IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 5, Ser. I (Sep. –Oct. 2022), PP 38-47
www.iosrjournals.org

Text Detection and Object Recognition from Scene Images Using CNN and YOLOv3

KaushikDas[1], Arun KumarBaruah[2]
[1]*(Computer Science and Engineering, Dibrugarh University, India)*
[2]*(Department of Mathematics, Dibrugarh University, India)*

Abstract
Object detection and text recognition, which is otherwise called Optical Character Recognition (OCR), is an emerges as an active area of research because of the quick development with many existing applications. With the fast improvement in the Deep Learning (DL),various powerful tools which can able to learn semantic, high-level, deeper features to tackle the problems in the traditional methods. However, these methods are generally deterministic and gives deterministic output. In this paper, a new DL based object detection and text detection methods was introduced with a novel hybrid activation function. The proposed detection model detects the text and object with high precision rate.

Keywords: *Object detection, text detection, real-time images, Deep Learning (DL), hybrid activation function.*

Date of Submission: 12-10-2022 Date of Acceptance: 27-10-2022

I. Introduction

A recent advancement made in the computer vision technology such as licence plate recognition have made day to day life more convenient. With the prominence of cell phones like smartphones, anybody can now effectively acquire image and videos using their mobile phones and share them on the internet. Among many objects presents in these image and videos, the textual information plays a significant role. The textural information includes rich, precise and high-level data that gives meaning to the objects in the natural scenes which helps the people to get better access and grasp the data within the images and videos (Mahajan & Rani, 2021). Therefore, obtaining textural content from the images has turned into an essential task for Machine Learning (ML).The import of semantic or high-level text data present in the image is that it can undoubtedly describe an image with good clarity and can be extracted utilizing low-level features like colour, texture, etc. which in turn varies with language, font, style and background, thus making the task of text extraction challenging one (Zhu *et al.* 2016).The difficulties involved in the text detection attract numerous researchers to contribute in the area of text detection. Over the period, remarkable achievement is accomplished in the text detection method by numerous researchers.The paradigm of the text detection method is rapidly shifting from the usage of fundamental features to modern and more intelligent algorithm. However, the test detection and recognition is certainly not an easy task, where a ML technique cannot accomplish high accuracy (Busta *et al.* 2015; Neumann& Matas, 2015).Because of the tremendous development and success of the DL, a lot of DL models have been developed for the text detection and recognition with increased accuracy and efficiency.

Object detection is the technique utilized to detect to all instances of the objects like people, car in an image. The object recognition system comprises of two parts such as detection and recognition(Yeo *et al*. 1995). The detection deals with distinguishing the object form the background and recognition deals with the classification of the object into one of the predefined categories. An object recognition algorithm depends on matching, learning or pattern recognition algorithms utilizing the feature based techniques (Belongie *et al.* 2002). In past few years, the object detection in real time and image processing has become an active area of research and many new approaches have been introduced. A lot of research on object detection have been conducted in past days.The traditional ML algorithms cannot handle the grasp the complexity of the object detection problem statement because of the subject matters and complexity (Sharma *et al.* 2013). Generally, ML algorithms depend on the hand-crafted features by experts or practitioners, since they have hand-on experience in relevant subject matters. For that reason, conventional learning techniques were not dependable. Since, the machines cannot detect the objects in an image instantly as like humans, it is truly essential for the algorithm to be fast and accurate to detect the object in real-time (Redmon *et al.* 2016). Besides, these different ML algorithms of Logistics Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbour (KNN), Mean Shift Algorithm (MSA), Decision Tree (DT) are considering raw image data without any learning of hidden delegations (Dutta, 2020). Furthermore, the pre-processing and reshaping is also based on the knowledge of experts which eventually consumes a lot of time and labour-intensive. To solve these limitations, the DL has shown promising potential (Esteva *et al*. 2019; Shanahan & Dai, 2020). DL has shown a breakthrough in capturing the hidden pattern and extract features in the most dependable manner. It has the benefit of automatically learning the most significant features from the image data rather than features extracted from them like in ML. Besides, DL algorithm such as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Long-Short-Term Memory (LSTM) networks, Fully-connected Feed forward Deep Neural Network (FNN), Regional Convolutional Neural Networks (R-CNN), You Only Look Once (YOLO), etc. which does not need manual pre-processing or hand crafted feature extraction on the raw data.Zhang*et al.* (2020) introduced an object and text detection method named as “DetReco” for the detection of objects and texts. The YOLOv3 algorithm is utilized for the detection of text and CRNN is used to recognition of text. It was concluded that the proposed method can detect and recognize the texts with robustness. The experimental results show that proposed method accomplishes the 78.3 mAP for detection of objects and 72.8 Map for the detection of texts. Baimukashev*et al*. (2019) introduced a DL-based object detection framework utilizing the synthetic depth dataset images of 22 objects randomly placed in 0.5m x 0.5 m x0.1 m box. The R-CNN was employed for the training of the dataset and detection accuracy of 40.96% and 93.5% was accomplished for real depth and synthetic depth images.Suho*et al*. (2018) employed R-CNN and improved RPN network to detect the various types of vehicle which are common in traffic scene. Masita *et al.* (2018) introduced a DL technique of R-CNN for the detection of pedestrian using the two various pedestrian detection dataset. Nath &Behzadan (2020) presented a novel DL based approach for the detection of lane in the road. The vehicle detection algorithm utilizing the YOLO is implemented to avoid the accident for autonomous vehicle systems. It was concluded that the proposed approach accomplished promising outcomes for both urban and rural roads. Ju *et al.* (2020) developed a DL based object detection model for the detection of old loess landslides using the Google Earth images. The RetinaNet, YOLOv3 and Mask R-CNN algorithm was used for the automatic detection landslides. The Mask R-CNN accomplished the high accuracy with AP of 18.9%, F1 score of 55.31%. Jiang *et al*. (2021) introduced DL based method for the damage detection and classification of concrete using the dataset comprises of 5000 images. The object detection algorithm was optimized by depth wise separable convolution, inverse residual network and linear bottleneck structure. It was observed that the inference speed of the proposed detection method was 24.1%-53.5% higher than the original network.Kohli*et al.* (2020) introduced a model named J&M for the detection of text form the handwritten images. The experimentation of proposed model with MNIST database in python and achieved the training accuracy of 99.5%, testing accuracy of 99% and training loss of 1.5%.

