1. **Introduction**

Text, as one of the most influential inventions of humanity, has played an important role in human life, so far from ancient times. The rich and precise information embodied in text is very useful in a wide range of vision-based application [1], therefore text detection and recognition in natural scenes have become important, active research topics in computer vision and document analysis. Especially in recent years, the community has seen a surge of research efforts and substantial progresses in these fields, though a variety of challenges (e.g. noise, blur, distortion, occlusion and variation) still remain.

The work on text recognition has begun many years ago but text recognition in images and video is an active research area so that a computer application with the ability to automatically read the text from images and video can be developed.

Optical Character Recognition (OCR) is a traditional technique of recognizing the text from the documents, and the accuracy of this technique is good in the scanned documents; but when the same technique is applied to scene images, the performance of this method was not up to the mark. Recognition of text from the scene needs special features because the character present in the scene may differ in size, shape, color, writing style, orientation, aspect ratio, quality of the image due to different lighting conditions, and blurred and complex background. These are the various challenges of text detection and text recognition [2].

The reason for text recognition is due to easy availability of large amount of digital information from videos and scene images which contain very useful information like street name, location address, traffic warning etc. Therefore, text extraction and recognition from this information are very effective and important in different text based applications.

The number of real time applications on text detection and recognition is increasing rapidly. The text information embedded in scene images suffers from the background complexity, font, font size, color, orientations and variant alignment when compared with the normal text. In scene images text are often a brief extract written in different fonts and languages.

Applications of text localization and recognition in real world images range from automatic tagging of image databases based on their textual content (e.g. Flickr or Google Images), assisting the visually impaired to reading labels in business map applications (e.g. Google Street View)

Some of the applications of the text recognition system include banking, legal, health care, other fields like education, finance and government agencies [3].

1. **Literature Review**

Asghar Ali Candio et.Al (2022), proposes the cursive text recognition in natural scene images using which is segmentation free using Deep CRNN. This work is based on the three components: i) A deep Convolutional Neural Network (CNN) to extract and encode the features with shortcut connections, ii) A Recurrent Neural Network (RNN) for decoding of the convolutional features, iii) A Connectionist Temporal Classification (CTC) for mapping of the predicted sequences and the target labels. Text recognition accuracy is increased by exploring VGG-16, VGG-19, ResNet-18 and ResNet-34, which are deeper CNN architectures. When compared to all the networks the deep CRNN network, it works well and shows appreciable results than other architectures in recognition of cursive texts [1].

Hamam Mokayed et.Al (2021), Proposes a method which uses post-processing for improving the results of text detection and recognition. Combination of SVM and Gaussian distribution is used to determine weights, which represents text here. Finally the weights are multiplied with the features to detect the defects. The bounding boxes are redrawn, where the text from natural scene images are detected without defects. MSRA-TD-500 and SVT are the popular standard datasets used for this research work [4].

Shirvale S.S et.Al (2020)**,** discusses the existing text detection methods, low-level feature-based techniques, high level feature based techniques and text region verification techniques , wherein finds that SVM is highly recommended classifier for text and non-text classification and CNN for region classification. And also summarizes that SWT and MSER are more stable for feature-based techniques. SVM and DNN are efficient and finds great success [5].

It consists of two main stages, MSER computation and SWT. Edges in the images are enhanced by canny edge detector. The combination of SWT image and geometric information of the images worked well in improving the non-text regions in the image. Experimental results showed that by using IITR datasets has ended up with encouraging results. Due to the presence of the similarity in stroke width of text and Non-text portions error has occurred [6].

Xuejian Rong et.Al (2020), discusses the extracting of texts and unambiguous text information from natural images. DTLN and CRTR model are used to localize and detect the text. To extract the text information from images the combination of vision based localization and language-based contextual description is used, by this approach the experimental results are appreciable [7].

Jian Ye et.Al, (2020) TextFuseNet exploits the use of richer features for text detection based on three levels of features. Character, word and global-level features. A novel framework TextFuseNet to detect arbitrary shape text by investigating on three levels of features, i.e., character-, word- and global-level features. Different level features are fully and finely explored to learn richer fused features, which are beneficial to text detection. Experimental results show that TextFuseNet achieves state-of-the-art performance in detecting text with arbitrary shape [8].

