ISSN 2347 - 3983
Volume 8. No. 7, July 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter140872020.pdf
https://doi.org/10.30534/ijeter/2020/140872020

An Extensive Review on Recent Evolutions in Object Detection Algorithms

S. Nandhini[1], D. Easwaramoorthy[2,*], R. Abinands[3]
[1,2]Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Vellore - 632 014, Tamil Nadu, India.
[3]Department of Electrical Engineering, School of Electrical Engineering, Vellore Institute of Technology, Vellore - 632 014, Tamil Nadu, India.
[1]nandhini.s@vit.ac.in, [2]easandk@gmail.com, [3]abinands.ramshanker2019@vitstudent.ac.in
*Corresponding Author.

ABSTRACT

The domain of Computer Vision Algorithms has achieved exemplary development in the past few years. Owing to its wide range of applications, especially in object detection, image classification and image segmentation, it has gained its own significance in this technological era. This paper provides a significant remark on fractal and multifractal methods for object identification in images and also a keen review of various Computer Vision algorithms and their features for detecting objects. We have also significantly analyzed the previous contributions on various algorithms discussed for the object detection process.

Key words: Object Detection, Fractal Analysis, Support Vector Machine, Histogram of Oriented Gradients, Deformable Parts Model, Convolutional Neural Networks, You Only Look Once.

1. INTRODUCTION

In the current scenario of enormous requirement of object detection process, Computer Vision (CV) plays a major role. CV is one of the most popular research topics in the world, since it is been employed in various aspects of human life which involves object detection. Object detection is an important area of computer vision and has important applications in scientific research and practical industrial production, such as face detection, text detection, pedestrian detection, logo detection, video detection, vehicle detection, medical image detection and other object detections. In the field of object detection, recently, tremendous success is achieved, but still it is a very challenging task to detect and identify objects with speedy and accuracy. Human beings can detect and recognize multiple objects in images or videos with ease regardless of the object's appearance, but for computers it is challenging to identify and distinguish between things. In line with this scenario, enormous computer vision algorithms with applications have been considered and analyzed for detecting objects effectively by numerous researchers [1].

This article will serve as a guide to afford some scientific note on fractal and multifractal approaches for object identification in complex level images also to implement the Computer Vision techniques across various applications in detail. We have presented the overview of various Object Detection algorithms Support Vector Machine (SVM), Histogram of Oriented Gradients (HOG) and Deformable Parts Model (DPM). After analyzing the traditional Computer Vision algorithms, we look at Convolutional Neural Networks (CNN). We also analyze how object detection has evolved from Spatial Pyramid Pooling to Faster Region Convolutional Neural Network (R-CNN). Then we present the study of You Only Look Once (YOLO) algorithm and also highlight the Multi Scale Deformable Algorithm which stands to be the most efficient algorithm.

The organization of this paper is as follows. Section 2 gives an introduction and a summary of the object detection. Fractal and multifractal features for object detection is discussed shortly in Section 3. Section 4 presents a detailed review on various object detection algorithms, its working methodology and the past work done on those algorithms. Finally Section 5 stretches the concluding remarks with the future directions.

2. OBJECT DETECTION

Object detection, which is to determine and locate the object instances either from a large number of predefined categories in natural images or for a given particular object is an important and challenging task in computer vision. Object detection and image classification share a similar technical challenge: both of the technologies handle a large number of highly variable objects. However, object detection is more difficult than image classification, as it must identify the accurate localization of the object of interest. Generic object detection aims at locating and classifying existing objects in any one image and labeling them with rectangular BBs to show the confidences of existence. The frameworks of generic object detection methods can mainly be categorized into two types. One follows the traditional object detection pipeline, generating region proposals at first and then classifying each

proposal into different object categories. The other regards object detection as a regression or classification problem, adopting a unified framework to achieve final results (categories and locations) directly [2,3,4].

Abundant approaches have been proposed to solve the problem of identifying the objects, mainly inspired by methods of computer vision and deep learning. However, existing approaches always perform poorly for the detection of small, dense objects, and even fail to detect objects with random geometric transformations. Image understanding is a difficult problem even in its simplest form because, objects, based on a variety of factors, can have a wide range of intra-class variability. Beyond recognition, detection requires the localization of the object which can become a costly search problem of the image if not given any heuristics. Several computer scientists have vastly investigated the problem of detecting objects within in a realistic image and explored innumerable algorithms for the same problem in various applications [5,6,7,8,9].

3. FRACTAL MEASURES FOR OBJECT DETECTION

Object detection methods aim to identify all target objects in the representative image; and determine the categories and position information in order to achieve the understanding of machine oriented vision. Generally images are obtained by photo electronic or photochemical methods. Transmission process of acquired objects tends to corrupt the quality of the digital images by introducing noise. The existence of noise in an image may be a drawback in any subsequent processing to be done over the noisy image such as object detection, image segmentation, image classification or pattern recognition. As a consequence, restoring the image to reduce or remove the noise without degrading its quality is a major step in any computer vision application. Because the corrupted images have high complexity and irregularity in nature or in its pixel values, it is very difficult to identify and quantify the restoring images or noise free images by using quantitative measures.

Fractals have broad applications in non-linear dynamical systems, computer graphics, biomedicine and other applied sciences. The complexity and irregularity that can be found in many physical and biological non-linear systems naturally and which has been analyzed by the tools of fractal theory and computed by the measure called fractal dimension. In the literature, when fractal technique has been applied to the complex signals and images, the dimensional measure has mainly been used to analyze the chaotic nature in different conditions [10,11,12,13,14,15,16,17].

Typically natural images, especially color or multi component images, are complex information-carrying signals. To contribute to the characterization of this complexity, we have to investigate the possibility of multiscale organization in the colorimetric structure of natural images. This is realized by means of a multifractal analysis applied to the gray scale and the color images. Fractal and multifractal features so far have been applied essentially by the scientific researchers to the spatial organization of gray scale and color images [18,19,20].

Especially in 2010, Liu et al. [21] have stepped into a review of man-made object detection algorithms is presented based on various fractal features which are derived from the blanket covering method. These fractal features include fractal dimension, fractal model fitting error, D-dimension area, multi-scale fractal feature related with D, and multi-scale fractal feature related with K. The gained results have revealed that different fractal features have different capability in discriminating between natural and man-made objects, and MFFK has the highest detection accuracy among all evaluated fractal features. In addition, the non-integer fractal dimension has employed in lot of object detection in natural images, mobile video images, daytime land fog images, atomic force microscopy images, surface anomaly images and other satellite and thermal images [22,23,24,25,26].

