SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Object Detection by a Hybrid of Feature Pyramid and Deep Neural Networks

S. M. Notghimoghadam, H. Farsi*, S. Mohamadzadeh

Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 05 May 2022
Reviewed 12 June 2022
Revised 22 July 2022
Accepted 30 August 2022

Keywords:
Object recognition
Deep learning
Convolutional neural networks
Object classification

*Corresponding Author's Email Address: *hfarsi@birjand.ac.ir*

Abstract

Background and Objectives: Object detection has been a fundamental issue in computer vision. Research findings indicate that object detection aided by convolutional neural networks (CNNs) is still in its infancy despite having outpaced other methods.

Methods: This study proposes a straightforward, easily implementable, and high-precision object detection method that can detect objects with minimum least error. Object detectors generally fall into one-stage and two-stage detectors. Unlike one-stage detectors, two-stage detectors are often more precise, despite performing at a lower speed. In this study, a one-stage detector is proposed, and the results indicated its sufficient precision. The proposed method uses a feature pyramid network (FPN) to detect objects on multiple scales. This network is combined with the ResNet 50 deep neural network.

Results: The proposed method is trained and tested on Pascal VOC 2007 and COCO datasets. It yields a mean average precision (mAP) of 41.91 in Pascal Voc2007 and 60.07% in MS COCO. The proposed method is tested under additive noise. The test images of the datasets are combined with the salt and pepper noise to obtain the value of mAP for different noise levels up to 50% for Pascal VOC and MS COCO datasets. The investigations show that the proposed method provides acceptable results.

Conclusion: It can be concluded that using deep learning algorithms and CNNs and combining them with a feature network can significantly enhance object detection precision.

Introduction

Object detection is a term to describe a subset of computer vision and machine learning techniques highly influenced by deep learning and CNNs. Many studies have been conducted in this area. Object detection aims to identify and locate the types of objects in an image and categorize them into humans, animals, or vehicles [1]. As humans easily identify objects in an environment, an object detection system is designed to be trained like humans to be able to identify objects of different categories in fed images. A lower semantic distance between the machine and man evidently improves system performance. The conventional object detection methods are developed on the basis of handcrafted and shallow trainable architecture features. The design of intricate sets combining the features of many low-level images can easily impede their performance. The swift advancement in deep learning robust has led to the introduction of robust tools with semantic learning capabilities and deeper features for problem-solving in conventional architectures [2]. The emergence of neural networks influenced object detection methods. A neural network is a computational model formed by a large

Doi: 10.22061/JECEI.2022.9012.567

number of interconnected nodes or neurons, each indicating an output function called the activation function [3]. Such networks have wide-ranging AI applications, such as in signal processing and automatic control [3], as well as various image processing areas [4]-[6], radar images processing [7], [8] and mobile telecommunications [9]. It should be stated that despite the wide application of neural networks, they generally fail to provide high precision in computer vision. Therefore, attempts were directed toward increasing the number of layers and the depth of these networks. However, there was a problem: increasing the number of layers would lead to the vanishing gradient problem, thus drastically slowing down the training process. Further, in more severe cases, it would halt the training process. Accordingly, the attempts to study deep learning increased. Deep learning, still under development, is a subset of machine learning that aims to learn from data using hierarchical architectures. It is extensively employed in artificial intelligence and machine vision [10]. Deep learning algorithms generally fall into four classifications: convolutional neural networks, restricted Boltzmann machine (RBM), autoencoders, and sparse encoders. These four categories are compared in this study. The computations of the RBM method are lengthier and time-taking. Autoencoders are not resistant to changes in the image, such as image rotation.

The training of features is impossible in sparse coding. CNNs are applicable in two-dimensional data. They are resistant to image changes, which makes them excellent options for object detection, allowing them to perform optimally. A critical factor that has recently directed the attention toward deep learning is the success of these algorithms in the ILSVRC challenge held by ImageNet every year [10]. In deep-learning-based methods, the object features should be extracted from the image. In other words, the data should be processed to identify the explicit and determining features of the objects. The higher the number of the extracted features results in the higher the precision of object detection. Research shows that using deep learning algorithms can significantly improve the precision of object detection systems and help achieve close-to-human precision. Deep learning also facilitates feature extraction from images without human intervention, which is an outstanding contribution and a critical advantage. As mentioned, the convolutional neural networks are perfect for object detection. The proposed method is a hybrid method with a core founded on the ResNet [11] deep neural network. This network has been designed and implemented in various depths. The details can be found in [11].

ResNet50, which is 50 layers deep, was selected among all ResNet layers. To improve the precision and quality of feature extraction, an FPN was designed based on [12], which was a hybrid of ResNet50 and a set of convolutional layers.

