|| Volume 5 || Issue 12 || April 2021 || ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH & ENGINEERING TRENDS

Multidisciplinary Journal

Double-Blind Peer Reviewed Refereed Open Access International Journal

OBJECT DETECTION USING MACHINE LEARNING

Funde Priyanka[1], Smita Jagtap[2], Varsha Nijave[3]

[1] **Student of Computer Engineering, Adsul Technical Campus Chas Ahmednager. priyankafunde8888@gmail.com**

[2]**Student of Computer Engineering, Adsul Technical Campus Chas Ahmednager.smitajagtap1212@gmail.com**

[3]**Student of Computer Engineering, Adsul Technical Campus Chas Ahmednagar.varshanijave@gmail.com**

Abstract

Computer vision is excelling in the field of segmentation, feature extraction, and object detection from image data. The object detection is gaining immense interest from a application such as healthcare, traffic monitoring, surveillance, robotics etc. The ability to detect the object more precisely is an important factor due to its application in sensitive domains. Over the past few years, researchers have strived to cope up with this challenge. This study presents a review of object detection approach considered using Convolutional Neural Network (CNN). The CNN is used in all three methods (salient, objectness, and category-specific) of object detection. Deep learning frameworks and the platforms that are popular for the object detection task are also reviewed.

Keywords: Salient, Objectness, Object Detection, F-RCNN, CNN.

INTRODUCTION

According to Gartner Technology, computer vision is one of the trending areas in a different domain such as medical, automation, surveillance, defense, and customer markers [1]. Moreover, it is estimated to reach a USD 17.38 billion by 2023 [2]. Computer vision dates back to the 1960's with an inspiration of human ability to vision and recognize the object [3]. However, with the development of information technology, optimum computer processing capabilities and affordable hardware's (camera, data storage, etc.) attracts the researchers to emerge in the field of analyzing high dimensional images and videos.

The deep learning boosted the growth of computer vision and produced state-of-the-art results in image recognition, feature extraction, and object detection. The deep learning requires a significant amount of dataset to train and a high computation power. The graphical processing units fulfill the requirement of computer vision to process efficiently. In the field of computer vision, object detection is an important task. The object detection is a process in which the instances of the objects are detected for a particular class in an image. The object detection is trending due to its applicability in a broader area [4].

The object detection in an image requires attention to many factors such as the luminous conditions, the scale of the object, and the orientation. The object detection is the primary step in the classification of the image, and many researchers have proposed different ways to detect and locate the object. The deformable part model was introduced using a sliding window to search for the object and sub-window to assign a score. The scores act as an indicator to detect the future unseen objects [5]. Further, the improvements have been proposed in the deformable part model using the Convolutional Neural Network (CNN). It generates the pooling layers that can efficiently handle the properties of the deformed object detection [6].

With the groundbreaking results, the convolutional networks became the backbone for the object detection models. The object detection process was time-consuming as it performs the forward pass convolution for each proposal. Therefore, several enhancements to the CNN have been proposed in the existing studies to minimize the training time of the object detection. This study reviews the literature and analyzes it critically, the algorithms and techniques used for the detection of objects in the images. Moreover, the study intends to find out the frameworks and models that are efficient by analyzing the results and to provide the insights for future works and direction on object detection.

This paper is organized as follows: the second section reviews the popular deep learning frameworks and platforms and types of convolutional networks. The literature is reviewed in section
2.3. Section 3 comprise of the detailed discussion on the literature. In the last, the paper is concluded in section 4.

OBJECT DETECTION

This section reviews the most popular deep learning frameworks that are used in the academics and the industry, also the interface used by the deep learning frameworks. Moreover, the section covers a comprehensive review of the object detection taxonomy.

A. Deep Learning Framework

Tech giants and opensource community have developed many deep learning frameworks. For example, Caffe is developed by Berkeley Vision & Learning Center, it uses the CNN and RNN deep learning models and provides the features of data storage, plain text schema for modeling, the speed and modular structure. The interface is C, C++, command line interface, Python, and MATLAB [7]. Whereas the TensorFlow works only in C++, Python, Java and Go.

The features that TensorFlow possess are the image classification, portability, mathematical computation with the help of a data flow graph, inception, and differentiation [8]. Keras works on the Python platform and is wrapped for the R platform as well; it offers arbitrary connection schemes, fast prototyping, minimalistic modules. Keras uses CNN, RNN, Deep Belief Network (DBN) and Boltzmann Machine (RBM). Also, it supports the multi-node parallel execution [9].