4. COMPUTER VISION ALGORITHMS FOR OBJECT DETECTION

In this section, we review the algorithms of object detection with applications using the computer vision concepts. A widespread study of different algorithms using SVM, HOG,DPM, CNN and YOLO has also been presented for detecting the objects accurately.

4.1. Support Vector Machine (SVM)

Object detection process involves lots of factors that have to be taken into account for the object detection. SVM is one of the effective algorithms, which helps in classification of objects. SVM is a learning technique developed by V. Vapnik and his team (AT&T Bell Labs., 1985) that can be seen as a new method for training polynomial, neural network, or Radial Basis Functions classifiers. It looks at the extremes of the dataset and draws a decision boundary. This decision boundary is known as a hyperplane. It segregates the dataset into two groups. The problem arises in drawing the decision boundary; we can draw it in many ways using di erent angles. The optimal decision boundary is important to classify the different types of elements. All these boundaries are called as support vectors. In this scheme, *D*+ represents the vectors towards the positive direction from the hyperplane and *D*- represents the vectors towards the negative direction [27]. In the SVM algorithm, the margin between the data points and the hyperplane are expected to be maximized.

$$f(x) = (wx + b) \quad (1)$$

Here, $f(x)$ is the function of the hyperplane, w represents the slope of the line and b represents the y intercept.

There might be cases where it is almost impossible to separate the two classes. In these cases, Linear Support Vector Machine algorithm (LSVM) is used in which, we convert the one–dimensional plane to two–dimensional plane. We can also convert 2–*D* to 3–*D* and draw a hyperplane. This is

Non-Linear Support Vector Machine. The only disadvantage in this method is the requirement of high computational power.

The significant application of Support Vector Machines (SVMs) in computer vision was investigated by Osuna et al. in 1997 [28]. They have presented a decomposition algorithm that guarantees global optimality, and can be used to train SVM's over very large data sets. The main idea behind the decomposition is the iterative solution of sub-problems and the evaluation of optimality conditions which are used both to generate improved iterative values, and also establish the stopping criteria for the algorithm. The experimental results of their implementation of SVM and the feasibility of their approach on a face detection problem have been demonstrated with the experimental datasets.

Support vector machines (SVMs) were originally designed for binary classification of objects. How to effectively extend it for multiclass classification is still an ongoing research issue in the computer vision field. Several methods have been proposed where typically we construct a multiclass classifier by combining several binary classifiers. Some authors also proposed methods that consider all classes at once. As it is computationally more expensive to solve multiclass problems, comparisons of these methods using large-scale problems have not been seriously conducted. Especially for methods solving multiclass SVM in one step, a much larger optimization problem is required so up to now experiments are limited to small data sets. The multi-level decomposition implementations for two such all-together methods have been sculpted by Hsu et al. in 2002 [29].

In 2004, Melgani et al. [30]have addressed the problem of the classification of hyperspectral remote sensing images by support vector machines (SVMs). First, they have proposed a theoretical discussion and experimental analysis aimed at understanding and assessing the potentialities of SVM classifiers in hyper-dimensional feature spaces. Then, they have accessed the effectiveness of SVMs with respect to conventional feature-reduction-based approaches and their performances in hyper-subspaces of various dimensionalities. To sustain such an analysis, the performances of SVMs were compared with those of two other nonparametric classifiers (*i.e*., radial basis function neural networks and the K-nearest neighbor classifier). Based on the results obtained on a real Airborne Visible/Infrared Imaging Spectroradiometer hyperspectral dataset, the authors have concluded that, whatever the multiclass strategy adopted, SVMs are a valid and effective alternative to conventional pattern recognition approaches for the classification of hyperspectral remote sensing objects.

In the rapid growth of bio-metric technology, the face detection got much attention over the past few years. Face recognition describes a biometric technology that attempts to establish an identity. A facial recognition system using machine learning especially, using support vector machines has been reviewed by Riyazuddin et al. in 2020 [31].

4.2. Histogram of Oriented Gradients (HOG)

The histogram of oriented gradients (HOG) is a feature descriptor used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image. This algorithm works by having an input of an image with an aspect ratio of $1{:}2$. Then the image is converted into 64×128 by dimension. In some cases, Gamma correction can be used to improve the performance gain. Then the horizontal and vertical gradients are calculated by the functions $m$and $\emptyset$ given respectively in Eqns. (2) and (3) in terms of intensity functions f_u and f_v with the corresponding directions of vectors u and v.

$$m(u,v) = \sqrt{f_u(u,v)^2 + f_v(u,v)^2} \quad (2)$$

$$\emptyset(u,v) = tan^{-1}\frac{f_u(u,v)}{f_v(u,v)} \quad (3)$$

The magnitude of the gradient is taken for the entire image. After that, the gradients are found in each cell after splitting the image into 8×8 cells. Furthermore, an image patch would contain 192 pixels and it would be robust to noise. Then the histogram of gradients is created in these 8×8 cells. The histogram contains 9 bins corresponding to angles. The gradient magnitude is grouped corresponding to the gradient directions. The next step is to perform 16×16 block normalization. It will be an evident that the algorithm is not a□ected by lighting process. Finally the HOG feature vector is calculated. The obtained histogram demonstrates that the pixel from background give much lower accumulating result than the pixel from the object.

In 2005, Dalal et al. [32] have studied the feature sets for robust visual object recognition and adopted the linear SVM based human detection as a test case. After reviewing existing edge and gradient based descriptors, they have shown experimentally that grids of HOG descriptors significantly outperform existing feature sets for human detection. This approach has given the near-perfect separation on the original MIT pedestrian database over 1800 annotated human images with a large range of pose variations and backgrounds.

The cascade-of-rejectors approach with the HOG features has been integrated by Zhu et al. [33] in 2006 to achieve a fast and accurate human detection system. The features used in their method are HOGs of variable-size blocks that capture salient features of humans automatically. In their system, they have utilized the integral image representation and a rejection cascade which significantly speed up the computation.

In 2009, Chaudhry et al. [34] have proposed a technique to represent each frame of a video using a histogram of oriented optical flow (HOOF) and to recognize human actions by classifying HOOF time-series. For this purpose, they have exposed a generalization of the Binet-Cauchy kernels to nonlinear dynamical systems (NLDS) whose output lives in a non-Euclidean space, *i.e*., the space of histograms. They have examined their approach in the recognition of human actions in several scenarios.

