Scene Based Text Recognition From Natural Images and Classification Based on Hybrid CNN Models with Performance Evaluation

Original Scientific Paper

Sunil Kumar Dasari
Department of ECE, SOE,
Presidency University, Bangalore, India-560054
sunilkumar.d@presidencyuniversity.in

Shilpa Mehta
Department of ECE, SOE,
Presidency University, Bangalore, India-560054
shilpamehta@presidencyuniversity.in

Abstract *– Similar to the recognition of captions, pictures, or overlapped text that typically appears horizontally, multi-oriented text recognition in video frames is challenging since it has high contrast related to its background. Multi-oriented form of text normally denotes scene text which makes text recognition further stimulating and remarkable owing to the disparaging features of scene text. Hence, predictable text detection approaches might not give virtuous outcomes for multi-oriented scene text detection. Text detection from any such natural image has been challenging since earlier times, and significant enhancement has been made recently to execute this task. While coming to blurred, low-resolution, and small-sized images, most of the previous research conducted doesn't work well; hence, there is a research gap in that area. Scene-based text detection is a key area due to its adverse applications. One such primary reason for the failure of earlier methods is that the existing methods could not generate precise alignments across feature areas and targets for those images. This research focuses on scene-based text detection with the aid of YOLO based object detector and a CNN-based classification approach. The experiments were conducted in MATLAB 2019A, and the packages used were RESNET50, INCEPTIONRESNETV2, and DENSENET201. The efficiency of the proposed methodology - Hybrid resnet -YOLO procured maximum accuracy of 91%, Hybrid inceptionresnetv2 -YOLO of 81.2%, and Hybrid densenet201 -YOLO of 83.1% and was verified by comparing it with the existing research works Resnet50 of 76.9%, ResNet-101 of 79.5%, and ResNet-152 of 82%.*

Keywords*: CNN, Scene based text detection,RESNET50,Text Detection,YOLO*

1. INTRODUCTION

Now day's digital world has increased its digital image or video database sizes, which is used for social media, digital education, commercial purpose, etc. Image-based retrieval has become a dynamic research such as computer vision, artificial intelligence, pattern recognition, etc. The CBIR method is of two types which are 1. Text-based 2. Content-based systems. The text-based method used for Google, Yahoo and Bing, etc., focuses on text keyword searches. The text-based image recovery method has numerous confines. For instance, once the keyword 'Apple' is given, retrieved results for the Apple laptop, company logo, apple fruits, etc. However, this text keyword search does not achieve good performance and does not describe the visual content of an image's properties.

The above problem is rectified by proposing a hybrid method; currently, many academics focus on hybrid features with re-ranking. The hybrid model is defined as the feature level that provides a stabilized order to retrieve similar images. The CBIR technique has resolved this problem, which uses hybrid features centered on low-level aspects like color, texture, shape, etc. Image characteristics are defined by low-level features. Those features that describe visual content might refer to color, texture, shape, etc., and hence, it helps to retrieve similar results and achieves better performance. In today's society, people learn various subject matter aspects that are related to different domains and industries. The subject matter learned consists of varied amounts of knowledge and concepts. With the different types of subjects that are available on a global scale, there exists a requirement to assess the people to see whether the supplied knowledge is encompassed within the people.

This section will give a brief introduction to the significance of scene-based text recognition and classification techniques. This part will deliver a momentary summary of the state of the art of text classification techniques used for several applications. This section provides the background of the research paper, including the overview of deep learning models for text recognition, along with the challenges associated with it. The section will further elaborate on the objectives of the study and research significance. In the 1950s, the automatic form of text summary instigated an impression, and in the late 1950s, Hans Peter Luhn issued an article named "Automatic creation of literature abstracts" [1] which could deploy features like word and sentence frequency to excerpt elementary stretches from the text for summary purposes.

Conferring to Radev et al. Summary is demarcated as "a text produced from one or more texts that convey the important information in the original text(s), and that is no longer than half the length of the original text(s) and usually significantly less than that." [2] Abstracts assist in realizing the implication of the text. Text summarization supports consumers in accomplishing a huge quantity of data by summarizing documents and integrating additional appropriate evidence.

