IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 2, April, 2023, pp. 407-414
Journal homepage: http://iieta.org/journals/ria

A Comprehensive Survey on Object Detection Using Deep Learning

Bhagyashri More[1], Snehal Bhosale[2]*

[1] Department of Computer Science, Symbiosis International (Deemed University) Lavale, Pune 412115, Maharashtra, India
[2] Department of Electronics and Telecommunication, Symbiosis Institute of Technology (SIT), Symbiosis International (Deemed University) (SIU) Lavale, Pune 412115, Maharashtra, India

Corresponding Author Email: snehal.bhosale@sitpune.edu.in

https://doi.org/10.18280/ria.370217

Received: 20 January 2023
Accepted: 10 February 2023

Keywords:
object detection, computer vision, deep learning, one-stage object detection, two-stage object detection

ABSTRACT

One of the common and difficult issues in computer vision is to detect the object. Researchers have widely experimented and contributed to the performance improvement of object detection and associated tasks including object classification, localization, and segmentation over the way of the last decade of deep learning's rapid evolution. Object detectors can be broadly categorized into two groups: two stage and single stage object detectors. Two stage detectors primarily focus on selected region proposals via sophisticated architecture whereas single stage detectors concentrate on all feasible spatial region proposals for object detection via relatively easier architecture in one go. Any object detector's performance is assessed using inference time and detection accuracy. In regards to detection accuracy, two stage object detectors surpass single stage object detectors. In this survey, we present a deep literature survey on object detection methods. We also provide a summary of the comparison between two-stage and single-stage object detectors along with suggestions for further research in real-world.

1. INTRODUCTION

Object detection has recently gained popularity due to its vast range of applications. Object detection is the most important aspect of computer vision. It is employed in real-life applications such as security, autonomous driving, video surveillance, remote sensing target detection, robotics, and so on [1, 2]. The main goal of object detection is to recognize visual items in images or videos of a given class, such as humans, cats, dogs, books, vehicles, etc., and subsequently highlight those objects by drawing boxes and sort out them into the classes of that specific object. Deep Learning algorithms have been widely employed in all aspects of computer vision in recent years. Object detection was designed using standard approaches until 2014 before Deep Learning methods were introduced [3-5].

Traditionally it works on SIFT, HOG, Haar, DPM and VJ detector. As SIFT algorithm has not worthy at illumination changes and high computational cost because it is very slow. In HOG, object identifying time is large due to it uses sliding window approach for feature extraction. Training duration is very large in VJ detector [6]. To overcome the problems of traditional methods Convolution Neural Network (CNN) was reintroduced with Deep Learning for object detection which has been recognized as a base for future approaches to video object detection tasks. Due to the fantastic performance of CNN, it is broadly used in image processing and computer vision fields; it generates accurate performance in image classification and detection tasks [7]. Figure 1 depicts the classification of object detection approaches, whereas Figure 2 depicts the fundamental architecture of CNN. In this survey paper, we have given more attention to deep learning methods: Two stage and Single or One stage.

The key challenges in object detection are as follows:

i. Intra class variation: Intra class variation across examples of the alike object is widespread in nature. This change could be produced by several reasons, including occlusion, illumination, position, and viewpoint. These uncontrolled externals can have a significant impact on the appearance of the thing [8]. It is expected that the components will deform non-rigidly may be rotated, resized, or obscured. Some items and may be surrounded by unobtrusive surroundings, which makes the extraction process hard.

ii. Number of categories: This is a stimulating task due to the large number of object classes to categories. It also demands more high-quality annotated data, which is currently in short supply.

A research question is if it is possible to train a detector with fewer examples.

iii. Efficiency: Modern prototype requires a large amount of computer power to provide decent detection results. With the development of edge and mobile devices, good object detectors are important for furthering computer vision innovations.

The essential topologies of deep learning-based object detection models are shown in Figure 3. In general, deep learning-based object detection models have a backbone and a head network. The head network uses the extracted features to locate the bounding boxes of the identified objects and classify them after the backbone network extracts information from the input images (See Figure 3).