In spite of the fact that there have been several attempts with the development of DL based object and text detection model, it remains an open subject of exploration for the researchers. One of the significant drawbacks as observed is the time taken to arrive an optimal solution; these necessitates a further investigation. The primary objective of this study is to develop a framework based on existing algorithms for the detection of objects and extraction of text from images using the YOLO V3 method using novel hybrid activation function.

II. Research Methodology

2.1 Preliminaries

2.1.1 CNN

CNN has become very popular DL model in the field of image processing, because of its high performance in detection of image patterns. This has opened up various application opportunities in our day to day activities such as object recognition, image classification, traffic monitoring, facial recognition, etc. CNN are sparse, feed-forward neural network, which is formed of artificial neurons and have a self-optimization and learning property as like as human brain (Kavitha *et al.* 2022). Because of this self-optimizing property, it can extract and more precisely classify the features extracted from the images. In addition, it requires very less pre-processing of the input data to yield high accurate and precise outcomes. CNNs are tremendously utilized in object detection and image classification. In image classification, every pixel is considered as a feature for the neural network. CNN attempts to understand and differentiate among the images relying upon these features. Conventionally, first few convolutional layers can capture very low-level features like edges, gradient orientation, colour, etc. However, with increased number of convolution layers, it starts extracting high-level features. The fig 1 presents the CNN architecture which contains convolutional layer, pooling layer and fully connected layer. The convolutional and pooling layers are generally altered and the depth of every filter increases from left to right, while the output size are reducing. The fully connected layer is the last stage which is like the last layer of the CNN.

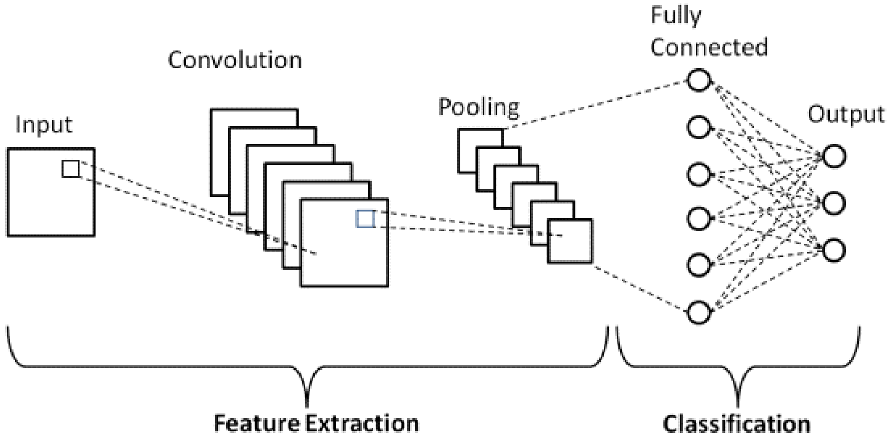


Fig. 1 CNN architecture

2.1.2 YOLO V3

YOLO V3 algorithm is a real-time detection algorithm proposed by Joseph Redmon & Ali Farhadi in 2018 (Zhang, 2021), it was based on regression technique. It is a CNN, which can predict the position and category of the multiple target frames simultaneously. It is an improvement of YOLO V1. It utilizes the residual neural network as the basic network of the feature extraction, on this premise, a convolution layer is added for the prediction of images of three various scales to get higher semantic data. Furthermore, taking into account the class labels, the YOLO V3 utilizes the logistics rather than the softmax classifier. It uses the FPN network to detect the targets of various sizes on multiple scales. The target data can be detected, when the cell are finer. The size of the feature map of each prediction task is as follows in equ: 1;

$$N\ X\ N\ X\ [3 * (4 + 1 + class\_num)] \qquad (1)$$

where N denotes the target size; 3 denotes the number of bounding boxes acquired from the every target; 4 denotes the number of bounding boxes coordinates; 1 denotes the predicted value of the target and $class\_num$ denotes the number of categories.