Microsoft Cognitive Toolkit (CNTK) is capable of running on the Command line interface, C#, C++, and Python. The main features of CNTK are batch normalization, automatic hyperparameter tuning, and Multi-dimensional dense data handling. It uses CNN and RNN; it also gives multi-nodes parallel execution functionality [10].

Pytorch is the enhanced version of the torch, which is developed under Facebook. It works on Python platform, it has minimal framework overhead and is faster as it's Neural Network

backends are written as independent libraries with a C99 API.
The memory is also customized and made efficient hence deeper, and more significant network models are possible to implement [11].

B. Convolutional Neural Network and its Variants

Deep learning is a branch of machine learning where the neural networks layers go in depth. This approach has achieved astonishing results in every field. A primary neural network consists of three layers, one input, the second hidden layer where the weights are set, and third is the output. With the addition of more layers, it is considered to be a deep neural network. The deep learning models have multiple level representation, this representation is obtained through the non-linear and simple modules. These modules change the representation from the level one to a higher level of representation. With such transformation, deep learning models learn more complex functions.

The most popular neural network in the field of computer vision is a convolutional network, it is in use for a long time and got popularity in recent years as the development of hardware have empowered machines with more computational power the convolutional networks are going towards deep learning providing better results. For the object detection purpose the most popular and used deep learning models consist of the recurrent convolutional network, fast recurrent convolutional network. These neural networks have their own advantages and are explained in the following subsections.

The convolutional neural network (CNN) can consist of one or more than one convolutional layers. The convolutional neural network architecture designed in such a way that it takes advantage of a 2D image, which is achieved via local connections and pooling layers to get the invariant features. Generally, there are two ways of convolution, one is with the fully connected convolutional layers, and the other is locally connected layers.

The fully connected layer convolutional network has all the input layers to the hidden layers. This method is feasible for smaller images as the resolution of the image increase, it processes becomes more computationally expensive as the number of parameters increase. Whereas in the locally connected convolutional network the connection of neurons to input is restricted this way the hidden neuron is connected to the limited number of input units. Through this method, the issue of extensive computation is reduced.

Naturally, a particular image has the same statistic in all regions of the image. This means that if the features of one region are known or calculated the same can be applied to the rest of the regions in the same image [12]. It is like parsing the convolutional (features) window on the image to get the convolved features of an image.

1) Recurrent Neural Network

Among the artificial neural networks, the recurrent neural networks are widely used due to its enhanced feature. The recurrent neural network has internal memory for the processing of the input. This memory unit on each neuron helps the neural net and form a loop within a neuron. The memory unit also helps to learn the information from the previous input, which allows it to predict the next outcome more accurately.

The application of a recurrent neural network is widespread in the object detection models as it is combined with feedforward convolutional network and recurrent multi-layer perceptron with the local connections. Since the final model is a combination of RNN and CNN, the final model is named as a Recurrent convolutional network [13]. This combination is illustrated in the following Fig. 1 where the convolutional net and RNN is combined to form the recurrent convolutional neural network. The benefit of RCNN is that there are few training parameters and the results are better than that obtained from the traditional CNN based models. Recurrent Convolutional neural networks are widely used for the feature extractions and generate the feature maps which are further used for the detection of the objects through the features.

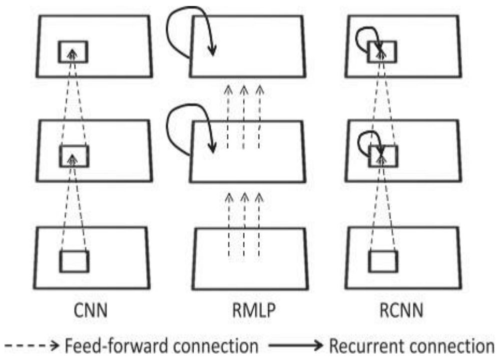

Fig. 1: Representation of a recurrent convolutional network

2) Fast Recurrent Convolutional Neural Network

The fast-recurrent neural network (fast RCNN) is an improved version of RCNN and Spatial Pyramid Pooling network. The architecture is shown in Fig. 2, where the image is fed to the fully connected convolutional neural net along with the region of interest. The region of interest is pooled into the feature map of fixed size, these are then mapped over the feature vector through a fully connected convolutional layer. The fast RCNN has two output vectors for each region of interest one is the softmax probabilities, and the other is the bounding box regression [14].