The text summarization procedure consists of 3 main phases such as analysis, transformation, and combination [3]. The authors of [4] examined and related the performance of 3 diverse processes. The difficulty of scene text recognition compacts with identifying text in natural scene images. In conventional means, text recognition was engrossed in identifying printed text in documents, and such kind of schemes projected images to be black and white and in a document style layout embracing text lines.

In this research, an efficient technique is presented to perceive texts in natural scene images. Also constructs the "CRF model" by relating "CNN scores of MSERs" and multiple neighborhood information. Moreover, missing text mechanisms were additionally recovered using perspective information. Additionally, it incorporates gray and binary features and designs shape-specific classifiers to exactly validate text lines. The suggested system has been estimated on four public benchmarks and succeeds in hopeful performance, representing the efficiency and robustness of this method. The most apparent drawback of CRF is the high computational intricacy of the training phase of the system.

This research work focuses on scene image text detection with the aid of a novel approach. A combination of YOLO and RESNET is the first case, combined INCEPTIONRESNET, AND YOLO and combined DENSENET201 AND YOLO approach. The precision value for all those methodologies will be evaluated for better output.

The other sections of the work are as follows: Section 2 explains the Literature Review that compares the existing research on video-based database systems. Section 3 defines the research methodology - a novel scene-based text recognition and classification model using hybrid CNN. Section 4 explains the analysis of simulation results and performance evaluation of the suggested methodology. Finally, section 5 describes the conclusion of the research.

2. LITERATURE REVIEW

In video-based database systems, every video is tagged manually with the support of several keywords to assist with the searching and retrieval process. This system is found to be protracted and unreliable in nature; for instance, any two users could use various keywords in order to look for a similar video. Another scheme is to extract keywords from text which appears in the frame [5]. Mainly, any video text is divided into two groups such as graphics text and scene text. During the editing section, the graphics text is added to the video content separately. Scene text occurs logically in the prospect captured by the camera. By means of the rapid development of the Internet, there exists an aggregate demand for text detection from video.

Several techniques have been established since earlier research, but text detection still faces several challenges. The major aspects which show the disadvantage is unconstrained colors, size, and alignment of characters. Also, the scene text gets exaggerated by lighting environments and perspective falsifications [6]. Text detection could be mainly categorized into three different means such as "component-based, edge-based and texture-based approaches" [7,8]. Owing to color draining and short divergence of text lines, connected components or elements might not preserve the complete shape of the character, and hence, these approaches cannot be applicable to video images.

There is a vast amount of data found online, so obviously, summarizing the information has special importance. Text summarization generates an overall summary of the given document. There are two different summarization techniques that can be used to generate the summary, extractive and abstractive. But the majority of Indian language summarization works focused on the extractive approach since the grammar rules are a little more complex in such languages to generate an abstract one.

The uses of image processing are developing each day, and some of its applications are: -

1. Pattern recognition - Pattern recognition involves studying and recognizing various patterns like handwriting. It is integrated with artificial intelligence like computers. Assisted diagnostics and handwriting can be done easily.
2. Video Editing - It is also a digital image processing field. A collection of frames or photos is arranged in such a way that it makes the photo flow faster. Includes reduced motion detection of sound movement and color space modification etc.

3. Image sharpening – In this, the appearance and field of an image could be changed. Image sharpening manages the image and gives the desired effect. It includes individual modification, blurring, sharpening, retrieval, and image recognition.
4. Robot Vision - Many robotic machines operate with the use of digital images, where they can see their paths, for example, the roots of obstacle detection and the next line of robots.

Edge-based systems were proposed in order to overcome the low contrast issues. Arithmetical topographies from Sobel edge maps were extracted [9] in 4 directions, and K-means were deployed to determine the pixels into text and non-text clusters. Though this method was found vigorous alongside the composite environs, it could not meet up with detecting low contrast text and text of a lesser font. Moreover, it was more exclusive to accomplish due to the huge feature set. Two filters were deployed in [10] to develop the edges in the text areas, and here, several threshold values were employed to choose whether to progress the edges in a particular area. Thus, it could not simplify in a better way for different datasets. To discover the potential line segments, research [11] analyzed the maximum gradient difference. The multi-orientation problem was addressed in [12] with the support of a system based on the laplacian and skeletonization concepts. This technique ended in a high false positive rate and misdirection ratio for multi-oriented text. This is owing to the heuristics entailed in the segmentation approach [13]. A novel method based on a 2-level classification system and two sets of features specifically developed for capturing equally inherent features of subjectively oriented video texts were suggested in [14]. Gradient Vector Flow based method was proposed to detect the scene text which deeds character gap [15]. This approach did not meet the necessary general symmetry points due to the small font, low contrast, and complex background in the video. The Discrete Cosine Transform (DCT) factors of the intensity images have been extensively deployed as the texture elements for text detection [16-18].

