Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 153–163

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Object Detection on Scene Images: A Novel Approach

Kaushik Das[a], Arun Kumar Baruah[b]

[a]*Dibrugarh University, Computer Science and Engineering, Dibrugarh 786004, India*
[b] *Dibrugarh University, Mathematics Department, Dibrugarh 786004, India*

Abstract

Convolutional Neural Networks (CNN) have played an important contribution to the significant development of Computer Vision. An important aspect can be the recognition of objects as it can be an essential part of Computer Vision. CNN's provides a one-step implementation for the detection and classification stages with an improved result. In both these stages, proper monitoring of object temporal will be attempted by observing the sequences in video, shape, size, presence, and location. In this paper, the authors have attempted to provide an implementation of the Object Detection method with a novelty along with text extraction from a scene image. A comparative analysis study of the performance evaluation of the proposed method has been made with various techniques at the end of the paper.

Keywords: Convolutional Neural Networks, ANN, CNN layers, Object Detection, FCN, YOLO

1. Introduction

Convolutional Neural Network or CNN or ConvNet is a popular algorithm in the area of Machine Learning. Digital Image Processing is the technique of manipulating digital photos with the help of a computer. The origin is from the area of signals and systems which focuses on processing images.

Object Detection is an approach similar to computer vision for identifying and locating things that are available in scene images and videos. The creation of bounding boxes around identified objects by object detection in particular enables us to determine whether they are present in a scene. One of the most prominent deep learning based object identification algorithms is the YOLO V3 methodology. YOLO V3 forecasts the objectless of each bounding frame using logistic regression. The YOLO V3 technique treats object detection as a regression concern. It estimates class chances and bounding box distances from full photos using a single feed-forward CNN. In contrast to region proposal-based algorithms, YOLO can employ universal knowledge about the complete picture while creating forecasts about the full picture which decreases background error.

10.1016/j.procs.2022.12.411

One may extract text from any image with the text extractor. An image can be uploaded and the tool will extract text from it. The goal of automatic text extraction without character recognition is to extract sections with text only. Image recognition and optical character recognition technologies have become an essential part of our daily lives as a result of the ever-increasing computing power and accessibility of scanning equipment. Optical Character Recognition can swiftly turn printed documents into digital text files, which can then be modified by the user. As a result, digitizing documents takes very less time which is especially useful for storing large amounts of printed materials.

Contribution and Motivation of the Paper-Investigation of the proposed system are to know the elements in an image and study the formation of the hierarchical structure. The paperwork is proposing a purely visual representation that allows the image content to be examined through a unified framework for object or stuff detection. The paper has described the existing methods related to object detection and text extraction.

The paper starts with an overview of the research made with YOLO V3, and the fundamental steps involved in Image Processing. The second heading is devoted to the Preprocessing of the Scene Images with an explanation of image cropping, filtering, intensity adjustment, histogram equalization, and image categorization. The next heading is devoted to the Segmentation of Scene Images followed by a heading with Object Detection and Text Extraction from Scene Images. The paper concludes with the Results and Discussions of the study based on the performance of the proposed framework which is explained in detail. The last heading deals with the Conclusion of the study in detail with the Future Scope of the proposed Image Processing Techniques.

2. Preprocessing of Scene Images

Pre-processing involves reducing undesirable distortions or enhancing certain features of the image that are crucial for subsequent processing. It may also include methods for manipulating image data as well as geometric transformations of images (such as rotation, scaling, and translation).

The brightness p from the scale $[p_0, p_k]$ is converted to the brightness q from the scale $[q_0, q_k]$ (p) in a grayscale transformation known as q = T, which is independent of the location of the pixel in the image. Clipping is to be used for values below p_0 and above p_k. Values below p_0 are referred to as q_0, and values above p_k are referred to as q_k. The shape of the curve showing the correlation between the values in the input and output an image is specified by the Alpha argument. The mapping is weighted toward higher output values if alpha is less than 1 and toward lower, darker output values if alpha is greater than 1. The value defaults to 1 if the parameter is not specified. A user can utilize graphical controls to modulate brightness, contrast and alpha correction [1].