Fast RCNN has some advantages over the traditional RCNN and Spatial Pyramid Pooling network. It provides higher quality detection in terms of mean absolute precision (mAP) than the other two. The training is done in one-step by using the multi- task loss since the training is done in one-step it requires less training time. Another advantage of Fast RCNN over RCNN is that it does not require the memory storage to store the features which are very useful for object detection as it increases the detection speed [16].

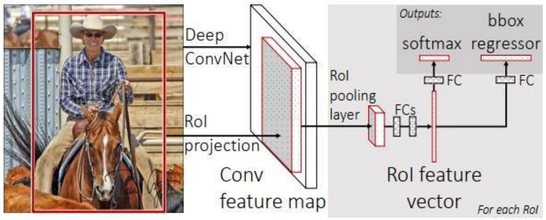

Fig. 1: Fast RCNN architecture [14]

C. Taxonomy of Object Detection

The deep learning is making remarkable contributions in computer vision. The model based on deep learning are performing well for face detection, image segmentation, and object detection [15]. The amount of data required to train deep learning algorithms may reach up to 100 terabytes and to process this huge data possess a significant challenge. Nowadays with the advanced GPUs tools, the process of image analytics has reached real-time processing with 30 to 60 frames per second.

The object detection is a basic computer vision problem, researchers have explored and tried to resolve this by proposing many methods to detect the objects. These methods are categorized into three main categories such as Objectness detection (OD), Salient object detection (SOD), and category- specific object detection (COD), refer Fig. 3 [17]. Apart from these, some other methods are proposed for object detection that is highlighted.

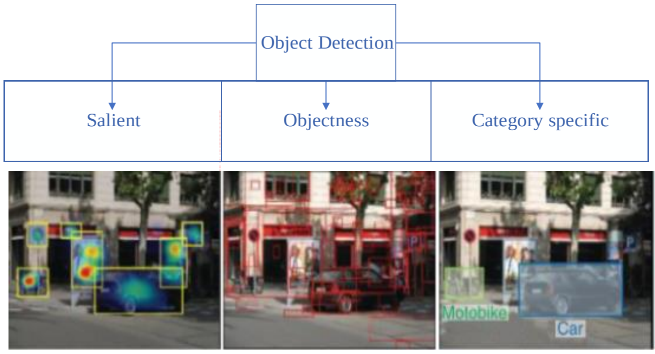

Fig. 2: Object Detection Taxonomy [17]

1) Salient Object Detection

The salient object detection method is inspired by the human vision's ability to distinguish the objects based on the highlights, which draw the attention towards the specific object and ignoring the background. The salient model should capture the attention of grabbing objects and complete segmentation of the objects [18].

In Salient object detection, there are two methods for the detection, top-down and bottom-up. The bottom-up method focuses on discriminating the objects in the background from the foreground in the visual scenes. Whereas the top-down method emphasizes on highlighting the category specific objects in the visual scenes. Zhang [19] proposed a model for salient object detection, the approach itself contains two parts which include the exploration of patch level and object level cues. In the first step of their model, the objectness algorithm is used for the coarsely localized positions of the salient object in the image. Then the variance is used to calculate the compactness of the spatial color distribution. Although the model worked well for images with a more straightforward background. It failed to perform well when the salient object and background of the image have a similar color.

For better salient object detection, it should be able to extract regions and objects that are distinctive in the image. In [20], deep learning was used to focus on the skip layered structure. They introduced a new method by adding short connections in the skip-layer for the salient in the holistically-nested edge detector architecture (HED). Their proposed architecture was based on the VGGNet model along with the HED. They combined both features of the deeper side output (salient regions) and shallow side output for low-level features. The architecture consists mainly of two stages connected together i.e. salient locating and details refinement stage. The salient stage looks for the salient regions in the image and the details, in the next stage a top to the bottom method is introduced. Here the short connections are made between both layers, by doing so the features of both layers can be used for better prediction of the salient objects, and it gives an accurate and dense saliency map.

2) Objectness Detection

Objectness detection creates multiple bounding boxes in an image for every possible object without considering the category of the object. The basic idea of objectness is to create an objectness measure for generating candidate proposals. This measure is the confidence score that determines the presence of the object in the generated proposal.