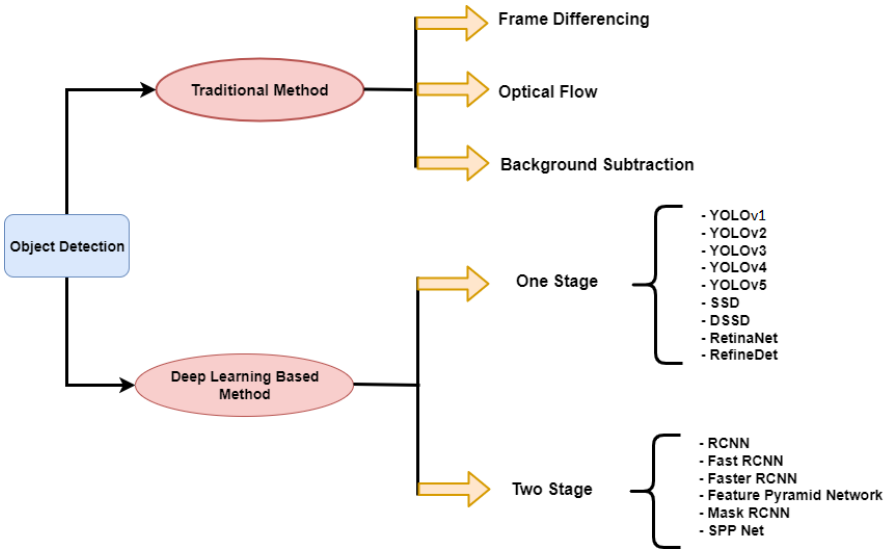

Figure 1. Object detection techniques

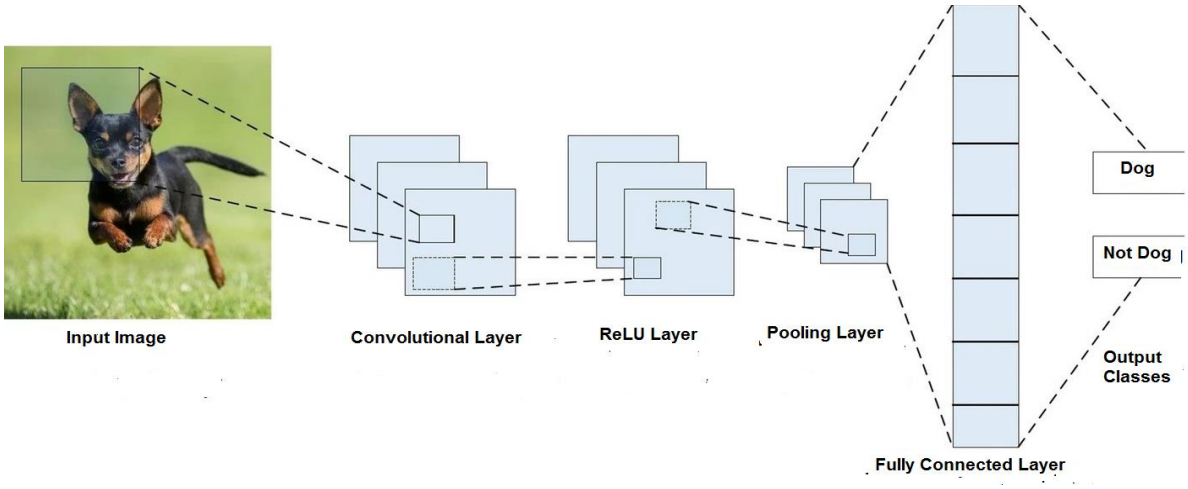

Figure 2. CNNarchitecture [9]