The histogram equalization approach is another option for improving visual contrast. It involves altering an intensity image's values so that the output image's histogram matches a predefined histogram. The statistical gray-level qualities of the region are used to describe it [2]. A one-dimensional array with n items is used to represent an image with n grey levels. The array's nth entry contains the number of pixels with a grey level of n.

There are several ways to smooth the picture in OpenCV. SMOOTH filters are utilized in the image smoothing process. The image is smoothed by utilizing an LPF kernel to convolve it. It can be used to reduce noise and lead to blurred edges [3].

3. Segmentation of Scene Images

Segmenting images based on their semantic content is essential for the creation of vision sensors for self-driving cars. Other methods, like object detection, only produce the bounding boxes of an object. Deeper information, such as the item's border and shape, is provided by semantic segmentation [4].

Convolution layers are substituted for the fully connected layer in a conventional CNN in Fully Convolutional Networks (FCN) [5], which are then up-sampled to return the output to the original image size. In terms of image semantic segmentation, an FCN is a pixel-wise FCN [6].

Consecutive pooling procedures during the encoding step, on the other hand, frequently result in a drop in featureresolution, which degrades final segmentation performance.

4. Related Work

Modern image conceptual segmentation methods frequently employ FCN [5], which trains over an edge network for image prediction and produces cutting-edge outcomes. SegNet [7] and U-Net [8] are two encoder-decoder networks that have been suggested to enhance image segmentation outcomes. In pixel-level prediction tasks, contextual information is critical. Sizeable changes occur in complex scenes, posing significant hurdles for sophisticated feature representations. To increase the receptive field and encode multi-scale context information, dilated convolutions were used. A network that aggregates feature maps from various resolutions was suggested by Zhao et al. [9] for parsing pyramidal scenes. DeepLabv3+ [10] used parallel Atrous Spatial Pyramid Pooling (ASPP), which joins simultaneous distended convolutions of different rates on the feature map, to successfully encode multi-scale information. A multi-scale profound context convolution layers network was developed by Zhou et al. [11] and includes feature maps from various network levels.

Using the non-local operator [13], DANet [12] examines orthogonal interplay in both the spatial and the channel dimensions. Long-range dependencies can be extracted directly from any two points in an image using non-local procedures. The DANet position attention module collects attributes of each location selectively using a total weighted sum of all locations. To determine total scene statistics, CFNet [14] adds an extra global average grouping path. Both models' networks do not incorporate local multi-scale properties. To enhance context aggregation, OCNet [15] employs consciousness to learn pixel-level object relevant information. In [16] and [17], researchers attempted to include more depth information to improve semantic segmentation.

5. Feature Extraction of Scene Images

A step in the dimensional reduction procedure is feature extraction. It segregates a large amount of raw data into more manageable groups. Processing will be easier as a result. The most important characteristic of these enormous data sets is the high number of variables. Therefore, feature extraction helps to extract the most suitable feature from those massive data sets, effectively reducing the volume of data, by choosing and combining variables into features.

The method of feature extraction is used to decrease the amount of resources needed without sacrificing any important or pertinent data needed for large data set analysis. The learning and generalisation phases of the machine learning process can be accelerated by feature extraction. Additionally, it facilitates the model's construction with less machine labour and expedites the data-reduction procedure.

5.1. Image Feature Representation

Few earlier works have looked at both image feature extraction and image feature representation, which are important aspects of multimedia processing. We give a complete overview of recent advances in image feature extraction and image feature representation in this study. We look at the usefulness of combining global and local characteristics in image processing. The bag-of-visual-word is then developed, followed by another sort of feature representation.

Finally, there are a number of intriguing concerns that should be investigated further in the future. First, investigate the relationship between the amount of characteristics and the ultimate performance. Intuitively, it is impossible that the higher the number of features, the better the overall performance. Then, investigating the link between feature representation and end. It features techniques for presenting features (global, block-based and region- based features). The partition or segmentation size has an impact on the final performance, particularly for block-based and region-based features. Third, it's also intriguing to investigate the relationship between their proper mix and ultimate performance to determine if the combination can boost the performance even more.