Deep network object detection frameworks can be divided into two branches of object detection, the region free and region- based methods. With the success of these methods [20] proposed a way to combine the best features of both region free and region based methods. In order to achieve, these two main problems of multiscale localization and negative space mining were focused. In multi-scale localization, the objects may be at any location on the image, and all of the locations need to be considered for the object detection. A reverse connection is proposed so that on the corresponding network the objects can be detected. Negative mining should be in every object detector, as there is an imbalance of non-object and object sample ratio.

An objectness prior phase is applied to the convolutional feature-map to reduce the object search space and to optimize the model in the training phase. The Reverse connection with objectness prior networks framework contributes towards end-to-end object detection, high accuracy, resource and time efficient. With the help of reverse connections semantic information is collected from the convolutional layers then the objectness prior provides the roadmap for the object search in the image. Finally, the optimization is done with the multitask loss function. Whereas the detection accuracy of the object detection is increased by the data augmentation and negative mining [21].

There is an increasing demand for fast object detection such as moving car, which requires a fewer amount of candidate windows to avoid exhaustively searching for massive sliding windows. One study have proposed Deep Objectness Representation and Local Linear Regression (DORLLR) to improve the quality of the blind proposals using Intersection- Over-Union (IOU). It is a metric used in the evaluation of the quality of the sliding window generating the proposals in real- time object detection [22]. Many efforts have been made to improve the proposal quality by using the blind estimation. The blind quality proposal assessment can be explained in two ways. The blind proposal quality proposal is taken as a background and foreground segmentation problem as the foreground areas are considered to have more information.

The segmentation method only identifies the proposal quality as background or foreground. While the other method focuses on the scores and rank of the window function in terms of the particular image cues. Considering these a generic blind proposal quality assessment (BPQA) a model is formulated which was able to select multiple proposals based on the (BPQA). The model trains in two phases, one is the deep objectness representation, and the other is a local linear regression. The CNN bases feature extraction is used to explain the deep objectness and to predict the quality of each proposal a local linear regression model is utilized.

In a study [23] proposed the hierarchical objectness network model which is capable of object detection and proposal generation. In their model, they considered the most crucial points in object detection such as precision, multi-scale, and the computation cost. The model works in three steps, the CNN extract the features from the image. Then a saliency map is predicted to give the plane for searching objects, and in the end, the potential proposals are refined by the stripe objectness. The precision of the model depends on the stripe objectness, which gives border objectness and in-out objectness. These provide the object border and confidence score on the proposal locations. The proposal marked with this information is divided into the vertical and horizontal strips, these strips do not overlap and show the probability of the object border or object itself.

The deep and shallow convolution layers are connected reversely to get the high-level semantic features. On multiple scales these features contain many degrees of resolution information for objects. The involvement of single saliency map deceases the memory overhead and computation time.

3) Category-Specific Object Detection

The category-specific object detection does not consider it as a human vision instead it converts it into a multiclass classification problem in which the image is discriminated into pre-trained classes.

In recent years there the field of computer vision is focusing on the object classification and detection. New models are proposed by researchers for better accuracy and less computation time. Leibe and Hutchison [24] proposed a model, Single Shot MultiBox detector (SSD) which is a feed-forward convolutional network. The model works in two steps. First, the convolutional network assigns the scores on the fixed sized bounding boxes to check for the presence of the object. Then to get the final result for detection non-maximum suppression step is applied.

The SSD contributes in a way which is faster and more accurate than the other models who produce pooling layers and region proposals. The model uses the small convolutional filters on the feature maps which gives the box offset and category score for each bounding box. The model achieves high accuracy by using the feature maps to predict the different scales and aspect ratio separately. These contributions lead the model to have the end- to-end learning and achieve high accuracy on the input of low- resolution images. The training of the model includes the selection of default boxes, data augmentation strategies, scales for detections and the hard-negative mining. Also, the output of a fixed set of detectors is supplied with the ground truth information. The computation time of the model is also surprising with the 70% mAP, and the speed is 22 FPS, which is very close to the real-time detection.

An object proposal method of object detection is presented by [25] in which they demonstrated that Convolutional neural networks could achieve dramatic results in detecting the objects. To do so, they focused on training a model with high capacity, provided with less quantity of annotated data and a deep network for the localization of the objects. The model essentially has 3 three modules for the object detection. At first region, proposals are generated which are independent of the object categories. These region proposals define the objects that may be detected by the object detector. This information then passes to the next module where a fixed-length feature vector is extracted from the large convolution neural network. The last step contains the linear support vector machine which gives scores to the feature vector obtained from the CNN.

