Deep Learning-Based Object Detection: An Investigation

Kanojia Sindhuben Babulal and Amit Kumar Das

Abstract Computer vision has one most important and challenging problem of object detection because of its wide application in field of deep learning such as medical image analysis and security monitoring autonomous driving. Object detection tasks have been greatly improved as object detection has compact association with video evaluation and image processing, and it has enticed the notice of researchers in adjunct years and describe the reference datasets at the beginning. This paper provides a complete review of a range of object detection technique, in a structured way detailing about the two-stage and one-stage detector, including the algorithms used both in detectors and in R-CNN, fast R- CNN and faster R-CNN. R-CNN, YOLO, SSD mask, etc. Also, we list the traditional (only detect object and its type) and new app (object detection with analysis and learning). Few indicative divisions of object detection are also discussed, and eventually, the performance of all models used in a one-stage and two-stage detector is discussed. We too in short examine various distinct jobs, together with projecting, face detection, object detection and pedestrian detection. Finally, various budding orientation and trends are furnished that assist as challenges or recommendation for upcoming prospective job.

Keywords Object detection · Deep learning · Computer vision · R-CNN · Fast R-CNN · Faster R-CNN · Mask R-CNN · YOLO · SSD

1 Introduction

Today's techniques for detection of objects have attracted more and more recognition in current years because of its broad variety of applications and current hi-tech advancements. This assignment is under expansive examination, inspection or real-world application for instance security reconnoiter, independent driving, transport surveillance, drone scene investigation and robotic vision. At the moment, the deep

K. S. Babulal (✉) · A. K. Das
Department of Computer Science and Technology, Central University of Jharkhand, Ranchi, Jharkhand, India
e-mail: sindhukanojia@gmail.com

P. K. Singh et al. (eds.), *Futuristic Trends in Networks and Computing Technologies*,
Lecture Notes in Electrical Engineering 936,
https://doi.org/10.1007/978-981-19-5037-7_50

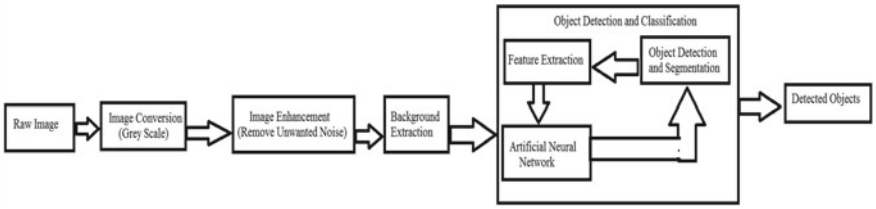

Fig. 1 Object detection workflow

learning model is broadly adopted throughout the area of computer vision, together with popular object detection as well as domain-specific object detection. Latest generation object detectors use detection and in background network utilize deep learning to extract characteristics out of provided images or videos, classification along with location separately. Object detection is an advance computer mechanism associated to image processing and computer vision in such a way that explores along with identifying occurrence of meaningful objects of a hundred of classes like as humans, cars, animal, buildings, etc. Well-investigated domains of object detection in digital images and videos comprise multi-category detection, edge detection as well as protruding object, pose, scene text, face detection, etc. [1]. Object detection is commonly applied in variety of domains of present life for instance security area, military area, transportation area, medical area and area of the life as a principal part of scene understanding. Furthermore, many milestones have played a vital part in the object detection field until now. Figure 1 illustrates the glimpse of work flow of object detection. This survey paper describes the use of different models and algorithm used for object detection and their performance including their working. And how object detection will help to build such application which can facilitate service to the human being.

2 Kind of Object Detector

Models in object detection approach is widely classified in two broad types (a) single-stage detector like YOLO and SSD and (b) two-stage detectors, for example, faster R-CNN, etc. Two-stage detector attains highest object detection accuracy but is slow while single stage detector highlights on inference speed and are very fast. Faster R-CNN (two-stage detector) executes in two steps, where first step builds RPN (Region Proposal Network) which produce region proposal that is provided to object detection model. In the second phase, the characteristics are obtained using the RoIPool (RoI grouping) functioning from every candidate box for the upcoming classification tasks as well as bounding box regression. Fundamental framework of two-stage detectors is exhibited in Fig. 2. Additionally, box prediction from the input images with the help of single-stage detector can be done without proposal of step

region making SSD extra time efficacious and further may be implemented in real-time appliances such as live monitoring via Raspberry Pi and vision by computer. Figure 3 represents the architecture of one-stage detector.

The primary benefaction of this work is recapitulated as follows: This research paper focuses on narrating and examining object detection technique based on deep learning in which object detection is major concern. This survey highlights on limited areas of object detection and do not include intricate method that might provide extra ordinary answers to progress of the computer vision analysis.