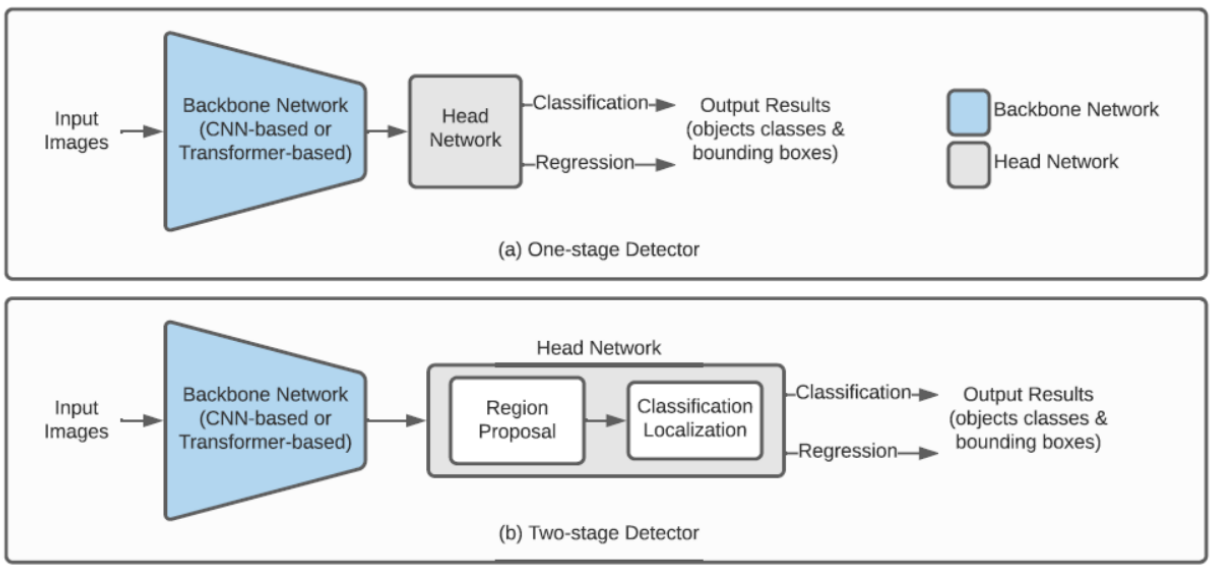

Figure 3. Basic object detection model architectures based on one-stage vs. two-stage deep learning. The backbone network can be employed as a CNN or transformer-based network, and depending on the head network's structure, it can be divided into one-stage or two-stage networks. The one-stage detector operates in the brain network's object localization and classification processes concurrently, as shown in (a). However, after obtaining the region proposals, the two-stage detector performs localization and classification on the regions, as demonstrated in (b)

We have thoroughly examined several object detection architectures and related technologies in this paper. The rest of the paper is constructed as: Section II gives a comprehensive literature review of advanced deep learning methodologies systems. Section III gives an analysis of the systems studied in the literature survey is done and future suggestions are provided. Section IV concludes the paper.

2. LITERATURE SURVEY

Object detection based on deep learning is broadly categories into two modules: Anchor-Based and Anchor-Free. Anchor-Based module can be further divided into two detectors according to the different training methods: Two-stage detectors (based on region proposals) and One-Stage detectors (based on regression) [10].

This review has two sections: two-stage detectors and single-stage detectors. A system called a two-stage detector has a second module that forecasts region borders. Object proposals are located in an image in the first stage of the model, then categorize and localize them in the second stage. Because they use two different procedures, these technologies generate proposals more slowly, have more complex architectures, and lack global context. In single-stage detectors, semantic objects are classified and identified by dense sampling. In order to locate items, they use predetermined boxes or keypoints having different sizes and dimensions.

2.1 Two stage detectors

1) R-CNN: Using CNNs to increase detection performance is possible with the first study of region-based convolutional neural networks (R-CNNs) [11, 12]. This study's authors provide a scalable and simple detection technique that enhances mean average precision (mAP) by greater than 30% over the former best result, which produced a mAP of 53.3%. When tagged training data is minimal, supervised pre-training for a supplementary structured form accompanied by domain-specific refinement yields a noteworthy performance improvement. This work [13] addresses the problem of producing plausible object positions for use in object recognition. The researchers recommend selective search, which combines the advantages of segmentation and comprehensive search The image structure, like segmentation, guides our sampling technique. They seek to capture all conceivable object positions, similar to an exhaustive search. Instead of depending on a single method to generate believable object placements, authors broaden their search and use a variety of complementary picture partitioning to deal with as many image scenarios as is practical. The trained, class-specific Support Vector Machines (SVMs) are then fed the feature vectors to compute confidence scores, according to the authors [8]. Despite being slow, time- and space-consuming, R-CNN introduced a new era in object detection [14]. Even when some computations were shared, the training process was difficult and required days to complete on tiny datasets.

2) Fast R-CNN: Girshick replaced the SPP-pyramidal net's structure of pooling layers with a unique spatial bin termed the RoI pooling layer. The researchers used variations of the current latest pre-trained models as a foundation, such as [15, 16]. In a one-step using stochastic gradient descent (SGD), a mini-batch of two photos was used to train the network. Back-propagation distributed computations between the two pictures' ROIs, enabling the network to converge faster.

3) Faster R-CNN: An RPN (region proposal network) with fully twisted networks accepts any input image and returns windows that can be used [17, 18]. All of these windows are associated with an objectness score, which indicates the likelihood of an object appears. RPN incorporates Anchor boxes, as opposed to its predecessors [19], which employed image pyramids to deal with object size variation. It regressed over many bounding boxes with varying aspect ratios to localize an object.

4) FPN: The authors created a Feature pyramid network (FPN) [20], a DCNN with an inherent multi-scale, pyramidal hierarchy that may be used to construct feature pyramids at a reasonable price. This algorithm accepts any image size as input and outputs attribute maps of the same size at different levels. This approach has an extensive range of applications. Here Faster R-CNN is based on ResNet-101. FPN has the potential to provide high-level semantics at all sizes, minimizing detection error rates.

5) Mask R-CNN: Mask R-CNN [21] is an approach that improves Faster R-CNN by introducing an object mask prediction branch alongside the conventional bounding box detection branch. Mask R-CNN is easy to train and requires less overhead than Faster R-CNN, which works at 5 frames per second. On every objective, Mask R-CNN exceeds all previous single-model entrants. Mask R-CNN training is the same as faster R-CNN training. Mask R-CNN surpassed existing single-model designs while also introducing instance segmentation with minimum overhead calculations.