For every class non-maximum suppression is applied which removes the area where the intersection of union score is high and overlaps with the selected region. Out of two challenges of localization and training, the training has two sections one is supervised pre-training of the CNN where ILSVRC 2012 data set is used with no bounding-box labels. The other is the fine-tuning of the CNN parameters. The model is then tested on the PASCAL VOC 2010-12 dataset and compared with the baseline model of SegDPM. The results of experimentations show that the proposed model gives 53.7% mAP whereas the baseline have the mAP of 40.4%.

4) Other Object Detection Methods

With the help of the deep neural networks [26] proposed a method to train the object detector. The approach they adopted was to train the deep neural network which has no prior information about the object class. The proposed model was named as DeepMultiBox, and it contributes towards the object detection problem in several ways. They explained object detection as a regression problem, by taking the coordinates of the bounding boxes and assigned the confidence value of each confidence box. This confidence value is the likelihood that the object is present in a particular box or not. Another main contribution of their model is a loss which is calculated from the ground truth boxes and the predicted ones. This loss is important as it carries the features information which is feed to the DNN through the back-propagation method. The last feature of their model is the scalability since all the object detection methods train the predictors for each class. However, this method doesn't train concerning class information, due to this the model can be used for any set of images. The advantage of DeepMultiBox has scalability over this one box per class.

A convolutional neural network framework is proposed by [27] which is referred to as DeepBox which is four lightweight layered architecture. The process of object detection in this framework is completed in two steps. The 1st step aims to generate the pool of Bottom-up proposals to get rid of the additional windows. 2nd step re-ranks the proposals obtained from the 1st step based on the scores that are generated by DeepBox. Two datasets PASCAL VOC 2007 and COCO which are publicly available were used in the experiments for the object detection. The model improves the bottom-up proposal approach in terms of the AUC over the Edge Box when applied on the VOC 2007 by 26%. The execution time of the bottom-up was closer to the edge box i.e. 250ms per image. It is observed from their experimentation that the DeepBox approach in Fast R-CNN with 500 proposals works better by 4.5 points as compared to the edge detection approach with the same number of proposals per image.

Although deep learning is achieving the remarkable performance in the object detection. But as the neural network gets deep the accuracy start getting saturated, and degradation starts in the performance. [28] proposed the solution to this problem in object detection by using the residual learning of the deep neural network. The advantages of using the residual learning method are, unlike the plain net whose error increases with the depth the deep residual nets are more natural to optimize. With the depth, issue resolved the residual nets can achieve better accuracy gain. For maximum depth of 152 layers, the network gave 3.57 % of top 5 error when applied on to the ImageNet dataset. This approach already surpassed and managed a 1st place in ImageNet detection, COCO detection, and other localization competitions in 2015.

A Large Scale Semi-Supervised Object Detection method was proposed by [29]. They incorporated the semantic and visual knowledge transferred to a weak category object detector. A weak category is the one which has image level annotations, whereas strong category contains object and image level labels. Their study aims to exploit the visual and semantic knowledge about the objects and to use the differences of the category specific image classifier and object detector. This way the performance of the model can be improved without using the bounding box annotations. Two models are integrated with the large-scale detection through adaption framework (LSDA), the integrated models are knowledge transfer through the visual similarity, and the other is knowledge transfer through the semantic relatedness. The final model is the linear composition of the two-knowledge transfer model.

The model was trained and tested on the semi-supervised and large-scale dataset of ILSVRC2013. The dataset has 200 object categories and 200,000 images. The experiments contained LSDA as a baseline model, and the model is trained through word2vec on Google news dataset which includes 100 billion words for the semantic representation. Both models performed better than the baseline LSDA model. The visual knowledge model showed steady improvement and achieved 19.02%mAP, which proves that the assumption that visual similarity improves the detection. The semantic model also performed better as compared to the LSDA, it achieved 19.04%mAP. The combined knowledge semi- supervised detection model achieved the state of the art result by improving the mAP score by 3.88% on weakly labeled categories. Although there is an improvement in the results it confuses the categories that are visually similar.