Here, the survey presents thorough study and discussion in several aspects, to the best of our understanding are the newly introduced in this area. Furthermore, non-identical from preceding surveys on object detection, this research article present methodical as well as diverse report on deep learning-based object detection technique and very crucial the modern detection results and group of notable research

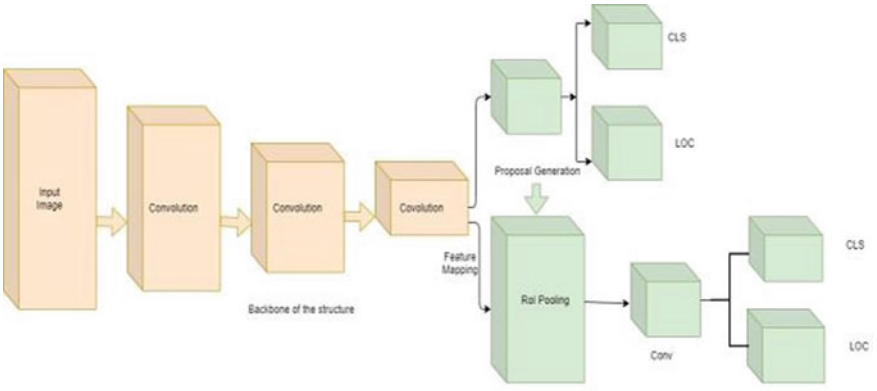

Fig. 2 Display the fundamental framework of two-stage detector that comprise region proposal network to supply region proposal within classifier plus regressor

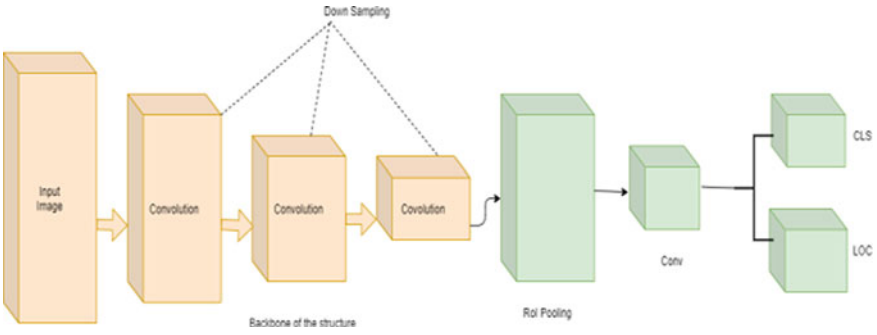

Fig. 3 Represent the fundamental framework of one-stage detector, predicting bounding boxes from provided input images instantly

inclination too. So, in this paper, our objective is to furnish an outline of how divergent deep learning techniques are utilized instead of a whole outline of all associated research articles. For better understanding of this topic, we endorse readers refer to [1, 2, 3].

The remaining of the paper is organized as depicted: Sect. 3 explains the backbone network as we know strong backbone network is important to extricate abundant attributes. Section 4 deals with the dataset; Sect. 5 discusses in detail about the applications and branches; Sect. 6 talks about the results, and Sect. 7 ends with conclusions and trends.

3 Backbone Network

Normally, domain-specific imaging object detectors are divided into two groups: one-stage detector, for example, YOLO, SSD, and other is two-stage detector, for example, faster R-CNN. High precision of location and object identification, while the one-stage detector attains elevated speed of inference. And two-stage detectors have two stages which can be divided by the grouping layer of region of interest (ROI), for example, first stage within faster R-CNN is termed RPN, which suggests frames candidate object delimiters. In the next stage, from ever single candidate box features are extracted in upcoming classification along with bounding box regression jobs using the ROIPOOL operation (ROI pooling). Figure 2 exhibits the primary framework of two-stage detectors. In addition, one-stage detectors advocated anticipated boxes out of input images straight way lacking a region suggestion phase. Therefore, these approaches are time saving as well as implemented with real-time devices, e.g., for live monitoring by Raspberry Pi and computer vision. Figure 3 exhibits primary framework of single-stage detectors. This detector was developed as backbone to accommodate detection layout but these are poor at identifying uneven shaped objects or a batch of small objects. The newly developed highly efficient classification networks can enhance the application as well as diminish the complication of the object detection mission. This is an effectual procedure to further improve network performance due to the backbone network acting as a feature extractor.

3.1 Typical Baseline

With expeditious growth of deep learning and the side-by-side upgradation of computing power, an enormous advancement has been done through area of normal object detection. Accompanied with the progress of earliest CNN-based object detector, R-CNN was introduced, and a sequence about notable advancement is carried that encourages expansion of common object detection with a massive limit. For beginners, some representative object detection framework has been initiated to obtain commencement within said field.