6) SPP Net: SPP-net just shifted CNN's convolution layers afore the object proposals module and added a pooling layer, which made the network size/aspect ratio independent and reduced calculations. Spatial Pyramid Pooling (SPP) is proposed as a method of analyzing photos regardless of size or aspect ratio [15] because they are reserved for technical editing by editors.

2.2 One stage detector

Figure 4. YOLO timeline

1) YOLO, YOLO v2, YOLO v3, YOLO v4, YOLO v5: The YOLO (You Only Look Once) method was the first regression-based technique, and it was put forth by authors [22] in 2016. YOLO subsequent versions are depicted in Figure 4. In an end-to-end neural network, it forecasted the coordinates of the bounding boxes while also categorizing the items. Despite the fact that YOLO permitted real-time object

detection, it was still hard to identify small-sized objects, and the bounding box coordinate inaccuracy was substantial.

Authors [23] presented the YOLOv2 approach, which is more precise and faster than the YOLO method. Although this method partially utilised the multi-scale region features and continued to use the Darknet19 backbone network, which had poor feature extraction performance, it was limited in its ability to further increase detection accuracy. To improve the detection accuracy of existing systems, the deep residual network (ResNet) was used as the core network in DSSD (Deconvolutional Single Shot Detector) [24] and YOLOv3 [25, 26]. However, the detection speed of these approaches is greatly hampered by the more complex network. The authors [27] provide a face mask recognition and standard wear detection algorithm built on a modified YOLO-v4 to address the problems posed by the challenging environment, such as low accuracy, low real-time performance, poor resilience, andothers. The YOLOv5 model is available in five different sizes: nano, small, medium, large, andextra large. The dataset determines the type of model. In addition, with version 6.0, the frivolous model of the YOLOv5 model is out, with a bettered inference speed of 1666 fps [28, 29].

2) SSD: SSD is a rapid single-shot multi-box detector for multiple classes developed by the authors [30]. It constructs a unified detector framework that is as quick as YOLO and as precise as Faster-RCNN. SSD's architecture incorporates the regression concept from the YOLO model as well as the anchoring method from the Faster R-algorithm CNN's.

3) RetinaNet: A distinct Focal Loss concentrates drill on a small set of difficult situations, avoiding the detector from being overloaded by a significant number of easy negatives during training. To assess the efficacy of loss [31], the authors created and trained RetinaNet, a basic dense detector. If trained with the focus loss, the results show that RetinaNet could match the speed of prior one-stage detectors while beating all current detectors in terms of precision. The researchers [32] proposed that for the trivial Retina NetmAP-FLOPs trade-off, the heaviest bottleneck layer be lowered. The suggested solution consistently improves the mAP-FLOPs trade offline with a linear downfall trend, whereas the input image scaling method reduces more exponentially. The proposed strategy improves mAP by 0.1 percent at 1.15x FLOPs reduction, 0.2% at 1.15x FLOPs reduction, and 0.3% at 1.8x FLOPs reduction.

4) Refinenet: The approach has been shown in the study [33] to knowingly improve efficiency on a wide variability of datasets, scene settings, and camera viewpoints, resulting in superior quality object boxes at a low computational cost. The design, in particular, achieves considerable performance gains while asserting a fast run-time speed. It is shown that iterative refining influences on future vision tasks like object tracking in both the image and ground planes. As per authors in the study [11] in order to localize and segment objects, convolutional neural networks can be applied to bottom-up region proposals and segment objects and when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost.

Figure 5 depicts number of publications in object detection.

Few of the object detections methods in videos are discussed below.

A semi-automatic algorithm is presented by Park et al. in the study [34]. Intra-frame object extraction and inter-frame object tracking are the two procedures involved.

Homogeneous region segmentation is used in intra-frame object extraction to reduce the need for human interaction, while 1-D projected motion estimate is used in inter-frame object tracking to speed up processing. Additionally, a more effective flooding technique for the traditional water shed algorithm is suggested

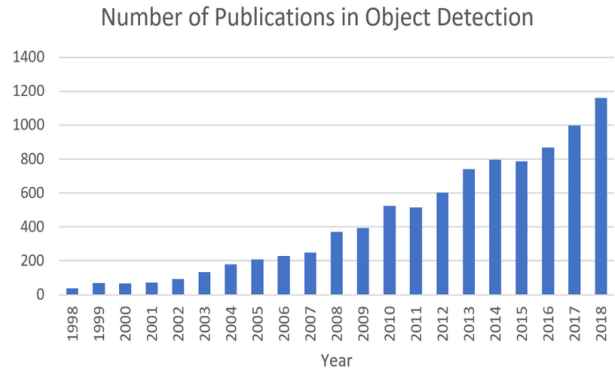

Figure 5. 1998-2018 object detection (Data from google scholar advanced search)