Tang [30] proposed another model, which focus on the deformable part-based weakly supervised learning for the object detection. In their study, they focused on enhancing the weakly supervised deformable part-based models, by emphasizing the on the size and location of class root filter. The weakly supervised learning model uses objectness approach to get the class-independent object proposal scores. The object detection is done by fusing the objectness and the deformable part model. In the first step, the objects are estimated using the objectness and salient region reference. In the next step, the multiple regions latent classifications are used to estimate the non-target and target object category. The classification and detection score is matched and rescored for the results. The detection is done using 2 models one with a single layer and other with multiple layers weakly supervised DPMs. In experimentation, PASCAL VOC 2007 and MS COCO 2014 datasets are used to train and evaluate the models. The result is compared with the weakly supervised DPM random, which initializes a random filter window. The overall result of the multilayer WPDM was 4.3 points better than the baseline model which is 17%. The results on MS COCO were lower as it contains smaller sized objects than the PASCAL data.

In [31] presented a model which is the mix of Region Proposal Networks who shares the convolutional layers with the object detection network models. The region Proposal network used in their study was a fully-convolutional network which can be trained for the generation of the detection proposal. This Regional Proposal network is combined with the fast RCNN to gain maximum information from the feature maps for the object detection. The training of both the regional proposal network and FRCNN was done separately, in this process of training and fine-tuning the proposals were fixed for both networks. This training is done in 4 brief steps which follow the training of the RPN in the 1st step which generates the proposal, in the second step the proposal generated from the first step is used in the training of the fast R-CNN. In the third step, the convolution layer is fixed and fine tuning of the RPN is done on the unique layers of convolution. In the last step, the same process is performed for the fast RCNN.

After these steps, the model is unified which shares the same convolution layers. This model is applied to the benchmark dataset provided by PASCAL VOC 2007, it contains 5000 test and training images of 20 different categories. The results of the two- stage model were better than the single stage by 4%, the overall model efficiency in terms of mAP is 58.7%. This method allows object detection based on the deep-learning and enables the system to give 5 to 17 frames per second.

In a study, scale transferable object detection method is proposed by [32]. In this approach, the scale-transfer module is introduced to get the high level meaningful and multi-scale feature maps. This scale-transfer module can be embedded into the DenseNet. The DenseNet contains the dense blocks, and the output of each block becomes the input of the next block. Through this process, the predictions are combined from the previous feature maps. The Dense net is combining the low and high-level features, whereas the scale-transfer module is comprised of the scale transfer layer and pooling layer. The pooling layer is used to get the feature maps of small scales while the scale transfer layer for the feature maps of large scales. With these two embedded, the scale-transfer module and DenseNet a one stage object detector is formed as Scale-transfer detection network (STDN). As a whole, this STDN is divided into 3 parts, the base, and two prediction subnets. The 1st subnet is responsible for the object classification, and the 2nd is the box regression. Both of these subnets have a similar structure of 1x1 and 3x3 convolution layer.

In addition to convolution layers in the box, regression contains the 4A filter in its last layer. For the training purpose, the PASCAL VOC and MS COCO datasets with 20 and 80 categories were selected. The evaluation of the model was done through the mean average precision (mAP). The model performed with 80.9% mAP on PASCAL dataset. This score is a bit less than that achieved by DSSD was 81.5%. Although DSSD had better accuracy, it didn’t perform well in testing speed as it had more residual parameters. The proposed model achieved the test speed of 28 frames per second.

The presence of an object within a context makes it easier to detect the object correctly, [33] uses both the contextual information and the local appearance for the object detection. The proposed model works in 3 phases, in the first phase object proposal are generated through the faster R-CNN. Two faster R- CNN architecture is used, the Simonyan and Zisser- man model, is called VGG whereas Zeiler and Fergus model (ZF). The VGG has 13 while ZF has five shareable convolutional layers. The second phase employs a conditional random field framework (CRF) which model the contextual information with the object detection. In the last and final phase mean field approximation approach is used to conclude and maximize the object level agreement.

For the evaluation is a benchmark object detection dataset of PASCAL VOC2007 is used, that contains 20 distinct categories of the objects. The model with ZFnet is trained on PASCAL VOC 2007 and 2012, which performed better on 13 classes out of 20 than the baseline model by 0.75% mAP. The model with the VGGnet performed 73.51% that improved 0.70% from the baseline model. This performed well on 14 classes out of the 20. The results show that the approach of using the contextual information greatly improve the object detection performance.