Randheer Bagi et. Al (2020), Proposed a light-weight scene text spotter that can address the issue of cluttered environment of scene images. It is an end-to-end trainable deep neural network that uses local part information, global structural features, and context cue information of oriented region proposals for spotting text instances. It helps to localize in scene images with background clutters, where partially occluded text parts, truncation artifacts, and perspective distortions are present. They mitigated the problem of misclassification caused by inter-class interference by exploring inter-class separability and intra-class compactness. Also incorporated multi-language character segmentation and word-level recognition in a light-weight recognition module and have used six publicly available benchmark datasets in different smart devices to illustrate the efficacy of the network [9].

Syed Yasser Arafa et. Al(2020), Proposed a methodology that covers detection, orientation prediction, and recognition of Urdu ligatures in outdoor images. As a first step, the custom Faster RCNN algorithm has been used in conjunction with well-known CNNs like Squeezenet, Googlenet, Resnet18, and Resnet50 for detection and localization purposes for images of size 320 \_ 240 pixels. For ligature Orientation prediction, a custom Regression Residual Neural Network (RRNN) is trained/tested on datasets containing randomly oriented ligatures. Recognition of ligatures was done using Two Stream Deep Neural Network (TSDNN). Urdu-text using average-precision(AP). Resnet50 features based Faster RCNN was found to be the winner detector with AP of.98. While Squeeznet, Googlenet, Resnet18 based detectors had testing AP of.65, .88, and .87 respectively. RRNN achieved and accuracy of 79% and 99% for 4k and 51K images respectively. Similarly, for characters classification in ligatures, TSDNN attained a partial sequence recognition rate of 94.90% and 95.20% for4k and 51K images respectively. Similarly, a partial sequence recognition rate of 76.60% attained for real world-images [10].

Asghar Ali Chandio et. Al,2020 proposed a multi-scale feature aggregation (MSFA) and a multi-level feature fusion (MLFF) network architecture to recognize isolated Urdu characters in natural images. The network first aggregates multi-scale features of the convolutional layers by up-sampling and addition operations and then combines them with the high-level features. Finally, the outputs of the MSFA and MLFF networks are fused together to create more robust and powerful features. A comprehensive dataset of segmented Urdu characters is developed for the evaluation of the proposed network models. Synthetic text on the patches of images with real natural scene backgrounds is generated to increase the samples of infrequently used characters. The proposed model is evaluated on the Chars74K and ICDAR03 datasets. To validate the proposed model on the new Urdu character image dataset, compared its performance with the histogram of oriented gradients (HoG) method. The experimental results show that the aggregation of multi-scale and multilevel features and their fusion is more effective, and outperforms other methods on the Urdu character image and Chars74K datasets [11].

Shangbang Long et. Al 2019, With the rise and development of deep learning, computer vision has been tremendously transformed and reshaped. As an important research area in computer vision, scene text detection and recognition has been inevitably influenced by this wave of revolution, consequentially entering the era of deep learning. In recent years, the community has witnessed substantial advancements mindset, methodology and performance. This survey is aimed at summarizing and analyzing the major changes and significant progresses of scene text detection and recognition in the deep learning era. This article is devoted to: (1) introduce new insights and ideas; (2) highlight recent techniques and benchmarks; (3) look ahead into future trends. Specifically, will emphasize the dramatic differences brought by deep learning and the grand challenges still remained. It is expected that this review paper would serve as a reference book for researchers in this field. Related resources are also collected and compiled in our Github repository [12].

Jinjin Zhang et.Al (2019), aims to propose a universal framework for STR with the combination of object detection and attention based language model. The competition results in ICDAR2019 including LSVT, ArT, ReCTS demonstrate the effectiveness and robustness of our algorithm. In addition, it is convenient to extend the current algorithm to stateof- the-art method such as replacing detection or attention mechanism with better architectures, such as fast and accurate detector EfficientDet or Transformer based self-attention mechanism. Moreover, it is feasible to synthesize text images from natural language process corpus for data augmentation and helpful for attention-based language model. In future, it is imperative to build an end-to-end differentiable STR algorithm with both detection and recognition requires GPUs with large memory like V100. It is essential to eliminate the detection failure case with semantic information. Language model based methods like BERT should be beneficial in our framework which takes the context of whole sentences into consideration instead of the previous word merely. Furthermore, visual based question answering or semantic analysis modules can be integrated with the framework for text based high-level semantic applications [13].