Related Work

Many studies address object detection, more specifically, to improve the precision of the existing systems and approximate neural networks to the human recognition system. Modern object detectors rarely can reach high-speed and precise inference with a short training. The TTFNet network has been proposed to balance precision, speed of inference, and training time [13]. Detectors with high-speed inference operators directly need less training time. Highly accurate detectors fall into detectors with low inference speeds and detectors requiring long training time. Authors in [13] stress the importance of shortening training time while maintaining the performance of well-known detectors. Since the time required for feature encoding and loss calculation is insignificant compared to feature extraction time. They adopt the training sample encoding approach to help enhance the learning rate and accelerate the training process. Authors in [14] propose a simple yet effective method called progressive self-knowledge distillation (PS-KD). This method employs its predictions as a model of trainer knowledge to strengthen the generalization performance of deep neural networks. In this method, the system plays the role of a learner that turns into a trainer over time. In this regard, the objectives are adjusted by combining the main background and the previous model predictions. The extensive research on image classification, object detection, and machine translation indicate improved performance by using this method. CNNs often encode an input image into a set of intermediate features. Although this structure is suitable for classification tasks, it fails to perform optimally in tasks requiring simultaneous detection and localization, such as object detection. The encoder-decoder architectures have been proposed as a solution. They are used by applying a decoder network on a backbone model. It has been noted that due to the decrease in the scale of the backbone, encoder-decoder architectures do not influence the establishment of robust multi-scale features. Accordingly, SpineNet, a backbone with scale-permuted features, is introduced [15]. The authors seek the answer to this question: Is the scale-permuted model a suitable backbone architectural design for simultaneous detection and localization? Intuitively, scale-permuted backbone architecture exterminates spatial information with down-sampling, challenging the retrieval of a decoder network. Object detection [16] is defined by estimating a very large but extremely sparse bounding box dependent probability distribution. This article introduces two new concepts: a corner-based region-of-interest estimator and a deconvolution-based CNN model. Most object detectors are based on the anchor mechanism. To evaluate the alignment between the anchors and objects, they depend on calculating intersection over union parameters between the predefined anchor and real bounding boxes of objects. Authors in [17] question this type of using the intersection of union (IOU) and propose a new anchor alignment criterion. Anchors are a set of predefined reference boxes of a certain height and width. They are tiled across the image and help the network manages scale and object form changes by converting the object detection problem to a regression problem and classifying the anchor boundary box. This article proposes a mutual guidance mechanism that establishes an adaptive alignment between anchors and objects. The most modern object detection convolutional architectures are designed manually. In [18], the aim is to achieve a better architecture than an FPN for object detection. This article explores neural architectures and finds a new FPN in a new scalable searching space. This architecture is known as NAS-FPN, which is a combination of bottom-up and top-down connections. One-stage detectors simultaneously predict the classification time of objects and changes in the regression of the predefined boxes. This structure suffers from several flaws despite its good performance: The result of the predefined classification is inappropriately assigned to the regression during reasoning. Also, only one-time regression does not suffice for precise object detection. The present study first proposes a new module known as Reg-Offset-Cls (ROC) to solve the problem. The proposed module consists of three stages: bounding box regression, predicting the feature sampling location, and bounding box regression classification. Also, the hierarchical shot detector (HSD) detector was proposed to solve the second problem. This detector consists of two ROC modules and a feature-enhanced module [19]. Another study presents a systematic investigation of neural networks architecture designed for object detection and proposes several major optimizations to improve system performance. This article first proposes a bi-directional feature pyramid network (BiFPN) that allows the convenient and fast combination of multi-scale features. Next, a hybrid scaling method is proposed. The authors use the EfficientNet backbone to develop a new family of object detectors known as EfficientDet [20] . Humans perceive the world through sight, hearing, touch, and past experiences. Human experiences are taught by normal or unconscious learning. Authors in [21] propose an integrated network called YOLOR to encode implicit and explicit knowledge. Similar to the human brain, which is capable of conscious and unconscious knowledge acquisition, this integrated network can set up a display of simultaneous performance of tasks. This article summarizes how to design an integrated network that interpolates implicit knowledge into explicit knowledge. A systematic investigation of copy-paste indicates the sufficiency and acceptable performance of the simple mechanism of -randomly pasting the objects [22]. In [23], a large set of untagged images have been utilized for object detection. Authors study only several tagged images in each class. This method is known as multi-sample object detection. This procedure is iterated between the training model and the highly reliable sampling. In the training process, convenient samples are created first. Then, the initial weak model can improve. Over time, more reliable samples are selected, followed by another round of model improvement. The introduced framework is referred to as multi-modal self-paced learning for detection (MSPLD) [23]. An accurate, flexible, and completely anchor-free framework for object detection has been introduced [24].

Table 1: Related works

Method	Brief description of the method
TTFNet [13]	The TTFNet network has been proposed to balance precision, speed of inference, and training time.
PS-KD [14]	This method employs its predictions as a model of trainer knowledge to strengthen the generalization performance of deep neural networks.
SpineNet [15]	In this method SpineNet, a backbone with scale-permuted features, is introduced.
DeNet 101 [16]	In this method Object detection is defined by estimating a very large but extremely sparse bounding box dependent probability distribution.
Localize [17]	This article proposes a mutual guidance mechanism that establishes an adaptive alignment between anchors and objects
NAS-FPN AmoebaNet [18]	This article explores neural architectures and finds a new FPN in a new scalable searching space.
HSD [19]	Since the one-time regression does not suffice for precise object detection, HSD detector was proposed to solve this problem.
EfficientDet-D7x [20]	This article proposes a bi-directional feature pyramid network and a hybrid scaling method.
YOLOR-D6 [21]	Authors in this paper propose an integrated network called YOLOR to encode implicit and explicit knowledge.
Cascade Eff-B7 NAS-FPN [22]	A systematic investigation of copy-paste indicates the sufficiency and acceptable performance of the simple mechanism of -randomly pasting the objects.
MSPLD [23]	In this article a large set of untagged images have been utilized for object detection.
FoveaBox [24]	In this article an accurate, flexible, and completely anchor-free framework for object detection has been introduced.