6. Object Detection and Text Extraction of Scene Images

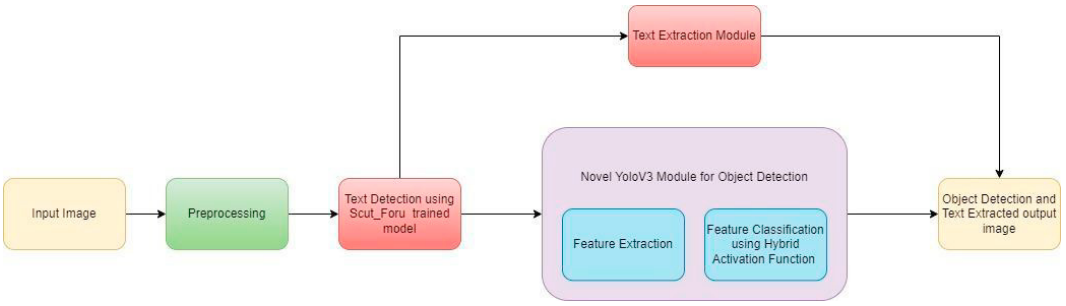

Fig 1. implies the entire flow diagram of the new network framework.

Convolutional layers are utilized for text detection maps initially, and feature maps are used by the object and text detection component. Object detection is employed in a variety of applications, employing intelligent transportation systems, Advanced Driver Assistance Systems (ADAS), and automated driving structures. One-stage and two-stage object detection techniques are the two different types. [18].

Scene detection and item detection have seen rapid progress in recent years that might employ another DL approach and also operate with NN security.
In YOLOv3, the FPN (Feature Pyramid Networks) concept is used to combine the output of the later layers with that of the intermediate layers. The YOLOv3 framework uses an FCNN to find objects and messages in real-scene photos. The image`s features in several scale feature maps are extracted utilizing a convolutionalnetwork.

7. Text Extraction from Image

Three steps are involved in extracting text from an image: (1) pre-processing the image taken as input, (2) marking each location for the text, and (3) text extraction. Our recommended method accepts a colour image as input. The provided image must therefore undergo preprocessing. The image is initially divided into two smaller versions. If the image is not divided into smaller pieces, some very small text will be lost, and the final recovered text image will still contain some noise. The next step is to convert two sub-images into two grayscale images, which are then translated into two binary images. Before creating two sub-images, we applied the text extraction procedure to each sub-image. The text is then copied and pasted into a second grayscale image. The preparation procedures are not necessary if the supplied image is grayscale. However, this is not very effective when it deals with image that is not clearly visible.

Based on earlier research on OCR-based systems, text detection and recognition for natural settings and character recognition for commercial promotion photos are made. To reduce the number of lawful challenges that could arise from the exercise of incorrect words in advertisements, an unacceptable words detection system will be added after the completion of the essential character detection system.

Text identification and localization in natural scene photos remains difficult due to a variety of parameters such as font, size, color, and orientation changes. Image distortion and deformation can also be caused by the intricacy of the backdrop or variations in illumination.

7.1. Faster R-CNN Object Detection

A categorized schema GloVe-FRCNN is built by integrating the GloVe text categorization technology with the Faster-RCNN. This model can forecast the chance of erroneous categorization [19]. The fundamental aspect of ontology is how to appropriately manage various types of data in order to retrieve the possible common information hidden behind various types of data and how to increase model identification and classification accuracy. Through multi-level screening, a suitable categorization practice is to actualize the diverse application of large data.

Additionally, the report is evaluated using GloVe text categorization [14] and Faster-RCNN [17] image detection technologies in order to obtain the Vespa mandarinia categorization results.

