International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 3S, 2025
https://www.theaspd.com/ijes.php

A Review of the Single-Stage vs. Two-Stage Detectors Algorithm: Comprehensive Insights into Object Detection

Sara A.Alhashmi[1]* , Adel Al-azawi[2]
[1]Scicomphd222305@uodiyala.edu.iq, [2]adil_alazzawi@uodiyala.edu.iq
[1,2]University of Diyala, College of Science, Computer Science Department, Iraq

Abstract

Object Detection is the most common and tough issue in the field of computer vision. Deep learning has advanced enormously in the last 10 years; it has encouraged researchers to use very basic deep models to explore the effective improvement of object detection and correlated tasks. These tasks include classification, localization, and segmentation. One can broadly categorize object detectors into two categories: two-stage and single-stage detectors. Two-stage detectors take most of their support from designs that first propose regions. On the other hand, single-stage detectors focus all their attention on the use of simple settings to propose all regions at once for object detection. Single-stage detectors, however, are faster in terms of computation time. For accuracy, the YOLO algorithm and its variants sometimes outperform two-stage detectors, which is largely influenced by the Mean Average Precision (mAP) metric. YOLO is popular because it is very fast in processing rather than accurate in its detection. This paper pushes forward full-fledged one-stage object recognizers many incarnations of YOLO two-stage recognizers various flavors of YOLO and some alternative approaches in the realm of object detection.

Keywords: Object detection, Computer vision, CNN, YOLO, and Deep learning

Article history:

1. INTRODUCTION

The identification of visual features associated with one specific sector visualized within photographic images is known as object detection, e.g., humans, animals, or automobiles. The task of object detection forms some of the basic elements responsible for supporting all advanced tasks in the field of computer vision. Object detection is a base supply method of performing all later computer vision works which include instance segmentation [1], image labeling [2], and object tracking [3]. Object detection research can be divided into two categories: "general object detection" and "recognition applications." While the former general object detection tries to apply a single framework to mimic the ability of humans to perceive and understand, in this case, it will differentiate objects of several types [4]; the latter concentrate on specific situations concerning pedestrian, face, and text recognition. One of today's most interesting hot topics is object recognition, applied in a broad range of practical applications: e.g., autonomous driving, robotic vision, video surveillance, etc. People often claim that object recognition seems easy. Humans acquire the ability to recognize familiar objects typically within the first 3 or 4 months of life. Teaching computers this task was not trivial until fairly recently in this century. The earliest models for performing object recognition were based on hand-building feature extractors. One notable example is the Viola-Jones detector [5], itself based on the HOG (Histogram of Oriented Gradients) detector [6]. Such models were relatively slow, inaccurate, and performed only averagely on untrained datasets. The reappearance of CNNs and other deep-learning approaches for image categorization radically transformed the field of visual recognition. The use of AlexNet [7]in ILSVRC 2012 stimulated further inquiry into its applicability for computer vision. Object detection is now so widespread in applications as diverse as self-driving cars, verification processing, security, and medical applications that it has, until recently, not been markedly growing within a framing that has resulted in such

* Corresponding author: Scicomphd222305@uodiyala.edu.iq

dramatic growth. From the progression of object categorization, that is the recognition of objects in an image, object detection naturally follows. Its primary task is to determine all of the occurrences of a pre-specified class and to indicate their approximate locations within the image. The bounding boxes are aligned with the coordinate axes. It has to be able to identify every instance of a particular class and place a box around it. This has been described as supervised learning. State-of-the-art object recognition systems are developed using large repositories of digitized images and tested against a variety of popular metrics. This study is part of the ongoing debate on deep-learning-based object detection. Another paper could also summarize that, "Computer vision has had incredible growth over the last decade. And yet it remains extremely hard." Problems exist w.r.t. the network concerning its maturity for application:

• Intra-class variety within the same item is prevalent in nature. These disparities may be related to occlusion, illumination, location, and perspective[8]. These impacts that are unrestrained by the law may have a considerable influence on the look of an item. Objects are meant to suffer non-intuitive deformations, rotations, scaling, or blurring. The setting of anything may be subtle and so difficult to bring out.

• The varied variety of item types available for categorisation makes it a challenging challenge. More effectiveannotated data is need, which is not simple to acquire by. The training of detectors utilising a decreasing number of instances is still an unsolved scientific challenge.

• Current methodologies require a significant amount of computer resources to generate reliable detection results. Good object recognizers are very important when it comes to the progress of computer vision, considering the fact that mobile and edge devices are becoming popular more and more.

2. OBJECT DETECTION

Object detection is the process of identifying and arranging items in an image or video. It has lately grown increasingly relevant due of its vast variety of applications [9][37]. Object recognition is crucial and challenging in the computer area of vision. In the past decade, the increased development of deep learning has led to a lot of interest in the field, and the purpose of optimizing the effectiveness of object recognition and related tasks, including the classification of objects, their location, and the segmentation of objects using basic deep models. On the other hand, single-stage detectors argue for a complete approach to space that will provide a simple design for the future in a single operation. The efficiency of any item detector is defined by its precision and the length of time it takes to infer. In terms of detector accuracy, two-stage detectors are often more accurate than single-stage detectors. However, the time needed to form an inference about single-stage detectors is larger than that of comparable detectors. Currently, object recognition is applied in autonomous automobiles, identification verification, security standards, and medicinal applications. Recently, its rise has been enormous because of the quick development of new technologies[10].

This project is designed as a lesson on object recognition that is followed by explanations of datasets like MSCOCO and Flickr. Other techniques of deep learning are also explored.