Although almost all the highly advanced object detectors use predefined anchors, their proposed method directly learns the probability of the existence of an object and the bounding box coordinates without the anchor. This procedure involves two stages: the - prediction of class-sensitive semantic maps to measure the mentioned probability and the production of class-bounding boxes for each location with a potential object. In Table 1, related works are briefly stated.

The Proposed Method

Object classification and location are two critical steps in an object detection system design. The proposed method in this study is CNN-based and one-stage, with a ResNet50 core that can localize and classify the objects in the image. To achieve higher efficiency and better feature extraction, the FPN used here was combined with ResNet50. With the usage of a one-stage detector in this study, this question may arise: Is a simple one-stage detector able to achieve the same precision as a two-stage detector, as they have been known for their good performance? It should be noted that one-stage - detectors are applied to the regular and dense samplings of objects' locations scales. Recent research on one-stage detectors indicates that new one-stage detector designs are ten to forty percent more accurate than advanced two-stage methods [25]. Accordingly, the proposed method is a one-stage system with a ResNet50 core [11]. Moreover, the method involves an FPN to improve the feature extraction process and enhance the precision of the object detection system.

In general, the proposed method consists of two stages:

A) The training stage: Before the proposed system begins object detection, a convolutional neural network should be trained on a dataset.

B) The object detection stage: at this stage, the desired images (test images) are fed to the system. The system then detects the objects based on the training received in the previous stage. Below is the pseudocode of the proposed method.

Network ResNet

It is one of the deepest available architectures. It was introduced by Microsoft in 2015. With a depth of 152 convolutional layers and one fully-connected layer, this architecture has been recognized as the superior architecture in many competitions. It has also been able to reduce the ISLVRC challenge error by 3.57% [11]. This architecture has various layer depths discussed in [11]. The proposed method uses the Resnet50 structure. The number (50) indicates the layer depth of this architecture. (See Table 2). The ResNet architecture is one of the deepest architectures, which is presented in different depths (18, 34, 50, etc.) [11].

Algorithm 1: The pseudocode for proposed method

```
START
Downloading the dataset.
Implementing utility functions.
    coordinates of the corners
    coordinates of the center and the box
Computing pairwise Intersection Over Union (IOU.
Implementing Anchor generator.
Resizing The Input IMAGE (1333*800)
Encoding labels.
Building the ResNet50 backbone.
Building Feature Pyramid Network(FPN).
Building the classification and box regression head.
Implementing decode predictions.
    confidence_threshold=0.05,
    nms_iou_threshold=0.5,
Implementing losses.
Setting up training parameters.
Initializing and compiling model.
Setting up callbacks.
Load the dataset.
Training the model And Loading weights.
Building inference model.
Generating Object detections.
    END.
```

Table 2: The layer details of ResNet50 [11]

Layer name	Output size	50-layer
Conv1	112×112	7×7, 64, stride 2
Conv2_x	56×56	3×3 max pool, stride 2 $\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
Conv3_x	28×28	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$
Conv4_x	14×14	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$
Conv5_x	7×7	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
	1×1	Average pool, 1000-d fc, softmax

Before this architecture was introduced, other architectures had shown that increasing network depth has a direct effect on network efficiency and increases network efficiency but the problem was that in those architectures due to the problem of vanishing gradient. The depth of the network could be increased to a certain extent. In order to increase the depth of a shallow network, we need the identity layer. ResNet is a residual deep learning framework whose architecture is such that the identification of the layers is easily performed.

Therefore, by using this architecture, the depth of the network can be increased. In addition, ResNet's architectural structure is designed in such a way that there is an input from the previous step in each step, and this causes that better feature maps to be produced. Since in the proposed method we combine a feature pyramid network with the ResNet architecture and this increases the depth of the network, therefore, we choose the depth of 50 (ResNet50) among the different structures of the ResNet architecture so that the training of the network does not take too long and object detection is faster performed.

Feature Pyramid Network

FPNs can be applied to extract feature maps of images more efficiently. The proposed FPN used in [12] has a top-down architecture with lateral connections to create feature maps. This network significantly improves the feature extraction process. Multi-scale object detection is a major challenge in computer vision. This pyramid can enable a given model to detect objects in a wide range of scales by scanning locations and pyramid surfaces, hence the use of feature pyramids in the proposed method. The reported method in [12] purposes using the pyramidal and hierarchical structure of a convolution network. To achieve this goal, a structure has been relied on that combines low-resolution and semantically strong features with high-resolution and semantically weak features in a top-down manner. This pyramid can enable a given model to detect objects in a wide range of scales by scanning locations and pyramid surfaces, Therefore, following the ResNet network, we implemented a feature pyramid network to extract features more accurately.

Eight two-dimensional convolutional layers have been incorporated into the proposed method to develop this network (See Table 3).