The text conveys a lot of information succinctly and quickly. Because it has so many uses in computer vision, retrieval of this information is crucial for human learning and comprehension [20]. Over the past few decades, a number of techniques for text identification from documents and natural scene text have been proposed. The authors have implemented, analyzed, and evaluated two of these architectures, namely Faster R-CNN and Effective and Accurate Scene Text Detector. It was intended to compare them based on the types of pipelines, the quantity of layers, the rate at which they were executed, and the degree of accuracy. Despite the fact that neither network is flawless, the EAST detector demonstrated exceptional performance when compared to other architectures. The model demonstrated nearly 78% accuracy for the DenseNet201's high receptive field, deeper network, and addition of the dice loss function. This method lacks a dataset. A new constructed dataset will be useful and the concept integration of recognition with detection can be done.

7.2. YOLO

As a target detection system, YOLO has a quick detection speed and is appropriate for target identification in a real-time settingWhen tried to compare to other desired detection systems of a similar nature, it provides a greater detection accuracy and a quicker detection time. Even in a complicated setting, good detection accuracy can be guaranteed. Deep CNNs have shown outstanding performance in the areas of semantic segmentation, object detection, and image classification. The standard technique did not produce an accurate outcome in terms of face detection accuracy. To increase the precision of face detection, the deep learning model is used. In the study, the efficiency of face detection is compared to the old method's accuracy. The recommended model employs the CNN as a deep learning strategy for recognizing faces in videos.

7.3. HISTOGRAM OF ORIENTED GRADIENTS

The main difficulties are human detection techniques, classifier design, and feature representation. The majority of human detection frameworks make use of either local or global feature techniques. Linear SVM is a common and well-built classifier for human detection. The suggested technique detects multi-posture humans by combining the Hog Bo feature with a swift stabilized non- linear SVM [21].

In the current study, a method for categorising distinct text and non-text sections in a document image is developed. As a feature descriptor, a modified form of histogram of oriented gradient (HOG) is employed [22]. Text is recognized using Corner-HOG and Extremal Regions (ERs). The experimental findings show that the Corner HOG based pruning approach can eliminate 83 on average [23]. The leftover ERs are then arranged into text lines, and potential text lines are examined using the black-white transition feature and the HOG covariance descriptor. The results of the experiments indicate that the suggested method for trimming non-text components is successful.

7.4. FAST RCNN OBJECT DETECTION

The whole image is transmitted to the network throughout the training stage to extract CNN features and suggestions [24]. During testing, the whole normalized picture is instantly fed through the CNN, and the feature: map created by the final convolution layer includes suggestions. In the classification process, FastRCNN uses a deep network to evaluate categorization and arrive at position regression.CNN first transforms the picture into a strong feature convolution diagram preferably a graph with high dimension. This system combines the over improvement on a quick RCNN with automotive categorization and counting on live streaming footage. Additionally, the suggested system was developed and trained using parameters for getting improved information on the Datasets like Stanford Vehicle and the Myanmar Vehicle.

7.5. FASTER RCNN TEXT EXTRACTION

Unlike classic feature extraction methods, the Faster RCNN retrieves features using the kernel with convolution, every elements of neuron, and the preceding layer`s neighboring accessible field. Because shared weights are produced by supervised learning, the improved RCNN network which works faster is mostly applied to recognize double dimensional pictures. As a result of avoiding artifacts, the RCNN with faster execution capacity

has the advantage of knowledge to share weights with the training data. The planned and executed model can recognize a number of distinguished types of text and logos from usual seen sign boards with a background and light forefront text.

The fusion technique underlies its distinctiveness for image identification and text categorization as a latest derivation approach from applied mathematics that combines various models using systematic linear fitting. We employ a novel fitting strategy that is based on the notion of adjustment to fit the model with variable weights inorder to increase the training data. Combination of GloVe and FRCNN leaded to a classification model created by integrating GloVe text classification technology with a FasterRCNN. Additionally, the report is examined and results for Vespa mandarinia classification are obtained using GloVe document classification and FasterRCNN picture recognition technologies.

8. Experimental Results

The standard datasets like CoCo, VOC2007 + VOC2012 and SCUT FORU database were read for experimental purpose based on which the following explanations of evaluation were being made in the sections below:

8.1. COCO Dataset

One of the widely used benchmark data sets for image detection and segmentation was the CoCo data set which was written by Microsoft. This dataset comprises of natural images of complex scenes that includes multiple objects. It has 91 type of objects with more than million labeled instances in 328K images. COCO has a fewer number of categories compared 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).