The research works on action recognition has focused on adapting hand-designed local features, such as HOG, from static images to the video domain. The unsupervised feature learning as a way to learn features directly from video data has designed by Le et al. in 2011 [35]. More specifically, they presented an extension of the Independent Subspace Analysis algorithm to learn invariant spatio-temporal features from unlabeled video data. The ease of training and the efficiency of training and prediction have also been dealt as a benefit of this technique with respect to the realistic database.

In 2013, Oreifej et al. [36] have presented a new descriptor for activity recognition from videos acquired by a depth sensor. They have described the depth sequence using a histogram capturing the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates. To build the histogram, 4D projectors, which quantize the 4D space are created and represented the possible directions for the 4D normal. Through extensive experiments, the authors have demonstrated that their descriptor captures the joint shape-motion cues better in the depth sequence.

A concrete technique for human detection from video has been studied by Surasak et al. [37] in 2018, which is HOG by developing a piece of application to import and detect the human from the video. They have used the HOG Algorithm to analyze every frame from the video to find and count people. From this research work, the obtained results including the detection of people in the video and the histogram generation are the evidence to show the appearance of human detected in the video file.

Traffic signs are important markers in two-wheeled and four-wheeled vehicles. However, there is a change in direction or arrangement on the road that cannot be opened on a map which can cause incorrect information, which can cause traffic jams. In 2019, Reinaldo et al. [38] have used a camera mounted on a car that provides a solution for drivers who issue problems that occur on the road that show directions or arrangements that are not directly updated using the Histogram of Oriented Gradients (HOG) and Max Margin Object Detection (MMOD) methods. They have suggested that the information received can be sent directly to an electronic map so that it can be accessed automatically by the driver's information and assistance and other information finds the right path, so that it can help the driver and can avoid traffic jams.

4.3. Deformable Parts Model (DPM)

Typically, SVM for classification and HOG for feature extraction can be used in the object detection methods. The main drawback for this image based algorithm with HOG and SVM is that it does not catch the object in certain poses or deformations. Humans are deformable and have many poses unlike non-living objects. Thus, the algorithm may fail to detect humans in certain poses. DPM is the algorithm which takes care of this aspect.

A deformable part model (DPM) is a method used for object detection in images that leverages the fact that objects are inherently made up of a collection of parts. Each part of an object is connected to one or more other parts in a treelike structure. These parts can vary in distance, orientation, or pose with respect to one another but, within some reasonable range, still be considered the skeleton of the same object. DPM compensate for this property of various objects by utilizing HOG features for object representation at coarse and fine scales, pictorial structures, and application of a deformation cost on that pictorial structure. As such, these models can allow for variations in object pose, shape, and viewpoints while still remaining a very specific representation of that object describing not only the object as a whole, but also each of its distinct parts and their spatial relationships.

DPM is a model that came about to solve the problem of object detection during different poses in images. To take the different poses into account, each body part is detected. But there could be a possibility that there may be multiple legs, arms or other body parts in the crowd. For solving this, the idea of penalty scores was introduced. If the body part is closer, the penalty is lesser. The score for the whole body is taken, then the score for each body part is added and finally the penalties are subtracted from the resultant value. SVM and HOG are typically used together for each body part and then they are summed up. There is a course filter for the entire object detected and there will be multiple higher resolution part filters for each part.

The model has root filter F_0 and n part models (F_i, v_i, d_i). The score of the hypothesis depends on the root filter and part filter. Eqn. (4) represents the sum of the root filter and part filters in the first term. The second term represents the penalties and b represents the bias. The root occurs at p_0 and the part occurs at $p_{1,}, \dots, p_n$. F_i represents the filters. $\phi(H, p_i)$represents the features of subwindows at location p_i. d_i represents the deformation parameters. ϕ_d is the displacement of the part *i* relative to its anchor position. The first term is referred to as the data term and the second term is referred to as the spatial prior.

$$\text{score}(p_0, p_1, \dots, p_n) = \sum_{i=0}^{n} F_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot \phi_d(dx_i, dy_i) + b \quad (4)$$

$\beta$is the set which contains all the unknowns which are the filters, deformation and the bias. $\phi$is the known term as mentioned above. It is the response of the algorithm for each part filter and the root filter.

$$\beta = (F_0, \dots, F_n, d_1, \dots, d_n, b) \quad (5)$$

$$\psi(H, z) = (\phi(H, p_0), \dots, \phi(H, p_n), -\phi_d(dx_1, dy_1), \dots, -\phi_d(dx_n, dy_n), 1) \quad (6)$$

We multiply the two terms $\beta$and $\psi(H, z)$ to compute the score.

$$\text{score}(z) = \beta \cdot \psi(H, z) \quad (7)$$

$\phi$is the known term as above andβ is the term which will be computed.

If we notice the deformation cost. It is a four dimensional vector. $\phi_d$is also a four dimensional vector.

$$d_i = (0,0,1,1) \quad (8)$$

$\emptyset_d(dx, dy) = (dx, dy, dx^2, dy^2)$ (9)

Initially when we take, the distance would be given by $dx^2 + dy^2$.

Depending on the values *dx* and *dy*, the locus of displacement would be a circle or ellipse.

The overall score of a root location is computed by the best possible placement of the parts. After training this model, wewant to make sure that the human detection is confirmed by the parts.

$\text{score}(p_0) = \max_{p_0,\dots,p_n} \text{score}(p_0, \dots, p_n)$(10)

Eqn. (10) represents the score at the root location p_0.

There might be cases in which a body part of another person might be considered by mistake. For avoiding this error, a dynamic programming and generalized distance transform is implemented. It is the process of finding the closest part, to avoid errors.

In 1991, Terzopoulos et al. [39] have developed a physically based approach to fitting complex three-dimensional shapes using a novel class of dynamic models that can deform both locally and globally. They have formulated the deformable superquadrics which incorporate the global shape parameters of a conventional superellipsoid with the local degrees of freedom of a spline. The authors have fitted a model to visual 2-D monocular image data and 3-D range data by transforming the data into forces and simulating the equations of motion through time to adjust the translational, rotational, and deformational degrees of freedom of the models.