Analysis of visuals like images and videos and their evolving systems is a stimulating and significant task to be enhanced and accomplished on benchmark datasets. This issue is resolved by deploying the STN-OCR model, including deep neural networks and spatial transformer networks. This research network design model consists of 2 phases such as localization and recognition network. In the localization network, it advocates text regions and creates the sampling grid [19]. A novel approach based on wavelet medium moments is presented in [20], along with a new indication of angle projection for perceiving multi-oriented text in the video.

A “deep neural network” for perceiving text in natural scene images was suggested in [21], which uses the drawbacks of arbitrarily oriented texts and intricate background images. This approach mentioned is constrained to natural scene images then not applicable to video as it includes several actions in it. A novel method was suggested in [22] for identifying scene text with the aid of deep reinforcement learning. Here, an agent is provided with a state and learns to predict future return aspects and develops sequential-based decisions to identify scene texts. For reading scene text in the wild centered on scene text proposal, the new method is presented in [23]. For natural scene text detection, [24] leveraged color prior centered MSER which extracts stroke features with support of strokes width distance depending on segmented edges. Several drawbacks still exist, such as multi-type texts, graphics, and scene texts. In [23], “Fourier laplacian filtering and hidden Markov model” is suggested for text and non-text classification. A two-staged approach [25] is suggested by means of a quadrilateral scene text detector, but these structural aspects might not hold for the irregular form of text. An adaptive bezier curve-based network is deployed in [26] for text detection in scene images. Similarity prediction among the textual elements of several views of natural scene images was presented in [27] for attaining advanced performance. A scene text detection-based segmentation system is proposed in [28], which introduced the text-mountain model. An edge descriptor model is used in [29] to discover local binary patterns. CNN model is explored in [30] for multi-lingual based text detection for detecting natural scene images. To enhance the capability of recognition-based methods, research in [31] introduced fractals, wavelet transforms, and optical flow for undertaking the drawback of video and natural scene images. Dependencies among word tokens in a sentence are extracted in [31], which assist 2D spatial dependences among two characters in a scene text image.

3. RESEARCH METHOD

The research methodology includes the proposed research flowchart given in Figure 1 for designing a novel scene-based text recognition and classification model using hybrid CNN. The following steps, such as data collection, data preprocessing, feature extraction, and classification, are discussed as follows:

3.1 DATA COLLECTION

In the object detection phase, "ground truth" states data gathered at a particular location. Ground truth permits image sequence data to be connected to physical features and materials on the ground. The assembly of ground truth data empowers the discovery of the entities in the image or video format.

3.2 DATA PREPROCESSING

Data preprocessing techniques include text filtering, binarization, image segmentation, etc. Image preposssing are steps taken to format images before they are used by model training and inference systems. This contains re-sizing, orienting, and color corrections. Bounding box on all regions in the image and selects the point with the highest score as the center of the crop.

3.3 FEATURE EXTRACTION

Feature extraction is described as extracting relevant features from data. Here, the function of max pooling could minimize feature map dimension and solves the fitting issue. By depending upon these feature values, the network could be trained and tested.

K-medoids Clustering is an unsupervised means of clustering system that cluster objects in labeled data. In "K-medoids Clustering," as an alternative to taking the centroid of objects in a cluster as a location point as in k-means clustering, this research considers medoid as a reference point.

In ResNet50, layers are unequivocally reformulated as learning residual functions with respect to layers input. The resnet50 could categorize images into 1000 object groups, including keyboard, mouse, pencil, and many animals. As an effect, the network has learned rich feature depictions for an inclusive array of images. The network has an image input size of "224-by-224" for further pre-trained networks in "MATLAB."