Table 3: The convolutional layer details used in creating the FPN used

Padding	Stride	Dimensions	Number of filters	Layer name
Same	1	1×1	256	Conv 1
Same	1	1×1	256	Conv 2
Same	1	1×1	256	Conv 3
Same	1	3×3	256	Conv 4
Same	1	3×3	256	Conv 5
Same	1	3×3	256	Conv 6
Same	2	3×3	256	Conv 7
Same	2	3×3	256	Conv 8

Fig. 1 shows the structure of the core of the proposed system, and as it is known, the core of the system in our proposed method is created by combining a deep convolutional neural network called ResNet [11] with a feature pyramid network. This structure makes the extraction of features better and ultimately improves the accuracy of the object recognition system.

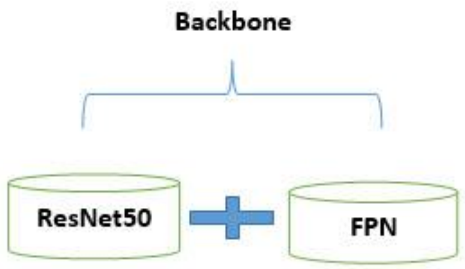

Fig. 1: The proposed system's core.

Image Resizing

The dimensions of the image change in a way that the length of the shorter side of the image equals 800 pixels. If the longer side of the image equals 1333 pixels after image resizing, the image size changes so that the longer side length equals 1333 pixels. In other words, in this stage, a minimum image side of 800 pixels and a maximum side length of 1333 pixels are defined.

Raw labels

Raw labels, including bounding boxes and class attributes, are applied in network training. This operation consists of the following stages:

- Creating anchor boxes in proportion to the database image dimensions.
- Assigning the ground truth box of objects (the main box refers to the actual box of every object in the image) to anchor boxes.

Here, the ground truth box of objects is assigned to anchor boxes based on overlapping. To that end, the IOU should be calculated among all anchor boxes during the training time and the ground truth box of the objects. The comparison based on IOU as a block diagram depicted in Fig. 2.

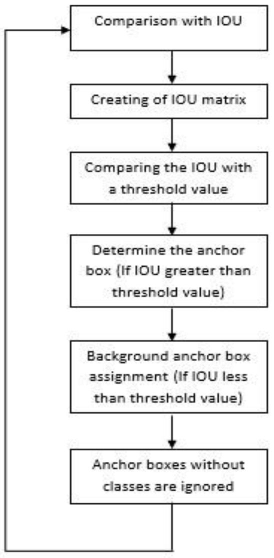
Comparison with IOU
Creating of IOU matrix
Comparing the IOU with a threshold value
Determine the anchor box (If IOU greater than threshold value)
Background anchor box assignment (If IOU less than threshold value)
Anchor boxes without classes are ignored

Fig. 2: The IOU block diagram.

System Training and Feature Extraction

In the training process, the system seeks to identify the best unknown parameters, such as the weight of convolutional filters and coefficients of the fully-connected layers, to achieve the least classification error rate. The proposed method uses the back propagation of errors and stochastic gradient descent (SGD) methods to update the weights in each iteration.

In stochastic gradient descent, the weights in each iteration are updated according to (1) . Equation (2) is the simpler re-expression of (1).

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i)}; y^{(i)}) \quad (1)$$

$$x = x - (\text{Learning\_rate}) * dx \quad (2)$$

where x is the order of parameters and weight of the filters and Learning_ rate indicates the rate at which the network is trained, determined at the beginning of the training process . It should be noted that the training rate parameter is modifiable. After a sufficient number of iterations, the network is trained to classify the database images.

After the training, the features of the dataset images should be extracted, which is a critical step as classification is performed based on these features and the rectangular boxes of objects (called windows, showing the location coordinates of objects). Finally, when an image is fed to the convolutional network, it crosses the convolutional layers, leading to a vector output. This output is known as the feature vector. By feeding all the images in the dataset to the network, there will be a set of feature vectors. Each vector is provided to the fully-connected layer for classification. The fully-connected layer in the proposed method here consists of 1000 neurons. Therefore, there will ultimately be 1000 specific classifications of database images. Generally, to summarize the first stage, the designed system is trained based on a set of datasets in the first stage. The optimum weights of filters and network parameters of the convolutional network, which are set to the optimal state to minimize classification errors, are then calculated to finalize the classification of the dataset images. This data is used to detect the existing objects in a new image. Next, for object detection, the test images are fed to the system. With the passing of images from the trained system, the feature vector of each image is extracted. Based on the classification set in the previous stage, the fully-connected layer determines the class of each object.

Activation Functions

Activation functions are essential in neural networks, playing a vital role in the network. The role of these functions is to make the optimization problem non-linear. More specifically, the activation functions decide whether or not a specific neuron in neural networks should be activated, determining to which category or class the output of the neural network belongs. Activation functions fall into various types, such as sigmoid, Tanh hyperbolic, and ReLU. The proposed method uses the ReLU and sigmoid functions as given by (3) and (4), respectively.

$$f(x)=\max(0,x) \quad (3)$$

$$\sigma(x) = 1/(1+e^{-x}) \quad (4)$$

Results and Discussions

This section discusses the evaluation criteria, datasets, and implementation results. Also, it compares the numerical results of the proposed method with some recent methods.