2.1.3 Swish Activation function

The swish is a new type of activation function introduced by Google in 2017, which has created a huge sensation in its presence (Ramachandran *et al.* 2017). The swish function expressions are not obtained through the theoretical resonating, yet through experimental application of small-scale exhaustive search and large-scale RNN controller application. An enormous number of investigations demonstrate that its impact is much better than the Relu function. Its expression is shown in equ: 2;

$$Swish\ (x) = \frac{x}{1+e^{\beta x}} \qquad (2)$$

Numerous experiments have affirmed that when the value of is 1, the gradient is consistent with Relu. This is most appropriate for the reinforcement learning training. The swish derivative expression is shown in equ: 3;

$$Swish'(x) = \frac{1-swish\ (x)}{1+e^{-x}} + swish\ (x) \qquad (3)$$

2.1.4 Leaky ReLU

Leaky ReLUis one of the variants of ReLU by assigning a non-zero output for the negative input (Xu *et al.* 2015), that is $f(x) = \max(\alpha x, x)$, where α indicates a predefined parameters in the range of $(0, 1)$. It is an improved version of ReLU function, where for negative values of x, rather than characterizing the value of ReLU functions as zero,. The ReLU maps the negative input to zero, whereas the LeakyReLU utilizes a predefined linear function to compress negative input. The compression of LeakyReLU empowers the negative part of the feature data retained. Thus, the LeakyReLU compromises the sparsity of the network and its input data. Mathematically, it's expression was shown in equ: 4 and 5;

$$f(x) = 0.01x, x < 0 \qquad (4)$$

$$f(x) = x, x \geq 0 \qquad (5)$$

2.2 Object detection and text detection model

The fig. 2 presents the Architecture of object detection and text detection model. The proposed network architecture composed of the two parts such as object detection and text detection. The CNN is used to for the text detection and YOLO V3 is used for the object detection. The YOLO V3 which uses a fully CNN to detect the objects in the image. The input image data was pre-processed using the smooth filtering. The smooth filtering is used for blurring reduction of noise present in the input image. The blurring is the pre-processing steps for the removal of small details and noise reduction is achieved by blurring. Thus, the filtering process is performed to improve the quality of images. The input layer of the CNN network feeds the input images to the convolutional layers, where the features are convolved.The convolution network is utilized to extract the features in multi scales feature maps form the input image. The classification and bounding box regression networks directly outputs the objectness score, object classes and the coordinate offsets of the object at multiple feature maps. The text detection model was trained with the SCUT FORU dataset to prepare the model. The text images in the SCUT FORU dataset are cropped form the original images corresponding to the coordinates in the annotations. The text detection is done by executing the trained CNN model to predict the text and it matrix location. After the features extracted from the CNN layer, the feature map sizes of 13x13, 26x26, 52x52, 104x104 are acquired and the bottom feature layer is up-sampled form the bottom to the neighbouring feature are used in the YOLO V3 for the object detection.

2.3 Hybrid Activation Function

The activation unction is also known as non-linear mapping and serve as decision function. It is used to increase the expressive ability of the network and helps in learning the intricate pattern. While utilizing a Neural Network, it is essential to choose which activation function to be used on the hidden layer and output node. The selection of suitable activation function can accelerate the learning process and improve the performance of network for a specific task. The most commonly utilized activation functions are Sigmoid, Tanh, Swish,ReLU, Leaky ReLU, etc. (Gu *et al*. 2018; Nwankpa *et al.* 2018). The majority of these activation functions has some drawbacks. The drawbacks of the swish activation are more slow to compute as compared to ReLU. ReLU is one of the most commonly utilized activation function in the DL algorithm, because it is computationally effective (Haque *et al.* 2020). The drawbacks of theReLU activation function is that if the input is less than 0, then its output is also zero, thus the network cannot continue the backpropagation. The drawbacks of the Leaky ReLU activation function is that α- value is always constant and hyper parameter. Hence, in this proposed network, the hybrid activation function is used by combining swish and Leaky ReLU activation function. The main advantages of hybrid activation function less time consumptions and more accuracy.