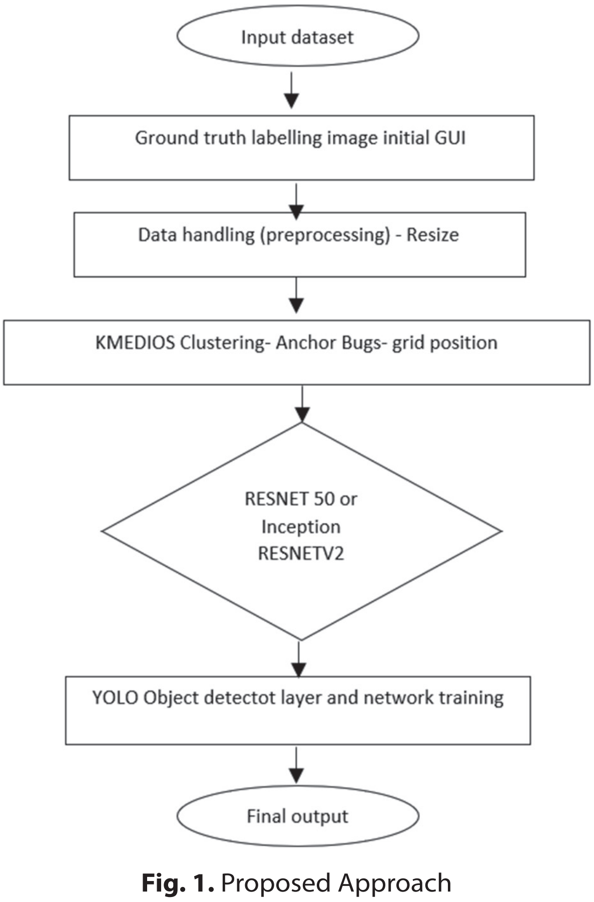
Input dataset
Ground truth labelling image initial GUI
Data handling (preprocessing) - Resize
KMEDIOS Clustering- Anchor Bugs- grid position
RESNET 50 or Inception RESNETV2
YOLO Object detectot layer and network training
Final output

Fig. 1. Proposed Approach

"Inception-ResNet-v2" is a convolutional neural network, and it is qualified on further than a million images from the ImageNet database. The network is 164 layers deep and can categorize images into 1000 object groups like the keyboard, mouse, pencil, and various animals; as a consequence, the network has learned rich feature depictions for a varied series of images and has an image input size of "299-by-299".

DenseNet-201 is a convolutional neural network that is 201 layers deep, and a pre-trained form of network proficient on more than a million images from the "ImageNet database" could be loaded. The pre-trained network could categorize images into "1000 object" types such as keyboard, mouse, pencil, and many animals. As an effect, the network has learned rich feature demonstrations for a comprehensive series of images.

The proposed model encompasses a neural network along with Res Net as a feature extractor and YOLO v2 for classification. The selection of these two systems is to progress the efficacy of the object discovery scheme by increasing accuracy and identifying even smaller objects exactly. A neural network is fashioned by using the ResNet network, which does the feature extraction manner. ResNet is deployed for its enhanced learning and accuracy with deeper networks. The deprivation problem occurs in deeper networks which saturates the accuracy of the model by restating the similar process again and again at greater levels. To overcome this problem, the ResNet model is used, which evades a step that is administered more than twice, so that accuracy may not saturate even with deeper networks. Mainly for classification, a fully connected network is transformed into YOLO v2 network.

YOLO is a single-stage entity-based recognition network, and it has a fast recognition speed. It is qualified by dense and consistent sampling above locations, scales, and an end-to-end flow. YOLO is a system that uses neural networks to afford instantaneous object detection. This system is standard for its speed and accuracy. It has been deployed in innumerable applications to perceive traffic signs, people, parking meters, and animals. In our project, we used to detect the Devanagari images.

Anchor Box:

The "YOLOv2" model fragments the input image into N×N grid cells, and all grid cell has the task of restricting an object if the midpoint of that particular object falls in a grid cell. If the midpoint of 2 objects corresponds with one another, the detection system could merely prefer any one of the objects. To resolve this problem, the conception of "Anchor Boxes" is used.

The YOLO algorithm is significant owing to subsequent details:

This system advances promptness of recognition as it can envisage objects in real time. YOLO is a prognostic system that offers precise effects with insignificant background errors. The system has excellent learning competencies that empower it to learn depictions of objects and relate them to object discovery.