3.2 Two-Stage Detector

R-CNN: Region-based detector (R-CNN) has four modules. Category-independent region is produced by first module region, while the second module separates each region proposal from fixed-length feature vector. The third section acts as a group of class-specific linear SVMs for categorizing the objects present inside single image, and fourth and last section showcases bounding-box regressor to accurately predict bounding box. This uses prior-training on huge dataset accompanied by adjustment on particular dataset is a commendable technique to reach rapid convergence.

Fast R-CNN: The drawback of R-CNN taking long processing time on SVM's classification leads Ross Girshick [4] to propose a speedy variant of R-CNN, termed as fast R-CNN after one year of R-CNN. Fast R-CNN separates features out of a complete photograph as input as well as subsequently proceeds the region of interest (ROI) pooling layer to acquire the stilled size characteristics serving as input toward classification along with bounding box regression are completely connected layers. The feature is separated by the complete image one time and as well as is forwarded to CNN toward classification and localization. While in R-CNN complete inputs image every region is forwarded to CNN that consumes excess time which is stored in fast R-CNN because CNN's processing time and huge disk depot to accumulate a prominent trade in features is saved.

Faster R-CNN: Faster R-CNN [5] proposed that after 3 months of fast R-CNN additionally enhances the region-based CNN baseline. Selective search is utilized by fast R-CNN to recommend region of interest that is indeed stolid as well as requires the identical execution time as the detection network. RPN is substituted in faster R-CNN that is completely convolutional network to accurately anticipate region proposal along with a vast scope of scales as well as aspect ratios. RPN improves speed of region proposals generation because it dispenses full-image convolutional characteristics along with a general group of convolutional layers including the detection network. Assessment demonstrates precision along with detection efficiency exceeds in faster R- CNN. Fast R-CNN outperforms faster R-CNN 10 times in total running time with identical backbone.

Mask R-CNN: This is expansion of faster R-CNN basically designed for instance segmentation task. Mask R-CNN is more precise object detector. He et al. utilize faster R-CNN accompanied by ResNet [6]-FPN [7] (feature pyramid network, a backbone separates ROI characteristic by various level about the feature pyramid relative for its scale).

3.3 One-Stage Detector

YOLO: You Only Look Once (YOLO) was introduced by Redmon et al. [8], which predicts both together confidence for numerous groups as well as bounding boxes. Working of You Only Look Once is displayed in Fig. 4. YOLO splits the given image within S × S grid; moreover every grid cell remains accountable toward estimating the object focused within particular grid cell. Every grid cell estimates B bounding boxes along with their equivalent confidence scores. This confidence scores actually interpreted as Pr(Object) * IOU $\text{pred}^{\text{truth}}$ that demonstrate how probably there occur objects (Pr(Object) > = 0) as well as present confidence of its prediction ($\text{IOU}^{\text{truth}}$pred).

There are 24 convolutional layers (CONV) including two fully connected layers (FC) in YOLO. Few CONV layers devise groups of initiation section with 1*1 reduction layers accompanied through 3*3 CONV layers. Speed in which network could operate on images in real time at 45 frame per second (FPS) as well as fast YOLO extends by 155 FPS with preferable outcomes as compared to remaining contemporaneous detectors. Moreover, YOLO construct scares false positives on backdrop that helps in coordination through fast R-CNN. Later an advanced edition, YOLOv2 was suggested in [9], which supports various magnificent techniques for instance (batch normalization) BN, anchor boxes, and dimension cluster including multi-scale training.

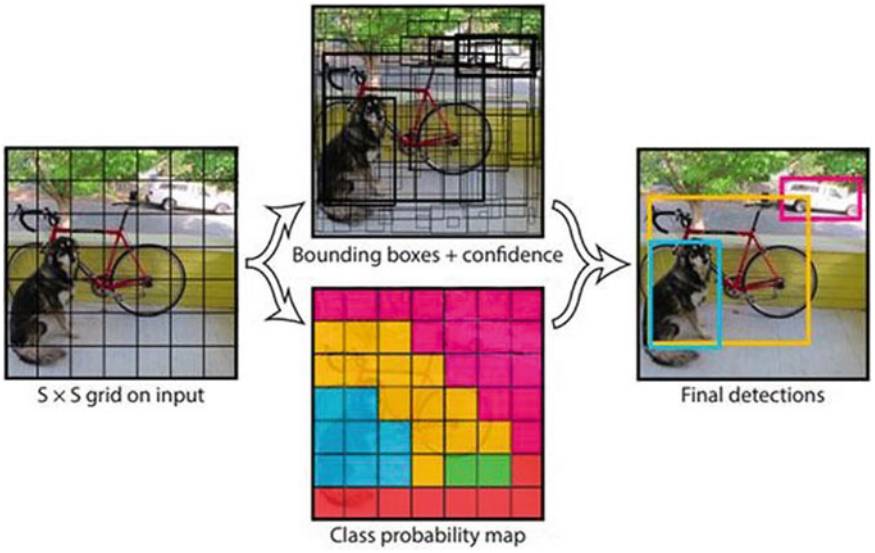

Fig. 4 Working of YOLO [8]

