE Transactions 2015

[8] Sami Romdhani, Jeffrey Ho,ThomasVetterand David J. Kriegman, "Face Recognition Using 3-D Models: Pose and Illumination Novel face recognition algorithms, based on three-dimensional information, promise to improve automated face recognition by dealing with different viewing and lighting conditions", IEEE Transactions 2006

[9] Ali Borji "What is a Salient Object? A Dataset and a Baseline Model for Salient Object Detection", IEEE Transactions 2015

[10] W.Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.

[11] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in IEEE Conference on Computer Vision and Pattern Recognition, June 2015, pp. 5325–5334.

[12] C.Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," IEEE Trans. Image Process., vol. 19, no. 1, pp. 185–198, Jan. 2010.

[13] S.Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[14] H.Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in Proc. IEEE Int. Conf. Autom. Face Gesture Recognit., Jun. 2017, pp. 650–657.