3.4 CLASSIFICATION USING HYBRID CNN MODEL

The classification process is executed by using a hybrid CNN-RNN model, and a detailed analysis of the proposed system architecture is given and evaluates the methodology involved in the text recognition and classification process. And finally, the network is trained with the help of trainYOLOv2ObjectDetector with SGDM (stochastic gradient descent with momentum) optimizer and detects the output likewise. In a convolution neural network, the convolutional filters remove the noise images. Filters detect spatial patterns such as edges in an image by detecting the changes in intensity values of the image.

Convolutional layers are strong feature extractors in which the convolutional filters are capable of finding features of images. The function of max-pooling layers is to reduce the size of feature maps and solve over fitting problems. Based on these features value, the data is trained and tested. After this PYTHON is integrated in order to recognize the character which is given as input. Here Easy-OCR is used for recognizing the character of the object. Optical Character Recognition is a piece of technology that can identify text in digital images. Text in scanned documents and photos is frequently recognized using this technique.

An easy-to-use Python tool called Easy-OCR makes optical character recognition possible for computer vision engineers. Optical character recognition is by far the easiest to use when it comes to OCR, from that an Easy-OCR is the most popular method. Easy-OCR is a straightforward, lightweight library to use in comparison to others, as the name suggests. It supports a variety of languages. Additionally, it may be accomplished to perform better for particular use cases by adjusting various hyper-parameters. There is also another OCR which is Tesseract-OCR but it will not provide better result based on speed and accuracy when compared to Easy-OCR, so Easy-OCR is chosen. After Object Character Recognition it will be converted to ASCII key and then given to MATLAB GUI to finally identify the text from the natural image provided as input.

4. RESULTS

The analysis of simulation results and performance evaluation of the suggested methodology is described as follows.

1) Initial GUI

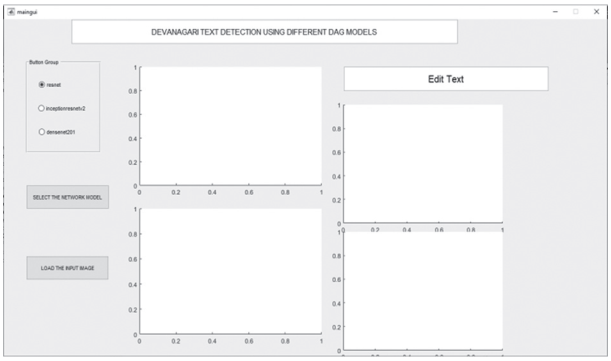

Fig. 2. Initial GUI

Figure 2 shows the initial GUI when the input image is taken. Figure 3 denotes the combined form of RESNET and YOLO output, where the text detected is shown separately. The detected text is shown separately by means of the EC-OCR coding model.

2) Combined RESNET and YOLO output

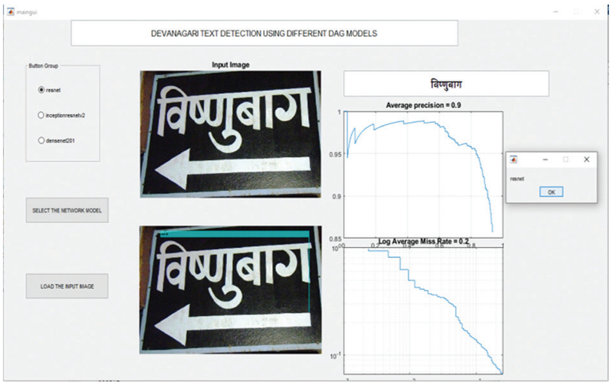

Fig. 3. Combined RESNET and YOLO output

3) Combined INCEPTIONRESNET and YOLO output

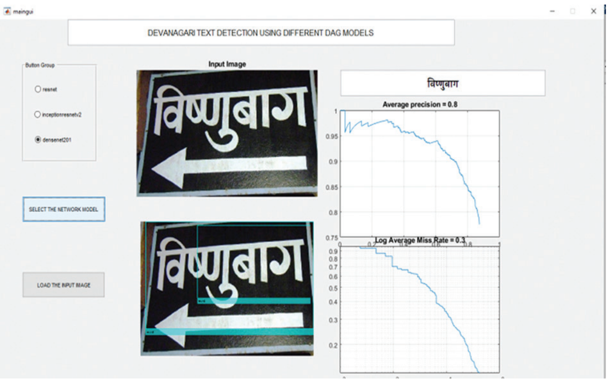

Fig. 4. Combined INCEPTIONRESNET and YOLO output

Figure 4 represents the combined form of INCEPTIONRESNET and YOLO output, where the text detected is shown separately in a white box. Here, the detected text is shown separately by means of the EC-OCR coding model.