Evaluation Criteria

Here, the mAP criterion was used, as it is one of the most common and important evaluation criteria in object detection.

Intersection Over Union (IOU)

IOU is an evaluation criterion used to examine the precision of the predicted bounding box according to the ground truth box of the objects. Its value indicates the overlapping area between the predicted box and is a number that is the ground truth box of the object. In the case that this value is greater than a specific threshold, the system recognizes that anchor box as a predicted object. Therefore, this criterion determines the location precision by comparing the overlap between the ground truth box and the predicted bounding box. The determined value is between zero and one. The detection of an object is regarded as true when its IOU value exceeds a predefined threshold value. The usual threshold value is 0.5. When the predicted box overlaps with a ground truth box of more than 50%, the diagnosis made is considered valid. IOU is expressed as the follows:

$$\text{IOU} = (\text{area of overlap})/(\text{area of union}) \quad (5)$$

where the area of overlap is the intersection between the two boxes that also indicates the level of overlap. Area of union indicates an area with zero overlap between the two boxes [26].

The Precision and Recall Criteria

Precision (P) and recall (R) are used to evaluate the -classification ability of a method in question. In that regard, the values of the following parameters should be determined: true positive (TP) diagnosis, false positive (FP) diagnosis, and false negative (FN) diagnosis. These parameters are estimated based on location precision, -which is calculated based on IOU. Based on the IOU threshold values, then the values of TP and FP are calculated for the detected objects. The values of P and R are determined per each detection (true or false) based on TP and FP values using the Equations below [26]:

$$P = \frac{TP}{TP + FP} = \frac{TP}{\text{all detections}} \quad (6)$$

$$R = \frac{TP}{\text{all ground truths}} = \frac{TP}{TP + FN} \quad (7)$$

Average Precision (AP) and Mean Average Precision (Map)

The average precision criterion is calculated according to the values of P and R. This parameter is the average precision obtained per recall. The value of mAP is then -calculated by determining the mean of the AP value in different classifications [26], [27].

Datasets

To assess the performance of every object detection system, they should be run on suitable databases, and the standard criteria should be calculated accordingly. In this study, the Pascal VOC and COCO datasets were used for this purpose.

The Pascal Visual Object Classes (VOC) Dataset

This dataset was introduced in 2005. It comprises two sections: the first section consists of various image classes. The second section includes the annual tournaments. It consists of five challenges: classification, diagnosis, segmentation, action classification, and person layout. The various solution methods used in different competitions are discussed and compared in a workshop. The tournament was held from 2005 to 2012. Fig. 3 shows several sample images of this dataset [28].

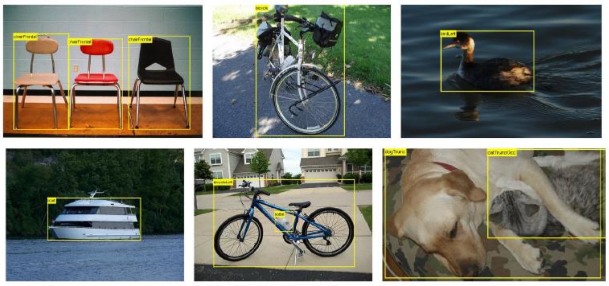

Fig. 3: Sample images of the Pascal VOC dataset [28].

The Microsoft Common Objects in Context (MS COCO) Dataset

Microsoft researchers introduced this database in 2014. It consists of 91 classes of objects and many samples and tagged images. Compared to Pascal VOC, it has fewer feature categories and samples in each category.

This dataset addresses three core problems: detection of non-iconic view of objects, contextual reasoning between objects, and precise two-dimensional location of objects.

Fig. 4 presents a sample image of images in this dataset [29].

Fig. 4: Sample images of the MS COCO dataset [29].

The Implementation Results

This section is a discussion of simulation results performed on two important object detection datasets. First, the average precision criterion is separately applied as a sample to each class of the existing objects in both datasets. The mAP of the proposed method is then compared with some recent methods.

Table 4 shows that the proposed method achieved a better average precision in 10 sample classes, such as Bird, Dog, Cat, and Person. Then, by calculating the precision of all classes and obtaining their means, the mAP value of the proposed method is calculated. This criterion indicates the performance of the object detection system.

Table 4: The average precision of the proposed method in ten different classifications in the Pascal VOC 2007 and MS COCO datasets

Class	Average precision (%)	
	Voc 2007 dataset	MS COCO dataset
Person	97.43	85.08
Cat	96.91	71.42
Dog	94.43	73.80
Bird	98.40	87.43
Train	91.23	67.42
Car	98.29	90.06
Motor	93.55	79.32
Bus	95.14	76.83
Cow	92.43	72.18
Sheep	94.27	85.1

Table 5 compares the proposed method with several recent methods based on the values of the mAP criterion.

As shown, the proposed method displayed a good object detection performance. The method proposed yielded a mean average precision (mAP) of 41.91 in the Pascal Voc2007 and 60.07% in COCO. Fig. 5 and Fig. 6 display samples of the detected objects by the proposed method in Pascal VOC and MS COCO datasets.