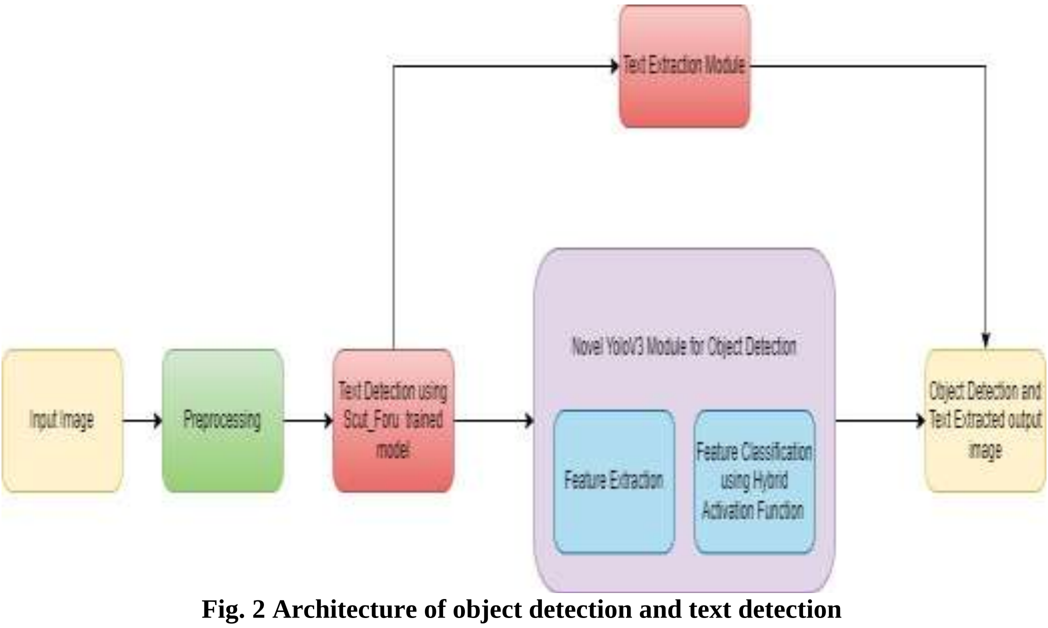

Text Extraction Module
Input Image
Preprocessing
Text Detection using Scut_Foru trained model
Novel YoloV3 Module for Object Detection
Feature Extraction
Feature Classification using Hybrid Activation Function
Object Detection and Text Extracted output image

Fig. 2 Architecture of object detection and text detection

Text Extraction Algorithm

Step 1: Load the input image data
Step 2: Install the required libraries like cv2, matplotlib etc.
Step 3: Pre-process the data using SMOOTH filter
Step 4: Text Detection by CNN using SCUT FORM dataset
Text detection is done by executing the trained CNN model to predict the text and it matrix location
Step 5: Drawing the Bounding boxes according to detected and extracted Text images.
Step 6: Feature extracted in Convolution conv2D layer used in Yolov3
Step 7: Detecting or Classifying the Object using YOLO V3 method.
In this, a novel Hybrid Activation function added for YoloV3 classification.

III. Experimentation and Results

3.1 Dataset

The proposed method was evaluated on the Coco Dataset, VOC2007 + VOC2012 dataset and SCUT FORU Database. The detail description about the dataset are explained in the section: 3.1.1-3.1.3.

3.1.1 COCO Dataset

The Microsoft COCO data set was one of the popular benchmark data used for the task of the image detection and segmentation. This dataset comprises of natural images of complex scenes that includes multiple objects. It has 91 type so objects with more than million labelled instances in 328K images. COCO has a fewer number of categories comparted with the popular ImageNet dataset yet has far more pictures in each one the categories. The COCO dataset addresses the issue of past dataset by giving non-iconic views, precise 2D localization of objects and multiple objects per image (Lin *et al.* 2014).

3.1.2 VOC2007 + VOC2012

The VOC2007 dataset is the challenge to detect objects from a various visual objects classes in a realistic scene. This database comprises of 9963 annotated images. The VOC2012 dataset is the same challenge as VOC2007 which increases the size of the training set.

3.1.3 SCUT_FORU Text

The SCUT_FORU database was introduced by the South China University of Technology. This dataset contains Chinese 2k and English 2k. The English 2k is only used for the performance evaluation. The English 2k dataset includes character annotations and word annotations. The character of the dataset contains 52 upper lower case letters and 10 Arabic numerals. The label format of the dataset is $\{x, y, w, h, label\}$. The $\{x, y\}$ are the top-left coordinates of the rectangular box. The $\{w, h\}$ are width and height of the rectangular box. The $\{label\}$ is the word label of the text region. There are a total of 1715 images of which 1200 images are training and 515 images are testing images. The dataset has an average of 18.4 characters and 3.2 words per image.

3.2 Pre-processing

The proposed framework employs smooth filter to remove the noise pixels of the input image. This smooth filtering applies blurring effect to blur out the noisy pixels of small details in the image. The pre-processing stage greatly helps in the reduction of detection error since it aims to eliminate the pixels that might cause misconceptions.

3.3 Training and testing of network

The experiments were performed on the datasets such as SCUT-FORU, VOC2007+VOC2012. The SCUT-FORU is the large-scale detection dataset comprises of 4405 images. The VOC2007 datasets comprises of 9963 annotated images. Rom that, around 5011 images are used as training dataset and 4952 images are used as testing dataset. The VOC2012 datasets comprises of 17125 images training dataset. The COCO dataset is a large-scale detection dataset comprises of 330K images and 200K labels. The dataset is integrated into a comprehensive dataset of 29265 images with 23565 training and 5700 testing images. The network model training runs for 2000000 epochs. The initial learning arte is 0.01 with exponential decay of 0.1 for each 500000 epochs. The experiments use the gradient descent with momentum to train the network (Ruder, 2016).

3.4 Performance Evaluation

The performance of the proposed pre-processing stage is evaluated in terms of Mean Sqaure Error (MSE), Peak Signal to Noise Ratio (PSNR), and Signal to Noise Ratio (SNR).

- MSE can be defined as the measure of the average squares of error values. The MSE value obtained for the filtered image is 18.088. MSE can be calculated using eq. (6).

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \check{y}_j\right)^2 \qquad (6)$$

Where n is the number of input pixels, y_j are the pixels of original image, and $\check{y}_j$ are the pixels of filtered image.

- PSNR can be defined as the ratio between the maximum possible power of an image to the power of the corrupting number of noise pixels. The PSNR value obtained for the filtered image is 35.556 dB. PSNR can be calculated using eq. (7).