Minghui Liao et.Al (2019), Mask TextSpotter, a novel framework of end-to-end text recognition in the wild. Different from the previous text spotters that consider learning-based text recognition as a one-dimensional sequence prediction problem, the proposed method is very easy to train and able to read irregular text, profiting from its two-dimensional representation for both detection and recognition. The state-of- the-art results achieved by Mask TextSpotter in the tasks of scene text detection, scene recognition, and end-to-end text recognition on the standard benchmarks including horizontal text, oriented text, and curved text, validate its generality and effectiveness in reading scene text. In the future, we would like to improve the efficiency of Mask TextSpotter, especially exploring to replace the detection stage with more elegant detection method, which is the most time-consuming part of the proposed method [14].

Shangbang Long et. Al (2019), with the rise and development of deep learning, computer vision has been tremendously transformed and reshaped. As an important research area in computer vision, scene text detection and recognition has been inevitably influenced by this wave of revolution, consequentially entering the era of deep learning. In recent years, the community has witnessed substantial advancements mindset, methodology and performance. This survey is aimed at summarizing and analyzing the major changes and significant progresses of scene text detection and recognition in the deep learning era. This article is devoted to: (1) introduce new insights and ideas; (2) highlight recent techniques and benchmarks; (3) look ahead into future trends. Specifically, will emphasize the dramatic differences brought by deep learning and the grand challenges still remained [15] .

Minghui Liao, et. Al (2018) presents an end-to-end trainable fast scene text detector, named TextBoxes++, which detects arbitrary-oriented scene text with both accuracy and efficiency. Text detection is done directly by predicting word bounding boxes with quadrilaterals via a single neural network that is end-to-end trainable. Here the rectangular box representation is replaced by quadrilateral representation to achieve better receptive field that covers text regions which are usually long. The final outputs are the non-maximum suppression outputs on all boxes. A single forward pass in the network densely detects multi-scale scene text boxes all over the image. Further they combined TextBoxes++ with CRNN an open source text recognition module. The recognizer not only produces extra recognition outputs but also regularizes text detection with its semantic-level awareness, thus further boosts the accuracy of word spotting considerably. The combination of TextBoxes++ and CRNN yields state-of-the art performance on both word spotting and end-to-end text recognition tasks, which appears to be a simple yet effective solution to robust text reading in the wild. In all experiments, TextBoxes++ has achieved state-of-the-art performance with high efficiency for both horizontal text datasets and oriented text datasets. Future scope of this work is to, investigate the common failure cases (*e.g.*, large character spacing and vertical text) faced by almost all state-of-the-art text detectors [16].

Baoguang Shi, et.Al (2018) introduces ASTER, which stands for **A**ttentional **S**cene **TE**xt Recognizer with Flexible **R**ectification, for scene text recognition. The rectification network adaptively transforms an input image into a new one, rectifying the text in it. It is powered by a flexible Thin-Plate Spline transformation which handles a variety of text irregularities and is trained without human annotations. The recognition network is an attentional sequence-to-sequence model that predicts a character sequence directly from the rectified image. End-to-end text recognition is addressed in a two-stage manner, meaning the detection is performed in another network using separate features. They have demonstrated that, ASTER performs implicit text detection. But this detection capability is limited to a small range near the target text. Expanding this range to the full image would yield a one-stage, end-to-end recognition system, and is a direction worthy of further investigation [16].

Shu Tian, et.Al (2018) proposes a generic Bayesian-based framework of Tracking based Text Detection And Recognition (T2DAR) from web videos for embedded captions, which is composed of three major components, i.e., text tracking, tracking based text detection, and tracking based text recognition. A few of recent methods, which pay attention to text tracking using multiple frames, however, do not effectively mine the relations among text detection, tracking and recognition. In this unified framework, text tracking is first conducted by tracking-by-detection. Tracking trajectories are then revised and refined with detection or recognition results. Text detection or recognition is finally improved with multi-frame integration. Moreover, a challenging video text (embedded caption text) database (USTB-VidTEXT) is constructed and publicly available. A variety of experiments on this dataset verify that this proposed approach largely improves the performance of text detection and recognition from web videos [17].