Table 5: The results of the mAP criterion of the proposed method and several recent methods in the Pascal VOC 2007 and MS COCO datasets

Method Name	Pascal VOC mAP%	Method Name	MS COCO mAP%
PS-KD [14]	79.7	TTFNet [13]	35.1
DeNet-101 [16]	77.1	SpineNet-49 [15]	45.3
Localize [17]	81.50	UniverseNet-20.08s [30]	47.4
HSD [19]	83.00	NAS-FPN AmoebaNet [18]	48.3
ReCoR [31]	83.90	EfficientDet -D7x [20]	55.1
Cascade Eff-B7 NAS-FPN [22]	88.6	YOLOR-D6 [21]	57.3
FoveaBox [24]	76.60	MSPLD [23]	56.6
proposed method	91.41	proposed method	60.07

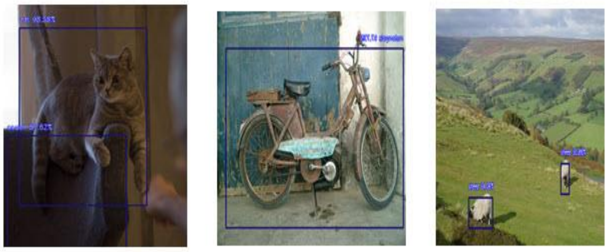

Fig. 5: Samples of objects detected in Pascal VOC.

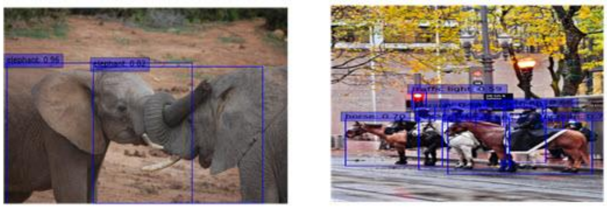

Fig. 6: Samples of objects detected in MS COCO.

The Proposed System Under Noise

To better evaluate the proposed model, some noise was introduced to the proposed system to examine its performance under noise. The test images of the datasets were combined with the salt and pepper noise [32] to obtain the value of mAP for different noise levels. Fig. 7 shows the mAP value against noise levels.

As shown, even at 0.3 noise level (for instance, when 30% of pixels of the image have been ruined by noise), the detection precision is above 50%. In real imaging situations, usually, the noise generated on the images is less than 0.05., hence the acceptable performance of the proposed model even in dealing with this level of noise. To obtain an intuitive understanding of the noise level, Fig. 8 display the output of the proposed model under different noise levels.

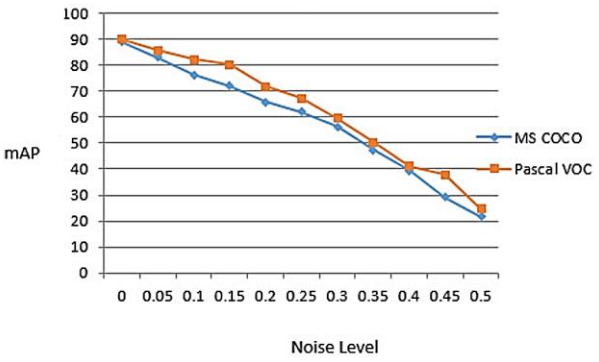

Fig. 7: The mAP diagram based on the noise level.

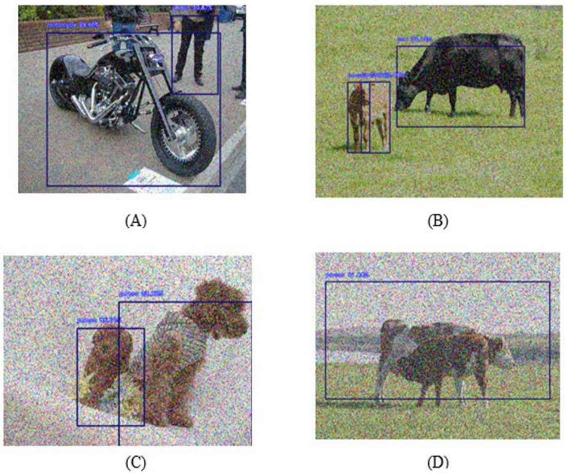

Fig. 8: Object detection with different noise levels: (A) 0.1, (B) 0.25, (C) 0.35 and (D) 0.5.

Conclusion

Since object detection is widely used in various fields such as medicine, self-driving cars, radar images processing and etc, the aim of this article is to implement a method for object detection with high accuracy. In this research, we studied the new articles which were presented in the topic of object recognition in the last few years and we found this reality that the object detection by convolutional neural networks is superior to other methods, but still the existing methods can be improved in terms of accuracy and speed. Therefore, our most important goal in this article was to present a method based on deep learning, using convolutional neural networks to recognize objects so that the proposed method detects objects with the least error in the shortest possible time. In general, object detection methods are classified into single-stage and two-stage detectors, and two-stage detectors are often more accurate but slower than single-stage detectors. Here, a one-stage object detection method was implemented using CNNs with a ResNet50 core. An FPN was used to improve the quality of the feature extraction process. The proposed system was then trained and evaluated with MS COCO and Pascal VOC2007 datasets. In the end, the value of mAP, one of the most notable criteria for performance evaluation of object detection systems, was calculated.