YOLOv2: [10] is advanced edition of YOLO [11] that makes arrogate sequence of blueprint conclusion by previous tasks with novel idea to enhance YOLO's precision and speed.

YOLOv3: It is an enhanced edition of YOLOv2. Initially, YOLOv3 utilize multi-label classification to accommodate additional difficult, dataset comprising overlapping labels. YOLOv3 utilizes three distinct scale feature maps to forecast the bounding box. 3-D tensor encoding class prediction, objectiveness and bounding boxes are forecasted by the extreme convolutional layer. YOLOv3 recommends an in depth as well as vigorous feature extractor, named darknet-53, influenced through ResNet.

SSD: YOLO is strenuous in tackling small-sized objects in set because of strong spatial constraints forced on bounding box predictions [8]. At the same time, YOLO scuffles to approximate to objects in current/uncommon aspect ratios/layout and assemble, respectively, substantial features because of numerous down sampling functioning.

Liu at el. discuss a single-shot multi-box detectors (SSD) [12], motivated by the anchors embraced in multi-Box [13], RPN [4] including multi-scale portrait [14]. Provided particular feature map, SSD captures the benefit concerning group of conventional anchor boxes with separate aspect ratio along with scales to separate the resultant space of bounding boxes different from YOLO which adopts fixed grids. To manage items with numerous measurements, the network merges prediction from numerous feature maps with distinct resolution. Figure 5 demonstrates the architecture of SSD. Provided with VGG16 backbone framework, SSD appends various feature layers with verge of the network. As a result, anticipating the offsets to conventional boxes to variable scales, aspect ratios including related confidence become easier. The network is upskilled with a weighted sum of localization loss as well as confidence loss. Terminating detection solutions are gathered by performing NMS using multi-scale filtered bounding boxes. The SSD 300 runs at 59 FPS and is extra precise as well as well-structured than YOLO, and at the same time SSD is not expert at dealing with objects with smaller size.

DSSD: Deconvolutional single-Shot detector (DSSD) [15] is an enhanced update of single-shot detector (SSD) that appends prediction module as well as deconvolution module and further embrace ResNet-101 as backdrop network. Toward prediction module, a residual block is appended to every predicting layer, and subsequently component-wise inclusion of the results to prediction layer and residual block is done. Deconvolution module increases the resolution of feature maps to strengthen features. Every deconvolution layer accompanied by a prediction module to estimate numerous objects escorted by varied measurement.

Retina Net: In February 2018, Lin et al. [16] introduced focal loss as classification loss function in one-stage object detector, whereas two-stage object detector R-CNN is classical. In the initial phase produces a scanty group of region proposals helping next phase categorize every candidate placement. First stage filters out the majority of negative locations. When compared to one-stage detector, two-stage detector can achieve higher precision which proposes a compact set of candidate locations.

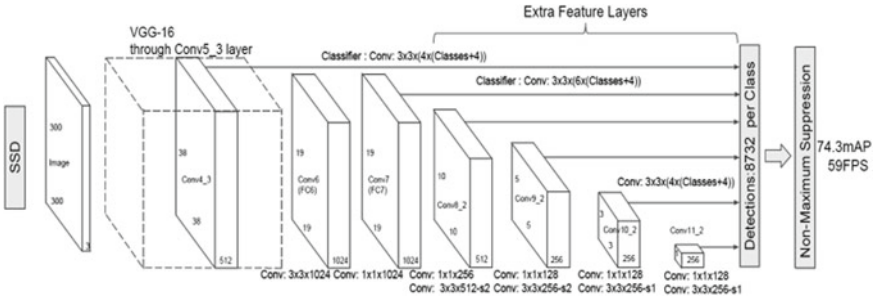

Fig. 5 Architecture of SSD [12], At the end of VGG16 backbone network, SSD appends various feature layers to forecast the offset to conventional anchor boxes including corresponding confidence. Concluding detection outcomes are procured due to performing NMS upon multi-scale refined bounding boxes

The primary reason to this higher precision is utmost forefront–background class imbalance, while one-stage detector trains networks to obtain confluence. Retina Net inherits a rapid speed of preceding one-stage detectors with disadvantages of strenuousness to teach unbalanced positive and negative illustration.