Savita Choudhary (2018) proposes an approach for detection of text area from natural scene images using Maximally Stable Extremal Regions (MSER) and recognizing the text using a self-trained Neural Network. Canny edge detector is used to locate the smaller areas that may more likely contain text. This approach successfully predicts the characters with very less error rate. Due to this new approach no character is left out or dropped in the prediction process. Practical use of this system can be in mapping work where addresses and landmarks need to be read from images. The banners and posters on roads can be perfectly read by this system. It can also read graffiti and sign boards without much difficulty. Blurred images still create trouble for text recognition systems with no solid solution. Some more feature detection methods can be added to include extremely blurred text present in the images [18].

Ong Yi Ling (2018), discusses the automatic recognition of vertical texts in natural scene images, this work consists of two major process i.e, Text localization and segmentation and Text recognition. For detecting the text gray scaling is followed by MSER. To obtain the binary image Binarization and dilation is performed. False positives are eliminated by connected component segmentation. By using Optical Character Recognition (OCR) the text part from the natural scene images are recognized. After the recognition of true texts from natural images, identification of vertical texts as horizontal stacked vertical text, Top-to-bottom vertical text is carried out by this model. 90degrees rotation of the detected text and bottom-to-top vertical texts rotation at -90 degrees id performed. Finally the oriented text undergoes OCR , detection and recognition results are appreciable [19].

Ling-qun zuo et. Al, this method converts the natural text recognition into a sequence mark by combining the connection time classification (CTC) and attention mechanism under the encoder and decoder framework, in order to overcome the problem of character segmentation, using the correlation between image and text sequence. First of all, a convolutional neural network (CNN) is used to generate an ordered feature sequence from the entire word image. Then, the generated feature sequence is feature-coded using the bidirectional long short-term memory (Bi-LSTM) network. Finally, an integrated module of the CTC and attention mechanism is designed to decode and output the text sequence. The experimental results show that the CTC-Attention joint mechanism significantly improves the recognition performance of the model and has an advantage in identifying scene text images. Dataset used are Synth90k dataset containing 9 million synthetic scene text images as a training set, using SVT dataset , IIIT5K dataset , ICDAR 2003 dataset , ICDAR 2013 dataset was tested as a test set [20].

1. **Objectives**
2. Analyze the literature review on text detection and recognition from natural images, which can explain the methods used and identify the features of each method. Review and compare traditional and current deep learning-based algorithms for detecting and recognizing text in natural images.
3. Identify the most significant challenges and applications.
4. Develop algorithms to detect and localize the text in natural images and videos.
5. To experiment a new set of features/ feature-combination for efficient classification.
6. Develop algorithms which can separate text and non-text regions.
7. Develop algorithms to efficiently recognize the detected and localized text.
8. To improvise accuracy rates recorded in literature.

1. **TEXT RECOGNITION SYSTEM FROM NATURAL SCENE IMAGES**

Descriptors

Input Images

(Training set)

Training Phase

Apply Feature Extraction

MSER GLCM, HoG, LBP

SVM

Candidate Regions

Text Regions

Build

Classifers

Non-Text Regions

Remove Non Text Regions

Character Grouping

Bounded texts by boxes

Classifiers

(From training set)

Word

Or

Text lines

Map text regions on original images

Input Images

Testing Set

Testing Phase

Apply Feature Extraction

MSER MSER GLCM, HoG, LBP

Candidate

Regions

1. **Work carried out so far**
2. Analyze the literature review on text detection and recognition from natural images, which can explain the method used and identify the features of each method. Review and compare traditional and current deep learning-based algorithms for detecting and recognizing text in natural images.
3. Identify the most significant challenges and applications.
4. **Further work to be carried out**
5. Develop algorithms to detect and localize the text in natural images and videos.
6. To experiment a new set of features/ feature-combination for efficient classification.
7. Develop algorithms which can separate text and non-text regions.
8. Develop algorithms to efficiently recognize the detected and localized text.
9. To improvise accuracy rates recorded in literature.

**.**