8.2. VOC2007 + VOC2012

The VOC2007 dataset is the challenge to identify objects from a various visual objects categories in a scene image. This database comprises of 9963 annotated images. The VOC2012 dataset presents the same difficulty as the VOC2007 dataset, expanding the training set.

8.3. SCUT_FORU Text

The SCUT_FORU database contains Chinese 2k and English 2k. The dataset's label format is {x,y,w,h,label}. The rectangular box's {x,y} coordinates are top-left. The dimensions of the rectangular box are w and h. The text region's word label is referred to as the "label." There are 1715 total images, 1200 of which are used for training and 515 for testing. The dataset averages 3.2 words and 18.4 characters per image.

8.4. Experimentation

The proposed method was experimented with the Coco Dataset, VOC2007 + VOC2012 dataset and SCUT FORU Database as discussed in the previous section. The concept is to conduct an experiment that explains how well the proposed object detection method can perform under the harsh circumstances.

8.5. Metrics for Evaluation

The performance of the suggested detection method was assessed using the mean Average Precision (mAP) as the performance metric.

The portion of a predicted region that corresponds to the actual data is what is meant by precision. Equation contains the Precision expression: 1.1

$$Precision = \frac{TP}{TP+FP} \tag{1.1}$$

The metrics True Positive (denoted as TP), True Negative (denoted as TN), False Positive (denoted as FP), and False Negative (denoted as FN) are used to calculate precision.

TP: A situation where the predicted value and actual value are both positive.
TN: A situation where the predicted value is negative and the actual value is negative.
FP: A situation when the predicted value is positive but the actual value is negative.
FN: A situation when the predicted value is negative but the actual value is positive.

Average Precision or shortly AP is a calculation that averages precision values across different recall levels. When the AP values were high, the performance was better. The expression for AP is given in equation: 1.2.

$$AP = \sum(R_n - R_{n-1})P_n \tag{1.2}$$

where Pn and Rn are the precision and recall at nth threshold.

Recall calculated as the percentage of the ground truth that is present in the predicted region. The expression forrecall is given in equation: 1.3.

$$\text{Recall} = \frac{TP}{TP+FP} \tag{1.3}$$

Mean Average Precision or shortly mAP is the widely used metric for performance evaluation in case of object recognition. It represents the average precision determined across all classes. Following equation contains the definition of mean Average Precision (mAP).: 1.4.

$$\text{mAP} = \frac{1}{N}\sum_{l=1}^{N} AP_{\text{i}} \tag{1.4}$$

where AP_i stands for average precision, and N stands for all classes combined..

8.6. Performance testing of the suggested approach

The proposed detection method's detection precision (mAP) performance was tabulated in the table 1 and tested against the COCO dataset. In Fig. 4, a graphic comparison of the performance of the various detection methods' detection precision (mAP) is shown.

Table 1. mAP values trained on COCO Dataset.

Method	**References**	**Estimated mAP (%)**
PG-PS-FR-CNN	Cheng *et al.*(2020)	20.7
DETR	Carion*et al.*(2020)	44.9
POTO-ResNext- 101-CN	Wang *et al.*(2021)	47.6
CBNET	Liu *et al.*(2020)	53.3
YoloV3	Zhao & Li,(2020)	53.2
Weighted Boxes Function	Solovyevet 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 Weighted Boxes Function detection technique's mAP value, as tested using the COCO dataset, was 56.4%. It was discovered that the suggested detection technique's mAP value was 4.07% -183% better than the ones currently in use.