M2Det: M2Det contributes to a huge diversity of scale differences beyond object example, and a multi-level feature pyramid network building additional effectual feature pyramids was proposed by authors [17]. For acquiring final enhanced feature pyramid, authors choose three steps. In the first step like FPN, M2Det adopts multi-level features extraction by several layers at backdrop that acts as the foundation attributes. Next step involves feeding into the block the base feature, comprising changing joint thinned U-shape sections in addition to feature fusion sections, to acquire decoded layers of TUM for feature in succeeding step. Third and final step, a feature pyramid comprising multi-level features is assembled by merging each decoder layers of homogenous scale. Up to aforementioned point features in company of multi-scale as well as multi-level are available. Subsequently rest portion functions to pursue the SSD framework for procuring bounding box localization with classification outcome in a successive mode.

Refine Det: Complete network of Refine Det [18] comprises paired interrelated sections, first the anchors refinement section with second object detection section. Both the sections are attached closed to transfer connection block in order to shift with increasing features out of previous section to improve foretell objects within final section. Each training procedure is an end-to-end method, managed in trio phases, preprocessing, detection (two inter-connected sections), as well as NMS.

SSD, YOLO and Retina Net are classical one-stage detector which uses one-phase regression procedure for acquiring the last output. Furthermore researchers discover allowing usage of two-step cascaded regression procedure will preferably foretell higher quality hard detected objects, particularly those of tiny size objects including furnish extra precise point of objects in image (Table 1).

Table 1 Comparison of various models

Sl. no	Parameters	YOLO	SSD	Retina Net	DSSD	Refine Det	M2Det
1	Model name	You Only Look Once	Single-shot multi-box detector	Retina Net	Deconvolutional single-shot detector	Refine Det	M2Det
2	Backbone network	DarkNet, COSA, DeiT-Ti	VGG MobileNet	ResNet	Feature pyramid network (FPN)	VGG-16 or ResNet-101	VGG-16 or ResNet-101
3	Speed	Low	High	High	High	High	High
	Accuracy (%)	80.3	72.1	83.3 [18]	73.2	81.8	64.6
4	Time	0.84–0.9 sec/frame	0.17–0.23 sec/frame	0.62–0.75 sec/frame	0.20–0.30 sec/frame	0.62–0.75 sec/frame	0.34–0.84 sec/frame
5	Frame per second	45	59	122	50	37	12
6	Mean average precision	0.358	0.251	0.786	0.293	0.29	0.481
7	Number of boxes	~1 k	~8–26 k	~100 k	~8–30 k	~100 k	Default

4 Datasets

Object detection indicates that the object consists in a particular class and locates it in the image. Bounding box is used for localization of an object. Implementing challenging dataset as benchmark is noteworthy in numerous sectors of research, as they sketch a benchmark correlation between various algorithms and place objective for results. Former algorithms concentrate on face detection utilizing numerous ad hoc data records. Subsequently, additional practical and difficult face detection dataset were produced. Other popular obstacle within face detection datasets is to create the dataset. The common object detection datasets, e.g., PASCAL VOC [5], MS COCO [19] and ImageNet-loc [3], are popular guideline for object detection exercise. Furthermore, authorized benchmark is typically assumed to check the functioning of detectors in company of relative dataset. All the fashions to be had at the tensor flow item detection model can be practiced on coco dataset (commonplace objects in context) [20]. This dataset comprise 120,000 photos with a complete 880,000 labeled objects in these images. These models are clever to come across the 90 different varieties of items classified on this dataset. An entire listing of all these different gadgets is available in the statistics part of the skilled model. This listing of items consists of a vehicle, someone, a desk, etc.

5 Applications and Branches

Various areas widely use object detection applications to enable humans to complete tasks, like military domain, security area, transportation area, medical area, etc. We demonstrate the traditional and latest mechanism deployed in this area in detail.

5.1 Safety Field

Face Detection focuses at diagnosing human faces in a photograph. Due to illumination and resolution variation and extreme poses, face detection is still a complex goal. Numerous implementations pay attention on accurate detector designing. To uplift the performance of distinct task, Ranjan et al. [21] discover corresponding work (gender identification, face identification, face landmarks localization and head posture evaluation). To discover unvarying characteristic among near-infrared (NIR) as well as visual (VIS) face images, He et al. [22] instigated a new Wasserstein convolutional neural network technique. Relevant designing of loss functions will improve discriminating ability of DCNNs built on wide-ranging face identification. Researchers in [23, 24, 25, 26] state that the cosine-based SoftMax losses attains considerable victory in deep learning-deployed face identification. Deng et al. [27] propose an additive angular margin loss (ArcFace) toward achieving immensely different attribute aimed at face identification. GUO et al. [16] introduces a fuzzy sparse auto-encoder structure aimed at solo input picture each individual face identification. Readers can consult to [15] for better understanding.