3. THE STAGES OF OBJECT DETECTION

Supervised machine learning covers Sub-Issues: (i) Of Regression And (ii) Of Classification. These two are demonstrated in Figure 1..

Fig 1. Some samples of photographs and captions from MS-COCO, Flickr8k

However, the semantics of image labeling is more akin to traditional problems with categorization. After recognizing an object in an image, the following step is to find the object as many times as possible. A sufficiently deep neural network should be implemented to cover the object with a bounding box. Normally the object detection problem requires feature computation followed by classification and/or localization; the two-stage object detector is a variant of this theme as shown in Table 2. The first step is to generate the region of interest (RoI)

through a region proposal network (RPN) and the second step reveals the target region along with its bounding box. The prime consideration during the learning process is therefore to gather possible domain prompts with diverse methodologies that may also involve negative sampling of prompts. Ones of its classes which bear popular models include region-based convolutional neural networks (RCNNs), Fast RCNNs, and Faster RCNNs. Single-stage object detectors have their specific structures for detecting objects in a single stage. These detectors may produce bounding boxes covering underground objects and certain classes of confidence, all the dimensions of an image are analyzed in a single shot, as illustrated in Figure 2.

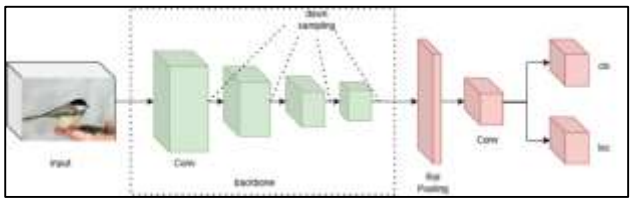

Fig 2. The typical form of a single-stage object detector.

Nonetheless, two-stage object detectors have greater capacity than the one-stage object detector. This is because the former only identifies items at plausible places, while the latter only works in certain regions. One the other hand, liability for the development of YOLO and its sequels can be credited with a significant increase in recognition of single-stage approaches because localization was perceived as a regression issue using deep neural networks. As shown in Figure 3, YOLO is not the first approach to have adopted a single detector (SSD) for the loca.

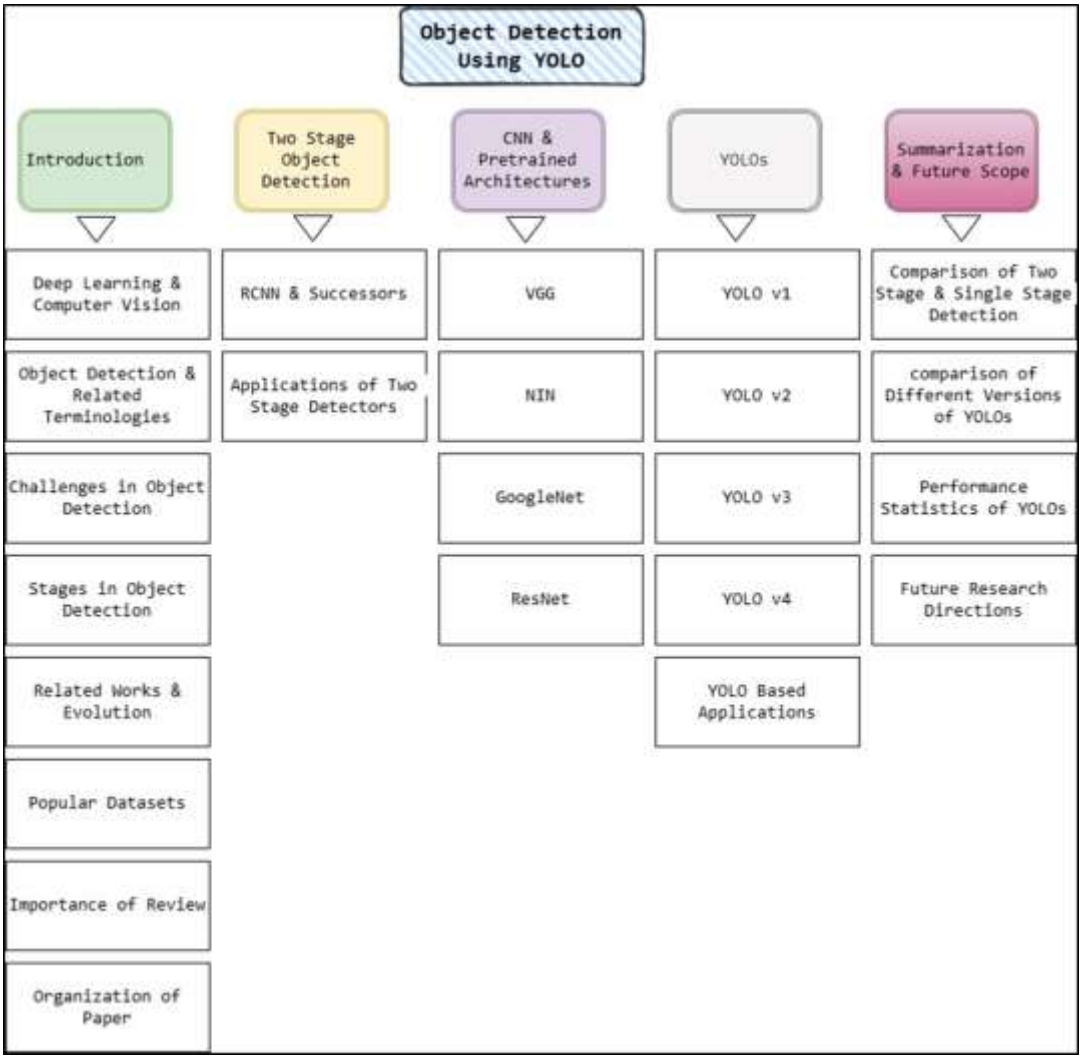

Fig 3. Object detection using the YOLO technique

To date, several more methods are formulated including single-shot detectors (SSD) [11], deconvolutional single-shot detectors (DSSD) [12], RetinaNet [13], M2Det [14], and RefineDet++ [15], employing a single-stage approach of object detection. Two-stage detectors prove to be flexible and powerful and are generally preferable than single-stage detectors. YOLO successfully competed with the two-stage and earlier single-stage detectors with respect to accuracy and speed of approximation. Therefore, because of its very simple conceptualization, this method is one of the favorites in implementation.