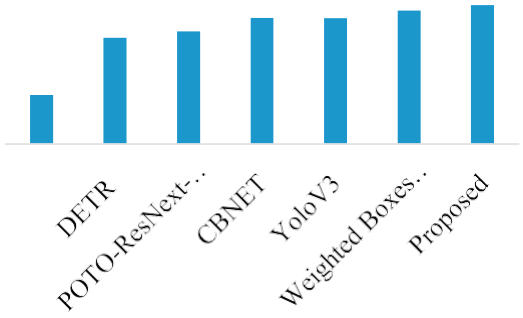

Fig. 2. Detection precision (mAP) performance of the proposed detection method

The proposed detection method's recognition precision (mAP) effectiveness was tabulated in the table 2 and tested using the VOC2007 + VOC2012 dataset. Fig. 2 provides a graphic comparison of the detection precision (mAP) performance of the different detection methods.

Table 2. mAP values trained on VOC2007 + VOC2012 Dataset.

Object Detection Frame Work	References	Estimated mAP (%)
Fast R-CNN	Zhang *et al.*(2020)	70.0
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

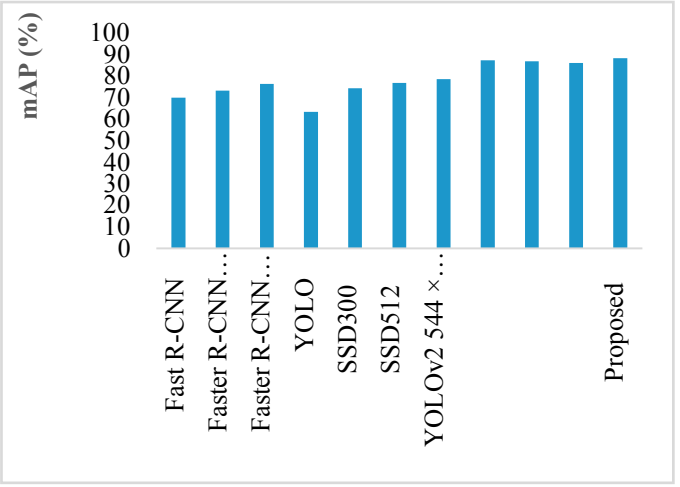

Fig. 3. Detection precision (mAP) effectiveness of the proposed detection method.

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.

The proposed detection method's detection precision (mAP) effectiveness was tabulated and tested using the SCUT FORU Text dataset. 3. Fig. 3 provides a graphic comparison of the detection precision (mAP) effectiveness of the different detection methods.

Table 3. mAP values trained on SCUT_FORU TextDataset.

Methods	References	Estimated 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
ProposedYOLOv3	-	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.

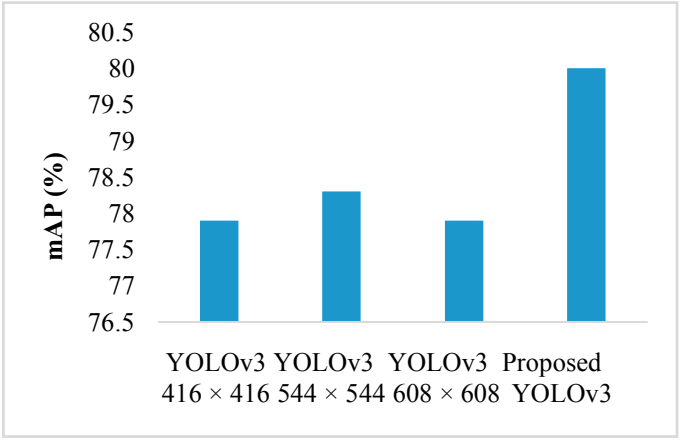

Fig. 4. Detection precision (mAP) effectiveness of the proposed detection method

9. Conclusion

The suggested approach in this study has been found to be effective and can be used to extract text from scene images and detect historical objects without the use of optical character recognition. The proposed technique combines each item detection and textual content throughout the work. The paper has included implementation of a few scenarios where the diagnosed textual content contexts round the items and make capable of use to differentiate the item. The proposed technique will be applicable in multi disciplinary datasets thereby building its capacity in large programs together with various structures.