Pedestrian detection emphasizes at identifying walker/pedestrians inside the neural view. Braun et al. [17] emancipate a European City Persons dataset comprising pedestrians, bicyclists and different riders in city busy locations. Real-time walker recognition dedicated to complicacy-conscious cascaded pedestrian spotter.

Anomaly detection assists in swindle identification, weather study, as well as healthcare observance. Present abnormality identification procedures examine the details on a point-wise rationale. To point the expert analyst interesting regions (anomalies) of the data a novel unsupervised procedure termed “Maximally Diver gent Intervals” (MDI) that hunts for adjacent intervals of time including regions in space by Barz et al. [28].

5.2 Defense Area

In defense area, automatic target detection, topographic survey, remote sensing object identification, surveillance mission, flyer identification, intelligent airport security system, anomalous behavior detection, etc., are illustrative applications, which are

using object detection mechanism to combat against criminal activity, public safety, terrorism, etc. So, in this kind of application, high-speed computing service molds it to be more impressive, and hence, amid various object detection techniques YOLO is the most appropriate in terms of functioning. For such critical areas, symmetry between network lifetime and object detection techniques is primary concern, and therefore, better approach which focuses on reliability of link becomes important in any wireless network [29].

From a provided video or images identification and classification of military vehicle with assistance of unmanned aerial system is still a challenge and object detection approach can prove to be beneficial in this aspect. During battle automatic target detection in battlefield from captured images with the help of object detection can be helpful.

6 Results

A system's or algorithm's result analysis is based on a set of parameters. Performance, time spent, resources required, accuracy and other factors are commonly used in practically all analyses. The performance is a parameter that indicates how well the algorithm works. The time consumed to evaluate the algorithm and obtain the output is represented by the parameter time taken. The quantity of resources required by the algorithm is defined as resources needed. Accuracy is the algorithm's promising factor, defined as the percentage of accurate output produced by the algorithm.

When the common conditions are applied to Ross Girshick's R-CNN techniques of object detection, the outcome reveals that R-CNN is significantly speedy than the conventional techniques based on classification techniques [30, 31]. RCNN uses a selective search to extract only 2000 regions per image, rather than a large number of regions. As a result, the feature extraction will only cover 2000 regions. R-CNN still has its limitations after such a substantial reduction in computation. To begin with, training the network takes a long time because each image is classified into 2000 regions. Second, it is not applicable in real time because each test image takes around 40 s to process. Finally, because it employs a preset selective search strategy, it is unable to learn from past experiences.

Ross Girshick later designed a latest type of R-CNN entitled fast R-CNN to address the R-CNN's flaw. Unlike R-CNN implementation the input picture is supplied to the CNN instead of the region proposals in this method. To locate image's proposal region, CNN creates a convolutional feature map. For each object, a feature vector is produced from the feature map, and the softmax layer is utilized to forecast the class of the proposed area based on this vector increasing the effectiveness of this technique.

It is not mandatory to provide CNN 2000 area proposals every single time, which is considerably superior to R-CNN. Instead, each image is subjected to a single CNN process. Another method, alike R-CNN including fast R-CNN, is suggested. The approach is implemented same way as the former techniques, but rather than selective

search algorithm, an independent network is utilized to forecast the proposed regions. The ROI polling layer is used to reshape the proposed regions, which are subsequently utilized to separate and identify the classes and borders. Because an independent network rather than a fixed technique is utilized to anticipate the proposed region, this technique is substantially speedy than fast R-CNN.

For object recognition, a recent method YOLO is presented. Because previous methods rely on proposed areas to recognize the object in the image and never whole image is considered. Object detection is performed on regions with a high possibility of containing items. However, in YOLO, there is simply one convolutional network, which analyzes the entire image. The image is divided into a S $\times$ S grid, accompanied by m bounding boxes. The network produces a class probability for each box, including the classes with greater probability compared to the threshold value which are utilized to find the object. Due to its single convolutional neural network, this technique has numerous advantages. The whole image is assessed once, and bounding boxes and class probability are calculated. Second, the whole detection procedure is conducted in a single network, making network optimization simple. Because it just has one convolutional neural network, it is substantially speedy than the R-CNN, fast R-CNN and faster R-CNN.