4. INTRODUCTION TO DEEP LEARNING

Deep learning features a network of interconnected sensors based on linear regression plus a few activation functions. This is the big heart behind this technique — the model for plain statistical regression WTX+b. The main difference is that in deep learning, many more parts of the brain are activated, as opposed to the single region in classic statistical learning, such as that used in linear regression. A decision-maker is also described as a sensing unit or neuron. Such a term greatly helps in naming these neural parts "neural networks." Another difference is that between input and output, there are many layers. There could be anything like several hundred to a few thousand neurons in one layer. The nodes from input to output referred to as hidden nodes, and the layers between the input and output layer are known as the hidden layers. Standard machine learning classifiers need assumptions that are hard to explain. Yet deep neural networks can figure them out themselves, which makes them ideal for investigation within complex relationships [44].

The idea of using convolutional neural networks started in the 80s. At that time the cat cortex was taken as a model. LeNet-5 is the typical implementation of a CNN. The system's performance on the MNIST dataset was a mere 0.9% accuracy. It is oftentimes applied at financial institutions for deciphering handwritten checks, though it is not able to work with large images. There has since been devised very good software that takes advantage of GPUs to crack the Image Net challenge. GPUs have played a significant role in this evolution, leading to a resurgence of interest in CNNs. One of the issues with deep neural networks is very high training time due to too many hidden nodes in the network. Today, fast GPUs brought a decline to this limitation. The most famous applications of convolutional neural networks are in image recognition and speech processing. Convolutional neural networks are similar to biological brain networks because of their net architecture and the way information flows through them. This can reduce the model complexity dramatically and the number of weights used in the network [45]. All in all, it's more effective for processing high-dimensional images since the whole network has the capacity to take an image as input, at once, avoiding the feature extraction and reconstructions that had to be done in the past technologies. When images are identified using convolutional neural networks, these networks show a surprising degree of invariance for visual distortion which is fairly common (scale, tilt, translation, etc.) [46].

4.1 Conventional Layer

The convolutional layer applies a filter, which is also known as a kernel, of any dimension— height, width, and number of channels— over the input image or most recently produced feature map of that input through convolution [45]. The equation for the feature map F_K is:

$$F_K = \left(\sum_k^K W_{ki} * X_i\right) + b_k \qquad 1$$

Equation (1) gives the formulation of the sub-kernel W_{ki} of the 1^{st} channel with the input X_i of the ith channel. The operator * denotes a two-term convolution operator, with the first term, b_k, representing its distortion. Every kernel within the family of Wki shall have $N \times N \times L$ weights because the convolutional layer has L input channels, X has M×M×L values, and every kernel is W_ki having N×N×L weights. The total number of parameters in a convolution block with K feature maps is K×M×M×L..

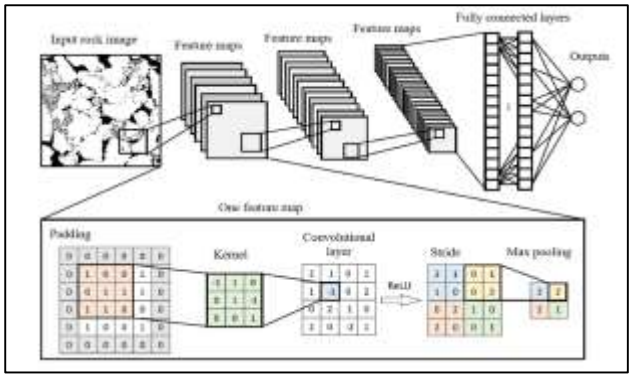

Fig 4. Convolutional Neural Network Model

5. OBJECT DETECTION ALGORITHM

As one example of the aesthetic advantages of deep learning, the present degree of object identification has a contrast to 20 years ago that illustrates the "wisdom of the cold weapon era". Many of the early recognition algorithms for things were constructed using hand-crafted characteristics. Because there was a lack of excellent visual representation at the time, individuals were driven to build complicated feature combinations and different techniques of optimization in order to optimise the exploitation of limited computer resources.

P. Viola and M. Jones launched the pioneering of real-time face recognition that removed any constraints linked with skin color[5] (such as segmentation by color). The detector employed a 700 MHz Pentium III CPU and was many times more quick than other current approaches, while keeping the same degree of detection accuracy. The technique of detection is thereafter referred to as the "Viola-Jones (VJ) detector", since the authors committed the detector to the identification of its substantial influence. The VJ detector offers a systematic approach for detection that uses a sliding window to investigate all potential scales and locations in the picture and decides whether a face is there. Despite the seemingly simple simplicity of the technique, the computations that underpin it are significantly more than the capability of the technology throughout the time. The VJ detector accelerates the speed of detection by integrating three main approaches: "integral image", "feature selection", and " Detection cascade".