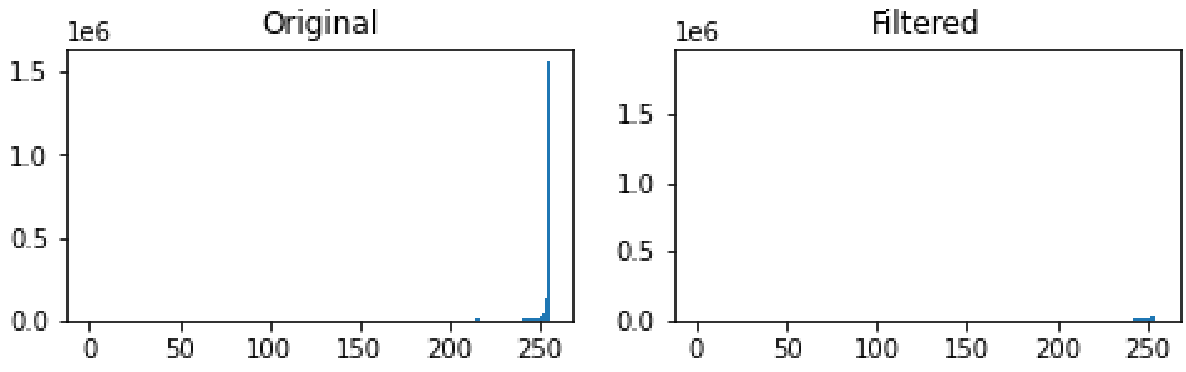
$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right) \qquad (7)$$

- SNR can be defined as the quality of the image with respect to the total measured pixel pixels. The SNR value obtained for the filtered image is 2.73 dB. SNR can be calculated using eq. (8).

$$SNR = \frac{mean\,(I_f)}{std\,(I_f)} \qquad (8)$$

Where I_f represents filtered image.

The histograms of the input image and the filtered image are shown in figure 3. It can be seen that the peak is found in around 250 in the original image (figure 3 (a)). This uneven distribution is equalized and improved in the filtered image (figure 3 (b)).



(a) **(b)**

Fig 3. Histogram patterns. (a) Original Image, (b) Filtered Image

The comparison on the values of MSE, PSNR, and SNR for original and filtered images are shown in table 1.

Table. 1 Performance of Pre-processor

Parameter	Original Image	Filtered Image
MSE	46.92	18.088
PSNR (dB)	20.69	35.5568
SNR (dB)	2.43	2.738

The proposed Object detection and text extraction from the image model was simulated in the python to assess the performance and it was compared with th existing methods under the following performance metrics.

- The mean Average Precision (mAP) was used as the performance metrics to evaluate the performance of the proposed detection method. The mAP is the widely used performance metric for the object detection. It is the mean of average precision calculated over all classes. The expression for mAP is given in equ: 9.

$$mAP = \frac{1}{N}\sum_{I=1}^{N} AP_i \quad (9)$$

where AP_i denotes the average precision and N denotes the total number of classes.

- Precision is defined as the percentage of a predicted region that belongs to the ground truth. The expression for Precision is given in equ: 10

$$Precision = \frac{TP}{TP+FP} \qquad (10)$$

where TP is an instance for which both actual and predicted values are positive; TNis an instance for which both actual and predicted values are negative, TP s an instance for which actual value is negative and predicted value is positive and FNis an instance for which actual value is positive and predicted value is negative.

- Average Precision (AP) is calculates the average value of precision over various level of recall. If the values of AP were high, it denotes that the performance was better. The expression for AP are given in equ: 11.

$$AP = \sum_n (R_n - R_{n-1}) P_n (11)$$

where P_n and R_n are the precision and recall at n^{th} threshold.

3.5 Performance of the proposed method

The detection precision (mAP) performance of the proposed detection method tested under COCO dataset was tabulated in the table.2. The comparison of the detection precision (mAP) performance of the various detection method are graphically presented in Fig. 4.

Table. 2mAP values trained on COCO Dataset

Method	References	mAP (%)
PG-PS-FR-CNN	Cheng *et al.* (2020)	20.7
DETR	Carion *et al.* (2020)	44.9

POTO-ResNext-101-DCN	Wang *et al.* (2021)	47.6
CBNET	Liu *et al.* (2020)	53.3
YoloV3	Zhao & Li, (2020)	53.2
Weighted Boxes Function	Solovyev *et al.* (2021)	56.4
Proposed YOLO V3	-	58.7

The mAP value of the proposed detection method tested under COCO dataset was 58.7%. The mAP value of the PG-PS-FR-CNN detection method tested under COCO dataset was 20.7%. The mAP value of the DETR detection method tested under COCO dataset was 44.9%. The mAP value of the POTO-ResNext-101-DCN detection method tested under COCO dataset was 47.6%. The mAP value of the CBNET detection method tested under COCO dataset was 53.3%. The mAP value of the YoloV3 detection method tested under COCO dataset was 53.2%. The mAP value of the Weighted Boxes Function detection method tested under COCO dataset was 56.4%. It was found that the mAP value of the proposed detection method was 4.07% -183% higher than the existing detection methods.