The extraction of part of the visual information presented in streets, roads, and motorways plays a vital role in highways. This information, provided by traffic, road signs or route-guidance signs, is extremely important for safe and successful driving. An automatic system that is capable of extracting and identifying these signs automatically would help human drivers enormously; navigation would be easier and would allow the driver to concentrate on driving the vehicle. The system would indicate to the driver the presence of a sign in advance, so that some incorrect human decisions could be avoided. A deformable model scheme has been built to include the knowledge used for designing the signs in the algorithm and also used for their detection and identification by addressing some problems, such as uncontrolled lighting conditions; occlusions; and variations in shape, size, and color by Escalera et al. in 2004 [40].

A discriminatively trained, multiscale, deformable part model for object detection has been established by Felzenszwalbet at. in 2008 [41]. The developed scheme has achieved a two-fold improvement in average precision over the best performance in the 2006 PASCAL person detection challenge and also outperformed the best results in the 2007 challenge in ten out of twenty categories. It has been believed that their training methods will eventually make possible the effective use of more latent information such as hierarchical (grammar) models and models involving latent three dimensional pose.

In 2020, Felzenszwalb et al. [42] have described an object detection system based on mixtures of multiscale deformable part models. The outlined system is able to represent highly variable object classes and achieves state-of-the-art results in the PASCAL object detection challenges. The authors presented a new method for discriminative training with partially labeled data.

A general method has been presented for building cascade classifiers from part-based deformable models such as pictorial structures by Felzenszwalb et al. in 2010 [43]. The authors have mainly focused on the case of star-structured models and shown that how a simple algorithm based on partial hypothesis pruning can speed up the object detection by more than one order of magnitude without sacrificing detection accuracy. Finally, they have charted a cascade detection algorithm for a general class of models defined by grammar formalism. Weakly supervised discovery of common visual structure in highly variable, cluttered images is a key problem in object recognition. This problem using deformable part-based models with latent SVM training has been addressed by Pandey et al. in 2011 [44]. These models have been introduced by the authors for fully supervised training of object detectors, but they have demonstrated that it has capable of more open-ended learning of latent structure for such tasks as scene recognition and weakly supervised object localization.

In 2018, Pool et al. [45] have explored the state of the art, deformable part models (DPMs), and their applicability for complex object detection in very high-resolution satellite images. The authors have investigated the landscape of research regarding DPM, how this class of methods for object detection has evolved, and what remains to be explored to make the method more suitable for high-level, complex geospatial object understanding.

4.4. Convolutional Neural Networks (CNN)

Convolution Neural Networks (CNNs) have shown impressive performance in various vision tasks such as image classification, object detection and semantic segmentation. In the previous computer vision techniques, HOG algorithms acts as feature extractors while SVM acts as a classifier. In a Convolutional Neural Network, the convolutional pooling layers would act as feature extractors. The fully connected and softmax layers act as classifiers (Figure 1).

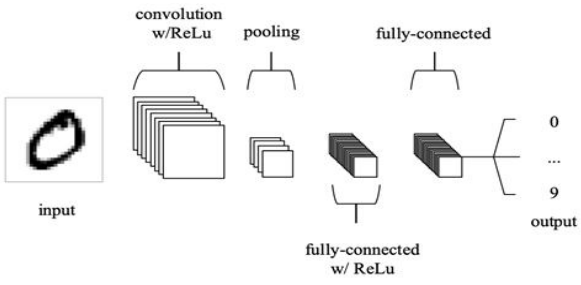

Figure 1: Convolution Neural Network Method

The classifier could also be modified to perform localization which is drawing a bounding box around the detected object. To draw the bounding box we need to get the coordinates of one point and get the height and width. We would have 4 parameters. The last layer of the convolutional layer consists of one fully connected layer followed by softmax. The layer would have the scores for each class. Softmax basically converts the scores to probabilities. In CNN, Bounding Box Regression Training is used to train to get the bounding box coordinates. We input an image along with the 4 coordinates required for building a bounding box. In the beginning, the layers of CNN would have the weights assigned as 0.1 for every layer. We pass the initial vector of features through the layers to obtain the four coordinates. We calculate the loss which is the difference between the squares of the expected coordinates and the generated coordinates. The value we get is back propagated through the layer. This leads to the value of the weights changing. The neural network is again processed to get 4 coordinates. Then L_2 loss is calculated again. This is done till we get the sum of the losses of each coordinate to be zero. After this, the CNN would be trained to draw a bounding box for the image (Figure 2).

We use the sliding Window technique to detect multiple and crop whenever the object is detected. This would be included as a preprocessing step. Then we convert that image into 224×224 to draw the bounding box. Finally we merge the cropped images to form the complete images with bounding boxes.

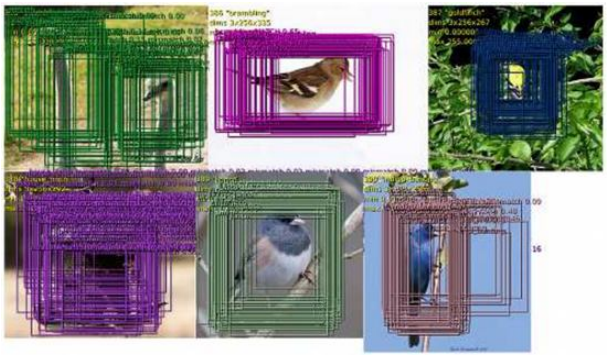

Figure 2: Object Detection by CNN Algorithm on Sample Images

In cases where objects are of different sizes and overlap, we use the sliding window combined with the image pyramid technique. The image pyramid represents the process of resizing the image continuously so that an object of different sizes gets detected at different scales. A confidence score is given to each of the detection. This would ensure that partial detections do not get cropped or fed to the CNN. Usually convolutional Neural Networks require a fixed image as the feature vectors would get multiplied at the final layer - the fully connected layer. Due to a lot of advancement, the fully connected layer is used as a convolutional layer. Instead of an array, it is considered as a matrix. So the sliding window technique can be used to detect the objects and it can be fed to the CNN without the requirement of cropping. Furthermore, repeated pixels will not run again and this would save computational power. We use image pyramid to get spatial output. Spatial output is the matrix of the confidence scores at each sliding window box. This process is the idea of the overfeat algorithm. The effective stride is the number of pixels in which the algorithm moves if 1 pixel is shifted in the spatial output. The effective stride should be as low as possible.