A technique for unsupervised segmentation in both photos and videos is proposed in the study [35] by Deng and Manjunath. The JSEG algorithm uses color-texture regions in both video and picture data to operate. The proposed approach consists of two steps: colour quantization and spatial segmentation. The first stage involves quantizing the image's colours into a number of representative classes that can be used to distinguish distinct parts of the picture. Then the labels for the matching colour classes are placed in place of the pixels. They are now left with an image's class map. Authots create the "J-image" by applying the suggested "excellent" segmentation criterion to the class-map using local windows, where high values represent potential color-texture area boundaries and low values represent interiors. The image is then segmented using a region-growing technique based on multi-scale J-images. In the case of video, a second region tracking approach is utilized in addition to the previously indicated method to ensure consistency in results even in the case of nonrigid object motion. This method's drawback is that the algorithm oversegments each colour when a smooth colour change (such going from red to orange) takes place. However, even if we manage to solve this issue by looking for smooth transitions, we still run into the issue that a smooth transition might not necessarily signify a homogeneous zone. In the case of video, an error created in one frame affects the following frames.

A technique for automatically segmenting moving objects in MPEG-4 films is presented by Tsaig and Averbuch (2002) in the research [36]. Each frame of an image sequence is divided into video object planes by MPEG-4 (VOPs). In the scene, each VOP represents a single moving object. The fundamental step is to categorize areas based on motion information into foreground or background. The formulation of the segmentation problem is the detection of moving items against a static background. By using an eight parameter perspective motion model, camera motion is adjusted. The watershed algorithm is used to first obtain a spatial split. The spatial gradient in the colour space is then estimated using Canny's gradient. Region matching in a hierarchical framework is used to estimate the mobility of each foreground region. The regions are initially classified using a Markov

random field model that incorporates the predicted motion vectors and is optimized using highest confidence first (HCF). Information from the preceding frame is included in MRF. A dynamic memory is used in the last stage to guarantee the segmentation process' temporal coherency.

An strategy for automatically counting the number of objects and extracting independently moving video objects from MPEG-4 films is presented by Venkatesh Babu et al. [37]. Since compressed MPEG films have sparse motion vectors (i.e., one motion vector per macro-block), a technique to enhance the motion information from a few frames on each side of the present frame is suggested. A motion vector is assigned to each pixel in the frame by the use of median filter interpolation. egmentation is then completed. The Expectation Maximization (EM) approach is utilized since there aren't enough data points to estimate the motion parameters. There is a suggested algorithm for calculating the number of motion models. Following initial segmentation, tracking produces Video Object Planes (VOPs). The pixels at the edges are assigned to the appropriate VO during the edge refining phase, which takes place after the VOs.

Segmenting a video sequence into the objects is the goal of Mezaris et al. [38]. This algorithm is divided into three phases: First-frame segmentation using colour, motion, and position data, followed by a temporal tracking process, and then a region-merging approach. The K-means-with-connectivity-constraint algorithm is used for segmentation. Utilizing a Bayes classifier, tracking is carried out. Reassigning modified pixels to existing regions and handling new regions added to the sequence are both done via rule-based processing. Instead of using motion at the frame level, region merging is done using a trajectory-based method. One benefit is that it successfully tracks moving objects or new things that emerge on the scene.

An incremental Log- Euclidean Riemannian subspace learning approach is proposed by Hu et al. in the research [39]. The log-Euclidean Riemannian metric is used to first convert the co-variance matrices of the image features into a vector space. A log-Euclidean block-division appearance model captures both global and local spatial layout information. Particle filtering-based Bayesian state inference is utilized for both single-object and multi-object tracking with occlusion reasoning. The log-Euclidean block division appearance model is incrementally updated to incorporate changes in object appearance.

An approach that uses frame-by-frame target detection results as the input is suggested by Huang et al. in the study [40]. The first batch of target tracklets (tracking fragments) is produced using a dual-threshold method that is cautious. Only trustworthy detection replies are linked as a result. Until further data is gathered, associations that are in doubt are postponed. The Maximum A Posteriori (MAP) problem, which in addition to initializing, tracking, and terminating them, hypothesizes a trajectory of being a false alarm, is achieved via hierarchical association using many passes. This problem is resolved using the Hungarian algorithm. These associations' subsequent ranking is viewed as a bagranking problem that can be solved by a bag-ranking boosting algorithm. In order to simplify the optimization of the released objective loss function, this paper also provides a soft max/min.