Authors in paper [32] compares various models in terms of delay, mean average precision (mAP), frames per second (FPS) and usage in real-time applications. Aforementioned paper ahead without doubt illustrates that YOLO outperforms R-CNN-based algorithms in terms of latency and frame per second (FPS). It is evident that a precision trade-off was made in order to attain this speed. Despite its small mean average precision (mAP), YOLO has an acceptable for real-time applications, and once combined with its high FPS and latency, it is evident that it is the finest algorithm in its class.

7 Conclusion and Trends

Deep learning-centered object detection technology has matured swiftly along with increase the regular upgrade of strong and intensive computing equipment. To implement more dynamically and precise applications, there is a requirement of a high-precision real-time system. Researchers have developed various directions to attaining highly efficient precision detectors like building a current architecture, squeezing rich features, utilizing good representation, upgrading processing speed, train from beginning, techniques without anchors and troubleshooting complex 40 to 4 scenes such as small objects and occluded objects. Implementation of object detection is slowly expanded in the security area, military area, transportation area, medical area and life field with the increasingly influential object detectors. In recent days, but there is still a lot to achieve in future development.

Due to the importance of speed and accuracy in object detection applications, it is crucial to maintain a small computation time and to process the input speedily in order to give the user with the output. Because it only has one neural network

and works in real time, YOLO is the ideal option for real-time object detection. Quick computation time. It can produce results faster than other approaches, and the precision of the technique can be handled according to the system's needs.

7.1 Trends

Hybrid stage detector: Hybrid stage detector is combination of various stage detectors like one-stage detector and two-stage detector. Such detector helps researcher gain higher precision and maximum throughput in real-time applications.

Video object detection: *Detection* of object in real-time video, operational videos are important research areas and are extremely challenging due to problems like motion blur challenging to detect object in moving video, hazy motion, defocus of video, motion target vagueness, extreme target movement, tiny targets, obstruction and abridged. To achieve a better performance is really a difficult task.

Multi-domain object detection: High detection interpretation regularly requires domain-specific detection on the specified dataset. Therefore, a general detector which is efficient on functioning on numerous multi-domain detector, image domains can resolve this issue without previous understanding of current discipline. Domain shift is a bothering assignment for future research.

3-Dimension object detection: When juxtaposed to 2D image-based detection, 3-D object detection has become demanding research orientation with the innovation of 3-D sensor and manifold application of 3D knowledge. "Light Detection and Ranging" (LiDAR) point cloud produces genuine depth details which can be utilized to precisely discover objects and personify their forms. In 3-D space, LiDAR also permits correct localization of objects.

Salient Object Detection: Salient object detection (SOD) has a difficulty in spotlighting major object regions in pictures while in video object identification classification as well as location of objects of interest in a continual arena is done. SOD is executed in a vast scope of object-level application in so many disciplines. Precise "object detection" in videos is offered by SOD by extracting object regions of interest in every frame. Hence, for prominent identification and difficult detection mission, spotlighting target detection is a pivotal preparatory procedure.

Multi-source information assistance: Nowadays, with the growth of big data techniques including widespread usage of Internet community, multi-source details are becoming effortless to approach. Numerous social media details can supply both images and videos, and their representation in textual form can assist identification duty. Blending multi-source data is becoming apparent study inclination with the emerging of numerous techniques.

Medical diagnosis and imaging: With the encouragement, artificial intelligence deployed medical gadget in April 2018, Food and Drug Administration (FDA) in US initially recommended a diabetic retinopathy identifier having correctness of 87.4% which was based on artificial intelligence software named IDx-DR. Amalgamation of image identification systems as well as mobile devices can prepare cell phone as strong family investigative device for anybody. Aforementioned inclination is packed with obstacles at the same time with too much hopefulness.

Real-time detection and remote sensing airborne: Precise investigation of remote sensing image, automated detection software and integrated hardware are essential in both agriculture fields and military services, and this will escort unmatched development in this area. System on chip (SoC) realizes real-time high-altitude detection while supplying to deep learning-focused object detection methodology.

Advanced medical biometrics: Researchers started studying utilization of deep neural network, by using neural networks to recognize the possibility of heart disease by inspecting the retinal images and speech pattern. Medical biometrics and its application will be used for passive monitoring in near future.