- Integral image: Integral image computes a feature which makes box filtering or convolution very efficient. Much like other object recognition algorithms of its time [18], Haar wavelets are employed as features to describe the image in the VJ detector. Integral image is the computation cost of each window in the VJ detector, irrespective of the size of the window.
- Feature selection: The authors chose the Adaboost algorithm from more than a dozen manually picked weak classifiers. [19] From a massive pool of random features (approximately 180,000 dimensions), a subset of features, very good at face finding, are chosen.
- **Detection cascades**:The VJ detector utilises a multiple-stage detection process dubbed a "detection cascade" that is meant to lessen the computational effort associated with background areas and concentrate on the face objects in order to maximise attention.

5.1. HOG Detector

A typical feature descriptor which is applied in computer vision for edge-based or pattern-based gradient descent in images is Histogram of Oriented Gradients (HOG). The first stage in building a HOG descriptor uses the grayscale image that makes the computation for gradient directions simpler. Normally byquantizing the changes in intensity in the Y and X axes with the help of the Sobel filter. This step gives the major gradients which delineate the boundaries in the image. Then, it computes the angle of the gradient for each pixel, which is the angle of the line with respect to the horizontal, and the magnitude of the gradient at pixels, typically between -90 and +90 degrees. The magnitude of the gradient relates to the intensity of each pixel; the orientation of the gradient relates to the orientation of the bin. This kind of directional binning happens inside what is termed a cell, which comes to be a small region of the image that is oriented in some specific way [20]. The Histogram of Oriented Gradients (HOG) feature description was first introduced by N. Dalal and B. Triggs in 2005. [21]. HOG constitutes a major break from the current approach of scale-invariant feature transformation [22] as well as shape context [23]. The

HOG description has been formulated to combine characteristic invariance in a linear detection setup (such as translation, scaling, and illumination effects) and object nonlinearity for distinguishing between different types of objects. The computation is on a dense grid of uniformly spaced cells with a form of local contrast normalization applied over "blocks" in order to enhance detectability. While HOG is indeed targeted at differentiating many types of objects, its major driving application is pedestrian detection. The HOG detector finds objects of different sizes by scanning the input image at various scales while keeping the size of the detection window constant. The HOG detector serves as the inspiration for other object detectors [24] and even computer vision throughout history.

5.2. CNN-based Two-stage Detectors

Two-stage detectors are mostly used because of the multiple-stage approach to detection. This can improve the detection accuracy, especially when detecting complex and highly crowded objects. Region proposed is the key phase of this approach, wherein objects are detected and classified or else modified within it[25]. The challenge in object detection was recognized after 2010, when the handcrafted feature performance began to degrade. R. Girshick reported that from 2010 to 2012, there was a slow progress with only minor gains in the development of integrated systems and little change to the previously successful approaches. Only by deep convolutional networks did 2012 mark a significant explosion in popularity. Modulo the question for whether or not object recognition is indeed suitable for deep convolutional networks because of their ability to acquire rather complex and important image information. R Girshick, etc., Object Recognition with CNN-Provided Regions (RCNN). ground to a halt in 2014 [26]. Since then, object detection has evolved rapidly. In the deep learning era, two fundamental approaches exist for object detection: "two-stage detection" and "single-stage detection." Critics of the former school say that recognition is a " coarse-to-fine" process, while champions of the latter contend that it is a "one-step" procedure.

5.3 Regions with Convolutional Neural Networks (RCNN)

Roughly speaking, what it does is first to take a set of proposals appropriate for the object that is being looked for (selective bounding boxes around objects-these are the candidates) as illustrated in Figure 3 [25]. Each proposal is then warped to 1,000 dimensions and then fed into a CNN model such as AlexNet [27], which has been pre-trained on the ImageNet database for feature extraction, to obtain features. Finally, a linear SVM detector is trained to detect whether an object is present in each region and to say the category or not. RCNN improved the mAP on VOC07 by a very large margin, jumping from 33.7% (DPM-v5 [28]) to 58.5%. This is because, in handwriting, the speed and precision with which the writing the phrases are great.

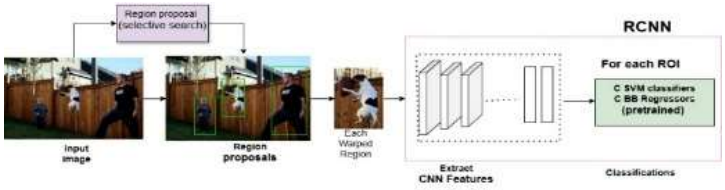

Fig 5. Two-step object detector, traditional networked RCNN

5.4. Spatial Pyramid Pooling Networks SPPNet

This style of Convolutional Neural Network is called the Spatial Pyramid Pooling Network (SPPNet). K. He et al. built it in 2014. [29] It was created to address the inefficiencies of normal CNNs in the handling of images of different dimensions. The goal of the network is to enhance object detection and improve image categorization. While standard CNNs resize all input images to a single dimension, which harms the image by losing spatial information and such information, normaCNNnt is due to the need for fixed-size input for fully connected layers, this obstacle is surmountable by a Spatial Pyramid Pooling (SPP) layer such that SPPNet can process images of different dimensions without resizing. The representation length of the layer is the same no matter the input image size and mixes features of different sizes since, for example, 1x1, 2x2, 4x4 grids can be combined. It significantly improves the performance of CNNs in object recognition and other tasks where spatial variance is critically important since SPPNet preserves a greater quantity of spatial information. Thus, this guarantees the fine-grained picture information to be preserved and, at the same time, even the computational effort cost for