The standard CNN can be used to generate proposals and classify an image in different regions. The problem is that the object that is required to be found could have a different aspect ratio. Furthermore, it might have different spatial locations. To solve this, R-CNN can be used. Initially, the Selective Search Algorithm is used to extract just 2000 region proposals. So we work with just 2000 regions instead of an infinite number of different regions. These 2000 regions are then merged into a square. It is then fed into a CNN which generates 4096-dimensional feature vectors. It is then fed into an SVM algorithm to find if the object is present or not. The offset values adjust the boundary boxes of the region proposal. The disadvantage of this method is that, the Selective Search Algorithm is fixed and so no learning takes place. Now when we have image proposals classified on every object class, we bring the entire image back using greedy non-maximum suppression. Non-maximum suppression is just the process where the computer takes the intersection of union of each proposal and selects the region with the higher score.

In the case of building large convolutional neural networks, signal propagation speed is one of priority factors. Training large neural structures requires enormous time for achieving satisfying accuracy. Unlike the R-CNN, in the Fast R-CNN, object detection is made more efficient by avoiding feeding the region proposals to the CNN. Instead a convolutional feature map is obtained by feeding the image. Then RoI (Region of Interest) pooling layer is used to reshape them into fixed size and then we feed them into a fully connected layer. It is called Fast R-CNN because it is not required to feed 2000 region proposal to CNN. Faster CNN is set out to find a way to replace the techniques of Selective Search and Edge boxes with a Dense Sampling technique like sliding windows. The objects obtained from these CNN are either squares or rectangular.

Recent literature indicate that the generic descriptors extracted from the convolutional neural networks are very powerful. In 2014, Razavian et al. [46] have mounted a report on a series of experiments conducted for different recognition tasks using the publicly available code and constructed a model of the OverFeat network which was trained to perform object classification on ILSVRC13. The obtained results by the authors have strongly suggested that features obtained from deep learning with convolutional nets should be the primary candidate in most visual recognition tasks.

A Fast Region-based Convolutional Network method (Fast R-CNN) for object detection has been explored by Girshick in 2015 [47]. Compared to existing research works,

it has been proven that Fast R-CNN employs several innovations to improve training and testing speed while also increasing detection accuracy.

In 2016, Korytkowski et al. [48] have uncovered a fast computing framework with some methods to optimize the signal propagation speed and also compared their implementation with the original OverFeat implementation.

While deep learning based methods for generic object detection have improved rapidly, most approaches to face detection are still based on the R-CNN framework, leading to limited accuracy and processing speed. In 2017, Jiang et al. [49] have applied the Faster R-CNN, which has recently demonstrated impressive results on various object detection benchmarks, to face detection. By training a Faster R-CNN model on the large scale WIDER face dataset, the authors have reported that the state-of-the-art results on the WIDER test set as well as two other widely used face detection benchmarks, FDDB and the recently released IJB-A.

Sometimes the detection performance tends to degrade with increasing the intersection over union (IoU) thresholds. Two main factors are responsible for this: 1) overfitting during training, due to exponentially vanishing positive samples, and 2) inference-time mismatch between the IoUs for which the detector is optimal and those of the input hypotheses. A multi-stage object detection architecture, the Cascade R-CNN, has been proposed by Cai et al. in 2018 [50] to address these problems. A simple implementation of the Cascade R-CNN has shown to surpass all single-model object detectors on the challenging COCO dataset. The experimented research work has also revealed that the Cascade R-CNN is widely applicable across detector architectures, achieving consistent gains independently of the baseline detector strength.

Nowadays, many question answering systems adopt deep neural networks such as convolutional neural network (CNN) to generate the text features automatically, and obtained better performance than traditional methods. But the traditional CNN is unable to extract the variable length n-gram features and non-consecutive n-gram features. In 2019, Liu et al. [51] have established a multi-scale deformable convolutional neural network to capture the non-consecutive n-gram features by adding offset to the convolutional kernel, and also proposed to stack multiple deformable convolutional layers to mine multi-scale n-gram features by the means of generating longer n-gram in higher layer. Furthermore, they have applied the proposed model into the task of answer selection.

Currently 3D shape recognition becomes essential due to the popularity of 3D data resources. The new method, hybrid deep learning network convolution neural network and support vector machine (CNN–SVM), for 3D recognition has been introduced by Hoang et al. 2020 [52]. They have obtained and stored the 2D projection of this 3D augmentation data in a matrix form, the input data of CNN–SVM. The proposed method has worked with both the 3D model in the augmented/virtual reality system and in the 3D Point Clouds, an output of the LIDAR sensor in autonomously driving cars.

Pedestrian detection and tracking is a critical task in the area of smart building surveillance. Pedestrian detection in smart building is greatly challenged by the image noises by various external environmental parameters. The advancements in deep learning algorithms perform exponentially well in handling the huge volume of image data. In 2020, Kim et al. [53] have analyzed about a pedestrian detection model based on deep convolution neural network (CNN) for classification of pedestrians from the input images. They have proposed a optimized version of VGG-16 architecture is evaluated for pedestrian detection on the INRIA benchmarking dataset consisting of 227×227 pixel images.

A new multi-scale convolution model based on multiple attentions has been unveiled by Yang et al. in 2020 [54]. It has introduced the attention mechanism into the structure of a Res2-block to better guide feature expression. First, they have adopted a channel attention to score channels and sort them in descending order of the feature's importance (Channels-Sort). Then, they have implemented channel attention on the residual small blocks to constitute a dual attention and multi-scale block (DAMS-block). The experimental results have shown that the convolution model with an attention mechanism and multi-scale features is superior in image classification.

In 2020, Cao et al. [55] have compared compare and analyzed mainstream object detection algorithms and proposed a multi-scaled deformable convolutional object detection network to deal with the challenges faced by current methods. Their analysis demonstrated a strong performance on par, or even better, than state of the art methods. The authors have used deep convolutional networks to obtain multi-scaled features, and add deformable convolutional structures to overcome geometric transformations and then fused the multi-scaled features by up sampling, in order to implement the final object recognition and region regress.

4.5. You Only Look Once (YOLO)

Deep learning technology has been widely used in object detection. Although the deep learning technology greatly improves the accuracy of object detection, the lot of challenges are often occurred in a high computational time. You Only Look Once (YOLO) is a network for object detection in images. In YOLO algorithm, a matrix is created as in Eqn. (11). The anchor boxes are represented by p_c. The coordinates required for drawing the bounding boxes is represented by b_x , b_y, b_h and b_w; and c_1, c_2 and $c_3$represents the classes.

$$y = \begin{bmatrix} p_c \\ b_x \\ b_y \\ b_h \\ b_w \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} (11)$$

Usually more than one anchor box will present for the algorithm. In that case, the arguments in the target vector y are followed by the second anchor box followed by the boundary boxes and classes again. The anchor boxes are used to detect more than one image in the same grid cell. If the anchor box is zero, the specified boxes are given with dummy values. Initially, a training set is constructed. If two anchor boxes are used, then the output would be of the size 3×3. Then the target vector y would be $3 \times 3 \times 8$. We multiply by 8 as the target vector y is of the size 8. We go through 9^{th} target grid cells and form the target vector y. Also the class in which the detected object belongs, is denoted by 1 and the other classes are denoted by 0. For each of the three classes, non-max suppression is used to generate final predictions.