The proposed method was evaluated against seven other methods based on the mAP value obtained in each dataset. Moreover, a specific noise level was added to further investigate the system's performance in different noise level scenarios and mAP values.

The results indicated the better performance of the proposed method. Therefore, since the proposed system was trained using a deep-learning algorithm, the features of images were extracted hierarchically from the input images. The extracted data were then combined with an FPN. This combination improved the proposed method's detection precision. Accordingly, it can be concluded that using deep learning algorithms and CNNs and combining them with a feature network can significantly enhance object detection precision. And finally, considering the importance and many applications of object recognition a lot of research can be done in this field. Some examples of future work are as follows:

- Implementation of the proposed method presented in the article with other architectures and comparison with the results of our method.
- Investigating the use of noise reduction methods in object recognition.
- Implementation of the proposed method in order to object detection in the video.
- Designing a system by combining current object detection methods which may lead to improved performance.

Author Contributions

Prof. Hassan Farsi and Dr. were the supervisor and co-supervisor of the current research plan. They sketched the research framework and the roadmap. Also, they analyzed the results and tabulated the outcome derived from excerpted literatures. In this line, Seyed Mojtaba Notghimoghadam searched in authentic journals to gather all relevant papers. In addition to, he prepared the blueprint of the research plan.

Acknowledgment

This work is completely self-supporting, thereby no any financial agency’s role is available.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviation

CNN	Convolutional Neural Network
FPN	Feature Pyramid Network
MAP	Mean Average Precision
ResNet	Residual Network
COCO	Common Objects in Context
MS COCO	MicrosoftCommon Objects in Context
VOC	Visual Object Classes
RBM	Restricted Boltzmann Machine
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
TTFNet	Training Time Friendly Network
PS-KD	Progressive Self-Knowledge distillation
IOU	Intersection Of Union
HSD	Hierarchical Shot Detector
BIFPN	Bi-directional Feature Pyramid Network
YOLOR	You Only Learn One Representation
MSPLD	Multi-modal Self-Paced Learning for Detection
DeNet	Danish Ethernet Network
AR	Average Precision

References

[1] Z. Zou, Z. Shi, Y. Guo, J. Ye, "Object Detection In 20Years: A Survey," arXiv preprint arXiv: 1905.05055, 2019.

[2] Z. Zhao, P. Zheng, S. Xu, X. Wu, "Object detection with deep learning: A Review," IEEE Trans. Neural Networks Learn. Syst., 30(11): 3212-3232, 2019.

[3] Y. Wu, J. Feng, "Development and application of artificial neural network," Wireless Pers. Commun., 102(2): 1645-1656, 2018.

[4] R. Nasiripour, H. Farsi, S. Mohamadzadeh, "Visual saliency object detection using sparse learning," IET Image Proc., 13(13): 2436-2447,2019.

[5] S. Pasban, S. Mohamadzadeh, J. Zeraatkar-Moghaddam, A. Shafiei, "Infant brain segmentation based on a combination of VGG-16 and U-Net deep neural networks," IET Image Proc., 14(17): 4756-4765, 2021.

[6] Z. Dorrani, H. Farsi, S. Mohamadzadeh, "Image edge detection with fuzzy ant colony optimization algorithm," Int. J. Eng., 33(12): 2464-2470, 2020.

[7] C. Seale, T. Redfern, P. Chatfield, C. Luo, k. Dempsey, “Coastline detection in satellite imagery: A deep learning approach on new benchmark data,” Remote Sens. Environ., 278: 113044, 2022.

[8] K. Zeng, Y. Wang,” A deep convolutional neural network for oil spill detection from spaceborne SAR images,” Remote Sens., 12(6): 1015, 2020.

[9] H. Aliakbari, A. Abdipour, A. Costanzo, D. Masotti, R. Mirzavand, P. Mousavi, "ANN-Based design of a versatile millimetre-wave slotted patch multi-antenna configuration for 5G scenarios," IET Microwaves Antennas Propag, 11(9): 1288-1295, 2017.

[10] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M. Lew, "Deep learning for visual understanding: A review," Neurocomputing, 187: 27-48, 2016.

[11] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in Proc. the IEEE conference on computer vision and pattern recognition: 770-778, 2016.

[12] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature pyramid networks for object detection," in Proc. the IEEE conference on computer vision and pattern recognition: 2117-2125, 2017.

[13] Z. Liu, T. Zheng, G. Xu, Z. Yang, H. Liu ,D. Cai, "Training-time-friendly network for real-time object detection," in Proc. the AAAI Conference on Artificial Intelligence, 34(07): 11685-11692, 2020.

[14] K. Kim, B. Ji, D. Yoon, S. Hwang, "Self-Knowledge distillation with progressive refinement of targets," in Proc. the IEEE/CVF International Conference on Computer Vision: 6567-6576, 2021.

[15] X. Du, T. Lin, P. Jin, G. Ghiasi, M. Tan, Y. Cui, Q. Le, X. Song, "SpineNet: Learning scale-permuted backbone for recognition and localization," in Proc. the IEEE/CVF Conference on Computer Vision And Pattern Recognition: 11592-11601, 2020.