processing all images at one resolution. SPPNet presents an important innovation in the deep learning space aimed at quality and efficiency improvements for the CNN-based models while working with images of different sizes. For example, the previous CNN model, AlexNet, required images at a size of 224x224 pixels [27]. The main breakthrough of SPPNet is the spatial pyramid pooling (SPP) layer that permits the CNN representation to be computed over the entire extent of an image or object of interest, at any scale. To avoid repeatedly computing convolutional features, the SPPNet can be used for image object recognition. SPPNet paragraphs map features for the entire image first. Then the detector learns to detect objects with fixed-length representations of sub-regions (regions of the image) without computation repeated across sub-regions. SPPNet is more than 20 times faster than R-CNN yet achieves mean average precision at the level of 59.2% on VOC07. The detection speed is therefore improved, but SPPNet has its own drawbacks as well: the first being the training is still multi-stepped and second only the layers that are newly appended will be updated by SPPNet and mostly ignores all the previous layers. A later iteration from the same group, Fast R-CNN [30], was published in the same year..

5.5 Fast Region-based Convolutional Neural Network (RCNN)

Fast R-CNN, presented by Ross Girshick in 2015 [30], marks a milestone in the field of object detection through improved speed and accuracy over its predecessor R-CNN. Object detection, the act of localizing and classifying objects in an image, is typically extremely computationally expensive in earlier systems. R-CNN requires bottom-up region proposals, applies a CNN to each of them, and then extracts a vector of features from the CNN; this careful approach is disadvantageous in terms of computation time. Fast R-CNN works to rectify this shortcoming by introducing a unified architecture for that simultaneously identifies object proposals and computes a corresponding bounding box, which increases relatively computational efficiency and accuracy of the system [31] as shown in Figure 5.

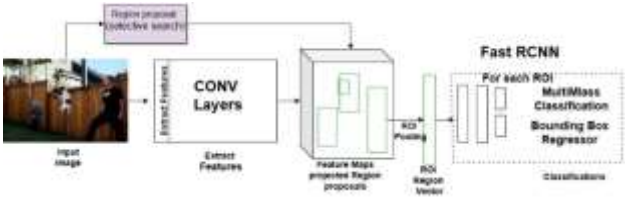

Fig 5. Two-step object detector, classical Fast R-CNN

It is Fast R-CNN's region of interest (RoI) pooling layer, which is of benefit as it essentially instantiates fixed-sized maps of possible areas of interest in the image from the entire map of it. This makes it unnecessary to adapt and fine-tune the size of the proposals for each region, a major headache in R-CNN. Fast R-CNN unifies feature extraction, region classification, and bounding box regression into a single network that advances ease of training and faster inferencing. The model would be capable of improving both the classification and localization powers, causing a considerable increase in the accuracy of benchmark datasets like PASCAL VOC and MS COCO. Fast R-CNN goes over an image at about the same speed as R-CNN, while Fast R-CNN has done this many times faster than the others [32]. Fast R-CNN minimizes redundant computations by computing the convolutional feature map for the entire image and then sharing it with all region proposals. Others have applied CNN to each individual region. This is achieved with an operation called RoI pooling that, in essence, symmetrically solves the size problem with ample supplies since feature maps for proposals would be of the same size. This would ensure the easy acceptance of input sizes by the model, as illustrated in Figure 6.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 3S, 2025
https://www.theaspd.com/ijes.php