Among the many convolutional layers, the final layer predicts class probabilities and the bounding box coordinates. A linear activation function is used for the final layer and all other layers use the following leaky rectified linear activation. This leads to specialization between the bounding box predictors. Each predictor gets better at predicting certain sizes, aspect ratios, classes of object, or improving overall recall (Figure 3).

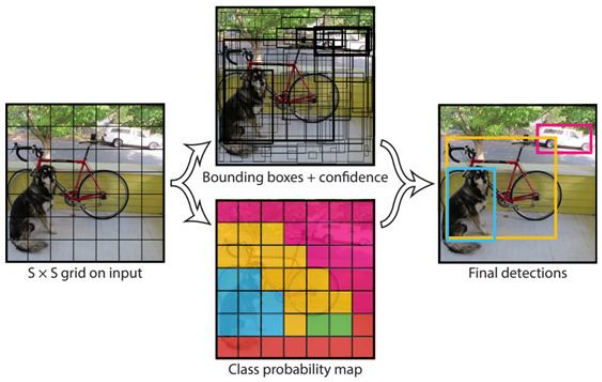

Figure 3: Object Detection by YOLO Algorithm on Sample Image

In 2016, Redmon et al. [56] have presented a new approach for object detection named YOLO. In this algorithm, object detection has been framed as a regression problem to spatially separated bounding boxes and associated class probabilities. The YOLO model has processed images in real-time at 45 frames per second. The authors have also discussed about another version, Fast YOLO, which processed 155 frames per second. They have also concluded that YOLO outperforms other detection methods such as DPM and R-CNN, when converting from natural images to other domains like artwork.

In 2017, Redmon et al. [57] have introduced YOLO9000, a real-time object detection algorithm that can detect over 9000 object categories. The authors have proposed a method to train YOLO9000 on the COCO detection dataset and the ImageNet classification dataset. They have concluded that the YOLO9000 predicts detections for more than 9000 different object categories, in real-time.

In 2018, Laroca et al. [58] have exhibited a sophisticated Automatic License Plate Recognition (ALPR) system based on the YOLO object detector. The authors have developed a two-stage approach employing simple data augmentation tricks such as inverted License Plates (LPs) and flipped characters. They have implemented ALPR approach on two data sets, first being SSIG dataset and second being UFPR-ALPR dataset; and they have also concluded that ALPR approach performs better for both the datasets.

In 2018, Xie et al. [59] have proposed CNN-based MD-YOLO framework for multi-directional car license plate detection. They have discussed that the sketched method can manage rotational problems using accurate rotation angle prediction and a fast intersection-over-union evaluation strategy. The authors have concluded that, the various experimental results shows that the proposed method outperforms over other existing methods in terms of better accuracy and lower computational cost.

In 2019, Nguyen et al. [60] have established a Tera-OPS streaming hardware accelerator implementing a YOLO-CNN. The parameters of YOLO-CNN are retrained and quantized with the PASCAL VOC data set using binary weight and flexible low-bit activation. In the proposed design, all convolutional layers are fully pipelined for enhanced hardware utilization. The authors have concluded that this design outperforms the one-size-fits-all designs in both performance and power efficiency.

A real-time object detection algorithm for videos based on the YOLO network has been developed by Lu et al. in 2019 [61]. They have eliminated the influence of the image background by image pre-processing, and then they have trained the Fast YOLO model for object detection to obtain the object information. Based on the Google Inception Net (GoogLeNet) architecture, the improvised YOLO network has been presented by using a small convolution operation to replace the original convolution operation, which can reduce the number of parameters and greatly shorten the time for object detection.

A modified YOLOv1 based neural network is proposed for object detection by Ahmad et al. in 2020 [62]. The new neural network model has been improved by using YOLOv1 network, adding a spatial pyramid pooling layer; and also including an inception model with a convolution kernel, which reduced the number of weight parameters of the layers.

In 2020, Alsanad et al. [63] has flowed with a new approach by training and improving a convolutional neural network (CNN) based on You Only Look Once version 2 (YOLOv2) to efficiently detect the fuel trucks from images in

embedded systems. The proposed method has considered the entire image area for strong object detection compared with existing methods that only focus on the image area where the class object exists to predict its probability to be in a class. The authors have recommended that the proposed method is suitable to monitor long country borders using unmanned drones.