In the research, Farah et al. [41] describe a reliable tracking technique to remove a rodent from a frame in an uncontrolled laboratory environment. It operates in two phases: First, the target is crudely tracked by combining three weak traits. The tracker's boundaries are then modified to remove the rodent. Overlapping Histograms of Intensity (OHI), a new segmentation methodology, and edglet-based built pulses are a few of the recently introduced methodologies. Edge fragments known as edglets are broken edges. To coarsely localize the target, a sliding window approach is employed.

Two significant contributions of Chien et al. [42] are found in: First, a multi-background model video object segmentation threshold choice technique is proposed. Then, diffusion distance measuring colour histogram similarity and motion cue from video object segmentation are combined to create a video object tracking framework based on particle filter with probability function. This framework is capable of handling abrupt changes in lighting, background noise, and non-rigid moving objects. The ideal threshold value for segmentation is chosen by the threshold decision algorithm. For improved tracking of non-rigid objects, a color-based histogram is added. In order to reduce computing complexity, a 1-D colour histogram is employed instead of a 3-D colour histogram.

3. COMPARATIVE ANALYSIS

Tables 1 and 2 outline the properties of Two-stage and One-stage object detection models.

In Table 1, we present the features of models detecting objects in two stages. By considering the size of the input image, region proposal method, optimization technique, and loss function, it gives a brief explanation of the object detector.

A model for detecting objects with one stage is presented in Table 2. It gives a brief introduction of every object detector with the parameter input size of the image, optimization technique and loss functions. Analysis of object detectors is on fixed input size of image with SGD optimization technique. Here, sum error, sum square error, binary cross-entropy, confidence loss and Logits loss functions with binary cross-entropy loss functions are used.

Object detectors consider the size of input to be either fixed or arbitrary. The difference between predicted and the expected output is measured by using loss functions such as Bounding box, regressor loss, hinge loss categorization loss, etc. The Comparative analysis highlights Region Proposal Method used by object detectors. The bounding box loss function is used in each object detector.

4. DISCUSSION

Studying the state-of-the-art system in literature survey we found the following shortcomings separately for two stage and single stage detectors.

1) Shortcomings of Two stage detectors: Because of the vast amount of space and time necessary for RCNN training, it is costly. Because features are recovered for each image region, image area extraction is a difficult task. Fast RCNN is slow due to selective search, and region proposal calculation is a bottleneck in SPP net. For real-time applications, faster RCNN training is inefficient, and performance for small and multiscale objects is inadequate. To address multiscale problems in FPN, a pyramid representation is required, which influences the performance of object detection. When it comes to real-time applications, the detection speed of mask RCNN is slow.

Detecting things in the real-time video is becoming increasingly important. It has certain challenges, such as low image quality, which leads to poor accuracy. Different video detectors are planned with temporal variables in mind to link objects over multiple frames and understand the object's behaviors. For spatial-temporal suggestions, tubelet networks are used in video detectors for pre-processing. Deep feature flow, flow-guided feature aggregation, STMNs and flow-guided feature aggregation are some of the techniques used for the same.

Table 1. Comparative analysis of two stage detectors

Model	Year	Input Size of image	Region Proposal Method	Optimization technique	Loss/ Cost Function
RCNN [11]	2014	Fixed	Selective Search	SGD, BP	Bounding box regressor loss, Hinge loss
SPP-NET [15]	2014	Arbitrary	Selective Search	SGD	Bounding box Regressor loss, Hinge loss
Fast RCNN [14]	2014	Arbitrary	Selective Search	SGD	Bounding box Regression loss categorization loss
Faster RCNN [18]	2015	Arbitrary	RPN	SGD	Bounding Box regression loss, Categorization Loss
FPN [20]	2017	Arbitrary	RPN	Synchronised SGD	Bounding box regression loss categorization Loss
Mask RCNN [21]	2017	Arbitrary	RPN	SGD	Categorization loss, Mask loss, Bounding box regression loss

Table 2. Comparative analysis of one stage detectors

Model	Research Year	Input Size of Image	Optimization technique	Loss/tion	Cost	unc-
YOLO [22]	2016	Fixed	SGD	Sum Error		
YOLOv2 [23]	2017	Fixed	SGD	Sum Squared Error		
YOLOv3 [25]	2018	Fixed	SGD	Binary entropy	cross	
YOLOv4 [43]	2020	Fixed	SGD	Binary entropy	cross	
YOLOv5 [28, 29]	2020	Fixed	SGD	Logits function with binary cross entropy	Loss	
SSD [30]	2016	Fixed	SGD	Confidence loss		

2) Shortcomings of one stage detectors: Low-resolution objects are difficult to localize in Yolo v1 and v2. Without anchor boxes, they cannot anticipate more than one box for a given region. There may be a problem with using YOLOv3 to train niche models if it is difficult to collect large datasets, and it may also be ineffective for recognizing small objects. Only higher resolution layers in SSD allow for the detection of small objects, but these layers also contain low-level features like edges that are useless for categorization.