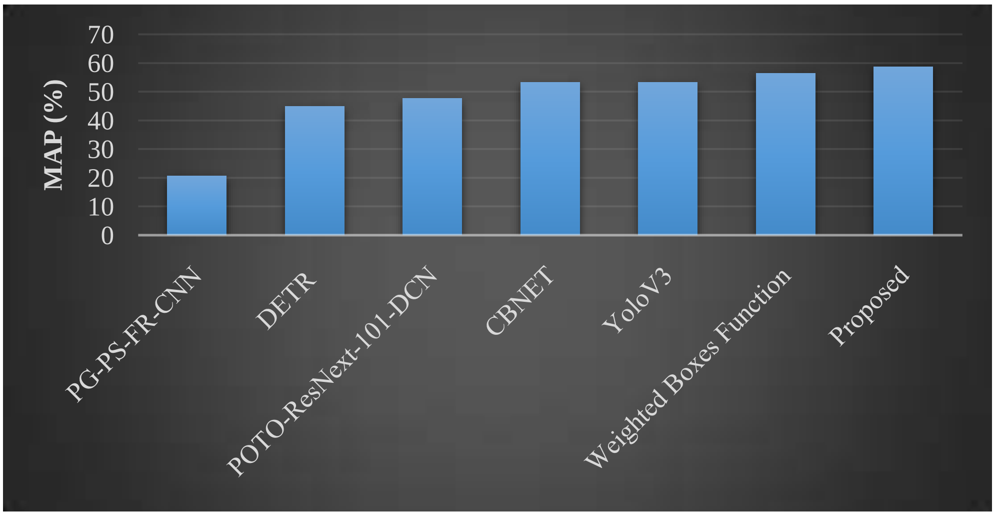


Fig. 4mAP values of proposed and existing models trained on COCO Dataset

The detection precision (mAP) performance of the proposed detection method tested under VOC2007 + VOC2012 dataset was tabulated in the table. 3. The comparison of the detection precision (mAP) performance of the various detection method are graphically presented in Fig. 5.

Table. 3mAP values trained on VOC2007 + VOC2012 Dataset

Object Detection Frame Work	**References**	**mAP (%)**
Fast R-CNN	Zhang *et al.* (2020)	70
Faster R-CNN VGG-16	Zhang *et al.* (2020)	73.2
Faster R-CNN ResNet	Zhang *et al.* (2020)	76.4
YOLO	Lechgar*et al.* (2021)	63.4
SSD300	Abas *et al.* (2021)	74.3
SSD512	Sharif *et al.* (2021)	76.8
YOLOv2 544 × 544	Wang *et al.* (2021)	78.6
YOLOv3 416 × 416	Wang *et al.* (2021)	87.4
YOLOv3 544 × 544	Wang *et al.* (2021)	86.8
YOLOv3 608 × 608	Wang *et al.* (2021)	86.1
Proposed YOLO v3	-	88.2

The mAP value of the proposed detection method tested under VOC2007 + VOC2012 dataset was 88.2%. The mAP value of the Fast R-CNN detection method tested under VOC2007 + VOC2012 dataset was 70%. The mAP value of the Faster R-CNN VGG-16 detection method tested under VOC2007 + VOC2012 dataset was 73.2%. The mAP value of the Faster R-CNN ResNet detection method tested under VOC2007 + VOC2012 dataset was 76.4%. The mAP value of the YOLO detection method tested under VOC2007 + VOC2012 dataset was 63.4%. The mAP value of the SSD300 detection method tested under VOC2007 + VOC2012 dataset was 74.3%. The mAP value of the SSD512 detection method tested under VOC2007 +

VOC2012 dataset was 76.8%. The mAP value of the YOLOv2 544 × 544 detection method tested under VOC2007 + VOC2012 dataset was 78.6%. The mAP value of the YOLOv3 416 × 416 detection method tested under VOC2007 + VOC2012 dataset was 87.4%. The mAP value of the YOLOv3 544 × 544 detection method tested under VOC2007 + VOC2012 dataset was 86.8%. The mAP value of the YOLOv3 608 × 608 detection method tested under VOC2007 + VOC2012 dataset was 86.1%. It was found that the mAP value of the proposed detection method was 0.91% -39% higher than the existing detection methods.

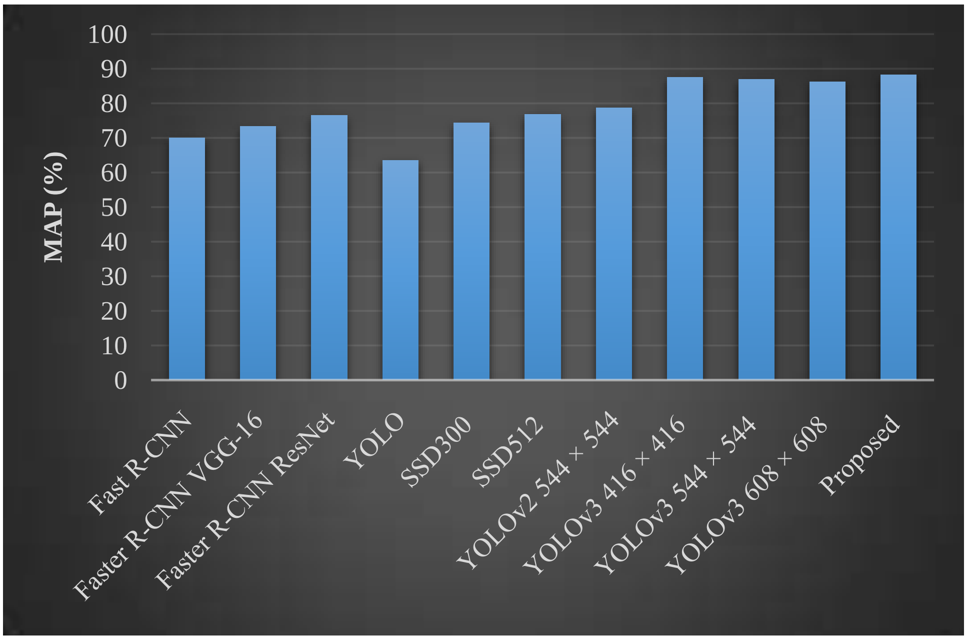


Fig. 5mAP values of proposed and existing models trained on VOC2007 + VOC2012 Dataset