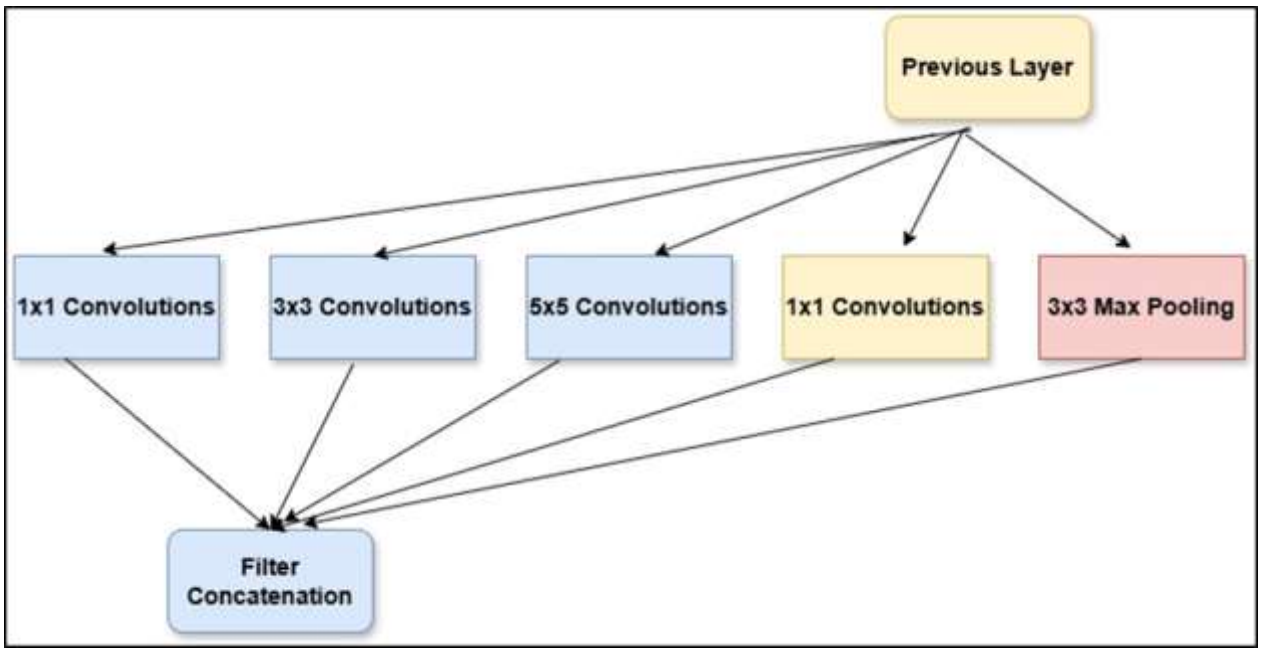

Fig 6. Getting Started with the GoogLeNet Theory of Architecture

However, Fast R-CNN relies on external methods for region proposal that are conservative in their search, hence making its application not ideal in real-time applications. The dependence, however, has been abandoned in the successor, Faster R-CNN, an RPN that has a network specialized in proposing regions inside the neural network, simplifies the process, and further increases the performance during the real-time application. Even with this limitation, Fast R-CNN remains to be a very significant advance in the research for object detection because it has an architecture that is different from all others and has the possibility of making the trade-off between speed and accuracy.

The impact of Fast R-CNN transcends the technological benefits. Most of the time it is applied in various applications, which require Object detection, for instance in autonomous driving, accurate and quick object detection is essential for the detection of cars, pedestrians, and traffic signs. In the field of medical imaging, it is applied to the detection of abnormal formations such as tumors or lesions sur[33]veillance systems that also allow monitoring and detecting activities in real time. These methods have led to major influences on some very crucial advancements in computer vision, of which one is creating Mask R-CNN, which builds upon segmentation capabilities of Fast R-CNN because it includes multiple branches predicting masks at the pixel level. In the end, Fast R-CNN marks a big step in the development of object detection algorithms concerning its capability to process images quickly, yet with high accuracy and can therefore be considered a leading model for research and applications based on images, as demonstrated in Figure 7.

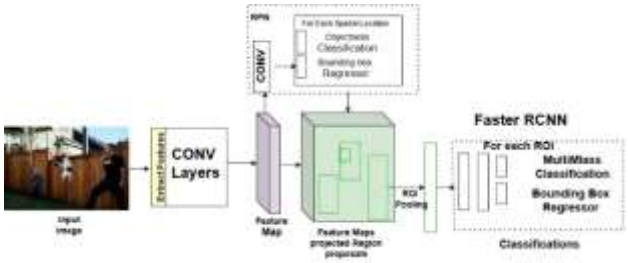

Fig 7. Two-step object recognizer, classic Faster R -CNN

Though overshadowed by Faster R-CNN and other recent approaches, original contributions, mainly RoI pooling and end-to-end architectures have had a lasting impact on the development of state-of-the-art object detection frameworks [34].

5.6 Feature Pyramid Networks (FPNs)

FPN was proposed by Lin et al. in 2017. The innovation has marked a giant leap in the sphere of deep learning, particularly for computer vision applications whereby objects of varying sizes are usually very difficult to be detected [35]. The FPN architecture is designed to generate multiple-scale maps that amalgamate the rich semantic information of top-down layers with high spatial details of bottom-up layers, hence achieving good performance for objects of various sizes. The network performs top-down inference, i.e., it first infers high-level semantic feature maps from deep layers to shallower ones. These, in turn, are combined with strong but meaningless feature maps from other layers in the bottom-up convolutional hierarchy through lateral connections. This would make the feature pyramid representation very powerful since each layer is predicted independently and hence provides a very strong grasp over different object sizes. This method eases learning for hierarchical features and offers versions of semantic and spatial information at any level.

Since its inception, FPN has undergone a series of improvements and modifications. For example, PANet, proposed in 2018, introduces a new solution based on the bottom-up path other than that of FPN, which will make it possible to propagate low-level information relatively quickly and boost the precision of localization. Similarly, NAS-FPN uses neural architecture search to improve the construction of a multi-scale feature pyramid, acting through the dynamic change of feature hierarchy to enhance detection performance. Core to EfficientDet is BiFPN, a methodology of weighted feature fusion that selects and sorts features in multiple scales so as to detect perpetrators with maximum efficiency and accuracy. Such top adaptations as HrFPN increase the functional scope of FPN to tasks requiring high spatial resolutions such as semantic segmentation [36].

Table 1. Differences that are noteworthy between one-stage and two-stage CNN-based models for object identification.

Aspect	One-Stage Detectors	Two-Stage Detectors
Architecture	Single step (no proposal)	Two-step (proposal + refinement)
Speed	Faster, suitable for real-time	Slower, typically not real-time
Accuracy	Generally lower, but improving	Higher, better for small/overlapping objects
Complexity	Simple, computationally efficient	Complex, requires more computation
Class Imbalance	Faces imbalance; uses techniques like Focal Loss	Reduced imbalance due to proposal filtering
Examples	YOLO, SSD, RetinaNet	Faster R-CNN, Mask R-CNN, R-FCN