5. CONCLUSION AND FUTURE SCOPE

Despite tremendous advancements in object detection over the previous years, most of the detectors are still a long way from the best performance. As its real-world applications grow, there will be a greater need for compact algorithms that can be used on mobile and embedded devices. This industry has drawn more attention, but the issue has not yet been resolved. In this article, we discussed how single-stage and two-stage detectors descended from one another. Two-stage detectors are slower and can be used for real-time applications, although typically more accurate. However, one stage detectors have grown similarly precise and substantially faster in recent years. In the future, lightweight and much more precise object detectors for video can be produced.

REFERENCES

[1] Kaur, J., Singh, W. (2022). Tools, techniques, datasets and application areas for object detection in an image: A review. Multimedia Tools and Applications, 81: 38297-38351.https://doi.org/10.1007/s11042-022-13153-y

[2] Jiao, L.C., Zhang, R.H., Liu, F., Yang, S.Y., Hou, B., Li, L.L., Tang, X. (2022). New generation deep learning for video object detection: A survey. IEEE Transactions on Neural Networks and Learning Systems, 33(8): 3195-3215. https://doi.org/10.1109/TNNLS.2021.3053249

[3] Kaur R., Singh S. (2022). A comprehensive review of object detection with deep learning. Digital Signal Processing, 132. https://doi.org/10.1016/j.dsp.2022.103812

[4] Zou, Z., Shi, Z., Guo, Y., Ye, J. (2019). Object detection in 20 years: A survey. ArXiv, abs/1905.05055. https://doi.org/10.48550/arXiv.1905.05055

[5] Guan, Y.R., Aamir, M., Hu, Z.H., Dayo, Z.A., Rahman, Z., Abro, W.A., Soothar, P. (2021). An object detection framework based on deep features and high-quality object locations. Traitement du Signal, 38(3): 719-730.https://doi.org/10.18280/ts.380319

[6] Zaidi, S.S.A., Ansari, M.S., Aslam, A., Kanwal, N., Asghar, M., Lee, B. (2022). A Survey of modern deep learning based object detection models. Digital Signal Processing,126: 103514. https://doi.org/10.1016/j.dsp.2022.103514

[7] Galvez, R.L., Bandala, A.A., Dadios, E.P., Vicerra, R.R.P., Maningo, J.M.Z. (2018). Object detection using convolutional neural networks. TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea, pp. 2023- 2027. https://doi.org/10.1109/TENCON.2018.8650517

[8] Liu, L., Ouyang, W., Wang, X. (2020). Deep learning for generic object detection: A survey. International Journal of Computer Vision, 128: 261-318. https://doi.org/10.1007/s11263-019-01247-4

[9] Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al- Shamma, O., Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8(1): 1-74. https://doi.org/10.1186/s40537-021-00444-8

[10] Liu, S., Zhou, H., Li, C., Wang, S. (2020). Analysis of anchor-based and anchor-free object detection methods based on deep learning. 2020 IEEE International Conference on Mechatronics and Automation (ICMA), Beijing, China, pp. 1058-1065. https://doi.org/10.1109/ICMA49215.2020.9233610

[11] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2014.81

[12] Zhao, Z.Q., Zheng, P., Xu, S.T., Wu, X. (2019). Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 30(11), 3212-3232. https://doi.org/10.1109/TNNLS.2018.2876865

[13] Uijlings, J.R.R., Sande, K.E.A., Gevers, T., Smeulders, A.W.M. (2013). Selective search for object recognition. International Journal of Computer Vision, 104(2): 154-171. https://doi.org/10.1007/s11263-013-0620-5

[14] Girshick, R. (2015). Fast R-CNN. IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[15] He, K., Zhang, X., Ren, S., Sun J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

[16] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. https://ui.adsabs.harvard.edu/link_gateway/2014arXiv1409.

[17] Shelhamer, E., Long, J., Darrell, T. (2017). Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4). https://doi.org/10.1109/TPAMI.2016.2572683

[18] Ren, S.Q., He, K.M., Girshick, R., Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems -1 (NIPS'15). MIT Press, Cambridge, MA, USA, 91-99. https://doi.org/10.1109/tpami.2016.2577031

[19] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), NV, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[20] Lin, T.Y., Dolla´r, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyra-mid networks for object detection. arXiv:1612.03144, pp. 2117-2125. https://doi.org/10.48550/arXiv.1612.03144

[21] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask R-CNN. IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2980-2988. https://doi.org/10.1109/ICCV.2017.322

[22] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 779-788.