The Salient maps are essential in object detection, [34] proposed a novel approach of detecting the object using saliency estimation and hierarchical spectral partition. The proposed model transforms the image to superpixels and at the same time detects the edge of the image. Superpixels are used to enhance the saliency estimate, whereas the edge detector is used to support superpixels further as the centroid may lie outside in the superpixels. Both of the transformed forms of images are used to form a graph with edge the and color information. Ncut, a spectral partition technique is used on the obtained graph, which divides the superpixels into two partitions at the same time saliency estimation is done on the respective hierarchies. To retain the object connections, all the saliency maps are added as they have equal probability. In the final stage, the results are further refined by retaining the superpixels, this is done using semi-supervised graph-based learning. The limitation of this model is that it fails to grasp the smaller objects in the presence of large objects, this is due to the biased split of superpixels by Ncut. Also, in some cases where the small objects attract more attention are highly likely to be detected.

DISCUSSION

Deep learning has brought great improvements in the field of computer vision through processing multiple layers to learn representations of data with several levels of abstraction. The deep learning framework simplifies and improves development time, model building, feature extraction, and efficient resource allocation.

CNN is designed to process the image and any data modalities that can be represented in the form of multiple arrays. CNN and its extensions such as R-CNN and Faster R-CNN are applied with great success to detection, segmentation, and recognition of objects and regions in images. In the reviewed literature, the use of CNN is reflected as a core function, and many proposed models have their version of filters attached to the convolutional network. From the review summary mentioned above in most of the proposed the datasets used be the benchmark and being tested by many researchers but the adoption of approach yield different results.

Few factors must be kept in mind that every object detector model has two parts, one is to create the filter, and the other one on which that filter will work. For example, the salient object detection uses supervised learning, the CNN generates the salient maps and then a top-down or bottom-up technique is used for the object detection. However, the other methods of object detection have a similar approach, but the in objectness based detection the feature maps are used (object proposal) then a ranking window rank based on the presence of object present in the particular window.

The category-specific object detection works on the localization, which is done through the region proposals. It takes object detection as a classification task as the location with the object and location with the no-object present.

Apart from these methods researchers have proposed other similar methods, for example, a Deep box method is introduced which uses the bottom-up proposal which is reranked and then a bounding box locate the object. In the same way, a Dense Net model is proposed which combines the low-end and high-end features along with the pooling layer and knowledge transfer layer. These models are the optimized version of the basic object detection methods.

The summary of the reviewed articles is presented in Table I. The existing models are evolved using multiple techniques such as the combination of object cue and salient features, region-based and region free based detection, F-RCNN with different detecting architecture and going further deep in residual learning [21], [23], [28]. With the evolution of the models and approaches, better results are achieved in the object detection.

Apart from the models, the data itself is very important as new dataset are huge in size and the manual labeling is time- consuming. There is a need for unsupervised techniques to automatically group the objects based on similarity. Finally, it is a significant challenge to build the model on imbalanced data [21], [24]. From the above discussion, it can be observed that the performance is improved by the development of new hybrid models which are more robust and can detect the large and small objects alike. For the future perspective, the features of the region proposal, salient proposals, and the edge information can be combined efficiently to increase the performance of the object detection.

Table I: Review Summary

Papers	Dataset	Techniques Used	Performance Metrics	Score
[33]	PASCAL VOC2007	Faster R-CNN + VGG and ZF	mAP	73.50%
[32]	PASCAL VOC + MS COCO	Scale Trasnfer densenet	mAP	80.90%
[23]	PASCAL VOC 2007 + 2011, MS COCO 2017	hierarchical objectness	mAP	78.60%
[20]	ECSSD, PASCALS, MSRA-B, HKU-IS, and SOD	VGGnet	MAE, F-measure, precision-recall curve	-
[34]	ECSSD, PASCAL-S, ASD, DUT-OMRON, and MSRA+	saliency estimation and hierarchical spectral partition	MAE, F-measure, precision-recall curve	-
[30]	PASCAL VOC 2007 + MS COCO 2014	deformable part-based weakly supervised learning	mAP	17%
[31]	PASCAL VOC 2007	Region Proposal Network + FR-CNN	mAP	58.70%
[22]	PASCAL VOC 2007	Deep Objectness Representation and Local Linear Regression	-	-
[21]	PASCAL VOC 2007, MS COCO and PASCAL VOC 2012	combined region free and region-based object detection	mAP	81%
[18]	PASCALS-S, JuddDB, ECSSD, MSRA10K, THUR15K, SED2 and DUT-OMRON	comparative study	-	-
[26]	ILSVRC 2012	DeepMultiBox	mAP	61%
[25]	PASCAL VOC 2010-12	Rich feature hierarchies for accurate object detection and semantic segmentation	mAP	53.70%
[19]	ASD, ECSSD, THUSK10K, and MSRA5K	salient object detection	PR curve, OR scores, AUC and F-measure	PR-curve 0.8459, AUC 0.9826, MAE score is 0.0753, OR score is 0.8157
[29]	ILSVRC2013	Large Scale Semi-Supervised Object Detection	mAP	23%
[28]	Imagenet, MS COCO, PASCAL 2007, 2012	residual learning of the deep neural network	mAP	76.40%
[24]	PASCAL VOC 2007, and MS COCO	Single shot MultiBox detector	mAP	80%
[27]	PASCAL VOC 2007, and MS COCO	DeepBox	mAP	26.00%