References

1. Khan A, Sohail A, Zahoora U, Qureshi AS (2019) A survey of the recent architectures of deep convolutional neural networks. arXiv preprint arXiv:1901.06032
2. Zou Z, Shi Z, Guo Y, Ye J (2019) Object detection in 20 years: A survey. CoRR, abs/1905.05055
3. Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikainen M (2018) Deep learning for generic object detection: A survey. arXiv preprint arXiv:1809.02165
4. Jiang H, Learned-Miller E (2017) Face detection with the Faster R-CNN. In: 12th International Conference on Automatic Face & Gesture Recognition. IEEE, Washington DC USA
5. Yang Z, Nevatia R (2016) A multi-scale cascade fully convolutional network face detector. arXiv:1609.03536v1[cs.cv]
6. Lin T, Dollar P, Grilshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid network for object detection. In: Conference on Computer Vision and Pattern Recognition (CVPR), 936–944. IEEE, Hawaii
7. Girshick R, Donahue J, Darrell T,Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 14th International Conference on Computer Vision, IEEE, Ohio
8. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: International Conference on Computer Vision and Pattern Recognition. IEEE, Las Vegas
9. Redmon J, Farhadi A (2016) YOLO9000: better, faster, stronger, 7263–7271. arXiv:1612.08242
10. Li Y, Lu Y, Che J (2021) A deep learning approach for real-time rebar counting on the construction site based on YOLOv3 detector. Automation Const 124:1–14
11. Liu Z, Li J, Shu Y, Zhang D (2018) Detection and recognition of security object based on Yolo9000. ICSAI. IEEE, Nanjing
12. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) SSD: Single shot multi-box detector. In: European Conference on Computer Vision. Springer, Cham
13. Erhan D, Szegedy C, Toshev A, Anguelov D (2014) Scalable object detection using deep neural networks. In: International Conference on Computer Vision, IEEE, Ohio

14. Bell S, Lawrence Zitnick C, Bala K, Girshick R (2016) Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In: CVPR, IEEE. Las Vegas
15. Wang M, Deng W (2018) Deep face recognition: a survey. https://arxiv.org/abs/1804.06655
16. Guo Y, Jiao L, Wang S, Wang S, Liu F (2017) Fuzzy sparse autoencoder framework for single image per person face recognition. IEEE Trans Cybernatics 48(8):2402–2415
17. Li X, Flohr F, Yang Y, Xiong H, Braun M, Pan S, Li K, Gavrila DM (2016) A new benchmark for vision-based cyclist detection. In: Proc IEEE Intell Vehicles Symp (IV), pp 1028–1033. IEEE, Sweden
18. Tetila EC, et al (2020) Detection and classification of soybean pests using deep learning with UAV images. Comp Elect Agri 179:1–11
19. Chen C, Seff A, Kornhauser AL, Xiao J (2015) Deep driving: learning affordance for direct perception in autonomous driving. In: 15th International Conference on Computer Vision, IEEE, Chile
20. https://cocodataset.org
21. Ranjan R, Patel VM, Chellappa R (2019) 'HyperFace: a deep multitask learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans Pattern Anal Mach Intelligence 41(1):121–135
22. He R, Wu X, Sun Z, Tan T (2019) Wasserstein CNN: learning invariant features for NIR-VIS face recognition. IEEE Trans Pattern Anal Mach Intelligence 41(7):1761–1773
23. Zhang X, Zhao R, Qiao Y, Wang X, Li H (2019) AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations. arXiv:1905, 00292
24. Liu Y, Li H, Wang X (2017) Rethinking feature discrimination and polymerization for large-scale recognition. arXiv:1710.00870
25. Ranjan R, Castillo CD, Chellappa R (2017) L2-constrained softmax loss for discriminative face verification. arXiv:1703.09507. https://arxiv.org/abs/1703.09507
26. F. Wang, X. Xiang, J. Cheng, and A. L. Yuille.: NormFace: L2 hypersphere embedding for face verification. In: Proc. 25th ACM Int. Conf. Multimedia, ACM, 1041–1049 (2017).
27. Deng J, Guo J, Xue N, Zafeiriou S (2018) ArcFace: additive angular margin loss for deep face recognition. arXiv:1801.07698. Available: https://arxiv.org/abs/1801.07698
28. Bjorn B et al (2018) Detecting regions of maxima divergence for spatio-temporal anomaly detection. IEEE Trans Pattern Anal Mach Intell 41(5):1088–1101
29. Babulal KS, Tewari RR (2011) Cross layer design with link and reliability analysis for wireless sensor network. In: Proceedings of 2nd International Conference on Current Trends in Technology, IEEE. Nirma University Ahmedabad
30. Khare M, Thanh Binh N, Srivastava RK (2014) Human object classification using dual tree complex wavelet transform and Zernike moment. In: Transaction on large scale data and knowledge centered system XVI, LNCS, 87 101
31. Kumar P, Thakur RS (2021) An approach using fuzzy sets and boosting techniques to predict liver disease. CMC-Computers Mat Cont 68(3):3513–3529
32. Kumar P et al (2019) A comparative study of object detection algorithm in a scene. Int J Eng Res Tech 8(5):1–3