[23] Redmon, J., Farhadi, A. (2017). YOLO9000: Better, faster, stronger. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 6517-6525. https://doi.org/10.1109/CVPR.2017.690

[24] Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C. (2017). DSSD: Deconvolutional single shot detector. arXiv:1701.06659. https://doi.org/10.48550/arXiv.1701.06659

[25] Redmon, J., Farhadi, A. (2018). YOLOv3: An incremental improvement. ArXiv180402767. https://doi.org/10.48550/arXiv.1804.02767

[26] Padmanabula, S.S., Puvvada, R.C., Sistla, V., Kolli, V.K.K. (2020). Object detection using stacked YOLOv3. Ingénierie des Systèmes d'Information, 25(5): 691-697. https://doi.org/10.18280/isi.250517

[27] Yu, J., Wei Z. (2021). Face mask wearing detection algorithm based on improved YOLO-v4. Sensors, 21(9): 3263. https://doi.org/10.3390/s21093263

[28] Thuan, D. (2021). Evolution of yolo algorithm and yolov5: The state-of-the-art object detection algorithm. Thesis. https://urn.fi/URN:NBN:fi:amk-202103042892, accessed on Dec. 12, 2022

[29] Solawetz, J. (2020). YOLOv5 new version - improvements and evaluation. https://blog.roboflow.com/yolov5 -improvements -and -evaluation/, accessed on Nov. 1, 2022.

[30] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016). SSD: Single shot multi box detector. arXiv: 1512.02325. https://doi.org/10.1007/978-3-319-46448-0_2

[31] Lin, T., Goyal, P., Girshick, R., He, K., Dollar, P. (2020). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 318-327. https://doi.org/10.48550/arXiv.1708.02002

[32] Li, Y., Dua, A., Ren, F. (2020). Light-weight retinanet for object detection on edge devices. In IEEE 6th World Forum on Internet of Things (WF-IoT), LA, USA, pp. 1-6. https://doi.org/10.1109/WF-IoT48130.2020.9221150

[33] Rajaram, R.N., Ohn-Bar, E., Trivedi., M.M. (2016). Refinenet: Refining object detectors for autonomous driving. IEEE Transactions on Intelligent Vehicles 1(4): 358-368. https://doi.org/10.1109/TIV.2017.2695896

[34] Park, D.K., Yoon, H.S., Won, C.S. (2000). Fast object tracking in digital video. IEEE Transactions on Consumer Electronics, 46(3): 785-790. https://doi.org/10.1109/30.883448

[35] Deng, Y., Manjunath, B.S. (2001). Unsupervised segmentation of color-texture regions in images and video. IEEE Transactions on Pattern Analysis and

Machine Intelligence, 23(8): 800-810. https://doi.org/10.1109/34.946985
[36] Tsaig, Y., Averbuch, A. (2002). Automatic segmentation of moving objects in video sequences: A region labeling approach. IEEE Transactions on Circuits and Systems for Video Technology, 12(7): 597-612. https://doi.org/10.1109/TCSVT.2002.800513
[37] Venkatesh Babu, R., Ramakrishnan, K., Srinivasan, S. (2004). Video object segmentation: A compressed domain approach. IEEE Transactions on Circuits and Systems for Video Technology, 14(4): 462-474.https://doi.org/10.1109/TCSVT.2004.825536
[38] Mezaris, V., Kompatsiaris, I., Strintzis, M.G. (2004). Video object segmentation using bayes-based temporal tracking and trajectory-based region merging. IEEE Transactions on Circuits and Systems for Video Technology, 14(6): 782-795. https://doi.org/10.1109/TCSVT.2004.828341
[39] Hu, W.M., Li, X., Luo, W.H., Zhang, X.Q., Stephen, M., Zhang, Z.F. (2012). Single and multiple object tracking using Log- Euclidean Riemannian subspace and block-division appearance model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(12): 2420-2440. https://doi.org/10.1109/TPAMI.2012.42
[40] Huang, C., Li, Y., Nevatia, R. (2013). Multiple target tracking by learning-based hierarchical association of detection responses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(4): 898-910. https://doi.org/10.1109/TPAMI.2012.159
[41] Farah, R., Langlois, J.M.P., Bilodeau, G.A. (2013). Catching a rat by its Edglets. IEEE Transactions on Image Processing, 22(2): 668-678. https://doi.org/10.1109/TIP.2012.2221726
[42] Chien, S.Y., Chan, W.K., Tseng, Y.H., Chen, H.Y. (2013). Video object segmentation and tracking framework with improved threshold decision and diffusion distance. IEEE Transactions on Circuits and Systems for Video Technology, 23(6): 921-934. https://doi.org/10.1109/TCSVT.2013.2242595
[43] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint, arXiv:2004.10934. https://doi.org/10.48550/arXiv.2004.10934