5. CONCLUSION AND FUTURE DIRECTIONS

In this study, we have remarkably discussed about fractal and multifractal techniques for image classification and also elaborately deliberated various algorithms of object detection such as SVM, HOG, DPM, CNN, R-CNN, Faster R-CNN, Multi-Scale Deformable R-CNN, and YOLO systems. Object detection algorithms have wide applications in many fields such as optimal character recognition, tracking objects, face detection, face recognition, object extraction from an image, object extraction from a video, medical imaging and also in sports. Object detection algorithms and fractal theory can be implemented in airports using thermal images in various phases such as screening, crowd control, aviation and aircraft maintenance. In the future, we can focus on how effectively we can hybrid the fractal features and the object detection algorithms in various applications such as airports to avoid aircraft accidents, bio-medical image analysis to segment the abnormal portions accurately; and also agricultural fields where we can detect the animals which intrudes the growth of the crops.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. X. Feng, Y. Jiang, X. Yang, M. Du, and X. Li**, Computer vision algorithms and hardware implementations: A survey**, *Integration, the VLSI Journal*, vol. 69, pp. 309–320, 2019. https://doi.org/10.1016/j.vlsi.2019.07.005
2. J.D. Pujari, D.K.Bhadangkar, and R. Yakkundimath, **Identification and Recognition of Facial Expressions Using Image Processing Techniques: A Survey**, *International Journal of Emerging Trends in Engineering Research*, vol. 5, no. 5, pp. 1–10, 2017.
3. N. Nawaz, **Artificial Intelligence Face Recognition for applicant tracking system**, *International Journal of Emerging Trends in Engineering Research*, vol. 7, no. 12, pp. 895–901, 2019.
4. A.S. Alon, E.D. Festijo, and C.D. Casuat, **Tree Extraction of Airborne LiDAR Data Based on Coordinates of Deep Learning Object Detection from Orthophoto over Complex Mangrove Forest**, *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. 2107–2111, 2020. https://doi.org/10.30534/ijeter/2020/103852020
5. M.I. Chacon-Murguia, A. Guzman-Pando, G. Ramirez-Alonso, and J.A. Ramirez-Quintana, **A novel instrument to compare dynamic *object detection algorithms***, *Image* and Vision Computing, vol. 88, pp. 19–28, 2019.
6. Z. Zhao, P. Zheng, S. Xu, and X. Wu, **Object Detection With Deep Learning: A Review**, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.
7. X. Wu, D. Sahoo, and S.C.H. Hoi, **Recent advances in deep learning for object detection**, *Neurcomputing*, vol. 396, pp. 39–64, 2020.
8. K. Tong, Y. Wu, and F. Zhou, **Recent advances in small object detection based on deep learning: A review**, *Image and Vision Computing*, vol. 97, Article No.: 103910, 2020.
9. Y. Xiao, Z. Tian, J. Yu, Y. Zhang, S. Liu, S. Du, and X. Lan, **A review of object detection based on deep learning**, *Multimedia Tools and Applications*, 2020.
10. D. Easwaramoorthy, and R. Uthayakumar, **Estimating the Complexity of Biomedical Signals by Multifractal Analysis**, *Proceedings of the IEEE Students' Technology Symposium*, pp. 6–11, 2010.
11. D. Easwaramoorthy, and R. Uthayakumar, **Analysis of EEG Signals using Advanced Generalized Fractal Dimensions**, *Proceedings of the International Conference on Computing*, Communication and Networking Technologies, pp. 1–6, 2010.
12. D. Easwaramoorthy, and R. Uthayakumar, **Analysis of Biomedical EEG Signals using Wavelet Transforms and Multifractal Analysis**, *Proceedings of the IEEE International Conference on Communication Control and Computing Technologies*, pp. 544–549, 2010. https://doi.org/10.1109/ICCCCT.2010.5670780
13. D. Easwaramoorthy, and R. Uthayakumar, **Improved Generalized Fractal Dimensions in the Discrimination between Healthy and Epileptic EEG Signals**, *Journal of Computational Science*, vol. 2, no. 1, pp. 31–38, 2011.
14. R. Uthayakumar, and D. Easwaramoorthy, **Multifractal-Wavelet Based Denoising in the Classification of Healthy and Epileptic EEG Signals**, *Fluctuation and Noise Letters*, vol. 11, no. 4, Article No.: 1250034, 2012.
15. R. Uthayakumar, and D. Easwaramoorthy, **Epileptic Seizure Detection in EEG Signals using Multifractal Analysis and Wavelet Transform**, *Fractals*, vol. 21, no. 2, Article No.: 1350011, 2013.
16. R. Uthayakumar, and D. Easwaramoorthy, **Fuzzy Generalized Fractal Dimensions for Chaotic Waveforms**, *Chaos, Complexity and Leadership 2012*, *Springer Proceedings in Complexity*, pp. 411–422, 2014.
17. D. Easwaramoorthy, P.S. EliahimJeevaraj, A. Gowrisankar, A. Manimaran, and S. Nandhini, **Fuzzy Generalized Fractal Dimensions Using Inter-Heartbeat Interval Dynamics in ECG Signals for Age Related Discrimination**, *International Journal of Engineering and Technology (UAE)*, vol. 7, no. 4.10, pp. 900–903, 2018.
18. R. Uthayakumar, and D. Easwaramoorthy**, Generalized Fractal Dimensions in the Recognition of Noise Free**

Images, *Proceedings of the International Conference on Computing, Communication and Networking Technologies*, pp. 1–5, 2012.

19. R. Uthayakumar, and D. Easwaramoorthy, **Multifractal Analysis in Denoising of Color Images**, *Proceedings of the International Conference on Emerging Trends in Science, Engineering and Technology*, pp. 228–234, 2012.
20. P.S. EliahimJeevaraj, P. Shanmugavadivu, and D. Easwaramoorthy, **Fuzzy Cut Set-Based Filter for Fixed-Value Impulse Noise Reduction**, *Advances in Algebra and Analysis, Trends in Mathematics*, vol. 1, pp. 205–213, 2018. https://doi.org/10.1007/978-3-030-01120-8_24
21. J. Liu, and H. Wei, **Optimal selection of fractal features for man-made object detection from infrared images**, *2nd International Asia Conference on Informatics in Control, Automation and Robotics*, pp. 177–180, 2010.
22. D.J. Holliday, and A. Samal, **Object recognition using L-system fractals**, *Pattern Recognition Letters*, vol. 16, no. 1, pp. 33–42, 1995.
23. X. Wen, D. Hu, X. Dong, F. Yu, D. Tan, Z. Li, Y. Liang, D. Xiang, S. Shen, C. Hu, and B. Cao, **An object-oriented daytime land fog detection approach based on NDFI and fractal dimension using EOS/MODIS data**, *International Journal of Remote Sensing*, vol. 35, no. 13, pp. 4865–4880, 2014.
24. Y. Zhou and J. Liang, **Fractal features for object recognition,** *12th International Conference on Signal Processing***,** Article No.: 14868730, 2014.
25. P. Shivakumaraa, W. Liang, L. Tong, C.T. Lim, M. Blumenstein, and B.S. Anami, **Fractals based multi-oriented text detection system for recognition in mobile video images**, *Pattern Recognition*, vol. 68, pp. 158–174, 2017.
26. M. Shiran, M.A.Z. Asadi, P. Mozzi, H. Adab, and A. Amirahmadi, **Detection of surface anomalies through fractal analysis and their relation to morphotectonics (High Zagros belt, Iran)**, *Geosciences Journal*, 2020.
27. T. Evgeniou, and M. Pontil, **Support Vector Machines: Theory and Applications**, In: Paliouras G., Karkaletsis V., Spyropoulos C.D., (Eds.), *Machine Learning and Its Applications, ACAI-1999, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg*, vol. 2049, pp. 249–257, 2001. https://doi.org/10.1007/3-540-44673-7_12
28. E. Osuna, R. Freund, and F. Girosit, **Training support vector machines: an application to face detection**, *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 130–136, 1997.
29. C. Hsu, and C. Lin, **A comparison of methods for multiclass support vector machines**, *IEEE Transactions on Neural Networks*,vol. 13, No. 2, pp. 415–425, 2002.
30. F. Melgani, and L. Bruzzone, **Classification of hyperspectral remote sensing images with support vector machines**, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778 –1790, 2004.
31. Y.M. Riyazuddin, S. MahaboobBasha, J. Krishna Reddy, and S. NaseeraBanu, **Effective Usage of Support Vector Machine in Face Detectio**n, *International Journal of Engineering and Advanced Technology*, vol. 9, no. 3, pp. 1336–1340, 2020.
32. N. Dalal, and B. Triggs, **Histograms of oriented gradients for human detection**, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 886–893, 2005.
33. Q. Zhu, M. Yeh, K. Cheng, and S. Avidan, **Fast Human Detection Using a Cascade of Histograms of Oriented Gradients**, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1491–1498, 2006.
34. R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, **Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions**, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1932–1939, 2009.
35. Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng, **Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis**, *Conference on Computer Vision and Pattern Recognition*, pp. 3361–3368, 2011.
36. O. Oreifej, and Z. Liu, **HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences**, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716–723, 2013.
37. T. Surasak, I. Takahiro, C. Cheng, C. Wang, and P. Sheng, **Histogram of oriented gradients for human detection in video**, *5th International Conference on Business and Industrial Research, Bangkok*, pp. 172–176, 2018.
38. Reinaldo, N. Manurung, J.I. Simbolon, and Christnatalis, **Traffic sign detection using histogram of oriented gradients and max margin object detection**, *Journal of Physics: Conference Series, IOP Publishers*, vol. 1230, Article ID: 012098, 2019.
39. D. Terzopoulos, and D. Metaxas, **Dynamic 3D models with local and global deformations: deformable superquadrics**, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 7, pp.703–714, 1991. https://doi.org/10.1109/34.85659
40. A. Escalera, J.M. Armingol, J.M. Pastor, and F.J. Rodriguez, **Visual sign information extraction and identification by deformable models for intelligent vehicles**, *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 2, pp. 57–68, 2004.
41. P.F. Felzenszwalb, D. McAllester, and D. Ramanan, **A discriminatively trained, multiscale, deformable part model**, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.
42. P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, **Object Detection with Discriminatively Trained Part-Based Models**, *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627–1645, 2010.