CONCLUSION

The literature on object detection is vast and many researchers have achieved benchmark performance in this field. This article reviews benchmark datasets that are used and the different directions of the object detection i.e. Salient object detection, Objectness object detection, and Category-Specific object detection. The salient detection is based on the distinguished objects which are highlighted and detected. The objectness goes with all the possible chances of the object in the image from the feature map, whereas the category-specific based detection is based on the classification, it discriminates the image into classes through the pre-trained functions.

The feature extraction based on the Convolutional network and the popular deep learning frameworks that are in practice, their interface which is usually used for the image data processing are discussed. Moreover, it also reviews the other advanced hybrid models for object detection with improved algorithms.

REFERENCES

1.K. Panetta, “Gartner Top 10 Strategic Technology Trends for 2018 - Smarter With Gartner,” https://www.gartner.com, 2017.
2.MarketsandMarkets, “Digital Signage Market worth 32.84 Billion USD by 2023,” https://www.prnewswire.com, 2016.
3.T. Huang, “Computer Vision: Evolution And Promise,” 19th Cern Sch. Comput., pp. 21–25, 1996.
4.S. Kamate and N. Yilmazer, “Application of Object Detection and Tracking Techniques for Unmanned Aerial Vehicles,” in Procedia Computer Science, 2015, vol. 61, pp. 436–441.
5.P. Models, “Object Detection with Discriminatively dpm,” in IEEE transactions on pattern analysis and machine intelligence, 2014, pp. 6–7.
6.W. Ouyang et al., “DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 7, pp. 1320– 1334, 2017.
7.N. Pittaras, F. Markatopoulou, V. Mezaris, and I. Patras, “Comparison of fine-tuning and extension strategies for deep convolutional neural networks,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, vol. 10132 LNCS, pp. 102–114.
8.“TensorFlow.”[Online].Available: https://www.tensorflow.org/. [Accessed: 18-Aug-2018].
9.“Keras.” [Online]. Available: https://keras.io/. [Accessed: 18-Aug-2018].
10.“MicrosoftCognitiveToolkit.”[Online].Available: https://www.microsoft.com/en-us/cognitive-toolkit/. [Accessed: 18-Aug-2018].
11.“PyTorch.” [Online]. Available: https://pytorch.org/about/. [Accessed: 18-Aug-2018].
12.stanford, “Unsupervised Feature Learning and Deep Learning Tutorial,” 2018. [Online]. Available: http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/. [Accessed: 29-Aug-2018].
13.M. Liang and X. Hu, “Recurrent convolutional neural network for object recognition,” Comput. Vis. Pattern Recognit. (CVPR), 2015 IEEE Conf., no. Figure 1, pp. 3367– 3375, 2015.
14.R. Girshick, “Fast R-CNN,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, vol. 2015 Inter, pp. 1440–1448.
15.Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.
16.I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning,” MIT Press, p. 1, 2016.
17.J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, “Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection: A Survey,” IEEE Signal Process. Mag., vol. 35, no. 1, pp. 84–100, 2018.
18.A. Borji, M.-M. Cheng, H. Jiang, and J. Li, “Salient Object Detection: A Benchmark,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 5706–5722, 2015.
19.Q. Zhang, J. Lin, W. Li, Y. Shi, and G. Cao, “Salient object detection via compactness and objectness cues,” Vis. Comput., vol. 34, no. 4, pp. 473–489, 2017.
20.Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, “Deeply Supervised Salient Object Detection with Short Connections,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 1, no. XX, pp. 1–14, 2018.
21.T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, “RON: Reverse connection with objectness prior networks for object detection,” in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, vol. 2017–Janur, pp. 5244–5252.