References

[1] Mordvintsev, A., & Abid, K. (2014). "Opencv-python tutorials documentation." *Obtenido de https://media.readthedocs. org/pdf/opencv-python-tutroals/latest/opencv-python-tutroals.pdf.*
[2] Long, J., Shelhamer, E., Darrell, T. "Fully convolutional networks for semantic segmentation." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.*
[3] Noh, H., Hong, S., Han, B. "Learning deconvolution network for semantic segmentation." *In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1520–1528.*
[4] Badrinarayanan, V., Kendall, A., Cipolla, R. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 2481–2495.*
[5] Chen, L.-C., Papandreou, G., Schroff, F., Adam, H. "Rethinking atrous convolution for semantic image segmentation." *arXiv 2017, arXiv:1706.0558.*
[6] Yang, M., Yu, K.., Zhang, C., Li. Z., Yang, K. "DenseASPP for semantic segmentation in street scenes." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.*
[7] Zhao, H., Shi, J.; Qi, X., Wang, X., Jia, J. "Pyramid scene parsing network." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.*
[8] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. "Encoder-decoder with atrous separable convolution for semantic image segmentation." *In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 801–818.*
[9] Wang, X., Girshick, R., Gupta, A., He, K. "Non-local neural networks." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.*
[10] Zhang, H., Zhang, H., Wang, C., Xie, J.. "Co-occurrent features in semantic segmentation." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 548–557.*
[11] Yuan, Y., Wang, J.. "OCNet: Object context network for scene parsing." *arXiv 2018, arXiv:1809.00916.*
[12] Chen, C. H.. (2020) "A cell probe-based method for vehicle speed estimation." *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 103(1), 265-267.*
[13] 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).*
[14] Shien Sun, Zheng Gao, Chenyu Huang, Haitao Yu. (2021) "GloVe-FRCNN: Comprehensive network algorithm for Vespa mandarinia image-text extraction and classification." *International Conference on Communications, Information System and Computer Engineering (CISCE).*
[15] Mayank, Swapnamoy Bhowmick, DishaKotecha, Priti P Rege. (2021) " Natural Scene Text Detection using Deep Neural Networks." *6th International Conference for Convergence in Technology (I2CT).*
[16] Yuanyuan Feng, Yonghong Song, Yuanlin Zhang. (2015) "Scene text localization using extremal regions and Corner-HOG feature." *IEEE International Conference on Robotics and Biomimetics (ROBIO).*
[17] Lei Quan, DongPei, BinbinWang, WenbinRuan. (2017) "Research on Human Target Recognition Algorithm of Home Service Robot Based on Fast-RCNN." *10th International Conference on Intelligent Computation Technology and Automation (ICICTA).*
[18] Wang Yang, Zheng Jiachun. (2018) "Real-time face detection based on YOLO." *1st IEEE International Conference on Knowledge*

Innovation and Invention (ICKII).
[19] Reagan L. Galvez, Elmer P. Dadios, Argel A. Bandala, Ryan Rhay P. Vicerra. (2011) "YOLO-based Threat Object Detection in X-ray Images." *11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM).*

[20] Stoble B, Jain;M.Sreeraj. (2015) "Multi-posture Human Detection Based on Hybrid HOG-BO Feature." *Fifth International Conference on Advances in Computing and Communications (ICACC).*
[21] Lei Quan, Dong Pei, Binbin Wang, Wenbin Ruan. (2017) " Research on Human Target Recognition Algorithm of Home Service Robot Based on Fast-RCNN." *10th International Conference on Intelligent Computation Technology and Automation (ICICTA).*
[22] Boya Wang, Jianqing Xu,Junbao Li, Cong Hu, Jeng-Shyang Pan. (2017) "Scene text recognition algorithm based on faster RCNN." *First International Conference on Electronics Instrumentation & Information Systems (EIIS).*
[23] Latha H N, Sadhan Rudresh, Sampreeth D, Sangamesh M Otageri, Saurabh S Hedge. (2018) "Image understanding: Semantic Segmentation of Graphics and Text using Faster-RCNN." *International Conference on Networking, Embedded and Wireless Systems (ICNEWS).*
[24] Zuge Chen, Kehe Wu, Yuanbo Li, Minjian Wang, Wei Li. (2019) "SSD-MSN: An Improved Multi-Scale Object Detection Network Based on SS." *IEEE Access (Volume: 7).*

***