43. P.F. Felzenszwalb, R.B. Girshick, and D. McAllester, **Cascade object detection with deformable part models**, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2241–2248, 2010.
44. M. Pandey, and S. Lazebnik, **Scene recognition and weakly supervised object localization with deformable part-based models**, *International Conference on Computer Vision*, pp. 1307–1314, 2011.
45. N. Pool, and R.R. Vatsavai, **Deformable Part Models for Complex Object Detection in Remote Sensing Imagery**, *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pp. 57–62, 2018.
46. A.S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, **CNN Features Off-the-Shelf: An Astounding Baseline for Recognition**, *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 512–519, 2014.
47. R. Girshick, **Fast R-CNN**, *IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.
48. M. Korytkowski, P. Staszewski, P. Woldan, and R. Scherer, **Fast Computing Framework for Convolutional Neural Networks**, *IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*, pp. 118–123, 2016.
49. H. Jiang, and E. Learned-Miller, **Face Detection with the Faster R-CNN**, *12th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 50–657, 2017.
50. Z. Cai, and N. Vasconcelos, **Cascade R-CNN: Delving Into High Quality Object Detection**, *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, 2018.
51. D. Liu, Z. Niu, C. Zhang, and J. Zhang, **Multi-Scale Deformable CNN for Answer Selection**, *IEEE Access*, vol. 7, Article No.: 19144224, 2019.
52. L. Hoang, S. Lee, and K. Kwon, **A 3D Shape Recognition Method Using Hybrid Deep Learning Network CNN–SVM**, *Electronics*, vol. 9, Article ID: 649, 2020. https://doi.org/10.3390/electronics9040649
53. B. Kim, N. Yuvaraj, K.R. Sri Preethaa, R. Santhosh, and A. Sabari, **Enhanced pedestrian detection using optimized deep convolution neural network for smart building surveillance**, *Soft Computing*, 2020.
54. Y. Yang, C. Xu, F. Dong, and X. Wang, **A New Multi-Scale Convolutional Model Based on Multiple Attention for Image Classification**, *Applied Sciences*, vol. 10, no. 1, Article ID: 101, 2020.
55. D. Cao, Z. Chen, and L. Gao, **An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks**, *Human-centric Computing and Information Sciences*, vol. 10, Article No.: 14, 2020.
56. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, **You Only Look Once: Unified, Real-Time Object Detection**, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.
57. J. Redmon, and A. Farhadi, **YOLO9000: Better, Faster, Stronger**, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6517–6525, 2017.
58. R. Laroca, E. Severo, L.A. Zanlorensi, L.S. Oliveira, G.R. Gonçalves, W.R. Schwartz, and D. Menotti, **A Robust Real-Time Automatic License Plate Recognition Based on the YOLO Detector**, *International Joint Conference on Neural Networks*, pp. 1–10, 2018.
59. L. Xie, T. Ahmad, L. Jin, Y. Liu, and S. Zhang, **A New CNN-Based Method for Multi-Directional Car License Plate Detection**, *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, pp. 507–517, 2018. https://doi.org/10.1109/TITS.2017.2784093
60. D.T. Nguyen, T.N. Nguyen, H. Kim, and H. Lee, **A High-Throughput and Power-Efficient FPGA Implementation of YOLO CNN for Object Detection**, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 8, pp. 1861–1873, 2019.
61. S. Lu, B. Wang, H. Wang, L. Chen, M. Linjian, and X. Zhang, **A real-time object detection algorithm for video**, *Computers and Electrical Engineering*, vol. 77, pp. 398–408, 2019.
62. T. Ahmad, Y. Ma, M. Yahya, B. Ahmad, S. Nazir, and A. Haq, **Object Detection through Modified YOLO Neural Network**, *Scientific Programming*, Article ID: 8403262, 2020.
63. H.R. Alsanad, O.N. Ucan, M. Ilyas, A.U.R. Khan, and O. Bayat, **Real-Time Fuel Truck Detection Algorithm Based on Deep Convolutional Neural Network**,*IEEE Access*,vol. 8, pp. 118808–118817, 2020. https://doi.org/10.1109/ACCESS.2020.3005391