The detection precision (mAP) performance of the proposed detection method tested under SCUT_FORU Text dataset was tabulated in the table. 4. The comparison of the detection precision (mAP) performance of the various detection method are graphically presented in Fig. 6.

Table. 4mAP values trained on SCUT_FORU TextDataset

Methods	References	mAP (%)
YOLOv3 416 × 416	Wang *et al.* (2021)	77.9
YOLOv3 544 × 544	Wang *et al.* (2021)	78.3
YOLOv3 608 × 608	Wang *et al.* (2021)	77.9
Proposed YOLOv3	-	80

The mAP value of the proposed detection method tested under SCUT_FORU Text dataset was 80%. The mAP value of the YOLOv3 416 × 416 detection method tested under SCUT_FORU Text dataset was 77.9%. The mAP value of the YOLOv3 544 × 544 detection method tested under SCUT_FORU Text dataset was 78.3%. The mAP value of the YOLOv3 608 × 608 detection method tested under SCUT_FORU Text dataset was 77.9%. It was found that the mAP value of the proposed detection method was 2.17% -2.69% higher than the existing detection methods.

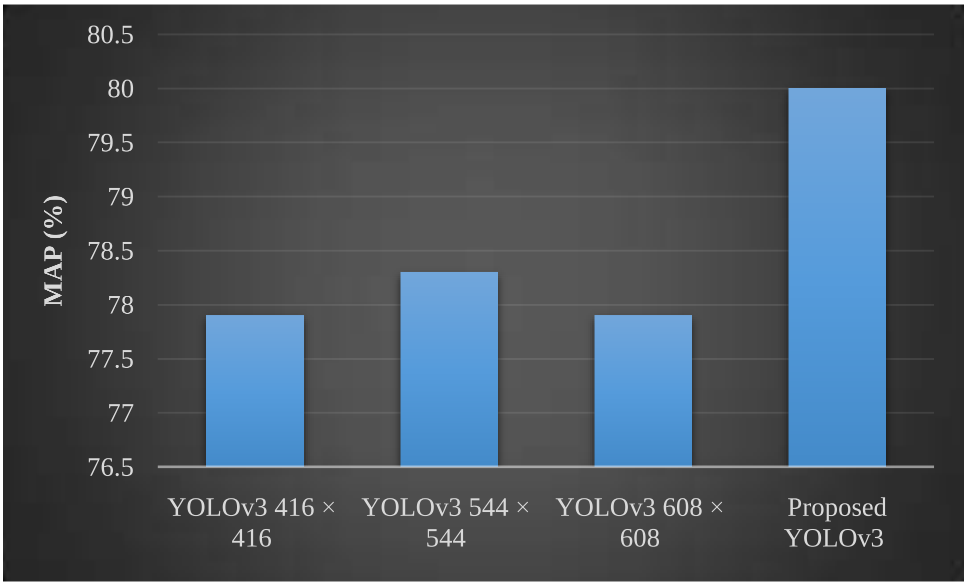


Fig. 6mAP values of proposed and existing models trained on SCUT_FORU TextDataset

IV. Conclusion

This paper presented a novel DL based object and text model for the detection of real-time objects. The performance of the proposed model was simulated and assessed using the datasets such as COCO, VOC2007 + VOC2012 and SCUT_FORU Text datasetsunder metrics such as precision (mPA). The mAP value of the proposed model was 58.7%, which is almost 4.07%-183% higher than the existing methods for the COCO dataset. The mAP value of the proposed model was 88.2%, which is almost 0.91% -39% higher than the existing methods for the VOC2007 + VOC2012dataset. The mAP value of the proposed model was 80%, which is almost 2.17% -2.69% higher than the existing methods for the SCUT_FORUdataset. It was concluded that the hybrid activation function used in this proposed helps in fast learning of the network, thus results in better precision results. The proposed model can detect the text and objects in few seconds, even though it was measured in this study which can be investigated in future.

References

[1]. Abas, S. M., & Abdulazeez, A. M. (2021). Detection and Classification of Leukocytes in Leukemia using YOLOv2 with CNN. *Asian Journal of Research in Computer Science*, 64-75.

[2]. Baimukashev, D., Zhilisbayev, A., Kuzdeuov, A., Oleinikov, A., Fadeyev, D., Makhataeva, Z., & Varol, H. A. (2019). Deep learning based object recognition using physically-realistic synthetic depth scenes. *Machine Learning and Knowledge Extraction*, *1*(3), 883-903.

[3]. Belongie, S., Malik, J., &Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, *24*(4), 509-522.

[4]. Busta, M., Neumann, L., & Matas, J. (2015). Fastext: Efficient unconstrained scene text detector. In *Proceedings of the IEEE international conference on computer vision* (pp. 1206-1214).