6. 4. Popular dataset

The MSCOCO dataset has a proper reference model for computer vision in general. The dataset was developed mainly to be used for the evaluation of the performance of image and object detection and segmentation using machine learning and deep learning techniques. The dataset has fewer classes compared to the number of instances in each class. It includes 91 types of entities; people, animals, transports, and many other common objects. Moreover, multiple instances per category and several different pictures with various characteristics for each example are considered [16]. Similar to previously introduced datasets in the pursuit of benchmarking visual recognition, the Pascal Visual Object Classes (Pascal VOC) is concerned with the classification, differentiation, and identification of visual objects. It started with 4 classes in 2005, expanding to 20 classes in 2007, with its

community interested annually in keeping and updating the dataset by augmenting it with more classes. The Flickr dataset [2][17] plays a major role in the field of computer vision and natural language processing since it supports the training as well as the evaluation of machine learning systems meant for tasks such as image captioning and retrieval[9]. As shown in Table 1, since there are multiple releases of Flickr, thus Flickr8k is a dataset of 8,092 photographs where each one has up to five different descriptions; it is generally used as a source of references for the performance of image description systems, such as KAGGLE. Flickr30k is a dataset of 31,000 photographs, with each one having five human-annotated sentences to facilitate a better evaluation of models regarding image captioning and image retrieval. Source Paper. Meanwhile, the Yahoo Flickr Creative Commons 100 Million Collection (YFCC100M) contains approximately 99.2 million photos plus 800,000 videos, all of which are available under a Creative Commons license, together with a comprehensive description of the content provided by users over the span of a decade. These datasets have driven the picture captioning, scene, and cross-modal learning research owing to availability of diverse quality datasets well documented whose utility was useful for training and evaluation of models. Lately, mean average precision (AP) has emerged as the prime metric for assessing object detection accuracy, which gives the typical accuracy at diverse recall levels and has multiple versions for every theme while the average class precision (mAP) is used as a single metric to gauge object recognition performance across all categories...

Table 2. The expansion of the Flickr dataset in terms of size, description, and academic use.

Dataset	Size	description	Academic use
Flickr8k	8,092 images	Five captions per photo.	Image captioning, creating models that generate written descriptions from visual content.
Flickr30k	31,000 images	Five captions per photo.	Image-sentence alignment, visual question answering, and multiple modes of learning.
Flickr30k Entities	Increase in the number of Flickr users 30k.	Spatially-appropriate correspondences that link written phrases to specific regions of images.	Visual anchoring, sentence localizing
Flickr-Faces-HQ (FFHQ)	70,000 exceptional images with a resolution of 1024×1024.	High-quality images that demonstrate age, ethnicity, and diversity in background.	Generative Adversarial Network (GAN), Face Recognition Research
YFCC100M	99.2 million photos and 800,000 videos	Metadata that includes user comments, descriptions, and geographic information.	Large-scale multimedia investigations, computer vision, multimedia exploration, and social media analysis.
Flickr Material Database	1,000 images divided into 10 material categories	Categories of materials: fabric, foliage, glass, leather, metal, paper, plastic, stone, water, wood.	Recognizing materials, creating algorithms that differentiate the types of material based on visual attributes.
Flickr Style Dataset	About 80,000 photos annotated with 20 curated style labels	Image labels that describe the style of each image.	The computational arts, style recognition, and the retrieval of images based on style are all part of the computational aesthetics.
Flickr Cropping Dataset	3,413 images with cropping and rankings associated with them.	Cropping notes and rankings for visual favorites	Reviewing automatic image cropping methods, aesthetic evaluation, and composition of images.
Flickr Image Relationships	Web data representing the Flickr social network	Nodes that represent users, and edges that represent the connections between users.	Network analysis, which studies the way images are connected, understands the visual data network, and models the relationship between images.
Flickr Africa Dataset	Flickr images of geotagged African countries	Metadata that includes geographic information and user-supplied information	Investigating the diversity of geo-datasets, recognizing the biases in these datasets, and creating models that are generalizable across geographic regions.

CONCLUSION

Objects have been detected indeed much over the last 20 years; it has become a fundamental element of visual research. This paper reviews those detectors that have driven the drive toward the invention of object detection. These include VJ and HOG which set the scene on which real-time face or human detection could be achieved. it was Equally important, later improvements admitted more great flexibility in the modeling process for deformations of objects Advanced Learning R-CNN, YOLO, and SSD had further changed the field, increasing its accuracy and efficiency.

Apart from the background details, it will give the major technologies and optimization techniques that in one way or another have increased the speed of detection and its accuracy. This will involve feature extraction, neural network design, and hardware acceleration. The work will cover various fields of practical application: object detection in autonomous cars, medicine imaging, surveillance, or robotics, and in the end, shall provide the description of the benchmark datasets and evaluation metrics. Among these metrics are precision curves, mean average precision (mAP), and other valuable metrics that made research feasible by enabling comparisons among methodologies to be conducted consistently.

This study reviews the issues currently challenging the scientific community, including the need to handle occlusions, recognize small objects, achieve high real-time performance on resource-constrained hardware systems, and ensure that they are reliable in diverse environmental conditions. It discusses potential future directions that may alleviate these challenges, such as transformer-based architectures (with various modalities), increased unsupervised learning, and more generalizable methods to be applied to new class objects. This work tries to collect the works done in the past and proposes a guide to the future, intended as a guide for researchers and implementers who want to advance the field of object recognition.

REFERENCES:

[1] R. Kaur and S. Singh, “A comprehensive review of object detection with deep learning,” Digit. Signal Process., vol. 132, p. 103812, 2023.

[2] A. Karpathy, “Deep Visual-Semantic Alignments for Generating Image Descriptions”.

[3] K. Kang et al., “T-CNN: Tubelets with Convolutional Neural Networks for Object Detection from Videos,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 10, pp. 2896–2907, 2018, doi: 10.1109/TCSVT.2017.2736553.

[4] N. Ejaz, T. Bin Tariq, and S. W. Baik, “Adaptive key frame extraction for video summarization using an aggregation mechanism,” J. Vis. Commun. Image Represent., vol. 23, no. 7, pp. 1031–1040, 2012.