4) Combined DENSENET201 and YOLO output

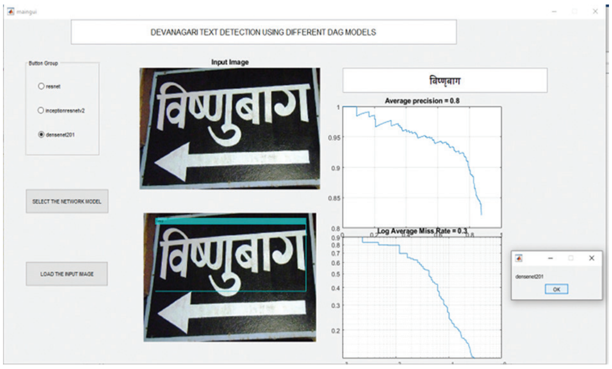

Fig. 5. Combined DENSENET201 and YOLO output

Fig. 5. represents the combined form of DENSENET 201 and YOLO output, where the text detected is shown separately in a white box. Here, the detected text is shown separately by means of the EC-OCR coding model and, the accuracy graph is extracted for each model.

Accuracy is a fraction of true positive occurrences to all positive examples of objects in the detector, centered on ground truth. For a multiclass indicator, average precision is a vector of average precision scores for all object classes.

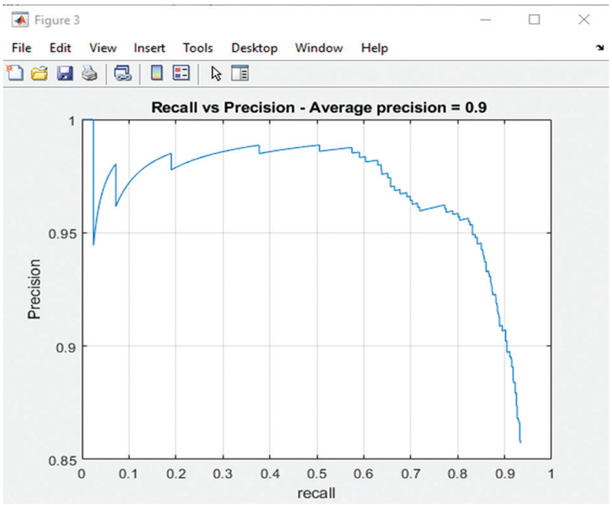

Fig. 6. Representation of the total word accurately detection rate in each and every images

Log miss rate, resumed as either one vector of numeric scalars or as cell arrangement. For the multiclass indicator system, FPPI and log miss rate were cell arrays, where every cell covers data points for every object class.

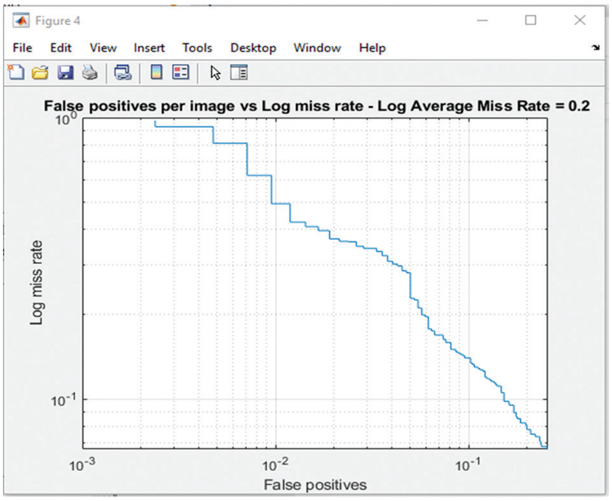

Fig. 7. Representation of the total word detection miss rate in each and every images

4.1. PERFORMANCE EVALUATION

The performance of the anticipated system is estimated with the aid of diverse performance metrics like "accuracy, precision, recall, f1 score, and support", and respective outcomes are discussed.