[16] L. Tychsen-Smith, L. Petersson, "Denet: Scalable real-time object detection with directed sparse sampling," in Proc. the IEEE international Conference on Computer Vision: 428- 436, 2017.

[17] H. Zhang, E. Fromont, S. Lefèvre, B. Avignon, "Localize to classify and classify to localize: Mutual guidance in object detection," in Proc. the Asian Conference on Computer Vision, 2020.

[18] G. Ghiasi, T. Lin, Q. Le, "Nas-Fpn: Learning scalable feature pyramid architecture for object detection," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 7036-7045, 2019.

[19] J. Cao, Y. Pang, J. Han, X. Li, "Hierarchical shot detector," in Proc. the IEEE/CVF International Conference on Computer Vision: 9705-9714, 2019.

[20] M. Tan, R. Pang, Q. Le, "Efficientdet: Scalable and efficient object detection," in Proc. the IEEE/CVF conference on computer vision and pattern recognition: 10781-10790, 2020.

[21] C. Wang, I. Yeh, H. Liao, "You only learn one representation: unified network for multiple tasks," arXiv preprint arXiv: 2105.04206, 2021.

[22] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T. Lin, E. Cubuk, Q. Le, B. Zoph., "Simple copy-paste is a strong data augmentation method for instance segmentation," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2918-2928, 2021.

[23] X. Dong, L. Zheng, F. Ma, Y. Yang, D. Meng, "Few-Example object detection with model communication," IEEE Trans. Pattern Anal. Mach. Intell., 41(7): 1641-1654, 2018.

[24] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, J. Shi, "Foveabox: Beyound anchor-based object detection," IEEE Trans. Image Proc., 29: 7389-7398, 2020.

[25] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, "Focal loss for dense object detection," in Proc. the IEEE International Conference on Computer Vision: 2980-2988, 2017.

[26] P. Rafael, S. L. Netto, E. Silva, "A survey on performance metrics for object-detection algorithms," in Proc. 2020 International Conference on Systems, Signals and Image Processing (IWSSIP): 237-242, 2020.

[27] X. Wu, D. Sahoo, S. Hoi, "Recent advances in deep learning for object detection," Neurocomputing, 396(7): 39-64, 2020.

[28] M. Everingham, S. Eslami, L. Gool, C. Williams, J. Winn ,A. Zisserman, "The pascal visual object classes challenge: A retrospective," Int. J. Comput. Vision, 111(1): 98-136, 2015.

[29] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C. Zitnick., "Microsoft COCO: common objects in context," in Proc. European Conference on Computer Vision: Springer, Cham: 740-755, 2014.

[30] Y. Shinya, "USB: Universal-scale object detection benchmark," arXiv preprint arXiv: 2103.14027, 2021.

[31] Z. Chen, J. Zhang, D. Tao, "Recursive context routing for object detection," Int. J. Comput. Vision, 129(1): 142-160, 2021.

[32] J. Azzeh, B. Zahran, Z. Alqadi, "Salt and pepper noise: Effects and removal," JOIV: Int. J. Inf. Visualization, 2(4): 252-256, 2018.

Biographies

Seyed Mojtaba Notghimoghadam received the B.Sc. degree in electrical engineering from the Islamic Azad University of Birjand, Birjand, Iran, 2012. He received the M.Sc. degree in telecommunication engineering from University of Birjand, Birjand, Iran, in 2022. He is currently Ph.D. student in university of Birjand, Birjand, Iran. His research interests in digital image processing, visual signal processing, deep learning and artificial intelligence.

- Email: m.notghimoghadam@birjand.ac.ir
- ORCID: 0000-0002-7320-929X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Hassan Farsi received the B.Sc. and M.Sc degrees from Sharif University of Technology, Tehran, Iran, in 1992 and 1995, respectively. Since 2000, he started his Ph.D in the Centre of Communications Systems Research (CCSR), University of Surrey, Guildford, UK, and received the Ph.D degree in 2004. He is interested in speech, image and video processing on wireless communications. Now, he works as professor in communication engineering in department of Electrical and Computer Eng., University of Birjand, Birjand, IRAN.

- Email: hfarsi@birjand.ac.ir
- ORCID: 0000-0001-6038-9757
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://cv.birjand.ac.ir/hasanfarsi/en

Sajad Mohamadzadeh received the B.Sc. degree in electrical engineering from Sistan & Baloochestan, University of Zahedan, Iran, in 2010. He received the M.Sc. degree in telecommunication engineering from university of Birjand, Birjand, Iran, in 2012. Now, he works as associate professor in department of Electrical and Computer Eng., University of Birjand, Birjand, IRAN. His area research includes image processing, deep learning, pattern recognition, digital signal processing and sparse representation.

- Email: s.mohamadzadeh@birjand.ac.ir
- ORCID: 0000-0002-9096-8626
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://cv.birjand.ac.ir/mohamadzadeh/en

How to cite this paper:
S. M. Notghimoghadam, H. Farsi, S. Mohamadzadeh, “Instructions and formatting rules for authors of journal of electrical and computer engineering innovations, JECEI,” J. Electr. Comput. Eng. Innovations, 11(1): 173-182, 2022.

DOI: 10.22061/JECEI.2022.9012.567

URL: https://jecei.sru.ac.ir/article_1769.html