[5] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 1, no. February 2001, 2001, doi: 10.1109/cvpr.2001.990517.

[6] N. Dalal, B. Triggs, N. Dalal, and B. Triggs, “Histograms of Oriented Gradients for Human Detection To cite this version : Histograms of Oriented Gradients for Human Detection,” IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 886–893, 2005, [Online]. Available: http://lear.inrialpes.fr

[7] N. L. W. Keijsers, “ImageNet Classification with Deep Convolutional Neural Networks,” Encycl. Mov. Disord. Three-Volume Set, pp. V2-257-V2-259, 2012, doi: 10.1016/B978-0-12-374105-9.00493-7.

[8] L. Liu et al., “Deep Learning for Generic Object Detection: A Survey,” Int. J. Comput. Vis., vol. 128, no. 2, pp. 261–318, 2020, doi: 10.1007/s11263-019-01247-4.

[9] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, “A survey of modern deep learning based object detection models,” Digit. Signal Process. A Rev. J., vol. 126, pp. 1–18, 2022, doi: 10.1016/j.dsp.2022.103514.

[10] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," Multimed. Tools Appl., vol. 82, no. 6, pp. 9243–9275, 2023, doi: 10.1007/s11042-022-13644-y.

[11] W. Liu et al., "SSD: Single shot multibox detector," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2.

[12] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional Single Shot Detector," 2017, [Online]. Available: http://arxiv.org/abs/1701.06659

[13] A. N. Azhar and M. L. Khodra, "Fine-tuning Pretrained Multilingual BERT Model for Indonesian Aspect-based Sentiment Analysis," 2020 7th Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2020, pp. 2980–2988, 2020, doi: 10.1109/ICAICTA49861.2020.9428882.

[14] P. Yun, L. Tai, Y. Wang, C. Liu, and M. Liu, "Focal Loss in 3D Object Detection," IEEE Robot. Autom. Lett., vol. 4, no. 2, pp. 1263–1270, 2019, doi: 10.1109/LRA.2019.2894858.

[15] S. Zhang, L. Wen, Z. Lei, and S. Z. Li, "RefineDet++: Single-Shot Refinement Neural Network for Object Detection," IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 2, pp. 674–687, 2021, doi: 10.1109/TCSVT.2020.2986402.

[16] T. Lin, "Focal Loss for Dense Object Detection," arXiv Prepr. arXiv1708.02002, 2017.

[17] Acharya, Image Processing: Principles and Applications [book review], vol. 18, no. 2. 2007. doi: 10.1109/tnn.2007.893088.

[18] R. Luo, A. Peng, H. Yap, and K. Beard, "Joint Moment Retrieval and Highlight Detection Via Natural Language Queries," 2023, [Online]. Available: http://arxiv.org/abs/2305.04961

[19] J. Trotter and P. Agrawal, "A multiprocessor architecture for circuit simulation," in 1991 IEEE International Conference on Computer Design: VLSI in Computers and Processors, IEEE Computer Society, 1991, pp. 621–622.

[20] Y. Munian, A. Martinez-Molina, D. Miserlis, H. Hernandez, and M. Alamaniotis, "Intelligent system utilizing HOG and CNN for thermal image-based detection of wild animals in nocturnal periods for vehicle safety," Appl. Artif. Intell., vol. 36, no. 1, p. 2031825, 2022.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Ieee, 2005, pp. 886–893.

[22] D. G. Lowe, "Object recognition from local scale-invariant features," in Proceedings of the seventh IEEE international conference on computer vision, Ieee, 1999, pp. 1150–1157.

[23] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 4, pp. 509–522, 2002.

[24] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in 2011 International conference on computer vision, IEEE, 2011, pp. 89–96.

[25] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," Proc. IEEE, vol. 111, no. 3, pp. 257–276, 2023.

[26] H. Wang and B. Johansson, "Deep Learning-Based Connector Detection for Robotized Assembly of Automotive Wire Harnesses," IEEE Int. Conf. Autom. Sci. Eng., vol. 2023-Augus, 2023, doi: 10.1109/CASE56687.2023.10260619.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[28] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Discriminatively trained deformable part models, release 5," 2012.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8691 LNCS, no. PART 3, pp. 346–361, 2014, doi: 10.1007/978-3-319-10578-9_23.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "September. Spatial pyramid pooling in deep convolutional networks for visual recognition. In european conference on computer vision (pp. 346-361)." Springer, Cham, 2014.

[32] R. Girshick, "Fast R-CNN," Proc. IEEE Int. Conf. Comput. Vis., vol. 2015 Inter, pp. 1440–1448, 2015, doi: 10.1109/ICCV.2015.169.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks, p 91–99," Adv. Neural Inf. Process. Syst. NIPS, Montr. Canada, 2015.

[34] X. Li, M. Shang, H. Qin, and L. Chen, "Fast accurate fish detection and recognition of underwater images with Fast R-CNN," Ocean. 2015 - MTS/IEEE Washingt., pp. 1–5, 2016, doi: 10.23919/oceans.2015.7404464.

[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE Conference on computer vision and Pattern Recognition, 2017, pp. 2117–2125.

[36] M. Turk and A. Pentland, "E i g e d c e s for Recognition," vol. 3, no. 1.

[37] Al-Azzawi, A., Ouadou, A., Max, H., Duan, Y., Tanner, J. J., & Cheng, J. (2020). DeepCryoPicker: fully automated deep neural network for single protein particle picking in cryo-EM. BMC bioinformatics, 21, 1-38.