Table 1. Proposed model

S. No	Models name	Precision	Detection miss rate
1	Hybrid resnet - YOLO	0.907	0.152
2	Hybrid inceptionresnetv2 -YOLO	0.819	0.307
3	Hybrid densenet201 -YOLO	0.831	0.283

4.2 PERFORMANCE COMPARISON

A comparative analysis is provided by comparing the efficiency of the anticipated method with the methodology presented in the existing works. Maximum accuracy of 0.907 is attained from the hybrid resnet –YOLO approach, which shows the efficiency of the proposed method.

Results attained while considering the existing research are given as follows:

Model names resnet50 and Region Proposal Network (RPN) with feature pyramid network to generate bounding boxes for the input image are presented.

Table 2. Existing research work

S. No	Models name	Precision
1	Resnet50	0.769
2	ResNet-101	0.795
3	ResNet-152	0.82

5. CONCLUSION

The current research work focused on the scene-based text detection from any such natural images with the aid of comparative methodology, which includes a combination of YOLO and RESNET as the first case, combined INCEPTIONRESNET AND YOLO and combined DENSENET201 AND YOLO approach as a third case. In the first case, the model comprehends a neural network with Res Net as a feature extractor and YOLO v2 for the classification stage. The choice of these two algorithms is to advance the efficacy of the object detection system by increasing accuracy through detecting even smaller objects in an accurate manner. This approach could improve the speed of detection as it could foresee objects in real time. The research was implemented and tested using MATLAB 2019A, and the packages used were RESNET50, INCEPTIONRESNETV2, and DENSENET201. The effectiveness of the suggested technique was verified by relating accuracy value with the existing research works. Maximum accuracy of 0.907 is attained from the hybrid resnet –YOLO approach, which shows the efficiency of the proposed method.

6. REFERENCES

[1] H. Butt, M. R Raza, M. J. Ramzan, M.J. Ali, M. Haris, “Attention-based CNN-RNN Arabic text recognition from natural scene images", Forecasting, Vol. 3, No. 3, 2021, pp. 520-540.

[2] Y. S. Arafat, M. J. Iqbal, "Urdu-text detection and recognition in natural scene images using deep learning", IEEE Access, Vol. 8, 2020, pp. 96787-96803.

[3] A. N. Joseph, C. Junmin, R. Nersisson, V.G. Mahesh, Z. Zhuang, "Bilingual text detection from natural scene images using faster R-CNN and extended histogram of oriented gradients", Pattern Analysis and Applications, 2022, pp. 1-13.

[4] H. Lin, P. Yang, F. Zhang, "Review of scene text detection and recognition", Archives of Computational Methods in Engineering, Vol. 27, No. 2, 2020, pp. 433-454.

[5] L. T. Akin, T. Jaya, "Improved firefly algorithm-based optimized convolution neural network for scene character recognition Signal", Image and Video Processing, Vol. 15, No. 5, 2021, pp. 885-893.

[6] T. Khan, R. Sarkar, A. F. Mollah, "Deep learning approaches to scene text detection: a comprehensive review", Artificial Intelligence Review, Vol. 54, No. 5, 2021, pp. 3239-3298.

[7] T. Khan, A. F. Mollah, "AUTNT-A component level dataset for text non-text classification and benchmarking with novel script invariant feature descriptors and D-CNN", Multimedia Tools and Applications, Vol. 78, No. 22, 2019, pp. 32159-32186.

[8] Z. Zhong, L. Sun, Q. Huo, "Improved localization accuracy by LocNet for Faster R-CNN based text detection in natural scene images", Pattern Recognition, Vol. 96, 2019, p. 106986.

[9] X. Chen, L. Jin, Y. Zhu, C. Luo, T. Wang, "Text recognition in the wild: A survey", ACM Computing Surveys, Vol. 54, No. 2, 2021, pp. 1-35.

[10] J. Wang, H. Zhang, C. Zhang, W. Yang, L. Shao, J. Wang, "An effective scheme for generating an overview report over a very large corpus of documents", Proceedings of the ACM Symposium on Document Engineering, 2019, pp. 1-11.

[11] D. Kumar, S. Bhalekar, H. Disle, S. Gorule, K. Gotarane, "Automatic Text Summarization Using Local Scoring and Ranking", International Journal for Research Trends and Innovation, Vol. 4, No. 5, 2019.