[5]. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., &Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213-229). Springer, Cham.

[6]. Cheng, G., Yang, J., Gao, D., Guo, L., & Han, J. (2020). High-quality proposals for weakly supervised object detection. *IEEE Transactions on Image Processing*, *29*, 5794-5804.

[7]. Dutta, S. (2020). A 2020 guide to deep learning for medical imaging and the healthcare industry. *Nanonets. com*.

[8]. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, *25*(1), 24-29.

[9]. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern recognition*, *77*, 354-377.

[10]. Haque, K. F., Haque, F. F., Gandy, L., &Abdelgawad, A. (2020, August). Automatic detection of COVID-19 from chest X-ray images with convolutional neural networks. In *2020 international conference on computing, electronics & communications engineering (iCCECE)* (pp. 125-130). IEEE.

[11]. Jiang, Y., Pang, D., & Li, C. (2021). A deep learning approach for fast detection and classification of concrete damage. *Automation in Construction*, *128*, 103785.

[12]. Ju, Y., Xu, Q., Jin, S., Li, W., Su, Y., Dong, X., & Guo, Q. (2022). Loess Landslide Detection Using Object Detection Algorithms in Northwest China. *Remote Sensing*, *14*(5), 1182.

[13]. Kavitha, M., Gayathri, R., Polat, K., Alhudhaif, A., &Alenezi, F. (2022). Performance evaluation of deep e-CNN with integrated spatial-spectral features in hyperspectral image classification. *Measurement*, *191*, 110760.

[14]. Kohli, H., Agarwal, J., & Kumar, M. (2022). An Improved Method for Text Detection using Adam Optimization Algorithm. *Global Transitions Proceedings*.

[15]. Lechgar, H., Bekkar, H., &Rhinane, H. (2019). Detection of cities vehicle fleet using YOLO V2 and aerial images. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*, 121-126.

[16]. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... &Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham.

[17]. Liu, Y., Wang, Y., Wang, S., Liang, T., Zhao, Q., Tang, Z., & Ling, H. (2020, April). Cbnet: A novel composite backbone network architecture for object detection. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 11653-11660).

[18]. Mahajan, S., & Rani, R. (2021). Text detection and localization in scene images: a broad review. *Artificial Intelligence Review*, *54*(6), 4317-4377.

[19]. Masita, K. L., Hasan, A. N., & Paul, S. (2018, November). Pedestrian detection using R-CNN object detector. In *2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI)* (pp. 1-6). IEEE.

[20]. Nath, N. D., &Behzadan, A. H. (2020). Deep convolutional networks for construction object detection under different visual conditions. *Frontiers in Built Environment*, *6*, 97.

[21]. Neumann, L., & Matas, J. (2015). Efficient scene text localization and recognition with local character refinement. In *2015 13th international conference on document analysis and recognition (ICDAR)* (pp. 746-750). IEEE.

[22]. Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*.

[23]. Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

[24]. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[25]. Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

[26]. Shanahan, J. G., & Dai, L. (2020, August). Introduction to computer vision and real time deep learning-based object detection. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3523-3524).

[27]. Sharif, M. I., Li, J. P., Amin, J., & Sharif, A. (2021). An improved framework for brain tumor analysis using MRI based on YOLOv2 and convolutional neural network. *Complex & Intelligent Systems*, *7*(4), 2023-2036.

[28]. Sharma, A., Vans, E., Shigemizu, D., Boroevich, K. A., & Tsunoda, T. (2019). DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific reports*, *9*(1), 1-7.

[29]. Solovyev, R., Wang, W., &Gabruseva, T. (2021). Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, *107*, 104117.

[30]. Suhao, L., Jinzhao, L., Guoquan, L., Tong, B., Huiqian, W., & Yu, P. (2018). Vehicle type detection based on deep learning in traffic scene. *Procedia computer science*, *131*, 564-572.

[31]. Wang, J., Song, L., Li, Z., Sun, H., Sun, J., & Zheng, N. (2021). End-to-end object detection with fully convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15849-15858).

[32]. Wang, K., Liu, M., & Ye, Z. (2021). An advanced YOLOv3 method for small-scale road object detection. *Applied Soft Computing*, *112*, 107846

[33]. Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

[34]. Yeo, B. L., & Liu, B. (1995). Rapid scene analysis on compressed video. *IEEE Transactions on circuits and systems for video technology*, *5*(6), 533-544.

[35]. Zhang, D. (2021, April). Object Detection Algorithm Based on YOLOv3 Model to Detect Occluded Targets. In *Journal of Physics: Conference Series* (Vol. 1881, No. 4, p. 042043). IOP Publishing.

[36]. Zhang, F., Luan, J., Xu, Z., & Chen, W. (2020). DetReco: object-text detection and recognition based on deep neural network. *Mathematical Problems in Engineering*, *2020*.

[37]. Zhu, Y., Yao, C., & Bai, X. (2016). Scene text detection and recognition: Recent advances and future trends. *Frontiers of Computer Science*, *10*(1), 19-36.

XXXXXXXXXX, et. al. “Text Detection and Object Recognition from Scene Images Using CNN and YOLOv3.” *IOSR Journal of Computer Engineering (IOSR-JCE),* 24(5), 2022, pp. 38-47.