[12] X. Zhang, X. Gho, C. Tian, "Text detection in natural scene images based on color prior guided MSER", Neurocomputing, Vol. 307, 2018, pp. 61-71.

[13] D. Van Nguyen, S. Lu, S. Tian, N. Ouarti, M. Mokhtari, "A pooling-based scene text proposal technique for scene text reading in the wild", Patten Recognition, Vol. 87, 2019, pp. 118-129.

[14] A. Sain, A. K. Bhunia, P. P. Roy, U. Pal, "Multi-oriented text detection and verification in video frames and scene images", Neurocomputing Vol. 1549, No. 275, 2020, p. 531.

[15] S. Wang, Y. Liu, Z. He, Y. Wang, Z. Tang, "A quadrilateral scene text detector with two-stage network architecture", Pattern Recognition, Vol. 102, 2020, p. 107230.

[16] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, L. Wang, "ABC-Net: real-time scene text spotting with adaptive Bezier curve network", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13-19 June 2020.

[17] S. H. Katper, A. R. Gilal, A. Alshanqiti, A. Waqas, A. Alsughayyir, J. & Jaafar, "Deep neural networks combined with STN for multi-oriented text detection and recognition", International Journal of Advanced Computer Science and Applications, Vol. 11, No. 4, 2020.

[18] H. Wang, S. Huang, L. Jin, "Focus on scene text using deep reinforcement learning", Proceedings of the 24th International Conference on Pattern Recognition, Beijing, China, 2018, pp. 3759-3765.

[19] S. Albahli, M. Nawaz, A. Javed, A. Irtaza, "An improved faster-RCNN model for handwritten character recognition", Arabian Journal for Science and Engineering, Vol. 46, No. 9, 2021, pp. 8509-8523.

[20] J. Diaz-Escobar, V. Kober, "Natural scene text detection and segmentation using phase-based regions and character retrieval", Mathematical Problems in Engineering, Vol. 2020, 2020.

[21] X. Liu, G. Meng, C. Pan, "Scene text detection and recognition with advances in deep learning: a sur-

vey", International Journal on Document Analysis and Recognition, Vol. 22, No. 2, 2019, pp. 143-162.

[22] M. Vidhyalakshmi, S. Sudha, "Text detection in natural images with hybrid stroke feature transform and high performance deep Convnet computing", Concurrency and Computation: Practice and Experience, Vol. 33, No. 3, 2021, p. e5271.

[23] Y. Tang, X. Wu, "Scene text detection using superpixel-based stroke feature transform and deep learning based region classification", IEEE Transactions on Multimedia, Vol. 20, No. 9, 2018, pp. 2276-2288.

[24] F. Cong, W. Hu, Q. Huo, L. Guo, "A comparative study of attention-based encoder-decoder approaches to natural scene text recognition", Proceedings of the International Conference on Document Analysis and Recognition, 2019, pp. 916-921.

[25] D. Pandey, B. K. Pandey, S. Wairya, "Hybrid deep neural network with adaptive galactic swarm optimization for text extraction from scene images", Soft Computing, Vol. 25, No. 2, 2021, pp. 1563-1580.

[26] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, X. Bai, "Textfield: Learning a deep direction field for irregular scene text detection", IEEE Transactions on Image Processing, Vol. 28, No. 11, 2019, pp. 5566-5579.

[27] L. T. Sherly, T. Jaya, "An efficient indoor scene character recognition using Bayesian interactive search algorithm-based adaboost-CNN classifier", Neural Computing and Applications, Vol. 33, No. 22, 2021, pp. 15345-15356.

[28] F. Jiang, Z. Hao, X. Liu, "Deep scene text detection with connected component proposals", arXiv:1708.05133, 2017.

[29] A. Ali, M. Pickering, "A hybrid deep neural network for Urdu text recognition in natural images", Proceedings of the IEEE 4th International Conference on Image, Vision and Computing, 2019, pp. 321-325.

[30] N. Gupta, A. S. Jalal, "Text or non-text image classification using fully convolution network (FCN)", Proceedings of the International Conference on Contemporary Computing and Applications, Lucknow, India, 2020, pp. 150-153.

[31] X. Wang, X. Feng, Z. Xia, "Scene video text tracking based on hybrid deep text detection and layout constraint", Neurocomputing, Vol. 363, 2019, pp. 223-235.